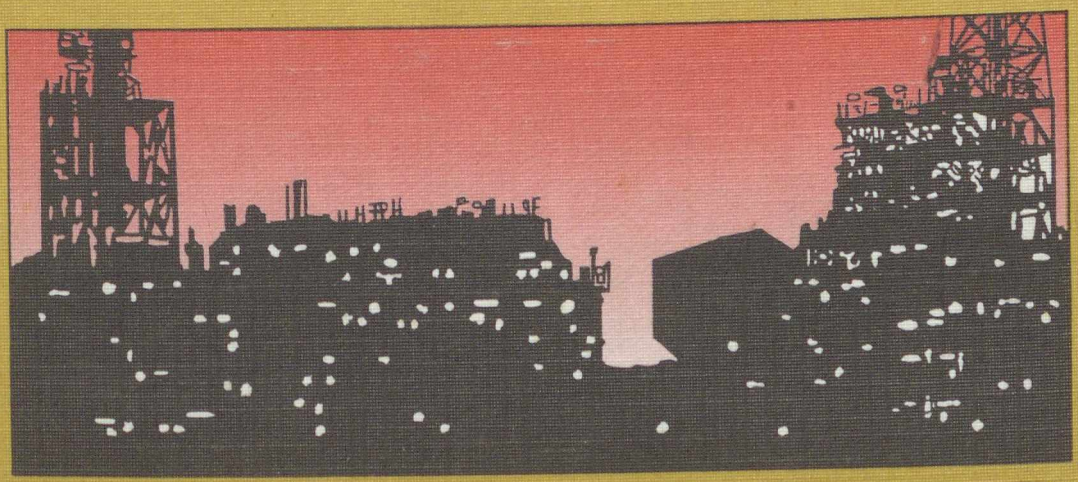
# STUDIES IN COMPUTER-AIDED MODELLING, DESIGN AND OPERATION

Part B

## Systems

I. PALLAI
G. E. VERESS
(Editors)



AKADÉMIAI KIADÓ · BUDAPEST

# STUDIES IN COMPUTER-AIDED MODELLING, DESIGN AND OPERATION

## PART B

## Systems

3362

# STUDIES IN COMPUTER-AIDED MODELLING, DESIGN AND OPERATION

## PART B

## Systems

**Edited by**

**I. PALLAI**

*Institute for Science Management and Informatics, Budapest, Hungary*

**and**

**G. E. VERESS**

*National Committee for Technological Development, Budapest, Hungary*

MTAK 0 00002 98129 1

**AKADÉMIAI KIADÓ, BUDAPEST 1992**

507954

Translated and revised by

GY. JALSOVSZKY

10029

This book is published as Number 5B in the series
Computer-Aided Chemical Engineering
Joint edition with Elsevier
Amsterdam—London—New York—Tokyo

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

© I. Pallai—Z. Fonyó, 1992

© English translation — Gy. Jalsovszky, 1992

ISBN 963 05 6084 4 (Vols A—B)

ISBN 963 05 6086 0 (Vol B)

All rights reserved. No part of this book may be reproduced by any means, or transmitted, or translated into machine language without the written permission of the copyright owner.

Printed by Akadémiai Kiadó és Nyomda Vállalat, Budapest, Hungary

M. TUD. AKADÉMIA KÖNYVTÁRA
Könyvleltár. 9564./19. 92. sz.

# LIST OF CONTRIBUTORS

*Hajdú, H.*
Department of Chemical Engineering
Technical University of Budapest
Műegyetem rkp. 3/9
H–1111 Budapest, Hungary

*Mizsey, P.*
Department of Chemical Engineering
Technical University of Budapest
Műegyetem rkp. 3/9
H–1111 Budapest, Hungary

*Németh, A.* Prof.
Department of Chemical Technology
Technical University of Budapest
Műegyetem rkp. 3/9
H–1111 Budapest, Hungary

*Soltész, P.*
Department of Chemical Technology
Technical University of Budapest
Műegyetem rkp. 3/9
H–1111 Budapest, Hungary

*Tátrai, F.*
Department of Chemical Technology
Technical University of Budapest
Műegyetem rkp. 3/9
H–1111 Budapest, Hungary

*Vajda, S.*
Laboratory for Chemical Cybernetics
Eötvös Loránd University
Pázmány P. sétány 2
H–1117 Budapest, Hungary

*Valkó, P.*
Laboratory for Chemical Cybernetics
Eötvös Loránd University
Pázmány P. sétány 2
H–1117 Budapest, Hungary

*Veress, G. E.* Prof.
National Committee for Technological Development
Martinelli tér 8
H–1052 Budapest, Hungary

# CONTENTS


FOREWORD . . . . . XI

PREFACE . . . . . XIII

1 THE SYSTEMS APPROACH TO CHEMICAL ENGINEERING (A. NÉMETH) 1

1.1 Introduction to the systems approach to chemical engineering . . . . . 1
1.2 Elements of chemical technological systems and the connections between them 10
1.2.1 The element: definition and types . . . . . 10
1.2.2 Connections in chemical technology . . . . . 17
1.3 Hierarchy of a chemical technological system . . . . . 23
1.4 References . . . . . 27
1.5 Literature suggested for further reading . . . . . 27

2 MODELLING AND REGRESSION (S. VAJDA and P. VALKÓ) . . . . . 28

2.1 Modelling and its uses . . . . . 28
2.1.1 Why should we want a model? . . . . . 29
2.1.2 What kind of model? . . . . . 31
2.2 Parameter estimation . . . . . 32
2.2.1 Steps of parameter estimation . . . . . 33
2.2.2 Linear algebraic model: linear regression . . . . . 35
2.2.3 Nonlinear algebraic model: the Gauss–Marquardt algorithm. . . . . 41
2.2.4 Indirect least squares method of fitting dynamic models. . . . . 51
2.2.5 Direct least squares method of fitting dynamic models . . . . . 61
2.3 Model reduction . . . . . 77
2.3.1 Principal component analysis of nonlinear regression models . . . . . 77
2.3.2 Nearly-singular estimation problems and model reduction in chemical kinetics . . . . . 82
2.4 References . . . . . 94
Appendix (at the end of the book, pp. 493—518)

3 ANALYSIS OF CHEMICAL ENGINEERING NETWORKS
(F. TÁTRAI and P. SOLTÉSZ) . . . . . 96

3.1 Introduction . . . . . 96
3.2 Structure analysis of chemical engineering systems. . . . . 96
3.2.1 Basic notions . . . . . 98

3.2.2 Partitioning and related questions . . . 102
3.2.3 Tearing . . . 119
3.3 Flowsheeting . . . 124
3.3.1 Degrees of freedom analysis . . . 125
3.3.2 How to define the flowsheeting problem . . . 147
3.3.3 Iteration methods . . . 151
3.3.4 Approaches to flowsheeting . . . 159
3.3.5 Some typical flowsheeting problems and their solutions . . . 167
3.4 References . . . 186
3.5 Literature suggested for further reading . . . 188
3.5.1 Flowsheeting systems and related topics . . . 188
3.5.2 Structure analysis and iteration procedures . . . 188

**4 ECONOMIC ANALYSIS OF CHEMICAL ENGINEERING SYSTEMS**
(F. TÁTRAI and P. SOLTÉSZ) . . . 190

4.1 Profitability analysis . . . 190
4.1.1 Definition of terms . . . 191
4.1.2 Measures of profitability . . . 201
4.1.3 Special aspects of profitability calculations . . . 217
4.2 Capital cost estimation . . . 230
4.2.1 Estimation of fixed capital investment . . . 231
4.2.2 Working capital estimation . . . 253
4.3 Manufacturing cost estimation . . . 255
4.4 Capital cost estimation by CAPCOS . . . 264
4.4.1 The program CAPCOS . . . 264
4.4.2 A complex example to illustrate capital cost estimation by CAPCOS . . . 265
4.5 References . . . 277
Appendix (at the end of the book, pp. 519—565)

**5 PRINCIPLES OF CHEMICAL PROCESS CONTROL**
(H. HAJDÚ and P. MIZSEY) . . . 280

5.1 Introduction . . . 280
5.2 Control dynamics . . . 285
5.2.1 Basic concepts of feedback control . . . 285
5.2.2 Control of a dead time process . . . 292
5.2.3 Control of capacity processes . . . 303
5.2.4 The *PI*-controller . . . 306
5.2.5 *PD*- and *PID*-controllers . . . 310
5.2.6 Tuning the controller. Cycling method . . . 313
5.3 Linear control theory . . . 314
5.3.1 Classification of processes . . . 315
5.3.2 Typical input signals . . . 316
5.3.3 Linearization . . . 317
5.3.4 The transfer function method . . . 319
5.3.5 The frequency function method . . . 335
5.3.6 Simple linear systems . . . 341

5.3.7 Control loop behaviour . . . . . 361
5.3.8 Controller tuning . . . . . 372
5.4 The state space method . . . . . 379
5.4.1 The state of a system . . . . . 379
5.4.2 State equations of multivariable systems . . . . . 380
5.4.3 State equations of higher-order elements . . . . . 382
5.4.4 The solution of the state equations . . . . . 386
5.5 Notation . . . . . 386
5.6 Literature . . . . . 387
Appendix (at the end of the book, pp. 567—598)

**6 MEASURING SUBSYSTEMS FOR THE PROCESS CONTROL OF CHEMICAL ENGINEERING SYSTEMS** (G. E. VERESS) . . . . . 388

6.1 Chemical measuring systems . . . . . 388
6.1.1 Definition of chemical measuring systems . . . . . 388
6.1.2 Mathematical modelling of chemical measuring systems . . . . . 393
6.2 Signal processing methods . . . . . 396
6.2.1 Signal transformations . . . . . 396
6.2.2 Fourier transformation . . . . . 401
6.2.3 Walsh and Haar transformations . . . . . 405
6.2.4 Convolution and correlation . . . . . 410
6.2.5 Curve fitting . . . . . 413
6.2.6 Factor analysis . . . . . 418
6.2.7 Processing of density function-like signals . . . . . 426
6.3 Signal interpretation methods . . . . . 435
6.3.1 Qualitative analysis by supervised classification . . . . . 435
6.3.2 Quantitative analysis by linear analytical investigation . . . . . 441
6.4 Performance characteristics of measuring systems . . . . . 445
6.4.1 Types of performance characteristics . . . . . 446
6.4.2 Statistical characteristics . . . . . 448
6.4.3 Information theoretical characteristics . . . . . 462
6.4.4 Decision theoretical characteristics . . . . . 471
6.4.5 Characteristics for credibility . . . . . 479
6.4.6 Selection of analytical chemical measuring techniques for quality control 483
6.5 Literature . . . . . 488
6.5.1 Chemical measuring systems . . . . . 488
6.5.2 Signal processing methods . . . . . 489
6.5.3 Signal interpretation methods . . . . . 489
6.5.4 Characterization of chemical measuring systems . . . . . 489
Appendix (at the end of the book, pp. 599—621)
Appendices (to Chapters 2, 4, 5 and 6) . . . . . 491
Subject index . . . . . 623

# FOREWORD

The use of computer-aided simulation and design has seen rapid expansion during the past two decades due to the availability of high speed computers. In the chemical process industry large, realistic, nonlinear problems are now being solved routinely via simulation. The increasing use of computer-aided design techniques has broadened the usefulness of the scientific approach to engineering. Since design relies upon a scientific rather than empirical approach to engineering, the competency in process simulation requires

— a sound understanding of engineering fundamentals; i. e. the process cannot be viewed as a black box,
— the ability to set mathematical relations which adequately describe the significant process behaviour,
— the capability of rapid and inexpensive computational solutions.

The text was compiled as a result of the teaching activity of the authors during the last decade to follow in the undergraduate course as a senior course to link the theory in standard courses to industrial practice. Thus the book contains computer-aided studies in modelling based on the available software. The set of programs was operations based on the available software and developed during the period 1982–1987 within the framework of the CHEMISYS project in order to compile chemical engineering programs and include them in a unified system. This may serve as a suitable computer background for a modern approach to chemical engineering training.

Several CHEMISYS programs were incorporated in the international EURECHA (European Committee for Computers in Chemical Engineering Education) project. These programs are referred to in the corresponding chapters, and their manuals are appended. The programs are available from EURECHA Secretariat, ETH Zürich, Technisch-Chemisches Laboratorium, ETH Zentrum, CH–8092 Zürich, Switzerland.

The book is written both for engineers in industry who are interested in applying computer techniques to the design and analysis of plants, and for senior chemical engineering students for use in design and simulation classes.

Many people and organizations assist in a project of this kind. We are especially grateful to Dr. L. M. Rose, who inspired this book and encouraged us to compile it. We wish to acknowledge the aid for the CHEMISYS project, of the Hungarian Ministry of Culture and Education in developing the educational materials and set of programs. We express our deep appreciation to Akadémiai Kiadó and Elsevier Science Publishers. Without their encouragement and support, this textbook could not have been undertaken.

*Iván Pallai*

# PREFACE

The complex character of chemical engineering demands a rich description of chemical engineering problems. Besides the mathematical modelling of chemical engineering systems, which is the topic of Part A of this work, a number of important questions must be considered: how should we model the chemical engineering systems? What kind of model identification method should we use? How should we characterize the economic character of the chemical engineering systems? How should we control them? How should we characterize the measuring subsystems? And so on.

This volume deals with several important aspects of modern methods for solving chemical engineering problems, using original, new points of view.

The aim of Chapter 1 is to emphasize the importance of the systems theoretical characterization of chemical engineering problems. The mathematical modelling of chemical engineering systems must be based on their systems theoretical description. To model the functional behaviour of a system the application of Kalman's or Zadeh's idea is extremely useful. The structure of a system can be described by graph theoretical tools.

Based on the systems theoretical description, the functional model of chemical engineering systems must be identified. Chapter 2 deals with the very modern aspects of model identification and parameter estimation using the new methods of model reduction. These topics are very important not only in chemical engineering but also in all fields of mathematical modelling.

Chapter 3 discusses the structural analysis of chemical engineering networks. Applying graph theoretical and matrix representations, this chapter deals with the solution of flowsheeting problems.

Though chemical engineering systems are thermodynamic systems, the discriminating feature of chemical engineering systems is their economic character. The subject of Chapter 4 is the economic analysis of chemical engineering systems, estimating the profitability and the different types of costs.

Chapter 5 summarizes the basic ideas of process control methods which are useful in the practice of chemical engineering. The didactical character of this chapter is valuable not only for chemical engineers but also for other specialists in process control applications.

Chapter 6, which describes measuring systems, bridges the gap between chemical

engineers and analytical chemists. The unified characterization of the various measuring systems discussed here is based on their systems theoretical description. This chapter emphasizes the importance of the performance characteristics of chemical measuring systems in selecting the most suitable measuring system for solving a chemical engineering control problem.

Based on the above methodological perspectives this volume contains some useful computer programs developed within the framework of the CHEMISYS project in order to compile chemical engineering programs and include them in a unified system serving as a suitable background for a modern approach to chemical engineering education. Program REPROCHE is a very useful tool for parameter estimation, CAPCOS is a program for calculating capital costs, TACS is a valuable program system for teaching control methods, and program STATCHAR calculates the statistical characteristics of measured data. In addition some other programs are also mentioned and examples and case-studies are given.

This volume is written for those interested in chemical engineering problems; thus for engineers in industry, for students and for teachers in classes.

Although, with the rapid development of computational techniques, the numerical methods and computer programs presented in this volume may become obsolescent, we hope that the point of view of the book will be of lasting value.

*Gábor E. Veress*

# 1 THE SYSTEMS APPROACH TO CHEMICAL ENGINEERING

by

A. NÉMETH

## 1.1 Introduction to the systems approach to chemical engineering

The process used in the chemical industry usually involve several pieces of apparatus, linked to one another, which perform physico-chemical changes on the raw material or on intermediates. The link between the pieces of apparatus is provided by flows of raw material, fine product, intermediates, by-products, auxiliary products, and by energy and information flows. Consequently, the preparation of a given product takes place in a complex apparatus interconnected by various flows in a predetermined sequence of chemical and technological transformations, the working as a uniform entity. A process of the chemical industry may therefore be regarded as a system.

In order to illustrate the systematic nature of such processes, let us study a typical operation, using the example shown in *Fig. 1.1.* The components of the process (chemical engineering processes, unit operations) are the reactor responsible for the chemical transformation, the absorber for the separation of the product(s) leaving the reactor, and the distillation equipment. If the operation of this relatively simple part of a real process is studied, it is easy to see that a lower chemical conversion coupled with a better absorption and a more perfect distillation may yield the same product(s) as a higher conversion with less perfect separation.

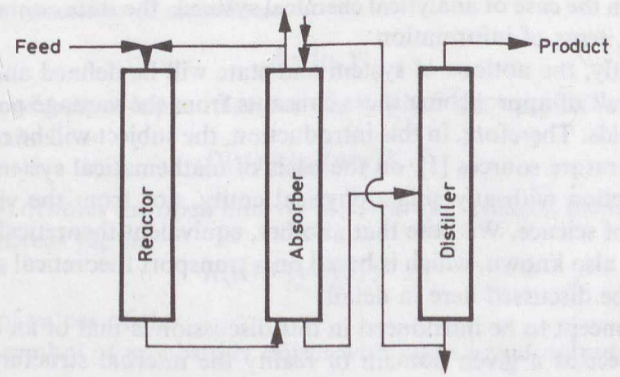



Fig. 1.1. A characteristic unit operation of chemical industry

The choice between the two possibilities with respect to the entire process, i.e. with respect to the economic result, is determined by the analysis of investment and operating costs. The first variant requires a smaller reactor and a more complex separation device, whereas the latter requires the opposite. This example shows that the optimum operation of a chemical industrial process is not simply a combination of the individual optima of its component processes, but is rather a property of the entire process, which may be different from the component parts.

Consequently, the processes of chemical industry should be considered from a systems perspective. A system, in a completely general sense, is one of the most widespread notions in human cognition. This is also shown by its definition, according to which a system is a set of objects linked to one another by mutual relationships. We start this discussion deliberately from the most general definition in order to be able to derive from it the special properties of chemical technological systems including networks, process control and chemical measuring systems.

One property that may be assigned to a system is its state, which may be expressed simply, to a first approximation, as the pieces of information, pertaining to a given instant, which allow one to determine the behaviour of the system from this instant; given a known input, of course. Such pieces of information are, e.g. in the case of technological networks, for instance, the chemical composition, mass rate, pressure and temperature of technological flow (using the most common variables in chemical engineering practice for the description of the flow), and the parameters of unit operations, such as the conversion, in the case of the reactor, the partition coefficient or the vapour–liquid equilibrium constant in the separation, the pressure difference in the case of pumps and valves, or the transmitted heat in the case of heat exchangers. The systems discussed in this book may have several types of state. In the control of the processes of chemical technology, the time dependence of the state is relevant. The investigation of chemical technological networks involves the treatment of both dynamic and static (steady) states, when time dependence or the state before and after the change is relevant. In the former case one is concerned with transition, while in the latter case transmission is discussed. In the case of analytical chemical systems, the state contains analytical and chemical items of information.

Consequently, the notions of system and state will be defined and used in this book as a result of approaching these concepts from the vantage point of several specialized fields. Therefore, in this introduction, the subject will be reviewed, with the use of literature sources [1], on the basis of mathematical system theory, and not in connection with any single physical entity, nor from the viewpoint of a narrow field of science. We note that another, equivalent theoretical treatment of this subject is also known, which is based on a transport theoretical approach, but this will not be discussed here in detail.

The first concept to be introduced in this discussion is that of an object. Let us define an object as a given domain of reality the internal structure of which is unknown, as expressed in the case of network approach by the concept of a 'black box' as used in cybernetics.

Let us now consider physical object $P$ together with the set of its properties. Let us restrict our discussion to the continuous case and assume that the range $T$ of variable $t$ is restricted to the real numbers. The values of the properties in instant $t$ are $v_1(t), \ldots, v_n(t)$.

The physical identity of these functions is then irrelevant mathematically, only the relationships between them being of interest. Thus, one arrives at the abstract object formed by these relationships.

The set of pairs $\{t, \underline{v}(t)\}$, $t \in I$, is called the segment of $v$ over observation interval $I$ where the expression in brackets denotes a set of multitude (occasionally the symbol $\{x|P\}$ will also be used – referring to property $P$ of the elements of set $x$). The symbol $\in$ shows $t$ is an element of interval $I$.

Let us denote a set of segments for all $t_0$ and $t_1$ in the interval $(-\infty, \infty)$ where $t_1 \geqq t_0$, by $V \triangleq \{\underline{v}[t_0, t_1]\}$. It is also possible that more than one segments $\underline{v}[t_0, t_1]$ are assigned to all intervals $t_0$, $t_1$. (The symbol $[t_0, t_1]$ denotes, as usual, the closed interval $\{t|t_0 \leq t \leq t_1\}$, whereas the symbol $\triangleq$ means 'equal by definition'.)

Let us denote a point in the interval $[t_0, t_1]$ by $\tau$. The set of segments $V = \{\underline{v}[t_0, \tau], \underline{v}[\tau_1, t_1]\}$ is called closed with respect to segmentation if, for any $\tau$ within the interval $[t_0, t_1]$ both $\underline{v}[t_0, \tau]$ and $\underline{v}[\tau, t_1]$ belong to $V$. This is one possible way of expressing the condition closedness. An equivalent expression is that set $V$ is closed with respect to segmentation if and only if any segment of a segment belonging to $V$ also belongs to $V$. By means of the condition of closedness, abstract objects may be defined in a particular but still general way. Accordingly, abstract object $A$ is the set

$$A = \{\underline{u}[t_0, t_1], \underline{y}[t_0, t_1]\} \tag{1.1}$$

of ordered segment pairs $\{\underline{u}[t_0, t_1], \underline{y}[t_0, t_1]\}$ which satisfies the condition of closedness.

Let us call segment $\underline{u}[t_0, t_1]$ the input of $A$ and segment $\underline{y}[t_0, t_1]$ the output of $A$. Then $\{\underline{u}[t_0, t_1], \underline{y}[t_0, t_1]\}$ is the input-output pair belonging to $A$. This has led us to the general definition that an abstract object is a set of input-output pairs satisfying the condition of closedness:

$$A = \{\underline{u}, \underline{y}\} \tag{1.2}$$

The group of output-input relations which gives the outputs as a function of inputs, i.e. the set

$$D(A) \triangleq \{\underline{u}|(\underline{u}, \underline{y}) \in A\} \tag{1.3}$$

where $\underline{u}|(\underline{u}, \underline{y})$ denotes the open interval $\{u|u < u < y\}$, is called the range of definition of $A$, whereas the set

$$R(A) \triangleq \{\underline{y}|(\underline{u}, \underline{y}) \in A\} \tag{1.4}$$

is the range of values of $A$.

The usual symbol of an abstract object with input $\underline{u}$ and output $\underline{y}$ is shown in *Fig. 1.2.*

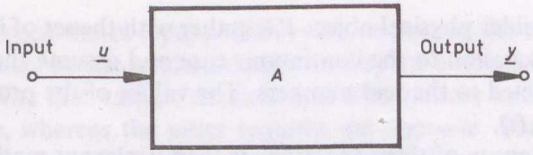



Fig. 1.2. An abstract object

Returning briefly to the connection between abstract and physical objects, it follows from the above definition that input consists of properties which may change (other) properties of the physical object. The latter properties comprise the output, which indicates the effect of the change for the observer.

Since any interaction with a physical object is observed through changes in properties, it is justifiable to identity objects with observable properties.

We should like to draw attention to the fact that owing to the concept of abstract object outlined above, the input-output relationship has been defined as a causal relation, i.e. input and output are not necessarily identical to the physical flows entering and leaving the system, respectively. By way of proof, let us consider, for instance, that in the control of a technological network the properties not only of the input but also of the output flows may be the input values of the abstract object. This is utilized when a so-called planning simulation is performed, i.e. the parameters of the equipment are determined by a search using fixed values of technological input and output flows, such as amount and quality of product, feed and composition of raw material, etc.

It should also be noted here that the definition of abstract object permits a whole range of outputs to belong to each input. This can be illustrated by considering a differential equation used for the description of control systems, which can be regarded as a typical input-output relationship:

$$dy/dt + y = u \tag{1.5}$$

The general solution of Eq.(1.5) is

$$y(t) = \alpha \exp[-(t-t_0)] + \int_{t_0}^{t} \exp[-(t-\zeta)]\, u(\zeta)\, d\zeta \tag{1.6}$$

where $t >= t_0$

In this solution $\alpha$ is an arbitrary real number. This will be called a state of $A$.

As can be seen from this example, the reason for the ambiguity of response to a given input is that output depends not only on the input but also on the state of $A$.

Since this monograph is concerned with the practical application of system theory, it is necessary to point out that the number of input-output pairs defining an object is usually infinite, and thus no exhaustive list can be given. Therefore, instead of a list of pairs, one should give relationships, equations or algorithms for

the description of relations between input and output. These methods can be used to generate output values belonging to arbitrary input values.

In chemical engineering practice, real systems are often approximated by linear mathematical models. The practical advantage of this approach is also underspread in the treatment of chemical technological systems. As an illustration, the design of chemical technological networks may be mentioned. In the course of design and development, a significant role is played by the simplified calculation of material and energy balances based on linear systems of equations. A more advanced variant of the linear approximation method is when the value of the dependent variable vector is approximated under the assumption of linearity, and this is then refined by a nonlinear relationship. Such a combined method is, for instance, the simultaneous–sequential strategy used in the simulation of chemical technological networks, which may significantly speed up the necessary calculations.

According to the usual definition, object $A$ is linear if and only if a linear combination of any two of its input-output pairs is also an input-output pair of $A$.

If this condition is fulfilled, then the linear combination of two cases of the differential equation discussed above as an example

$$dy_1/dt + y_1 = u_1 \tag{1.7a}$$

and

$$dy_2/dt + y_2 = u_2 \tag{1.7b}$$

can be written as

$$d/dt(k_1 y_1 + k_2 y_2) + k_1 y_1 + k_2 y_2 = k_1 u_1 + k_2 u_2 \tag{1.8}$$

and

$$\{(k_1 y_1 + k_2 y_2), (k_1 u_1 + k_2 u_2)\} \tag{1.9}$$

is also an input-output pair of $A$.

Let us now link several abstract objects in such a way that the inputs or outputs of the individual objects coincide with the inputs or outputs of the other objects of the set. Such a combination of objects is called a system.

The chemical technological process shown in *Fig. 1.1* is shown as a system consisting of three objects in *Fig. 1.3*. In the figure, $\underline{u}$ denotes an input and $\underline{y}$ an output. The system communicates with its environment through constraints $\underline{u} = \underline{u}_{11}$ and $\underline{y} = \underline{y}_{31}$, whereas the connections between the objects of the system are determined by constraints $\underline{y}_{11} = \underline{u}_{21}$, $\underline{y}_{21} = \underline{u}_{31}$, etc.

The question is how the input-output pairs of system $A(A_1, \ldots A_n)$ can be determined if the input-output pairs of the objects of the system and the connections between the objects of the system are known.

One possibility is that all input-output pairs of components $A_1, \ldots A_n$ of the given system and the connections between the inputs and outputs of component systems $A_1, \ldots A_n$ are known from a list or an explicit description. In this case the set of input-output pairs forming system $A(A_1, \ldots A_n)$ can be determined.

In the other case the input-output pairs of component systems $A_i$ are unknown,

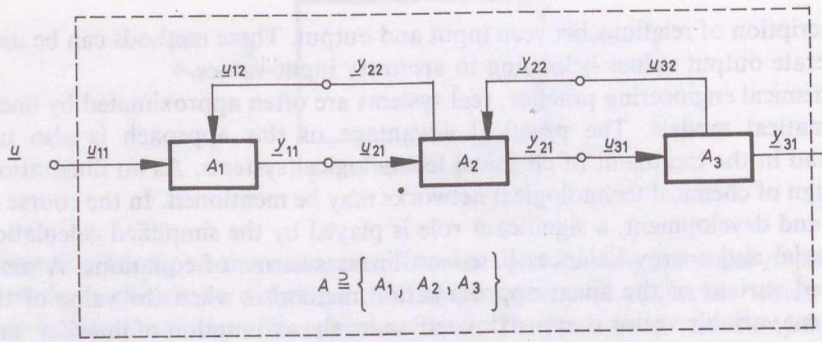



Fig. 1.3. A chemical technological system

only the relations determining them being available. In this case the equations defining these relations should be solved first, in order to determine the input-output pairs. In the treatment of chemical technological systems the majority of cases are of the latter type.

Therefore, we should return to the problem mentioned previously that the notion of an abstract object permits several possible outputs to belong to a single input, i.e. the response to a given input is not unique. The reason has also been given: the output depends not only on the input but also on the actual state of the object.

An approximate definition of state has already been given, and a possible mathematical description was illustrated by Eq.(1.6). Now the precise and general definition of the problem will be discussed.

Let us consider a system which satisfies the condition of closedness and the input-output relation of which is determined by Eq.(1.5). The general solution (Eq.1.6) of this differential equation gives the set of the input-output pairs explicitly. Let us denote the subset containing all input-output pairs starting with $t_0$ by

$$A(t_0) = \{\underline{u}_{[t_0,\,t]}\ \underline{y}_{[t_0,\,t]}\}$$

where $t_0 <= t$. If the pairs having some common property are selected from this subset (let us use label $\alpha_0$ to denote this common property), then a special subset of input-output pairs is obtained:

$$A_{\alpha_0}(t_0) = \left\{\underline{u}_{[t_0,\,t]}\ \alpha_0 \exp\left[-(t-t_0)\right] + \int_{t_0}^{t} \exp\left[-(t-\zeta)\right] u(\zeta)\,\mathrm{d}\zeta\right\} \tag{1.10}$$

The characteristic of this special subset is that it contains the input-output pairs holding for value $\alpha_0$. This special set is usually called a bundle, and $\alpha_0$ is the label of the bundle. Thus, a bundle contains those input-output value pairs which are generated by $u$ for all $t_0$ under a given value of $\alpha_0$. The symbol of such a bundle is

$$\{A_{\alpha_0}(t_0)\}_{\alpha_0}$$

Our next task is to investigate how to input-output pairs can be grouped into bundles so that their labels give the state of the system. Without going into a detailed analysis here, we note that under certain conditions the set of input-output pairs is consistent.

From among all the conditions of consistency, two may be mentioned with the intention of future application. One is uniqueness, which requires that to each input $u$ may belong only one output $y$. The second condition is that of continuation. This can be stated briefly so that if a system in a state $\alpha_0$ at initial moment $t_0$ passes into state $\alpha_1$ at moment $t_1$ as an effect of input $u_0$, then $\alpha_1$ may be regarded as an initial state of an input starting at $t_1$. The bundle $A_{\alpha_0}(t_0)$ of a consistent set is called an aggregate.

The label $\alpha_0$ of a system satisfying these conditions is called the state of system $A$ belonging to time $t_0$. If it is necessary to emphasize the time-dependent nature of the state, for instance, when referring to the dynamic behaviour of the system, the state is denoted by $x(t)$. Set $X_t$ consisting of numbers $x_t$ or, using the $\alpha_0$ notation set $\alpha_0 \in \Sigma_t$ is called the state space of the system.

Let us recall that an aggregate was defined as a bundle of input-output value pairs which satisfies, among other things, the condition of uniqueness. Thus, in the range of definition of the aggregate in any moment $t$ a single element $\underline{y}$ corresponds to each element $\underline{u}$:

$$y(t) = g(t, x_{t_0}, u_{[t_0, t]}) \tag{1.11}$$

This equation is called the equation of state of the input-output.

We arrive now at a relation, which generates one output value for one input. This enables us, as a result of the above considerations, to obtain a basic relationship of system theory which can also be used in engineering practice. Accordingly, one of the most important properties of the concept of state is that if the system is initially, i.e. at moment $t_0$, in a given state $\alpha_0$, then the input-output relation determines a unique response to input $\underline{u}$ through label $\alpha_0$. In other words, if the state of system in $t_0$ is known, the output of the system generated by an input starting from this time can be determined uniquely.

It may also be said that the state of the system contains all information of its past history that is necessary for determining its future behaviour.

Moreover, it may be added that $x_t$ determines the state of the system at time $t$ just as $\alpha_0$ determines its state at time $t_0$, i.e. it can be regarded as an initial state of an input starting at time $t$.

At this point let us summarize the most important concepts: a system is a set of input-output pairs, an aggregate is a bundle of input-output pairs satisfying certain conditions of consistency, and a state is the label of an aggregate.

Since state is only a label assigned to the aggregate, the same aggregate may be described by means of various labels, but only one aggregate belongs to each label. Thus, it is necessary to discuss the equivalence of states and the problem of matching system and states. Mathematical system theory has treated all these problems; for details we refer the reader to the literature (e.g. Ref. [2]).

Of the concepts and relationships of mathematical system theory reviewed here, later sections of this monograph will employ the concept of input-output state relation. This assigns, according to Eq.(1.11), the value of output $y(t)$ at time $t$ under consideration to given input segment $u$ and initial state $x_{t_0}$. Let us look at some examples as illustrations.

In connection with the control of chemical industrial processes, Eq.(1.4) has already been discussed. This form of linear differential equation is a usual input-output state relation in the investigation of the dynamic behaviour of a control system. To illustrate this, let us consider e.g. a control task according to which the outlet temperature of a substance fed continuously into a container with varying mass rate and temperature must be maintained at a constant value *(Fig. 1.4)*.

This can be achieved by changing the mass rate of a heating medium of constant temperature as a function of the temperature of the feed. Let us denote the mass rate of feed by $W_1$, its inlet temperature by $T_1$, outlet temperature by $T_2$ and the mass rate of the heating medium by $W_2$.

The dynamic energy balance of a container under conditions of perfect mixing is

$$Mc_p \mathrm{d}T_2/\mathrm{d}t = W_1 c_p T_1 + W_2 \lambda - W_1 c_p T_2 \tag{1.12}$$

where $M$ is the mass of liquid in the container,
$c_p$ is the specific heat of the liquid, and
$\lambda$ is the heat of condensation of heating steam.

Let us denote the nominal values of the variables in the steady state before perturbation by $W_{1,0}$, $W_{2,0}$, $T_{1,0}$ and $T_{2,0}$. Then the energy balance for the nominal steady state has the form

$$0 = W_{1,0}\, c_p (T_{1,0} - T_{2,0}) + W_{2,0}\lambda \tag{1.13}$$

As a consequence of perturbation the state changes. Let us use as an approximation by a Taylor polynomial in the vicinity of the nominal values, and introducing difference variables:

$$\Delta T_2 = T_2 - T_{2,0}$$

$$\Delta W_1 = W_1 - W_{1,0}$$

$$\Delta W_2 = W_2 - W_{2,0}$$

one obtains

$$Mc_p \mathrm{d}T_2/\mathrm{d}t = c_p (T_{1,0} - T_{2,0}) \Delta W_1 - W_{1,0} c_p \Delta T_2 + \lambda \Delta W_2 \tag{1.14}$$

By subtracting Eq.(1.14) from Eq.(1.12), introducing the notation $\Delta T_0 = T_{2,0} - T_{1,0}$ and since $W_{2,0}\lambda = 0$; arranging according to the change of the control variable, one has

$$M\mathrm{d}\Delta T_2/\mathrm{d}t + W_1 \Delta T_2 = \Delta T_0 \Delta W_1 + W_{1,0} \Delta T_0 + \lambda/c_p \Delta W_2 \tag{1.15}$$

which is the same as the differential equation, Eq.(1.5).

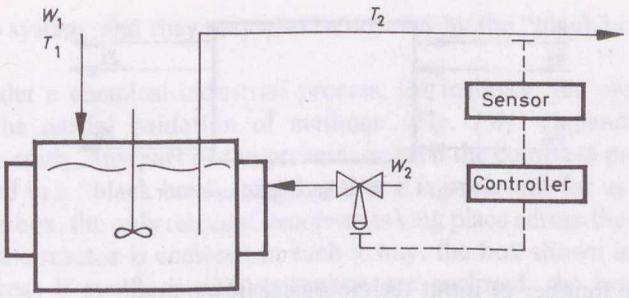



Fig. 1.4. Control scheme of a tank

Another example is a chemical technological network. In this case, too, the description of the elements of the system by means of input-output state relations is of central importance. The elements of the mathematical model of chemical technological systems may also be called technological operators, since they transform the variables of input technological flows into those of output technological flows. A heat exchanger is, for example, a technological operator which changes the temperature and energy of the input, a distillation column changes the chemical composition of inputs without producing new chemical components, whereas a reactor performs the latter change and may also affect the temperature and energy of the input.

An important case, from a practical point of view is when the process is modelled as a steady-state operation. The input-output relation is then time-independent.

Under such circumstances, the relationship between the inputs and outputs of the technological operator shown in *Fig. 1.5* has the form

$$\underline{Y}=F(\underline{U}) \tag{1.16}$$

where $\underline{U}$ is the column matrix of the variables of input technological flows,
$\underline{Y}$ is the column matrix of the variables of output technological flows, and
$F(\underline{U})$ is a vector–vector function of the variables of input technological flows.

If function $F$ is given in matrix form, a linear model for the system or element of the system is assumed.

Let us denote the column matrix of input flow by

$$\mathbf{U}=\begin{bmatrix} U_1 \\ \vdots \\ U_m \end{bmatrix}$$

where $U_1, U_2, \ldots, U_m$ are, themselves, column matrices:

$$U_1=\begin{bmatrix} U_{11} \\ \vdots \\ U_{1k} \end{bmatrix} \quad U_2=\begin{bmatrix} U_{21} \\ \vdots \\ U_{2k} \end{bmatrix} \ldots, \; U_m=\begin{bmatrix} U_{m1} \\ \vdots \\ U_{mk} \end{bmatrix}$$

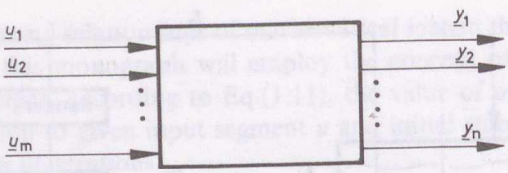



Fig. 1.5. A technological operator

Here $m$ is the number of input technological flows, and,

$k$ is the number of chemical component flows in the input technological flows.

If there is a linear relationship between input and output flows, the linear model of the technological operator can be given as

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} P_{11} & P_{21} & \dots & P_{m1} \\ \vdots & \vdots & & \vdots \\ P_{1n} & P_{2n} & \dots & P_{mn} \end{bmatrix} \cdot \begin{bmatrix} U_1 \\ \vdots \\ U_m \end{bmatrix} \tag{1.17}$$

where $n$ is the number of output technological flows,

$Y_1, \dots, Y_n$ are column matrices:

$$Y_1 = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1k} \end{bmatrix}, \quad Y_2 = \begin{bmatrix} y_{21} \\ \vdots \\ y_{2k} \end{bmatrix} \dots, \quad Y_n = \begin{bmatrix} y_{n1} \\ \vdots \\ y_{nk} \end{bmatrix}$$

and hypermatrix $\mathbf{P}$ is the transformation matrix of the technological operator. The size and elements of this matrix depend on the type of technological operator. Its particular forms for simple operations will be shown in Section 1.2.

## 1.2 Elements of chemical technological systems and the connections between them

### 1.2.1 The element: definition and types

The definition of a system given in Section 1.1 is suitable for the discussion of the specific properties of systems of chemical technological networks. As a result of this discussion, a system has been interpreted as a set of objects connected by input-output relationships. A unique description of this system may be given by the input-output state relation which, satisfying the given conditions, generates an output value for the given input value.

In the investigation of the processes used in the chemical industry, those components of the system will be regarded as "objects" for which a further subdivision is unnecessary under the given circumstances. These objects will be taken as the

elements of the system, and they may also be covered by the "black box" concept of cybernetics.

Let us consider a chemical industrial process; for instance, the production of acetylene by the partial oxidation of methane *(Fig. 1.6)*. Depending on the objective of the study, any part of the process, or even the complete process itself, may be enclosed in a "black box", implying that it is irrelevant for us what takes place within the box, the only relevant processes taking place across the boundaries of the box. If the reactor is enclosed in such a way, the box shown in *Fig. 1.6* is obtained, whereas if synthesis and separation are enclosed, the box shown in *Fig. 1.7* is obtained. The main point of the black-box concept is that, using it, a complex problem may be reduced arbitrarily into simpler partial problems, and thus the whole of a complex chemical technological system need not be investigated simultaneously.

An element of a chemical technological system may thus be a unit operation, but it may also be a unit having a higher rank in the hierarchy of the system, e.g. a step of a process or a complete process, for instance, the acetylene plant in a complex petrochemical plant.

To illustrate the input-output state relation of an element, let us take the case when a simple unit operation is the element of a chemical technological system. In the simplified case of the mass balance level description of a chemical technological system, the input-output relations are linear. In the general form, as a characteristic example of input-output state relations, this was illustrated by Eq.(1.17) in Section 1.1. In the subsequent discussion, the same example will be discussed for the particular cases, i.e. the basic ways of handling chemical technological systems at network level.

The transformation matrix of a mixer *(Fig. 1.8, a)* is a row matrix consisting of two unit matrices:

$$\mathbf{Y}=[\mathbf{E}_1,\ \mathbf{E}_2]\times\begin{bmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{bmatrix}=\begin{bmatrix}1\ldots\mathbf{0} & \mathbf{1}\ldots\mathbf{0}\\ \mathbf{0}\ldots 1 & \mathbf{0}\ldots 1\end{bmatrix}\times\begin{bmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{bmatrix} \tag{1.18}$$

where the unit matrix is a $k\times k$ matrix ($k$ is the number of chemical components).

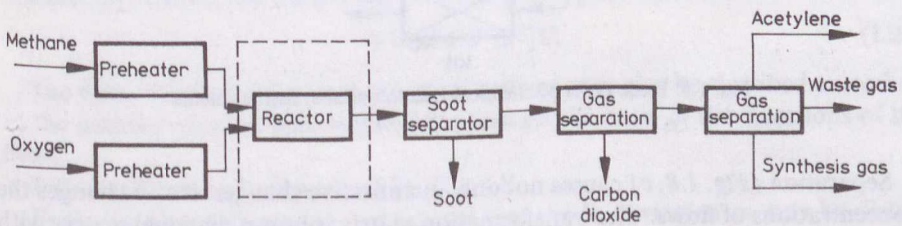



Fig. 1.6. The relative nature of system and its elements

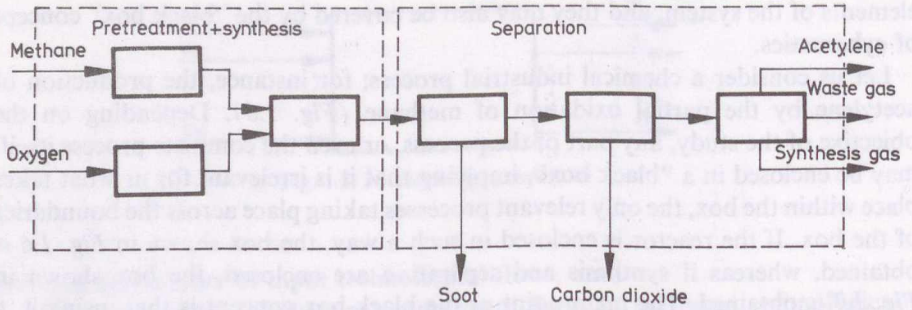



Fig. 1.7. The relative nature of system and its elements

The input-output relation of a branch *(Fig. 1.8, b)* is

$$\mathbf{Y}_1 = [a_1\ \mathbf{EU}] = \begin{bmatrix} a_1 & & \mathbf{0} \\ & \ddots & \\ & a_1 & \\ & & \ddots \\ \mathbf{0} & & a_1 \end{bmatrix} \mathbf{U} \tag{1.19}$$

and

$$Y_2 = [(1-a_1)\,\mathbf{EU}] = \begin{bmatrix} (1-a_1) & & \mathbf{0} \\ & \ddots & \\ & (1-a_1) & \\ & & \ddots \\ \mathbf{0} & & (1-a_1) \end{bmatrix} \mathbf{U} \tag{1.20}$$

where $a_1$ is a scalar and denotes the fraction directed into flow 1 ($a_1$ may also be called the cutting or branching ratio).

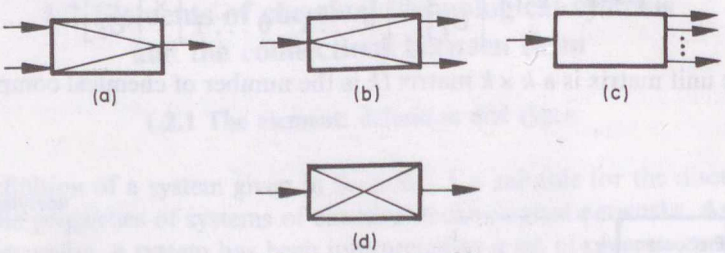



Fig. 1.8. Basic types of chemical technological system elements

Separation *(Fig. 1.8, c)* causes not only quantitative changes: it also changes the concentrations of flows. The transformation matrix is now a diagonal matrix with different terms in the main diagonal. Accordingly, the input-output relation is

$$\underline{Y}=\begin{bmatrix} a^{j}_{11} & & & \\ & \ddots & & \\ & & a^{j}_{ii} & \\ & & & \ddots \\ & & & & a^{j}_{kk} \end{bmatrix} \underline{U} \tag{1.21}$$

The chemical transformer (reactor) (*Fig. 1.8, d*), by definition, is a transformer of chemical components. Its transformation matrix is, for $r$ independent stoichiometric equations,

$$\mathbf{P}=\mathbf{E}+\mathbf{RX} \tag{1.22}$$

where $\mathbf{X}$ is the conversion matrix, and $\mathbf{R}$ is the reaction matrix

$$\mathbf{X}=\begin{array}{c} \\ 1 \\ 2 \\ i \\ r \end{array}\begin{array}{c} \begin{array}{ccc} 1 & 2 & k \end{array} \\ \begin{bmatrix} 0 & \dfrac{X_{12}}{-\nu_{12}} & 0 \\ \dfrac{X_{21}}{-\nu_{21}} & 0 & 0 \\ 0 & 0 & \dfrac{X_{ik}}{-\nu_{ik}} \\ 0 & 0 & 0 \end{bmatrix} \end{array} \qquad \mathbf{R}=\begin{array}{c} \\ 1 \\ i \\ k \end{array}\begin{array}{c} \begin{array}{ccc} 1 & j & r \end{array} \\ \begin{bmatrix} \nu_{11} & \nu_{1j} & \nu_{1r} \\ \nu_{i1} & \nu_{ij} & \nu_{ir} \\ \nu_{k1} & \nu_{kj} & \nu_{kr} \end{bmatrix} \end{array}$$

In these matrices $X_{ij}$ is the conversion of reaction $j$ for component $i$, and $\nu_{ij}$ is the stoichiometric coefficient of component $i$ in reaction $j$. The positions of the elements in the matrix depend on the initial conditions (a conversion value should be assigned to all reactions). The calculation of conversion may also be based on kinetic data. Thus, for instance, if a first order, irreversible reaction proceeds in a tank reactor with perfect mixing, the conversion is

$$X=1-\frac{1}{1+k\Theta} \tag{1.23}$$

where $\Theta$ is the residence time and $k$ is a rate constant.

From the above, the input-output relation of the reactor is

$$\mathbf{Y}=[\mathbf{E}+\mathbf{X}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}]\mathbf{U} \tag{1.24}$$

The elements of chemical technological systems may also be classified according to the number of input and output flows and on the basis of the directions of the flows.

This classification is illustrated by *Fig. 1.9*.

From the above discussion, we can see that the following cases may be distinguished:

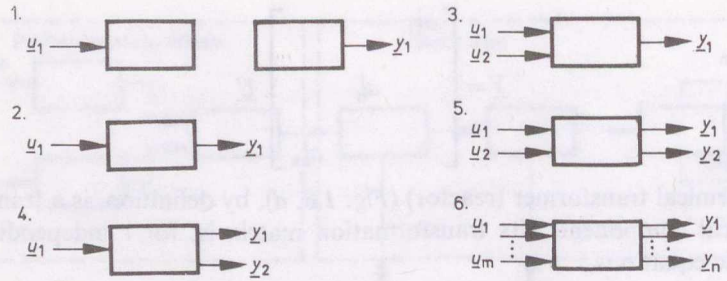



Fig. 1.9. Classification of elements according to number of flows

1. element with one input or one output flow,
2. element with one input and one output flow,
3. element with two input and one output flow,
4. element with one input and two output flows,
5. element with two input and two output flows,
6. elements with arbitrary number of input and output flows.

If the flows are restricted to mass and energy flows only, the above cases may be illustrated by the following examples:

**Case 1**

— filling or emptying of a tank, if mass flow only is taken into account,
— filling or emptying of an adiabatic tank, if energy flow is also considered,
— adiabatic mixing in a mixer with closed feed and removal, taking account of energy flow.

**Case 2**

— tubing, considering the mass flow in it,
— flow in adiabatic tubing, taking account of energy flow connected to mass flow,
— adiabatic tube or continuous tank reactor with respect to energy flow,
— grinder mill if mass flows are considered.

**Case 3**

— mixing of two mass flows, taking account of mass flows only,
— adiabatic mixing of two mass flows, with respect to the energy flows
— substance flowing in heated tube, with respect to its energy flows,
— compression with respect to energy flows,
— grinder mill if energy flows are regarded.

**Case 4**

— flow splitter with regard to mass flows,

— separation of substance in filter, centrifuge or cyclone, taking account of mass flows,

— expansion turbine with respect to energy flows.

**Case 5**

— heat exchange between two mass flows with regard to mass flows,

— heat exchange between two mass flows in an adiabatic heat exchanger, taking account of energy flows.

Despite its formal character, this classification is also useful in practice. In the investigation of the network structure of real chemical technological systems, one of the important tasks is the enumeration of input and output flows, which is connected with the above classification.

A further advantage is that this classification gives assistance in recognizing the basic flow modes: concurrent, countercurrent and cross-flows. As is known, these types of flow involve different driving force relationships.

Another possibility for the classification of elements is a classification on the basis of the mathematical form of the state relations. The simplest principle is classification on the basis of the linearity or nonlinearity of the equations used for their description. In the case of chemical technological systems, the elements are usually nonlinear in their behaviour. Nevertheless, it is a general practice, e.g. in the initial phase of design or development, when material balance level variants are analyzed, to approximate these nonlinear elements by linear equations. In such cases this is a sufficient and useful approximation. When higher accuracy is required, the elements of chemical technological systems may be modelled by algebraic or transcendental equations.

A further classification principle arises from the distinction of elements according to the type of equation used for their description (i.e. ordinary or partial differential equations). Usually, ordinary differential equations are used to describe the behaviour of elements with distributed parameters in one space coordinate, in the steady state, and those with concentrated parameters in non-steady states. Partial differential equations are used in the case of steady state elements if the parameters are distributed among several space coordinates, and for non-steady-state systems with distributed parameters.

It should be noted that the classification of elements according to mathematical relationships is useful because the elements described by the same mathematical relationship can be treated as models with the same mathematical methods, having no regard to their widely divergent technological roles.

On the basis of time dependence, steady state and non-steady-state elements may be distinguished. In the former, the values of parameters at any points in space are constant in time. The element is in a non-steady-state if its state varies with time

at all points in space. Examples of the latter are startup or shutdown of systems or transitions of systems from one steady state to another one. If the change in time is relatively slow, the system is in a quasi-steady-state.

In certain cases, classification may be based on the extent of changes in state variables upon perturbation. Perturbations manifest themselves in the fact that for known input or control variables the state variables are not unambiguously determined. According to the effect of the perturbations, deterministic and stochastic elements may be distinguished. An element is deterministic if the initial state and the input uniquely determine the output and the future state of the system. This also includes the possibility for stochastic input variables. In contrast, if for a given input, a range of outputs and future states exists, all with a probability that may be determined in advance, then the element is stochastic.

A different manner of classification is possible, based on the effect of an element on the substances. If classification is based on mass flows, the following categories are possible:

— elements for separation (splitting),
— elements for mixing (merging),
— elements for dispersing, comminution,
— elements for agglomeration,
— elements for shaping,
— elements for heat exchange,
— elements for compression and expansion,
— elements for chemical transformation (reaction).

All of these groups include several types of equipment. Thus, for example, separation includes distillation, drying, centrifuging, absorption, etc.

A further characteristic of the classification of elements by their function is when the properties of matter do not change within the element. They are:

— elements for storage,
— elements for transport.

A basis for classification according to function may be the type of energy transformation. The most frequent types of energy in chemical technology are:

— mechanical energy,
— electric energy,
— heat,
— concentration energy,
— chemical energy.

During energy transformations, any of the above energies may be an input or an output. A real element may be a mixture of a number of basic types. Thus, for example, the input of a cooled, mixed tank reactor includes mechanical and chemical energies, and the output is thermal and chemical energy, whereas the input and output of distillation are concentration and thermal energies.

### 1.2.2 Connections in chemical technology

In chemical technological systems the interconnection of elements is often complicated. Nevertheless, on the basis of a closer examination, the characteristic connections and the resulting characteristic structures which may occur in a chemical technological system may be recognized. They are called basic connections or basic structures, respectively.

It may be added that these basic structures and connections can also be applied if the interconnections of subsystems consisting of a number of elements are studied.

The basic connections and their properties will therefore be discussed here at some length. The following basic connections can be distinguished:

— serial,
— cascade,
— parallel,
— serial–bypass, and
— feedback connections.

When these connections are defined, no restrictions are imposed on the elements within the subsystem. This classification according to structure is based on the model-type reproduction of system structure. Accordingly, similarly to the way models aim generally at a description of the various properties of the system, a structural classification gives particulars on a given property of the system, too. This will be illustrated using the example of mass and energy transport; the conclusion drawn from these examples, however, are of more general validity. This will be pointed out where appropriate.

A connection is serial if the output of an element is the input of another element, with the restriction that in the sequence of flows all elements are encountered only once *(Fig. 1.10)*. The number of elements may be arbitrary. Of course, a subsystem consisting of serially connected elements may be connected to other elements, subsystems or the environment in a way that is other than serial. This also holds, of course, for the other connections discussed below.

By its function, serial connection is applied to perform consecutive changes in properties or to create special conditions for the process. Accordingly, it is applied when the required change cannot be performed in a single step, when some limiting values should be taken into account for certain process parameters (e.g. maximum temperature), or when the multistep process has certain advantages (e.g. higher efficiency), over the single-step one.

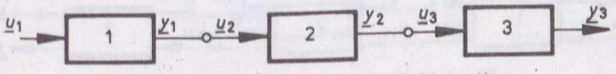



Fig. 1.10. Serial technological connection

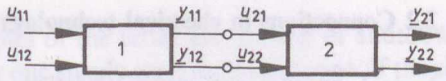



Fig. 1.11. Concurrent technological connection

Subcategories of serial connection can be distinguished if the elements have two input and two output flows. Such elements often occur in chemical technological systems, hence the distinction. These subcategories are:

— concurrent,
— countercurrent, and
— jump connection.

Concurrent connection is characterized by having the same flow direction between the two connected elements *(Fig. 1.11)*. In this case both output flows of element 1 are the input flows of element 2.

Countercurrent connection is seen when the flows between the two elements are opposite in direction *(Fig. 1.12)*.

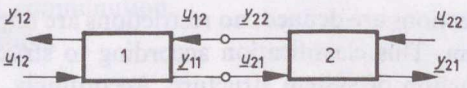



Fig. 1.12. Countercurrent technological connection

Concurrent and countercurrent connections are mainly applied in heat exchanger systems.

An extension of the serial connection is the jump connection, in which one or more further elements are inserted between two elements connected concurrently or countercurrently. According to the direction of flow, two basic cases can be distinguished *(Fig. 1.13)*. Such connections are also very frequent, mainly in heat exchanger systems.

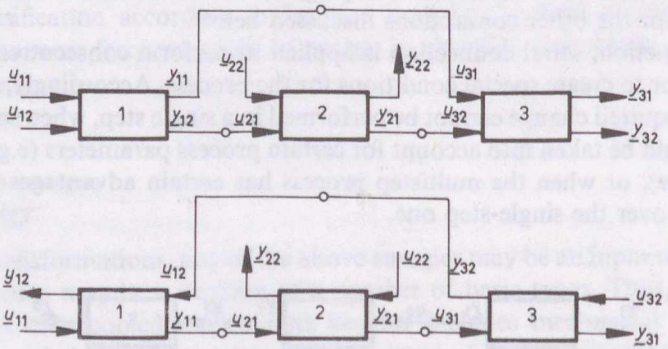



Fig. 1.13. Technological jump connections

Cascade connection is similar in many respects to serial connection, therefore, the two terms are often used as synonyms for a single category. Cascade connection is, more precisely, a connection in which the flow of useful product of one element is fed into the next element. The flow of useful product is not necessarily identical to the output flow; this is the reason for the distinction between cascade and serial connections. A characteristic example of a cascade is when operational units in equilibrium are connected, such as the plates of rectification columns or continuous steady-state tank reactors. Further practical examples of cascades will be found in the case of energy supplying circular flows connected with chemical transformation.

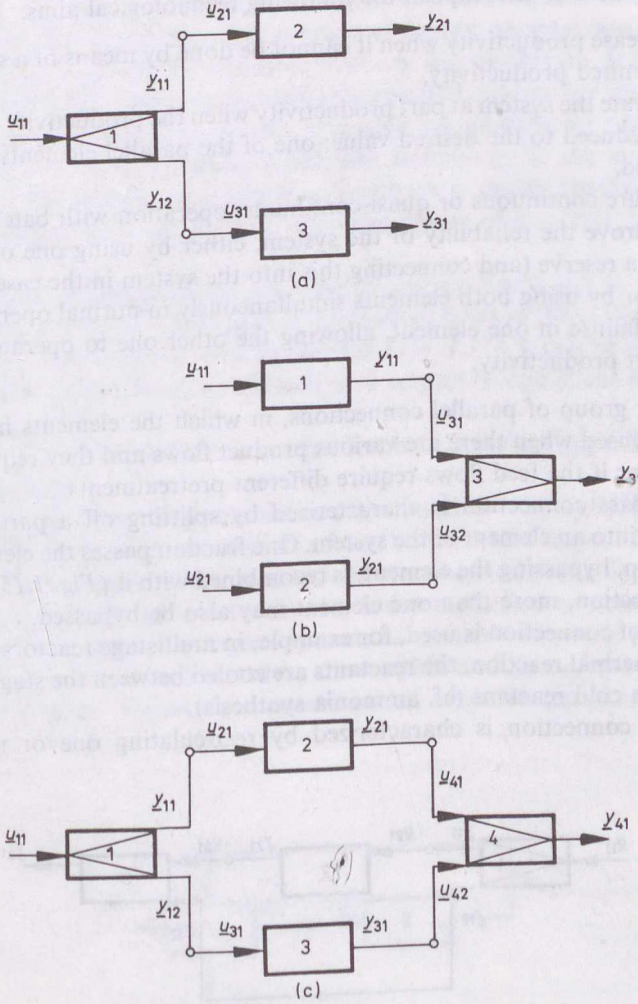



Fig. 1.14. Parallel technological connections

The connection of elements shown in *Fig. 1.14* is a parallel connection, for which three subclasses can be distinguished: parallel connection for the input flows *(a)*, the output flows *(b)* or both *(c)*. It can be seen from the figures that a parallel connection consists of a mixer and/or splitter and the parallel elements. It is stressed that subsystems consisting of several elements may also be connected in parallel.

Further classification is possible if it is based on the function of the elements. Accordingly, parallel connection may exist between elements with the same or different functions.

The former system, characterized by elements in parallel having the same function may be used to accomplish the following technological aims:

— to increase productivity when it cannot be done by means of a single element due to its limited productivity,

— to operate the system at part productivity when the productivity of an element cannot be reduced to the desired value; one of the parallel elements may be shut down instead,

— to ensure continuous or quasi-continuous operation with batch elements,

— to improve the reliability of the system, either by using one of the parallel elements as a reserve (and connecting this into the system in the case of failure in the other), or by using both elements simultaneously in normal operation and, in the case of failure in one element, allowing the other one to operate the system, albeit at part productivity.

The other group of parallel connections, in which the elements have different functions, is used when there are various product flows and they require different treatments, or if the feed flows require different pretreatments.

Serial–bypass connection is characterized by splitting off a part of the flow before entry into an element of the system. One fraction passes the element and the other fraction, bypassing the element, is recombined with it *(Fig. 1.15)*. Of course, in this connection, more than one element may also be bypassed.

This type of connection is used, for example, in multistage reactors when, in the case of exothermal reaction, the reactants are cooled between the stages by mixing the feed with cold reactant (cf. ammonia synthesis).

Feedback connection is characterized by recirculating one or more output

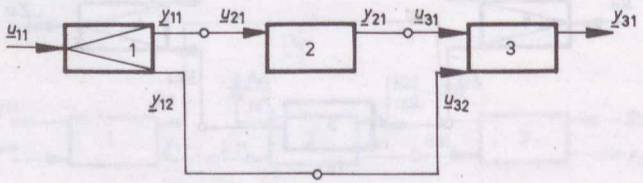



Fig. 1.15. Serial–bypass technological connection

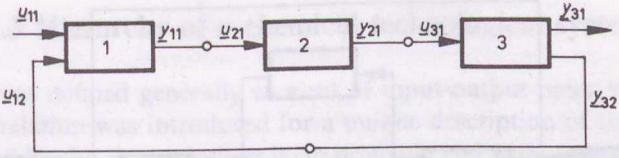



Fig. 1.16. Feedback technological connection

flow(s) of an element, possibly after passing them through other elements, into the given element *(Fig. 1.16)*. A synonym for feedback connection that is often used is recirculation connection. This connection may be necessary for various reasons; e.g. in reactor technology if the full conversion of the raw material into the product is impossible in a single step for thermodynamic or kinetic reasons. It may then be advantageous to separate the product from the reaction mixture leaving the reactor and to feed it back, if necessary after changing its enthalpy or physical state. Similar considerations have led in distillation technology to the application of reflux for the improvement of separation. Feedback is widely applied in the field of auxiliary materials; as an example the recirculation of catalysts may be mentioned.

Feedback is also significant in the case of energy flows and is applied, for instance, to the utilization of the heat of by-product flows in the system. This type of recirculation is often called cross connection *(Fig. 1.17)*. It can be seen from the scheme that this connection is a feedback with respect to one of the flows (in this case the energy flow).

The variety of these examples shows that feedback is widely applied in chemical technological systems.

The connection types discussed above correspond to a classification of technological connections from the aspects of one of the technological flows. In the general case, the character of the connections of a chemical technological system, i.e. the way of the elements are connected, is a complicated combination of the typical connections. All technological systems have their own technological structure (topology), which is determined by the number and type of the components of the system and by the required combinations of serial, parallel, bypass and feedback connections.

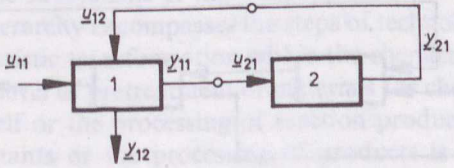



Fig. 1.17. Technological cross connection

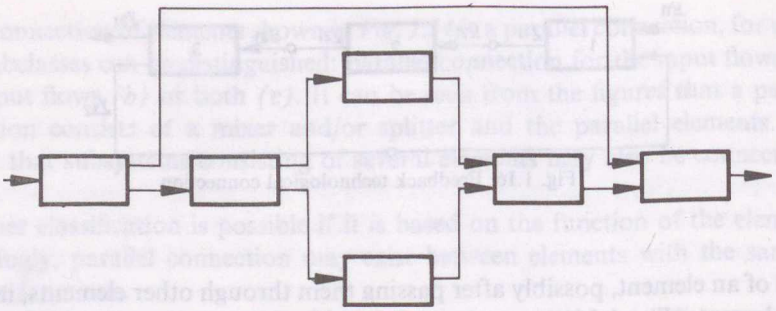


Fig. 1.18. Open chemical technological system

From the aspects of technological topology, chemical technological systems are classified as open and closed systems.

The topology of an open chemical technological system is characterized by serial, parallel and bypass connections between the elements. In such cases all technological flows pass any element of the technological system only once *(Fig. 1.18)*.

In a closed chemical technological system, a group of elements is closed into a loop by at least one recirculated mass or energy flow. A closed system may contain simple or composite feedback *(Figs 1.19* and *1.20)*. In fact, any recirculating subsystem is a closed subsystem of the chemical technological system.

The development of modern chemical technological systems is characterized by an increasing number of closed subsystems and the number of elements within each subsystem.

For the treatment of system structures in chemical technology, advanced computer methods have been developed, which will be discussed in a later chapter of this work.

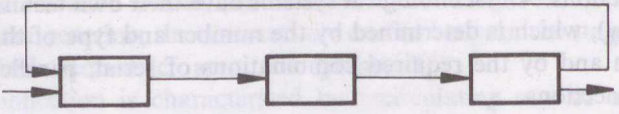


Fig. 1.19. Closed chemical technological system with simple feedback

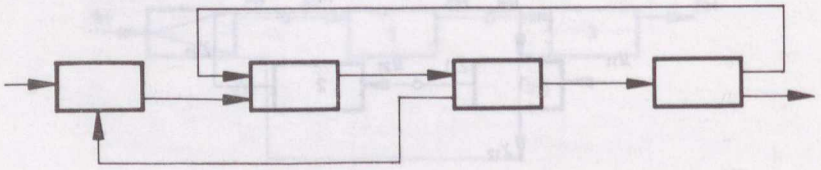


Fig. 1.20. Closed chemical technological system with composite feedback

## 1.3 Hierarchy of a chemical technological system

A system was defined generally as a set of input-output pairs, and the input-output state relation was introduced for a unique description of the system. The system defined by the state relations is, in principle and with general validity, not identical to the original system, since the state relations were set up under various assumptions, and since they involve physico-chemical data subject to errors of measurement. This line of thought leads us to realize that the system described in this manner is a model of the real system.

Models are always created with a particular aim, by applying abstractions and simplifying assumptions in harmony with the stated goal.

It is reasonable, in the case of chemical technological systems, to construct and investigate the model on the basis of a hierarchical strategy. The variability and complexity of chemical technological systems make it necessary to apply a modelling strategy which assists the designer in setting up the model and in assessing relationships and data. In the choice of the individual levels of hierarchy, depending on the degree of abstraction, several variants are possible. This section deals with a multilevel hierarchy of chemical technological systems which, on the one hand, illustrates by its detail the complexity of real chemical technological systems and, on the other hand, is abstract enough as not to lose the advantages of abstraction *(Fig. 1.21)*. It should be noted, however, that the hierarchy discussed here is only one possibility, and other types of classification are also known. Moreover, like any other classification, it should also be handled with due flexibility.

The lowermost level of the hierarchy comprises the simple operational units. A unit is called simple if the input flows are mixed once at most. A characteristic example is a mixer or a single plate of a distillation column.

The next level of the hierarchy is the composite operational unit. In this unit, mass flows are mixed several times with one another, consecutively, in separate but interconnected parts of the unit. It is a condition that the flows leaving the separate parts are in physical equilibrium or quasi-equilibrium. Such a composite operation unit is a distillation or absorption column.

The next highest level is the complex operational unit. This consists of composite operational units consisting of identical stages. This level is generally characterized by medium complexity and sophistication. The limit of its magnitude is at most 5 to 10 elements. As example, systems of heat exchangers and distillation columns can be mentioned. Complex units are very frequent in chemical technological systems, they form the subsystems of the whole system.

The next level of hierarchy encompasses the steps of technological process, each of which is a characteristic transformation within the chemical industrial process. Generally, this is the level of pretreatment of materials for chemical reactions, the chemical reaction itself or the processing of reaction products. Of course, if the pretreatment of reactants or the processing of products is a complex process, further process steps may be defined within these processes.

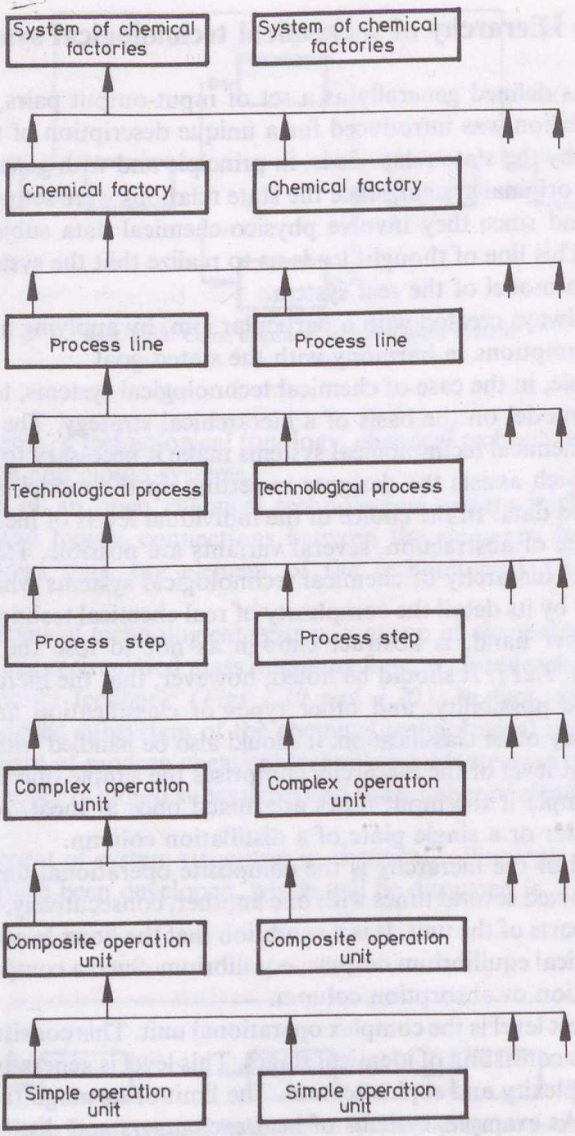



Fig. 1.21. Hierarchy of chemical technological systems

In chemical processes, the utilities (energy, auxiliary materials) are often complex and sophisticated systems, therefore further steps may also be defined here.

By aggregating process steps, a higher level of hierarchy is obtained: the process. The main characteristic of the process is that it encompasses the steps of reactant pretreatment, chemical transformation and product processing as well as the necessary utilities and supply of auxiliary materials. Of course, the above bound-

aries of process level may not be taken rigidly. Let us consider a common situation in chemical industry, where processes are connected to one another in such a way that the processing of the product of one process is the reactant pretreatment step of the next process. Process steps consisting of composite and complex operation units and technological processes composed of these steps are shown in *Fig. 1.22,* illustrated by the example of acetylene production.

This note leads us to the next level of hierarchy, at which level processes are organized into a system with the aim of producing a marketable product from a raw material. This unit is called process line. An example is polyethylene manufacture, consisting of the processes of petroleum pyrolysis, polyethylene production and polyethylene processing. A process line may also be such that the marketed product is produced from a number of raw materials. An example may be the manufacture of polyvinyl chloride from chlorine and ethylene, in which ethylene is produced by petroleum pyrolysis, chlorine by alkali chloride electrolysis, then from these raw materials the polymer is produced and finally it is processed.

The next level of hierarchy is the chemical factory or the chemical combine. The characteristics of this level are:

— generally several plants are included, with various connections among them,
— they generally have central utilities and systems for auxiliary materials, and central waste disposal and sewage systems,
— their important subsystem is the storage of raw materials, intermediates and final products, with respect to both the technological and the energy and auxiliary material flows.

The uppermost level of chemical technological systems comprises the systems obtained by connecting chemical factories and combines. An example is an oil refining and petrochemical combine or the set of such combines. An important feature of this system is the link between the factories, achieved by means of pipelines and intermediate storage systems.

This uppermost hierarchial level of chemical technological systems plays an increasing role nowadays. One reason for this is that the construction of larger plants is advantageous economically because of the declining character of the specific investment costs vs. capacity function, which is one of the preconditions of economic investment. On the other hand, this level promotes the gradual development of vertical production. It cannot be neglected, either, that this system allows the plants to be economically located: the one processing raw material my be located close to the source of raw material, the one processing the product may be located in an appropriate position with respect to the market. For these and other reasons, such links of substance flows are being extended to an increasing range of substances, such as coal, petroleum, natural gas, ethylene, propylene, higher hydrocarbons, oxygen, nitrogen, ammonia, methanol, chlorine, etc.

It may be noted in connection with the figures that the blocks used in the flowsheets illustrate the most important hierarchical features. Accordingly, the

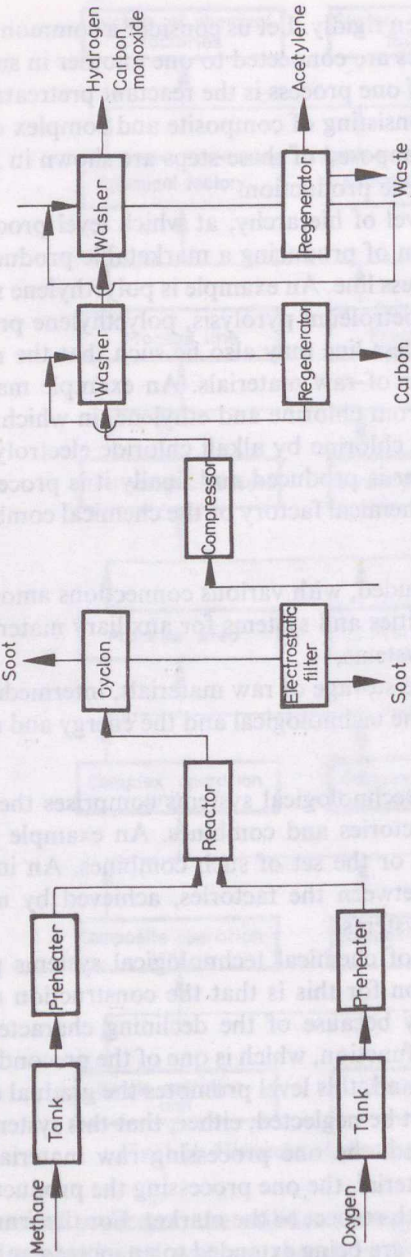



Fig. 1.22. Process steps consisting of composite and complex operation units; the process steps comprise a technological process

types of flows representing the connection between the elements of the system are not specified. If this is also done, further classifications become possible within the hierarchy levels discussed so far. Generally, from the aspects of the investigation, the analysis of chemical technological systems, the system may be built up gradually from mass, energy and information flows and from their combinations.

Two important features of hierarchic design are evident from the above discussion. First, the components of the system corresponding to the given level of hierarchy are the systems of the hierarchy one level below. Second, the detail and accuracy of the models of the elements forming the system should always match the given level of the hierarchy. Accordingly, during a given calculation, adequate modelling of the element is required for all elements or subsystems.

## 1.4 References

[1] Zadeh, L. A. and Polak, E. (1969): *System Theory*. Inter-University Electronics Series **8.** McGraw Hill, New York.

[2] Zadeh, L. A. and Desoer, P. (1967): *Linear System Theory*. Series in System Science. McGraw Hill, New York.

## 1.5 Literature suggested for further reading

Benedek, P. and László, A. (1967): *Grundlagen der Chemie-Ingenieurwesen*. Verlag für Grundstoffindustrie, Leipzig.

Crowe, C. M., Hamielic, A. F., Hoffman, T. W., Johnson, A. I., Woods, D. R. and Shannon, P. T. (1971): *Chemical Plant Simulation, An Introduction to Computer Aided Steady State Process Analysis*. Prentice Hall, Englewood Cliffs, New Jersey.

Danzer, K., Than, E. and Molch, D. (1976): *Analytik—Systematischer Überblick*. Akad. Verlag, Leipzig.

Evans, L. B. (1980): *Chemical Process System Analysis*. MIT, Cambridge.

Kardos, L. et al. (1981): *Systemverfahrenstechnik*. I–II. Akademie Verlag, Leipzig.

Henley, E. J. and Rosen, E. M. (1978): *Material and Energy Balance Computations*. Wiley, New York.

Himmelblau, D. M. and Bischoff, K. B. (1965): *Process Analysis and Simulation*. Deterministic Systems. Wiley, New York.

Himmelblau, D. M. (1978): *Fault Detection and Diagnosis in Chemical and Petrochemical Processes*. Elsevier, Amsterdam.

Kafarov, V. (1976): *Cybernetic Methods in Chemistry and Chemical Engineering*. MIR Publishers, Moscow.

Kafarov, V. (1985): *Wasteless Chemical Processing*. MIR Publishers, Moscow.

Levenspiel, O. (1972): *Chemical Reaction Engineering*. Wiley, New York.

Rose, L. M. (1974): *The Application of Mathematical Modelling to Process Development and Design*. Applied Science Publishers, London.

Rose, L. M. (1985): *Chemical Reactor Design in Practice*. Elsevier, Amsterdam.

Rudd, D. F., Powers, G. J. and Siirola, J. J. (1973): *Process Synthesis*. Prentice-Hall, Englewood Cliffs, New Jersey.

Seinfeld, J. H. and Lapidus, L. (1974): *Mathematical Methods in Chemical Engineering*. Vol. 3. Prentice-Hall, Englewood Cliffs, New Jersey.

Westerberg, A. W., Hutchison, P., Motard, R. L. and Winter, P. (1978): *Process Flowsheeting*. Cambridge University Press, Cambridge.

# 2. MODELLING AND REGRESSION

by

S. VAJDA and P. VALKÓ

## 2.1 Modelling and its uses

The word 'model' or 'mathematical model' in this chapter will mean a system of algebraic and/or differential equations, which are held to be a representation of a physical system. From this general point of view any relationship among different quantities is a model, and the principles of modelling are the principles of chemical engineering science. Process models are constructed by application of the three fundamental conservation principles: conservation of mass, momentum, and energy. Application of these conservation laws to the modelling or particular processes is discussed in many excellent works such as Bird, Stewart and Lightfoot [10], Aris [1, 2, 3], Russel and Denn [44], Luyben [35] and Denn [15].

The conservation principles alone are never sufficient to provide an adequate model, however. Additional constitutive relationships are required which are not derivable from the conservation principles, are unique to the particular system being considered, and in most cases, are directly or indirectly based on experimental observations. It should also be emphasized that models usually represent an essential compromise between the simplicity which is desirable for application and the complexity which is representative of the real physical phenomena. Units such as chemical reactors and distillation columns often contain complex flows of streams undergoing phase changes and large numbers of chemical reactions. Detailed descriptions of the phenomena will rarely be available, and experimental observations necessarily play an important role in formulating adequate, yet mathematically tractable models.

In this work we are concerned with the problem of linking models, based on an a priori knowledge of chemical engineering, with available experimental data. This stage of modelling primarily involves parameter estimation, i. e. adjusting the values of some unknown parameters in the model to give a best match to the observations.

In most cases we restrict consideration to some well-established parameter estimation methods included in the special program REPROCHE (Regression Programs for Chemical Engineers, 1986). After solving some relatively simple test examples demonstrating the power of the selected methods we present some complex case studies. These latter show that fitting models in practice to observations practically never involves a 'pure' parameter estimation problem which can be solved automatically in one go, even with a fairly good estimation program

available. In fact, parameter estimation in chemical engineering usually requires additional assumptions and good knowledge of the underlying processes, and one thus ends up with an engineering rather than a statistical problem.

Parameter estimation as a stage of mathematical modelling raises quite a few questions. Why should we want a model? What kind of model? What kind of prior knowledge and observations are necessary? How do we fit the model to our observations? How do we decide whether the model is any good? How can we measure the reliability and precision of the resulting model? If the estimation fails, how can we find a better model? The first questions will be answered fairly quickly in this section, while the remainder of the chapter is devoted to a partial answer of the others, discussing only methods and computational algorithms which perform well on a variety of problems, even if they fail to solve some other problems involving different assumptions on the statistical structure of errors.

### 2.1.1 Why should we want a model?

To make any sense, modelling must have some definitive purpose which, in turn determines our expectations at the stage of parameter estimation. The most telling reasons for modelling can be listed as follows.

**Scientific curiosity.** The classical aim of modelling is the satisfaction of scientific curiosity, thus increasing our understanding of some mechanism by testing how the ideas work and how they are related to observations. Assumptions resulting in a model that is incapable of fitting good data with reasonable values of the parameters are at least questionable, and thus model adequacy is of primary concern. In most cases, however, the derived parameters are of fundamental value in the theory (e. g., rate constants in reaction kinetics) and hence their accuracy is important. The fact that we have a well-specified model is certainly advantageous. Unfortunately, the observations are frequently insufficient to uniquely confirm or deny the validity of the model and to estimate its parameters with reasonable accuracy, and the estimation problem may become very difficult.

*Prediction 1: Information reduction.* A desire for prediction is a powerful motive for modelling, common to a number of applications. We frequently want a model simply to avoid interpolation in extensive tables and to store and use less numerical data. From a utilitarian viewpoint, a predictive model should be judged solely on the accuracy of its predictions. The plausibility and simplicity of the prediction model and its power to provide insight are all incidental, although they help in model construction and acceptance. Thermodynamic state equations are good examples to emphasize this last point.

*Prediction 2: Indirect measurement.* In chemical analysis we frequently observe functions of the quantity of interest, e. g., pH or extinction, rather than concentration. Calibration curves are genuine mathematical models, and their construction has received due attention in the statistical literature.

*Prediction 3*: *Design.* To keep the design procedure tractable, the model should be simple, and high accuracy is not of prime concern. It is, however, important to know the variability of the prediction and its sensitivity to the design variables, so that the system may be designed to have an acceptable worst-case performance.

*Prediction 4*: *Optimization.* For a long time the major task of modelling in chemical engineering was that of finding the optimum operating conditions. A lot of experience has been gained by solving such static optimization problems, contributing to the development of flowsheeting methodology.

Although the models used were usually more complex than those used in design, progress was often hampered by the lack of really adequate models. It has also become clear that it is not completely sufficient to view the problems simply as static ones, and dynamic models are thus needed. This gave rise to the field of chemical process control.

*Prediction 5*: *Control.* Prediction by a dynamical model is also important in control-system design. Design of any control scheme more ambitious than the traditional trial-and-error tuned two-term controller requires a model. There is, however, a considerable conflict of views in this area. On one hand, detailed dynamic models have been formulated for the most important chemical processes; see, e. g., Aris [3] and Denn [15]. On the other hand, control engineers seem to pay little attention to these models. They argue that control theory can handle only simple linear models, and an accurate model is not a realistic goal due to variability in the system to be controlled. In fact, most computer controlled systems are based on the use of low-order linear difference equations as models, with parameters and even the structure estimated on-line by well developed identification methods; see Åström and Wittenmark [4]. Perhaps the future will see more interactions between the two approaches.

*Prediction 6*: *Simulation for teaching and operator training.* While detailed models of chemical processes are not really exploited in control applications, they are valuable for 'what if' analysis in teaching and operator training. While accuracy and completeness may be less crucial when qualitative outcomes are being explored rather than precise numerical consequences, the model should reflect the essentials of behaviour over a very large domain of input variables and parameters. The continuously updated simple models used in control applications are clearly unsuitable here.

**Diagnosis of faults and inadequacies.** A great benefit of modelling, now acknowledged in chemical engineering, is its ability to uncover shortcomings and anomalies in measurements. For example, when an attempt is made to write out the balance equations, the measurements are often found to be inadequate. In these applications we look for deviations from the predicted behaviour, though we rarely formulate complete prediction models.

### 2.1.2 What kind of model?

Models in general can be classified in many ways. There are verbal, physical, graphical and mathematical models, and it is not possible to separate these aspects entirely. Here we confine our attention to mathematical models as sets of equations, ignoring any background images, and list some basic dichotomies important forparameter estimation.

**Deterministic versus stochastic.** When formulating a model on the basis of conservation principles and constitutive relations, we usually assume that the input, initial condition and parameters of the model determine its response $y$, and thus the approach is a deterministic one. In fact, there are unmodelled disturbances in both the input and the system behaviour. This is a prime concern in control, but is rarely taken into account in other applications. To make model fitting meaningful, we may, however, never neglect the fact that observations of the response are corrupted by noise. For simplicity, the observed quantity is usually regarded as $\tilde{y}=y+\varepsilon$, where $\varepsilon$ is a random variable representing measurement error. Therefore, models for parameter estimation become stochastic, in spite of the deterministic approach, originally adopted, and further assumptions on the statistical properties of $\varepsilon$ are needed.

**Algebraic versus dynamic.** The simplest model of a system is the algebraic expression

$$\underline{y}=\underline{f}(\underline{x}, \underline{p}) \tag{2.1}$$

which is assumed to predict the dependent variables $\underline{y}=(y_1, \ldots, y_{ny})^T$ in terms of the independent variables $\underline{x}=(x_1, \ldots, x_{nx})^T$ and parameters $\underline{p}=(p_1, \ldots, p_{np})^T$. Though one of the components of $\underline{x}$ may be the time $t$ and hence Eq. (2.1) may predict the evolution of the system with time, we call a model dynamic if it is given by ordinary or partial differential equations.

**Lumped versus distributed.** In case of a lumped dynamic model the only independent variable is $t$, and the model is given by the ordinary differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\underline{x}=\underline{g}(\underline{x}, t, \underline{p}), \qquad \underline{x}(0, \underline{p})=\underline{x}^0(\underline{p}) \tag{2.2}$$

and by some observation function

$$\underline{y}=\underline{h}(\underline{x}, t, \underline{p}) \tag{2.3}$$

where $\underline{x}=(x_1, \ldots, x_{nx})^T$ and $\underline{y}=(y_1, \ldots, y_{ny})^T$.

In chemical engineering applications, models frequently involve observed quantities only, and hence $\underline{y}=\underline{x}$. On the other hand, a distributed system depends on some space coordinates in addition to time, and its model is a set of partial

differential equations. Such models are needed to describe many chemical processes. Their solution, however, frequently involves elimination of all but one independent variables by lumping or discrete approximations, and hence we restrict our considerations to lumped dynamic models given by Eqs (2.2) and (2.3).

**Linear versus nonlinear.** The algebraic model (2.1) is linear if

$$y = \sum_{i=1}^{np} p_i x_i \tag{2.4}$$

Under some further assumptions on the observation errors, parameters of the linear model (2.4) can be estimated linear regression, which is a very simple procedure. Furthermore, exact statistical interpretation of the estimates can be given only for linear models. It is, however, sufficient to require linearity in the parameter, and linear regression applies to models of the form

$$y = \sum_{i=1}^{np} p_i f_i(x) \tag{2.5}$$

where $f_1, \ldots, f_{np}$ are known functions of $x$. Linearity of $f$ and $h$ in Eqs (2.2) and (2.3) confers a similar advantage only if special estimation methods are used, one of which will be discussed in this chapter.

## 2.2 Parameter estimation

Whereas linear regression is a well developed area of mathematical statistics (see Kendall and Stuart [33], Hocking [28]), nonlinear parameter estimation requires a practical rather than a theoretical approach. As is emphasized in the basic reference book by Bard [6], nonlinear parameter estimation uses mainly a variety of computational algorithms which perform well on a class of problems but which may fail on some other problems. In addition, most statistical tests and estimates of variability apply exactly only to linear models and in the nonlinear case we frequently cannot do anything but apply these linear results as approximations. However, chemical engineering deals almost exclusively with nonlinear models and cannot afford to throw away useful tools just because they are not perfect.

Such a practical point of view is explicitly adopted in this chapter. Though, when listing the steps in the performance of parameter estimation we briefly discuss the variety of problems and different approaches, our considerations will be restricted to the simple least squares method. Furthermore, the only computational method discussed here is the classical Gauss–Newton–Marquardt minimization algorithm, the efficiency and robustness of which have been proved in many applications. Though most of our examples simply show how the methods work, we also try to demonstrate that the more advanced approaches, requiring fewer assumptions but also more calculations, do not necessarily yield better results.

### 2.2.1 Steps of parameter estimation

**Model selection.** While the model is correct only in rare situations, it should be at least tentatively appropriate and it should be validated simultaneously with the estimation of its parameters. It is, however, emphasized that models firmly based on the principles of chemical engineering are not necessarily good models for fitting, since the available data may be insufficient for estimating the parameters with any reasonable accuracy. In such cases a simplified model with fewer free parameters may be superior. From the point of view of the estimation procedure the simplest model reduction procedure is the fixing of the values of some parameters on the basis of *a priori* assumptions. This problem is frequently encountered in scientific applications, and we say that the original model is overparametrized (see Jennrich and Sampson [31]).

**Error structure selection.** As we have remarked estimating the parameters of the model also implies estimating the unknown measurement errors. A meaningful solution of this problem requires assumptions on the statistical properties of these errors, hence the error structure is an important part of any model for parameter estimation. A relatively simple parameter estimation problem is formulated if we consider a single-response algebraic model, thus $ny=1$ in Eq. (2.1). With additional error the observed response $y_i$ in the $i$th sample point is given by

$$\tilde{y}_i = f(\underline{x}_i, \underline{p}) + \varepsilon_i \tag{2.6}$$

and we assume that

(a) the independent variables $x_i$ are error-free;
(b) the error $\varepsilon_i$ has zero mean, thus $E\{\varepsilon_i\}=0$;
(c) the variance of $\varepsilon_i$, $D^2\{\varepsilon_i\}=\sigma_i^2$, is known at least up to a scalar factor;
(d) the error $\varepsilon_i$ is independent of $x_i$;
(e) the errors $\varepsilon_i$ and $\varepsilon_j$, $i \neq j$, are independent.

A zero mean of $\varepsilon_i$ implies that there is no systematic error in the measurement and the model is correct. This is a very general assumption, since a nonzero mean $\bar{\varepsilon}_i = E\{\varepsilon_i\}$ cannot be separated from the deterministic response $f(\underline{x}_i, \underline{p})$ of the system. As we shall see, assumption of known variances gives weighting coefficients representing the accuracy, and hence the importance, of each observation when estimating the parameters.

Some of the assumptions may be relaxed, but then it is more difficult to solve the estimation problem. For example, allowing for errors in $x$ implies that we also have to estimate the errors in these variables by so-called error-in-variables estimation methods. Similarly, we can relax conditions (c) and (e), and estimate the covariance matrix simultaneously with the parameters. This clearly means that we introduce additional parameters (i. e. elements of the covariance matrix) and hence the problem is more difficult to solve. This latter approach is called multiresponse estimation. We will show that in spite of its theoretical superiority, considerable difficulties can be faced when applying it.

**Selection of the estimation criterion.** The estimation criterion is determined by the assumptions made concerning the error structure. In case of the model just discussed, a natural goal is to minimize the norm

$$Q(p) = \sum_{i=1}^{nm} \varepsilon_i^2 w_i = \sum_{i=1}^{nm} [\tilde{y}_i - f(x_i, p)]^2 w_i \tag{2.7}$$

where $nm$ is the sample size and $w_i$ represents the weighting coefficients. This is the method of least squares.

The least squares method has a number of advantageous properties. First, for a linear model (2.4) or (2.5) with weighting coefficients $w_i = 1/\sigma_i^2$ the estimates obtained are unbiased, and thus their expected values are equal to the true parameters. Second, these estimates have the minimum variance within the class of all unbiased estimates (Kendall and Stuart [33]). Although in the nonlinear case these properties hold only assymptotically, thus increasing the sample size unboundedly (Jennrich [30]), the method produces acceptable estimates in many situations (Bard [6]). Third, assuming that $\varepsilon_i$ is a normally distributed random variable in addition to conditions (a) to (d), the important maximum likelihood principle of statistical estimation theory is reduced to the least squares estimation criterion.

The assumption of normally distributed errors is not a necessary condition for meaningful application of the least squares method. Any violation of conditions (a) to (e) means, however, that the method is not being correctly applied and at least, a different estimation criterion should be used, as noted in the previous point.

**Parameter estimation.** In the strict sense parameter estimation is the procedure of calculating the estimates by localizing the extremum point of the objective function. The least squares estimation is very simple for models that are linear in their parameters, since it can be performed by linear regression in a single step. In the nonlinear case iterative minimization algorithms are required and the calculations involved are far from simple.

**Goodness of fit.** As has been discussed above parameter estimation involves a number of assumptions, which should be validated *a posteriori*, after having thus computed the estimates $\hat{p}$. The analysis is mainly based on the residuals $r_i = \tilde{y}_i - \hat{y}_i$, where $\hat{y}_i$ is the predicted response $\hat{y}_i = f(\tilde{x}_i, \hat{p})$. If the residuals are so large or so nonrandom in nature that they cannot be ascribed to random observation errors, then this constitutes a strong reason for rejecting one of the assumptions. The analysis of residuals is completed through several statistical procedures such as calculation of their variance and serial correlation.

**Interpretation of the estimates.** It is not enough to compute the estimates $\hat{p}$ of the parameters, we must also examine their reliability and precision. The covariance matrix of $\underline{\hat{p}}$ can be estimated on the basis of the assumption (a) to (e). In order to

answer such questions as 'what are the chances that $\hat{p}$ deviates from the true value $p$ by no more than 1%', and thus to compute confidence limits, we need, however, the additional assumption of normally distributed errors. It is also emphasized that most statistical tests and estimates of variability apply only approximately to nonlinear models, and even for linear models they are exact only if the errors do indeed follow whatever distribution was assumed for them. Nevertheless, even these approximate results provide some feeling of the uncertainties associated with the estimates.

**Simulation.** Since parameter estimation and interpretation of the estimates is far from simple, it is useful to see how the selected procedure solves problems with *a priori* known results, involving data generated on the computer. Initially it is advisable to experiment with error-free data, and then to add errors having the assumed structure and to check the variability of the estimates by performing several runs.

### 2.2.2 Linear algebraic model: linear regression

We fit model (2.4) to the set $\{(x_{i1}, \ldots, x_{i,\,nx}, \tilde{y}_i),\ i=1, 2, \ldots, nm\}$ of observations, where

$$\tilde{y}_i = p_1 x_{i1} + p_2 x_{i2} + \ldots + p_{nx} x_{i,\,nx} + \varepsilon_i \tag{2.8}$$

Under the assumptions (a) to (e) the least squares method will be applied with the weighting coefficient satisfying

$$D^2\{\varepsilon_i\} = \sigma_i^2 = \sigma^2 / w_i \tag{2.9}$$

where $\sigma^2$ is an unknown scalar. Introducing the notations

$$\underline{Y} = \begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_{nm} \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1,\,nx} \\ x_{21} & x_{22} & \ldots & x_{2,\,nx} \\ \vdots & & & \\ x_{nm,\,1} & x_{nm,\,2} & \ldots & x_{nm,\,nx} \end{bmatrix} \tag{2.10}$$

and

$$\underline{\varepsilon}^{\mathrm{T}} = (\varepsilon_1, \ldots, \varepsilon_{nm}), \qquad \underline{p}^{\mathrm{T}} = (p_1, \ldots, p_{nx}) \tag{2.11}$$

Equation (2.8) is reduced to

$$\underline{Y} = \mathbf{X}\underline{p} + \underline{\varepsilon} \tag{2.12}$$

whereas the sum of squares objective function, Eq.(2.7), is given by

$$Q(\underline{p}) = (\mathbf{Y} - \mathbf{X}\underline{p})^{\mathrm{T}}\, \mathbf{W} (\mathbf{Y} - \mathbf{X}\underline{p}) \tag{2.13}$$

where $\mathbf{W}$ is an $nm \times nm$ diagonal matrix with nonzero entries $w_1, \ldots, w_{nm}$. The least squares estimate $p$ is the minimum point of Eq.(2.13) and hence the solution

$$\hat{\underline{p}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \underline{Y} \tag{2.14}$$

of the linear equations $\partial Q(p)/\partial p_i = 0$, $i = 1, 2, \ldots, nx$.

The goodness of the fit is measured by the residual sum of squares $Q(\hat{\underline{p}})$ and the residual variance

$$s_r^2 = Q(\hat{\underline{p}})/(nm - nx) \tag{2.15}$$

Under our conditions $s_r^2$ is the unbiased estimate of the factor $\sigma^2$ in Eq.(2.9). If we possess another estimate $s^2$ of $\sigma^2$ (e.g., from replicates), than the $F$-test involving the ratio $F = s_r^2/s^2$ can be used to check the validity of the model.

It is sometimes advantageous to measure the goodness of fit on a 0–1 scale. For this purpose the multiple correlation coefficient $R^2$ is defined as

$$R^2 = 1 - \frac{Q(\hat{\underline{p}})}{\underline{Y}^T \mathbf{W} \underline{Y}} \tag{2.16}$$

if none of the columns of matrix $\mathbf{X}$ is constant. In case of a constant column the definition

$$R^2 = 1 - \frac{Q(\hat{\underline{p}})}{(\underline{Y}^c)^T \mathbf{W} \underline{Y}^c} \tag{2.17}$$

is used, where $\underline{Y}^c$ is the vector of centred responses, i.e. the vector of deviations of the observations from their weighted arithmetic mean.

In the linear case the least squares estimates are unbiased, thus $E\{\hat{\underline{p}}\} = \underline{p}$, whereas the covariance matrix $E\{(\hat{\underline{p}} - \underline{p})(\hat{\underline{p}} - \underline{p})^T\}$ of $p$ is estimated by

$$\mathbf{C}_{\hat{\underline{p}}} = s_r^2 (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \tag{2.18}$$

The least squares estimator is the best linear unbiased estimator in the sense that other estimators of the same kind cannot yield a better covariance matrix of the parameters. The covariance matrix $C_{\hat{\underline{p}}}^{(1)}$ is considered to be better than $\mathbf{C}_{\hat{\underline{p}}}^{(2)}$ if the matrix $\mathbf{C}_{\hat{\underline{p}}}^{(2)} - \mathbf{C}_{\hat{\underline{p}}}^{(1)}$ is positive definite.

Statistical dependence between $\hat{p}_i$ and $\hat{p}_j$ is measured in term of the correlation coefficient

$$r_{ij} = [C_{\hat{\underline{p}}}]_{ij} / ([C_{\hat{\underline{p}}}]_{ii} [C_{\hat{\underline{p}}}]_{jj})^{1/2} \tag{2.19}$$

The diagonal entry $[C_{\hat{\underline{p}}}]_{ii}$ estimates the variance $s_{\hat{p}_i^2}$ of $\hat{p}_i$ and the confidence interval for $p_i$ is given by

$$\hat{p}_i - s_{\hat{p}_i} t_{\alpha,\, nm-nx} \leq p_i \leq \hat{p}_i + \hat{s}_{\hat{p}_i} t_{\alpha,\, nm-nx} \tag{2.20}$$

where $t_{\alpha,\, n}$ is value of the $t$-distribution with $n$ degrees of freedom at the probability $\alpha$. When the parameters are strongly correlated, however, the individual confidence intervals are rather approximate and hence it is better to evaluate the joint confidence region, given by

$$(\underline{p} - \hat{\underline{p}})^T \mathbf{C}_{\hat{\underline{p}}}^{-1} (\underline{p} - \hat{\underline{p}}) \leq \chi^2_{\alpha,\, nx} \tag{2.21}$$

where $\chi^2_{\alpha,\,nx}$ denotes the value of the $\chi^2$ distribution with $nx$ degrees of freedom at the probability $\alpha$. The confidence region is an $nx$-dimensional ellipsoid around $\hat{p}$ which contains $p$ in $100(1-\alpha)\%$ of all possible data samples.

As discussed by Bard [6], the principal axes of the confidence ellipsoid are along the eigenvectors $\underline{u}_1, \ldots, \underline{u}_{nx}$ of the matrix $\mathbf{X}^T\mathbf{W}\mathbf{X}$ and the lengths of the axes along $\underline{u}_i$ are proportional to $\lambda_i^{-1/2}$, where $\lambda_i$ denotes the $i$th eigenvalue of the matrix. Let us introduce the parameters called principal component defined by $\psi_i = \underline{u}_i^T p$. Then, in terms of $\Delta\psi_i = \underline{u}_i^T(\underline{p} - \underline{\hat{p}})$, Eq.(2.21) is expressed in the canonical form

$$\frac{1}{\sigma^2}\sum_{i=1}^{nx}\lambda_i(\Delta\psi_i)^2 \leq \chi^2_{\alpha,\,nx} \tag{2.22}$$

If the matrix $\mathbf{X}^T\mathbf{W}\mathbf{X}$ is nearly singular, there thus exists an eigenvalue $\lambda_i \approx 0$, and the corresponding principal component $\psi_i$ (i.e., a linear combination of the original parameters) is estimated with large variance. Near-singularity is an inherent difficulty in many parameter estimation problems, and hence principal component analysis of the $\mathbf{X}^T\mathbf{W}\mathbf{X}$ matrix is used as a routine tool to investigate its sources.

Recent linear statistical program packages usually apply singular value decomposition *(SVD)* of the matrix $\mathbf{X}\mathbf{W}^{1/2}$:

$$\mathbf{X}\mathbf{W}^{1/2} = \mathbf{H}\boldsymbol{\Lambda}^{1/2}\mathbf{G}^T \tag{2.23}$$

where $\mathbf{H}$ is an $nm \times nx$ matrix with the property $\mathbf{H}^T\mathbf{H} = \mathbf{I}$, the $nx \times nx$ diagonal matrix $\boldsymbol{\Lambda}^{1/2}$ contains the square roots of the eigenvalues $\lambda_i$ (here called singular values) and the columns of the $nx \times nx$ matrix $\mathbf{G}$ are the eigenvectors $u_1, \ldots, u_{nx}$ (Golub [23]). In terms of *SVD* we can write Eq.(2.14) as

$$\underline{\hat{p}} = \mathbf{G}\boldsymbol{\Lambda}^{-1/2}\mathbf{H}^T\underline{Y} = \mathbf{G}\underline{c} = \sum_{i=1}^{nx}\underline{u}_i c_i \tag{2.24}$$

and Eq.(2.18) as

$$\mathbf{C}_{\hat{p}} = s_r^2\mathbf{G}\boldsymbol{\Lambda}^{-1}\mathbf{G}^T = s_r^2\sum_{i=1}^{nx}\lambda_i^{-1}\underline{u}_i\underline{u}_i^T \tag{2.25}$$

*SVD* gives the same estimates as the traditional approach but it has better numerical properties. Moreover, it automatically supplies the additional information otherwise obtained from the principal component decomposition of the $\mathbf{X}^T\mathbf{W}\mathbf{X}$ matrix

**Example 2.1.** *Decomposition of the rate constant of an acid-catalyzed reaction by linear regression*

The hydrolysis

$$CH_3C(OC_2H_5)_3 + H_2O \rightarrow CH_3COOC_2H_5 + 2C_2H_5OH$$

of orthoacetic acid ethyl ester is a typical acid-catalyzed reaction. As has been shown by Schwetlick [45], when performing the reaction in the presence of the weak acid $NO_2C_6H_4OH$ and at constant ionic strength, the rate constant $k$ of the process can be decomposed as

$$k = k_0 + k_H[H^+] + k_{HA}[HA] \tag{2.26}$$

where $k_0$ is the rate constant of the uncatalyzed reaction, whereas $k_H$ and $k_{HA}$ are the catalysis constants representing the effects brought about by $[H^+]$ and the concentration [HA] of the undissociated acid. In our case HA is $NO_2C_6H_4OH$. *Table 2.1*, originally published by Brönsted and Wynne-Jones [11] lists rate constants observed at different values of $[H^+]$ and $[NO_2C_6H_4OH]$.

We estimated the parameters $k_0$, $k_H$ and $k_{HA}$ by the unweighted least squares method, thus $w_i = 1$. The results are shown in *Table 2.2*. The standard error $s_r$ is approximately 2% of the observed response, which is certainly comparable with the error of the experiment. Thus the fit is acceptable, though there is some nonrandomness in the residuals.

Although the confidence intervals of the parameters are reasonably small, at 95% the confidence interval of $k_0$ contains the value $k_0 = 0$ and hence the hypotheses $k_0 = 0$ cannot be rejected at the 95% significance level. Furthermore, the estimate $\hat{k}_0 = 3.43\text{E-}6$ is almost within the range of the assumed 2% measurement error, and hence the simplified model $k = k_H[H^+] + k_{HA}[HA]$ is also acceptable from a practical point of view. Fitting this latter model gives the standard error $s_r = 3.222\text{E-}6$, thus the goodness of the fit is practically unchanged, which can be supported by applying the $F$-test. Therefore, the rate constant $k_0$ of the uncatalyzed reaction can be neglected under the experimental conditions considered. The parameters derived from the simplified model with their standard errors (i.e., $s_{\hat{p}_i}$) are $\hat{k}_H = 2.28\text{E}4 \pm 2.3\text{E}2$ and $\hat{k}_{HA} = 1.61\text{E-}3 \pm 1.1\text{E-}4$.

Table 2.1. Rate constant of acid-catalyzed reaction

| $[H^+]$, mol/l | [HA], mol/l | $k$, $s^{-1}$ |
|---|---|---|
| 4.8 E−9 | 2.42 E−3 | 1.21 E−4 |
| 4.8 E−9 | 5.66 E−3 | 1.20 E−4 |
| 4.8 E−9 | 1.60 E−2 | 1.35 E−4 |
| 4.8 E−9 | 2.121 E−2 | 1.44 E−4 |
| 6.5 E−9 | 3.84 E−3 | 1.54 E−4 |
| 6.5 E−9 | 1.025 E−2 | 1.61 E−4 |
| 6.5 E−9 | 1.83 E−2 | 1.77 E−4 |
| 1.02 E−8 | 3.1 E−3 | 2.37 E−4 |
| 1.02 E−8 | 1.03 E−2 | 2.47 E−4 |
| 1.02 E−8 | 3.09 E−2 | 2.84 E−4 |

Table 2.2. Results of Example 2.1

| | |
|---|---|
| Sum of squares of residuals $Q(\hat{p})$ | 7.252E−11 |
| Standard error $s_r$ | 3.219E−6 |
| Multiple correlation coefficient $R^2$ (2.16) | 0.99758 |
| $t_{0.05,7}$ (tabulated value) | 2.373 |

| Parameter | Estimate | $s_{\hat{p}_i}$ | 95% confidence interval | |
|---|---|---|---|---|
| $k_0$ | 3.434E−6 | 3.41E−6 | (−4.65E−6, | 1.15E−5) |
| $k_H$ | 2.240E−4 | 4.59E−2 | (2.13E−4, | 2.35E−4) |
| $k_{HA}$ | 1.584E−3 | 1.17E−4 | (1.31E−3, | 1.86E−3) |

*Correlation matrix of the parameter estimates*

| | $k_0$ | $k_H$ | $k_{HA}$ |
|---|---|---|---|
| $k_0$ | 1 | −0.861 | −0.257 |
| $k_H$ | | 1 | −0.173 |
| $k_{HA}$ | | | 1 |

This example is also suitable to illustrate some problems connected with principal component analysis. The eigenvalue-eigenvector decomposition of the $\mathbf{X}^T\mathbf{X}$ matrix is shown in *Table 2.3(a)*. The eigenvalue $\lambda_3 = 5.077\text{E}{-17}$ is very small, but this is due to the low order of magnitude of column 2 of matrix $\mathbf{X}$. Obviously, a more realistic picture of the problem can be obtained if the columns of matrix $\mathbf{X}$ are scaled (multipled by suitable positive numbers). Scaling does not affect the final estimates if the results are correctly rescaled.

A straightforward way of scaling is to multiply the columns of matrix $\mathbf{X}$ to yield the unit length. The eigenvalue-eigenvector decomposition of the resulting problem is shown in *Table 2.3(b)*. As may be seen, the problem is not singular.

Another type of scaling is based on the relative importance of the variable in the fit. This can be done after the estimates have been obtained.

The $i$th column is multiplied by the $i$th parameter estimate. Each scaled regression coefficient will be exactly unity. The eigenvalue-eigenvector decomposition of the problem scaled in this way is shown in *Table 2.3(c)*. The smaller eigenvalue is $\lambda_3 = 1.054\text{E}{-11}$. It is comparable with the variance of the residuals $s_r^2 = 1.036\text{E}{-11}$. Moreover, the only parameter responsible for this small eigenvalue is $k_0$, as can be seen from the corresponding eigenvector. Now we have an additional argument for dropping $k_0$ from the model. When discussing model reduction we will give further motives for using this latter type of scaling.

The traditional least squares approach regards unbiasedness as a key concept. As has been discussed by Vinod and Ullah [56], many investigators have considered biased estimators to be a better alternative. Biased estimators are useful if the

Table 2.3. Principal component analysis of the $\mathbf{X}^T\mathbf{X}$ matrix in Example 2.1

*(a) X not scaled*

| $i$ | $\lambda_i$ | Eigenvector component corresponding to | | |
|---|---|---|---|---|
| | | $k_0$ | $k_H$ | $k_{HA}$ |
| 1 | 1.000E−1 | 1.000 | 0.000 | 0.012 |
| 2 | 7.809E−4 | −0.012 | 0.000 | 1.000 |
| 3 | 5.077E−17 | 0.000 | 1.000 | 0.000 |

*(b) X scaled to yield unit column length*

| $i$ | $\lambda_i$ | Eigenvector component corresponding to | | |
|---|---|---|---|---|
| | | $k_0$ | $k_H$ | $k_{HA}$ |
| 1 | 2.710 | 0.589 | 0.589 | 0.554 |
| 2 | 2.410E−1 | −0.374 | −0.410 | 0.832 |
| 3 | 4.889E−2 | 0.716 | −0.698 | −0.022 |

*(c) X scaled to yield unit regression coefficients*

| $i$ | $\lambda_i$ | Eigenvector component corresponding to | | |
|---|---|---|---|---|
| | | $k_0$ | $k_H$ | $k_{HA}$ |
| 1 | 2.703E−7 | 0.020 | 0.993 | 0.117 |
| 2 | 2.007E−9 | 0.017 | −0.118 | 0.993 |
| 3 | 1.054E−11 | 1.000 | −0.018 | −0.019 |

problem is nearly singular, and hence some parameter estimates have very large variances. As a consequence, a particular realization of the unbiased estimate can be very far from the true parameter. On the other hand, a realization of the biased estimate with considerably smaller variance lies nearer to the true value, as is shown in *Fig. 2.1.*

Two biased estimators are commonly used: the principal component estimator and the ridge regression estimator.

The principal component estimator (Kendall [32]) deletes the terms from the sum (2.24) corresponding to the small eigenvalues $\lambda_i$ (consequently to the large coefficients $c_i$). Roughly speaking, the parameters are estimated only in a subspace of the parameter space where sufficient information is available. In this subspace there is an infinite number of solutions to the problem. The principal component estimator is the only one of these having the additional property of minimum length. The basic difficulty in applying the principal component estimator is determining which eigenvalues can be considered to be small in the given problem.

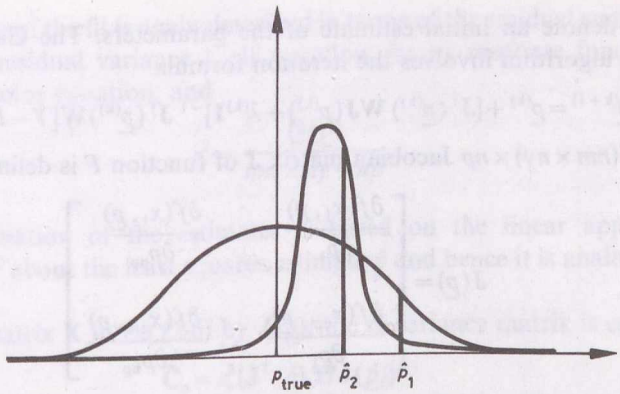



Fig. 2.1. Unbiased estimate with large variance ($\hat{p}_1$) and biased estimate with small variance ($\hat{p}_2$)

The ridge regression estimator of Hoerl and Kennard [27] is given by

$$\underline{p}(k)=(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{Y} \tag{2.27}$$

where $0\leq k\leq\infty$ is the ridge parameter. The ridge parameter should be selected on the basis of a compromise: it should be large enough to decrease the variance of the estimate but small enough to keep the bias below a reasonable $k$ for a given problem. Unfortunately, none of them is general enough. Ridge regression is also used as a graphical tool in the analysis of the stability of the estimates. For this purpose $\hat{p}(k)$ is plotted against $k$. This so called Ridge Trace reveals the heavily varying (i.e. unstable) components of the parameter vector. This information can be used either to select a good compromise value of $k$ or to reduce the model.

### 2.2.3 Nonlinear algebraic model: the Gauss–Marquardt algorithm

If the algebraic model, Eq.(2.1) is nonlinear in its parameters, a search algorithm is required to localize the minimum of the sum of squares objective function

$$Q(\underline{p})=[\underline{Y}-\underline{F}(\underline{p})]^{\mathrm{T}}\,\mathbf{W}[\underline{Y}-\underline{F}(\mathbf{p})] \tag{2.28}$$

where

$$\underline{Y}=\begin{bmatrix}\tilde{y}_1\\ \tilde{y}_2\\ \vdots\\ \tilde{y}_{\mathrm{nm}}\end{bmatrix}\qquad \underline{F}(\underline{p})=\begin{bmatrix}f(x_1,\underline{p})\\ f(x_2,\underline{p})\\ \vdots\\ f(x_{\mathrm{nm}},\underline{p})\end{bmatrix} \tag{2.29}$$

and $\mathbf{W}$ is the diagonal matrix of weight coefficients with entries defined by Eq.(2.9). We assume that assumptions (a) to (e) apply.

Let $\underline{p}^{(0)}$ denote an initial estimate of the parameters. The Gauss–Marquardt estimation algorithm involves the iteration formula

$$\underline{p}^{(k+1)}=\underline{p}^{(k)}+[\mathbf{J}^{T}(\underline{p}^{(k)})\,\mathbf{W}\mathbf{J}(\underline{p}^{(k)})+\lambda^{(k)}\mathbf{I}]^{-1}\,\mathbf{J}^{T}(\underline{p}^{(k)})\mathbf{W}[Y-F(\underline{p}^{(k)})] \tag{2.30}$$

where the $(nm\times ny)\times np$ Jacobian matrix $\mathbf{J}$ of function $F$ is defined by

$$\mathbf{J}(\underline{p})=\begin{bmatrix} \dfrac{\partial f(x_1,\underline{p})}{\partial p_1} & \cdots & \dfrac{\partial f(x_1,\underline{p})}{\partial p_{np}} \\ \vdots & & \\ \dfrac{\partial f(x_{nm},\underline{p})}{\partial p_1} & \cdots & \dfrac{\partial f(x_{nm},\underline{p})}{\partial p_{np}} \end{bmatrix} \tag{2.31}$$

and $\lambda^{(k)}$ is a scalar aimed at avoiding the near-singularity of the matrix to be inverted. At $\lambda^{(k)}=0$ each step of the iteration is a linear regression minimizing the error norm $\underline{\varepsilon}^{T}\mathbf{W}\underline{\varepsilon}$ in the expression

$$\underline{Y}-\underline{F}(\underline{p}^{(k)})=\mathbf{J}(\underline{p}^{(k)})\,(\underline{p}^{(k+1)}-\underline{p}^{(k)})+\varepsilon \tag{2.32}$$

based on the local linear approximation of $F$ at $p^{k}$. This may clearly be seen by comparing Eqs (2.14) and (2.30).

A variety of different rules has been proposed to select $\lambda^{(k)}$ in subsequent iteration steps and accordingly there exist several Marquardt-type algorithms, see, e.g., Bard [6] and Himmelblau [25].

The numerical efficiency of each algorithm depends on the selection rule for $\lambda^{(k)}$ and on a number of additional details. In particular, the algorithm used in this section contains the following simple, but practically important modifications.

(i) In the $(k+1)$th iteration step we look for the minimum of Eq.(2.22) in the space of normalized parameters $\beta_j=p_j/p_j^{(k)}$, $j=1, 2, \ldots, np$, where $p^{(k)}$ is the estimate from the previous step. Therefore, the starting point in each iteration is given by $\beta_j=1$, and the entries in the Jacobian matrix $J(\beta)$ are computed by

$$\frac{\partial f(x_i,\beta)}{\partial \beta_j}=\frac{\partial f(x_i,p)}{\partial p_j}\,p_j^{(k)} \tag{2.33}$$

Although $\beta_j$ is not defined if $p_j^{(0)}=0$, by Eqs (2.30) and (2.27) this implies $p_j^{(k)}=0$ for all $k$ and hence we allow for zero initial estimates of the parameters.

(ii) The cross-product matrix $\mathbf{J}^{T}\mathbf{W}\mathbf{J}$ is normalized to a correlation-type matrix through transformations of the form (2.19), and this enables us to use the following simple rules to select $\lambda^{(k+1)}$.

(iii) Initially $\lambda^{(0)}=0.01$, and $\lambda^{(k+1)}=0.1\lambda^{(k)}$, if $Q(p^{(k+1)})<Q(p^{(k)})$; $\lambda^{(k+1)}=10\lambda^{(k)}$ otherwise.

(iv) Since the sign of the parameter is usually known, we keep the estimates non-negative. If $\beta_j^{(k+1)}=\beta_j^{(k)}+\Delta\beta_j^{(k)}<0$, then the step size is restricted to ensure positive $\beta_j^{(k+1)}$.

The goodness of the fit is again described in terms of the residual sum of squares $Q(\hat{p})$ and the residual variance $s_r^2$. If we allow for $ny$ response functions, thus Eq.(2.1) is a vector equation, and

$$s_r^2 = \frac{Q(\hat{p})}{nm \times ny - np} \tag{2.34}$$

The interpretation of the estimates is based on the linear approximation [Eq.(2.32)] of $F$ about the least squares estimate $\hat{p}$ and hence it is analogous to the linear case.

Replacing matrix $\mathbf{X}$ in Eq.(2.6) by $\mathbf{J}(\hat{p})$, the covariance matrix is estimated by

$$\hat{\mathbf{C}}_p = s_r^2 [\mathbf{J}^T(\hat{p})\, \mathbf{W} \mathbf{J}(\hat{p})]^{-1} \tag{2.35}$$

Similarly, based on the local linear approximation, the joint confidence region of the parameters is given by Eq.(2.21). It must, however, be emphasized that this region may substantially differ from the exact confidence region of the parameters, since the linear approximation is sufficiently good only in a small neighbourhood of $\hat{p}$. As has been shown by Bard [6], the exact confidence region is determined by the inequality $Q(p) - Q(\hat{p}) \leq \varepsilon$ with $\varepsilon$ depending on the selected probability level, thus it is given by contours of the objective function. For $np > 2$, however, this result is difficult to deal with. Though the ellipsoid, Eq.(2.21) is an approximation, its use is more convenient, since the shape of the ellipsoid can easily be described in terms of the eigenvalues and eigenvectors of $\mathbf{J}^T(\hat{p})\, \mathbf{W} \mathbf{J}(\hat{p})$, as was done in the linear case. Here we perform the analysis in the space of the transformed parameters $\beta$, and the entries of $\mathbf{J}(\hat{\beta})$ are given by Eq.(2.33). This transformation is very important, since a mechanistic interpretation can be given to certain forms of eigenvectors, thereby identifying the sources of near-singularity. This procedure will be described in detail in Section 2.3.

**Example 2.2.** *Fitting a nonlinear rate expression*

Rational functions of the form

$$y = p_1 + x/(p_2 v + p_3 w) \tag{2.36}$$

are frequently encountered as rate expressions of catalytic processes. Model (2.36) and the set of data listed in *Table 2.4* form a test example in nonlinear parameter estimation, first given by Bard [5], and since used by several authors (e.g., Garcia-Peña *et al.* [20]) to compare different methods. Results are shown in *Fig. 2.2** as part of the output of the REPROCHE program. The starting point of the search is $p^{(0)} = (1, 1, 1)^T$ and, although we use a finite-difference approximation of the Jacobian matrix $\mathbf{J}(p)$ in order to avoid the need for analytic differentiation, the final estimate is obtained in five iteration steps with the termination condition

* At the end of this chapter.

Table 2.4. Data for Bard's test problem

| $i$ | Independent variables | | | Dependent variable |
|---|---|---|---|---|
| | $x$ | $v$ | $w$ | $y$ |
| 1 | 1 | 15 | 1 | 0.14 |
| 2 | 2 | 14 | 2 | 0.18 |
| 3 | 3 | 13 | 3 | 0.22 |
| 4 | 4 | 12 | 4 | 0.25 |
| 5 | 5 | 11 | 5 | 0.29 |
| 6 | 6 | 10 | 6 | 0.32 |
| 7 | 7 | 9 | 7 | 0.35 |
| 8 | 8 | 8 | 8 | 0.39 |
| 9 | 9 | 7 | 7 | 0.37 |
| 10 | 10 | 6 | 6 | 0.58 |
| 11 | 11 | 5 | 5 | 0.73 |
| 12 | 12 | 4 | 4 | 0.96 |
| 13 | 13 | 3 | 3 | 1.34 |
| 14 | 14 | 2 | 2 | 2.10 |
| 15 | 15 | 1 | 1 | 4.39 |

$||\Delta p^{(k)}||/||p^{(k)}|| < 1.E{-}4$, where $||\cdot||$ denotes the Euclidean norm of a vector. According to this result, the minimization algorithm of the REPROCHE program is comparable with the best ones studied by Bard [5] and Garcia-Peña *et al.* [20] in terms of numerical efficiency. The plot of the residuals clearly shows one outlying observation.

Even with a good program, fitting nonlinear models usually requires considerable computational efforts, since iterative function minimization and convergence can be attained only from sufficiently good initial estimates. Therefore, any transformation replacing the model by a linear one is of practical value.

**Example 2.3.** *Arrhenius parameters by linearization and weighting*

A well-known problem is the estimation of parameters in the Arrhenius expression

$$k = A \exp[-E/(RT)] \tag{2.37}$$

describing the temperature dependence of the rate constant $k$, where $R = 8.3144$ J/(mol K). The pre-exponential factor $A$ and the energy of activation $E$ are usually estimated by fitting a straight line

$$y = \ln k = ax + b = -E\left(\frac{1}{RT}\right) + \ln A \tag{2.38}$$

to the transformed observations $(1/RT_i, \ln k_i)$. Fitting functions (2.37) and (2.38) results, however, in different estimates, since the logarithmic transformation

changes the distribution of measurement errors, and the original assumption on variance does not apply to the transformed model.

First we study this problem more generally and consider the model (2.6) with the error structure $E\{\varepsilon_i\}=0$, $E\{\varepsilon_i^2\}=\sigma_i^2$. As has been discussed, for asymptotic unbiasedness, the least squares method should be applied with the weighting coefficients $w_i=1/\sigma_i^2$. We apply a transformation, however, and fit the model $z=g(y)=$ $=g(f(x,p))$ to the transformed observations $\tilde{z}_i=g(\tilde{y}_i)$, $i=1,\ldots,nm$, where

$$\tilde{z}_i=g(\tilde{y}_i)=g(y_i)+\tilde{\varepsilon}_i \tag{2.39}$$

with some new error $\tilde{\varepsilon}_i$. To find the variance $\tilde{\sigma}_i^2$ of $\tilde{\varepsilon}_i$ note that from Eq.(2.31) $\varepsilon_i$ is defined by

$$\tilde{\varepsilon}_i=g(\tilde{y}_i)-g(\tilde{y}_i-\varepsilon_i) \tag{2.40}$$

Using the linear approximation $g(\tilde{y}_i-\varepsilon_i)\approx g(\tilde{y}_i)-g'(\tilde{y}_i)\varepsilon_i$, where $g'=\partial g/\partial y$, we obtain $\tilde{\varepsilon}_i=g'(\tilde{y}_i)\varepsilon_i$, and hence

$$\tilde{\sigma}_i^2=E\{\tilde{\varepsilon}^2\}\approx[g'(\tilde{y}_i)]^2\sigma_i^2 \tag{2.41}$$

Therefore, fitting the transformed model with the weighting coefficients $\tilde{w}_i=\tilde{\sigma}_i^{-2}=$ $=[g'(\tilde{y}_c)\sigma_i]^{-2}$ reduces the bias introduced by the transformation.

As an application of this general result, consider the rate constants of the reaction

$$CH_3I+C_2H_5ONa \rightarrow CH_3OC_2H_5+NaI$$

observed at six temperatures and listed in *Table 2.5* (Newling and Hinshelwood [37]), where $k$ is assumed to have identical variance in each observation, thus $w_i=1$. Since $g(k_i)=\ln(k_i)$, according to Eq.(2.41) after linearization by Eq.(2.38) it is advisable to use the weighting coefficients $\tilde{w}_i=\tilde{k}_i^2$. *Table 2.6* shows the estimates of the parameters by nonlinear least squares and by linear regression both with weights $\tilde{w}_i=1$ and $\tilde{w}_i=\tilde{k}_i^2$. The latter choice of weighting coefficients clearly reduces deviations between linear and nonlinear estimates. *Table 2.4* also shows limits of 95% confidence intervals, which are now not symmetric about the estimate due to the estimation of $\ln A$. Notice that nonlinear estimation gives very large confidence interval for $A$. This is obviously an overestimation, since altering the parameters within this interval causes the sum of squares objective function to

Table 2.5. Temperature dependence of rate constant $k$

| $i$ | $T$, K | $k$, 1/(mol s) |
|---|---|---|
| 1 | 273.16 | 5.6 E−5 |
| 2 | 279.16 | 1.18 E−4 |
| 3 | 285.16 | 2.45 E−4 |
| 4 | 291.16 | 4.88 E−4 |
| 5 | 297.16 | 1.00 E−3 |
| 6 | 303.16 | 2.08 E−3 |

Table 2.6. Estimates of the Arrhenius parameters for Example 2.3

| | Unweighted linear regression | Weighted linear regression | Nonlinear regression |
|---|---|---|---|
| $A$, 1/(mol s)<br>95% confidence interval | 3.25E11<br>(8.75E10, 1.21E12) | 3.11E12<br>(5.90E11, 1.63E13) | 3.42E12<br>(−2.28E12, 9.12E12) |
| $E$, J/mol<br>95% confidence interval | 8.25E4<br>(7.94E4, 8.57E4) | 8.81E4<br>(8.39E4, 9.22E4) | 8.83E4<br>(8.40E4, 9.26E4) |

vary significantly. This illustrates that local linear approximation may be inadequate for the interpretation of the estimates. On the other hand, we use the same approximation when estimating the parameters by the Marquardt method, and hence it is really difficult to locate the value of $\hat{A}$.

We note, however, that the use of the weighting coefficients $\tilde{w}_i = 1$ when fitting the line of Eq.(2.38) is the best choice if we assume that $k$ is observed with constant relative variance, thus $\sigma_i^2/k_i^2 = \text{const}$ for every $i$. This assumption implies the weighting coefficients $w_i = 1/\tilde{k}_i^2$ in the original nonlinear estimation problem. In any case, however, it is important to select consistent weighting coefficients if the estimation involves the linearized model.

Unfortunately, the linearization transformations applied in chemical engineering often interchange independent and dependent variables and deviations cannot be reduced by simply adopting suitable weighting. In such cases the estimates from the linear fit can only give starting estimates for fitting the nonlinear model.

Applying good initial estimates for the parameters is not the only way of reducing computational efforts in nonlinear regression. A faster convergence can be expected if the extent of nonlinearity of a given model is decreased. Bates and Watts [7] have suggested a measure of the extent of nonlinearity. They found that nonlinear behaviour usually stems from intrinsic properties of the problem only to a slight degree, the major part of nonlinearity being due to the actual parametrization used. In such cases a simple transformation of some parameters may considerably improve the efficiency of the searching algorithm.

**Example 2.4.** *Effect of reparametrization*

The problem of fitting the model

$$y = a - b \exp(-cx^d) \tag{2.42}$$

to the data shown in *Table 2.7*, where $y$ is pasture regrowth yield and $x$ is time, was studied by Ratkowsky [42]. He found that parametric nonlinearity is high for model (2.42) and it can be reduced by introducing the transformation $c^* = -\ln c$. At a first glance this might be surprising, since the reparametrized model takes the form

$$y = a - b \exp[-\exp(-c^*)x^d] \tag{2.43}$$

Table 2.7. Pasture regrowth data

| $i$ | Time, $x$ | Yield, $y$ |
|---|---|---|
| 1 | 9 | 8.93 |
| 2 | 14 | 10.80 |
| 3 | 21 | 18.59 |
| 4 | 28 | 22.33 |
| 5 | 42 | 39.35 |
| 6 | 57 | 56.11 |
| 7 | 63 | 61.73 |
| 8 | 70 | 64.62 |
| 9 | 79 | 67.08 |

We compare the performance of the Gauss—Newton—Marquardt algorithm for models (2.42) and (2.39) starting from the initial estimate $a=100$, $b=100$, $c=0.0002$ ($c^*=8.51719$) and $d=1$. The unweighted least squares estimate is $\hat{a}=69.96$, $\hat{b}=61.69$, $\hat{c}=1.003\text{E}-4$ and $\hat{d}=2.377$.

Because of the highly nonlinear nature of model (2.42), the Marquardt parameter must be kept at the level of $\lambda=0{,}01$ for dozens of iterations, hence the convergence is only linear, as may be seen from *Table 2.8*. For model (2.43) the Mar-

Table 2.8. Iteration history for Example 2.4

| No. of iteration | Model (2.42) | | | Model (2.43) | | |
|---|---|---|---|---|---|---|
| | $Q(p)$ | Marquardt's $\lambda$ | Remark | $Q(p)$ | Marquardt's $\lambda$ | Remark |
| 0 | 1.744E4 | — | initial | 1.744E4 | — | initial |
| 1 | 5.759E3 | 1E1 | | 1.592E4 | 1E2 | |
| ⋮ | | | | | | |
| 9 | 3.318E1 | 1E−2 | | 1.699E1 | 1E−4 | |
| 10 | 3.246E1 | 1E−2 | | 8.397 | 1E−5 | |
| 11 | 3.173E1 | 1E−2 | | 8.378 | 1E−6 | |
| 12 | 3.099E1 | 1E−2 | | 8.378 | 1E−6 | converged |
| ⋮ | | | | | | |
| 20 | 2.480E1 | 1E−2 | | | | |
| ⋮ | | | | | | |
| 40 | 1.264E1 | 1E−2 | | | | |
| ⋮ | | | | | | |
| 60 | 8.941 | 1E2 | | | | |
| ⋮ | | | | | | |
| 70 | 8.376 | 1E−5 | | | | |
| 71 | 8.376 | 1E−6 | converged | | | |

quardt parameter can be decreased to its minimum, thus the convergence is superlinear.

Unfortunately, model reparametrization still remains a kind of art since no systematic approach has yet been found for it.

Hitherto we have considered single response models where only one dependent variable is observed at one measurement point. The least squares method can be extended to the multi-response case, as is shown in the following example, if a little information is available concerning the nature of measurement errors.

**Example 2.5.** *Fitting a multi-response model by least squares method*

*Figure 2.3* shows the mechanism assumed for the isomerization of α-pinene. Assuming first-order reactions in each step, the concentrations of the five species involved change according to

$$y_1 = y_1^0 \exp(-\Phi t)$$

$$y_2 = \frac{k_1 y_1^0}{\Phi}(1 - \exp(-\Phi t))$$

$$y_3 = c_1 \exp(-\Phi t) + c_2 \exp(\beta t) + c_3 \exp(\gamma t)$$

$$y_4 = k_3\left[\frac{c_1}{\varphi}(1 - \exp(-\varphi t)) + \frac{c_2}{\beta}(\exp(\beta t) - 1) + \frac{c_3}{\gamma}(\exp(\gamma t) - 1)\right]$$

$$y_5 = k_4\left[\frac{c_1}{k_5 - \varphi}\exp(-\varphi t) + \frac{c_2}{k_5 + \beta}\exp(\beta t) + \frac{c_3}{k_5 + \gamma}\exp(\gamma t)\right] \qquad (2.44)$$

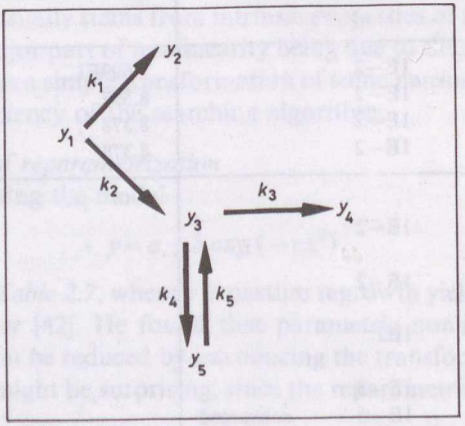



Fig. 2.3. Scheme of isomerization of α-pinene

where $y_1^0 = 100$ is the initial concentration of $y_1$ and

$$\Phi = k_1 + k_2$$
$$\alpha = k_2 + k_4 + k_5, \quad \beta = (-\alpha + \sqrt{\alpha^2 - 4k_3k_5})/2$$
$$\gamma = (-\alpha - \sqrt{\alpha^2 - 4k_3k_5})/2$$
$$c_1 = k_2 y_1^0 (k_5 - \Phi)/[(\Phi + \beta)(\beta + \gamma)]$$
$$c_2 = k_2 y_1^0 (k_5 + \beta)/[(\Phi + \beta)(\beta - \gamma)]$$
$$c_3 = k_2 y_1^0 (k_5 + \gamma)/[(\Phi + \gamma)(\gamma - \beta)]$$

The unknown parameters are the rate constants $k_1, \ldots, k_5$ and the data are listed in *Table 2.10*.

Table 2.9. Isomerization of α-pinene

| $i$ | Time, min | Composition, % | | | | |
|---|---|---|---|---|---|---|
| | | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
| 1 | 1.23E3 | 88.35 | 7.3 | 2.3 | 0.4 | 1.75 |
| 2 | 3.06E3 | 76.4 | 15.6 | 4.5 | 0.7 | 2.8 |
| 3 | 4.92E3 | 65.1 | 23.1 | 5.3 | 1.1 | 5.8 |
| 4 | 7.80E3 | 50.4 | 32.9 | 6.0 | 1.5 | 9.3 |
| 5 | 1.068E4 | 37.5 | 42.7 | 6.0 | 1.9 | 12.0 |
| 6 | 1.503E4 | 25.9 | 49.1 | 5.9 | 2.2 | 17.0 |
| 7 | 2.262E4 | 14.0 | 57.4 | 5.1 | 2.6 | 21.0 |
| 8 | 3.642E4 | 4.5 | 63.1 | 3.8 | 2.9 | 25.7 |

Since several concentrations are observed at the same sample point, it is implausible to assume uncorrelatedness of the errors. To avoid the assumption of a known covariance matrix, Box *et al.* [13] estimated the parameters by minimizing the objective function

$$Q(p) = \det(\mathbf{V}(p)) \tag{2.45}$$

where $\mathbf{V}$ denotes the empirical covariance matrix of the errors computed by

$$[\mathbf{V}(p)]_{ij} = \sum_{k=1}^{nm} [\tilde{y}_{ki} - f_i(x_k, p)][\tilde{y}_{kj} - f_j(x_k, p)] \tag{2.46}$$

This multi-response estimation method certainly has theoretical advantages, but its practical application is far from simple. In particular, it is not directly applicable in this problem, since the concentrations in *Table 2.9* are subject to affin linear relations and hence $\det(\mathbf{V}(p)) \approx 0$ for any value of the parameters. To select a linearly independent set of data Box *et al.* [13] applied principal component analysis to the matrix $\mathbf{ZZ}^T$, where $\mathbf{Z}$ represents the matrix of observations centred around their arithmetic means. There are two small eigenvalues, indicating linear dependences. One of the linear relationships stems from the normalization of data

$$y_1 + y_2 + y_3 + y_4 + y_5 = 100$$

since $y_1^0 = 100$. Explaining the second affine linear relationship Box *et al.* [13] found that $y_4$ was not actually measured, but simply taken as $y_4 = 0.03(100 - y_1)$. For linearly independent variables they selected the principal components corresponding to the three significant nonzero eigenvalues. The new variables are given by

$$\begin{aligned} y_1^* &= 0.8087 \times y_1 - 0.5404 \times y_2 - 0.0127 \times y_3 - 0.0241 \times y_4 - 0.2307 \times y_5 \\ y_2^* &= 0.0568 \times y_1 - 0.2236 \times y_2 - 0.6122 \times y_3 + 0.0038 \times y_4 + 0.7563 \times y_5 \\ y_3^* &= -0.2957 \times y_1 - 0.6108 \times y_2 + 0.6402 \times y_3 - 0.0100 \times y_4 + 0.3599 \times y_5 \end{aligned} \quad (2.47)$$

Obviously, both the response functions (2.44) and the observations in *Table 2.9* should be transformed according to Eq. (2.47).

*Table 2.10* shows the estimates of the rate constant derived by different methods. We first fitted Eq. (2.44) to all the five observed concentrations with unweighted least squares. The second row of the table was obtained by minimizing Eq. (2.45) using the transformed variables Eq. (2.47). The deviations are significant, and we might conclude that using the least squares method introduces unjustified assumption which severely influence the estimates. This conclusion, which has been suggested in the statistical literature (see Box *et al.* [13]; Bates and Watts [8]) is, however, not completely valid in this example, since the main source of deviations is simply the wrong assumption $y_4 = 0.03\ (100 - y_1)$, which was eliminated from the multiresponse method but used in the least squares one. This can easily be demonstrated by applying the least squares method to the transformed variables Eq. (2.47), which are partially independent of $y_4$. For simplicity, we can use the original variables $y_1, \ldots, y_5$ by introducing the weighting matrices $\mathbf{W}_i = \mathbf{W} = \mathbf{F}\mathbf{F}^T$ for each observation, where $\mathbf{F} = [u_1, u_2, u_3]$ and $u_i$, $i = 1, 2, 3$ are the eigenvectors with the large eigenvalues. The resulting weighting matrix is

$$\mathbf{W} = \begin{bmatrix} 0.745 & -0.269 & -0.234 & -0.016 & -0.250 \\ -0.269 & 0.715 & -0.247 & 0.018 & -0.264 \\ -0.234 & -0.247 & 0.785 & -0.008 & -0.230 \\ -0.016 & 0.018 & -0.008 & 0.001 & 0.005 \\ -0.250 & -0.264 & -0.230 & 0.005 & 0.755 \end{bmatrix}$$

Though $\mathbf{W}$ is not diagonal as in our previous examples, the least squares method applies and gives the estimates shown in row 3 of *Table 2.10*, which are close to the values obtained by Box *et al.* [13]. The analysis of this well known example emphasized that the least squares method can also give completely acceptable results in multiresponse situations, and a careful analysis of data is more important than the choice of the theoretically more appropriate determinant criterion. Furthermore, checking linear dependences among the data may be quite useful, though the least squares method does not require linear independence.

Table 2.10. Estimates of the rate constants in Example 2.5

| Method | Rate constant, 1/min | | | | |
|---|---|---|---|---|---|
| | $k_1$ | $k_2$ | $k_3$ | $k_4$ | $k_5$ |
| Least squares 5 components | 5.93E−5 | 2.96E−5 | 2.05E−5 | 2.75E−4 | 4.00E−5 |
| Box–Draper 3 principal components | 5.95E−5 | 2.84E−5 | 4.3E−6 | 3.13E−4 | 5.74E−5 |
| Least squares 3 principal components | 5.95E−5 | 2.87E−5 | 5.1E−6 | 2.96E−4 | 5.16E−5 |

### 2.2.4 Indirect least squares method of fitting dynamic models

In this section we estimate the parameters in lumped dynamic models given by Eqs (2.2) and (2.3). As is shown in Example 2.5, this is not a specific problem if the solution of the differential equations is available in analytical form. Since we fit this solution to the data instead of the dynamic model itself, the approach is called the indirect least sqares (ILS) method.

The indirect approach also applies when no analytical solution is available, simply by numerically integrating the differential equations in each iteration step of the estimation procedure at the current estimates of the parameters. An additional numerical difficulty is the computation of the Jacobian matrix Eq. (2.31), where $f$ represents the output of the model, thus $f(t,p)=h(x(t,p)t,p)$, $x(t,p)$ being the solution of the differential equations with the initial conditions $x(0,p)=$ $=x^0(p)$. The entries of $J$ are also called parameter sensitivity coefficients. The simplest way of computing these sensitivity coefficients is the finite difference approximation

$$\frac{\partial y_i(t,p)}{\partial p_j} \approx \frac{h_i(x(t,p+\Delta p_j),t,p+\Delta p_j)-h_i(x(t,p),t,p)}{\Delta p_j} \tag{2.48}$$

where $\Delta p_j=(0,\ldots,0,\Delta p_j,\ 0,\ldots,0)^T$ represents a perturbation vector of the parameters. To evaluate the derivatives $\partial y/\partial p_1,\ldots,\partial y/\partial p_{np}$ the differential equations should be solved $(np+1)$ times in each step of the iteration, and it is not easy to choose the suitable perturbation $\Delta p_j$.

A more advanced method of calculating the sensitivity coefficients is the solution of the sensitivity equations obtained by differentiating Eq.(2.2) with respect to the parameters. The sensitivity equations are given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial x}{\partial p_j}\right)=\frac{\partial g(x,t,p)}{\partial x}\left(\frac{\partial x}{\partial p_j}\right)+\frac{\partial g(x,t,p)}{\partial p_j} \tag{2.49}$$

where $\partial g(x, t, p)/\partial x$ is the $nx \times nx$ matrix of the partial derivatives of $g$ with respect to variables $x$. If the initial condition $x_i$ does not depend on parameter $p_j$, then the initial condition for Eq.(2.49) is given by $\partial x_i(0, p)/\partial p_j = 0$, whereas for $p_j = x_i^0$ (i.e., an unknown initial concentration $\partial x_i(0, p)/\partial p_j = 1$. Solving the $nx \times np$ differential equations (2.49) in conjunction with the original $nx$ equations, the sensitivity coefficients can be computed by

$$\frac{\partial y}{\partial p_j} = \frac{\partial h(x, t, p)}{\partial x} \frac{\partial x}{\partial p_j} + \frac{\partial h(x, t, p)}{\partial p_j} \tag{2.50}$$

where $\partial h(x, t, p)/\partial x$ is an $ny \times nx$ matrix. Though the computational effort required for solving Eq.(2.49) is the same as in the finite difference approximation, the sensitivity coefficients are more accurate. Furthermore, a variety of methods have been proposed for reducing the cost of sensitivity calculations. For example, once the solution $x(t, p)$ of Eq.(2.2) is available, Eq.(2.49) is a set of time-variable linear differential equations that is relatively easy to solve by quadrature. This is the basic idea of the Green's function-method proposed by Rabitz [41]. Another possibility is to solve Eq.(2.49) together with Eq.(2.2) and use the information already computed for solving the original differential equations. This approach is known as the decomposed or decoupled direct method of sensitivity calculation; see Dickinson and Gelinas [16].

Since the differential equations describing a chemical process are usually rather hard to solve, the success of parameter estimation depends significantly on the performance of the algorithm used to solve the ordinary differential equations (ODEs) and to compute sensitivities. In the area of chemical engineering, particularly in reaction engineering the most celebrated ODE-solver is the one proposed by Gear [21]. According to more recent studies, however, the relatively simple semi-implicit Runge-Kutta methods (Gottwald and Wanner [24], Prokopakis and Seider [40]) have comparable performance in the solution of ODEs in chemical engineering applications. In this work we use the algorithm ROW4A of Gottwald and Wanner [24] and its extension ROW4S (Valkó and Vajda [55]), which also computes the sensitivity coefficients simultaneously. Though the original program given by Valkó and Vajda [55] requires user-supplied subroutines to evaluate the partial derivatives $\partial g/\partial x$ and $\partial g/\partial p$, and even higher-order derivatives in addition to the right hand side $g$ of the ODEs, the version used in REPROCHE involves finite difference approximations in order to reduce the user's programming effort.

An additional problem encountered when estimation parameters in dynamic models is that of identifiability. Since the solution of the ODEs is not obtained in an analytic form, it is difficult to assess that the solution of the estimation problem is unique, or whether there exist several or even an infinite number of equivalent solutions, e.g. local minima of the objective function (2.28). The problem can be somewhat simplified by introducing the concept of deterministic identifiability,

thus assuming error-free and continuous observation of the output function $y$ in Eq.(2.3), see Park and Himmelblau [39] and Vajda [50]. We know that kinetic models are identifiable if all concentrations are observed, but this is not necessarily true if only some of them or their functions are observed [Eq.(2.3)]. Simple identifiability conditions are available for linear dynamic systems. In the nonlinear case, however, the calculations are so cumbersome that it is advisable to check deterministic identifiability only if matrix $\mathbf{J}^{\mathrm{T}}(\underline{p})\mathbf{W}\mathbf{J}(\hat{\underline{p}})$ is singular or nearly-singular, see Vajda [51]. We emphasize that deterministic unidentifiability is only one of the sources of near-singularity, which may also be caused by an unsuitable experimental design.

**Example 2.6.** *Test problem of Eakman* [17]

Comparing the least squares and the multiresponse method, Eakman considered the reactions

$$A+B \overset{k_1}{\rightarrow} C+D$$

$$A+D \overset{k_2}{\rightarrow} 2X$$

Assuming rate expressions $r_1=k_1[A][B]$ and $r_2=k_2[A][D]$, Eakman generated several sets of data with the rate constants $k_1=0.3$ and $k_2=0.1$. Here we consider his set of 'bad' data (Eakman [17]), shown in *Table 2.11*. The problem is easy to solve if all five concentrations are considered. The use of the least squares criterion is completely justified, since the additional errors selected by Eakman are uncorrelated, though not identical in variance. Nevertheless, very good estimates are obtained even through the use of the weighting matrix $\mathbf{W}=\mathbf{I}$ shown in the second numerical column of *Table 2.12*. These estimates are close to the ones that Eakman obtained by the multiresponse method. He was, however, unable to reach convergence when considering only $A$ as an observed variable, though the model is identifiable even in this case. In fact, using the REPROCHE program, we

Table 2.11. Eakman's 'bad' data

| $t$ | $A$ | $B$ | $C$ | $D$ | $X$ |
|---|---|---|---|---|---|
| 0.000E+0 | 0.1500E+1 | 0.1000E+1 | 0.0000E+0 | 0.0000E+0 | 0.0000E+0 |
| 0.100E+1 | 0.1115E+1 | 0.6198E+0 | 0.3111E+0 | 0.3033E+0 | 0.4920E−1 |
| 0.200E+1 | 0.9465E+0 | 0.5279E+0 | 0.5352E+0 | 0.4372E+0 | 0.1321E+0 |
| 0.300E+1 | 0.7022E+0 | 0.4256E+0 | 0.6145E+0 | 0.5058E+0 | 0.2048E+0 |
| 0.400E+1 | 0.6495E+0 | 0.2993E+0 | 0.7034E+0 | 0.5479E+0 | 0.2770E+0 |
| 0.500E+1 | 0.5133E+0 | 0.2468E+0 | 0.6494E+0 | 0.5916E+0 | 0.3294E+0 |
| 0.600E+1 | 0.5383E+0 | 0.1855E+0 | 0.7144E+0 | 0.5876E+0 | 0.4688E+0 |
| 0.700E+1 | 0.4501E+0 | 0.2103E+0 | 0.8014E+0 | 0.5738E+0 | 0.4837E+0 |
| 0.800E+1 | 0.3689E+0 | 0.1979E+0 | 0.8312E+0 | 0.5874E+0 | 0.4671E+0 |
| 0.900E+1 | 0.3961E+0 | 0.1489E+0 | 0.9286E+0 | 0.5725E+0 | 0.6203E+0 |
| 0.100E+2 | 0.3397E+0 | 0.1061E+0 | 0.8383E+0 | 0.5756E+0 | 0.6599E+0 |

Table 2.12. Estimates of the rate constants in Example 2.6

| Parameter | True value | $A, B, C, D, X$ observed | Only $A$ observed |
|---|---|---|---|
| $k_1$ | 0.3000 | $0.3041 \pm 0.0068$ | $0.3488 \pm 0.0287$ |
| $k_2$ | 0.1000 | $0.1042 \pm 0.0033$ | $0.0815 \pm 0.0157$ |

obtained reasonable estimates, which are also shown in *Table 2.12*, from the starting values $k_1^{(0)} = 0.5$ and $k_2^{(0)} = 0.5$ given by Eakman, though deviations from the true values are increased. *Figures 2.4* and *2.5* show the response function $X$ from model fitted to the observations of all concentrations and only $[A]$, respectively. In the latter case we still obtain reasonable estimates for the unobserved variables. In practical situations the predictions are rarely so good, since here we used a 'true' model (i.e., computer-generated data).

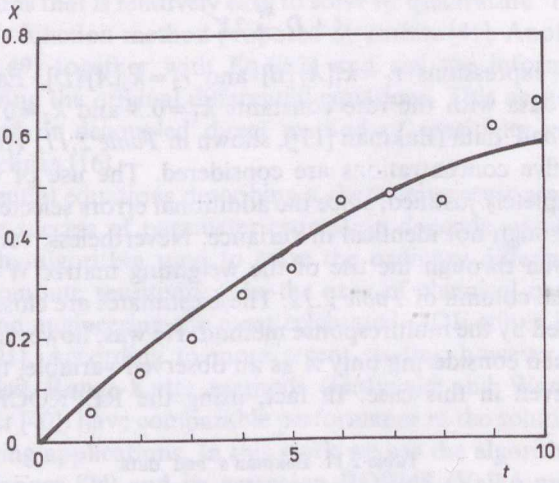



Fig. 2.4. Observed and computed response $X$ in Example 2.6. Model fitted to all observations

**Example 2.7.** *Aspirin metabolism and elimination*

The problem discussed here has its origin in pharmacokinetics. In spite of its apparent simplicity, in this estimation problem we encounter a number of difficulties that are more or less inherent in kinetic analysis.

We study concentrations of salicylic acid following the intake of 2400 mg Aspirin. Aspirin is very rapidly hydrolyzed to salicylic acid, which is absorbed in the blood plasma according to a first-order rate. From the plasma the drug is eliminated by two pathways, one of which is linear (excretion into urine) and the other follows the Michaelis—Menten rate law (transformation to a metabolite). *Figure 2.6* shows the so-called compartmental model (Godfrey [22]) corresponding

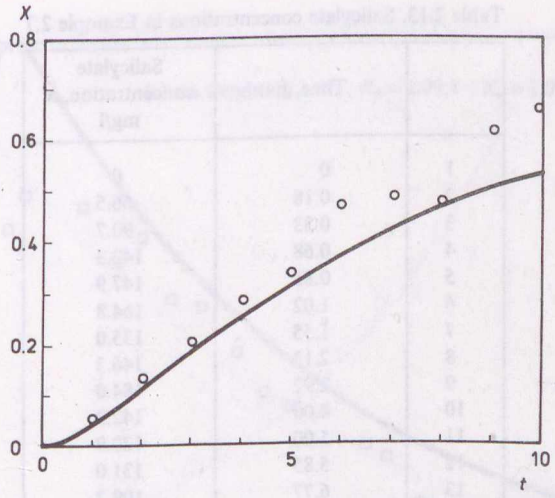



Fig. 2.5. Observed and computed response $X$ in Example 2.6. Model fitted to $[A]$ only

to this system, which results in the differential equation

$$V\frac{dx}{dt} = Dk_a e^{-k_a t} - \frac{V_m x}{K_m + x} - k_{el} x \tag{2.51}$$

where the notation is as follows:

$x$ is the salicylation concentration in blood plasma, mg $l^{-1}$,
$V$ is the plasma volume, l,
$D$ is the dose of salicylic acid, 1813 mg,
$k_a$ is the absorption rate constant, $h^{-1}$,
$k_{el}$ is the elimination rate constant, $h^{-1}$, and
$V_m$, $K_m$ are the Michaelis—Menten parameters, mg $h^{-1}$ and mg $l^{-1}$, respectively.

The only deviation from chemical kinetic models is the unknown volume $V$ of the 'reacting' system, and the parameters to be estimated are $p=(k_a, V_m, K_m, k_{el}, V)$.

The observed salicylate concentrations $\tilde{x}_i$ are shown in *Table 2.13*. As ad-

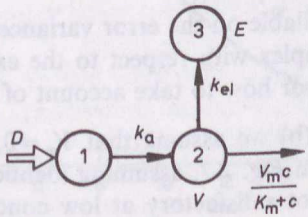



Fig. 2.6. Compartmental model of Aspirin metabolism and elimination

Table 2.13. Salicylate concentrations in Example 2.7

| $i$ | Time, h | Salicylate concentration, $x$ mg/l |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 0.18 | 66.5 |
| 3 | 0.33 | 90.7 |
| 4 | 0.68 | 142.3 |
| 5 | 0.85 | 147.9 |
| 6 | 1.02 | 164.8 |
| 7 | 1.55 | 135.0 |
| 8 | 2.13 | 146.3 |
| 9 | 2.92 | 184.0 |
| 10 | 4.00 | 142.3 |
| 11 | 5.00 | 122.9 |
| 12 | 5.85 | 131.0 |
| 13 | 6.77 | 109.2 |
| 14 | 7.77 | 106.8 |
| 15 | 8.89 | 89.9 |
| 16 | 9.78 | 76.1 |
| 17 | 10.77 | 70.5 |
| 18 | 11.73 | 68.9 |
| 19 | 13.06 | 56.8 |
| 20 | 13.77 | 52.0 |
| 21 | 14.77 | 35.0 |
| 22 | 15.73 | 33.2 |
| 23 | 23.83 | 2.8 |

ditional information we observe the total amount of drug excreted by the linear pathway, which gives the constraint

$$\tilde{E}=k_{el}\int_0^\infty x(\tau)\mathrm{d}\tau=877\text{ mg} \tag{2.52}$$

Interpolating the observed values $\tilde{x}_i$ by a cubic spline function and evaluating the integral in Eq.(2.52) we obtain the estimate $\hat{k}_{el}=0.4983$.

As will be discussed in this example we encounter the following difficulties:

(a) No information is available on the error variances.
(b) The model is too complex with respect to the experimental observations.
(c) It is not completely clear how to take account of Eq.(2.52).

To demonstrate (a) and (b) we assume that $V_m=0$, and thus consider only linear excretion. As shown in *Fig. 2.7*, assuming identical variances $\sigma_{c_i}^2=\sigma^2$, and thus that $\mathbf{W}=\mathbf{I}$, the fit is not satisfactory at low concentrations, whereas with relative errors $\sigma_{c_i}^2=\sigma^2 x_i^2$ the last observation has too much influence. The com-

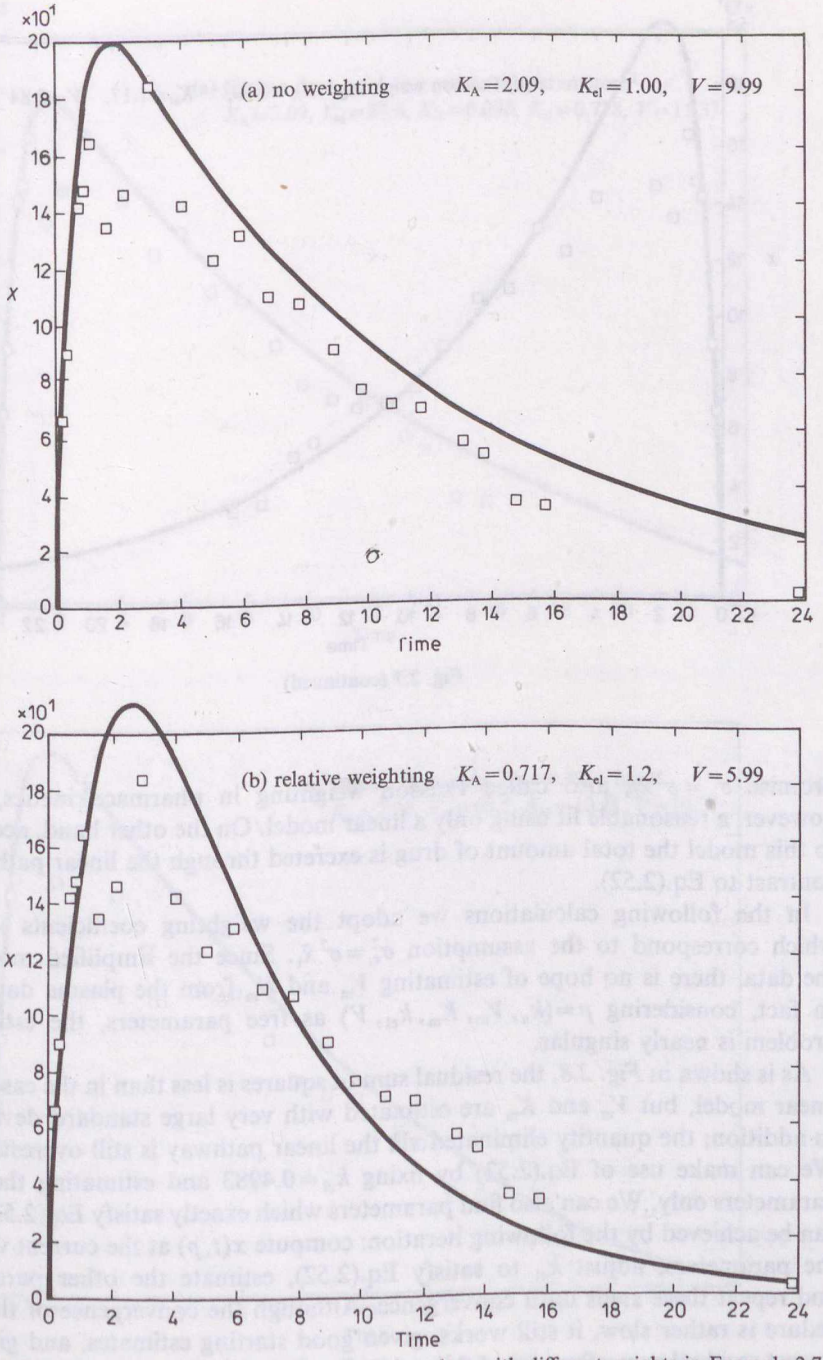



Fig. 2.7. Fit of the linear dynamic model to salicylate data with different weights in Example 2.7.

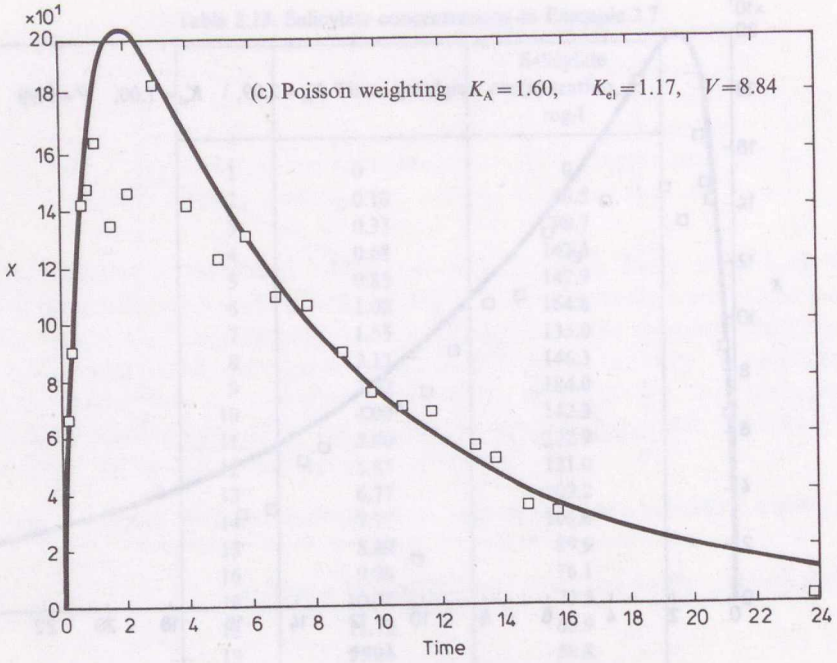
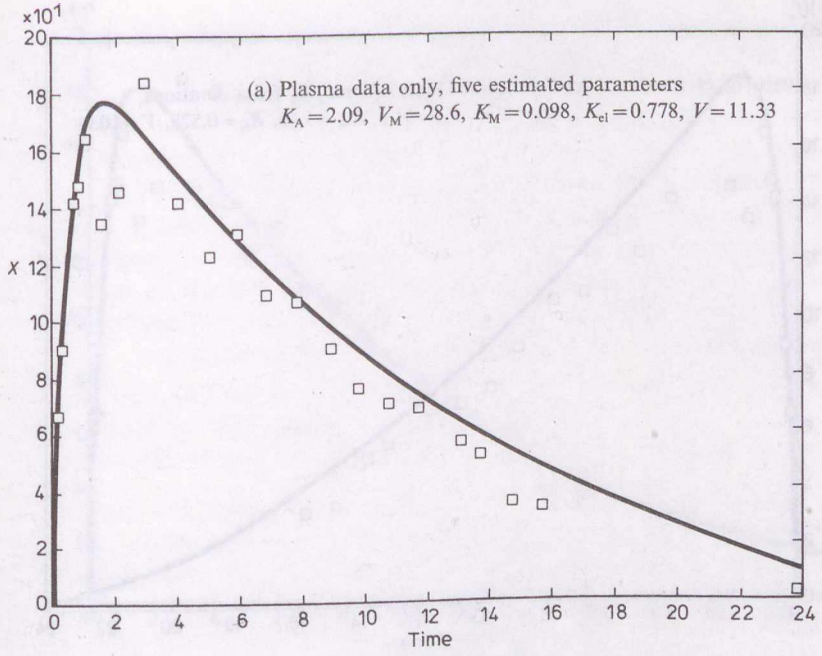



Fig. 2.7 (continued)

promise $\sigma_{c_i} = \sigma^2 \tilde{x}_i$, also called Poisson weighting in pharmacokinetics, gives, however, a reasonable fit using only a linear model. On the other hand, according to this model the total amount of drug is excreted through the linear pathway in contrast to Eq.(2.52).

In the following calculations we adopt the weighting coefficients $w_i = 1/\tilde{c}_i$ which correspond to the assumption $\sigma_{c_i}^2 = \sigma^2 \tilde{x}_i$. Since the simplified model fits the data, there is no hope of estimating $V_m$ and $K_m$ from the plasma data only. In fact, considering $p = (k_a, V_m, K_m, k_{el}, V)$ as free parameters, the estimation problem is nearly singular.

As is shown in *Fig. 2.8*, the residual sum of squares is less than in the case of the linear model, but $V_m$ and $K_m$ are estimated with very large standard deviations. In addition, the quantity eliminated via the linear pathway is still overestimated. We can make use of Eq.(2.52) by fixing $k_{el} = 0.4983$ and estimating the other parameters only. We can also find parameters which exactly satisfy Eq.(2.52). This can be achieved by the following iteration: compute $x(t, p)$ at the current value of the parameters, adjust $k_{el}$ to satisfy Eq.(2.52), estimate the other parameters and repeat these steps until convergence. Although the convergence of this procedure is rather slow, it still works, given good starting estimates, and gives the lowest residual sum of squares achieved so far. The last two approaches, however,



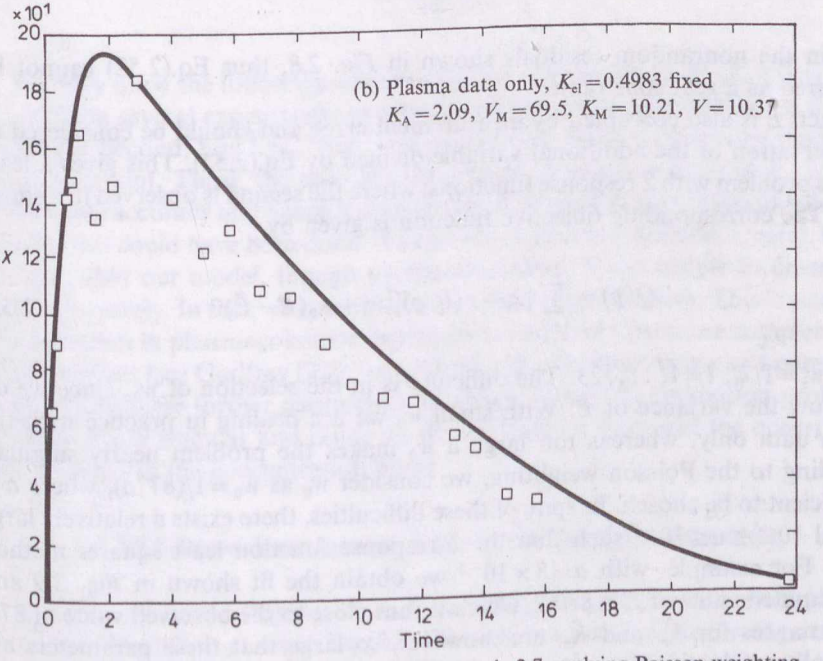




Fig. 2.8. Fit of the original model in Example 2.7 applying Poisson weighting

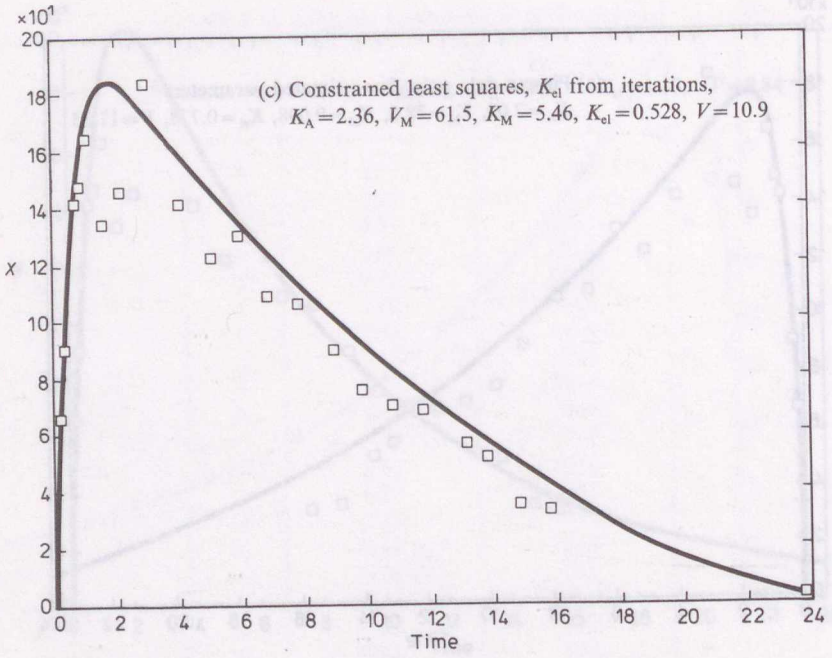



Fig. 2.8 (continued)

result in the nonrandom residuals shown in *Fig. 2.8*, thus Eq.(2.52) cannot be considered as a real constraint.

In fact, $\tilde{E}$ is also corrupted by measurement error and should be considered as an observation of the additional variable defined by Eq.(2.52). This gives a least squares problem with 2 response functions, where the second is observed in a single point. The corresponding objective function is given by

$$Q(p) = \sum_{i=1}^{23} [x_i - x(t_i, p)]^2 w_i + w_0 (E - \tilde{E}) \tag{2.53}$$

where $w_i = 1/\tilde{x}_i$, $i = 1, \ldots, 23$. The difficulty is in the selection of $w_0$, since we do not know the variance of $\tilde{E}$. With small $w_0$ we are dealing in practice with the plasma data only, whereas too large a $w_0$ makes the problem nearly singular. According to the Poisson weighting, we consider $w_0$ as $w_0 = 1/(877a)$, where $a$ is a coefficient to be chosen. In spite of these difficulties, there exists a relatively large interval $10^{-5} \leq a \leq 10^{-4}$ such that the 2 response function least squares method works. For example, with $a = 8 \times 10^{-5}$ we obtain the fit shown in *Fig. 2.9* and the computed value of $E$ is 880.2, which is thus close to the observed value of 877. The variances for $\hat{V}_m$ and $\hat{K}_m$ are, however, so large that these parameters are practically unidentifiable.

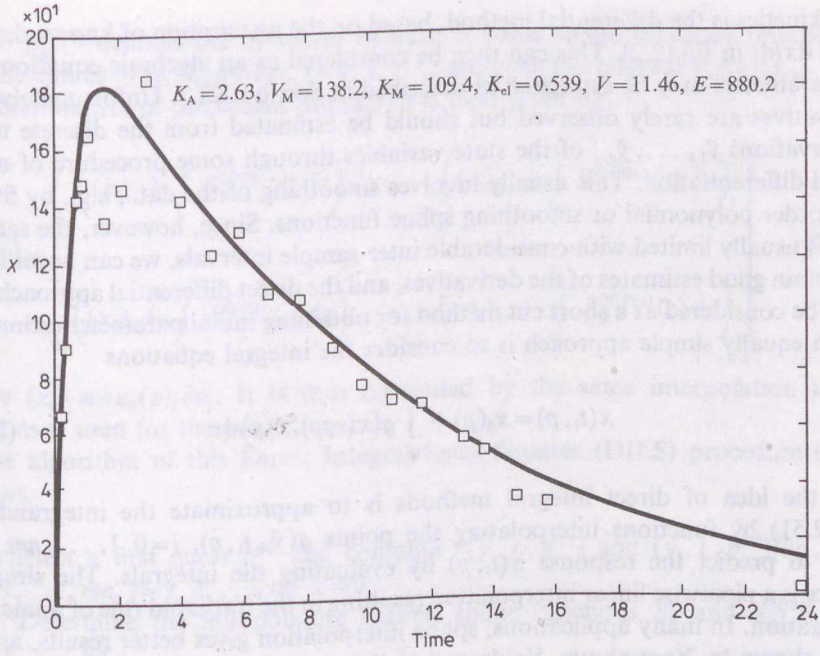



Fig. 2.9. Fit of the original model in Example 2.7 applying Poisson weighting using both plasma data and $E$

We may draw the following conclusions from this analysis. First, it is advisable to perform several experiments in order to obtain some information on the error structure. Second, Eq.(2.52) based on the observed value $\tilde{E}$, cannot be used simply as a constraint. Third, the data are insufficient to estimate $K_m$ and $V_m$ with any reasonable accuracy and hence it is advisable to observe $E$ also at intermediate time points (this could have been done, though with large measurement errors). Finally, we note that our model, though overparametrized, is too simple to describe the data adequately. In fact, we can observe two maxima in the data. This is a common phenomenon in pharmacokinetic experiments and stems from the complex nature of absorption (see Godfrey [22]), involving further compartments and parameters. Introducing these further parameters, however, causes the estimation problem to become nearly singular and hence we do not attempt to model the occurrence of two maxima in the concentration curve.

### 2.2.5 Direct least squares method of fitting dynamic models

Models in chemical engineering are frequently formulated in terms of observed variables, thus $h(t, x, p) = x$ in Eq.(2.3). In such cases the estimation of parameters can be simplified considerably. A traditional approach used in reaction engineering

and kinetics is the differential method, based on the assumption of known derivatives $dx/dt$ in Eq.(2.2). This can then be considered as an algebraic equation and its parameters may be estimated as described in Section 2.2.3. Unfortunately, the derivatives are rarely observed but should be estimated from the discrete noisy observations $\tilde{y}_1, \ldots, \tilde{y}_{nm}$ of the state variables through some procedure of numerical differentiation. This usually involves smoothing of the data, e.g., by fitting low-order polynomial or smoothing spline functions. Since, however, the sample size is usually limited with considerable inter-sample intervals, we can never hope to obtain good estimates of the derivatives, and the direct differential approach can only be considered as a short cut method for obtaining initial parameter estimates.

An equally simple approach is to consider the integral equations

$$x(t_i, p) = x_0(p) + \int_{t_0}^{t_i} g(x(\tau, p), \tau, p)\, d\tau \tag{2.54}$$

and the idea of direct integral methods is to approximate the integrands in Eq.(2.51) by functions interpolating the points $g(\tilde{y}_i, t_i, p)$, $i=0, 1, \ldots, nm$ and then to predict the response $x(t_i, p)$ by evaluating the integrals. The simplest choice is a piecewise linear interpolation resulting in the trapezoid rule of numerical integration. In many applications, spline interpolation gives better results, as has been shown by Yermakova, Vajda and Valkó [58] and Sinha and Quijie [46].

Let $S_p^f$ denote the $n$-vector of natural cubic splines interpolating the values $g(\tilde{y}_i, t_i, p)$, $i=0, 1, \ldots, nm$. Then the predicted output is given by

$$\tilde{f}(t_i, x_0, u, p) = x_0(p) + \int_{t_0}^{t_i} S_p^f(\tau)\, d\tau \tag{2.55}$$

In order to minimize the deviations between observed and predicted outputs, we use the criterion

$$Q_{\mathrm{DILS}}(\underline{p}) = (\underline{Y} - \underline{\tilde{F}}(\underline{p}))^{\mathrm{T}} \mathbf{W} (\underline{Y} - \underline{\tilde{F}}(\underline{p})) \tag{2.56}$$

where

$$\underline{\tilde{F}}(\underline{p}) = \begin{bmatrix} \tilde{f}(t_1, x_0, p) \\ \tilde{f}(t_2, x_0, p) \\ \ldots\ldots\ldots\ldots \\ \tilde{f}(t_{nm}, x_0, p) \end{bmatrix} \tag{2.57}$$

Evaluation of the Jacobian matrix $\partial \underline{F}/\partial \underline{p}$ is computationally the most expensive step of the indirect method, but by contrast the Jacobian matrix $\partial \underline{\tilde{F}}/\partial \underline{p}$ of the prediction $\underline{\tilde{F}}(\underline{p})$ can easily be computed. Changing the order of differentiation and integration, we obtain

$$\frac{\partial}{\partial p_j} \int_{t_0}^{t_i} S_p^f(\tau)\, d\tau = \int_{t_0}^{t_i} S_p^{\partial f/\partial p_j}(\tau)\, d\tau \tag{2.58}$$

where $S_p^{\partial f/\partial p_j}$ denotes the $ny$-vector of natural cubic spline functions interpolating the values $\partial f(\tilde{y}_i, t_i, p)/i\partial p_j$, $i=0, 1, \ldots, nm$, over the interval $t_0 \leq t \leq t_{nm}$. The Jacobian matrix of prediction [Eq.(2.57)] is then given by

$$\tilde{\mathbf{J}}(\underline{p}) = \begin{bmatrix} \left((x_0)_1 + \int_{t_0}^{t_1} S_p^{\partial f/\partial p_1}(\tau)\,d\tau\right) & \cdots & \left((x_0)_{np} + \int_{t_0}^{t_1} S_p^{\partial f/\partial p_{nm}}(\tau)\,d\tau\right) \\ \cdots\cdots\cdots & & \cdots\cdots\cdots \\ \left((x_0)_1 + \int_{t_0}^{t_{nm}} S_p^{\partial f/\partial p_1}(\tau)\,d\tau\right) & \cdots & \left((x_0)_{np} + \int_{t_0}^{t_{nm}} S_p^{\partial f/\partial p_{np}}(\tau)\,d\tau\right) \end{bmatrix} \tag{2.59}$$

where $(x_0)_j \equiv \partial x_0(p)/\partial p_j$. It is thus computed by the same interpolation technique as is used for the prediction.

The algorithm of this Direct Integral Least-Squares (DILS) procedure is as follows:

(*i*) Select a first guess $\underline{p}^{(k)}$ and compute $f(y_i, t_i, p^{(k)})$ and $(y_i, t_i, p^{(k)})/\partial p_j$, $i=0, 1, \ldots, nm$, $j=1, 2, \ldots, np$, $k=0$

(*ii*) Determine the interpolating natural spline functions $S_p^f$ and $S_p^{\partial f/\partial p_j}$, $j=1, 2, \ldots, np$

(*iii*) Compute the integrals (2.55) and (2.58) analytically and form $\tilde{F}(\underline{p}^{(k)})$ and $\tilde{J}(\underline{p}^{(k)})$ according to Eqs (2.47) and (2.49), respectively.

(*iv*) Compute the new estimate $\underline{p}^{(k+1)}$ by the Marquardt method:

$$\underline{p}^{(k+1)} = p^{(k)} + [\tilde{J}^T(\underline{p}^{(k)})\, W \tilde{J}(\underline{p}^{(k)}) + \lambda^{(k)} I]^{-1} \tilde{J}^T(\underline{p}^{(k)})\, W[\underline{Y} - \underline{\tilde{F}}(\underline{p}^{(k)})] \tag{2.60}$$

(*v*) Return to (*ii*) with $p^{(k+1)}$ in place of $p^{(k)}$ and continue until convergence.

The DILS method is numerically more efficient than the ILS method because a spline approximation followed by analytical integration is much faster than numerical integration of differential and sensitivity equations. If model (2.2) is linear in the parameters, then the Jacobian matrix does not depend on the parameters and the least-squares estimate is obtained in one step.

Applications have appeared in the literature for many years, e.g., Himmelblau, Jones and Bischoff [26].

The initial conditions can be eliminated from the predictions by the method of modulating functions (Loeb and Cohen [34]), but this is not necessary in the approach described here since the initial conditions can be regarded as free parameters. The nonlinear version of the DILS method has been studied by Yermakova, Vajda and Valkó [58]. The DILS approach is usually regarded as only a short-cut method to simplify parameter estimation (see, for example, Hosten [29]), but here we show that it may be more preferable than the ILS estimator from a statistical point of view.

The residual error $\underline{Y} - \underline{\tilde{F}}(\underline{p})$ in the DILS method is correlated with the predicted

response $\tilde{\underline{F}}(\underline{p})$ and the estimates are not consistent (i.e. they are asymptotically biased, Strejc, 1980), whereas the ILS estimates are consistent (Jennrich [30]). However, consistency or even unbiasedness does not necessarily imply practical superiority of an estimation procedure if the Jacobian matrix $\mathbf{J}(p)$ is nearly singular, and hence the conventional *LS* estimates tend to be inflated (Hoerl and Kennard [27], Hocking, [28]). In some applications, this problem can be avoided by using low-order models and appropriate parametrizations, but it remains an important problem in applications involving models with physically justified state variables and parameters. Jennrich and Sampson [31] showed that such models are often 'poorly parametrized' in the sense that the linear interdependence in the Jacobian is inherent.

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{np} > 0$ denote the eigenvalues of the approximate Hessian matrix $\mathbf{J}^T(\hat{p})\mathbf{W}\mathbf{J}(\hat{p})$ at the estimates $\hat{p}$. As was discussed in Section 2.2.3, the volume of the joint confidence region of the parameters is proportional to

$$S = \sigma^2 / \prod_{j=1}^{np} \lambda_j \tag{2.61}$$

Further, let $L^2 = (\underline{p} - \underline{\hat{p}})^T (\underline{p} - \underline{\hat{p}})$ denote the squared distance between the LS estimates $\underline{\hat{p}}$ and the true parameters $p$. Based on the linear approximation of the response function,

$$E(L^2) \simeq \text{trace}\,[\mathbf{J}^T(\underline{\hat{p}})\,\mathbf{W}\mathbf{J}(\underline{\hat{p}})]^{-1} = \sum_{j=1}^{np} \sigma^2/\lambda_j > \sigma^2/\lambda^{\min} \tag{2.62}$$

(Hoerl and Kennard [27]). Thus if near-singularity occurs, we may expect both large variances and a large mean-square error of the estimates. As the number of samples becomes small, these are of more serious concern than asymptotic bias.

To reduce the consequences to near-singularity, several biased estimators have been proposed that damp or shrink the least-squares estimator towards the origin by increasing the eigenvalues of the Hessian matrix (Hocking [28]). The eigenvalues of $\tilde{\mathbf{J}}^T(\underline{\hat{p}})\mathbf{W}\tilde{\mathbf{J}}(\underline{\hat{p}})$ in the DILS method are usually larger than those of $\tilde{\mathbf{J}}^T(\underline{\hat{p}})\mathbf{W}\tilde{\mathbf{J}}(\underline{\hat{p}})$ in the ILS method. To see the source of this difference, consider a change $\Delta p$ in the parameters. In the DILS method, the values $x(t_j, p)$ are approximated by observations $y_i$ which do not depend on the parameters. Hence $\Delta \underline{p}$ affects the DILS objective function Eq.(2.56) more directly than in the ILS case, for which the response function values $x(t_i, \underline{p})$ are also influenced by $\Delta \underline{p}$ and these cross-effects may result in nearly linear interdependence of the columns of the Jacobian matrix.

If there are large measurement errors, widely-spaced samples or a poor interpolating function, there is likely to be considerable error in the approximation of the right-hand side of Eq.(2.54). In this case, the poor fit and/or severely biased estimator may outweigh the advantage of larger eigenvalues and hence a further method for reducing the bias of the estimates is introduced.

Let $\hat{S}_p^f$ and $\hat{S}_p^{\partial f/\partial p_j}$ denote the natural cubic spline functions interpolating the values $f(\hat{x}(t_i, p), t_i, p)$ and $\partial f(\hat{x}(t_i, p), t_i, p)/\partial p_j$ respectively, where

$i=0, 1, \ldots, nm$. Define $\hat{\underline{F}}(\underline{p})$ replacing $S_p^f$ in Eq.(2.45) by $\hat{S}_p^f$. Similarly, let $\hat{J}(p)$ denote the matrix obtained from Eq.(2.46), replacing $S_p^{\partial f/\partial p_j}$ by $\hat{S}_p^{\partial f/\partial p_j}$. Then, according to Method IV (instrumental variable method, Young [59]), Eq.(2.60) is replaced by

$$\underline{p}^{(k+1)}=\underline{p}^k+[\hat{\mathbf{J}}^T(\underline{p}^k)\,\mathbf{W}\hat{\mathbf{J}}(\underline{p}^k)+\lambda^k\mathbf{I}]^{-1}\,\hat{\mathbf{J}}^T(\underline{p}^k)\,\mathbf{W}[\underline{Y}-\hat{\underline{F}}(\underline{p}^k)] \tag{2.63}$$

Since $\underline{\hat{x}}$ is an estimate of the undisturbed process signal $x$, Eq.(2.63) is the bootstrap estimator discussed by Söderström and Stoica [47]. This estimator is consistent under fairly mild conditions and one may expect global convergence for sufficiently large samples.

We shall, however, use a different type IV estimator. First, the IV-variable $\hat{x}$ is the solution of Eq.(2.2) and thus we can replace the predicted response $F$ by the indirect response function $F$ as defined by Eq.(2.29). Second, to avoid the inversion of a nonsymmetric matrix and later to form a multistage estimation procedure, we restrict our considerations to the symmetric bootstrap ($SB$) estimator

$$\underline{p}^{(k+1)}=\underline{p}^{(k)}+[\hat{\mathbf{J}}^T(\underline{p}^{(k)})\,\mathbf{W}\hat{\mathbf{J}}(\underline{p}^{(k)})+\lambda^{(k)}\mathbf{I}]^{-1}\,\hat{\mathbf{J}}^T(\underline{p}^{(k)})\,\mathbf{W}[Y-F(\underline{p}^{(k)})] \tag{2.64}$$

The SB estimator gives rise to the following estimation algorithm:

(*i*) Compute the DILS estimates.
(*ii*) Solve Eq.(2.2) at the parameter estimates to evaluate the instrumental variable $\hat{x}$.
(*iii*) Using the values $\hat{x}(t_i\,p^{(k)})$, construct $F(p^{(k)})$ and $\hat{J}(p^{(k)})$.
(*iv*) Form the new estimates by Eq.(2.64).
(*v*) Repeat steps (ii) to (iv) until the parameter estimates converge.

This algorithm can be interpreted in various ways. First, from a practical point of view, it minimizes the indirect least squares objective function [Eq.(2.28)] using the approximate Jacobian matrix $\hat{\mathbf{J}}(p)$ based on Eq.(2.59), with spline functions interpolating the values $\partial g(\hat{x}(t_i,p),t_i,p)/\partial p_j$, $i=0, 1, \ldots, nm$. Second, with $\lambda^{(k)}=0$, Eq.(2.64) is the continuous-time counterpart of the discrete-time pseudo-linear regression estimator, recently discussed by Stoica *et al.* [48]. In fact, let $s_j(t,p)=\partial\hat{x}(t,p)\,\partial p_j$ denote the vector of sensitivities of the solution of Eq.(2.2) with respect to $p_j$, then from the sensitivity equation [Eq.(2.49)] the entries of the Jacobian matrix $\mathbf{J}(p)$ are given by

$$\frac{dx_i}{dp}(t_i,p)=\int_{t_0}^{t_i}\frac{\partial g}{\partial x}(\hat{x}(\tau,p),\tau,p)\,s_j(\tau,p)\,d\tau+ \tag{2.65}$$

$$+\int_{t_0}^{t_i}\frac{\partial g}{\partial p_j}(\hat{x}(\tau,p),\tau,p)\,d\tau$$

Evaluating $\tilde{\mathbf{J}}(p)$, we neglect the first term on the right-hand side of Eq.(2.65), as in the pseudo-linear regression and we approximate the integrand in the second term by spline functions. As has been shown by Stoica *et al.* [48], for linear systems and in this basic form [i.e., with $\lambda^{(k)}=0$ in Eq.(2.45)], this algorithm converges to a consistent estimate under quite restrictive conditions, but its convergence can be improved considerably by a "step-variable" modification, multiplying the second term in Eq.(2.64) by a scalar $\alpha_k$ such that

$$\alpha_k \geq 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty, \quad \text{and} \quad \lim_{k \to \infty} \alpha_k = 0.$$

It should be emphasized that, when monitoring the convergence, the Marquardt algorithm automatically introduces a similar sequence of multipliers, with an additional effect of regularization. In fact, if $Q(p^{(k+1)}) \geq Q(p^{(k)})$, then $\lambda^{(k+1)} > \lambda^{(k)}$, thereby reducing the step length. The algorithm effectively eliminates oscillation of the estimates, although convergence is relatively slow and we do not expect to attain the minimum of $Q$ provided by the ILS estimates. Our aim, however, is not consistency, usually accompanied by large variance and mean square error, but a reasonable compromise between the DILS and ILS estimates, reducing the bias while avoiding consequences of near-linear interdependence. According to the simulation studies presented in this section, the SB estimator serves this purpose efficiently.

The DILS, SB and ILS estimators can be considered as stages, increasing in computational complexity, of a multistage procedure. In most cases, the DILS stage is a robust way of obtaining good initial estimates with low computational effort. The SB stage removes bias and minimizes the indirect objective function $Q$ with a reduced number of equivalent function evaluations (reduced because no parameter sensitivities are required). Finally, the ILS stage is the conventional estimator and provides the approximate covariance matrix $s_r^2[\mathbf{J}^T(\hat{p})\,\mathbf{W}\mathbf{J}(\hat{p})]^{-1}$ of the estimates, where $s_r^2$ is the residual variance.

For programming the procedure, we use four subroutines, as follows:

(*i*) Response function evaluation by solution of the differential equations.
(*ii*) Response function evaluation by spline approximation.
(*iii*) Evaluation of the Jacobian by solution of the sensitivity equations.
(*iv*) Spline-function approximation of the Jacobian.

In general-purpose Gauss–Marquardt procedures, subroutines (*ii*) and (*iv*) give the DILS, (*i*) and (*iv*) the SB and (*i*) and (*iii*) the ILS methods.

The result of the procedure is most statisfactory if the estimates in the different stages do not change very much. Significant differences between DILS and ILS estimates reveal uncertainties stemming from the data (large measurement errors and/or samples too widely spaced), or from the model (near-singularity of the

Hessian matrix in this ILS approach) or from the DILS method itself (due, for example, to an inadequate prediction of the response functions by the spline functions). The procedure offers statistics for selecting the best estimate and giving an indication of when it is possible to skip the most complex ILS stage.

Our examples show that the DILS method is usually preferable if the following considerations hold:

(a) $Q_{\mathrm{DILS}}(\underline{\hat{p}}) = Q\underline{\hat{p}})$, where $\underline{\hat{p}}$ denotes the DILS estimates. (The errors in interpolating noisy data by spline functions are then relatively small.)
(b) $S_{\mathrm{ILS}} > S_{\mathrm{DILS}}$, where $S$ is defined by Eq. (2.61). (An increased variance of the estimates in the ILS stage can then be expected.)
(c) $\lambda^{\min}_{\mathrm{ILS}} < \lambda^{\min}_{\mathrm{DILS}}$. (An increased mean square error in the ILS stage can then be expected.)

The DILS method has no statistical advantages if (b) and (c) are not satisfied and it is certainly not adequate if $Q(\hat{p})$ is considerably greater (by a factor of two) say than $Q_{\mathrm{DILS}}(\hat{p})$, which is an indication that the spline function predictions are unacceptably poor. This may occur due both to low signal-to-noise ratio and to the form of the true responses. The SB method can reduce the resulting bias of the estimates and may give the least variance and mean square error if (b) and (c) are satisfied. If $Q_{\mathrm{DILS}}(\hat{p}) \ll Q(\hat{p})$, the DILS and SB results can often be improved by using piecewise linear interpolation instead of spline functions, since the latter are more sensitive to large errors in the observations.

**Example 2.8.** *Solving Eakman's test problem by the DILS method*

Consider the problem discussed in Example 2.6. If all five components are observed, it can also be solved by the DILS method. (In fact, if we know the initial concentrations from the observation of two appropriately selected components, e. g., $A$ and $B$ we can compute all the other concentrations from the stoichiometry, and the DILS method applies, after some preliminary calculations.) Since the differential equation is linear in the parameters, the DILS method is reduced to a multiresponse linear regression and hence it is very easy to perform. With no weighting it yields the estimates $k_1 = 0.30826 \pm 0.00227$ and $k_2 = 0.10744 \pm 0.00227$, which differ less than by 1% from the indirect least squares estimates. Thus in this example the DILS method is appropriate and results in a considerable reduction of the computational effort.

The remainder of this section is devoted to a comparative analysis of DILS, SB and ILS estimators by solving from more examples. All examples are based on published sets of data. In order to study the statistical properties, further data sets were generated for each example by adding normally distributed errors to model response values, i. e. to the values of the response of the model with the published parameter values, at the sampling instants. For these sets, the DILS, SB and ILS

estimates of the individual parameters $p_i$ are compared in terms of the mean $\bar{p}_i$, bias $b_i$ and standard deviation $s_i$ defined by:

$$\bar{p}_i = \frac{1}{N}\sum_{j=1}^{N} \hat{p}_{ij}$$

$$b_i = \bar{p}_i - p_i$$

$$s_i = \left[\frac{1}{N-1}\sum_{j=1}^{N} (\hat{p}_{ij} - \bar{p}_i)^2\right]^{\frac{1}{2}}$$

where $\hat{p}_{ij}$ denotes the estimate of $p_i$ in run $j$, $j = 1, 2, \ldots, N$.

The entire vector $\hat{p}$ of parameters is described in terms of two mean square distances:

$$L_1^2 = \frac{1}{N}\sum_{j=1}^{N} (\hat{p}_j - \bar{p})^T (\hat{p}_j - \bar{p})$$

$$L_2^2 = \frac{1}{N}\sum_{j=1}^{N} (\hat{p}_j - p)^T - (\hat{p}_j - p)$$

where $\hat{p}_j$ is the vector of estimates in run $j$. $L_1^2$ is a measure of variance, whereas $L_2^2$ is an estimate of the mean square error, $E(L^2)$. In practical terms, $L_2^2$ is the more important of the two measures. In all examples, the number of runs, $N$, is 5.

**Example 2.9.** *Hydrogenation of fatty oils*

The data are generated from a model for the hydrogenation of fatty oils in a catalytic reactor (Yermakova and Umbetov) [57]:

$$\dot{x}_1 = p_6(p_1 + p_2)x_1/g$$

$$\dot{x}_2 = p_6[p_1 x_1 - (p_3 + p_4)x_2 + 1.5p_3 x_3]/g$$

$$\dot{x}_3 = p_6[p_2 x_1 + p_3 x_2 - (1.5p_3 + p_5)x_3]/g$$

$$\dot{x}_4 = -(\dot{x}_1 + \dot{x}_2 + \dot{x}_3)$$

where $g = p_6 + (p_1 + p_2)x_1 + p_4 x_2 + 1.5p_5 x_3$.

The data set is obtained by adding 3% normally distributed errors with initial conditions $x_0 = [1, 0, 0, 0]^T$ and is shown in *Table 2.14*, which is taken from Vajda, Valkó and Yermakova [54].

Parameter estimates using the three methods are given in *Table 2.15*; in all cases, it has been assumed that $W = I$. The fit of the model with DILS estimates of the parameters is shown in *Fig. 2.10*.

Conditions (a) to (c) are again all satisfied and $S_{ILS} \gg S_{DILS}$, the eigenvalues for the DILS estimator being 10.1, 3.19, 2.89, 0.31, 0.12 and $7.9 \times 10^{-4}$ and those for the ILS estimator being 3.87, 0.20, 0.18, 0.10, $1.52 \times 10^{-2}$ and

Table 2.14. Concentrations in Example 2.9

| $t$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|
| 0. | 1.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.25 | 0.8081 | 0.6979E−1 | 0.7802E−1 | 0.2219E−2 |
| 0.50 | 0.7052 | 0.1292 | 0.1548 | 0.8828E−2 |
| 0.75 | 0.5689 | 0.1820 | 0.2244 | 0.2129E−1 |
| 1.00 | 0.4614 | 0.2379 | 0.2714 | 0.3816E−1 |
| 1.50 | 0.2586 | 0.2968 | 0.3291 | 0.9069E−1 |
| 2.00 | 0.1424 | 0.3633 | 0.3583 | 0.1731 |
| 2.50 | 0.7002E−1 | 0.3335 | 0.3479 | 0.2503 |
| 3.00 | 0.3153E−1 | 0.3136 | 0.3117 | 0.3335 |
| 4.00 | 0.5361E−2 | 0.2593 | 0.2239 | 0.5225 |
| 5.00 | 0.8103E−3 | 0.1800 | 0.1498 | 0.6465 |
| 6.00 | 0.1096E−3 | 0.134 | 0.1089 | 0.7657 |
| 7.00 | 0.1376E−4 | 0.8930E−1 | 0.7447E−1 | 0.8255 |
| 8.00 | 0.1783E−5 | 0.6052E−1 | 0.4963E−1 | 0.9151 |
| 10.00 | 0.2262E−7 | 0.2783E−1 | 0.2471E−1 | 0.9816 |

$2.39 \times 10^{-4}$. Thus we expect that, for the simulated data, the DILS estimates will be significantly better than the ILS estimates, in terms of both $L_1^2$ and $L_2^2$.

This is confirmed in *Table 2.16*, which summarizes results from additional data sets with 5% and 10% relative errors added to the model response values. It can

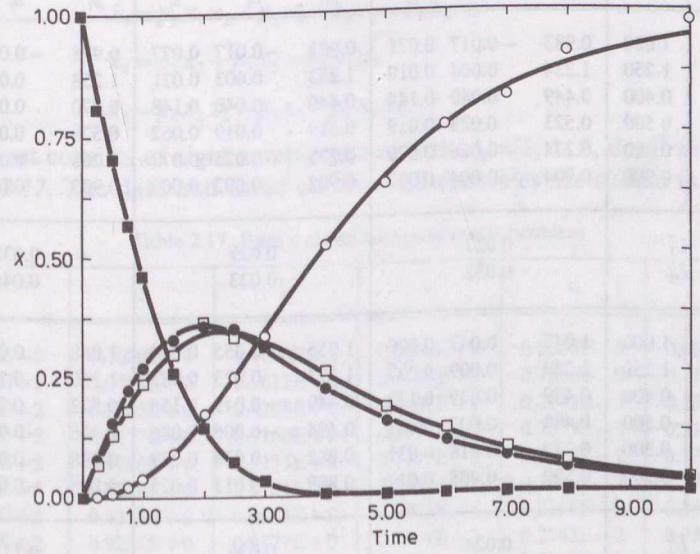



Fig. 2.10. Model fit (DILS and ILS) in Example 2.9. The observations are denoted by symbols: ■ – $X_1$, □ – $X_2$ ● – $X_3$ and ○ – $X_4$

Table 2.15. Estimates of rate constants in Example 2.9

| Parameter | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | $p_6$ | Objective function |
|---|---|---|---|---|---|---|---|
| Starting value | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | $Q_{\text{DILS}}=1.034\text{E}+1$ |
| DILS | 1.072 | 1.328 | 0.422 | 0.514 | 0.303 | 0.867 | $Q_{\text{DILS}}=4.884\text{E}-3$<br>$Q_{\text{ILS}}=4.850\text{E}-3$ |
| SB | 1.072 | 1.328 | 0.422 | 0.514 | 0.303 | 0.867 | $Q_{\text{ILS}}=4.850\text{E}-3$ |
| ILS | 1.094 | 1.344 | 0.405 | 0.514 | 0.303 | 0.860 | $Q_{\text{ILS}}=4.805\text{E}-3$ |
| Standard deviation | 0.071 | 0.061 | 0.194 | 0.082 | 0.089 | 0.025 | — |

also be seen from *Table 2.16* that the SB estimates are much better than the ILS estimates, but are not as good as the DILS estimates.

**Example 2.10.** *Bard's mass action kinetic test problem*

The data are from test problem 3 of a model of chemical reaction kinetics (Bard

Table 2.16. Estimate averages, biases and standard deviations of rate constants in Example 2.9 from 5 simulated data sets

| % Error | $p_i$ | True value | DILS | | | SB | | | ILS | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\bar{p}_i$ | $b_i$ | $s_i$ | $\bar{p}_i$ | $b_i$ | $s_i$ | $\bar{p}_i$ | $b_i$ | $s_i$ |
| 5 | $p_1$ | 1.000 | 0.983 | −0.017 | 0.071 | 0.983 | −0.017 | 0.077 | 0.978 | −0.022 | 0.074 |
| | $p_2$ | 1.250 | 1.254 | 0.004 | 0.019 | 1.253 | 0.003 | 0.021 | 1.258 | 0.008 | 0.039 |
| | $p_3$ | 0.400 | 0.449 | 0.049 | 0.144 | 0.440 | 0.040 | 0.148 | 0.470 | 0.070 | 0.174 |
| | $p_4$ | 0.500 | 0.523 | 0.023 | 0.059 | 0.519 | 0.019 | 0.062 | 0.528 | 0.028 | 0.073 |
| | $p_5$ | 0.300 | 0.274 | −0.026 | 0.070 | 0.275 | −0.025 | 0.072 | 0.268 | −0.032 | 0.071 |
| | $p_6$ | 0.900 | 0.904 | 0.004 | 0.007 | 0.902 | 0.002 | 0.009 | 0.903 | 0.003 | 0.008 |
| | $L_1^2$ | | 0.027 | | | 0.029 | | | 0.038 | | |
| | $L_2^2$ | | 0.032 | | | 0.033 | | | 0.046 | | |
| 10 | $p_1$ | 1.000 | 1.047 | 0.047 | 0.079 | 1.055 | 0.055 | 0.084 | 1.068 | 0.068 | 0.131 |
| | $p_2$ | 1.250 | 1.259 | 0.009 | 0.037 | 1.273 | 0.023 | 0.083 | 1.350 | 0.100 | 0.071 |
| | $p_3$ | 0.400 | 0.439 | 0.039 | 0.133 | 0.449 | 0.049 | 0.156 | 0.522 | 0.122 | 0.314 |
| | $p_4$ | 0.500 | 0.488 | −0.002 | 0.051 | 0.494 | −0.006 | 0.086 | 0.496 | −0.004 | 0.146 |
| | $p_5$ | 0.300 | 0.282 | −0.018 | 0.031 | 0.282 | −0.018 | 0.029 | 0.288 | −0.012 | 0.058 |
| | $p_6$ | 0.900 | 0.892 | −0.008 | 0.017 | 0.889 | −0.011 | 0.025 | 0.893 | −0.007 | 0.027 |
| | $L_1^2$ | | 0.024 | | | 0.036 | | | 0.117 | | |
| | $L_2^2$ | | 0.028 | | | 0.043 | | | 0.147 | | |

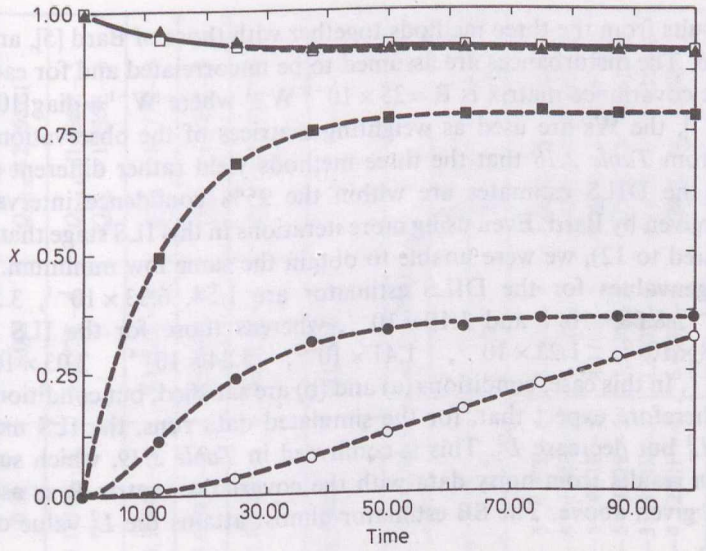



Fig. 2.11. Model fit (ILS) in Example 2.10. The observations are denoted by the symbols: □ – $X_1$, ▲ – $X_2$, ■ – $10X_3$, ○ – $10^4 X_4$ and ● – $10^3 X_5$

[5]) described by

$$\dot{x}_1 = -p_1 x_1 x_2 + p_2 x_3$$

$$\dot{x}_2 = -p_1 x_1 x_2 + p_2 x_3 - p_4 x_2 x_3 + p_5 x_5 - p_6 x_2 x_4$$

$$\dot{x}_3 = p_1 x_1 x_2 - p_2 x_3 - p_3 x_3 - p_4 x_2 x_3$$

$$\dot{x}_4 = p_3 x_3 + p_5 x_5 - p_6 x_2 x_4$$

$$\dot{x}_5 = p_4 x_2 x_3 - p_5 x_5 + p_6 x_2 x_4$$

The data set consists of eight sample points with $x_0 = [1, 1, 0, 0, 0]^T$ and is given in *Table 2.17*. The data and fitted curves (ILS estimate) are shown in *Fig. 2.11*.

Table 2.17. Bard's chemical reaction test problem

| $t$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| 0.000E+0 | 0.1000E+1 | 0.1000E+1 | 0.0000E+0 | 0.0000E+0 | 0.0000E+0 |
| 0.125E+2 | 0.9458E+0 | 0.9612E+0 | 0.4949E−1 | 0.1550E−3 | 0.1115E−3 |
| 0.250E+2 | 0.9265E+0 | 0.9288E+0 | 0.6905E−1 | 0.3145E−3 | 0.2363E−3 |
| 0.375E+2 | 0.9177E+0 | 0.9160E+0 | 0.7518E−1 | 0.7093E−3 | 0.3117E−3 |
| 0.500E+2 | 0.9290E+0 | 0.9175E+0 | 0.7716E−1 | 0.1192E−2 | 0.3331E−3 |
| 0.625E+2 | 0.9278E+0 | 0.9201E+0 | 0.7809E−1 | 0.1688E−2 | 0.3403E−3 |
| 0.750E+2 | 0.9253E+0 | 0.9123E+0 | 0.7905E−1 | 0.2195E−2 | 0.3568E−3 |
| 0.875E+2 | 0.9251E+0 | 0.9177E+0 | 0.7839E−1 | 0.2742E−2 | 0.3583E−3 |
| 0.100E+3 | 0.9173E+0 | 0.9075E+0 | 0.7793E−1 | 0.3200E−2 | 0.3620E−3 |

The results from the three methods together with those of Bard [5], are given in *Table 2.18*. The disturbances are assumed to be uncorrelated and for each sample point, the covariance matrix is $\mathbf{R}=25\times10^{-8}\,\mathbf{W}^{-1}$ where $\mathbf{W}^{-1}=\text{diag}[100, 100, 1, 10^{-2}, 10^{-4}]$; the **W**s are used as weighting matrices of the observations. It may be seen from *Table 2.18* that the three methods yield rather different estimates, although the DILS estimates are within the 95% confidence intervals of the estimates given by Bard. Even using more iterations in this ILS stage than Bard (15 as compared to 12), we were unable to obtain the same low minimum.

The eigenvalues for the DILS estimator are 1.54, $6.33\times10^{-1}$, $3.51\times10^{-3}$, $2.28\times10^{-3}$, $8.68\times10^{-6}$ and $1.19\times10^{-7}$, whereas those for the ILS estimator are $7.28\times10^{-2}$, $1.23\times10^{-2}$, $1.41\times10^{-3}$, $5.84\times10^{-4}$, $2.03\times10^{-6}$ and $2.85\times10^{-7}$. In this case, conditions (a) and (b) are satisfied, but condition (c) is not and we therefore expect that, for the simulated data runs, the ILS method will increase $L_1^2$ but decrease $L_2^2$. This is confirmed in *Table 2.19*, which summarizes simulation results from noisy data with the covariance matrix **R** as assumed by Bard and given above. The SB estimator almost attains the $L_2^2$ value of the ILS estimator.

**Example 2.11.** *Microbial growth process in a batch reactor*

The model in this example concerns a microbial growth process (Nihtilä and Virkkunen [38]) described by

$$\dot{x}_1=\frac{p_1(x_2-p_5)x_1}{p_2+x_2-p_5}-p_3x_1;\qquad x_1(0)=p_6$$

$$\dot{x}_2=-\frac{1}{p_4}\frac{p_1(x_2-p_5)x_1}{p_2+x_2-p_5};\qquad x_2(0)=p_7+p_5$$

where $p_5$ represents the constant bias in $x_2$. Parameters are estimated from 16 samples given in *Table 2.20*.

Results obtained by the three methods are given in *Table 2.21*, and are compared with those of Nihtilä and Virkkunen [38]. In this example, conditions (b) and (c) are satisfied but condition (a) is not. The spline functions fitted to noisy data are unable to predict the response function adequately, and we see from *Table 2.21* that the $Q(\hat{p})$ computed at the DILS estimates is more than twice as large as $Q_{\text{DILS}}(\hat{p})$. The model fit is shown in *Fig. 2.12*.

The additional data sets in this example are obtained with absolute errors having standard deviations of 0.2 and 0.6. This corresponds to additional data sets generated by Nihtilä and Virkkunen [38]. The results presented in *Table 2.22* confirm that, using spline predictions, the DILS estimator performs very poorly and even the SB stage is unable to remove bias, since the approximate Jacobian $\hat{\mathbf{J}}(p)$ is also poor.

The problem is avoided by using piecewise linear interpolation instead of splines, as illustrated in *Table 2.23*. For the data with standard deviation of 0.2, the DILS

Table 2.18. Results of Example 2.10 using Bard's data

| Parameter | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | $p_6$ | Objective function |
|---|---|---|---|---|---|---|---|
| Initial value | 0.01 | 0.01 | 0.001 | 0.001 | 0.02 | 0.001 | $Q_{DILS}=2.395E1$ |
| DILS | 6.1938E−3 | 6.5911E−2 | 9.4461E−6 | 5.5166E−4 | 1.1616E−1 | 9.1182E−4 | $Q_{DILS}=1.898E-5$<br>$Q_{ILS}=1.235E-5$ |
| SB | 6.3660E−3 | 6.7951E−2 | 9.4467E−6 | 5.4807E−4 | 1.1554E−1 | 9.1182E−4 | $Q_{ILS}=1.177E-5$ |
| ILS | 6.3612E−3 | 6.7783E−2 | 5.4754E−5 | 4.9963E−4 | 1.0350E−1 | 5.5108E−4 | $Q_{ILS}=1.075E-5$ |
| Reported by Bard | 6.3581E−3 | 6.7744E−2 | 5.9163E−5 | 4.9438E−4 | 1.0188E−1 | 4.2025E−4 | $Q_{ILS}=1.072E-5$ |
| Standard deviation | 9.4344E−5 | 1.1368E−3 | 2.8628E−5 | 3.0760E−5 | 8.3744E−3 | 4.8365E−4 | — |

Table 2.19. Estimate averages, biases and standard deviations from 5 simulation runs in Example 2.10

| $p_i$ | True value | DILS | | | SB | | | ILS | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\bar{p}_i$ | $b_i$ | $s_i$ | $\bar{p}_i$ | $b_i$ | $s_i$ | $\bar{p}_i$ | $b_i$ | $s_i$ |
| $p_1$ | 6.358E−3 | 6.180E−3 | −1.783E−4 | 1.17E−4 | 6.371E−3 | 1.265E−5 | 1.57E−4 | 6.342E−3 | −1.624E−5 | 1.43E−4 |
| $p_2$ | 6.774E−2 | 6.575E−2 | −1.986E−3 | 1.53E−3 | 6.800E−2 | 2.590E−4 | 1.94E−3 | 6.751E−2 | −2.336E−4 | 1.69E−3 |
| $p_3$ | 5.916E−5 | 1.838E−5 | −4.078E−5 | 2.58E−5 | 3.005E−5 | −2.911E−5 | 4.03E−5 | 4.116E−5 | −1.800E−5 | 4.05E−5 |
| $p_4$ | 4.944E−4 | 5.399E−4 | 4.550E−5 | 2.59E−5 | 5.233E−4 | 2.889E−5 | 3.92E−5 | 5.119E−4 | 1.750E−5 | 4.05E−5 |
| $p_5$ | 1.019E−1 | 1.121E−1 | 1.022E−2 | 5.28E−3 | 1.080E−1 | 6.163E−3 | 9.11E−3 | 1.063E−1 | 4.467E−3 | 9.95E−3 |
| $p_6$ | 4.203E−4 | 5.833E−4 | 1.631E−4 | 2.16E−4 | 5.708E−4 | 1.505E−4 | 2.81E−4 | 6.403E−4 | 2.200E−4 | 4.00E−4 |
| $L_1^2$ | | 2.425E−5 | | | 6.951E−5 | | | 8.169E−5 | | |
| $L_2^2$ | | 1.329E−4 | | | 1.076E−4 | | | 1.018E−4 | | |

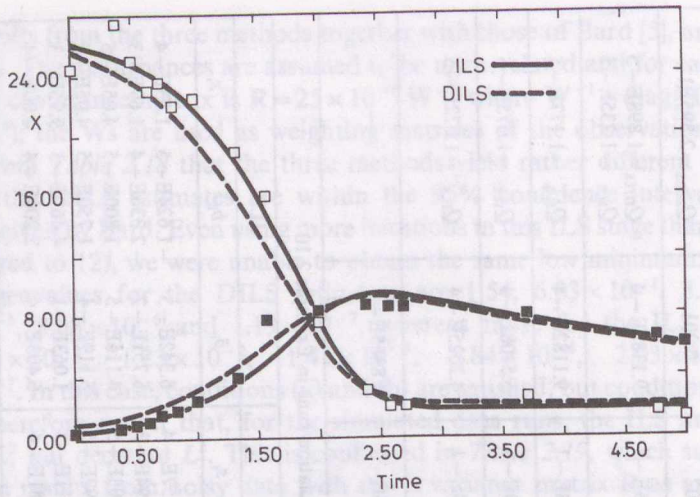



Fig. 2.12. Comparison DILS and ILS estimation in Example 2.11. The observations are denoted by the symbols: ■ – $X_1$ and □ – $X_2$

method gives the best estimates, whereas for the standard deviation of 0.6, the SB estimator is better than either the DILS or ILS estimators.

The computational effort required for estimating the parameters from the published sets of data in Examples 2.9 to 2.11 is summarized in *Table 2.24*. For all cases, the termination condition was $||\underline{p}^{(k+1)} - \underline{p}^{(k)}||/||\underline{p}^{(k)}|| < 10^{-4}$ and the number of iterations was restricted to 15.

Table 2.20. Concentrations in a microbial growth process

| Time | $x_1$ | $x_2$ |
|---|---|---|
| 0.000E0 | 0.4000E0 | 0.2450E2 |
| 0.375E0 | 0.7000E0 | 0.2750E2 |
| 0.500E0 | 0.9000E0 | 0.2500E2 |
| 0.625E0 | 0.1100E1 | 0.2300E2 |
| 0.708E0 | 0.1500E1 | 0.2350E2 |
| 0.833E0 | 0.1800E1 | 0.2200E2 |
| 0.100E1 | 0.2500E1 | 0.2250E2 |
| 0.133E1 | 0.3900E1 | 0.1900E2 |
| 0.158E1 | 0.7700E1 | 0.1600E2 |
| 0.200E1 | 0.8200E1 | 0.7500E1 |
| 0.240E1 | 0.8700E1 | 0.3000E1 |
| 0.258E1 | 0.8500E1 | 0.2000E1 |
| 0.271E1 | 0.8500E1 | 0.2000E1 |
| 0.371E1 | 0.7900E1 | 0.2500E1 |
| 0.477E1 | 0.6400E1 | 0.1600E1 |
| 0.500E1 | 0.6000E1 | 0.1200E1 |

Table 2.21. Results for Example 2.11

| Parameter | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | $p_6$ | $p_7$ | Objective function |
|---|---|---|---|---|---|---|---|---|
| Initial value | 2.50 | 9.00 | 0.200 | 0.460 | 1.80 | 0.500 | 22.0 | $Q_{DILS}=5.048E2$ |
| DILS | 2.09 | 8.48 | 0.206 | 0.469 | 2.06 | 0.886 | 23.6 | $Q_{DILS}=2.378E1$<br>$Q_{ILS}=5.558E1$ |
| SB | 2.33 | 9.28 | 0.207 | 0.461 | 1.87 | 0.589 | 24.5 | $Q_{ILS}=2.358E1$ |
| ILS | 2.50 | 9.30 | 0.202 | 0.468 | 1.81 | 0.468 | 24.3 | $Q_{ILS}=2.268E1$ |
| Reported by Nihtilä and Virkkunen | 2.51 | 9.34 | 0.203 | 0.468 | 1.81 | 0.468 | 24.3 | $Q_{ILS}=2.268E1$ |
| Standard deviation | 1.04 | 8.04 | 0.051 | 0.047 | 0.50 | 0.203 | 0.5 | — |

Table 2.22. Estimate averages, biases and standard deviation from five simulated data sets in Example 2.11

| SD | $p_i$ | True value | DILS | | | SB | | | ILS | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\bar{p}_i$ | $b_i$ | $s_i$ | $\bar{p}_i$ | $b_i$ | $s_i$ | $\bar{p}_i$ | $b_i$ | $s_i$ |
| 0.2 | $p_1$ | 2.51 | 2.80 | 0.29 | 0.20 | 2.80 | 0.29 | 0.20 | 2.71 | 0.20 | 0.20 |
| | $p_2$ | 9.34 | 10.96 | 1.62 | 1.41 | 10.84 | 1.50 | 1.38 | 10.40 | 1.06 | 1.38 |
| | $p_3$ | 0.203 | 0.206 | 0.003 | 0.012 | 0.207 | 0.004 | 0.01 | 0.207 | 0.004 | 0.01 |
| | $p_4$ | 0.468 | 0.473 | 0.005 | 0.01 | 0.473 | 0.005 | 0.01 | 0.473 | 0.005 | 0.01 |
| | $p_5$ | 1.81 | 1.86 | 0.05 | 0.17 | 1.84 | 0.03 | 0.14 | 1.77 | −0.04 | 0.12 |
| | $p_6$ | 0.468 | 0.388 | −0.080 | 0.06 | 0.408 | −0.060 | 0.03 | 0.426 | −0.042 | 0.02 |
| | $p_7$ | 24.30 | 24.36 | 0.06 | 0.29 | 24.32 | 0.02 | 0.15 | 24.29 | −0.01 | 0.11 |
| | $L_1^2$ | | 1.712 | | | 1.595 | | | 1.584 | | |
| | $L_2^2$ | | 4.426 | | | 3.922 | | | 2.754 | | |
| 0.6 | $p_1$ | 2.51 | 3.99 | 1.48 | 2.29 | 4.04 | 1.52 | 2.36 | 2.54 | 0.03 | 0.26 |
| | $p_2$ | 9.34 | 20.08 | 10.74 | 14.91 | 19.76 | 10.42 | 14.77 | 9.02 | −0.32 | 2.20 |
| | $p_3$ | 0.203 | 0.210 | 0.007 | 0.03 | 0.207 | 0.004 | 0.03 | 0.220 | 0.017 | 0.02 |
| | $p_4$ | 0.468 | 0.458 | −0.010 | 0.40 | 0.457 | −0.011 | 0.03 | 0.470 | 0.002 | 0.01 |
| | $p_5$ | 1.81 | 1.84 | 0.03 | 0.34 | 2.01 | 0.20 | 0.53 | 1.87 | 0.06 | 0.17 |
| | $p_6$ | 0.468 | 0.370 | −0.098 | 0.28 | 0.388 | −0.080 | 0.16 | 0.457 | −0.011 | 0.07 |
| | $p_7$ | 24.30 | 24.75 | 0.45 | 0.76 | 24.57 | 0.27 | 0.46 | 24.45 | 0.15 | 0.14 |
| | $L_1^2$ | | 182.70 | | | 179.48 | | | 3.989 | | |
| | $L_2^2$ | | 300.49 | | | 290.59 | | | 4.121 | | |

Table 2.23. Estimate averages, biases and standard deviations from five simulated data sets in Example 2.11 using piecewise linear interpolation

| SD | $p_i$ | True value | DILS | | | SB | | | ILS | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\bar{p}_i$ | $b_i$ | $s_i$ | $\bar{p}_i$ | $b_i$ | $s_i$ | $\bar{p}_i$ | $b_i$ | $s_i$ |
| 0.2 | $p_1$ | 2.51 | 2.64 | 0.13 | 0.18 | 2.65 | 0.14 | 0.19 | 2.71 | 0.20 | 0.20 |
| | $p_2$ | 9.34 | 9.29 | −0.05 | 1.03 | 9.69 | 0.35 | 1.29 | 10.40 | 1.06 | 1.38 |
| | $p_3$ | 0.203 | 0.208 | 0.005 | 0.01 | 0.208 | 0.005 | 0.01 | 0.207 | 0.004 | 0.01 |
| | $p_4$ | 0.468 | 0.473 | 0.005 | 0.01 | 0.474 | 0.006 | 0.01 | 0.473 | 0.005 | 0.01 |
| | $p_5$ | 1.81 | 1.89 | 0.08 | 0.14 | 1.81 | 0.00 | 0.13 | 1.77 | −0.04 | 0.12 |
| | $p_6$ | 0.468 | 0.399 | −0.069 | 0.06 | 0.424 | −0.044 | 0.03 | 0.426 | −0.042 | 0.02 |
| | $p_7$ | 24.30 | 24.34 | 0.04 | 0.23 | 24.27 | −0.03 | 0.12 | 24.29 | −0.01 | 0.11 |
| | $L_1^2$ | | | 0.928 | | | 1.391 | | | 1.584 | |
| | $L_2^2$ | | | 0.960 | | | 1.539 | | | 2.754 | |
| 0.6 | $p_1$ | 2.51 | 2.68 | 0.17 | 0.64 | 2.69 | 0.18 | 0.65 | 2.54 | 0.03 | 0.26 |
| | $p_2$ | 9.34 | 8.80 | −0.54 | 2.34 | 9.13 | −0.21 | 2.04 | 9.02 | −0.32 | 2.20 |
| | $p_3$ | 0.203 | 0.277 | 0.024 | 0.02 | 0.226 | 0.023 | 0.02 | 0.220 | 0.017 | 0.02 |
| | $p_4$ | 0.468 | 0.476 | 0.008 | 0.01 | 0.477 | 0.009 | 0.01 | 0.470 | 0.002 | 0.01 |
| | $p_5$ | 1.81 | 1.97 | 0.16 | 0.24 | 1.92 | 0.11 | 0.20 | 1.87 | 0.06 | 0.17 |
| | $p_6$ | 0.468 | 0.387 | −0.081 | 0.22 | 0.431 | −0.037 | 0.15 | 0.457 | −0.011 | 0.07 |
| | $p_7$ | 24.30 | 24.50 | 0.20 | 0.56 | 24.31 | 0.01 | 0.35 | 24.45 | 0.15 | 0.14 |
| | $L_1^2$ | | | 5.046 | | | 3.841 | | | 3.989 | |
| | $L_2^2$ | | | 5.434 | | | 3.932 | | | 4.121 | |

The left hand part of the Table shows the number of iterations in the separate steps of the multistage procedure. Example 2.10 involves a model that is linear in the parameters, so that the DILS estimates could, in principle, be obtained in a single iteration, but, gradually decreasing the parameter $\lambda^{(k)}$, the Marquardt algorithm requires a few more steps to attain the termination condition. The number of iterations in the SB stage can also be considered as an indicator for the applicability of the DILS method. For Example 2.9, the DILS estimator is definitely preferable. Using the published data, the SB stage is terminated in the first iteration and the estimates are unchanged as shown in *Table 2.15*. In some simulation runs for the same example, up to three steps have been taken.

The maximum allowed number of iterations was performed in both Example 2.10, where the DILS method is not preferable, and Example 2.11, where the number of iterations was reduced.

The columns in the right-hand part of *Table 2.24* show the number of equivalent function evaluations successively in the DILS and SB stages, in all three stages and in solving the problem by the ILS method alone. Using only the first two stages, the number of equivalent function evaluations is equal to the number of iterations

Table 2.24. Computational effort in Examples 2.9 to 2.11

| Example | Number of iterations | | | Number of equivalent function evaluations | | |
|---|---|---|---|---|---|---|
| | DILS | SB | ILS | DILS+SB | 3 stages | ILS only |
| 2.9 | 5 | 1 | 5 | 1 | 36 | 84 |
| 2.10 | 5 | 15 | 15 | 15 | 141 | 154 |
| 2.11 | 6 | 15 | 6 | 15 | 63 | 56 |

in the SB stage. From our results, the SB estimates were quite satisfactory in all examples, and comparison with the last column of *Table 2.24* emphasizes the numerical efficiency of the direct approach involving the DILS and SB estimators.

## 2.3 Model reduction

As a consequence of the experimenter's desire to incorporate all important physical phenomena, models based on fundamental principles are generally rather complex. On the other hand, a model is of practical value only if it can be validated and if its parameters can be estimated with reasonably low variances. Therefore, model reduction is a crucial step in the modelling process.

Scientific disciplines, such as chemical kinetics, have elaborated several intuitive tools for model reduction. However, it is always difficult to measure the validity of such approximations in qualitative terms. In this section we show how principal component analysis can serve as a quantitative basis for model reduction. This approach has been suggested by Vajda, Valkó and Turányi [53]. It is found to be particularly useful in chemical kinetic applications.

### 2.3.1 Principal component analysis of nonlinear regression models

#### Parameter transformations

First we recall some results of the previous sections, which will be useful in our further discussion. In Example 2.1 a linear regression model was investigated. The eigenvalue-eigenvector analysis of the $\mathbf{X}^T\mathbf{X}$ matrix was carried out without scaling and with two different scalings. The eigenvalues and eigenvectors provided a realistic picture of the relative importance of the parameters if the columns of $\mathbf{X}^T\mathbf{X}$ were scaled to yield unit transformed regression coefficients. In other words the parameter transformation

$$\beta_j = p_j/\hat{p}_j \tag{2.66}$$

was used before carrying out the principal component analysis, where $\hat{p}_j$ denotes the least squares estimate of $p_j$.

In Section 2.2.3, which dealt with the Gauss–Newton–Marquardt algorithm of nonlinear parameter estimation, we introduced a new set of parameters $\beta^{(k)}$ at each iteration step $k$:

$$\beta_j^{(k)} = p_j^{(k)} / p_j^{(k-1)}, \qquad j = 1, 2, \ldots, np \tag{2.67}$$

The Jacobian $\mathbf{J}(\underline{\beta})$ with respect to the new parameters could be obtained in a simple way by applying Eq.(2.33). When convergence was reached, all the parameters $\beta_j^{(k)}$ were unity. Therefore, we may expect similar information from the eigenvalue-eigenvector decomposition of the $\mathbf{J}^T(\underline{\beta})\,\mathbf{W}\mathbf{J}(\underline{\beta})$ matrix in the nonlinear case as we obtained from the same decomposition of the properly scaled $\mathbf{X}^T\mathbf{W}\mathbf{X}$ matrix in the linear case.

If we want to carry out model reduction independently of the parameter estimation procedure, it is advantageous to substitute Eq.(2.67) by

$$\beta_j = p_j / p_j^0, \qquad j = 1, 2, \ldots, np \tag{2.68}$$

where $p_j^0$ is a suitable nominal parameter value. Nominal values can be estimated from previous attempts to fit the model to observations or values taken from the literature, obtained by analogy, and so on.

Often we use the logarithm of $\beta_j$, i.e. the parameter transformation

$$\alpha_j = \ln (p_j / p_j^0), \qquad j = 1, 2, \ldots, np \tag{2.69}$$

where actual and nominal parameters are supposed to have the same sign.

We are interested in model behaviour in the vicinity of $\alpha_j^0 = 0$ and $\beta_j^0 = 1$, $(j = 1, 2, \ldots, np)$.

### Response surfaces

Let $y_{ij}(\alpha)$ denote the model response for observed variable $y_i$ at measurement point $j$. The effect on the calculated behaviour of the model brought about by a variation in the parameters may be expressed in terms of the response surface defined by

$$Q(\alpha) = \sum_{j=1}^{nm} \sum_{i=1}^{ny} [y_{ij}(\alpha) - y_{ij}(\alpha^0)]^2 \, w_{ij} \tag{2.70}$$

where the weights $w_{ij}$ are selected according to

$$w_{ij} = 1 / [y_{ij}(\alpha^0)]^2 \tag{2.71}$$

Since $Q(\alpha)$ and its gradient vector are zero at $\alpha^0$, expansion in a Taylor series up to second order terms yields

$$Q(\alpha) \approx \frac{1}{2} \underline{\alpha}^T \mathbf{H}(\alpha^0) \underline{\alpha} \tag{2.72}$$

where $\mathbf{H}$ is the Hessian of $Q$ with respect to $\alpha$, i.e. the matrix of second partial derivatives.

Similarly, using the transformed parameters [Eq.(2.68)] the response function can be approximated by

$$Q(\beta) \approx \frac{1}{2} \Delta\beta^{\mathrm{T}} \mathbf{H}(\underline{\beta}^0) \Delta\beta \tag{2.73}$$

At the nominal arguments the Hessian can be constructed from the Jacobian:

$$\mathbf{H}(\alpha^0) = 2\mathbf{J}^{\mathrm{T}}(\alpha^0)\,\mathbf{W}\mathbf{J}(\underline{\alpha}^0) \tag{2.74}$$

and

$$\mathbf{H}(\beta^0) = 2\mathbf{J}^{\mathrm{T}}(\beta^0)\,\mathbf{W}\mathbf{J}(\underline{\beta}^0) \tag{2.75}$$

since the other terms otherwise present are multiplied by zero at $\alpha = \alpha^0$ and $\beta = \beta^0$.

It is important to stress that at the nominal arguments

$$\frac{\partial y}{\partial \alpha_j} = \frac{\partial y}{\partial \beta_j} = \frac{\partial y}{\partial p_j} p_j^0 \tag{2.76}$$

thus locally both transformations yield the same Jacobian matrix $\mathbf{J}$.

Now we have two quadratic expressions

$$Q(\underline{\alpha}) \approx \bar{Q}(\underline{\alpha}) = \underline{\alpha}^{\mathrm{T}} \mathbf{J}^{\mathrm{T}}(\underline{\beta}^0)\,\mathbf{W}\mathbf{J}(\underline{\beta}^0) \tag{2.77}$$

and

$$Q(\underline{\beta}) \approx \bar{Q}(\underline{\beta}) = \Delta\underline{\beta}^{\mathrm{T}} \mathbf{J}^{\mathrm{T}}(\underline{\beta}^0)\,\mathbf{W}\mathbf{J}(\underline{\beta}^0)\,\Delta\underline{\beta} \tag{2.78}$$

with the same coefficient matrix. First we confine our attention to the behaviour of $\bar{Q}(\underline{\alpha})$.

### Canonical form

To see how $\bar{Q}(\alpha)$ changes with $\alpha$ we need to examine the eigenvalues of $\mathbf{J}^{\mathrm{T}}(\underline{\beta}^0)\,\mathbf{W}\mathbf{J}(\underline{\beta}^0)$. The eigenvalue-eigenvector decomposition gives

$$\mathbf{J}^{\mathrm{T}}(\underline{\beta}^0)\,\mathbf{W}\mathbf{J}(\underline{\beta}^0) = \mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}} \tag{2.79}$$

where the $np \times np$ diagonal matrix $\Lambda$ is formed from the eigenvalues $\lambda_i$, $i = 1, 2, \ldots, np$, and the $i$th column of the $np \times np$ matrix $\mathbf{U}$ is the $i$th normalized eigenvector $\underline{u}_i$ such that $\underline{u}_i^{\mathrm{T}} \underline{u}_i = 1$. Because of the orthogonality condition $\underline{u}_i^{\mathrm{T}} \underline{u}_j = 0$ if $i \neq j$. It is supposed that the eigenvalues are ordered: $\lambda_1 - \lambda_2 \geq \ldots \geq \lambda_{np} \geq 0$.

Define a new set of parameters

$$\underline{a} = \mathbf{U}^{\mathrm{T}} \underline{\alpha} \tag{2.80}$$

which are called the principal components. Rewriting (2.77) in terms of $a$ we obtain

$$\bar{Q}(\underline{a}) = \sum_{i=1}^{np} \lambda_i a_i^2 \tag{2.81}$$

Equation (2.81) is called the canonical form of the quadratic expression $\bar{Q}(\alpha)$.

If we restrict the length of the transformed parameter vector $\alpha$ to $d$, then the largest value of the response function is $\bar{Q}(a)=\lambda_1 d^2$ and it corresponds to the special selection of principal component values: $a_1=d$, $a_2=\ldots a_{np}=0$. Changing only the first principal component yields a movement along the corresponding eigenvector in the space of transformed parameters $\alpha$. This movement is more complicated to describe in the space of the original parameters. The original parameters change according to

$$p_j(d)=p_j^0 \exp(du_{j1}), \qquad j=1, 2, \ldots, np \tag{2.82}$$

On the other hand, if we are interested in the lowest value of the response function with a transformed parameter vector length $d$, then

$$a_{np}=d, a_1=\ldots=a_{np-1}=0$$

must be selected corresponding to $\bar{Q}(a)=\lambda_{np}d^2$.

From the aspects of model reduction small eigenvalues are of great importance.

### Interpretation of eigenvectors corresponding to small eigenvalues

The question of when an eigenvalue may be regarded as "small" will be discussed later. Let $\lambda_i=0$, and assume that the corresponding eigenvector is of the form $u_i=(u_{1i}, u_{2i}, 0, \ldots 0)^T$. Then $Q(\alpha)\approx 0$ along the line $\alpha_1=c\alpha_2$ where $c=u_{1i}/u_{2i}$. According to Eq.(2.69), this line defines the curve $p_1/p_2^c=\text{const}$ in the space of the original parameters. Thus the response function $Q(\alpha)$ practically depends on the parameter combination $p_1/p_2^c$ only and does not depend on the parameters $p_1$ and $p_2$ separately.

The most important particular cases are $c=1$ and $c=-1$. At $c=1$ we have the normalized eigenvector

$$(\pm 1)u_i=(0.707, 0.707, 0, \ldots, 0)^T \tag{2.83}$$

and at $c=-1$

$$(\pm 1)u_i=(-0.707, 0.707, 0, \ldots, 0)^T \tag{2.84}$$

According to our analysis, if $\lambda_i\approx 0$ and Eq.(2.83) holds, then the response function depends only on the ratio $p_1/p_2$. (However, in some cases this dependence may also be weak, as will be discussed later.) If $\lambda_i=0$ and Eq.(2.84) holds, then the response function depends only on the product $p_1p_2$. We note that a similar mechanistic interpretation can be given to more complicated forms of the eigenvectors.

### Noninfluental parameters

Up to now we have used Eq.(2.77) to obtain mechanistic interpretation of eigenvectors corresponding to small eigenvalues. In the following we will rather utilize Eq.(2.78). Obviously this change is only a matter of interpretation since the

eigenvalues and eigenvectors are the same. A counterpart of Eq.(2.80) is

$$\Delta \underline{b} = \mathbf{U}^{\mathrm{T}} \Delta \underline{\beta}, \tag{2.85}$$

where $\Delta b$ is the vector of principal components. We are interested in the question of whether a certain parameter $p_j$ can be set to zero without essentially affecting the computed values of the responses. We will tolerate only small (say $\leq 1\%$) average changes in the computed responses.

Assume that $\Delta\beta_j$ is a dominant element in the principal component $\Delta b_i$, that is $|\Delta b_i| \approx |\Delta \beta_j|$. Setting $p_j = 0$ implies $|\Delta\beta_j| \approx 1$, therefore $|\Delta b_j| \approx 1$ and $Q(b) \approx \lambda_i$. As may be seen, $\lambda_i$ can be considered as the measure of the effect brought about by setting $p_j = 0$ in the model. Let $\lambda_i \leq \lambda_{\min} = nm \times np \times 10^{-4}$, then by virtue of

$$Q(\beta) = \sum_{j=1}^{nm} \sum_{i=1}^{np} [y_{ij}(\beta) - y_{ij}(\beta^0)]^2 w_{ij} \tag{2.86}$$

which is a counterpart of Eq.(2.70), we may expect less than 1% average change in the computed responses of the $j$th original parameter is set to zero.

Usually a parameter is represented in several eigenvectors. To get a clearer picture, we set a component $u_{ij}$ to zero if its original value is less than 0.2, since then the given parameter has less than 4% contribution to the investigated effect. If the weights defined in Eq.(2.71) are used, the value $\lambda_{\min} = 10^{-4}$ can be selected as a threshold for 'small' eigenvalues. A parameter is considered to be noninfluential if it occurs only in eigenvectors corresponding to eigenvalues $\lambda_i - \lambda_{\min}$.

The selection of $\lambda_{\min}$ is arbitrary to some extent and several other considerations can also be taken into account. The expectation value of the squared distance of estimate $\hat{\underline{\beta}}$ from the true value $\beta$ is given by

$$E[(\hat{\underline{\beta}} - \underline{\beta})^{\mathrm{T}} (\hat{\underline{\beta}} - \underline{\beta})] = \sigma^2 \sum_{i=1}^{np} \frac{1}{\lambda_i} > \frac{\sigma^2}{\lambda_{np}} \tag{2.87}$$

where $\sigma^2$ denotes the variance of the measurement errors. According to Eq.(2.87), a 10% relative error limit for the parameters means that the eigenvalues less than $\lambda_{\min} = 100\sigma^2$ must be eliminated from the model. This threshold is advantageous if the primary goal of model reduction is to obtain a well parametrized, well identifiable model. In Eq.(2.87) $\sigma^2$ can be substituted by the weighted residual variance $s_r^2$ whatever type of weighting is used.

### Strategy for model reduction

Our general strategy for reducing nonlinear regression models can be formulated briefly as follows:

(*i*) Select a set of nominal parameters;

(*ii*) Determine the eigenvalue-eigenvector decomposition of matrix $\mathbf{J}^{\mathrm{T}}(\underline{\beta})\mathbf{W}\mathbf{J}(\underline{\beta})$;

(*iii*) Eliminate noninfluential parameters from the model;

(*iv*) Eliminate unidentifiable parameter combinations by applying justified assumptions, e.g. a quasi-steady-state approximation, etc.

### 2.3.2 Nearly-singular estimation problems and model reduction in chemical kinetics

Detailed models involving several dozens of elementary reactions are frequently encountered in chemical reaction kinetics. When such models are used to evaluate measurements, nearly-singular estimation problems usually arise. The sources of near-singularity can be identified by principal component analysis. Sensitivities are computed at some set of nominal rate constants $k^0$, as is discussed in Section 2.2.4. After applying the usual transformation $\beta_j = k_j/k_j^0$ and selecting some appropriate weighting factors, the matrix $\mathbf{J}^T(\underline{\beta})\,\mathbf{W}\mathbf{J}(\underline{\beta})$ is diagonalized.

We can select a relevant set of elementary reactions by inspecting the eigenvalues and corresponding eigenvectors. In other words, we just drop the noninfluential parameters of the model. The minimum set of reactions may be a small fraction of the original set of elementary reactions. In this rather straightforward step we eliminate most of the small eigenvalues of the matrix $\mathbf{J}^T(\underline{\beta})\,\mathbf{W}\mathbf{J}(\underline{\beta})$. At a certain point, however, we cannot drop more reactions from the model, because the rate constants responsible for the remaining few small eigenvalues are also represented in the eigenvectors corresponding to relative large eigenvalues. Nevertheless, we can find dependencies among the rate constants as discussed in the previous section.

The discovered dependences in the form of rate constant products and/or ratios may confirm or refute the validity of simplifying kinetic assumptions, namely the quasi-steady-state approximation and the rate-determining-step assumption. If such an assumption is accepted and introduced into the model, we further reduce the dimension of the actual parameter space, eliminating another small eigenvalue. By repeating the analysis one can always arrive at a reduced kinetic model with identifiable parameters.

The following example, which is also studied by Vajda [51], shows how principal component analysis is used to obtain an identifiable model for methane pyrolysis. Other applications can be found in Vajda, Valkó and Turányi [53] and Vajda and Turányi [52].

**Example 2.12**

Eisenberg and Bliss [19] employed a flow reactor to pyrolyze methane at several constant temperatures. The simple reaction mechanism shown in *Fig. 2.13* was proposed to explain the results. Denoting the concentrations of the species by $x_1 = [CH_4]$, $x_2 = [C_2H_6]$, $x_3 = [H_2]$ and $x_4 = [CH_3]$ and assuming mass-action kinetics, the proposed mechanism generates the following kinetic model:

1. $CH_4 \longrightarrow CH_3^{\bullet} + H^{\bullet}$
2. $CH_3^{\bullet} + CH_4 \longrightarrow C_2H_6 + H^{\bullet}$
3. $C_2H_6 \longrightarrow 2\ CH_3^{\bullet}$
4. $C_2H_6 \longrightarrow H_2 + C_2H_4$
5. $CH_3^{\bullet} + H_2 \longrightarrow CH_4 + H^{\bullet}$

Fig. 2.13. Methane pyrolysis reactions

$$\begin{aligned}
\dot{x}_1 &= -k_1x_1 - k_2x_1x_4 + k_5x_3x_4, & x_1(0) &= x_1^0 \\
\dot{x}_2 &= k_2x_1x_2 - (k_3 + k_4)x_2, & x_2(2) &= 0 \\
\dot{x}_3 &= k_4x_2 - k_5x_3x_4, & x_3(0) &= 0 \\
\dot{x}_4 &= k_1x_1 - k_2x_1x_4 + 2k_3x_2 - k_5x_3x_4, & x_4(0) &= 0
\end{aligned} \tag{2.88}$$

Two species were observed: methane and ethane. The data are shown in *Table 2.25 (a)–(c)*. The nominal values of the rate constants selected from the literature are shown in *Table 2.26*. At temperature $T = 1409$ K Eisenberg and Bliss [19] computed the least squares estimates making use of the methane data only. Their

Table 2.25. Methane pyrolysis data

(*a*) $T = 1371$ K, $[CH_4]^0 = 1.33E-3$ kmol/m$^3$

| $t$, s | 0.18 | 0.23 | 0.44 | 0.6 | 0.7 | 1.1 |
|---|---|---|---|---|---|---|
| $[CH_4]$ | 1.31E−3 | 1.28E−3 | 1.17E−3 | 1.037E−3 | .984E−3 | .825E−3 |
| $[C_2H_6]$ | 1.33E−6 | 1.99E−6 | 2.39E−6 | 1.064E−6 | 1.06E−6 | 0.13E−6 |

(*b*) $T = 1409$ K, $[CH_4]^0 = 1.3E-3$ kmol/m$^3$

| $t$, s | 0.2 | 0.35 | 0.53 | 0.8 | 0.97 |
|---|---|---|---|---|---|
| $[CH_4]$ | 1.19E−3 | 0.91E−3 | 0.791E−3 | 0.585E−3 | 0.52E−3 |
| $[C_2H_6]$ | 0.26E−5 | 0.137E−5 | 0.80E−6 | 0.26E−6 | 0.6E−7 |

(*c*) $T = 1458$ K, $[CH_4]^0 = 1.27E-3$ kmol/m$^3$

| $t$, s | 0.03 | 0.06 | 0.076 | 0.085 | 0.113 | .138 |
|---|---|---|---|---|---|---|
| $[CH_4]$ | 1.251E−3 | 1.189E−3 | 1.160E−3 | 1.120E−3 | 1.062E−3 | .939E−3 |
| $[C_2H_6]$ | 2.286E−6 | 2.540E−6 | 2.340E−6 | 2.286E−6 | 1.524E−6 | 0.635E−6 |

estimates are also shown in *Table 2.26*. The model fitted the methane data reasonably well at several sets of parameters, as was noticed by Eisenberg and Bliss. In view of this ambiguity the estimation was repeated by Denis and Daubert [14] using both methane and ethane data for all the three temperatures. Recognizing the near-singularity of the Hessian, Denis and Daubert applied a step-by-step strategy by keeping some parameters fixed in each step of the estimation procedure. As may also be seen from *Table 2.26*, their estimates at $T = 1409$ K are rather different from those derived by Eisenberg and Bliss, and at other temperatures they obtained negative values for $k_5$. In the following discussion we use the principal component approach to reveal the reasons for these unsatisfactory results.

Table 2.26. Literature values and previous estimates of the rate constants in Example 2.12

| $T$, K | Estimate | $k_1$, $s^{-1}$ | $k_2$, $m^3\,kmol^{-1}\,s^{-1}$ | $k_3$, $s^{-1}$ | $k_4$, $s^{-1}$ | $k_5$, $m^3\,kmol^{-1}\,s^{-1}$ |
|---|---|---|---|---|---|---|
| 1371 | estimate | 0.21E−2 | 4.3E6 | 38−680 | 300 | (8E6—210E6) |
| | D−D(2) | 0.90E−2 | 0.29E6 | 212.6 | 220.7 | −1.92E3 |
| 1409 | estimate | 0.59E−2 | 4.9E6 | 90−1580 | 500 | (8.7E6—240E6) |
| | E−B(1) | 0.45E−2 | 3.6E6 | 305.0 | 301.0 | 4E6 |
| | D−D(2) | 1.438E−2 | 0.275E6 | 1238.5 | 1237.5 | 3.35E3 |
| 1458 | estimate | 2.04E−2 | 5.8E6 | 260−4800 | 800 | (9.8E6—280E6) |
| | D−D(2) | 4.213E−2 | 3.188 | 765.4 | 787.9 | −3.02E4 |

(1) Eisenberg and Bliss [19]
(2) Denis and Daubert [14]

First, a nominal set of parameters is selected. Taking into account the literature values or limits together with the previous estimates, we select the nominal parameter vector (1.44E−2, 8.0E6, 3.5E12, 3.6E2, 2.0E5). Unfortunately, we have insufficient information concerning the standard deviations of the measurement errors. The weighting factors $w_{CH_4} = 1$ and $w_{C_2H_6} = 10^4$ are selected. The ratio of the weighting factors corresponds to the difference in the orders concentrations of the two species compensated to some extent due to the probably greater relative error of the ethane measurements. This choice might seem rather ad hoc but other reasonable weighting factors do not change the results of our analysis considerably.

The principal component analysis of the $\mathbf{J}^T(\underline{\beta})\,\mathbf{W}\mathbf{J}(\underline{\beta})$ matrix at $T = 1409$ K is shown in *Table 2.27*. The weighted residual sum of squares is of the order of $10^{-8}$ and therefore the estimated value of $\sigma^2$ is about $10^{-9}$. A reasonable threshold for small eigenvalues should be about $\lambda_{min} = 10^{-7}$. As can be seen from *Table 2.27*, almost all the eigenvalues are below this threshold, therefore the estimation problem is nearly singular. The smaller eigenvalue is due to rate con-

Table 2.27. Principal component analysis of model (2.88) at $T=1409$ K

| $N$ | Eigenvalue | Eigenvector component corresponding to | | | | |
|---|---|---|---|---|---|---|
| | | $k_1$ | $k_2$ | $k_3$ | $k_4$ | $k_5$ |
| 1 | 1.03E−3 | 0.02 | 0.01 | 0.71 | −0.71 | −0.01 |
| 2 | 6.87E−8 | 0.33 | −0.13 | 0.64 | 0.64 | 0.2 |
| 3 | 1.49E−8 | 0.14 | −0.61 | −0.20 | −0.22 | 0.71 |
| 4 | 6.66E−9 | 0.92 | 0.21 | −0.21 | −0.18 | −0.11 |
| 5 | 4.52E−12 | −0.10 | 0.75 | −0.01 | −0.01 | 0.65 |

stants $k_2$ and $k_5$. According to the eigenvector corresponding to $\lambda_5$ only the ratio $K=k_5/k_2$ can be identified.

The role of the ratio $K=k_5/k_2$ can easily be interpreted by applying the quasi-steady-state assumption to the radical $[CH_3]$. Converting the last equation of model (2.88) to an algebraic equation

$$0=k_1x_1-k_2x_1x_4+2k_3x_2-k_5x_3x_4 \tag{2.89}$$

the steady-state value of $x_4$ can be expressed as

$$x_{4,s}=\frac{k_1x_1+2k_3x_2}{k_2x_1+k_5x_3} \tag{2.90}$$

Substituting (2.90) into the model (2.88) we arrive at the reduced model

$$\begin{aligned}\dot{x}_1&=-k_1x_1+(Kx_3-x_1)(k_1x_1+2k_3x_2)/(Kx_3+x_1)\\ \dot{x}_2&=-(k_3+k_4)x_2+x_1(k_1x_1+2k_3x_2)/(Kx_3+x_1)\\ \dot{x}_3&=k_1x_2+Kx_3(k_1x_1+2k_3x_2)/(Kx_3+x_1)\end{aligned} \tag{2.91}$$

Table 2.28. Estimates of parameters in model (2.91)

| $T$, K | $k_1$ | $k_3$ | $k_4$ | $K=k_5/k_2$ | Weighted residual sum of squares $Q$ |
|---|---|---|---|---|---|
| 1371 | 0.995E−2<br>±0.659E−2 | 173.2<br>± 31.59 | 180.2<br>± 29.9 | 0.149E−1<br>±0.529E−1 | 2.41E−8 |
| 1409 | 1.329E−2<br>±0.726E−2 | 356.4<br>± 64.7 | 362.1<br>± 62.9 | 0.237E−1<br>±0.118E−1 | 1.65E−8 |
| 1481 | 1.120E−1<br>±0.987E−1 | 491.7<br>±145.0 | 552.0<br>±100.4 | 0.410E−1<br>±0.232E−0 | 2.54E−8 |

and $x_{4,s}$, if necessary, can be computed from Eq. (2.90). Note that we no longer need the initial condition $x_4^0$. This is not a disadvantage, since the initial value of the non-observed radical was rather hypothetical. As may easily be seen, model (2.91) contains only four parameters: $k_1, k_2, k_3$ and $K$. The observations can be interpreted within the framework of the quasi-steady-state approximation (QSSA) model as accurately as by application of the original model (2.88). Moreover, the Hessian of the QSSA model is better conditioned; hence nonlinear parameter estimation algorithms cause less trouble and give a better fit.

The parameter estimates are less sensitive to measurement errors. Results of parameter estimation from model (2.91) are shown in *Table 2.28*. As can be seen from the *Table*, the variance of $K$ is still large at least at $T=1371$ K and $T=1481$ K. The principal component analysis of model (2.91) at these two temperature values show that parameter $K$ is noninfluential. Therefore, the simple model

$$\dot{x}_1 = -2k_1x_1 - 2k_3x_2$$
$$\dot{x}_2 = (k_3 - k_4)x_2 + k_1x_1 \qquad (2.92)$$

Table 2.29. Parameter estimates from model (2.92)

| $T$, K | Parameter | | | Weighted sum of squares $Q$ |
|---|---|---|---|---|
| | $k_1$ | $k_3$ | $k_4$ | |
| 1371 | 1.207E−2 ±0.718E−2 | 169.0 ± 31.4 | 178.4 ± 29.8 | 2.72E−8 |
| 1409 | 1.145E−2 ±0.705E−2 | 441.9 ±132.8 | 451.9 ±131.9 | 3.47E−8 |
| 1458 | 1.314E−1 ±0.133E0 | 467.7 ±167.5 | 542.1 ±108.0 | 2.73E−8 |

can be used at $T=1371$ K and $T=1481$ K. The parameter estimates are shown in *Table 2.29*. Model (2.92) has an attractive simplicity and might be very useful in certain applications. If, however, the model is used to describe nonisothermal conditions, the form (2.91) should be kept because it is superior at $T=1409$ K.

We fitted the Arrhenius expression in the linearized form Eq. (2.38) to the estimates contained in *Table 2.28*. As has been discussed previously, a nonweighted straight line fit is suitable if a constant relative error is supposed for the rate constants. The recommended model of methane pyrolysis is therefore the QSSA

Table 2.30. Arrhenius parameters for model (2.91)

| Parameter | Estimates | |
|---|---|---|
| | ln $A$<br>$A$: m³/(kmol s) | $E*10^{-5}$<br>kJ/kmol |
| $k_1$ | 36.53±15.06 | 4.72±1.77 |
| $k_3$ | 22.46± 4.68 | 1.96±0.55 |
| $k_4$ | 23.83± 3.59 | 2.12±0.42 |
| $k_5/k_2$ | 12.75± 0.14 | 1.93±0.02 |

model (2.91) with parameters shown in *Table 2.30*. The model gives a reasonable interpretation of the measurements, as is seen from *Fig. 2.14*.

It is in order to note that, from a rigorous statistical point of view, the Arrhenius parameters obtained should be refined further, fitting the model to the three sets of data simultaneously. In chemical kinetics, however, we prefer to treat isothermal data separately since this treatment gives more insight into the kinetic process, and the Arrhenius plot can be used as a further tool in model validation. Principal component analysis is an extremely useful method for coping the near-singularity when processing isothermal kinetic experiments.

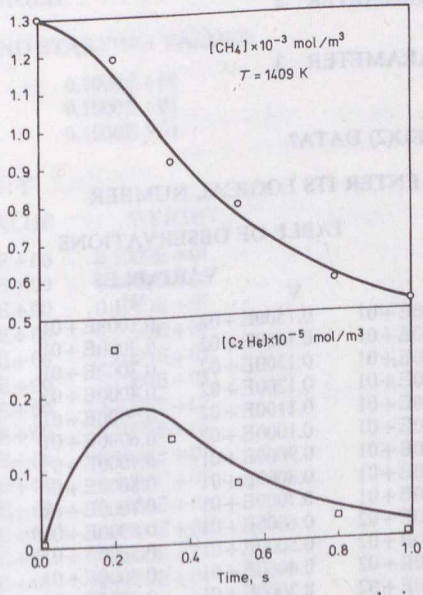



Fig. 2.14. Fit of model (2.91) to methane pyrolysis data

Fig. 2.2. Output of the REPROCHE program for Example 2.2

```
**********************************************************************************

        REPROCHE — REGRESSION PROGRAMS FOR CHEMICAL ENGINEERING

**********************************************************************************

ENTER NAME OF THE INPUT FILE
BARCOM
ENTER PROBLEM NAME(MAXIMUM 12 CHARACTERS)
BARD TEST
SELECT MODEL: ALGEBRAIC(1) OR DIFFERENTIAL(2)?
1
ENTER 4 INTEGERS TO SPECIFY NUMBER OF SAMPLE POINTS
                             NUMBER OF INDEPENDENT VARIABLES
                             NUMBER OF DEPENDENT VARIABLES
                             NUMBER OF PARAMETERS (KNOWN+UNKNOWN)
15 3 1 3

NAMES OF INDEPENDENT VARIABLES
ENTER NAME FOR VARIABLE  1
   X
ENTER NAME FOR VARIABLE  2
   V
ENTER NAME FOR VARIABLE  3
   W

NAMES OF DEPENDENT VARIABLES
ENTER NAME FOR VARIABLE  1
   Y

PARAMETER NAMES
ENTER NAME FOR PARAMETER  1

   P1
ENTER NAME FOR PARAMETER  2

   P2
ENTER NAME FOR PARAMETER  3

   P3

REAL(1) OR SIMULATED(2) DATA?
1
IF DATA INPUT FILE, ENTER ITS LOGICAL NUMBER
8
```

TABLE OF OBSERVATIONS

| SAMPLE POINT | VARIABLES | | | |
|---|---|---|---|---|
| | X | V | W | Y |
| 1 | 0.1000E+01 | 0.1500E+02 | 0.1000E+01 | 0.1400E+00 |
| 2 | 0.2000E+01 | 0.1400E+02 | 0.2000E+01 | 0.1800E+00 |
| 3 | 0.3000E+01 | 0.1300E+02 | 0.3000E+01 | 0.2200E+00 |
| 4 | 0.4000E+01 | 0.1200E+02 | 0.4000E+01 | 0.2500E+00 |
| 5 | 0.5000E+01 | 0.1100E+02 | 0.5000E+01 | 0.2900E+00 |
| 6 | 0.6000E+01 | 0.1000E+02 | 0.6000E+01 | 0.3200E+00 |
| 7 | 0.7000E+01 | 0.9000E+01 | 0.7000E+01 | 0.3500E+00 |
| 8 | 0.8000E+01 | 0.8000E+01 | 0.8000E+01 | 0.3900E+00 |
| 9 | 0.9000E+01 | 0.7000E+01 | 0.7000E+01 | 0.3700E+00 |
| 10 | 0.1000E+02 | 0.6000E+01 | 0.6000E+01 | 0.5800E+00 |
| 11 | 0.1100E+02 | 0.5000E+01 | 0.5000E+01 | 0.7300E+00 |
| 12 | 0.1200E+02 | 0.4000E+01 | 0.4000E+01 | 0.9600E+00 |
| 13 | 0.1300E+02 | 0.3000E+01 | 0.3000E+01 | 0.1340E+01 |
| 14 | 0.1400E+02 | 0.2000E+01 | 0.2000E+01 | 0.2100E+01 |
| 15 | 0.1500E+02 | 0.1000E+01 | 0.1000E+01 | 0.4390E+01 |

Fig. 2.2 (continued)

```
DO YOU WISH PARAMETER ESTIMATION(Y/N)?
  Y
ENTER NUMBER OF PARAMETERS TO BE ESTIMATED
3
ENTER 3 INTEGERS TO SPECIFY INDICES OF UNKNOWN PARAMETERS
1  2  3
ENTER 3 NUMBERS TO SPECIFY PARAMETER ESTIMATES
1  1  1
ENTER MAXIMUM NUMBER OF ITERATIONS AND RELATIVE STEP THRESHOLD
10 .001
SELECT WEIGHTS CORRESPONDING TO THE ASSUMPTION OF
                IDENTICAL VARIANCES(0)
                IDENTICAL RELATIVE VARIANCES(1)
                THE SAME USER-SPECIFIED VARIANCES IN EACH SAMPLE POINT(2)
                ARBITRARY USER-SPECIFIED VARIANCES(3)
0
                PARAMETER ESTIMATION PROBLEM: BARD TEST
                ***************************************************

NUMBER OF UNKNOWN PARAMETERS . . . . . 3
NUMBER OF SAMPLE POINTS . . . . . 15
NUMBER OF OBSERVED VARIABLES . . . . . 1
MAXIMUM NUMBER OF ITERATIONS . . . . . 10
RELATIVE STEP THRESHOLD . . . . . 1E-02

PARAMETER NAMES AND STARTING VALUES:

          P1          0.1000E+01
          P2          0.1000E+01
          P3          0.1000E+01

OBSERVED VARIABLE 1:Y

 SAMPLE POINT        VALUE          WEIGHT
        1       0.1400E+00      0.1000E+01
        2       0.1800E+00      0.1000E+01
        3       0.2200E+00      0.1000E+01
        4       0.2500E+00      0.1000E+01
        5       0.2900E+00      0.1000E+01
        6       0.3200E+00      0.1000E+01
        7       0.3500E+00      0.1000E+01
        8       0.3900E+00      0.1000E+01
        9       0.3700E+00      0.1000E+01
       10       0.5800E+00      0.1000E+01
       11       0.7300E+00      0.1000E+01
       12       0.9600E+00      0.1000E+01
       13       0.1340E+01      0.1000E+01
       14       0.2100E+01      0.1000E+01
       15       0.4390E+01      0.1000E+01
```

Fig. 2.2 (continued)

START OF ITERATION
**********

EIGENVALUE ANALYSIS
**********

| EIGENVALUE | P1 | P2 | P3 |
|---|---|---|---|
| 0.8775E+01 | 0.9092 | 0.2617 | 0.3239 |
| 0.4513E+02 | −0.4143 | 0.6477 | 0.6394 |
| 0.1419E−01 | −0.0424 | −0.7155 | 0.6973 |

ITERATION 1
MARQUARDT'S LAMBDA . . . . . 0.1E−01
RELATIVE STEP SIZE . . . . . 0.62E+00
OBJECTIVE FUNCTION AND PARAMETERS:

| | OLD VALUE | NEW VALUE |
|---|---|---|
| FUNCTION | 0.4168E+02 | 0.1325E+01 |
| P1 | 0.1000E+01 | 0.1060E+00 |
| P2 | 0.1000E+01 | 0.1426E+01 |
| P3 | 0.1000E+01 | 0.1434E+01 |

ITERATION 2
MARQUARDT'S LAMBDA . . . . . 0.1E−02
RELATIVE STEP SIZE . . . . . 0.21E+00
OBJECTIVE FUNCTION AND PARAMETERS:

| | OLD VALUE | NEW VALUE |
|---|---|---|
| FUNCTION | 0.1325E+01 | 0.3652E−01 |
| P1 | 0.1060E+00 | 0.9083E−01 |
| P2 | 0.1426E+01 | 0.1473E+01 |
| P3 | 0.1434E+01 | 0.1904E+01 |

ITERATION 3
MARQUARDT'S LAMBDA . . . . . 0.1E−03
RELATIVE STEP SIZE . . . . . 0.19E+00
OBJECTIVE FUNCTION AND PARAMETERS:

| | OLD VALUE | NEW VALUE |
|---|---|---|
| FUNCTION | 0.3652E−01 | 0.8231E−02 |
| P1 | 0.9083E−01 | 0.8346E−01 |
| P2 | 0.1473E+01 | 0.1141E+01 |
| P3 | 0.1904E+01 | 0.2335E+01 |

Fig. 2.2 (continued)

ITERATION 4
MARQUARDT'S LAMBDA . . . . . 0.1E−04
RELATIVE STEP SIZE . . . . . 0.87E−02
OBJECTIVE FUNCTION AND PARAMETERS:

| | OLD VALUE | NEW VALUE |
|---|---|---|
| FUNCTION | 0.8231E−02 | 0.8215E−02 |
| P1 | 0.8346E−01 | 0.8241E−01 |
| P2 | 0.1141E+01 | 0.1133E+01 |
| P3 | 0.2335E+01 | 0.2344E+01 |

ITERATION 5
MARQUARDT'S LAMBDA . . . . . 0.1E−05
RELATIVE STEP SIZE . . . . . 0.15E−04
OBJECTIVE FUNCTION AND PARAMETERS:

| | OLD VALUE | NEW VALUE |
|---|---|---|
| FUNCTION | 0.8215E−02 | 0.8215E−02 |
| P1 | 0.8241E−01 | 0.8241E−01 |
| P2 | 0.1133E+01 | 0.1133E+01 |
| P3 | 0.2344E+01 | 0.2344E+01 |

END OF ITERATION: STATISTICAL ANALYSIS
**********

QOODNESS OFF FIT
**********

WEIGHTED SUM OF RESIDUAL SQUARES . . . . . 0.821491E−02
NUMBER OF FREEDOM . . . . . 12
MEAN RESIDUAL SQUARE . . . . . 0.684575E−03
MEAN RESIDUAL . . . . . 0.261644E−01

OBSERVED VARIABLE: Y
SERIAL CORRELATION = −0.290300E+00

| SAMPLE | WEIGHT | OBSERVED | ESTIMATED | DEVIATION | REL.DEV. |
|---|---|---|---|---|---|
| 1 | 0.1000E+01 | 0.1400E+00 | 0.1341E+00 | 0.5879E−02 | 0.4200E−01 |
| 2 | 0.1000E+01 | 0.1800E+00 | 0.1797E+00 | 0.2651E−03 | 0.1473E−02 |
| 3 | 0.1000E+01 | 0.2200E+00 | 0.2203E+00 | −0.2742E−03 | −0.1247E−02 |
| 4 | 0.1000E+01 | 0.2500E+00 | 0.2565E+00 | −0.6541E−02 | −0.2616E−01 |
| 5 | 0.1000E+01 | 0.2900E+00 | 0.2892E+00 | 0.8236E−03 | 0.2840E−02 |
| 6 | 0.1000E+01 | 0.3200E+00 | 0.3187E+00 | 0.1300E−02 | 0.4061E−02 |
| 7 | 0.1000E+01 | 0.3500E+00 | 0.3455E+00 | 0.4463E−02 | 0.1275E−01 |
| 8 | 0.1000E+01 | 0.3900E+00 | 0.3700E+00 | 0.1996E−01 | 0.5118E−01 |
| 9 | 0.1000E+01 | 0.3700E+00 | 0.4522E+00 | −0.8222E−01 | −0.2222E+00 |
| 10 | 0.1000E+01 | 0.5800E+00 | 0.5618E+00 | 0.1821E−01 | 0.3140E−01 |
| 11 | 0.1000E+01 | 0.7300E+00 | 0.7152E+00 | 0.1481E−01 | 0.2029E−01 |
| 12 | 0.1000E+01 | 0.9600E+00 | 0.9453E+00 | 0.1471E−01 | 0.1533E−01 |
| 13 | 0.1000E+01 | 0.1340E+01 | 0.1329E+01 | 0.1122E−01 | 0.8371E−02 |
| 14 | 0.1000E+01 | 0.2100E+01 | 0.2096E+01 | 0.4221E−02 | 0.2010E−02 |
| 15 | 0.1000E+01 | 0.4390E+01 | 0.4397E+01 | −0.6767E−02 | −0.1542E−02 |

Fig. 2.2 (continued)

INTERPRETATION OF ESTIMATES
**********

CORRELATION MATRIX OF PARAMETERS

| | P1 | P2 | P3 |
|---|---|---|---|
| P1 | 1.0000 | | |
| P2 | 0.7533 | 1.0000 | |
| P3 | −0.7248 | −0.9974 | 1.0000 |

| PARAMETER | ESTIMATE | STAND.DEV. | LOWER BOUND | UPPER BOUND |
|---|---|---|---|---|
| P1 | 0.82415E−01 | 0.12375E−01 | 0.57664E−01 | 0.10717E+00 |
| P2 | 0.11331E+01 | 0.31064E+00 | 0.51182E+00 | 0.17544E+01 |
| P3 | 0.23437E+01 | 0.30002E+00 | 0.17436E+01 | 0.29437E+01 |

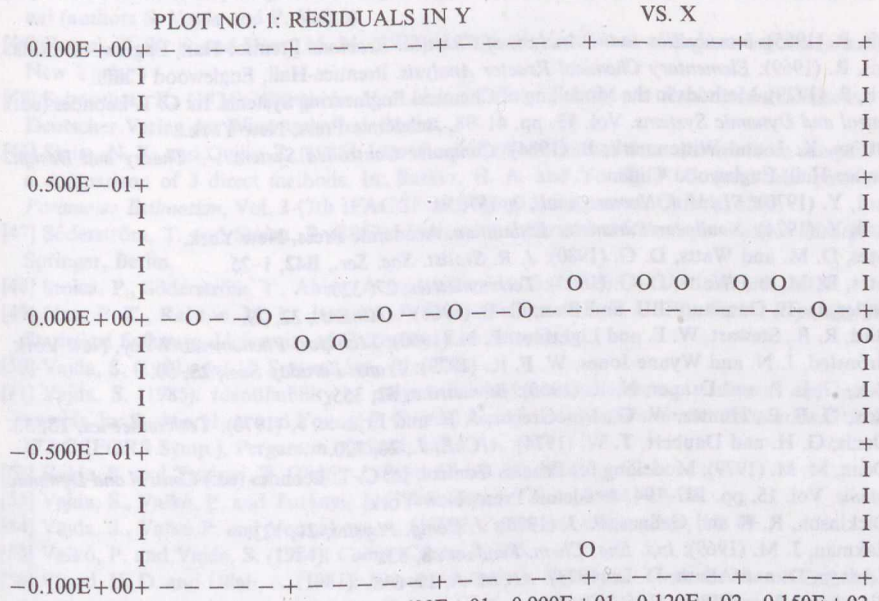
EIGENVALUE ANALYSIS
**********

| EIGENVALUE | P1 | P2 | P3 |
|---|---|---|---|
| 0.7935E−01 | 0.9239 | −0.3220 | 0.2065 |
| 0.6495E−02 | 0.3795 | 0.8393 | −0.3894 |
| 0.1443E+02 | −0.0480 | 0.4381 | 0.8976 |

SELECT PLOT 1 :NO(0), RESIDUALS(1), FIT(2), OR OFF-LINE(3)?
1
ENTER 2 INTEGERS TO SPECIFY INDEPENDENT AND DEPENDENT VARIABLES
1 1

VARIABLE RANGES ON THE PLOT
INDEPENDENT VARIABLE (X ): 0.1000E+01 − 0.1500E+02
RESIDUALS (Y ): −0.8222E−01 − 0.8222E−01
DO YOU WISH DIFFERENT RANGES(Y/N)?
Y
ENTER 4 NUMBERS TO RESPECIFY THESE RANGES
0 15 −.1 .1

Fig. 2.2 (continued)




O

**********

END OF RESULTS

**********

DO YUO WISH CHANGE PROBLEM DATA AND REPEAT ESTIMATION(Y/N)?

N

## 2.4 References


[1] Aris, R. (1965): *Introduction to the Analysis of Chemical Reactors.* Prentice-Hall, Englewood Cliffs.

[2] Aris, R. (1969): *Elementary Chemical Reactor Analysis.* Prentice-Hall, Englewood Cliffs.

[3] Aris, R. (1979): Methods in the Modelling of Chemical Engineering Systems. In: C. T. Leondes (ed.) *Control and Dynamic Systems.* Vol. **15**, pp. 41–98, Academic Press, New York.

[4] Åström, K. J. and Wittenmark, B. (1984): *Computer Controlled Systems — Theory and Design.* Prentice-Hall, Englewood Cliffs.

[5] Bard, Y. (1970): *SIAM J. Numer. Anal.*, **7**, 157.

[6] Bard, Y. (1974): *Nonlinear Parameter Estimation.* Academic Press, New York.

[7] Bates, D. M. and Watts, D. G. (1980): *J. R. Statist. Soc. Ser.*, **B42**, 1–25.

[8] Bates, D. M. and Watts, D. G. (1985): *Technometrics*, 27, 329.

[9] Biegler, L. T., Damiano, J. J. and Blan, G. E. (1968): *AIChE J.*, **32**, 29.

[10] Bird, R. B., Stewart, W. E. and Lightfoot, E. N. (1960): *Transport Phenomena.* Wiley, New York.

[11] Brönsted, J. N. and Wynne-Jones, W. F. K. (1929): *Trans. Faraday Soc.*, **25**, 59.

[12] Box, G. E. P. and Draper, N. R. (1965): *Biometrika*, **52**, 355.

[13] Box, G. E. P., Hunter, W. G., MacGregor, J. F. and Erjavec, J. (1973): *Technometrics*, **15**, 33.

[14] Denis, G. H. and Daubert, T. W. (1974): *AIChE, J.*, **20**, 720.

[15] Denn, M. M. (1979): Modelling for Process Control. In: C. T. Leondes (ed.) *Control and Dynamic Systems.* Vol. **15**, pp. 147–194. Academic Press, New York.

[16] Dickinson, R. P. and Gelinas, R. J. (1978): *J. Comp. Physics*, **21**, 123.

[17] Eakman, J. M. (1969): *Ind. Eng. Chem. Fundam.*, **8**, 53.

[18] Edelson, D. and Allara, D. L. (1973): *AIChE J.*, **19**, 638.

[19] Eisenberg, B. and Bliss, H. (1967): *Chem. Progr. Symp. Series*, **71**, 3.

[20] Garcia-Peña, J., Azen, S. P. and Bergman, R. N. (1983): *Appl. Math. Comput.*, **12**, 1.

[21] Gear, C. W. (1971): *Communication of the ACM*, **14**, 1976.

[22] Godfrey, K. R. (1983): *Compartmental Models and their Applications.* Academic Press, New York.

[23] Golub, G. (1969): Matrix Decomposition and Statistical Calculations. In: Milton, R. C. and Nelder, J. A. (eds) *Statistical Computation.* Academic Press, New York.

[24] Gottwald, B. A. and Wanner, G. (1982): *Simulation*, **37**, 169.

[25] Himmelblau, D. M. (1970): *Process Analysis by Statistical Methods.* Wiley, New York.

[26] Himmelblau, D. M., Jones, C. R. and Bischoff, K. B. (1967): *Ind. Eng. Chem. Fundamen.*, **6**, 539.

[27] Hoerl, A. E. and Kennard, R. W. (1970): *Technometrics*, **12**, 35.

[28] Hocking, R. R. (1983): *Technometrics*, **25**, 219.

[29] Hosten, L. H. (1979): *Comput. Chem. Engn.*, **3**, 117.

[30] Jennrich, R. I. (1969): *Ann. Num. Math. Stat.*, **40**, 633.

[31] Jennrich, R. I. and Sampson, P. F. (1968): *Technometrics*, **10**, 63.

[32] Kendall, M. G. (1957): *A Course in Multivariate Analysis.* Hafner, New York.

[33] Kendall, M. G. and Stuart, A. (1965): *The Advanced Theory of Statistics.* Vol. **2**: *Inference and Relationship.* Griffin, London.

[34] Loeb, J. H. and Cohen, G. M. (1965): *IEEE Trans. Automat. Control*, **10**, 359.

[35] Luyben, W. L. (1973): *Process Modelling, Simulation and Control for Chemical Engineers.* McGraw-Hill, New York.

[36] Nazareth, L. (1980): *SIAM Review*, **22**, 1.

[37] Newling, W. B. S. and Hinshelwood, C. N. (1936): *J. Chem. Soc.*, **23**, 1357.

[38] Nihtilä, M. and Virkkunen, J. (1977): *Biotech. Bioengn.*, **19**, 1831.

[39] Park, S. W. and Himmelblau, D. M. (1982): *Chem. Engn. J.*, **25**, 163.

[40] Prokopakis, G. J. and Seider, W. D. (1981): *Ind. Eng. Fundam.*, **20**, 255.

[41] Rabitz, H. (1980): *Computers and Chemistry*, **5**, 167.

[42] Ratkowsky, D. A. (1983): *Nonlinear Regression Modelling.* Marcel Dekker, New York.

[43] REPROCHE (1985): Regression Programs for Chemical Engineering. EURECHA program manual (authors S. Vajda and P. Valkó).
[44] Russel, T. W. F. and Denn, M. M. (1972): *Introduction to Chemical Engineering Analysis*. Wiley, New York.
[45] Schwetlick, K. (1974): *Kinetische Methoden zur Untersuchung von Reaktionsmechanismen*. VEB Deutscher Verlag der Wissenschaften, Berlin.
[46] Sinha, N. K. and Quijie, Z. (1985): Identification of continuous-time systems from sampled data: a comparison of 3 direct methods. In: Barker, H. A. and Young, P. C. (eds) *Identification and Parameter Estimation*, Vol. **1** (7th IFAC/IFORS Symp.), Pergamon, Oxford, 1575.
[47] Söderström, T. and Stoica, P. (1983): *Instrumental Variable Methods for System Identification*. Springer, Berlin.
[48] Stoica, P., Söderström, T., Ahlen, A. and Solbrand, G. (1985): *Int. J. Contr.*, **41**, 1421.
[49] Uno, F. K., Ralston, M. L., Jennrich, R. I. and Sampson, P. F. (1979): *Techn. Rep.*, **61**, BMDP Statistical Software, University of California, Los Angeles.
[50] Vajda, S. (1983): *Int. J. System. Sci.*, **14**, 1229.
[51] Vajda, S. (1985): Identifiability of polynomial and rational systems: structural and numerical aspects. In: Barker, H. A. and Young, P. C. (eds) *Identification and Parameter Estimation*. Vol. **1** (7th IFAC/IFORS Symp.), Pergamon, Oxford, 537.
[52] Vajda, S. and Turányi, T. (1986): *J. Phys. Chem.*, **90**, 1664.
[53] Vajda, S., Valkó, P. and Turányi, T. (1985): *Int. J. Chem. Kinet.*, **17**, 55.
[54] Vajda, S., Valkó P. and Yermakova A. (1986): *Comp. and Chem. Eng.*, **10**, 49.
[55] Valkó, P. and Vajda, S. (1984): *Comp. Chem.*, **8**, 255.
[56] Vinod, H. D. and Ullah, A. (1981): *Recent Advances in Regression Methods*. Marcel Dekker, New York.
[57] Yermakova, A. and Umbetov, A. S. (1980): *React. Kinet. Catal. Lett.*, **14**, 187.
[58] Yermakova, A., Vajda, S. and Valkó, P. (1982): *Appl. Catal.*, **2**, 139.
[59] Young, P. C. (1970): *Automatica*, **17**, 23.

# 3 ANALYSIS OF CHEMICAL ENGINEERING NETWORKS

by

F. TÁTRAI and P. SOLTÉSZ

## 3.1 Introduction

Structure analysis methods can be applied at every level of modelling and simulation of chemical engineering systems, in studies either of the steady state or of their behaviour. It is a valuable tool for the generation of models of complex interconnected systems, and also for the synthesis of chemical processes (synthesis of heat exchanger networks, distillation sequences, etc.).

A common basis for chemical engineering computations is the application of matter and energy conservation principles. If we apply these principles in finite form, the resulting balance equations form a system of algebraic equations. These systems of equations are generally very large but sparse. To solve them economically (in the sense of computer storage requirement and computing time) we have to use the information contained in their structure. Networking means the analysis of the structure of the problem by analyzing the network representing its structure. Therefore, in this chapter we will direct our attention to the structures and information flows associated with chemical engineering systems. The aim of structure analysis is to make computational problems easier to solve, i. e. to minimize the computing efforts.

## 3.2 Structure analysis of chemical engineering systems

Let us examine the meaning and application of the unit cell concept of Donati *et al.* [4]. A unit cell is a volume characterized by certain matter and energy fluxes entering and leaving it. The fluxes entering the unit cell are linked to the fluxes flowing out by certain rules, acting in accordance with the rules of stoichiometry, thermodynamics and the conservation principles. These rules, given the input fluxes, allow computation of the output fluxes. (Some authors call these sets of assignment rules 'procedures'.) They represent the balance equations and can be expressed in finite form. The unit cells are linked by connectivity equations.

The volume element may differ, depending on the aim of the study. It can be a whole plant, a part of a plant, a simple apparatus representing a unit operation, a small part of an apparatus, or an elementary volume portion of a small part of the apparatus, etc. (In this way differential balance equations appear to be particular cases of macroscopic balances.)

Therefore we can say that

— a process plant is a network of unit operations,
— a distillation column is a network of equilibrium stages,
— an absorption column is also a network of equilibrium stages,
— the so-called compartment model of chemical reactors is a network of completely mixed compartments (see Wen and Fan [5]).

There are some networks which appear physically to be networks in themselves: pipeline networks, gas distribution networks, pressure relief header networks, etc. The configuration of such networks is dictated by the real process.

Using the principles discussed above in model generation we obtain a system of equations. The equations are usually nonlinear and implicit. These systems of equations may contain thousands of unknown variables. Such an enormously large system cannot be solved by standard techniques. There is, however, a very important property of these systems which implies that significant simplifications may be made. These systems are sparse; that is, the number of variables appearing in any one equation is small compared with the total number of the variables. This sparsity is due to the structure of the unit cells and may be used to break up the system into significantly smaller subsystems (possibly into a multilevel hierarchy of subsystems) so that the solution to the whole problem may be assembled from some particular solutions to the individual subsystems. In this way the number of equations that need to be solved simultaneously may decrease drastically and the problem becomes much more tractable.

The models of chemical engineering systems are usually not completely determined. After a careful analysis of the degree of freedom we have to select the decision variables (for which an actual value will be given before starting a calculation) and design variables (which are the results of a calculation). In this section, however, we assume that all decision variables have already been specified.

There are two essentially different approaches to the solution of the large systems of equations describing chemical engineering systems. The difference is essentially in the information used for decomposition.

The most straightforward way is to consider the whole system of equations but to ignore the 'natural' structure of the unit cells. In this case the process simulator has to recognize a structure in the problem under consideration. This structure may differ from the 'natural' one. This approach provides the possibility of designing and optimizing arbitrary variables, but the size of the problem may be very large. This approach is frequently referred to as equation-oriented (EO).

However, in analyzing the structure of the system of equations we can also use the original unit cell network, which means that the equations are *a priori* grouped into subsystems called modules. Each module describes one unit cell, usually a chemical processing unit. The modules are hence built in into the process simulator program and the user cannot change them. The set of input, output and internal variables is fixed and also a solution method is given. Consequently, one cannot

select arbitrary variables as design or decision variables. The model of a unit cell may also be very large (e. g. rigorous distillation). This method has two advantages. First, we only have to analyze the unit cell structure, which is relatively small. Second, calculations involving the unit cell models may be rapid and robust, thus allowing the best method to be implemented. This approach is called sequential-modular (*SM*).

In either case, before starting the (usually iterative) calculations we have two tasks to perform:

1. Partitioning: select subsystems, as small as possible, so that they can be solved successively, i. e. without any iteration at this level.
2. Tearing: arrange the iterative calculations of the subsystems appropriately.

The scope of the remaining part of this chapter is as follows. Since the answers to the questions above require some graph theoretical results and methods, we first give the necessary basic notions of graph theory (Section 3.2.1). In Section 3.2.2 we deal with partitioning. It turns out that the partitioning problem for both approaches is essentially the same; the respective methods are then discussed. Thereafter the tearing problem will be solved for the *SM* approach (Section 3.2.1). Tearing in the EO simulation is based on purely mathematical considerations rather than on the analysis of the chemical engineering network. We do not discuss it in detail, see References [1, 2, 3, 16, 17].

### 3.2.1 Basic notions

In an analysis of the connections between the individual equations of an algebraic system of equations or between the units of a chemical engineering system, the mathematical concept of graphs seems to be a convenient common generalization. Graph theory is a widely investigated topic of mathematics. Almost all methods discussed here are applications of some graph theoretical results to our problems. Despite the fact that we do not intend to give even the broad outlines of this theory, we discuss our problems in a graph theoretical manner. First we define the necessary notions to be used later.

A network $N=(V, E)$ consists of a finite set $V$ of elementary units called points together with a set $E$ of directed links between the elementary units, i. e. 'named' ordered pairs of points called lines. The fact that line $e$ leads from $x$ to $y$ is denoted by $e$: $(x, y)$ and sometimes $e$ is said to be the input to $y$ and the output of $x$. Hence we allow 'parallel' lines between two points; for instance it is possible that $e_1$ and $e_2$ are two different lines but $e_1$: $(x, y)$ and $e_2$: $(x, y)$, i. e. their endpoints and orientation, are the same. 'Selfloops' may be also found in $E$, i. e. lines of form $e$: $(x, x)$. A network without parallel lines and selfloops is called simple. A network may be represented by a small circle for each point, with arrows directed from circle $x$ to circle $y$ for each line from $x$ to $y$. We shall denote the number of points by $n$

and the number of lines by $m$ of the network denoted by $N=(V, E)$. In the literature the points are variously called vertices, nodes, junction points or points; the lines are also referred to as arcs, links, branches or edges. Sometimes directed networks are called directed graphs or digraphs; note that these names are preferred in the mathematical literature.

A sequence of lines $P=(e_1, e_2, \ldots, e_k)$ is a directed walk from $x_0$ to $x_k$ of length $k$ (or simply a $k$-walk) if for some points $x_0, x_1, \ldots, x_k$, $e_1$: $(x_0, x_1)$, $e_2$: $(x_1, x_2)$, $\ldots$, $e_k$: $(x_{k-1}, x_k)$. If points $x_0, \ldots, x_k$ are distinct, then $P$ is a directed *path* of length $k$ (a $k$-path). The empty set 0 is sometimes considered as a 0-path from any point $x$ to the same point. If $x_0=x_k$ (but $x_1, \ldots, x_k$ are henceforward distinct), then $P$ is a (simple) *cycle*. We say that network $N$ is strongly connected if for any two distinct points $x$ and $y$ there exists a path $P$ from $x$ to $y$. It follows from the above definition (applying it for points $y$ and $x$) that a path $Q$ from $y$ to $x$ also exists. A network containing no cycle is *acyclic*.

By deleting some points and all lines starting from or leading to these points in network $N$ we obtain a subnetwork of $N$. We shall refer to subnetworks with their point set; e. g. if $X$ is a subset of $V$ then in subnetwork $X$ the lines are all lines of the original network with both endpoints belonging to $X$.

The number of input lines of $x$ is the indegree of $x$, denoted by $\text{in}(x)$. Similarly, the outdegree $\text{out}(x)$ is the number of output lines of $x$.

*Figure 3.1* illustrates the notions relevant to directed networks. *Figures 3.1, a* and *b* show parallel lines and a selfloop; the network in *Fig. 3.1, c* is strongly connected; lines $e_1, e_2, e_3, e_5, e_6$ are a walk of length 5; lines $e_3, e_5$ form a path of length 2; lines $e_3, e_5, e_6, e_4$ form a cycle. The indegree $\text{in}(x_1)=3$ and outdegree $\text{out}(x_3)=1$. In *Fig. 3.1, d* we see an acyclic network.

Similarly, an undirected network $N=(V, E)$ consists of a finite set $V$ and a set $E$ of unordered pairs of points; again the elements of $V$ and $E$ are referred to as points and lines, respectively. The notation $e$: $(x, y)$ means that line $e$ connects point $x$ to point $y$, or, equivalently, $y$ to $x$. Subnetworks, *walks*, *paths* and cycles can be

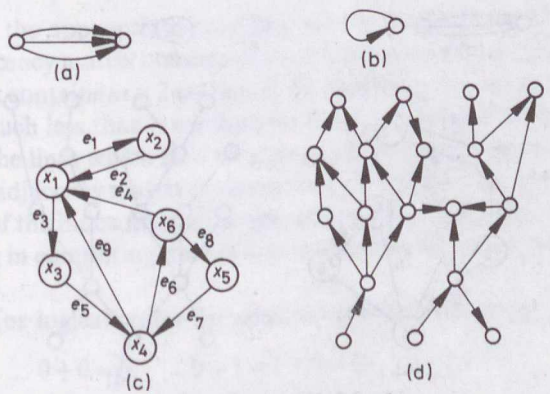



Fig. 3.1. Directed networks

defined on the analogy of the directed networks. The number of lines having point $x$ as their endpoint is the degree of $x$, denoted by $d(x)$. Parallel lines and selfloops are allowed; networks without parallel lines as well as selfloops are called simple networks. A network containing at least one path between any two distinct points is *connected*. An acyclic connected network is a *tree*.

In *Fig. 3.2* we illustrate the notions concerning undirected networks with the networks of *Fig. 3.1* but with undirected lines. *Figure 3.2, a* and *b* show parallel lines and a selfloop; the network in *Fig. 3.2, c* is connected; lines $e_1$, $e_2$, $e_3$, $e_6$ are a walk of length 5; lines $e_3$, $e_5$ form a path of length 2; lines $e_3$, $e_5$, $e_6$, $e_4$ form a cycle. The degree $d(x_4)=4$. In *Fig. 3.2, d* we see a tree; this is an acyclic and connected undirected network.

From the computational point of view it is necessary to encode the networks. There are several methods for doing this. In each case we have to enumerate all points $(1, \ldots, n)$ and all lines $(1, \ldots, m)$.

The adjacency matrix $\mathbf{A}$ has $n$ rows and $n$ columns; the entry $\mathbf{A}(i, j)$ is 1 if and only if there is a line from point $i$ to point $j$ and 0 otherwise. Consequently, the encoded network $N$ must be simple. We can also use the adjacency matrix for encoding undirected networks; in this case we define it as a symmetric matrix; i. e. $\mathbf{A}(i, j)=\mathbf{A}(j, i)$ for each $i, j$. For example, the adjacency matrix of the network in *Fig. 3.1, c* is the following (using the sequence of points shown there):

$$\mathbf{A}=\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

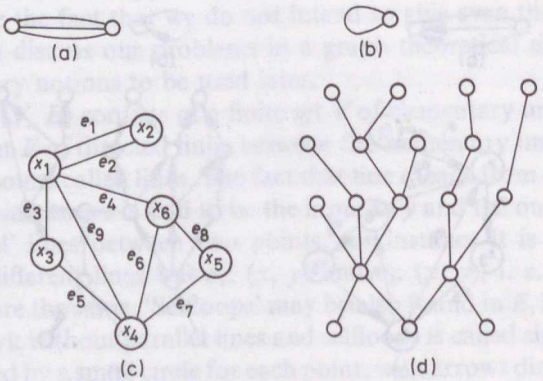



Fig. 3.2. Undirected networks

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

The symmetric adjacency matrix of the undirected network in *Fig. 3.2, c* is the following:

$$\mathbf{A}=\begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

The index matrix **I** of the directed network $N$ has $m$ rows corresponding to the lines and two columns. We write in positions $\mathbf{I}(i, 1)$ and $\mathbf{I}(i, 2)$ the serial numbers of the starting and end points of line $i$. This matrix is also able to encode networks with parallel lines. The index matrix of the directed network in *Fig. 3.1, c* is shown below, using the enumeration of points and lines shown there:

$$\mathbf{I}=\begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 1 & 3 \\ 6 & 1 \\ 3 & 4 \\ 4 & 6 \\ 5 & 6 \\ 6 & 5 \\ 4 & 1 \end{bmatrix}$$

The choice of the appropriate encoding system depends on the problem to be solved. The adjacency matrix consists of $n \times n$ numbers: $m$ ones and $n \times n - m$ zeros. The index matrix contains $m \times 2$ entries. If we have only limited computer memory, and if $m \times 2$ is much less than $n \times n$ then we choose the index matrix. On the other hand, selecting the lines which lead to a given point $x$ requires looking for ones in column $x$ of the adjacency matrix ($n$ operations) or looking for the number $x$ in the second column of the index matrix ($m$ operations). If $n$ is significantly smaller than $m$ and if a saving in computing time is important then we may prefer the adjacency matrix.

The Boolean (or logical) rules for addition are the following:

$$0+0=0; \qquad 0+1=1+0=1; \qquad 1+1=1$$

The identity matrix of size $n$ is denoted by **I** [i.e. $\mathbf{I}(i, j)=1$ if $i=j$; 0 otherwise].

### 3.2.2 Partitioning and related questions

In either *SM* or *EO* approaches we investigate the structure of connections between the 'natural' units of our system. More precisely, we analyze the precedence structure of the problem. If a unit yields some necessary information to another unit (whatever a 'unit' may be in this context), then obviously the first must precede the second. These connections have their own inherent directions in the *SM* case because they correspond to some material or information flows. On the other hand, the information is represented by the variables in the case of systems of equations, therefore a common variable implies a connection between two equations. As the input and output variables of an equation or a subsystem are not *a priori* determined by the problem, this connection has no 'natural' direction.

Despite of this essential difference between the sequential-modular and equation-oriented approach, the partitioning problem may be reduced to the same graph theoretical question: how can we identify the strongly connected components of the corresponding directed network? In the *EO* approach this directed network is less 'natural' than in the *SM* case; henceforth we shall need some special methods to construct it.

#### Structure network of unit models

First let us consider the *SM* case. The units in question are models of some chemical processing units in the form of procedures. A procedure always calculates its output variables provided that all its input variables are given. If an input variable of unit $A$ is an output from unit $B$, then we obviously have to calculate $B$ before $A$. Let us consider the units as points of a network and represent these relations by lines; i.e. if at least one input variable of unit $A$ is an output of unit $B$, then orient a line from $B$ to $A$. We may assume that this network is simple, because parallel lines and selfloops would have no meaning. A line shows that the second unit requires directly some outputs of the first one. A directed path shows an indirect precedence relation between the units corresponding to its end points. A strongly connected part of this network corresponds to models which use some data from the others, directly or indirectly. This means that these units must somehow be handled together. We naturally want to identify these unit groups so that they should be as small as possible.

#### Structure network of systems of equations

In a system of equations the 'natural' units are the individual equations and variables. Suppose that the number of equations ($n$) is equal to the number of variables. Consider the undirected network which has $n$ points representing the equations, another $n$ points representing the variables and in which there is a line from an equation point to a variable point if and only if the variable appears in the

equation. Thus, each line connects an equation point and a variable point; we shall refer to them as the equation and variable of the line.

Now, select an 'output assignment', i.e. assign an equation to each variable so that different variables get different equations. Each variable will be calculated from the equation assigned to it. Naturally, the variable must appear in 'its' equation. Such an assignment must exist if the system is not structurally singular (Steward [6]). In the undirected network the corresponding lines form a *matching*, that is, a set of $n$ lines with no end points in common.

Orient the matching lines towards the variable part. This shows that the variables are determined from the equations corresponding to the starting points of these lines. Direct all the remaining lines towards the equation part, which means that the variables at the starting point of the lines have to have a value before one can commence solving the equation at the end point. Both networks are called bipartite since their point set has two parts: the equation part and the variable part, and lines lead between the two parts only. To illustrate this, consider the following system of equations:

$$\begin{aligned} f_1(x_1, x_2) &= 0 \\ f_2(x_1, x_2) &= 0 \\ f_3(x_1, x_3, x_4, x_5) &= 0 \\ f_4(x_2, x_4, x_5) &= 0 \\ f_5(x_4, x_5) &= 0 \end{aligned} \tag{3.1}$$

This is a small sparse system. Let us assign variable $x_i$ to the $i$th equation $f_i$. The resulting undirected and directed networks are shown in *Fig. 3.3*; the matching lines are shown as thick lines.

In this network each path alternately uses lines belonging to the matching part and those that do not belong to it. A path from equation $B$ to equation $A$ shows that the computation of $B$ must precede the computation of $A$. Consequently, each equation of a strongly connected subnetwork should precede all the others. Of course, this is not possible, and hence these equations must somehow be solved simultaneously. We want to identify these subsystems, and do it naturally so that they do not contain any unnecessary unit.

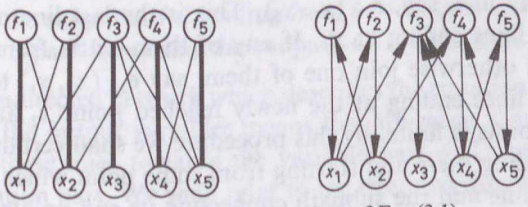



Fig. 3.3. Networks of the system of Eq. (3.1)

### Strong connectivity

In both *SM* and *EO* approaches we want to identify subsystems in such a way that they must be handled simultaneously but that they should be as small as possible. In graph theoretical language, each desired subnetwork $S$ must have the following properties: (1) $S$ is *strongly connected*, i.e. for any two points $x$ and $y$ of $S$, there exists a path from $x$ to $y$. (2) $S$ is maximum with respect to the strong connectivity, i.e. each subnetwork cannot be enlarged without violating (1).

These subnetworks are called the (strong) *components* of the original network. We briefly summarize some simple but useful properties of the components.

**Theorem 3.1**

The components of network $N=(V, E)$ are unique.

**Proof**: Let us call points $x$ and $y$ equivalent if $x=y$ or there exist paths $P$ and $Q$, with $P$ leading from $x$ to $y$ and $Q$ leading form $y$ to $x$. It is easy to check that this relation is an equivalence relation and its equivalence classes are just the components.

Now, contract the components, i.e. consider them as points of a new network, and orient a line from component $A$ to component $B$ if and only if there exists a line in $N$ from a point of $A$ to a point of $B$. We shall refer to this network as the component network of $N$. Note that it is simple by definition.

**Theorem 3.2**

The component network is acyclic.

**Proof**: If the component network contained a cycle, then the subnetwork defined by the union of the components lying on the cycle would be strongly connected. But this would contradict the maximality of the components on the cycle.

**Lemma 3.3**

If $N$ is acyclic then there exists a point $x$ for which $\text{in}(x)=0$.

**Proof**: We prove that if $\text{in}(x)>0$ holds for each point $x$, then the network contains a cycle. We construct a path backward: select any point $x_0$ and consider the lines ending at $x_0$. At least one such line surely exists as $\text{in}(x_0)>0$. If this line starts in $x_0$ then this line alone forms a cycle and we are done. If not, select any line, say, $e_1$: $(x_1, x_0)$. This is the last line of the path. Now consider the lines ending at $x_1$. If any of them starts from either $x_0$ or $x_1$, we are done; otherwise join one of them, say $e_2$: $(x_2, x_1)$ to the path. Now consider the lines ending at the newly reached point $x_2$ and so on. As the number of points is finite, by this procedure we shall certainly reach a point and a line ending there and starting from some point which we have reached earlier. This line and the subpath connecting its end points form a cycle.

**Theorem 3.4** (Precedence ordering)

Let $N$ be an acyclic network. Then there exists a sequence of its points $x_1$, $x_2, \ldots, x_n$ so that each line leads forward with respect to the sequence.

**Proof**: We determine the appropriate sequence. First select a point $x_1$ with the property that no line ends in $x_1$, $\text{in}(x_1)=0$. Such a point surely exists, applying Lemma 3.3. Delete $x_1$ from $N$. The remaining network is acyclic again. Select $x_2$ as a point with $\text{in}(x_2)=0$ in the remaining network, delete $x_2$, and so on. The sequence determined by this procedure fulfils the desired property.

In view of Theorem 3.2 and Theorem 3.4 we may conclude that both in the *SM* and *EO* approaches the component network gives just the desired partition of $N$ into subnetworks. In fact, applying Theorem 3.4 to the component network (which is acyclic by Theorem 3.2), we get a sequence of the components with no backward lines. This means that the corresponding subsystems can be computed in turn, according to this sequence.

### Output assignment and components

Let us investigate the case of systems of equations. We have started building the appropriate network by selecting an output assignment. The strong components of the corresponding network might depend on our choice. Our next aim is to prove that the partition is independent of the output assignment. This was first proved by Steward [6]. We need a simple lemma.

**Lemma 3.5**

In directed networks reversing the orientations of the lines along a cycle $C$ does not affect the reachability relations of the points.

**Proof**: We have to prove that for each pair of distinct points $x$ and $y$, if there exists a path $P$ from $x$ to $y$ then there also exists a path $Q$ from $x$ to $y$ in the new network. If path $P$ does not meet $C$, then $Q=P$ is suitable. Otherwise, travel from $x$ to $y$ along path $P$ in the original network; let $a$ and $b$ be the first and last points reached on $C$. Points $a$ and $b$ determine two subpaths on $C$ (one from $a$ to $b$ and the other from $b$ to $a$) both in the original and the new networks. Then an appropriate $Q$ can be assembled from the subpath of $P$ from $x$ to $a$, the subpath of $C$ leading from $a$ to $b$ in the new network (if $a=b$ then use the 0-path here) and the subpath of $P$ from $b$ to $y$.

Consider the undirected network which describes the equation-variable relationships. Assigning the output variables means that we select a matching $M$ in this network, orient these lines towards the variable part and orient all other lines towards the equation part. We shall refer to the resulting directed network as the network corresponding to matching $M$.

The next theorem expresses the very important fact that the partition of systems of equations using the methods discussed here does not depend on the particular choice of the output assignment.

**Theorem 3.6**

Let $M_1$ and $M_2$ be two matchings in an undirected bipartite network and let $N_1$ and $N_2$ be the corresponding directed networks. Then the point sets of components of $N_1$ and $N_2$ are the same.

**Proof**: Consider the symmetric difference $S$ of $M_1$ and $M_2$, i.e. the lines belonging to either $M_1$ or $M_2$, but not both. The line set $S$ can be partitioned into disjoint simple cycles. (To demonstrate this claim, note that for any point $x$ either no line of $S$ joins to $x$ or one input and one output line of $x$ belongs to $S$.) The corresponding cycles are oriented just the opposite way in $N_1$ and $N_2$. It follows that starting with $N_1$ and reversing the orientation along these cycles we obtain just $N_2$. By Lemma 3.5 $N_1$ and $N_2$ have the same components.

For example, select a new output assignment for the system of equations, Eq.(3.1):

$$\begin{aligned} f_1 &\to x_2 \\ f_2 &\to x_1 \\ f_3 &\to x_3 \\ f_4 &\to x_5 \\ f_5 &\to x_4 \end{aligned} \tag{3.2}$$

The corresponding undirected and directed networks are shown in *Fig. 3.4*. The symmetric difference of the original matching (shown in *Fig. 3.3*) and the new one consists of two cycles with points $f_1$, $x_1$, $f_2$, $x_2$ and $f_4$, $x_4$, $f_5$, $x_5$ *(Fig. 3.5)*. It is easy to check that reversing both cycles in the directed network in *Fig. 3.3* we get just the directed network in *Fig. 3.4*.

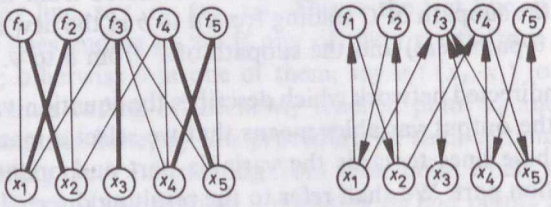



Fig. 3.4. Networks corresponding to Eq. (3.2)

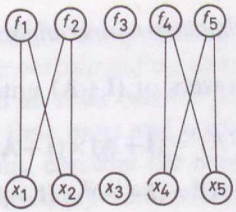



Fig. 3.5. The symmetric difference of two matchings consists of disjoint cycles

### Determining the components
### I. Reachability matrix

Now, the question arises: how can we determine the (strong) components of a given directed network $N$? We first present an algorithm based on the adjacency matrix $\mathbf{A}$. Let $\mathbf{R}$ be the reachability matrix of $N$, i.e. $\mathbf{R}(i,j)=1$ if there exists a path from point $i$ to point $j$; 0 otherwise. Then:

**Theorem 3.7** (Harary [7])
Using Boolean addition:

$$\mathbf{R}=(\mathbf{I}+\mathbf{A})^{n-1}=(\mathbf{I}+\mathbf{A})^{n}=(\mathbf{I}+\mathbf{A})^{n+1}=\ldots$$

**Proof**: First we prove that matrix $(\mathbf{I}+\mathbf{A})^k$ describes the $k$-reachability relations; that is $(\mathbf{I}+\mathbf{A})^k(x,y)=1$ if and only if there exists a path from $x$ to $y$ of a length at most $k$. The proof is inductive on $k$. For $k=1$ this is obvious. If $k>1$ then consider the expansion

$$(\mathbf{I}+\mathbf{A})^k(x,y)=\sum_{z=1}^{n}(\mathbf{I}+\mathbf{A})^{k-1}(x,z)\times(\mathbf{I}+\mathbf{A})(z,y)$$

That is to say, as a result of the Boolean addition, the entry $(\mathbf{I}+\mathbf{A})^k(x,y)=1$ if and only if there exists a point $z$ for which $(\mathbf{I}+\mathbf{A})^{k-1}(x,z)=1$ and $\mathbf{A}(z,y)=1$.

If such a point $z$ exists, then by the induction hypothesis there exist two paths: $P$ from point $x$ to point $z$ and $Q$ from $z$ to $y$, where $P$ contains at most $k-1$ lines and $Q$ contains at most one line. Then $P$ lengthened by $Q$ forms a path from $x$ to $y$ with a length of at most $k$.

Conversely, if there is a path $P$ from $x$ to $y$ with a length of at most $k$, then either $x=y$ or there is a point $z$ on $P$ just before $y$. Since the length of the subpath of $P$ between $x$ and $z$ is at most $k-1$, it follows by the induction hypothesis that $(\mathbf{I}+\mathbf{A})^{k-1}(x,z)=1$ and $\mathbf{A}(z,y)=1$.

Now, to finish the proof, observe that since each path has at most $n-1$ lines, the reachability relations are totally described by $(\mathbf{I}+\mathbf{A})^{k-1}$, and also by the higher powers of $(\mathbf{I}+\mathbf{A})$.

Theorem 3.7 yields a simple algorithm to compute the reachability matrix.

*Reachability matrix algorithm*

Calculate successively the powers of $(\mathbf{I}+\mathbf{A})$ until a power of at least $n-1$ is reached:

$$(\mathbf{I}+\mathbf{A})^2=(\mathbf{I}+\mathbf{A})\times(\mathbf{I}+\mathbf{A})$$

$$(\mathbf{I}+\mathbf{A})^4=(\mathbf{I}+\mathbf{A})^2\times(\mathbf{I}+\mathbf{A})^2$$

$$(\mathbf{I}+\mathbf{A})^8=(\mathbf{I}+\mathbf{A})^4\times(\mathbf{I}+\mathbf{A})^4$$

$$\text{,,}$$

$$\text{,,}$$

$$\text{,,}$$

Then the last computed matrix is $\mathbf{R}$. The running time of this algorithm is $o(n^3\times\log_2(n))$ since it consists of about $\log_2(n)$ multiplications of two matrices of order $n$ and the standard matrix multiplication algorithm takes $o(n^3)$ operations. There is a more sophisticated algorithm carrying out multiplication of Boolean matrices in time $o(n^3/\log_2(n))$ (Arlazanov *et al.* [8]). If we use this algorithm, the number of operations is reduced to $o(n^3)$. The number of memory addresses required is at most $n^2$ but this may also be reduced if we take advantage of the fact that each number stored is either 0 or 1. Furthermore, the use of bit-vector operations also accelerate these calculations.

*Algorithm for determining the components from* $\mathbf{R}$

We are now in a position to determine the components because $\mathbf{R}$ contains all the information necessary for this purpose. Take the first row and column of $\mathbf{R}$. $\mathbf{R}(1,x)=\mathbf{R}(x,1)=1$ if and only if point 1 and point $x$ belong to the same component. In fact, $\mathbf{R}(1,x)=1$ $(\mathbf{R}(x,1)=1)$ if and only if there exists a walk from 1 to $x$ (from $x$ to 1). Thus, the points of the component $S$ which contain point 1 are those points $x$ for which $\mathbf{R}(1,x)=\mathbf{R}(x,1)=1$.

Mark the rows corresponding to component $S$. Find the first unmarked row,

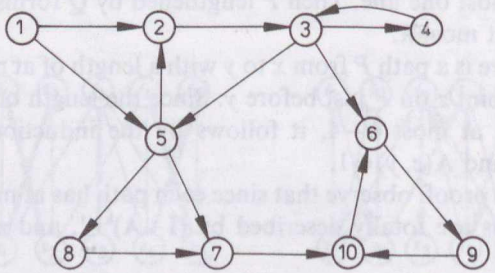



Fig. 3.6. Example for determining the component network

say, row $i$. Then the points of the component $S$ containing $i$ are points $x$ with $\mathbf{R}(i, x) = \mathbf{R}(x, i) = 1$. Mark the corresponding rows again.

Repeat this procedure until all rows become marked.

This algorithm takes $o(n^2)$ time units and $n$ space units for marking the rows.

To illustrate both algorithms, consider the network in *Fig. 3.6*.

The adjacency and the reachability matrices are as follows:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

To help identifying the components, change all entries $\mathbf{R}(i, j)$ to zero for which the symmetrically opposite entry $\mathbf{R}(j, i)$ is zero. The result matrix $\mathbf{W}$ and the adjacency matrix $\mathbf{A}'$ of the component network are as follows:

$$\mathbf{W} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix} \quad \mathbf{A}' = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Using matrix $\mathbf{W}$ one easily finds that the components are the following:

Component 1: 1

Component 2: 2, 3, 4, 5

Component 3: 8

Component 4: 7

Component 5: 6, 9, 10

Many authors have improved the speed of computing the reachability matrix; see, for instance, Mah [9]. This approach, however, has a major disadvantage, namely, that **R** may be a full matrix even if the original adjacency matrix is very sparse. For example, if the network is an $n$-cycle, the adjacency matrix contains $n$ nonzero entries but **R** consists of $n \times n$ nonzeros, i.e. **R** is totally full. For very large but sparse problems this method is not recommended.

### Determining the components
### II. Depth-first search

Now we discuss the more sophisticated algorithm discovered by Tarján [10]. As far as we know, this is the fastest algorithm for determining the strong components. Its running time and space requirement depend linearly on $n$ and $m$ (where $m$ is the number of lines), i.e. they are at most $a_1 \times n + b_1 \times m + c_1$ and $a_2 \times n + b_2 \times m + c_2$ with some implementation-dependent constants $a_1$, $b_1$, $c_1$, $a_2$, $b_2$, $c_2$. It is worth implementing this method if the expected size of problems is very large.

The base of Tarján's algorithm is a relatively simple and widely used enumeration algorithm which assigns to each point $x$ of the network a label $l(x)$ which is a nonnegative integer. First we describe this algorithm and then later we shall discuss how it works.

*Labelling algorithm*

Let the network in question be denoted by $N=(V, E)$. First, for the sake of simplicity, we add a new point $s$ to $V$ and $n$ new lines from $s$ to each point of $V$. This clearly does not affect the strong components, since no new reachability relation occurs, but we know that all old points are reachable from $s$. Then we systematically travel in the network, visiting all its points. During the procedure the points may be in one of two states:

$p1$. unlabelled
$p2$. labelled

Similarly, the lines have also two possible states:

$L1$. unscanned
$L2$. scanned

We shall always bear in mind the path leading from the starting point to the actual point $AP$.

START: Initially all points and lines are unlabelled and unscanned, respectively. Then $s$ receives label 0 $(l(s)=0)$ and from now on it is labelled. The path contains the single point $s$. The actual point $AP=s$; in other words, we are staying in point $s$. Continue with the scanning step.

SCANNING STEP: look for an unscanned line leading from the actual point $AP$ to an unlabelled point. Scan the lines leading out of $AP$ one after the other. For each line there are three cases.

CASE 1. The line is already scanned. Nothing to do.
CASE 2. The line is unscanned but its end point is labelled. This line becomes scanned.
CASE 3. The line is unscanned and its end point is unlabelled. Then:

STEP FORWARD: mark the line scanned, add its end point $y$ to the path, give the next label to $y$: $l(y):=l(x)+1$ and let the new actual point $AP=y$. Go on with the scanning step. In other words, the travel is continued by visiting a new point $y$.

If all lines leading out from the actual point are scanned (in other words, no unvisited point is reachable from $AP$), then:

STEP BACKWARD: along the path. Let the point just before $AP$ on the path be $x$. Then delete $AP$ from the path and let the new actual point be $x$. Go on with the scanning step.

If the backward step fails, i.e. there is no point before the actual point on the path, then TERMINATE.

This process uses the so called depth-first search. In fact, the travel is organized so that as soon as we find a line leading to an unvisited point, we visit it along that line. We turn back only if no other possibility exists. If the network represented a labyrinth with one entry at $s$ and we had to find something hidden in it, then this strategy would ensure success, as well as finding the way back. We examine first the depth of the labyrinth along one branch and consider the other branches later.

Now analyze the algorithm.

1. In implementing the algorithm, suppose that for each point we have a list containing the lines leading out. Then it is convenient to allocate a pointer to each point, which shows how many lines of the list have already been scanned. Then, after a backward step we can just go on scanning with the first unscanned line. This avoids unnecessary repetitions during the scanning procedure and each line may be considered only once. This ensures that the total number of scanning steps is $m+n$.

2. Each point becomes labelled, since all lines leading out of $s$ are also scanned. This is the point which was simplified by adding $s$ to the original network. In implementations this is not necessary.

3. We do not step back from a point $x$ until all points reachable from $x$ are labelled.

**Proof**: Consider a point $y$ and a path $P$ from $x$ to $y$. Let us investigate the moment when we step back from $x$. Then it is not possible that there exists a line $e$, say $e:(u, v)$ on $P$ for which $u$ is labelled but $v$ not. As $x$ is labelled, the next point of $P$ must also be labelled, and so on, until $y$.

In order to identify the strong components, it is a relatively straightforward idea to contract all points lying on a cycle into one 'pseudo-point' or 'group'. During

the scanning step we can find some cycles when a scanned line leads to a point of the actual path. If we contract the cycles during the labelling procedure, the final groups are just the components (Sargent and Westerberg [11]). Maintaining the groups, however, requires too much extra effort. Complicated list-processing methods may be avoided by Tarján's idea.

4. Consider a component of $N$. Let us call its root the point in which the component is first reached during our travel (i.e. the point with the smallest label). By note 3, all points of the component have been reached before stepping back from the root. We shall complete the labelling process so that the roots become recognizable when stepping back from them. For this purpose we assign a value $r(x)$ to each point $x$. This value will refer to the earliest point on the path which is known to belong to the same component as $x$. We add some simple operations to the steps to update these values.

START: $r(s)=0$
SCANNING STEP:
CASE 1: –
CASE 2: If the line is unscanned but leads to a labelled point $x$, check whether $x$ lies on the path or not. If not, go on. If it does, then the subpath from $x$ to $AP$ together with this line forms a cycle (see *Fig. 3.7*). Consequently, $x$ and $AP$ belong to the same component. Let $r(AP):=l(x)$ if this is new information i.e. if $l(x)$ is less than the old value of $r(AP)$.
CASE 3:
FORWARD STEP: let $r(AP)=l(AP)$ showing that we do not know anything about the newly reached point.

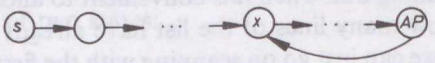



Fig. 3.7. Line leading from $AP$ to a point $x$ of the path closes a cycle

BACKWARD STEP: Denote by $x$ the point before the actual point $AP$ and by $y$ the point of the path with $l(y)=r(AP)$. Obviously, if $r(AP)<l(x)$, then $x$ lies on the actual path between $y$ and $AP$, therefore $y$, $x$ and $AP$ belong to the same component. Therefore let $r(x)=r(AP)\ (=l(y))$.

We now prove the property used for identifying the components.

5. After completing the algorithm, $r(x)=l(x)$ if and only if $x$ is the root of a component.

**Proof**: If $x$ is a root, then $r(x)$ received the value $l(x)$ when $x$ was reached. This does not change later since there is no other point in the same component with a label less than $l(x)$.

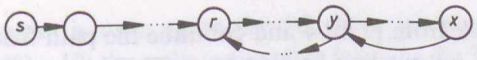



Fig. 3.8. Why $r(x)<l(x)$ if $x$ is not a root — case 1

On the other hand, if $x$ is not a root, then $r(x)<l(x)$. Denote the actual path at the moment of stepping back from $x$ by $P$ and the root of the component of $x$ by $r$. Then $r$ is reachable from $x$, say, on a path $Q$. We have already proved that the whole component of $r$ has been reached before stepping back from $r$; consequently, when stepping back from $x$, the root is still on the actual path $P$. Consider the first line of path $Q$ and denote its end point by $y$. If $y$ is on the actual path $P$ then $r(x)<=r(y)<=l(y)<l(x)$ *(Fig. 3.8)*. If not, then $Q$ must contain a line $e:(a, b)$ where $l(a)>l(x)$ and $b$ is on path $P$. Then $r(a)<<l(b)<l(x)$ and, during the backward steps from $a$ to $x$ this information is retained, i.e. $r(x)<=r(a)<l(x)$ again *(Fig. 3.9)*.

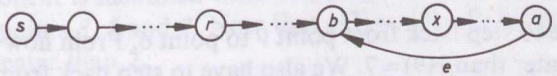
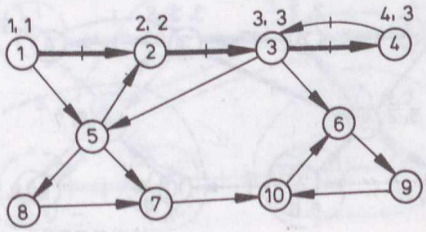



Fig. 3.9. Why $r(x)<l(x)$ if $x$ is not a root — case 2

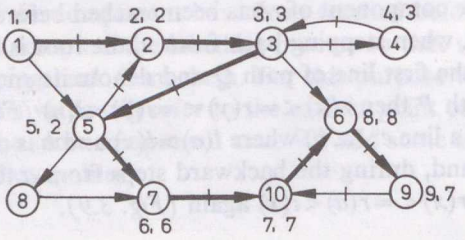
The roots of the components can be recognized on the backward steps. When leaving a point $x$ backwards there are two cases. If $r(x)<l(x)$ then $x$ is not a root, by *5*. If $r(x)$ is still equal to $l(x)$ then $x$ is a root. When we first find a root $r_1$, then mark all labelled points $x$ with $l(x)>=l(r_1)$ as belonging to the same component, by *3* and *5*. When we find the next root $r_2$, its component consists of the unmarked and labelled points $x$ with $l(x)>=l(r_2)$; and so on.

As an illustration we follow this procedure on the network shown in *Fig. 3.6*. To each point we give two numbers, the label and the $r$ value. Scanned lines are marked by a small dash across then. The thickened lines lie on the actual path. The additional point $s$ is omitted for the sake of simplicity.

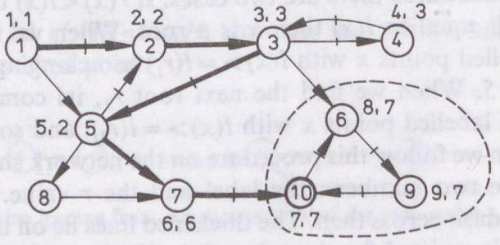
First we reach point *4* by successive forward steps. The labels and $r$ values are identical at the points passed. Scanning the only output line of point *4* we see that it leads to point *3* which lies on the actual path, therefore $r(4)=l(3)=3$:



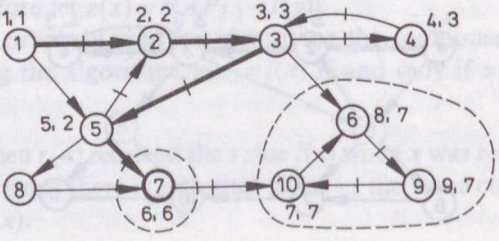
Now we step back from point *4* and continue the path from point *3* to point *5*. Examining its first output line leading to point *2* we see that point *2* has been labelled previously, and its $r$ value is 2, i.e. it is less than that of point *5*: points *5* and *2* must belong to the same component, therefore $r(5)=l(2)=2$. We continue our path through points *7*, *10*, *6* and *9*. Scanning the only output line of point *9*, leading to the labelled point *10*, we get $r(9)=l(10)=7$:



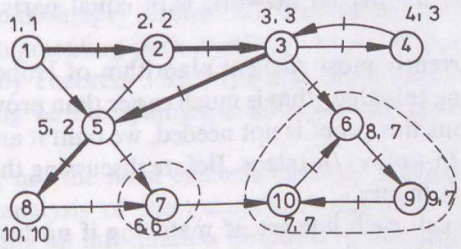

Then we have to step back from point *9* to point *6*. From now on $r(6)=7$, since $r(6)=8$ was greater than $r(9)=7$. We also have to step back from point *6* to point *10*, but $r(10)$ remains 7. Since all output lines of point *10* are scanned we have to step back from it; but $r(10)=l(10)=7$, therefore point *10* is a root and the points labelled later belong to its component. Thus the first component consists of points *10*, *6* and *9*:




During the stepping backwards from point *10* to point *7* the $r$ values remain unchanged, in particular $r(7)=l(7)=6$. As we also have to step back from point *7*, we can identify the second component containing the single point *7*:

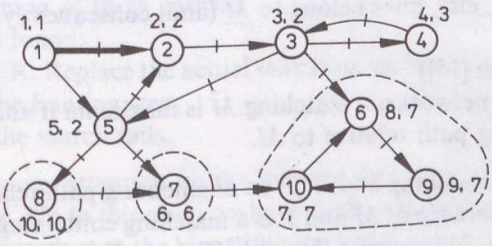


Thus, our actual point is once again the 5th. We visit point *8* along the line leading to it: $l(8)=r(8)=10$; the only output line leads to the labelled point *6* not lying on the actual path; therefore we cannot continue our travel here and the *r* values remain unchanged:

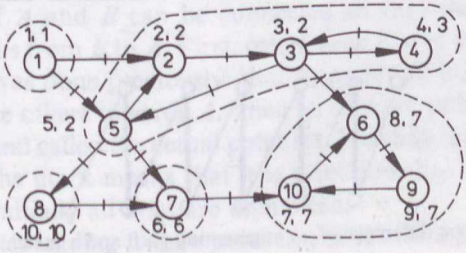



The third component is identified when stepping back from its only point *8*. We step back from points *8*, *5* and *3*. Note that $r(3)=r(5)=2$, because its previous *r* value (3) was greater than 2.

We also have to step back from point *2*; but $r(2)=l(2)$ shows that it is the root of the fourth component containing points *2*, *3*, *4* and *5*:




Finally, we scan the only unscanned line leading from point *1* to point *5*, but this causes no change. Now we step back from point *1* which forms a component on its own, and the procedure terminates:



*Output assignment*

Consider the case of systems of equations again. The last question related to the partitioning problem is how we may assign the output variables to the equations. We have already seen that in the graph theoretical formulation we ask the following: given a bipartite, undirected network with equal parts, how do we find a matching in it?

We present the currently most efficient algorithm of Hopcroft and Karp [12]. Fortunately, presenting this algorithm is much easier than proving its efficiency. As in practical applications this proof is not needed, we omit it and note only that the algorithm requires $o\left((n+m)\times\sqrt{n}\right)$ steps. Before discussing the algorithm we need some additional graph theory.

It is convenient to call each line set $M$ matching if no two lines of $M$ have a common end point, whatever the original network may be. A maximum matching is a matching of maximum cardinality. We say that the matching covers the end points of its lines; the remaining points are uncovered. Our aim is to find a maximum matching in bipartite networks. If this covers all points, then it determines a feasible output assignment, otherwise the original system of equations is structurally singular and so unsolvable.

Given a matching $M$, a simple path $P$ is called an augmenting path if its end points are both uncovered and its lines belong alternately to $M$ and $E$–$M$; that is, the second, fourth, etc., lines belong to $M$ (and, consequently, the others do not).

**Theorem 3.8** (Berge [15])

In undirected networks, a matching $M$ is maximum if and only if there exists no augmenting path relative to $M$.

**Proof**: Let $M$ be a matching and let $P$ an augmenting path relative to $M$. Then the symmetric difference of $M$ and $P$ is a matching containing one more line than $M$. (*Fig. 3.10*). This proves the 'only if' part of the theorem.

On the other hand, suppose that $M$ is not maximum and let $M'$ be a maximum matching. The symmetric difference $S$ of $M$ and $M'$ consists of disjoint, simple cycles and simple paths, because at most two lines of $S$ can join each point. The lines of these cycles and paths belong alternately to $M$ and $M'$; furthermore, the paths are augmenting paths relative to $M$. The question is whether such a path exists or whether $S$ contains only cycles. Each cycle

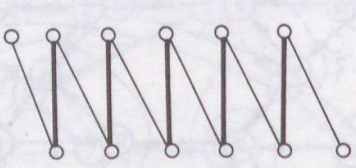


Fig. 3.10. The symmetric difference of an augmenting path and a matching in a larger matching

contains just as many lines of $M$ as of $M'$, and each path has one more line of $M'$ than of $M$. Since $M$ is not maximum, $M'$ contains more lines than $M$. Consequently, at least one path certainly exists in $S$. This completes the proof.

The first draft of the method could be the following: build up a matching $M$ line by line, starting with the empty set (which is a trivial matching) as actual matching $M$, try to find an augmenting path $P$ relative to $M$. If this is not possible, then terminate because, by Theorem 3.8, $M$ is a maximum matching. Otherwise, let the new actual matching be the symmetric difference of $M$ and $P$, and repeat the procedure.

This is, however, not the most efficient method. Seeking an augmenting path requires a careful analysis of the network. A significant improvement can be achieved by exploiting all information obtained in one search process. Indeed, it turns out that, when finding disjoint augmenting paths, only a little more effort is required to find as many as possible than finding a single one. Let $P'$ be the union of the disjoint augmenting paths; then the symmetric difference of the actual matching $M$ and $P'$ is again a matching and the cardinality increases by the number of paths in $P'$. Thus, the correct framework of the algorithm is as follows:

START: Let the actual matching $M$ be empty.
HUNTING STEP: Find a maximum set of disjoint augmenting paths relative to $M$. Let $P'$ be the union of these paths.
If such paths are found:
SWITCHING OVER: Replace the actual matching by the symmetric difference of $M$ and $P'$. Go to the hunting step.
TERMINATE, if the search fails.

We need to focus our attention on the hunting step since this is the nucleus of the procedure above. Up to this point we have dealt with arbitrary networks; from now on we restrict ourselves to the bipartite case. Let $A$ and $B$ denote the two parts of the network considered (in our application, the equation part and the variable part). The hunting step can be carried out in two stages.

*1. Extraction step*

We construct an auxiliary directed network which helps to find the augmenting paths. This takes $o(n+m)$ steps. We systematically map the way in which the uncovered points of $A$ and $B$ can be connected so that the augmenting paths become directed paths from $B$ to $A$: First, orient each line of the bipartite network in a similar way as was done previously: that is, orient all matching lines leading towards $B$ and all the others towards $A$. Then let us start with the uncovered part of $A$, denoted by $R$ and called the actual point set. We shall mark some lines of the directed network. The mark means that it is a relevant line for constructing the augmenting paths. Initially all lines are unmarked.

Now, mark the lines leading from a point not in $R$ into a point in $R$. Then add

all the starting points of these lines to $R$. Repeat this procedure until either any uncovered point of $B$ is reached or no line leads into $R$ from outside $R$.

If the uncovered part of $B$ cannot be reached by the method above then clearly no augmenting path exists with respect to the actual matching $M$.

Otherwise delete all unmarked lines. Add two new points $s$ and $t$ to the actual network and orient a line from $s$ to each uncovered point of $B$ and from each uncovered point of $A$ in $R$ to $t$. The auxiliary network is done. It is acyclic by virtue of its method of generation.

Obviously, any path from $s$ to $t$ corresponds to an augmenting path in the original network. Our next aim is to find a maximal set of disjoint directed paths from $s$ to $t$. Here 'disjoint' means that they have no common points but their end points, and 'maximal' means that the set cannot be enlarged. Note that the corresponding augmenting paths will be of the same length and disjoint: moreover, no shorter augmenting path exists. This property ensures that the whole hunting step can be repeated at most $2 \times \sqrt{n} + 2$ times. For the proof of this see Hopcroft and Karp [12].

*2. Maximal disjoint path set algorithm*

To find the desired directed paths in our directed, acyclic network we can use the labelling method discussed above. Indeed, label the points starting with $s$ (do not add any new point to this network because each point is reachable from $s$). Each time that $t$ can be reached, then the actual path is a new directed path from $s$ to $t$. Consequently, we have to modify only two steps of the labelling method:

SCANNING STEP/CASE 2: If the end point of the line actually scanned is $t$, then mark all the lines of the actual path together with the scanned lines as belonging to $P'$, the union of determined paths.

FORWARD STEP: If the newly reached point is $t$, then do the same as above (this is the first path leading from $s$ to $t$).

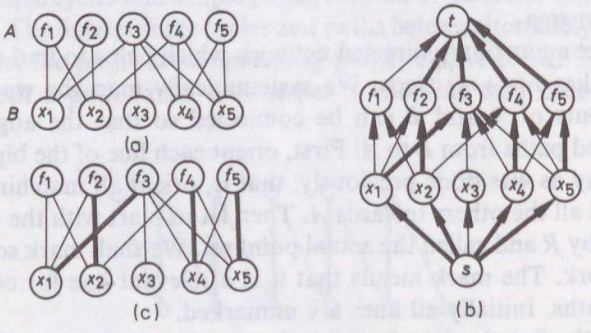



Fig. 3.11. Example: hunting step 1

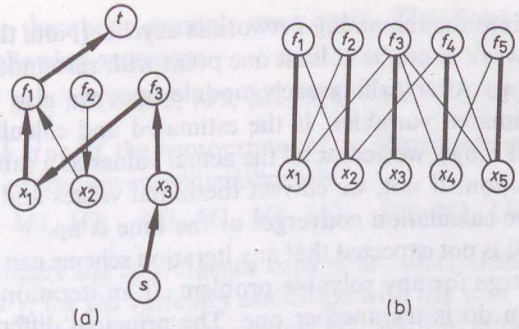



Fig. 3.12. Example: hunting step 2

Now the algorithm is totally described. The hunting step requires $o(n+m)$ operations for constructing the auxiliary network and another $o(n+m)$ operations to locate the maximal augmenting path set $P'$ by the labelling procedure. As we have mentioned (but not proved) above this can be carried out at most $o(\sqrt{n})$ times. Finally, we see that the total procedure requires $o((n+m)\times\sqrt{n})$ operations.

Let us illustrate the method by the system of equations Eq. (3.1). In *Fig. 3.11, a* we see the original undirected bipartite network. In the first extracting step all lines become marked because the initial $R$ is the whole equation part. All lines are oriented towards the equation part. The auxillary network is shown in *Fig. 3.11, b*. The disjoint paths found by the algorithm are shown as bold lines. In *Fig. 3.11, c* the matching lines are shown bold. Similarly, the result of the second (and final) hunting step is shown in *Fig. 3.12*.

### 3.2.3 Tearing

In the previous section we reduced the general problem to the case when the corresponding network is strongly connected. If it consists of more than one module then iterative calculation is usually used. A set of unknown variables corresponds to each line between the modules. All the variables associated with the same line may be handled similarly, allowing them to be simply represented by the line alone. We describe briefly how an iterative equation-solving procedure works in the *SM* approach. Throughout this section each network is supposed to be strongly connected.

First we select some lines called torn lines and give an estimate (or initial value) to each variable associated with them. The set of the torn lines is called the *tear set* and the corresponding variables are the *iterated variables*. The tear set must be selected so that all modules may be 'called' in turn, using the estimated values if necessary. Each untorn line must lead forwards with respect to the calling sequence. All directed cycles must contain at least one torn line. In other words, after

deleting the torn lines the remaining network is acyclic. (Note that by Lemma 3.3 the remaining network contains at least one point with zero indegree, that is, with all input lines torn.) After calling each module once we also obtain calculated values for the estimated variables. If the estimated and calculated values agree within a prescribed range, we accept all the actual values as a satisfactory approximation to the solution. If not, we correct the initial values and repeat the whole procedure until the calculation converges or the time is up.

Unfortunately, it is not expected that any iteration scheme can be developed that is certain to converge for any solvable problem. If an iteration scheme fails, the only thing one can do is try another one. The principal difference between the individual iteration schemes is how they determine the new initial values. These methods are discussed later except for the simplest and most commonly used one: direct iteration.

### Direct iteration scheme

Let the new initial values be equal to the last calculated values of the iterated variables.

By way of illustration consider a simple example consisting of four modules (*Fig. 3.13*). This network does not involve lines having no associated unknown variables.

Select the tear set $\{a, b\}$ and the calling sequence ($M2$, $M3$, $M1$, $M4$). One calculation proceeds as follows.

STEP 0. Determine the initial values of the iterated variables associated with lines $a$ and $b$.
STEP 1. Call module $M2$. Thus variables associated with $d$ and $f$ are calculated.
STEP 2. Call module $M3$. Now variables represented by $e$ and $c$ are calculated.
STEP 3. Call module $M1$. The iterated variables of line $a$ are re-calculated.
STEP 4. Call module $M4$. The variables of line $b$ also get re-calculated.

The convergence behaviour of any iterative solution to our problems depends on the tear set selection and the iteration scheme. Since the comparison of the different tear sets requires graph theoretical methods, we restrict ourselves to this problem in this section. The essential idea for comparing the tear sets is due to Upadhye and Grens [13].

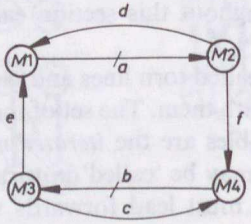



Fig. 3.13. Example network for direct iteration

Let us consider the above example once again. The direct iteration calls the modules in the following sequence:

$$M2, M3, M1, M4, \quad M2, M3, M1, M4, \quad M2, M3, M1, M4, \ldots$$

If we tear lines $b$, $d$ and $f$, the appropriate calling sequences may be $M3$, $M1$, $M4$, $M2$ and thus the procedure subsequently calls:

$$M3, M1, M4, M2, \quad M3, M1, M4, M2, \quad M3, M1, M4, M2, \ldots$$

In order to compare the convergence behaviour, select arbitrary initial values, say vectors $\underline{x}$ and $\underline{y}$, for the variables associated with the torn lines $a$ and $b$; then start an iteration procedure by calling module $M2$. We thus obtain some calculated values for the variables belonging to lines $d$ and $f$, say vectors $\underline{u}$ and $\underline{v}$. Now start a parallel direct iteration procedure, estimating variables of $b$ by $\underline{y}$ again and variables of $d$ and $f$ by $\underline{u}$ and $\underline{v}$. Obviously, the parallel calculations run identically, showing that there is no essential difference between the two calling sequences from our point of view. The reason for this is that the second sequence is a cyclic permutation of the first one. Generally, similar arguments prove that the convergence behaviour of the direct iteration procedure is not affected by a cyclical permutation of the calling sequence of the modules.

A precise investigation of the tear sets needs some generalization of the notion of a tear set. We allow the lines to be torn more than once. Of course, this would have no sense in practical applications (since it leads to a degraded convergence rate) but it turns out that some apparently reasonable tear sets have convergence properties that are identical with others containing multiple torn lines. It is appropriate to represent line sets by vectors, the $i$th value of which counts how many times the $i$th line is contained in the set. For example, vector $\underline{c}$ of a directed cycle has ones in the positions corresponding to the lines of the cycle, and zeros elsewhere. A vector $\underline{t}=(\underline{t}(e_1), \underline{t}(e_2), \ldots, \underline{t}(e_m))$ is a tear set, if all directed cycles contain at least one line $e$ with $\underline{t}(e)>0$. More formally, $t$ is a tear set if for each cycle $\underline{c}$:

$$\underline{c}^T \times \underline{t} > 0$$

holds. Now the base operation may be formulated:

**Replacement rule**

Let $N$ be a (strongly connected) network, $\underline{t}$ a tear set in it, and $x$ a point of $N$. Suppose that all input lines of $x$ are torn (we have already mentioned that such a point certainly exists). Then replace one copy of the input lines of $x$ in $\underline{t}$ by one copy of the output lines of $x$, that is, replace $\underline{t}$ by $\underline{t}'$, where:

$$\underline{t}'(e)=\underline{t}(e) \quad \text{if } e \text{ is not joined to } x$$

$$\underline{t}'(e)=\underline{t}(e)-1 \quad \text{if } e \text{ is the input of } x$$

$$\underline{t}'(e)=\underline{t}(e)+1 \quad \text{if } e \text{ is the output of } x.$$

**Theorem 3.9**

Applying the Replacement rule to a tear set $\underline{t}$, the result $\underline{t}'$ is again a tear set. Moreover, for each directed cycle $\underline{c}$ the numbers of times $\underline{c}$ is opened by $\underline{t}$ and $\underline{t}'$ are equal: $\underline{c}^{\mathrm{T}} \times \underline{t} = \underline{c}^{\mathrm{T}} \times \underline{t}'$.

**Proof**: If tear set $\underline{t}$ opens a cycle $\underline{c}$ at an input line $e$ of $x$, then $\underline{c}$ also contains an output line $f$ of $x$. Then $\underline{t}'$ opens cycle $\underline{c}\underline{t}(e)-1$ times at $e$ but $\underline{t}(f)+1$ times at $f$; therefore the total number of times $\underline{c}$ is opened remains unchanged.

The above theorem motivates the introduction of the following notion. Tear sets $\underline{t}$ and $\underline{t}'$ are said to be immediate relatives, denoted by $\underline{t}_{\bar{x}}\underline{t}'$, if $\underline{t}'$ can be obtained from $\underline{t}$ by applying the replacement rule at $x$. This involves that $\underline{t}(e)>0$ for each input line $e$ of $x$. Let us term two tear sets $\underline{t}$ and $\underline{t}'$ relatives, denoted by $\underline{t} \sim \underline{t}'$, if $\underline{t}'$ can be obtained from $\underline{t}$ by repeated applications of the Replacement rule, i.e. if there is a sequence of points $x_1, x_2, \ldots, x_k$ with:

$$\underline{t} = \underline{t}_0 \, \bar{x}_1 \, \underline{t}_1 \, \bar{x}_2 \, \underline{t}_2 \, \bar{x}_3 \cdots \underline{t}_k = \underline{t}'$$

The following theorem states that the above relationship between the tear sets is reversible:

**Theorem 3.10**

Tear sets $\underline{t}$ and $\underline{t}'$ are relatives if and only if $\underline{t}'$ and $\underline{t}$ are relatives.

**Proof**: Without loss of generality we may assume that $\underline{t}$ and $\underline{t}'$ are immediate relatives: $\underline{t}_{x_1}\underline{t}'$. Delete all the lines torn by $\underline{t}$. Since at least one line of each directed cycle is deleted, the remaining network is acyclic. By Theorem 3.4 there exists a precedence ordering of the points, i.e. a sequence $x=x_1, x_2, \ldots$ $\ldots, x_k$ so that each line leads forwards. Apply the replacement rule at $x=x_1$; the resulting tear set is $\underline{t}'$ and, obviously, a precedence ordering corresponding to it is $x_2, x_3, \ldots, x_k, x_1$. Applying the Replacement rule at the first point of any precedence ordering is always possible, since all its input lines must be torn, and shifting this point to the end of the sequence yields a precedence ordering for the newly obtained tear set. Consequently, we can apply the Replacement rule at $x_2, x_3, \ldots, x_k$. Thus $\underline{t}$ is reached again, proving that $\underline{t}'$ and $\underline{t}$ are relatives.

Let us define the *family* of a tear set $\underline{t}$ as the set containing all its relatives. The discussion at the beginning of this section showed that the convergence properties of relative tear sets are identical. A tear set $\underline{t}$ is redundant if there exists at least one line $e$, the value of which can be decreased so that the resulting set is again a tear set. A nonredundant family has only nonredundant members; a redundant family contains at least one redundant member. (This notion differs from that of Upadhye and Grens. They call a family redundant if all its members are redundant, and define the mixed families as those containing redundant and also nonredundant members. In our discussion there is no need to make such a distinction between families containing redundant tear sets.) A redundant family always has members

with multiple-torn lines. It follows that each tear set belonging to a redundant family has similar convergence properties as if it contained multiple-torn lines itself; therefore it is desired to select a tear set from a nonredundant family. The main problem concerning the tear sets arises from the fact that a redundant family may also contain nonredundant members.

For a careful analysis we need a further definition. Let us select an arbitrary point $x$ of the network. We say that a tear set $\underline{t}$ is $x$-canonical if $x$ is the only point for which each input line is torn.

In other words, $\underline{t}$ is $x$-canonical if the Replacement rule can be applied only at $x$. The most important fact is the following (Pretzel [14]):

**Theorem 3.11**

Each family contains exactly one $x$-canonical member, for arbitrary point $x$.

**Proof**: We omit the proof since it requires too deep an insight into the structure of directed cycles.

We have already mentioned that each redundant family has members with multiple-torn lines. This statement can be made slightly more strict: there always exists a point $x$ for which the $x$-canonical member contains multiple-torn lines. Consequently, we can easily check the nonredundancy of a family by determining all its canonical members. Now we are in a position to describe the algorithm.

**Nonredundant family algorithm *(NRF)***

First select an arbitrary enumeration of the points: $x_1, x_2, \ldots, x_k$ and any tear set $\underline{t}$ (for example, the trivial tear set $\underline{t}(e)=1$ for each line $e$). The algorithm is completed in $n$ steps; in step $i$ an $x_i$-canonical tear set is determined by the following method ($i=1, \ldots, n$):

1. Find a point other than $x_i$ where the Replacement rule can be applied. If no such point exists, then the step is complete. Otherwise:
2. Apply the Replacement rule to that point.
3. Let $\underline{t}(e)=\min(\underline{t}(e), 1)$ for each line $e$. That is, only one copy is kept torn from the multiple-torn lines. Go to 1.

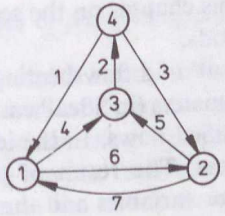



Fig. 3.14. Network to be torn

After completion of this algorithm, the actual tear set belongs to a nonredundant family. The proof of this claim is also omitted here.

Finally, the nonredundant family algorithm requires $o(n^3)$ running time and $o(n^2)$ space, where $n$ is the number of the points.

For the illustration of this procedure let us consider the network in *Fig. 3.14*. We start with the trivial tear set where all lines are torn. The algorithm yields the following canonical forms (note that they may possibly be all relatives):

1-canonical tear set: lines *1*, *3* and *7*
2-canonical tear set: lines *3*, *4* and *6*
3-canonical tear set: lines *4*, *5* and *7*
4-canonical tear set: lines *2* and *7*.

## 3.3 Flowsheeting

Flowsheeting is defined as the use of computers to perform steady-state heat and mass balancing, sizing and costing calculations [1]. So defined, flowsheeting may be seen to be an essential part of the chemical process design activity.

The whole process of chemical process design can be broken down into three basic steps:

process synthesis,
process analysis, and
optimization.

The aim of process synthesis is to select the process structure (flowsheet structure), the equipment to be used and its interconnection, and to specify the value of the decision (design) variables.

Process analysis means the simulation of the behaviour of the proposed flowsheet, sizing the necessary equipment and the profitability analysis of the process. This step gives an answer to the 'what if' type questions occurring during the design process.

The third step, optimization interacts with both steps described above, since different objects can be compared only if they are optimized in some sense using the same appropriate objective function. Both the process structure and the process parameters have to be optimized.

We focus our attention in this chapter on the second step: we deal with process analysis by flowsheeting methods.

In the simplest case, the 'input' of a flowsheeting calculation is a flowsheet with some specified equipment parameters (in 'ideal' cases all equipment parameters are specified), and with some specified flows. In the ideal case the specified flows are the feed streams of the process. The result of the flowsheet simulation is the steady-state value of the stream variables and the missing equipment parameters (if any). If (and only if) the flowsheet specification is correct can the steady-state

heat and mass balance be solved, and the sizing and costing calculations be performed.

In analyzing chemical engineering systems by flowsheeting methods, the flowsheet is characterized by basic elements of two types:

— unit operation module, and
— stream.

A unit operation module represents a device for changing one or more parameters of the inlet streams of a unit (equipment). The units are interconnected with material and/or energy and/or information flows, represented by the streams. The stream variables must not change between two unit operation modules. Concisely speaking, we can say that the unit operation module is the information transformer, while the stream maintains its information content between two units, its only task being to ensure that information is transferred ('communication') between the units.

In order to analyze whether the flowsheet definition is correct or not, one has to perform a so-called 'degree of freedom' analysis. After a brief discussion about the number of degrees of freedom, we shall establish its value for a process stream, for different unit operation modules, and for a complete flowsheet. Thereafter we shall try to find an appropriate means to satisfy the degrees of freedom, and we shall then introduce a method to reveal 'badly posed' flowsheeting problems, i.e. flowsheeting problems which have been incorrectly specified.

### 3.3.1 Degrees of freedom analysis

If we have to solve a set of equations with $n$ independent equations in $m$ unknown variables ($m>n$), the number of degrees of freedom, $d$ is simply the number of 'excess' variables:

$$d=m-n \tag{3.3}$$

In other words, $m-n$ variables must receive their values from 'external', independent sources before calculation starts. This is equivalent to adding $m-n$ independent equations to the set of original equations.

*Degrees of freedom of a stream*

Let consider a flowing system in the steady state. We limit our discussion to real material streams comprising $c$ chemical components. The streams of such a system are in complete (thermal, phase, chemical, etc.) equilibrium. The number of degrees of freedom indicates how many stream variables must be specified to completely describe the state of the stream in the steady state.

We know (or can conclude using common sense only) that if we specify the flow

of each chemical species ($c$ specifications) and fix the pressure and temperature (or, in the case of one-component streams, the enthalpy of the stream instead of temperature), then the state of the stream is completely and unambiguously specified.

So we can take the working hypothesis that

$$d = c + 2 \tag{3.4}$$

How can we deduce this result from basis chemical engineering principles?

Westerberg *et al.* [1] have presented a deduction using Duhem's theorem and supposing that $r$ independent reactions take place in the system. Since, in a flowsheeting calculation, reactions are not supposed to take place in a stream, we have introduced a slight modification to their deduction.

The calculation of the number of degrees of freedom simply means counting the independent equations describing the system, and counting the variables occurring in the equations. We too will follow this procedure.

The multiphase, multicomponent system in question has $c$ chemical components, which can be distributed among $p$ phases.

$T$ is used for temperature,
$P$ is used for pressure,
$F$ is used for total molar flow of a stream, and
$\mu$ is used for chemical potential.

Each variable has a maximum of two (lower) indices, the first referring to the chemical component, the second to the phase.

We will construct a table containing all of the necessary (independent) equations, the number of them and the 'new' variables associated with the equations (which have not occurred previously).

Let us consider a volume element of the stream in the steady state.

The flow of material through this volume is constant:

$$f_i = f_i^0 \qquad 1 = 1, 2, \ldots c \tag{3.5}$$

The stream is in thermal and pressure equilibrium: the temperature and pressure in each phase is the same

$$T_j = T \qquad j = 1, 2, \ldots, p \tag{3.6}$$

$$P_j = P \qquad j = 1, 2, \ldots, p \tag{3.7}$$

Since the system is in chemical and phase equilibrium, the chemical potentials are the same for any chemical species in each phase:

$$\mu_{ij} = \mu_i \qquad i = 1, 2, \ldots, c; \quad j = 1, 2, \ldots, p \tag{3.8}$$

The chemical potential can be expressed as a function of the state variables and the composition of the stream:

$$\mu_{ij} = \mu_{ij}(T, P, x_{kj}; k = 1, 2, \ldots, c) \qquad i = 1, 2, \ldots c; \quad j = 1, 2, \ldots, p \tag{3.9}$$

The sum of the mole fractions in each phase must be equal to unity:

$$\sum_{i=1}^{c} x_{ij} = 1 \qquad j = 1, 2, \ldots, p \tag{3.10}$$

The mole fractions are related to the molar flow rate of the species:

$$\sum_{j=1}^{p} x_{ij} F_j = f_i \qquad i = 1, 2, \ldots, c \tag{3.11}$$

If we collect Eqs (3.5) to (3.11) into one table, and count the number of equations and variables, we get *Table 3.1.*

Table 3.1. Equations and variables describing the state of a multiphase, multicomponent stream in the steady state

| Equations | Number of equations ($n$) | Number of variables ($m$) |
|---|---|---|
| Material balance: | | |
| $f_i = f_i^0 \quad i = 1, 2, \ldots, c$ | $c$ | $2c$ |
| Thermal and pressure equilibrium: | | |
| $T_j = T \quad j = 1, 2, \ldots, p$ | $p$ | $p+1$ |
| $P_j = P \quad j = 1, 2, \ldots, p$ | $p$ | $p+1$ |
| Chemical and phase equilibrium: | | |
| $\mu_{ij} = \mu_i \quad i = 1, 2, \ldots, c \quad j = 1, 2, \ldots, p$ | $cp$ | $(p+1)c$ |
| The chemical potential as a function of state variables and composition: | | |
| $\mu_{ij} = \mu_{ij}(T, P, x_{kj};\ k = 1, 2, \ldots, c) \quad i = 1, 2, \ldots, c \quad j = 1, 2, \ldots, p$ | $cp$ | $cp$ |
| Mole fraction sum equations: | | |
| $\sum_{i=1}^{c} x_{ij} = 1 \quad j = 1, 2, \ldots, p$ | $p$ | 0 |
| Mole fractions related to the molar flow rates: | | |
| $\sum_{j=1}^{p} x_{ij} F_j = f_i \quad i = 1, 2, \ldots, c$ | $c$ | $p$ |
| Total | $n = 2cp + 3p + 2c$ | $m = 2cp + 3c + 3p + 2$ |

Since $d=m-n=c+2$, our hypothesis holds true, the total number of the degrees of fredom associated with a stream (in steady state) is $c+2$. In addition to the molar flow rate of the stream one has to specify 2 stream variables to account for all the degrees of freedom.

*Degrees of freedom of a unit*

A unit operation module represents a device for changing one or more parameters of its inlet stream(s). In a real plant this transformation is realized by different process equipment. For instance pressure change can be achieved by pumps, compressors, valves or by any part of the process equipment having a pressure gradient. Temperature change can be achieved by heat exchangers, direct fired heaters, or in a reactor where an exothermic or endothermic reaction takes place, and so on. Compositional change occurs in a reactor, in a flash evaporator, in the multistage separation processes (distillation, extraction), etc.

The unit operation modules are represented in the flowsheeting system by unit operation models. In general, the model is a nonlinear equation describing the transformation taking place in the unit operation module. Functionally, the model can be described as

$$f(u, x, y, z)=0 \tag{3.12}$$

where $u$ is the vector of model parameters,
$x$ is the vector of inlet stream variables,
$y$ is the vector of outlet stream variables,
$z$ is the vector of internal variables, and calculated equipment parameters.

This representation of the unit models is clear-cut:

— The number of independent equations is equal to the sum of the number of outlet stream variables, internal variables (such as intermediate stage temperatures, compositions in a distillation column model) and calculated equipment parameters (such as power requirement in a pump or compressor, heat duty for a process heater).

— The number of degrees of freedom is equal to the sum of the model parameters and the inlet stream variables. A similar representation is given by Evans [18].

The formal description detailed above shows that the number of degrees of freedom depends greatly on the model describing the performance of the unit operation. This is an important conclusion, since the same unit operation can be described by different models which differ in the level of detail and which therefore have different model parameters. If we change the models in a flowsheeting calculation, the number of degrees of freedom of the unit models (and consequently that of the flowsheet) will change.

After this short introduction we shall now present some rather simple unit operation models, showing the way to ascertain the number of degrees of freedom of a unit.

Table 3.2. Unit operation models and their function

| Model | Function |
|---|---|
| Mixer | To produce a stream as a combination of two (or more) streams |
| Stream splitter | To split one stream into two (or more) streams without changing their composition |
| Flash drum | To divide a two-phase stream into a vapour and a liquid stream |
| Pump | Pressure rise |
| Compressor | To compress a gas stream |
| Valve | Pressure reduction or flow regulation |
| Heat exchanger | Heating or cooling a process stream with another stream |
| Process furnace | To elevate the heat content of a stream considerably |
| Reactor | To change the composition of a stream by chemical reaction |
| Separator | To separate one stream into two (or more) streams differing in composition. In most cases simple 'fraction separated' models |
| Distillation column | To separate one stream into two (or more) streams differing in composition. Ideal or 'rigorous' distillation models |

The unit operation models and their functions are compiled in *Table 3.2.*

Before going on to present the simple models themselves, we now write down the notation used, which will have the same meaning throughout the text of this chapter.

$c$ — number of components,

$n$ — number of streams attached to the same module, and number of equilibrium stages in case of the distillation column model,

$F_{ij}$ — molar flow of component $i$ in stream $j$,
$x_{ij}$ — mole fraction of component $i$ in stream $j$,
$T_j$ — temperature of stream $j$,
$P_j$ — pressure of stream $j$,
$H_j$ — the molar enthalpy of stream $j$,
$K_i$ — the vapour/liquid equilibrium coefficient of component $i$,
$Q$ — heat input to the stream.

The first unit will be a simple mixer, which combines two streams into one. The equations describing this unit and the variables associated with them are collected in *Table 3.3. Figure 3.15* is a schematic representation of this simple module. The degree of freedom is $d=2c+4$.

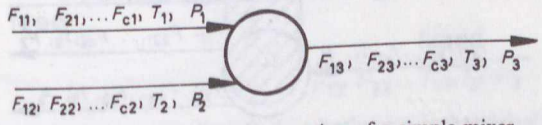



Fig. 3.15. Schematic representation of a simple mixer

Table 3.3. A simple mixer model

| Equations | Number of equations ($n$) | Number of variables ($m$) |
|---|---|---|
| Component mass balances: $F_{i3}-F_{i2}-F_{i1}=0$ | $c$ | $3c$ ($F_i j$) $j=1, 2, 3$ |
| Enthalpy balance: $H_3 \sum_{i=1}^{c} F_{i3}-H_2 \sum_{i=1}^{c} F_{i2}-H_1 \sum_{i=1}^{c} F_{i1}=0$ | 1 | 6 ($T_j, P_j$) $j=1, 2, 3$ |
| where $H_j=h_j(x_{1j}, \ldots, x_{cj}, T_j, P_j)$ $(j=1, 2, 3)$ and $x_{ij}=\frac{F_{ij}}{\sum_{i=1}^{c} F_{ij}}$ $(i=1, \ldots, c; j=1, 2, 3)$ | | |
| Pressure expression for the output stream: $P_3=\min(P_1, P_2)$ | 1 | |
| Total | $n=c+2$ | $m=3c+6$ |

Note that

(a) The result is trivial, since if the two inlet streams are specified, these specifications require $2(c+2)=2c+4$ specifications, and the unit module is completely specified.

(b) A similar deduction could be made for mixers with $n$ inlet streams, giving the result that the number of degrees of freedom is equal to $n(c+2)$.

The next unit to be considered is the stream splitter. The simple model can represent a blow-down, purge, pipe junction to distribute a stream into substreams, or a three-way valve. The schematic representation of the stream splitter is depicted in *Fig. 3.16*.

The equations describing this unit and the variables associated with them are collected in *Table 3.4*. The degree of freedom is $d=c+3$.

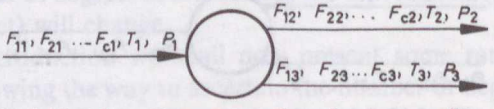



Fig. 3.16. Schematic representation of a stream splitter

Table 3.4. A simple stream splitter model

| Equations | Number of equations ($n$) | Number of variables ($m$) |
|---|---|---|
| Component mass balances: | | |
| $F_{i2}-aF_{i1}=0$ | $c$ | $2c+1\,(F_{ij}, a)\ j=1, 2$ |
| $F_{i3}+F_{i2}-F_{i1}=0$ $(i=1, \ldots, c)$ | $c$ | $c\,(F_{i3})$ |
| Equations for temperature and pressure: | | |
| $T_1=T_2=T_3$ | 2 | $3\,(T_j)$ |
| $P_1=P_2=P_3$ | 2 | $3\,(P_j)$ |
| Total | $n=2c+4$ | $m=3c+7$ |

Note that

(a) The result is trivial, since if the inlet stream and the split fraction are specified, these specifications require $c+3$ specifications, and the unit module is completely specified.

(b) A similar deduction could be made for splitters with $n$ outlet streams, giving the result that the number of degrees of freedom is equal to $c+n+2$. In this latter case, however, the relation $\sum_{j=1}^{n} a_j=1$ must hold.

The third unit to be discussed is the isothermal-isobaric flash separator. The schematic representation of the flash separator is depicted in *Fig. 3.17.*

The equations describing this unit and the variables associated with them are collected in *Table 3.5.* The degree of freedom is $d=c+2$.

Note that

(a) The result is trivial, since if the inlet stream is specified, these specifications require $c+2$ specifications, and the unit module is completely specified. (The stream 'knows' how to distribute its components between the two phases.)

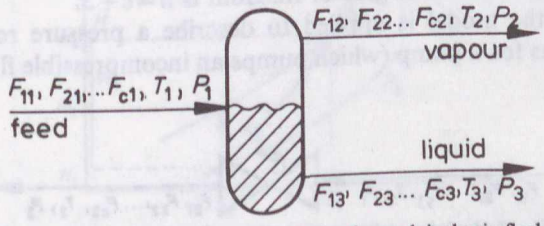



Fig. 3.17. Schematic representation of an isothermal–isobaric flash unit

Table 3.5. A simple isothermal-isobaric flash model

| Equations | Number of equations ($n$) | Number of variables ($m$) |
|---|---|---|
| Component mass balances: | | |
| $F_{i1}-F_{i2}-F_{i3}=0 \quad (i=1, \ldots, c)$ | $c$ | $3c\,(F_{ij})\ j=1,\ 2,\ 3$ |
| Performance equations: | | |
| $x_{i2}=K_i\,x_{i_3} \quad (i=1, \ldots, c)$ | $c$ | $2\,(T_3, P_3)$ |
| where | | |
| $x_{ij}=\dfrac{F_{ij}}{\sum_{i=1}^{c} F_{ij}} \quad (i=1, \ldots, c;\ j=1,\ 2)$ | | |
| and | | |
| $K_i=K_i\,(x_{12}, x_{22}, \ldots, x_{c2}, x_{13}, x_{23}, \ldots, x_{c3}, T_3, P_3)$ $(i=1, \ldots, c)$ | | |
| Equations for temperature and pressure: | | |
| $T_1=T_2=T_3$ | 2 | $2\,(T_1,\ T_2)$ |
| $P_1=P_2=P_3$ | 2 | $2\,(P_1,\ P_2)$ |
| Total | $n=2c+4$ | $m=3c+6$ |

(b) A similar deduction could be made for a flash drum to which some heat input, $Q$, is allowed giving the result that the number of degrees of freedom is $c+3$.

We will consider the simple model of the pump and that of the pressure relief valve together. These modules can be described by the same model, differing only in the sign of the pressure change. The schematic representation of a valve is depicted in *Fig. 3.18*.

The equations describing these units and the variables associated with them are collected in *Table 3.6*. The degree of freedom is $d=c+3$.

Note that if the model is applied to describe a pressure relief valve, $\Delta P$ is negative, whereas for a pump (which pumps an incompressible fluid) it is positive.

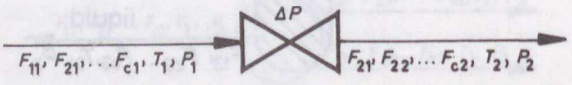



Fig. 3.18. Schematic representation of a pressure relief valve

Table 3.6. Model equations of a simple pressure relief valve or a pump

| Equations | Number of equations ($n$) | Number of variables ($m$) |
|---|---|---|
| Component mass balances: $F_{i2}-F_{i1}=0 \quad (i=1,\ldots,c)$ | $c$ | $2c(F) \quad j=1, 2$ |
| Equation for the pressure of the output stream: $P_2=P_1+\Delta P$ | 1 | $3(P_j, \Delta P) \quad j=1, 2$ |
| Enthalpy balance: $H_2-H_1=0$ | 1 | |
| where $H_j=h_j(x_{1j}, \ldots, x_{cj}, T_j, P_j) \quad (j=1, 2)$ | | $2(T_j) \quad j=1, 2$ |
| and $x_{ij}=\dfrac{F_{ij}}{\sum_{i=1}^{c} F_{ij}} \quad (i=1,\ldots,c;\ j=1, 2)$ | | |
| Total | $n=c+2$ | $m=2c+5$ |

The sixth unit to be considered is the (one stage) gas compressor. Multistage compressor models can be constructed from the one-stage compressor model and (if the interstage cooling effect is to be modelled) from the heat exchanger models.

We know from practical experience that such a unit has two unit parameters which must be specified: the pressure rise of the stream and the adiabatic efficiency of the compressor. In some cases, however, instead of the adiabatic efficiency one specifies directly the actual adiabatic compression work necessary to compress one mole of the fluid from $P_1$ to $P_2$. These two unit parameters are dependent upon each other. This interdependency is illustrated in *Fig. 3.19.*

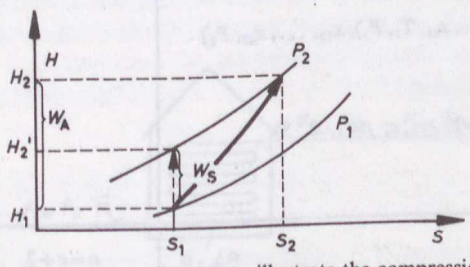



Fig. 3.19. Enthalpy–entropy diagram to illustrate the compression of a gas

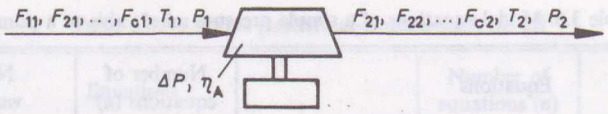



Fig. 3.20. Schematic representation of a gas compressor

The reversible work (which belongs to the 'ideal', isentropic compression) is $W_s$. The real process is irreversible, and therefore we can reach the isobaric line $P_2$ only with a substantial entropy rise (point $S_2, P_2, H_2$). The work attached to this change is $W_A = H_2 - H_1$, and is called actual adiabatic work. The adiabatic efficiency is a measure of the ratio of the isentropic work to the actual adiabatic work:

$$\eta_A = \frac{W_s}{W_A} = \frac{H_2' - H_1}{H_2 - H_1} \tag{3.13}$$

The equations describing this unit and the variables associated with the equations are collected in *Table 3.7. Figure 3.20* is a schematic representation of a gas compressor. The degree of freedom is $d = c + 4$.

Table 3.7. Simplified model of a gas compressor

| Equations | Number of equations ($n$) | Number of variables ($m$) |
|---|---|---|
| Component mass balances:<br>$F_{i2} - F_{i1} = 0$ | $c$ | $2c\,(F_{ij})\quad j=1,\ 2$ |
| Pressure change:<br>$P_2 = P_1 - \Delta P$ | 1 | $3\,(P_j, \Delta P)\quad j=1,\ 2$ |
| Enthalpy balance:<br>$H_2' - H_1 = \eta_A (H_2 - H_1)$<br>where<br>$H_j = h_j (x_{1j}, \ldots, x_{cj}, T_j, P_j)\quad (j=1,\ 2)$<br>$H_2' = h_2' \left(s(x_{11}, \ldots, x_{c1}, T_1, P_1), x_{12}, \ldots, x_{c2}, P_2\right)$<br>and<br>$x_{ij} = \dfrac{F_{ij}}{\sum_{i=1}^{c} F_{ij}}\quad (i=1, \ldots, c;\ j=1,\ 2,\ 3)$ | 1 | $3\,(T_1, T_2, \eta_A)$ |
| Total | $n = c+2$ | $m = 2c+6$ |

Table 3.8. A simple process furnace model

| Equations | Number of equations ($n$) | Number of variables ($m$) |
|---|---|---|
| Component mass balances: | | |
| $F_{i2}-F_{i1}=0 \quad (i=1, \ldots, c)$ | $c$ | $2c(F_{ij}) \quad j=1, 2$ |
| Pressure change: | | |
| $P_2=P_1-\Delta P$ | 1 | $3(P_j, \Delta P) \quad j=1, 2$ |
| Enthalpy balance: | | |
| $H_2 \sum_{i=1}^{c} F_{i2}-H_1 \sum_{i=1}^{c} F_{i1}-Q=0$ | 1 | $3(T_1, T_2, Q)$ |
| where | | |
| $H_j=h_j(x_{ij}, \ldots, x_{cj}, T_j, P_j) \quad (j=1, 2)$ | | |
| and | | |
| $x_{ij}=\dfrac{F_{ij}}{\sum_{i=1}^{c} F_{ij}} \quad (i=1, \ldots, c;\ j=1, 2, 3)$ | | |
| Total | $n=c+2$ | $m=2c+6$ |

It is clear from the above result that beside specifying the stream variables one has to specify two parameters, which is exactly what was expected.

The next unit to be considered is the process furnace. The process stream flows inside the furnace tube, and is heated by radiation and convective heat transfer through the wall of the furnace. The schematic representation of a process furnace is depicted in *Fig. 3.21.*

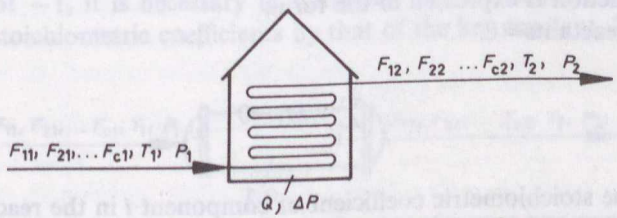



Fig. 3.21. Schematic representation of a process furnace

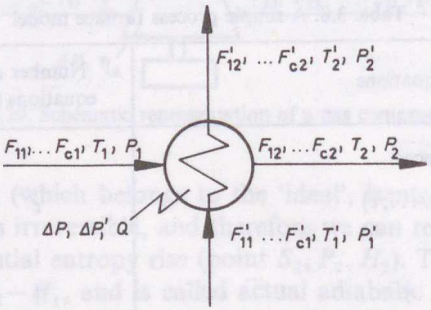



Fig. 3.22. Schematic representation of a heat exchanger

The equations describing this unit and the variables associated with them are collected in *Table 3.8*. The degree of freedom is $d=c+4$, which is equivalent to the statement that the furnace is completely defined if one of its streams and two other variables (e.g. the two unit parameters, $Q$ and $\Delta P$) are specified.

The eighth unit to be discussed is the heat exchanger. In this unit one stream exchanges heat with another (most frequently with a utility) stream through the tube wall. The schematic representation of the heat exchanger is depicted in *Fig. 3.22*. The variables of the utility stream are denoted by $c'$, $F'$, $H'$, $T'$, $P'$ and $x'$, respectively.

In writing the necessary equations, we assume that the heat absorbed by one of the streams is equal to the heat lost by the other, i.e. the heat loss to the environment is neglected. The equations describing this unit and the variables associated with them are collected in *Table 3.9*. The number of degrees of freedom is $d=$ $=c+c'+7$. These can be taken up (in the ideal case) by specifying the two input streams ($c+c'+4$ specifications), defining the pressure change on both sides of the exchanger, and specifying the value of $Q$.

Next we consider a simple 'extent of reaction' type reactor model with one reaction.

The schematic representation of this simple reactor model is depicted in *Fig. 3.23*. If the reactor is adiabatic, $Q$, the rate of heat addition to the reactor, may be specified as zero. The extent of a reaction is defined as the number of moles of a particular 'key' component with a stoichiometric coefficient $-1$ reacting in unit time. The reaction is expressed in the form:

Products-reactants $=0$.

$$\sum_{i=1}^{c} \nu_i M_i = 0 \tag{3.14}$$

where $\nu_i$ is the stoichiometric coefficient of component $i$ in the reaction, and
$M_i$ is the chemical formula of component $i$.

Table 3.9. A simple heat exchanger model

| Equations | Number of equations ($n$) | Number of variables ($m$) |
|---|---|---|
| Component mass balances: | | |
| $F_{i2}-F_{i1}=0 \quad (i=1, \ldots, c)$ | $c$ | $2c\,(F_{ij}) \quad j=1, 2$ |
| $F'_{i2}-F'_{i1}=0 \quad (i=1, \ldots, c')$ | $c'$ | $2c'\,(F'_{ij}) \quad j=1, 2$ |
| Pressure change: | | |
| $P_2=P_1-\Delta P$ | 1 | $3\,(P_j, \Delta P) \quad j=1, 2$ |
| $P'_2=P'_1-\Delta P'$ | 1 | $3\,(P'_j, \Delta P') \quad j=1, 2$ |
| Enthalpy balance: | | |
| $H_2 \Sigma F_{i2}-H_1 \Sigma F_{i1}-Q=0$ | 1 | $3\,(T_1, T_2, Q)$ |
| $H'_2 \Sigma F'_{i2}-H'_1 \Sigma F'_{i1}-Q=0$ | 1 | $2\,(T'_1, T'_2)$ |
| where | | |
| $H_j=h_j\,(x_{ij}, \ldots, x_{cj}, T_j, P_j) \quad (j=1, 2)$ | | |
| $H'_j=h_j\,(x'_{ij}, \ldots, x'_{cj}, T'_j, P'_j) \quad (j=1, 2)$ | | |
| and | | |
| $x_{ij}=\dfrac{F_{ij}}{\Sigma F_{ij}} \quad (i=1, \ldots, c;\ j=1, 2)$ | | |
| $x'_{ij}=\dfrac{F'_{ij}}{\Sigma F_{ij}} \quad (i=1, \ldots, c';\ j=1, 2$ | | |
| Total | $n=c+c'+4$ | $m=2c+2c'+11$ |

The equations describing this unit and the variables associated with them are collected in *Table 3.10*. Note that if the stoichiometric coefficient of the key reactant is not $-1$, it is necessary to normalize the stoichiometric equation by dividing the stoichiometric coefficients by that of the key reactant. The degree of

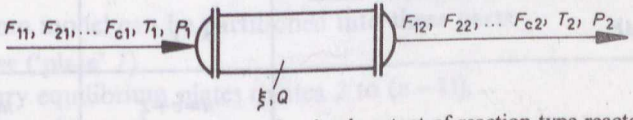



Fig. 3.23. Schematic representation of a simple extent of reaction type reactor model

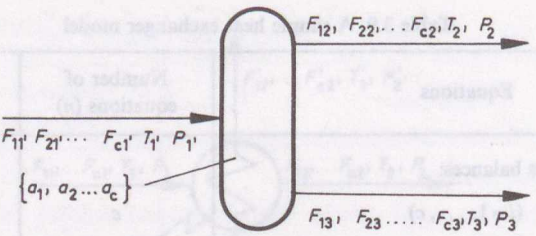



Fig. 3.24. Schematic representation of a multicomponent split unit

freedom is $d=c+5$. The equipment parameters of the reactor are the heat input ($Q$), the extent of reaction, $\xi$, and the pressure drop, $\Delta P$.

Note that $-\Delta H_R$, the molar heat of reaction, is an intermediate variable and can be calculated from basic physical properties of the species.

The simple separator, or multicomponent split unit is a useful means to model items of a process plant as fractionating columns in an approximate way. In effect it is an application of the 'fraction separated' method first proposed by Vela [19].

The schematic representation of the multicomponent split unit is depicted in *Fig. 3.24.*

Table 3.10. Model equations of a simple 'extent of reaction' type reactor model

| Equations | Number of equations ($n$) | Number of variables ($m$) |
|---|---|---|
| Component mass balances:<br>$F_{i2}-F_{i2}=\nu_i \xi \quad (i=1, \ldots, c)$ | $c$ | $2c+1\,(F_{ij}, \xi) \quad j=1, 2$ |
| Enthalpy balance:<br>$H_2 \Sigma F_{i2}-H_1 \Sigma F_{i1}-(-\Delta H_R)-Q=0$<br>where<br>$H_j=h_j(x_{1j}, \ldots, x_{ij}, T_j, P_j) \quad (j=1, 2)$<br>and<br>$x_{ij}=\dfrac{F_{ij}}{\Sigma F_{ij}} \quad (i=1, \ldots, c;\ j=1, 2, 3)$ | 1 | $5\,(T_j, P_j, Q) \quad j=1, 2$ |
| Equation for the pressure of the output stream:<br>$P_2=P_1-\Delta P$ | 1 | $1\,(\Delta P)$ |
| Total | $n=c+2$ | $m=2c+7$ |

Table 3.11. A simple multicomponent split unit model

| Equations | Number of equations ($n$) | Number of variables ($m$) |
|---|---|---|
| Component mass balances: | | |
| $F_{i2}-a_i F_{i1}=0 \quad (i=1,\ldots,c)$ | $c$ | $3c\,(F_{ij}, a_i) \quad j=1, 2$ |
| $F_{i3}-(1-a_i)F_{i1}=0 \quad (i=1,\ldots,c)$ | $c$ | $c\,(F_{i3})$ |
| Pressure equation: | | |
| $P_1=P_2=P_3$ | 2 | $3\,(P_j) \quad j=1, 2, 3$ |
| Temperature equilibrium: | | |
| $T_1=T_2=T_3$ | 2 | $3\,(T_j) \quad j=1, 2, 3$ |
| Total | $n=2c+4$ | $m=4c+6$ |

The equations describing this unit and the variables associated with them are collected in *Table 3.11*. The degree of freedom is $d=2c+2$. The equipment parameters of the multicomponent split unit are the fractions of the components.

*Degree of freedom analysis of a distillation column*

The model by which the degree of freedom analysis is carried out here is that of Naphtali and Sandholm [20]. We discuss this model for two reasons:

— This example illustrates the principle that the number of degrees of freedom is associated with the model, and is not unambiguously determined by the unit operation itself. The more rigorous the model, however, the closer the number of degrees of freedom is to the number of input stream variables and decision variables of the unit operation. The multicomponent split unit model and the Naphtali—Sandholm model both describe the distillation unit, but with different levels of rigour.

— This model is composed of three submodels. The calculation of the number of degrees of freedom for the whole column is based on a calculation of the number of degrees of freedom of the submodels. This method leads to the calculation of the number of degrees of freedom of a system based on a knowledge of the number of degrees of freedom of its subsystems. This mode of thought will be applied to the calculation of the number of degrees of freedom of a flowsheet.

The column model can be partitioned into three parts:

— reboiler ('plate' *1*),
— ordinary equilibrium plates (plates *2* to $(n-1)$),
— condenser (plate $n$).

For the sake of simplicity we deal only with a column equipped with a total condenser, having only one feed stream and no liquid or vapour side streams. Extended models with a partial condenser, any number of feed streams and side streams have been discussed in the literature, e.g. Lee [21].

The abbreviations used in describing the column model are as follows:

$L_{ij}$ — molar liquid flow of $j$th components from plate $i$,
$L_i$ — total molar liquid flow from plate $i$,

$$L_i = \sum_{i=1}^{c} L_{ij} \tag{3.15}$$

$V_{ij}$ — molar vapour flow of $j$th component from plate $i$,
$V_i$ — total molar vapour flow from plate $i$,

$$V_i = \sum_{i=1}^{c} V_{ij} \tag{3.16}$$

$D_j$ — molar flow of $j$th component of the distillate (liquid, from 'plate' $n$),
$D$ — total molar flow of the distillate,

$$D = \sum_{j=1}^{c} D_j \tag{3.17}$$

$K_{ji}$ — equilibrium constant for component $j$ on plate $i$,
$H_i$ — enthalpy of vapour stream leaving plate $i$,
$h_i$ — enthalpy of liquid stream leaving plate $i$,
$F_{jk}$ — molar flow of $j$th component of feed stream to plate $k$,
$F_k$ — total molar flow of feed stream to plate $k$,

$$F_k = \sum_{j=1}^{c} F_{jk} \tag{3.18}$$

$Q$ — heat input to the reboiler (plate 1) and to the condenser (plate $n$).

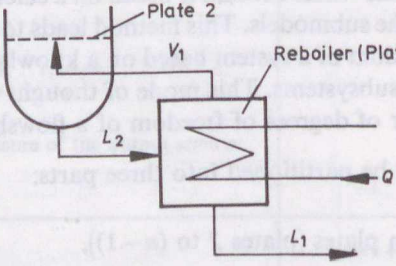



Fig. 3.25. Reboiler section of a distillation column

Table 3.12. A simple model of the reboiler section

| Equation | Number of equations ($n$) | Number of variables ($m$) |
|---|---|---|
| Component balances: | | |
| $L_{2j}=L_{1j}(B)+V_{1j}$ $(j=1, \ldots, c)$ | $c$ | $3c(L_{1j}(B), L_{2j}, V_{1j})$ $(j=1, \ldots, c)$ |
| Equilibrium equations: | | |
| $\frac{V_{1j}}{\Sigma V_{1j}}=K_{ij}\frac{L_{1j}}{\Sigma L_{1j}}$ $(j=1, \ldots, c)$ | $c$ | $2(P_1, T_1)$ |
| Enthalpy balance: | | |
| $H_1+h_1=h_2+Q$ | 1 | $3(T_2, P_2, Q)$ |
| Total | $n=2c+1$ | $m=3c+5$ |

Note that enthalpies, $K$ values and total molar flows are not included as variables in the analysis, since they are functions of pressure, temperature, liquid and vapour component flow rates.

The reboiler (see *Fig. 3.25*) can be described by the equations in *Table 3.12*. The degree of freedom of the reboiler is $d=c+4$. The pressure of the reboiler is ordinarily specified, and either the bottom rate ($B$), or the heat input is to be specified by the user. The $L_{2j}$, $T_2$ and $P_2$ values ($c+2$) should be supplied by the main section of the column (ordinary plates).

Let $k$ be the feed plate of the column (specified). The ordinary plate section (see *Fig. 3.26*) can be described by the following equations *(Table 3.13)*. The degree of freedom of this section of the column is $d=3c+n+4$. The pressure of each plate and the feed stream must be specified ($n-2+c+2=n+c$ specifications), and the remaining $2(c+2)$ values come from the solution of the reboiler and condenser models (the vapour stream of the reboiler and the liquid refluxing from the condenser).

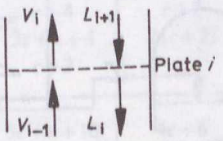



Fig. 3.26. Equilibrium plate of a distillation column

Table 3.13. A simple model of the ordinary plate section of a distillation column

| Equations | Number of equations ($n$) | Number of variables ($m$) |
|---|---|---|
| Component balances: | | |
| $L_{i+1,j}+V_{i-1,j}=L_{i,j}+V_{i,j}$ $(i=2,\ldots,n-1)$ $(i\neq k)$ $(j=1,\ldots,c)$ | $(n-3)\,c$ | $2c\,(n-3)\,(L_{ij}, V_{ij}, \quad j=1,\ldots,c)$ $+2c\,(L_{nj}, V_{nj}, \quad j=1,\ldots,c)$ |
| For the feed plate: | | |
| $F_{kj}+L_{k+1,j}+V_{k-1,j}=L_{kj}+V_{kj} \quad (i=k)\,c$ | | $3c\,(F_{kj}, L_{kj}, V_{kj}, \quad j=1,\ldots,c)$ |
| Equilibrium equations: | | |
| $\frac{V_{ij}}{\Sigma V_{ij}}=K_{ij}\frac{L_{ij}}{L_{ij}}$ $(i=2,\ldots,n-1)$ $(j=1,\ldots,c)$ | $(n-2)\,c$ | $2\,(n-2)\,(P_i, T_i, \quad 1=2,\ldots,n-1)$ |
| Enthalpy balance for ordinary plates: | | |
| $H_{i-1}+h_{i+1}=H_i+h_i$ $(i=2,\ldots,n-1)$ $(i\neq k)$ | $(n-3)$ | $4\,(T_1, T_n, P_1, P_n)$ |
| Enthalpy balance for the feed plate: | | |
| $H_k+h_k=h\,(\text{feed})+H_{k-1}+h_{k+1}$ | 1 | $2\,[T\,(\text{feed}),\ P\,(\text{feed})]$ |
| Total | $n=2cn-4c+$ $+n-2$ | $m=2cn-c+2n+2$ |

For the condenser we assume that the distillate is gained as a liquid at its ebullition point. The condenser (total condenser, see *Fig. 3.27*) can be described by the equations in *Table 3.14*. The degree of freedom of the condenser is $d=c+2$. The pressure of the condenser must be specified. The vapour entering the condenser represents $c$ variables, resulting from the solution of the 'ordinary plate section' of the column. There is an excess degree of freedom, which is a decision variable of the column:

— we can specify the distillate rate, $D$, or

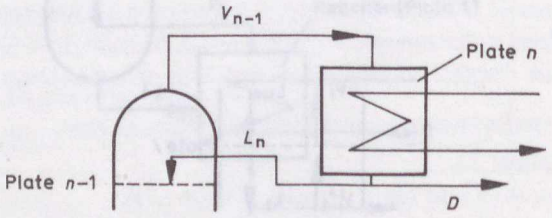



Fig. 3.27. Total condenser

Table 3.14. A simple model of the total condenser

| Equations | Number of equations ($n$) | Number of variables ($m$) |
|---|---|---|
| Component balances: $V_{n-1,j}=L_{n,j}+D_j \quad (j=1, \ldots, c)$ | $c$ | $3c(V_{n-1,j}, L_{nj}, D_j)$ |
| Composition function: $\frac{V_{nj}}{\sum_j V_{nj}}=\frac{L_{nj}}{\sum_j L_{nj}} \quad (j=1, \ldots, c)$ | $(c-1)$ | |
| Ebullition point function: $\sum_j K_{nj}(T_n, P_n)\frac{L_{n,j}}{\sum_j L_{n,j}}=1$ | 1 | $2(T_n, P_n)$ |
| Total | $n=2c$ | $m=3c+2$ |

— alternatively we can choose the reflux rate ($L/D$) as the decision variable to be specified, or

— we can specify the component flow of some arbitrarily chosen key component in the distillate, etc.

Up to this point we have developed expressions for the number of degrees of freedom for the three main parts (subsystems) of a simple distillation column model. The results are collected in *Table 3.15*.

The result is that the number of degrees of freedom of this model is $c+n+4$. To specify the problem unambiguously, we have to specify the pressure over each plate ($n$ specifications), the feed of the column ($c+2$ specifications), and the remaining two variables to be specified are the heat input to the reboiler and one decision

Table 3.15. The degree of freedom analysis of a simple distillation column model

| Section | *D. F.* | Common variables | Remaining *D. F.* |
|---|---|---|---|
| Reboiler | $c+4$ | $c+2$ | 2 |
| Ordinary plates | $3c+n+4$ | $2(c+2)$ | $n-2+c+2$ |
| Condenser | $c+2$ | $c$ | 2 |
| Total | $5c+n+10$ | $4c+6$ | $c+n+4$ |

variable of the total condenser (reflux ratio, distillate rate, distillate key component rate, etc.).

How did we obtain this result? The procedure was quite simple: We counted up the number of D.F. of the subsystems, and then reduced this sum by the number of the common variables of the subsystems. This method will be applied in calculating the number of degrees of freedom for the whole flowsheet.

*Degrees of freedom of a flowsheet*

The calculation of the number of degrees of freedom of a flowsheet is a straightforward task. It can either be done by simply counting all the unknown variables and all the independent equations of the models describing the flowsheet (the 'direct' method), or we can use the result of the degrees of freedom analysis of the unit models. If we follow this latter method, we have to take account of the fact that some variables are common in two models: the stream variables of the interconnecting streams ($c+2$ for each). Therefore, if we count the number of degrees of freedom of all unit models, we count these variables twice. The remedy is obvious: we have to reduce the total number of degrees of freedom by the number of variables counted twice. Let $k$ be the number of interconnecting streams and $u$ the number of units.

Then the number of degrees of freedom of the flowsheet (the plant to be modelled) is

$$D=\sum_{i=1}^{u} d_i - k(c+2) \tag{3.19}$$

This procedure will be illustrated by an example *(Fig. 3.28)*.

The above flowsheet can be transformed to a figure showing the number of degrees of freedom of the unit models (as we have outlined above) inside the circle

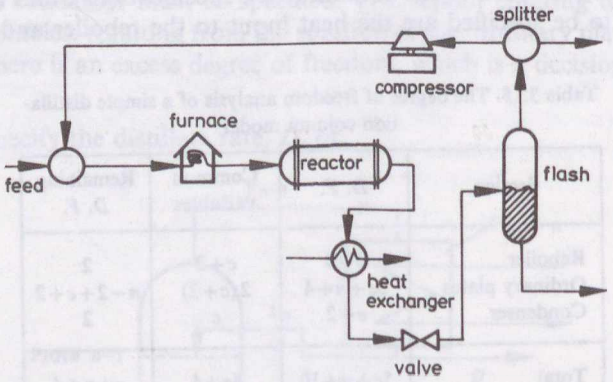



Fig. 3.28. Example demonstrating the degrees of freedom analysis of a flowsheet

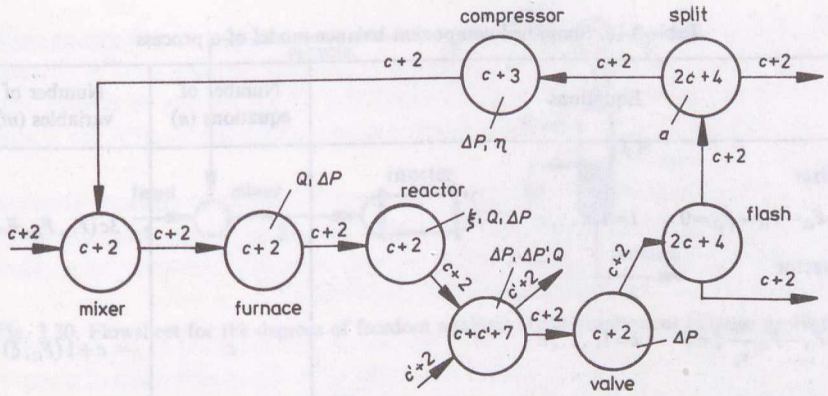



Fig. 3.29. Degrees of freedom analysis of the flowsheet depicted in *Fig. 3.28*.

representing the units, and their unit parameters beside the circles. The streams are 'labelled' with the number of independent stream variables *(Fig. 3.29)*. There are eight interconnecting streams in the figure. Therefore, we have to reduce the total number of degrees of freedom by $8(c-2)$:

$$d=2(c+2)+c+4+c+5+c+c'+7+c+3+c+2+c+3+c+4-8(c+2)=$$
$$=c+c'+16$$

The unit models were counted in the sequence

mixer→furnace→reactor→heat exchanger→valve→separator→
→stream splitter→compressor.

This result can be interpreted in a very simple way: If we sum the unit (equipment) parameters of all models of the flowsheet, and add this number to the number of stream variables of the feed streams of the process, we get the total number of degrees of freedom. In our case we have $c+2+c'+2=c+c'+4$ feed stream variables to the process, and 12 unit parameters.

The result of the analysis depends to a great extent on the models. If we aim to model only the component mass balances, and the heat balances and pressure equations are neglected, the models and even the flowsheet will change. The units that change only the enthalpy and pressure of the streams (heat exchanger, furnace, valve, compressor) will disappear from the flowsheet.

The flowsheet in *Fig. 3.28* will be transformed to the flowsheet depicted in *Fig. 3.30*.

The component balance equations of the above flowsheet are collected in *Table 3.16*.

Table 3.16. Simplified component balance model of a process

| Equations | Number of equations ($n$) | Number of variables ($m$) |
|---|---|---|
| Mixer | | |
| $F_{i2}-F_{i1}-F_{i7}=0 \quad i=1, \ldots, c$ | $c$ | $3c\,(F_{i1}, F_{i2}, F_{i7})$ |
| Reactor | | |
| $F_{i3}-F_{i2}\frac{\nu_i}{\nu_k}\xi=0 \quad i=1, \ldots, c$ | $c$ | $c+1\,(F_{i3}, \xi)$ |
| Separator | | |
| Component balance: | | |
| $F_{i3}-F_{i4}-F_{i5}=0 \quad i=1, \ldots, c$ | $c$ | $2c\,(F_{i4}, F_{i5})$ |
| Performance equation: | | |
| $x_{i5}=K_i\,x_{i4} \quad i=1, \ldots, c$ | $c$ | $2\,(T, P)$ |
| where | | |
| $x_{ij}=\frac{F_{ij}}{\sum_{i=1}^{c} F_{ij}} \quad i=1, \ldots, c;\ j=4,\ 5$ | | |
| and | | |
| $K_i=K_i\,(x_{14}, \ldots, x_{c4},\ x_{15}, \ldots x_{c5},\ T, P)$ | | |
| Splitter | | |
| $F_{i6}-aF_{i5}=0 \quad i=1, \ldots, c$ | $c$ | $c+1\,(F_{i6},\ a)$ |
| $F_{i7}-(1-a)\,F_{i5}-0 \quad i=1, \ldots, c$ | $c$ | |
| Total | $n=6c$ | $m=7c+4$ |

Hence, the degree of freedom is $d=c+4$, which is 'ideally' absorbed giving the $c$ feed component flow rates, and the unit model parameters ($\xi$, $T$, $P$, $a$). This is typically the case in solving the 'performance' or rating problems. If, however, the aim of the calculation is to design a process, we sometimes do not know some of the unit parameters, or, instead of the feed stream variables, the product flow rate is fixed. It may be rather taxing to define the problem correctly and unambiguously. This problem is discussed in the next section.

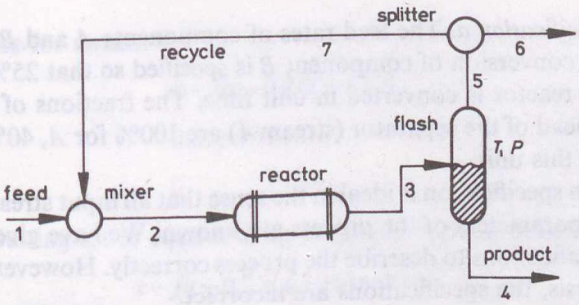



Fig. 3.30. Flowsheet for the degrees of freedom analysis of the component balance problem

### 3.3.2 How to define the flowsheeting problem

After analyzing the degrees of freedom of the flowsheet, we have to assign as many values as the number of degrees of freedom indicates. These specifications must be independent of each other. If some inconsistency exists, i.e. the specifications are not independent, the flowsheeting problem is termed 'badly posed' (Perkins [22]). Since badly posed problems with inconsistent specifications have a functionally singular Jacobian, this type of badly posedness can be identified by investigating the Jacobian matrix of the problem. (As we shall see in Section 3.3.3 all iteration methods use the Jacobian of the equation set to be solved, and problems with a singular Jacobian cannot be solved.)

The procedure is illustrated by the example in *Fig. 3.31* (Perkins [22]).

The flowsheet above consists of three units: a mixer, a reactor and a multicomponent split unit. There are three components in the system: *A*, *B*, and *C*.

One reaction takes place in the reactor:

$A+B \rightarrow C.$

For the sake of simplicity we model only the component balances of the system (each stream can be characterized by three variables).

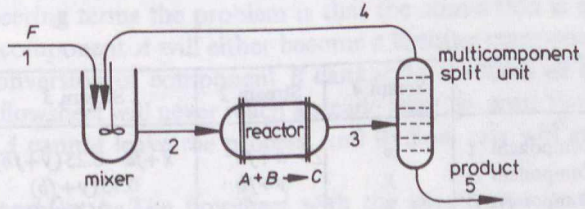



Fig. 3.31. Simple example to illustrate badly posed flowsheeting problems

*Problem specification a.* The feed rates of components *A* and *B* are specified as *fa* and *fb*. The conversion of component *B* is specified so that 25% of component *B* entering the reactor is converted in unit time. The fractions of stream *3* which appear in the head of the separator (stream *4*) are 100% for *A*, 40% for *B* and 0% for *C* entering this unit.

This problem specification is ideal in the sense that all input stream variables and all equipment parameters of the process are known. We have given the necessary number of specifications to describe the process correctly. However, as we shall see from the analysis, the specifications are incorrect.

*Problem specification b.* The molar feed rate of *B* is specified as *fb*, but the molar flow of *A* in this stream (indicated as *a*) is unknown. The conversion of component *B* in the reactor (called *b*) is also unknown. The split fractions of the separator (stream *4*/stream *3*) are:

for component *A*: 95%
for component *B*: 40%
for component *C*: 0%

Since *a* and *b* are not specified, we have to give two other specifications. Let the flows of *B* and *C* in the product stream be known: let them be *pb* and *pc*, respectively.

This problem specification is not ideal since one of the feed stream variables and one of the equipment parameters are not known. This is a typical design problem. It would be difficult to solve this problem with a sequential modular simulation package.

Supposing that stream *4* is torn in both problems, we shall write down the necessary equations to be converged simultaneously, and work out the Jacobian for each set of equations.

*Solution to problem a.* Denote the initial value of the tear stream variables (that of stream *4*) by *x*, *y*, and *z*, respectively.

The flowsheet with the specifications of problem *a* is depicted in *Fig. 3.32.*

We now write down the values of the stream variables of streams *2*, *3* and *4* in the first iteration step:

| | Sream *4* | Stream *2* | Stream *3* |
|---|---|---|---|
| Component *A* | $x$ | $x+fa$ | $x+fa-0.25(y+fb)$ |
| Component *B* | $y$ | $y+fb$ | $0.75(y+fb)$ |
| Component *C* | $z$ | $z$ | $z+0.25(y+fb)$ |

Calculated values for stream 4

$$x+fa-0.25(y+fb)$$

$$0.3(y+fb)$$

$$0$$

Hence the resulting set of equations to be solved (by any iteration methods) is

$$x-fa-0.25(y+fb)-x=0 \qquad \text{Eq.(1)}$$

$$0.3(y+fb)-y=0 \qquad \text{Eq.(2)}$$

$$0-z=0 \qquad \text{Eq.(3)}$$

Calculate the Jacobian for the above set of equations by differentiating the equations with respect to $x$, $y$, and $z$:

$$\mathbf{J}=\begin{array}{c} \\ \text{Eq.(1)} \\ \text{Eq.(2)} \\ \text{Eq.(3)} \end{array}\begin{array}{c} \begin{array}{ccc} (df/dx) & (df/dy) & (df/dz) \end{array} \\ \begin{bmatrix} 0 & -0.25 & 0 \\ 0 & -0.7 & 0 \\ 0 & & -1 \end{bmatrix} \end{array}$$

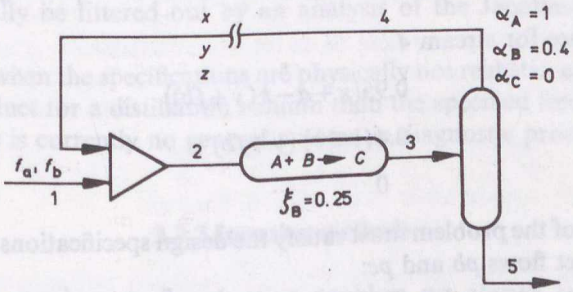



Fig. 3.32. Sample flowsheet with ideal but inconsistent specifications

This Jacobian is singular, because the equation set is independent of $x$. In chemical engineering terms the problem is that the conversion is related to component $B$, and component $A$ will either become a limiting component (in this case the specified conversion of component $B$ cannot be reached) or the process described by this flowsheet will never reach a steady state, because the unreacted part of component $A$ cannot leave the process, and its flow rate will grow to infinity.

*Solution to problem b.* The flowsheet with the specifications of problem $b$ is depicted in *Fig. 3.33.*

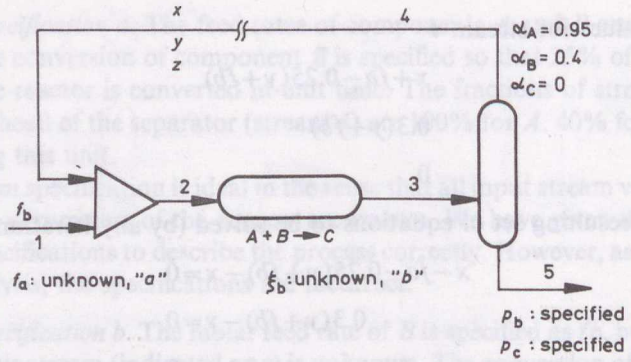



Fig. 3.33. Sample flowsheet with nonideal, inconsistent specifications

The values of the stream variables of streams *2*, *3*, *4* and *5* in the first iteration step are

| | Stream *4* | Stream *2* | Stream *3* | Stream *5* |
|---|---|---|---|---|
| Component *A* | $x$ | $x+a$ | $x+a-b(y+fb)$ | $0.05(x+a-b(y+fb))$ |
| Component *B* | $y$ | $y+fb$ | $(1-b)(y+fb)$ | $0.6(1-b)(y+fb)$ |
| Component *C* | $z$ | $z$ | $z+b(y+fb)$ | $z+b(y+fb)$ |

Calculated values for stream *4*

$$0.95(x+a-b(y+fb))$$
$$0.4(1-b)(y+fb)$$
$$0$$

The solution of the problem must satisfy the design specifications imposed by the specified product flows *pb* and *pc*:

$$0.6(1-b)(y+fb)-pb=0$$
$$z-b(y+fb)-pc=0$$

The resulting set of equations to be solved (by any iteration method) is:

$$0.95(x+a-b(y+fb))-x=0 \qquad \text{Eq.(1)}$$
$$0.4(1-b)(y+fb)-y=0 \qquad \text{Eq.(2)}$$
$$0-z=0 \qquad \text{Eq.(3)}$$
$$0.6(1-b)(y+fb)-pb=0 \qquad \text{Eq.(4)}$$
$$z-b(y+fb)-pc=0 \qquad \text{Eq.(5)}$$

Although we have 5 equations in 5 unknowns ($x, y, z, a, b$), Eqs (2) to (5) contain only three of them, $b$, $y$, and $z$. These are 4 equations in 3 unknowns: there must be some contradiction.

This fact is well indicated by the structure of the Jacobian:

$$\mathbf{J}= \begin{array}{c} \\ \text{Eq.(1)} \\ \text{Eq.(2)} \\ \text{Eq.(3)} \\ \text{Eq.(4)} \\ \text{Eq.(5)} \end{array} \begin{array}{c} \begin{array}{ccccc} (df/db) & (df/da) & (df/dx) & (df/dy) & (df/dz) \end{array} \\ \begin{bmatrix} -0.95(y+fb) & 0.95 & -0.5 & -0.95b & 0 \\ -0.4(y+fb) & 0 & 0 & -0(b-0.6) & 0 \\ 0 & 0 & 0 & 0 & -1 \\ -0.6(y+fb) & 0 & 0 & 0.6(1-b) & 0 \\ y+fb & 0 & 0 & b & 1 \end{bmatrix} \end{array}.$$

Columns 2 and 3 are linearly dependent, and therefore the Jacobian above is singular. The specifications were not independent of each other.

Perkins [22] emphasizes that there are three types of badly posed problems:

— trivial cases, when the number of specifications is not in accordance with the number of degrees of freedom,

— nontrivial cases, when the specifications are not independent. These cases can automatically be filtered out by an analysis of the Jacobian, as described above,

— and cases when the specifications are physically not realistic; e.g. if we specify more head-product for a distillation column than the specified feed rate. For this latter case there is currently no general automatic diagnostic procedure.

### 3.3.3 Iteration methods

To solve the steady-state flowsheeting problem we always have to solve a generally nonlinear equation set:

$$\underline{F}(\underline{x})=\underline{x} \tag{3.20}$$

$$\underline{f}(\underline{x})=\underline{F}(\underline{x})-\underline{x}=\underline{0} \tag{3.21}$$

The set of equations may be the complete description of the flowsheet (simultaneous, or equation oriented approach), or only the set of equations corresponding to the torn (stream) variables. The aim is the same: after assigning some initial value to the variables, we would like to approach the solution in a finite number of steps with a pre-specified tolerance, say *eps*.

$$\underline{F}(\underline{x})-\underline{x}<\underline{eps} \tag{3.22}$$

The solution procedure is terminated if the tolerance criterion (called the convergence criterion in some papers) is fulfilled. If not, then the difference or 'error vector' is used in some way to modify $x$ for the next iteration step.

The most frequently used convergence criteria are the following:

*Relative error*

The absolute values of the error ratios, $r$, are tested against *eps*:

$$r_i = \frac{y_i^k - y_i^{k-1}}{y_i^m} < eps \tag{3.23}$$

where $y_i^k$ is the value of the element $q_i$ after the $k$th iteration,
$y_i^m$ is the minimum of $y_i^k$ and $y_i^{k-1}$.

The relative error criterion is commonly used when component flows or concentrations are being determined.

*Absolute error*

In this case the absolute errors, $a_i$, are tested as:

$$a_i = |y_i^k - y_i^{k-1}| < eps \tag{3.24}$$

This criterion is commonly used when intensive variables (e.g. temperature) are being tested.

*Root mean square error*

The root mean square error criterion is defined as:

$$s = \frac{1}{n}\sqrt{\sum_{i=1}^{n} \frac{y_i^k - y_i^{k-1}}{y_i^k}} < eps \tag{3.25}$$

*Direct substitution*

This procedure is alternatively known as successive substitution, direct resubstitution or repeated substitution.

The idea of direct substitution can be well illustrated by its application to a simple flowsheeting problem. The flowsheet has only one simple recirculation loop, and the tear stream is that indicated in *Fig. 3.34*. Here we have introduced a new unit, the 'convergence unit'. This new unit represents the rule for the calculation of $x$ based on the result(s) of the previous iteration step(s).

In *Fig. 3.34* the initial value of the torn stream variables is indicated by $\underline{x}$, and

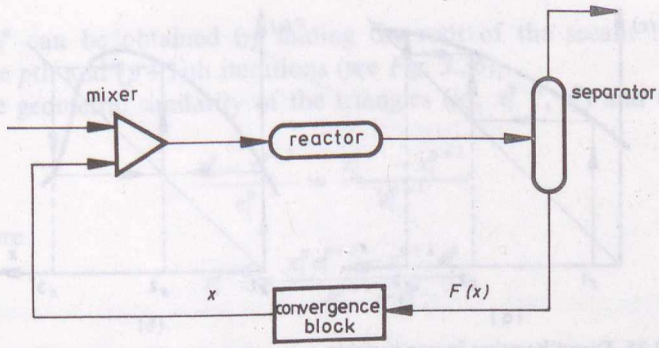



Fig. 3.34. Simple example to illustrate the idea of direct substitution

the value of the torn stream variables at the end of the iteration step by $\underline{F}(\underline{x})$. This notation also indicates that $\underline{F}(\underline{x})$ is a function of the initial value, $\underline{x}$.

Denote the estimated initial value at the $p$th iteration step by $\underline{x}^p$, and the value of the calculated stream variables by $\underline{F}(\underline{x}^p)$.

The rule for using direct iteration inside the convergence unit can be written as:

$$\underline{x}^{p+1} = \underline{F}(\underline{x}^p) \tag{3.26}$$

According, to the above rule, if the convergence criterion is not met, each element in the new initial value vector $\underline{x}^{p+1}$ will be equal to the value of the corresponding element of $\underline{F}(\underline{x}^p)$.

Let us examine how much nearer we can get to the solution in one iteration step.

The exact solution, $\underline{x}^*$ is characterized by the relation $\underline{x}^* = \underline{F}(\underline{x}^*)$.

First we write down the Taylor series expansion about the solution:

$$\underline{x}^{p+1} - \underline{x}^* = \underline{F}(\underline{x}^p) - \underline{F}(\underline{x}^*)\ \underline{F}(\underline{x}^p) - (\underline{F}(\underline{x}^p)) + \sum_j \frac{\partial F}{\partial x_j}(x_j^* - x_j^p) = \tag{3.26a}$$

$$= -\mathbf{J}(\underline{x}^* - \underline{x}^p) = \mathbf{J}(\underline{x}^p - \underline{x}^*)$$

In the above equation $\mathbf{J}$ is the Jacobian matrix, the elements of which are the partial derivatives of $\underline{F}(\underline{x})$ with respect to all elements of $\underline{x}$ point $\underline{x}^p$:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_1} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} & \cdots & \frac{\partial F_2}{\partial x_2} \\ \vdots & & & \vdots \\ \frac{\partial F_k}{\partial x_1} & \frac{\partial F_k}{\partial x_2} & & \frac{\partial F_k}{\partial x_k} \end{bmatrix}_{\underline{x}^p} \tag{3.27}$$

where $k$ is the dimension of $\underline{x}$.

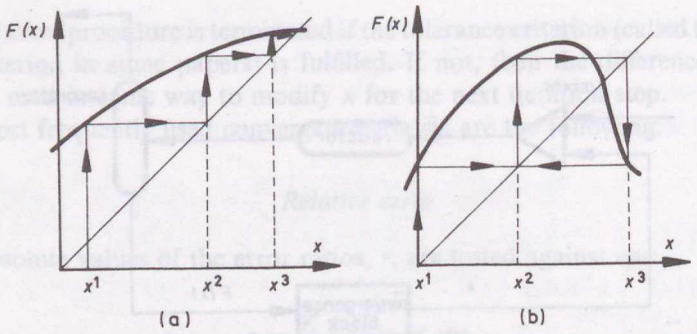



Fig. 3.35. Direct iteration in one variable: (a) convergent, (b) nonconvergent case

The main conclusion of the above result is that the approach to the solution in step $p+1$ is proportional to the approach in step $p$. This is called first order convergence. The proportionality 'constant' is the Jacobian. If the Jacobian matrix is equal to the unit matrix, E, then the direct substitution cannot converge. In effect, the Jacobian is constant only in case of linear equation sets. In nonlinear cases the values of its elements are largely dependent on $\underline{x}$.

For the one variable case the conclusions can be illustrated with the following figures *(Fig. 3.35)*.

*Notes*: The calculation of the Jacobian is done in most cases by numerical methods. It is a time-consuming task, but if we use only direct substitution, there is no need to calculate the Jacobian.

Direct substitution usually converges for mass balance problems. Heat balances may sometimes cause problems. According to the experience of most users, it is 'robust' (i. e. it converges for a vast majority of problems, no matter how crude the initial guesses), but slows down dramatically as the solution is approached.

### *Secant (Wegstein) method*

This method was first proposed for chemical engineering calculations by Wegstein [23], hence its name.

The aim of the method is to accelerate direct substitution by using not only the results of the last iteration step, but also those of the two previous steps.

Let $\underline{y}=\underline{F}(\underline{x}^p)$ be the calculated value after iteration step $p$, and let $\underline{e}^p$ be the error in the last direct iteration step.

$$\underline{e}^p=\underline{y}^p-\underline{x}^p=\underline{y}^p-\underline{y}^{p-1} \tag{3.28}$$

Chose $\underline{x}^{p+1}=\underline{y}^p$. Then

$$\underline{e}^{p+1}=\underline{y}^{p+1}-\underline{x}^{p+1}=\underline{y}^p-\underline{y}. \tag{3.29}$$

If we now consider the effect on a single variable in $\underline{x}$ (say $x_i$), a good esti-

mate for $x_i^*$ can be obtained by finding the root of the secant between the errors at the $p$th and $(p+1)$th iterations (see *Fig. 3.36*).

From the geometric similarity of the triangles $(x_i^p, x_i^{p+2}, e_i^p)$ and $(x_i^{p+1}, x_i^{p+2}, e_i^{p+1})$

$$\frac{x_i^p - x_i^{p+2}}{e_i^p} = \frac{x_i^{p+1} - x_i^{p+2}}{e_i^{p+1}} \tag{3.30}$$

and therefore

$$x_i^{p+2} = \frac{x_i^p e_i^{p+1} - x^{p+1} e_i^p}{e_i^{p+1} - e_i^p} \tag{3.31}$$

According to the above result, we need two direct iteration steps, we have to calculate the errors, and this information is used to give the new initial guess of the variables to be iterated.

This method usually accelerates the iteration, especially if $y_i$ depends mainly on $x_i$ and not strongly on the other $x_j$ variables $(j \neq i)$. In the unlikely event that $y_i = F(x_i)$ only, (i. e. the problem can be decomposed, and the Jacobian is diagonal), it can be shown [22] that

$$x_i^{p+1} - x_i^* \leqq K(x_i^p - x_i)^2, \tag{3.32}$$

i. e. the convergence is of second order.

If the system is not only decoupled, but also linear, a single secant step should give the exact solution.

In the literature the method has been applied in a different form. Introducing

$$s = \frac{e_i^p}{e_i^{p-1}} \tag{3.33}$$

it can be shown that Eq. (3.27) can be transformed to the following form:

$$x_i^{p+1} = x_i^p + \frac{s}{1-s}(x_i^p - x_i^{p-1}) \tag{3.34}$$

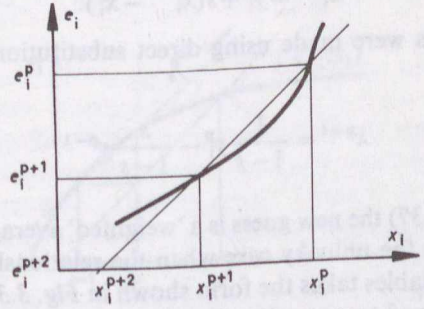



Fig. 3.36. Illustration of the secant method in a one variable

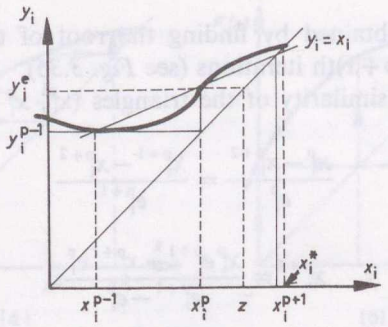



Fig. 3.37. Wegstein's method

Using this form, Wegstein's method can be interpreted as shown in *Fig. 3.37*. In *Fig. 3.37* $z$ indicates the next $x_i$ value obtained by direct substitution. The new value of $x_i$ guessed by Wegstein's method, $x_i^{p+1}$, is closer to the solution, $x$. To obtain this value, we approximated the function $y_i = F(x_i)$ by a straight line defined by the last two iterated points. The point is the intersection of the line defined by these two points and the line $y_i = x_i$. The new guess is the projection of this intersection point on the $x$ axis.

It is easy to show that the slope of the dotted line is just the same as the one denoted by $s$.

$$\frac{y_i^p - y_i^{p-1}}{x^p - x^{p-1}} = \frac{y_i^p - y^{p-1}}{y^{p-1} - y^{p-2}} = \frac{e_i^p}{e^{p-1}} = s \tag{3.35}$$

The new guess has to fulfil both the condition

$$y_i^{p+1} = x_i^{p+1}$$

and

$$y_i^{p+1} = y_i^p + s(x_i^{p+1} - x_i^p)$$

which means that

$$x_i^{p+1} = y_i^p + s(x_i^{p+1} - x_i^p) \tag{3.36}$$

If the last two steps were made using direct substitution, then $x_i^p = y_i^{p-1}$, and therefore

$$x^{p+1} = \frac{1}{1-s} y_i^p - \frac{s}{1-s} y^{p-1} \tag{3.37}$$

According to Eq. (3.37) the new guess is a 'weighted' average of the two previous $y_i$ values. However, in the unlucky case when the relationship between the computed and guessed variables takes the form shown in *Fig. 3.38*, a bad choice of the initial value of $x$ may lead to a step in which the predicted new value is very much

out. To guard against this possibility, the usual practice is to place some bounds on the acceleration factor.

Equation (3.37) can be written in the form

$$x_i^{p+1} = w y_i^p + (1-w)\, x^p \tag{3.38}$$

where

$$w = \frac{1}{1-s}$$

is the acceleration factor. In order to preserve stability $w$ has to remain in the range $1 < w < 6$.

The lower limit corresponds to direct iteration.

A very simple and widely accepted rule of thumb is to apply Wegstein's acceleration after each 3 to 5 direct substitution steps.

*Newton's method*

This is a widely used and basic method for the solution of nonlinear equation sets of the form

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_N) &= 0 \\ f_2(x_1, x_2, \ldots, x_N) &= 0 \\ &\vdots \\ f_N(x_1, x_2, \ldots, x_N) &= 0 \end{aligned}$$

or concisely, with vector notation

$$\underline{f}(\underline{x}) = \underline{0}$$

Since $\underline{f}(\underline{x})$ is nonlinear, there is no general analytical procedure for expressing the unknown variables in an explicit form. Therefore an iterative procedure has to be constructed which, starting from an initial guess $\underline{x}^1$ for the solution, generates a sequence of iteration $\underline{x}^1, \underline{x}^2, \ldots$ which may converge to the solution.

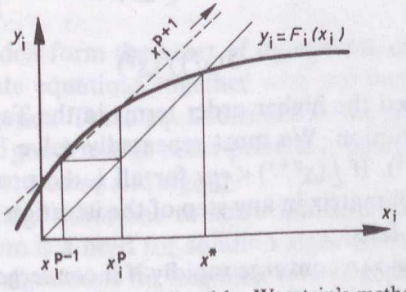



Fig. 3.38. The new value predicted by Wegstein's method is far out

The question is that, knowing a guessed value at the $p$th iteration, how do we obtain the next guess, $\underline{x}^{p+1}$, so that it is closer to the solution. Let $\underline{z}^p$ be the step from $\underline{x}^p$ to $\underline{x}^{p+1}$ i. e.

$$\underline{x}^{p+1} = \underline{x}^p + \underline{z}^p \tag{3.39}$$

$\underline{f}(\underline{x}^{p+1})$ can be expanded in a Taylor series about $\underline{f}(\underline{x}^p)$. For the sake of simplicity we present the idea for $N=2$ (i. e. two equations in two unknowns).

Then Eq. (3.39) becomes

$$\underline{x}^{p+1} = \begin{bmatrix} x_1^{p+1} \\ x_2^{p+1} \end{bmatrix} = \begin{bmatrix} x_1^p + z_1^p \\ x_2^p + z_2^p \end{bmatrix}$$

and

$$\underline{f}(\underline{x}^{p+1}) = \begin{pmatrix} f_1(x_1^{p+1}, x_2^{p+1}) \\ f_2(x_1^{p+1}, x_2^{p+1}) \end{pmatrix} = \begin{pmatrix} f_1(x_1^p + z_1^p, x_2^p + z_2^p) \\ f_2(x_1^p + z_1^p, x_2^p + z_2^p) \end{pmatrix}$$

Performing the Taylor series expansion for the two vector components independently, we get

$$f_1(x_1^p + z_1^p, x_2^p + z_2^p) = f_1(x_1^p, x_2^p) + \frac{\partial f_1}{\partial x_1}(x_1^p, x_2^p)\, z_1^p + \frac{\partial f_1}{\partial x_2}(x_1^p, x_2^p)\, z^p + \text{higher-order terms}$$

$$f_2(x_1^p + z_1^p, x_2^p + z_2^p) = f_2(x_1^p, x_2^p) + \frac{\partial f_2}{\partial x_1}(x_1^p, x_2^p) z_1^p + \frac{\partial f_2}{\partial x_2}(x_1^p, x_2^p) z_2^p + \text{higher-order terms}$$

To simplify the above equations we switch over to vector notation using the Jacobian, as introduced previously:

$$\underline{f}(\underline{x}^{p+1}) = \underline{f}(\underline{x}^p + \underline{z}^p) = \underline{f}(\underline{x}^p) + \mathbf{J}(\underline{x}^p)\underline{z}^p + \text{higher-order terms}$$

Neglecting the higher-order terms and making use of the fact that $\underline{x}^{p+1}$ should be the solution [i. e. $\underline{f}(\underline{x}^{p+1}) = \underline{0}$] we may derive a formula for $\underline{z}^p$:

$$\mathbf{J}(\underline{x}^p)\,\underline{z}^p = -\underline{f}(\underline{x}^p) \tag{3.40}$$

and therefore

$$\underline{z}^p = -\mathbf{J}(\underline{x}^p)\underline{f}(\underline{x}^p) \tag{3.41}$$

Since we have neglected the higher-order terms in the Taylor series expansion, $\underline{x}^{p+1}$ may not be the solution. We must repeatedly solve Eq. (3.41) for $\underline{z}^p$, and evaluate function $\underline{f}(\underline{x}^{p+1})$. If $\underline{f}(\underline{x}^{p+1}) < \underline{eps}$ for all $i$, the procedure is terminated. Naturally, if the Jacobian matrix in any step of the iteration becomes singular, the solution of the problem fails.

Newton's method is said to converge rapidly if it converges at all (Perkins [22]). In the vicinity of the method is of second order.

The most cumbersome step of the method is the repeated computation of the Jacobian matrix. If, for some reason explicit expressions for the functions $f(x)$ are not available (e. g. some subsystems are available in the form of procedures in sequential-modular flowsheeting systems) we have to use finite difference approximations for $\mathbf{J}(x)$.

$$\frac{\partial f_i}{\partial x_j} = \frac{f_i(x_1, x_2, \ldots, x_j + h, \ldots, x_N) - f(x_1, x_2, \ldots, x_N)}{h}$$

If $h$ is small enough, the approximation of the derivative is good. However, to compute the Jacobian we need to evaluate the function $N$ times. The method can be improved by using some skill to generate the Jacobian with fewer function evaluations (see e. g. Broyden [24]).

### 3.3.4 Approaches to flowsheeting

Flowsheeting systems are generally accepted software tools for solving flowsheeting problems, i. e. for solving steady-state heat and mass balancing, sizing and costing problems. These are computer programs which are capable of:

— asking the user for the correct definition of the (steady state) flowsheeting problem;
— interpreting the data, checking the input data for correctness, analyzing the problem;
— solving the problem; and
— giving the calculated results in a standard or user-defined format.

The teams developing flowsheeting systems intend to make their system as general as possible: capable of handling a wide range of materials, capable of calculating as many physical and thermodynamic properties as possible, and of handling very large and complicated flowsheets comprising a great variety of unit operations.

According to Evans [18] the basic building blocks of a flowsheeting system are as follows *(Fig. 3.39)*:

(*a*) *Models.* The models form the heart of the system, containing the conservation, transport and rate equations together with the methods of evaluating the desired physical properties. These equations are in the form of algebraic or differential equations. A good model must present a balance in respect of detail, validity, generality, correctness and rigour.

(*b*) *Algorithm.* The algorithm has to solve the equation set generated by the models. Therefore, there is a need for solution algorithms for both algebraic and differential equations, algorithms for handling very large and sometimes sparse set of equations and, in some cases, nonlinear programming methods.

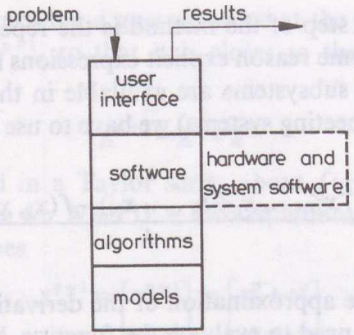



Fig. 3.39. The structure of flowsheeting systems (Evans [18])

The good algorithm is robust, clear-cut, elegant, as general as possible, and effective is terms of computing time and memory.

(*c*) *Software.* The software implements the above algorithm within the computational environment characterized by the actual hardware and systems software. Important parts of it are the system-architecture, the data structure, the programming language(s), in which the program(s) were written, the program itself in object and machine code, and the necessary documentations to install, use and develop the system.

Good software is easy to read, maintain, modify and is portable from one machine to another as possible.

(*d*) *User interface.* Most flowsheeting systems use some specially developed input language for defining the flowsheeting problem correctly. This input language, together with a report writer for presenting the results, and the user documentation may be regarded as the user interface.

A good user interface accepts the problem definition in the most natural way, and presents the results in the most easily interpretable manner in the desired (and sometimes user-defined) detail.

A good system gives the user the impression that he is working with the models directly themselves. In other words, the users need not know anything about algorithms and software: such parts of the system have to be 'transparent' to them.

Bluck *et al.* [25] gave the following practical criterion of effectiveness of a flowsheeting system:

"It should involve the minimum total expenditure of manpower and computing resources to derive a satisfactory solution to the problem within the timescale dictated by the project."

A good algorithm can contribute greatly to the effectiveness of the system. We now focus our attention on the solution algorihms, and in particular on the solution strategies incorporated into floowsheeting systems, bearing in mind the place and role of algorithms in the structure of systems.

Before beginning to categorize the strategies, the first thing to be said is that our aim is to solve large nonlinear sets of equations resulting in a correct description of the flowsheet in the steady state. The strategies may differ in the philosophy of decomposing the set of equations. If the equations are grouped according to the unit operation models involved, the resulting calculation sequence will be different, as if we decompose the system using only pure mathematical considerations.

According to many authors (Westerberg *et al.* [1], Rosen [26], Evans [18], Perkins [27], Clark and Reklaitis [28]), three basic flowsheeting strategies have been evolved over the last two decades, which appear under different names in the chemical engineering literature:

the sequential–modular (SM),
the equation oriented (EO) or global, and
the simultaneous–modular, or successive linearization or two-tier approach (SL).

Independent of the strategy used to decompose the system, we generally use some numerical (rather than analytical) procedures to find the solution. The numerical algorithms use some iterative procedures (see Section 3.3.3) and, due to the hierarchical nature of the problem, nested iterative loops may arise, generally on three levels:

1. in calculating the physical properties;
2. inside some unit operation models to find their steady state, or inside some equation subsystems to be solved simultaneously; and
3. in the level of recycle calculations.

Here we use the term 'recycle' in its most general meaning, regarding the iteration loops arising from design specifications as 'information recycles'.

Hence, flowsheeting systems using the same (e. g. SM, EO or SL) strategies may differ in the way they order the iteration loops, or in the selection of the variables to be converged in the same iterative cycle.

### The sequential–modular (SM) strategy

This is the most widely used strategy, having been introduced in the early sixties. Most industrial systems use this approach (PACER [29], CHESS [30], FLOWTRAN [31], FLOWPACK [32], PROCESS [33], CONCEPT [34], SIMUL [35], FORSIM [36], UNICORN [37], ASPEN [38], [42] etc.).

According to this strategy the unit operations are represented in the form of procedures or 'modules', which calculate the output stream variables and unit parameters as a function of the input stream variables and specified unit parameters. The procedures are called in a sequence that is determined by the process topology. If there is no recycle, the calculation is straightforward (it proceeds in downstream direction; i. e. if a unit is downstream of another unit in the flowsheet, it comes later in the sequence).

If, however, there is a recycle loop, then an iterative procedure is needed, i. e. we have to 'tear' the stream coming from a unit later in the sequence ('recycle' stream), and then guess the value of the torn stream variables in order to be able to proceed with the calculation. If we reach the 'source' unit of the torn stream, we evaluate the stream variables, and compare their new values with the initial (guessed) ones. If the difference is within a prespecified tolerance, the calculation is finished. If not, the calculated value is used (with the insertion of some appropriate numerical method) to produce a new guess. This approach is efficient if all input stream variables and all necessary unit parameters are known. This case is often called a 'performance', 'rating' or simple 'simulation' problem in the literature. In fact, this type of specification is ideal for simulation calculations with an *SM* simulator (see Section 3.3.2). Potentially, there are also three iteration levels of the simulation problems:

1. in calculating the physical properties;
2. inside some unit operation models; and
3. in the level of recycle calculations

If, however, some of the input stream variables and necessary equipment parameters are not known in advance, and the values of some output stream variables are specified in order to absorb the degrees of freedom, a new iteration level arises: we need to converge the calculated output stream variables to meet the specifications (see *Fig. 3.40*). This type of calculation is called the 'design' problem. One possible way to reduce the number of nested iteration loops is to apply the *EO* strategy to solve this type of flowsheeting problems.

The solution of a flowsheeting problem using the sequential–modular approach involves the identification of the maximum recycle loops, the tearing of each recycle loop, and an appropriate device for determining the hierarchy and sequence of calculations.

In *Fig. 3.40* a flowsheet is depicted, consisting of mixers (*M1, M2*), three reactors (*R*) and distillation columns (*D1, D2*). Distillation column *D1* has to produce a stream, *S5*, which is appropriate as feed to the reactor to give a stream, *S6*, with some prespecified property (e. g. the molar ratio of two components is specified). To meet this specification we have to control some parameters of distillation

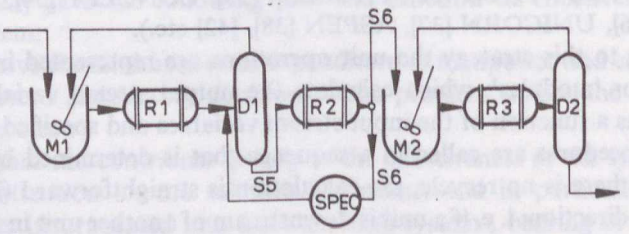



Fig. 3.40. Schematic representation of the sequential–modular strategy [18]

column *D1*. This gives rise to an 'information recycle', denoted by the dotted line in *Fig. 3.40*. The unit SPEC contain the rule according to which the calculated value of *S6* modifies the controlled parameter of column *D1*. From a computational point of view, this information recycle has the same effect as any other recycle 'stream'; i. e. a new iteration cycle will be created, and the first subsystem to be handled as a unit in the iteration level of recycles becomes larger: instead of (*M1, R1, D1*) it now contains (*M1, R1, D1, R2*, SPEC).

The procedure for identifying the maximum number cyclic subsystems which have to be handled together is called partitioning. The mathematical methods applied in partitioning of flowsheets were discussed in Section 3.2. In the above flowsheet there are two maximum cyclic subsystems.

Besides partitioning, the speed of calculation is largely dependent on:

— the choice of variables to be converged at the same iteration level and on the hierarchy of the iteration calculations; and

— on the choice of tear sets, which determines the calculation sequence.

There is no general rule for the organization of nested iterations. According to some authors, the number of iteration levels should possibly be kept to a minimum. Perkins [39] reported successful trials when converging the tear stream variables of recycle streams and design specifications in the same iteration loop. This approach is called a 'design-oriented modular strategy'.

The mathematical methods for determining the optimum tear sets of a maximum recycle loop were discussed in Section 3.2.

The iteration methods most frequently used in flowsheeting systems are:

direct substitution;
secant (Wegstein's) method;
Newton's method and its modifications (e.g. Broyden [24], [40]); and
the dominant eigenvalue method.

Rosen [26] emphasized that the Newton and Newton–Raphson type methods, which evaluate the Jacobian (or its inverse) at every iteration step, are faster in the vicinity of the solution, whereas far away from the solution the burden of re-evaluating the Jacobian gives them a disadvantage. In this region the 'diagonal' procedures (taking account of and evaluating only the main diagonal of the Jacobian) are preferable. Favourable combinations of these two types of methods can be constructed (e.g. Wegstein's method up to the vicinity of a certain 'loose' tolerance, then Broyden's method in the neighbourhood of the solution for final 'tuning').

### The equation oriented (EO) approach

Solving a design problem with a sequential–modular flowsheeting system means using a series of simulation runs to meet the design specifications. This indirect solution of the problem can be made direct by the use of *EO* methods. The *EO* flowsheeting systems contain procedures to generate the equation subset of the unit

operations (instead of giving the corresponding solution algorithms in the form of subroutines or procedures), and the model of the flowsheet as a whole is constructed from the subsystems of the unit operations using the information of the connectivity equations. The generated (nonlinear) system of equations will be analyzed by the use of various structure analysis methods (see Section 3.2) and they will be solved by various numerical methods (Section 3.3.3). Some systems using this approach are, for instance, SPEED-UP [41], QUASILIN [43], and GENDER [44]. There are two facts that must be stressed:

— the unit operation equation subsystems contain more variables than equations. The solution procedure cannot be constructed before specifying the design variables. The same model can be used in many different ways depending on the choice of design variables.

— The large equation system can be decomposed, or it can be solved simultaneously.

The philosophy of generating the system of model equations instead of a pre-determined solution procedure is a common feature of the *EO* simulators, but the practical solution of the problem can be different and varies from simulator to simulator. Usually a variety of the possible methods are combined into one system. This can also be observed in the model-generation phase:

(a) If the equation subset describing the unit also contains the equations to calculate the physical properties, the resulting equation system may become too large. According to some authors ([1], [45]) the number of the necessary physical property equations in such a model is approximately 4 times the number of unit model (performance and conservation) equations. Thus it is not recommended to integrate these physical property correlations into the model equations. A better choice is to write procedures for calculating the necessary physical properties, and then handling this part of the model separately. This philosophy results in a mixed system of equations and procedures. If, however, the flowsheeting system is capable of handling the mixed system of equations and procedures, it may be worthwhile to write as procedures some of the unit operation modules which jeopardize the convergence of calculation. If we write all unit operation modules in the form of procedures, we would arrive at the other end of the scale: the *EO* simulator turns into an *SM* simulator. As Perkins [27] has pointed out, it is highly improbable that the computational optimum would be at either end of the scale. Therefore, some *EO* systems are capable of handling mixed systems of equations and procedures (e.g. SPEED-UP [41]).

(b) Hence, most *EO* simulators use some variation of the quasi-Newton solution method, which needs the evaluation of the Jacobian or its inverse (the matrix of the first derivatives); the systems may differ in the way to evaluate the Jacobian or its inverse (whether analytical or numerical procedures are applied, how the initial value of the Jacobian is guessed, etc.)

(c) There are a number of methods for reducing the size of the equation set to be solved simultaneously. (Some of them were discussed in detail in Section 3.2.)

An interesting way of reducing the size of the equation set, is the so-called automatic output variable assignment. According to this method the size of the equation set to be handled simultaneously is reduced by searching an appropriate 'matching' of equations and variables to minimize the number of variables to be guessed and iterated. In the most fortunate case this method leads to a noniterative procedure, and the equations can each be solved one by one for their only (assigned) output variable.

Systems using this method are reputed to be very flexible but their robustness is problematic. Automatic output variable assignment is built into the widely used MA28 [46] general non-linear equation solving package (SPEED-UP [41], QUASILIN [43], ASCEND-II [47]).

(d) A well-known feature of the equation sets describing chemical engineering systems is their sparsity, i.e. the number of variables in one equation is relatively small; the number of zeros in the Jacobian is large. There are special methods for handling large, sparse equation sets (see Duff [48], [49]). Some systems make use of this sparsity to render the solution procedure more effective.

(e) The flowsheeting (in particular the design) problem can be formulated as an optimization problem, where the (equality) constraints are the model equations, and the argument of the objective function contains the decision variables:

$$\text{maximize } P(\underline{X}, \underline{U})$$

$$\text{with } \underline{F}(\underline{X}, \underline{U}) = 0$$

where $\underline{F}(\underline{X}, \underline{U})$ is the set of model equations,

$\underline{U}$ is the set of independent (decision) variables, e.g. unit parameters and feed stream variables.

$\underline{X}$ is the set of calculated (dependent), variables, e.g. intermediate and product stream variables, calculated equipment parameters, etc.

$P$ is the objective function.

The solution of the flowsheeting problem, formulated as an optimization problem, needs effective nonlinear programming methods. This approach is incorporated in the GENDER [44] flowsheeting system.

There are some objections to *EO* simulators in the literature. First, the great investment of industry in the development of unit operation models for *SM* simulators cannot be used (e.g. Evans [18]). *EO* simulators had been developed in academic institutions, and no industrial systems have been reported up to now. Fitting the standard thermodynamic packages to the *EO* simulators is also a difficult job. The root of both problems is the same unit operation models and thermodynamic property calculations were developed in the form of procedures. However, the problem of handling mixed systems of equations and procedures can be solved (Perkins [27]). Another argument against the global approach is the

enormous memory requirement for storage and handling the whole flowsheet model as a large set of equations. This can be reduced by using sparse matrix technique.

There are some numerical problems associated with this approach. These simulators are not reputed to be very robust. Sometimes there are multiple roots for some variables, which may cause trouble in the numerical procedure. In some cases the Jacobian becomes singular during the solution procedure. The initial guess and the scaling of the variables have a major influence on the convergence behaviour of the problem. And finally, if there is an error, and the calculation fails to converge, it is difficult to diagnose the cause of the failure.

### The simultaneous–modular or successive linearization (SL) method

Some of the disadvantages of the *EO* methods can be overcome by using *SL* methods. Such methods can be regarded as a combination of the *SM* and *EO* simulation strategies: the traditional unit operation models are used in the form of procedures as 'rigorous' models, and simple (most frequently linear) models are used to model the whole flowsheet in an *EO* fashion. The rigorous models are used to estimate the model parameters of the simple models, whereas the equation set generated using the simple models is solved by some efficient equation solving method. The results of the latter type of calculations are the stream variables necessary to evaluate the rigorous models once again. This 'alternating' use of the two levels is illustrated in *Fig. 3.41.*

The equation set generated using the simple models is solved simultaneously, whereas the original modular unit operation models are used in the same way as in the 8M simulators. Hence the name: 'simultaneous–modular'. The simple models can either be linear ones, with coefficients determined by numerical perturbation of the rigorous models, or approximate engineering models.

This approach was first proposed by Rosen [54], with the use of split fraction models as simple models. The simultaneous–modular approach takes advantage of the existing *SM* software, not only in the effectiveness of their calculation, but also by the use of the built-in heuristic procedures in generating initial guesses for the variables to be converged, and in error diagnostics. According to Evans [18], the inside-out algorithm of Boston and Brit [50] can be regarded as an extension of the two-tier approach, because it transfers the time consuming thermophysical calcula-

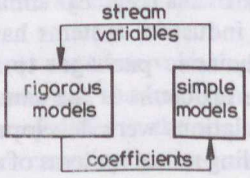



Fig. 3.41. Schematic representation of the simultaneous–modular approach

tions into the outside loop (which is executed less frequently) instead of leaving it in the inside loop. This approach is used e.g. in the MASTEP [51] system; and the ASPEN simulator has also been developed so that it will be easy to implement a simultaneous–modular algorithm in the future (Evans [18]).

Perkins notes that the most common nonlinear equation solving algorithms are always two-tier algorithms, since they generate an approximate linear model at each iteration step. This similarity has also been noted by other authors: Shacham [52] considered the split-fraction models as a possible nonlinear equation solving algorithm, and compared it to other equation solving methods.

### 3.3.5 Some typical flowsheeting problems and their solutions

The sample problems shown below are intended to give some overall impression of flowsheeting systems as generally accepted software tools for solving steady-state heat and mass balancing problems.

**Example 1**

This example was extracted from the Chess User Manual, and solved with CHESS by the author at the Imperial College of Science and Technology on the CDC 7600 Computer of the University of London. CHESS [30] is a typical sequential-modular flowsheeting system, now in use at several universities for demonstrating flowsheeting features.

**Simulation of a simple ethylene plant**

The process flow diagram of the plant is shown in *Fig. 3.42.*

The feed stream of the process is as follows:

| | |
|---|---|
| Pressure (bar) | 40 |
| Vapour fraction | 0 |
| Mole flow rates | (kmol/h) |
| Ethane | 80 |
| Propane | 120 |

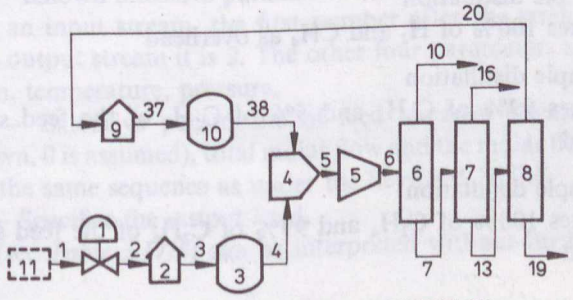



Fig. 3.42. Simple ethylene plant (CHESS 30)

In the system there are the following components:

Hydrogen
Methane
Ethylene
Ethane
Propylene
Propane
*n*-Butane

The plant consists of the following equipment:

Unit 1-valve
Reduces pressure to 6 bar

Units 2&9-Direct fired heaters

| | |
|---|---|
| Raise temperature to | 345K |
| Pressure drop (bar) | 0.7 |
| Heat duty (MJ/h) | 1,500,000 |

Unit 3-Reactor
The reaction taking place in the unit

$$3\,C_2H_6 + 6\,C_3H_8 \rightarrow 4\,H_2 + 4\,CH_4 + 5\,C_2H_4 + 2\,C_3H_6 + C_4H_{10}$$

The conversion of propane is 90%.

Unit 4-Mixer

Unit 5-Pump/Compressor

| | |
|---|---|
| No. of stages | 3 |
| Capacity (MJ/h) | 1,270 |
| Outlet pressure (bar) | 40 |
| Motive power | steam |
| Enthalpy of inlet steam | 2767 MJ/kg |
| Enthalpy of outlet steam | 2767 MJ/kg |

Unit 6-Simple distillation
Takes 100% of $H_2$ and $CH_4$ as overhead

Unit 7-Simple distillation
Takes 99% of $C_2H_4$ and 5% of $C_2H_6$ of the feed stream as overhead

Unit 8-Simple distillation
Takes 100% of $C_2H_4$ and 99% of $C_2H_6$ of the feed stream as overhead

Unit 9-Reactor

The reaction taking place in the unit:

$4\,C_2H_4 \rightarrow 2\,H_2 + 2\,CH_4 + 3\,C_2H_4$

with 80% conversion of $C_2H_6$

To solve the problem we need the component codes in the data bank of CHESS. In our case these are:

1 for hydrogen
2 for methane
22 for ethylene
3 for ethane
23 for propylene
4 for propane
6 for *n*-butane

CHESS has a simple, keyword-driven input language. The sequence of parameters in the parameter lists is fixed. The data line is generally (but not always) terminated by a semicolon, the sequence of keywords is fixed, spaces are used as separators between parameters.

The input data are shown in *Table 3.17/a*.

The main sections of the input file are:

TITLE-user defined name of the problem, which appears on the result sheet as a title.

LAYOUT-layout of the plant in the form of a process matrix. Input streams to the units have positive signs, output streams are denoted by negative numbers. If there is more than one input or output stream to the same unit, the sequence of the streams in the process matrix contains information about the role of the stream.

EQUIPMENT — Contains the unit parameters. For detailed description see the CHESS User Manual.

NUM — The data bank code numbers of the chemical components in the system.

SINTNSY — Known intensive parameters of the feed and product streams of the plant. For an input stream, the first number after the stream number is *1*, whereas for an output stream it is *2*. The other four parameters are: stream type, vapour fraction, temperature, pressure.

SEXTNSV — Extensive parameters of feed streams: stream number, heat content (if known, 0 is assumed), total molar flow and the molar flow rates for each component in the same sequence as under the keyword NUM.

OUTPUT — Specifies the output level.

The results (see *Table 3.17/b*) can be interpreted without further explanation.

Table 3.17/a. Input to Example 1.

```
NEW SAMPLES
TITLE
  ETHYLENE PLANT
LAYOUT
 1  VALV  VALV  1   -2;
 2  FHTR  FHTR  2   -3;
 3  REAC  REAC  3   -4;
 4  MIXR  MIXI  4    38    -5;
 5  PUMP  PUMP  5   -6;
 6  DIST  DIST  6   -10   -7;
 7  DIST  DIST  7   -16   -13;
 8  DIST  DIST  13  -20   -19;
 9  FH1R  FHT2  20  -37;
10  REAC  REA2  37  -38;
EQUIPMENT
1  6;
2  1500000    0.7    345;
3  0.9  6    0.6667    0.6667    0.333    -0.5    0.333    -1    0.1667;
4;
5    5    1270    40    2765    2635;
6  0  0  1  1    0  0    0  0    0;
7  0  0  1  1    0.99    0.05    0    0    0;
8  0  0  1  1    1  0.99    0    0    0;
9  1500000  0.7    345;
10  0.8  4  0.5  0.5  0.75    -1  0  0  0;
NUM  1  2  22    3    25    4  6
SINTNSV
1   1  0  0  298  40;
10  2;
16  2;
19  2;
SEXTNSV
1  0  200  0  0  0  80    0    120    0;
OUTPUT 0
EXECUTE
END
END
```

**Example 2**

This example was an EFCE (European Federation of Chemical Engineers) test example to test flowsheeting systems. The example was constructed to examine three basic features of the systems:

— to what extent is the system prepared for housing new unit models,

— what is the burden on the user of making some modifications on existing models, and

— what is the rate of convergence in case of nonlinear problems?

Table 3.17/b. Result of Example 1

```
.......................................................................
.......................... CHESS ...................................
.......................................................................
.......          UNIVERSITY OF HOUSTON                          .........
.......     CHEMICAL ENGINEERING SIMULATION SYSTEM              .........
.......          VER. 3.1 – 04 FEBRUARY 1980                    .........
.......................................................................

INPUT ROUTINE ENTERED 81/02/23.  15.21.59.
NEW PROBLEM — SAMPLE 1
SYSTEM HAS 7 COMPONENTS.

 ETHYLENE PLANT
                    "PROCESS VECTORS"
                                  STREAM NUMBERS
 ....... EQUIPMENT .......
NUMBER  INTERNAL  EXTERNAL
  1      VALV      VALV        1    -2    0
  2      FHTR      FHTR        2    -3    0
  3      REAC      REAC        3    -4    0
  4      MIXR      MIXI        4    38   -5
  5      PUMP      PUMP        5    -6    0
  6      DIST      DIST        6   -10   -7
  7      DIST      DIST        7   -16  -13
  8      DIST      DIST       13   -20  -19
  9      FHTR      FHT2       20   -37    0
 10      REAC      REA2       37   -38    0
 ETHYLENE PLANT
          "OTHER SYSTEM VARIABLES"
 NUMBER OF COMPONENTS                          7
 COMPONENT NUMBERS USED                    1,  2,  22,  3,  23  4,  6,
 RECYCLE EQUIPMENT LIST (KE2)              4,  5,   6,  7,   8  9,  10

 CONVERGENCE TOLERANCES . "ERROR"
          FLOWRATES                  .01000000
          VAPOUR FRACTION            .01000000
          TEMPERATURE                .01000000
          PRESSURE                   .01000000
          ENTHALPY                   .01000000
          FLASH CALC.S               .00000600

 MAX. LOOPS IN RECYCLE CALC.                5
          IN FLASH CALC.S                  30

 ***BEGIN TRIAL AND ERROR RECYCLE CALCULATIONS WITH EQUIPMENT LIST..
     4,  5,  6,  7,  8,  9,  10.
 ... BEGIN "LOOP  1" ...
 ... BEGIN "LOOP  2" ...
 ... BEGIN "LOOP  3" ...
 ... BEGIN "LOOP  4" ...
 ... BEGIN "LOOP  5" ...
****FBD CF RECYCLE CALCULATIONS
```

Table 3.17/b (continued)

********** FINAL RESULTS

ENHYLENE PLANT

| STREAM NUMBER | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| EQUIP. CONXION | FR 0 TO 1 | FR 1 TO 2 | FR 2 TO 3 | FR 3 TO 4 FR |
| VAPOR FRACTION | 0.0000 | .3378 | 1.0000 | 1.0000 |
| TEMPERATURE, K | 298.0000 | 258.1920 | 345.0000 | 345.0000 |
| PRESSURE, BARS | 40.0000 | 6.0000 | 5.3000 | 5.3000 |
| ENTHALPY, MJ | 2488.1187 | 2488.1187 | 5855.8055 | 8086.1760 |
| | | COMPOSITION, KG-MOLES/HOUR | | |
| HYDROGEN | 0.0000 | 0.0000 | 0.0000 | 72.0036 |
| METHANE | 0.0000 | 0.0000 | 0.0000 | 72.0036 |
| ETHYLENE | 0.0000 | 0.0000 | 0.0000 | 89.9964 |
| ETHANE | 80.0000 | 80.0000 | 80.0000 | 26.0000 |
| PROPYLENE | 0.0000 | 0.0000 | 0.0000 | 35.9640 |
| PROPANE | 120.0000 | 120.0000 | 120.0000 | 12.0000 |
| N-BUTANE | 0.0000 | 0.0000 | 0.0000 | 18.0036 |
| TOTAL | 200.0000 | 200.0000 | 200.0000 | 325.9712 |

| STREAM NUMBER | 5 | 6 | 7 | 10 |
|---|---|---|---|---|
| EQUIP. CONXION | FR 4 TO 5 | FR 5 TO 6 | FR 6 TO 6 | FR 6 TO 0 FR |
| VAPOR FRACTION | 1.0000 | 1.0000 | 0.0000 | 1.0000 |
| TEMPERATURE, K | 344.5585 | 344.5585 | 344.5585 | 344.5585 |
| PRESSURE, BARS | 5.3000 | 16.7160 | 16.7160 | 16.7160 |
| ENTHALPY, MJ | 9240.9885 | 9148.6773 | 4052.0493 | 3686.3819 |
| | | COMPOSITION, KG-MOLES/HOUR | | |
| HYDROGEN | 84.0358 | 84.0358 | 0.0000 | 84.0358 |
| METHANE | 84.0358 | 84.0358 | 0.0000 | 84.0358 |
| ETHYLENE | 109.1350 | 109.1350 | 109.1350 | 0.0000 |
| ETHANE | 32.0161 | 32.0161 | 32.0161 | 0.0000 |
| PROPYLENE | 35.9640 | 35.9640 | 35.9640 | 0.0000 |
| PROPANE | 12.0000 | 12.0000 | 12.0000 | 0.0000 |
| N-BUTANE | 18.0036 | 18.0036 | 18.0036 | 0.0000 |
| TOTAL | 375.1904 | 375.1904 | 207.1187 | 168.0716 |

| STREAM NUMBER | 13 | 16 | 19 | 20 |
|---|---|---|---|---|
| EQUIP, CONXION | FR 7 TO 8 | FR 7 TO 0 | FR 8 TO 0 | FR 8 TO 9 FR |
| VAPOR FRACTION | 0.0000 | 1.0000 | 0.0000 | 1.0000 |
| TEMPERATURE, K | 344.5585 | 344.5585 | 344.5585 | 344.5585 |
| PRESSURE, BARS | 16.7160 | 16.7160 | 16.7160 | 16.7160 |
| ENTHALPY, MJ | 1797.1755 | 2597.9170 | 1074.0770 | 874.0380 |

Table 3.17/b (continued)

| | | | | |
|---|---|---|---|---|
| | COMPOSITION, KG-MOLES/HOUR | | | |
| HYDROGEN | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| METHANE | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| ETHYLENE | 1.0914 | 108.0437 | 0.0000 | 1.0914 |
| ETHANE | 30.4153 | 1.6008 | .3042 | 30.1111 |
| PROPYLENE | 35.9640 | 0.0000 | 35.9640 | 0.0000 |
| PROPANE | 12.0000 | 0.0000 | 12.0000 | 0.0000 |
| N-BUTANE | 18.0036 | 0.0000 | 18.0036 | 0.0000 |
| TOTAL | 97.4743 | 109.6445 | 66.2718 | 31.2025 |
| STREAM NUMBER | 37 | 38 | | |
| EQUIP. CONXION | FR 9 TO 10 | FR 10 TO 4 | FR | |
| VAPOR FRACTION | 1.0000 | 1.0000 | | |
| TEMPERATURE, K | 345.0000 | 345.0000 | | |
| PRESSURE, BARS | 16.0160 | 16.0160 | | |
| ENTHALPY, MJ | 876.1446 | 1155.9959 | | |
| | COMPOSITION, KG-MOLES/HOUR | | | |
| HYDROGEN | 0.0000 | 12.0445 | | |
| METHANE | 0.0000 | 12.0445 | | |
| ETHYLENE | 1.0914 | 19.1580 | | |
| ETHANE | 30.1111 | 6.0222 | | |
| PROPYLENE | 0.0000 | 0.0000 | | |
| PROPANE | 0.0000 | 0.0000 | | |
| N-BUTANE | 0.0000 | 0.0000 | | |
| TOTAL | 31.2025 | 49.2692 | | |

The problem was launched prior to the CHEMCOMP 1982 Symposium on chemical process analysis and design using computers, and the results were shown at the same conference in Antwerp, 25—27 May, 1982.

The problem statement was as follows:

A methanol synthesis reactor is connected into a recycle loop with a heat exchanger/separator and a blow down to split some fraction of the recycle gases (see *Fig. 3.43*).

The results sheet had to contain

— the value of all process stream parameters in the steady state [all streams in kmol/h, temperature (Celsius), and pressure (bar)],
— the characterization of convergence for stream *2*,
— the details of building-in the new reactor module,
— the details of the calculation of $K$ value for the separator.

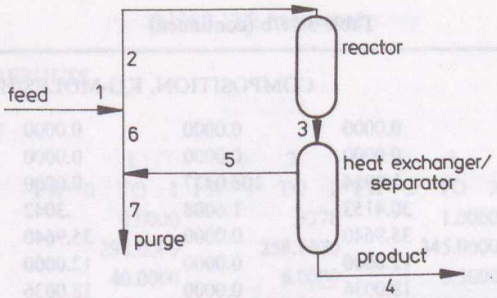



Fig. 3.43. Methanol synthesis loop

| | |
|---|---|
| Input data | |
| Feed rate | 6000 kmol/h |
| composition (vol%) | |
| carbon dioxide | 6 |
| methane | 3.2 |
| carbon monoxide | 14.6 |
| nitrogen | 14.1 |
| hydrogen | 72 |
| water | 0.1 |
| Recycling rate (moles of stream *6*/stream *1*) | 3.5 |
| Reactor pressure (bar) | 90 |
| Reactor output temperature (°C) | 270 |

Requirements to be compiled with by the models:

1. Reactor

The reaction equations to be considered

$$\text{methanol synthesis: } CO + 2\,H_2 \rightleftharpoons CH_3OH \tag{1}$$

$$\text{water gas reaction: } CO + H_2O \rightleftharpoons CO_2 + H_2 \tag{2}$$

$$Kp_1 = p(CH_3OH)/\left(p(CO) \times p(H_2)^2\right)$$

$$Kp_2 = \frac{p(CO) \times p(H_2O)}{p(CO_2) \times p(H_2)}$$

$$Kp_1 = Kp_1^0/Kj_1$$

$$Kp_2 = Kp_2^0/Kj_2$$

where

$$Kj_1 = 0.6 + (1.5e-3)\,(T-473)$$

$$Kj_2 = 0.89 + (4e-4)\,(T-473)$$

Table 3.18. Source code of the new reactor model

```
CHIYODA — FMX4 VSN=2.0                          DATE=5.17/82   PAGE=   1
                    ***LIST OF SOURCE TEXT***
( 10) ↄ= = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
( 20) ↄ *FILECLEAR '3608EFC3SNKI'
( 30) ↄ *FILESTORE '3608EFC3SNKI'  'CHEMCOMP PGM.'
( 40) ↄ= = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
( 50) ↄ MASSBALANCE CALC. IN A SIMPLIFIED METHANOL SYNTHESIS LOOP
( 60) ↄ= = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
( 70)       CDATA('INIT')
( 80)       CDATA('UNIT', 1, 2, 8)
( 90)       CDATA('COMP', 1, 8, 1011, 7, 5, 1, 9, 4211)
(100)       CDATA('REAC', 'RMET', 1411, 401)
(110)       CDATA('STRM', 1, 2, 1, 7, 'V', 4, 'L')
(120)       CDATA('EQUP', 'RE', 10, 'HE', 20)
(130)       A=ALOG(10.)
(140)       T1=260.
(150)       T2=270.
(160)       SREAC('RSYS', 'RMET', 'CK', 1411, 3971.*A, -7.492, 1.77E-3*A, s
(160)                                          -3.11E-8*A, 0., 9.218*A, s
(160)                                  'CK', 401, -1850.*A, 0., 0., 0., 0., 1.664*A)
(170)       ZK1=REAC('RMET', 1411, 'KP', T1)
(180)       ZK2=REAC('RMET',  401, 'KP', T2)
(190)       ZKP1=0.6+1.5E-3*(T1-200.)
(200)       ZKP2=0.89+4.E-4*(T2-200.)
(210)       AK1=ZK1/ZKP1
(220)       AK2=ZKP2/ZK2
(230)       SREAC('RSYS', 'RMET', 'AK', 1411, 'V', AK1, 'AK', 401, 'V', AK2)
(240) ↄ
(250)       XKM=3.8E-3
(260)       XKW=8.0E-4
(270) ↄ
(280)       EPS3=0.5E-3
(290)       TF=200.
(300)       PF=90.
(310)       F1=6000.
(320)       F6=3.5*F1
(330)       SFEED(1, TF, PF, F1, 1, 1, 6., 3.2, 14.6, 4.1, 72., 0.1, 0.)
(340)    10 MIXER(0, 1, 6, 0, 2)
(350)       SPRES(2, 90.)
(360)       STEMP(3, 270.)
(370)       REQIL(10, 'RMET', 2, 3, 0, 0, 1, 0., 0.)
(380)       STEMP(4, 40.)
(390)       STEMP(5, 40.)
(400)       F=XSUM(3)
(410)       FM=XMOL(3, 'VL', 4211)
(420)       FW=XMOL(3, 'V', 9)
(430)       FG=F-FM-FW
(440)       ZM=FM/F
(450)       ZW=FW/F
(460)       A=1, -XKM
```

Table 3.18 (continued)

```
(470)        B=1.-XKW
(480)        AA=A*B
(490)        BB=XKM*B+XKW*A-ZM*B-ZW*A
(500)        CC=XKM*XKW-ZM*XKW-ZW*XKM
(510)        ALF=(-BB+SQRT(BB*BB-4*AA*CC))/2/AA
(520)        SM=ALF/(ALF+(1-ALF)*XKM)
(530)        SW=ALF/(ALF+(1-ALF)*XKW)
(540)        SPBOX(20, 3, 4, 5, 1, 0., 0., 0., 9, SW, 4211, SM)

CHIYODA — FMX4  VSN=2.0                    DATE=5.17/82    PAGE=    2
                    ***LIST OF SOURCE TEXT***

(550)        BLEED(0, 5, 6, 7, F6, 1)
(560)        RCNTL(1, &10, 1, 20, EPS3,2)
(570)        CPRNT('STRM')
(580)        CPRNT('CNTL')
(590)        CPRNT('REAC')
(600)        STOP
```

and

$$^{10}\log Kp_1^0 = 9.218 + (3971/T - 7.492 \log T + (1.77e-3)T - (3.11e-8)T^2$$

$$^{10}\log Kp_2^0 = 1.664 - 1850/T$$

($T$ in $K$)

The equilibrium of the water gas reaction had to be calculated for the reactor outlet temperature, while that of the methanol reaction was to be calculated for a temperature 10 degrees (Celsius) below that.

2. Separator

One stage simple isothermal gas/liquid separator working at 40 °C. Pressure drop is neglected. The liquid phase consists of water and methanol only.

The equilibrium $K$ values at 40 °C:

$$K\,(\text{methanol}) = 3.8e-3$$

$$K\,(\text{water}) \quad = \quad 8e-4$$

In order to be able to compare the convergence behaviour of different systems, the initial values of the variables to be iterated on were fixed.

Initial conversion in the reactor is 0, the initial values of the stream variables of stream *2* are equal to the corresponding values of the feed stream.

Here we show the results obtained using CAPES (CHIYODA, Japan) as they presented it in the form of a poster at the Antwerp symposium. *Table 3.18* shows the source code of the new reactor model, *Table 3.19* is the stream data summary, *Table 3.20* is a demonstration of the convergence behaviour of stream 2.

Table 3.19. Stream data summary

CHIYODA — CAPES IV — 2.6 (PROP=SRK) DATE=5.17/82 PAGE= 0
**STREAM DATA** FORM= 1210

| STREAM NO. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| STREAM NAME | | | | | | | |
| TEMPERATURE .C | 200.0 | 76.1 | 270.0 | 40.0 | 40.0 | 40.0 | 40.0 |
| PRESSURE BAR | 90.000 | 90.000 | 90.000 | 90.000 | 90.000 | 90.000 | 90.000 |
| HEAT FLOW MCAL/H | −53242.1 | −106106.8 | −93082.1 | −92269.7 | −57950.7 | −52859.9 | −5090.8 |
| (JDX) | 100021 | 100021 | 100021 | 100021 | 100021 | 100021 | 100021 |
| MOLE FLOW KGMOL/H | (V) | (V) | (V) | (L) | (V) | (V) | (V) |
| 1 CBN-DIOXIDE | 360.00 | 495.46 | 148.45 | 0.0 | 148.45 | 135.41 | 13.04 |
| 2 METHANE | 192.00 | 2185.34 | 2185.34 | 0.0 | 2185.34 | 1993.37 | 191.98 |
| 3 CBN-MONOXIDE | 876.00 | 1008.35 | 145.07 | 0.0 | 145.07 | 132.33 | 12.74 |
| 4 NITROGEN | 246.00 | 2799.97 | 2799.97 | 0.0 | 2799.97 | 2554.00 | 245.97 |
| 5 HYDROGEN | 4320.00 | 20439.36 | 17671.78 | 0.0 | 17671.78 | 16119.37 | 1552.41 |
| 6 WATER | 6.00 | 9.81 | 356.81 | 352.64 | 4.17 | 3.80 | 0.37 |
| 7 METHANOL | 0.0 | 61.72 | 1272.01 | 1204.34 | 67.67 | 61.73 | 5.94 |
| — TOTAL — | 6000.00 | 27000.00 | 24579.43 | 1556.98 | 23022.45 | 21000.00 | 2022.45 |
| MOLE FRACTN. | (V) | (V) | (V) | (L) | (V) | (V) | (V) |
| 1 CBN-DIOXIDE | 0.06000 | 0.01835 | 0.00604 | 0.0 | 0.00645 | 0.00645 | 0.00645 |
| 2 METHANE | 0.03200 | 0.08094 | 0.08891 | 0.0 | 0.09492 | 0.09492 | 0.09492 |
| 2 CBN-MONOXIDE | 0.14600 | 0.03735 | 0.00590 | 0.0 | 0.00630 | 0.00630 | 0.00630 |
| 4 NITROGEN | 0.04100 | 0.10370 | 0.11392 | 0.0 | 0.12162 | 0.12162 | 0.12162 |
| 5 HYDROGEN | 0.72000 | 0.75701 | 0.71897 | 0.0 | 0.76759 | 0.76759 | 0.76759 |
| 6 WATER | 0.00100 | 0.00036 | 0.01452 | 0.22649 | 0.00018 | 0.00018 | 0.00018 |
| 7 METHANOL | 0.0 | 0.00229 | 0.05157 | 0.77351 | 0.00294 | 0.00294 | 0.00294 |
| | (V) | (V) | (V) | (L) | (V) | (V) | (V) |
| WEIGHT FLOW KG/H | 59168.8 | 206906.0 | 206905.9 | 44942.2 | 161963.7 | 147735.7 | 14228.0 |
| MOLE FLOW KGMOL/H | 6000.00 | 27000.00 | 24579.43 | 1556.98 | 23022.45 | 21000.00 | 2022.45 |
| AVG.M.W. | 9.861 | 7.663 | 8.418 | 28.865 | 7.035 | 7.035 | 7.035 |

Table 3.20. Illustration of the convergence behaviour

CHIYODA — CAPES IV — 2.6 (PROP=SRK) DATE=5.17/82
**TRACE BACK FOR CONVERGENCE (RCNTL)**

NCN= 1 MTD= 1
FLOW RATE: KGMOL/H

STRM. NO.= 2

| ITNO | KEY COMP. | FLOW RATE | RV. ERROR |
|---|---|---|---|
| 1 | CO2 | 360.00 | 0.0 |
| 2 | CH3OH | 11.642 | 1.0000 |
| 3 | CH3OH | 18.869 | 0.3830 |
| 4 | CH4 | 768.00 | 0.2500 |
| 5 | CH4 | 960.00 | 0.2000 |
| 6 | CH4 | 1152.0 | 0.1667 |
| 7 | CH4 | 1344.0 | 0.1429 |
| 8 | CH4 | 1536.0 | 0.1250 |
| 9 | CH4 | 1728.0 | 0.1111 |
| 10 | CH4 | 1920.0 | 0.1000 |
| 11 | CH4 | 2112.0 | 0.0909 |
| 12 | CO2 | 524.36 | 0.0619 |
| 13 | CO2 | 506.47 | 0.0353 |
| 14 | CO2 | 499.55 | 0.0139 |
| 15 | CO2 | 496.94 | 0.0053 |
| 16 | CO2 | 495.96 | 0.0020 |
| 17 | CO2 | 495.59 | 0.0007 |
| 18 | CO2 | 495.46 | 0.0003 |

CHIYODA — CAPES IV — 2.6 (PROP=SRK) DATE=5.17/82

**REACTION SYSTEM DATA**

RSNM =RMET
RS. TYPE (MRC) = 1
RS. KEY COMP. (MKY) =

| R.NO | R.EQUATION |
|---|---|
| 1411 | CO+2*H2=CH3OH |
| 401 | CO+H2O=CO2+H2 |

*EQIL.DATA (RK)

| R. NO | REAC. CODE | KEY COMP. | FRC. CONV | APR. EQIL | APR. KP | EQL. CONST |
|---|---|---|---|---|---|---|
| 1411 | 1231 | | 0.0 | 0.0 | 0.0 | .2150E−02 |
| 401 | 1231 | | 0.0 | 0.0 | 0.0 | 50.68 |

*JOBE

**Example 3**

The third problem to be considered here is a design problem: we have a stripper column, and there is a design specification imposed on the liquid product leaving the reboiler of the column. Such problems are handled by sequential modular packages using a so-called 'controller' unit to 'drive' the calculated value of the design variable to meet the design specification. In effect, the use of the controller unit means a new iteration level. Therefore, this specification type is called 'non-ideal' for sequential modular packages. On the other hand, for an *EO* simulator this type of design constraint is easy to handle: we simply add a new (independent) equation to the equation set. Although the problem was taken from the CHESS user manual, the solution shown below was generated using the QUASILIN package [43], which uses the equation oriented approach. (At that time the package was only under development and was installed at the University of Cambridge (U.K.) Computer Centre on a Cyber Computer.)

**Problem statement**

A feed mixture containing carbon dioxide, nitrogen and *C*1—*C*2 paraffinic hydrocarbons has to be separated in a stripper column into a hydrocarbon rich liquid and a vapour product containing carbon dioxide, nitrogen and methane. The molar ratio of methane to ethane in the liquid product has to be 0.02, with a tolerance of 0.001. The reboiler temperature can be used (in the range of 300—330°C) to control the *C*1/*C*2 ratio. *Figure 3.44* illustrates the situation.

The column has 11 plates. The feed stream is characterized by the following properties:

| | |
|---|---|
| Pressure (bar) | 29 |
| Temperature (°C) | 239 |
| Molar flow rates (kmol/h) | |
| Carbon dioxide | 0.088 |
| Nitrogen | 0.149 |
| Methane | 28.786 |
| Ethane | 20.103 |
| Propane | 24.977 |

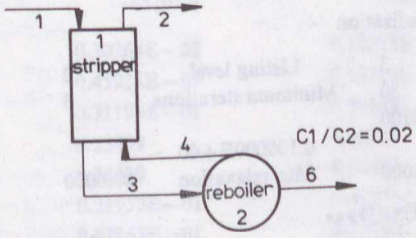



Fig. 3.44. Stripper column with temperature controlled reboiler [30]

| | |
|---|---|
| *i*-butane | 2.832 |
| *n*-butane | 10.221 |
| *n*-pentane | .5893 |

*Table 3.21* shows the input data file.

Table 3.21. Input data file of Example 3

```
THERMOPACK CON
COMPONENTS CARBON DIOXIDE, NITROGEN, METHANE, ETHANE, PROPANE,
I. BUTANE, N-BUTANE, N. HEPTANE

UNIT SOURCE TYPE SOURCE
Data  29  239  U  89.4  0.088  0.149  28.786  20.103  24.977  2.832  10.221  2.224

"SEQUENCE OF SIMPLE PLATES"
UNIT COLUMN TYPE COLUMN
DATA  11  0

"PARTIAL REBOILER WITH TEMPERATURE CONTROL"
UNIT REBOILER TYPE HEAT-CONTROLLED FLASH
DATA  0  3  0  I  1.1

STREAM FEED FROM SOURCE TO COLUMN: LIQUID IN
STREAM VAPOUR PRODUCT FROM COLUMN: VAPOUR OUT
STREAM LIQUID FROM COLUMN: LIQUID OUT TO REBOILER: FEED
STREAM VAPOUR FROM REBOILER: VAPOUR TO COLUMN: VAPOUR IN
STREAM LIQUID PRODUCT FROM REBOILER: LIQUID
STREAM HEAT FLOW TO REBOILER: HEAT

CONSTRAINTS  1
LIQUID PRODUCT NODE  1
IF STARTNODE < EQUATION NODE1:TEMP=330; >
ELSE          < EQUATION  50.0  NODE1:CONP (3)-NODE1:COMP (4)=0; >

sw qufet

QUASILIN RUN  11  PLATES  8  COMP WIDE RANGE CONSTR. 3/4
Input buffer is .CHESS:A11
Control file is .CONTROL
Time limit     60
Store limit    350

                Quasilin initalization

Trace level                  3        Listing level              2
Maximum iterations          30        Minimum iterations         3
Convergence tolerance   0.00100
Linear eqn parameters   0.0           0.100000E+00
Relaxation factor       0.80000       Min relaxation      0.40000

*** CCN THERMOPACK USED ***

G12SPS I=          -1 J=          1 K=          2581  L=                0
```

Table 3.22. Results of Example 3

CONVERGED IN TO ITERATIONS — ERROR 2.71797E−04

| Stream name | Feed | VAPOUR PRODUCT | LIQUID |
|---|---|---|---|
| Pressure (bar) | 29.000 | 29.000 | 29.000 |
| Temperature (K) | 239.000 | 239.105 | 312.201 |
| Enthalpy (kj/kgm) | −89577.9 | −79225.1 | −90395.4 |
| Mole fractions | | | |
| | 0.98456E−03 | 0.56826E−03 | 0.21211E−02 |
| | 0.16670E−02 | 0.45240E−02 | 0.25483E−07 |
| | 0.32206 | 0.86379 | 0.18496E−01 |
| | 0.22492 | 0.99307E−01 | 0.41583 |
| | 0.27945 | 0.27796E−01 | 0.38561 |
| | 0.31685E−01 | 0.11491E−02 | 0.35617E−01 |
| | 0.11435 | 0.28481E−02 | 0.12152 |
| | 0.24883E−01 | 0.20971E−04 | 0.20607E−01 |
| Mole flowrates | | | |
| | 0.88020E−01 | 0.18720E−01 | 0.23796 |
| | 0.14903 | 0.14903 | 0.28589E−05 |
| | 28.79243 | 28.45546 | 2.07500 |
| | 20.10748 | 3.27144 | 46.65187 |
| | 24.98259 | 0.91564 | 43.26056 |
| | 2.83263 | 0.37853E−01 | 4.01824 |
| | 10.22328 | 0.93834E−01 | 13.63300 |
| | 2.22450 | 0.69083E−03 | 2.31189 |
| Total flowrate | 89.39999 | 32.94266 | 0.11219E+03 |

| | VAPOUR | LIQUID PRODUCT | COLUMN..S001 |
|---|---|---|---|
| | 29.000 | 29.000 | 29.000 |
| | 329.762 | 329.762 | 239.220 |
| | −85103.9 | −82692.1 | −79423.6 |
| | 0.30264E−02 | 0.12275E−02 | 0.57138E−03 |
| | 0.45916E−07 | 0.53139E−08 | 0.19477E−02 |
| | 0.31190E−01 | 0.59642E−02 | 0.86586 |
| | 0.53499 | 0.29821 | 0.99698E−01 |
| | 0.34439 | 0.42629 | 0.27902E−01 |
| | 0.21953E−01 | 0.49503E−01 | 0.11541E−02 |
| | 0.62865E−01 | 0.17942 | 0.28610E−02 |
| | 0.15804E−02 | 0.39389E−01 | 0.21109E−04 |

Table 3.22. (continued)

| | VAPOUR | LIQUID PRODUCT | COLUMN..S001 |
|---|---|---|---|
| | 0.16866 | 0.69300E−01 | 0.16467E−01 |
| | 0.25589E−05 | 0.30001E−06 | 0.56132E−01 |
| | 1.73828 | 0.33672 | 24.95355 |
| | 29.81575 | 16.83611 | 2.87328 |
| | 19.19344 | 24.06711 | 0.80413 |
| | 1.22346 | 2.79479 | 0.33260E−01 |
| | 3.60354 | 10.12945 | 0.82456E−01 |
| | 0.88079E−01 | 2.22382 | 0.60836E−03 |
| | 55.73123 | 56.45732 | 28.81989 |
| Stream name | COLUMN..S002 | COLUMN..S003 | COLUMN..S004 |
| Pressure (bar) | 29.000 | 29.000 | 29.000 |
| Temperature (K) | 239.432 | 239.988 | 241.519 |
| Enthalpy (kj/kgm) | −79516.0 | −79575.5 | −79661.6 |
| Mole fractions | 0.57887E−03 | 0.60100E−03 | 0.66989E−03 |
| | 0.83701E−03 | 0.36898E−03 | 0.16328E−03 |
| | 0.86586 | 0.86334 | 0.85493 |
| | 0.10055 | 0.10289 | 0.10965 |
| | 0.28107E−01 | 0.28656E−01 | 0.30195E−01 |
| | 0.11632E−02 | 0.11876E−02 | 0.12657E−02 |
| | 0.28847E−02 | 0.29470E−02 | 0.31230E−02 |
| | 0.21346E−04 | 0.21959E−04 | 0.23720E−04 |
| Mole flowrates | | | |
| | 0.16714E−01 | 0.17372E−01 | 0.19396E−01 |
| | 0.24168E−01 | 0.10377E−01 | 0.44378E−02 |
| | 25.00053 | 24.95573 | 24.76261 |
| | 2.90314 | 2.97420 | 3.17468 |
| | 0.81156 | 0.82830 | 0.87423 |
| | 0.33587E−01 | 0.34326E−01 | 0.36356E−01 |
| | 0.83291E−01 | 0.85186E−01 | 0.90421E−01 |
| | 0.61629E−03 | 0.63476E−03 | 0.68676E−03 |
| Total flowrate | 28.87361 | 28.90613 | 28.95280 |
| | COLUMN..S005 | COLUMN..S006 | COLUMN..S007 |
| | 29.000 | 29.000 | 29.000 |
| | 245.495 | 254.309 | 268.856 |
| | −79887.5 | −80649.6 | −82148.8 |
| | 0.88574E−03 | 0.15108E−02 | 0.28778E−02 |
| | 0.64626E−04 | 0.26283E−04 | 0.97801E−05 |
| | 0.83048 | 0.76469 | 0.61920 |
| | 0.12903 | 0.18141 | 0.29761 |
| | 0.34455E−01 | 0.45447E−01 | 0.69267E−01 |
| | 0.14460E−02 | 0.19462E−02 | 0.30548E−02 |
| | 0.36168E−02 | 0.49299E−02 | 0.78962E−02 |
| | 0.28885E−04 | 0.44166E−04 | 0.86035E−04 |

Table 3.22 (continued)

| | COLUMN..S005 | COLUMN..S006 | COLUMN..S007 |
|---|---|---|---|
| | 0.25777E−01 | 0.44786E−01 | 0.91584E−01 |
| | 0.18808E−02 | 0.78069E−03 | 0.31125E−03 |
| | 24.16878 | 22.71312 | 19.70573 |
| | 3.75492 | 5.38836 | 9.47143 |
| | 1.00270 | 1.34989 | 2.20441 |
| | 0.42082E−01 | 0.57807E−01 | 0.97219E−01 |
| | 0.10526 | 0.14643 | 0.25130 |
| | 0.84062E−03 | 0.13119E−02 | 0.27381E−02 |
| | 29.10222 | 29.70288 | 31.82471 |
| Stream name | COULMN..S008 | COULMN..009 | COULMN..S010 |
| Pressure (bar) | 29.000 | 29.000 | 29.000 |
| Temperature (K) | 285.301 | 299.355 | 312.201 |
| Enthalpy (kJ/kgm) | −84610.1 | −86739.2 | −87575.8 |
| Mole fractions | | | |
| | 0.45233E−02 | 0.51577E−02 | 0.44772E−02 |
| | 0.31268E−05 | 0.85751E−06 | 0.21099E−06 |
| | 0.40260 | 0.20703 | 0.88502E−01 |
| | 0.46677 | 0.60038 | 0.62916 |
| | 0.10826 | 0.16084 | 0.23638 |
| | 0.48628E−02 | 0.72121E−02 | 0.11247E−01 |
| | 0.12806E−01 | 0.19080E−01 | 0.29751E−01 |
| | 0.17406E−03 | 0.30329E−03 | 0.49124E−03 |
| Mole flowrates | | | |
| | 0.16798 | 0.23364 | 0.23613 |
| | 0.11612E−03 | 0.38844E−04 | 0.11128E−04 |
| | 14.95147 | 9.37796 | 4.66761 |
| | 17.33467 | 27.19620 | 33.18187 |
| | 4.02038 | 7.28600 | 12.46649 |
| | 0.18057 | 0.32670 | 0.59315 |
| | 0.47555 | 0.86432 | 1.56909 |
| | 0.64641E−02 | 0.13739E−01 | 0.25908E−01 |
| Total flowrate | 37.13721 | 45.29867 | 52.74026 |
| | COLUMN..S011 | COLUMN..S012 | COLUMN..S013 |
| | 29.000 | 29.000 | 29.000 |
| | 239.106 | 239.220 | 239.432 |
| | −90145.6 | −90170.1 | −90186.2 |
| | 0.10057E−02 | 0.10080E−02 | 0.10153E−02 |
| | 0.65824E−03 | 0.28323E−03 | 0.12156E−03 |
| | 0.29657 | 0.29693 | 0.29629 |
| | 0.23112 | 0.23133 | 0.23207 |
| | 0.29165 | 0.29155 | 0.29164 |
| | 0.33163E−01 | 0.33146E−01 | 0.33142E−01 |
| | 0.11975 | 0.11968 | 0.11966 |
| | 0.26085E−01 | 0.26068E−01 | 0.26059E−01 |

Table 3.22 (continued)

| | COLUMN..S011 | COLUMN..S012 | COLUMN..S013 |
|---|---|---|---|
| | 0.85767E−01 | 0.86014E−01 | 0.86672E−01 |
| | 0.56133E−01 | 0.24168E−01 | 0.10377E−01 |
| | 25.29047 | 25.33742 | 25.29260 |
| | 19.79034 | 19.73921 | 19.81029 |
| | 24.87109 | 24.87858 | 24.89529 |
| | 2.82804 | 2.82837 | 2.82911 |
| | 10.21191 | 10.21275 | 10.21464 |
| | 2.22442 | 2.22442 | 2.22444 |
| | 85.27721 | 85.33093 | 85.36345 |
| Stream name | COLUMN..S014 | COLUMN..S015 | COLUMN..S016 |
| Pressure (bar) | 29.000 | 29.000 | 29.000 |
| Temperature (K) | 239.988 | 241.519 | 245.495 |
| Enthalpy (kj/kgm) | −90209.6 | −90268.0 | −90423.9 |
| Mole fractions | | | |
| | 0.10385E−02 | 0.11112E−02 | 0.13252E−02 |
| | 0.51963E−04 | 0.21986E−04 | 0.90644E−06 |
| | 0.29375 | 0.28642 | 0.26752 |
| | 0.23429 | 0.24066 | 0.25794 |
| | 0.29202 | 0.29301 | 0.29500 |
| | 0.33148E−01 | 0.33157E−01 | 0.33108E−01 |
| | 0.11966 | 0.11962 | 0.11927 |
| | 0.26045E−01 | 0.26001E−01 | 0.25826E−01 |
| Mole flowrates | | | |
| | 0.88695E−01 | 0.95077E−01 | 0.11418 |
| | 0.44381E−02 | 0.18811E−02 | 0.78099E−03 |
| | 25.08943 | 24.50557 | 23.04990 |
| | 20.01077 | 20.59102 | 22.22449 |
| | 24.94124 | 25.06975 | 25.41693 |
| | 2.83114 | 2.83687 | 2.85259 |
| | 10.21988 | 10.23472 | 10.27569 |
| | 2.22450 | 2.22465 | 2.22612 |
| Total flowrate | 85.41013 | 85.55954 | 86.15990 |

Table 3.22 (continued)

| COLUMN..S017 | COLUMN..S018 | COLUMN..S019 | COLUMN..S020 |
|---|---|---|---|
| 29.000 | 29.000 | 29.000 | 29.000 |
| 254.309 | 268.856 | 285.301 | 299.355 |
| −90763.1 | −91250.8 | −91665.9 | −91734.2 |
| 0.18224E−02 | 0.25352E−02 | 0.29771E−02 | 0.27870E−02 |
| 0.35290E−05 | 0.12439E−05 | 0.38469E−02 | 0.10465E−06 |
| 0.22703 | 0.16334 | 0.95470E−01 | 0.45828E−01 |
| 0.29799 | 0.36509 | 0.43273 | 0.45808 |
| 0.29759 | 0.30010 | 0.30812 | 0.33456 |
| 0.32759E−01 | 0.31790E−01 | 0.30676E−01 | 0.31026E−01 |
| 0.11759 | 0.11331 | 0.10804 | 0.10713 |
| 0.25221E−01 | 0.23829E−01 | 0.21989E−01 | 0.20602E−01 |
| 0.16088 | 0.23728 | 0.30294 | 0.30543 |
| 0.31155E−03 | 0.11642E−03 | 0.39144E−04 | 0.11428E−04 |
| 20.04247 | 18.28819 | 9.71467 | 5.00434 |
| 26.30766 | 34.17081 | 44.03233 | 50.01799 |
| 26.27147 | 28.08746 | 31.36307 | 36.53360 |
| 2.89200 | 2.97536 | 3.12148 | 3.38794 |
| 10.38076 | 10.60501 | 10.99378 | 11.69854 |
| 2.22655 | 2.23028 | 2.23755 | 2.24972 |
| 88.28203 | 93.59453 | 0.10176E+03 | 0.10920E+03 |

| Stream name | HEAT FLOW |
|---|---|
| Heat flow (kj/time) | 729804.37500 |

The last line of the input file contains the design specification in form of an equation:

$$50 : \text{comp}(3) - 1 : \text{comp}(4) = 0$$

The problem was solved using the thermodynamic data base of CONCEPT [34] and converged to the solution in 10 iterations. QUASILIN used the well-known MA28 nonlinear equation solving procedure.

*Table 3.22* shows the results. The streams S001–S020 are the intermediate streams of the column with the following notation:

S001–S010 vapour streams of plate 2–11
S011–S020 liquid streams of plate 1–10, respectively.

(*Note*: the plates of the column are numbered in a top-down direction. The vapour stream of plate 1 is called 'vapour product', and the liquid stream of plate 11 is called 'liquid'.)

## 3.4 References

[1] Westerberg, A. W., Hutchinson, H. P., Motard, R. L. and Winter, P. (1979): *Process Flowsheeting*. Cambridge University Press.

[2] Perkins, J. D. (1984): Equation oriented flowsheeting. In: *Proc. 2nd Int. Conf. on Foundations of Computer-Aided Process Design*, Westerberg, A. W. and Chien, H. H. (eds) *CACHE, Ann.Arbor*, p. 309.

[3] Shacham, M., Machietto, S., Stutzman, L. F. and Babcock, P. (1982): *Comput. Chem. Eng.*, **6** (2), 79.

[4] Donati, G., Marini, L. and Marziano, G. L. (1982): *Chem. Eng. Sci.*, **17** (8), 1265.

[5] Wen, C. Y. and Fan, L. T. (1975): *Models for Flow Systems and Chemical Reactors*. Marcel Decker, New York.

[6] Steward, D. V. (1962): *SIAM Rev.*, **4**, 321.

[7] Harary, F. (1959): *J. Math. Phys.*, **38**, 104.

[8] Arlazanov, V. L. E., Dinic, E. A., Kronrod, M. A. and Paradzeve, I. A. (1970): *Soviet Math. Dokl.*, **11** (5), 1209.

[9] Mah, R. S. H. (1974): *AIChE J.*, **20** (6), 1227.

[10] Tarján, R. E. (1972): *SIAM J. Comput.*, **1** (2), 146.

[11] Sargent, R. W. H. and Westerberg, A. W. (1964): *Trans. Instn. Chem. Eng.*, **42**, T190.

[12] Hopcroft, J. E. and Karp, R. M. (1973): *SIAM J. Comput.*, **2**, 225.

[13] Upadhye, R. S. and Grens, E. A. (1975): Selection of decompositions for process simulation. *AIChE J.*, **21**, 136.

[14] Pretzel, O. (1986): On reordering graphs by pushing down maximal vertices. *Order 3*, 135.

[15] Berge, C. (1962): *The Theory of Graphs and its Applications*. Wiley, New York.

[16] Stadtherr, M. A. and Wood, E. S. (1984): *Comp. Chem. Eng.*, **8** (1), 19.

[17] Stadtherr, M. A. and Wood, E. S. (1984): *Comp. Chem. Eng.*, **8** (1), 9.

[18] Evans, L. B. (1982): *Process flowsheeting — A state of the art review*. CHEMCOMP'82 Symp. Antwerp., Belgium, May 1982.

[19] Vela, M. A. (1961): Use split fraction for recycle balances. Parts I, II. *Petroleum Refiner*, **40** (5), 247; **40** (6), 189.

[20] Naphtali, L. M. and Sandholm, D. P. (1971): *AIChE. J.*, **17**, 148.

[21] Lee, N. B. M. (1980): *Simultaneous solution of design, torn stream and multicomponent distillation equations in a sequential modular flowsheeting package*. M. Sc. Thesis, Department of Chemical Engineering and Chemical Technology, Imperial College of Science and Technology, University of London.

[22] Perkins, J. D. (1981): *Lectures on flowsheeting, and personal communication*. Department of Chemical Engineering and Chemical Technology, Imperial College of Science and Technology, University of London.

[23] Wegstein, J. (1958): *Comm. ACM.*, **1**, 9.

[24] Broyden, C. G. (1965): *Math. Comp.*, **19** (92), 577.

[25] Bluck, D., Hughes, P., Mallin-Jones, A. K., Perris, F. A. and Seppard, A. J. (1976): *Flowpack II. — A third generation flowsheeting system*. Paper B-6 to EFCE Conference "Design Congress 76", Sept. 1976, Aston, Birmingham.

[26] Rosen, E. M. (1980): Steady-state chemical process simulation: a state of the art review. In: Squires, R. G. and Reklaitis, G. V. (eds) *Computer Applications to Chemical Engineering*. ACS Symp. Ser., 124.

[27] Perkins, J. D. and Sargent, R. W. H. (1982): A computer program for steady-state and dynamic simulation and design of chemical processes, *AIChE Symp. Ser.*, 214, Vol. **78**.

[28] Clark, S. M. and Reklaitis, G. V. (1984): *Comp. Chem. Eng.*, **8** (3/4), 205.

[29] Leesley, M. E. (ed.) (1982): *Computer-Aided Process Plant Design*. Gulf, Houston.

[30] Motard, R. L. and Lee, H. M. (1971): *CHESS user's Guide*, (3rd ed.) Sept. 1971. Dept. of Chem. Eng., University of Houston, Texas. 77004.

[31] Seader, J. D., Seider, W. D. and Pauls, A. C. (1977): *FLOWTRAN simulation — An introduction.* (2nd ed.) Ann Arbor, Michigan, CACHE Corp.
[32] Berger, F. and Perris, F. A. (1979): *Comp. Chem. Eng.*, **3**, 309.
[33] *PROCESS INPUT MANUAL*, Revision 1. 1980. Simulation Science Incorporated.
[34] *CONCEPT MARK IV. User's Guide.* CAD Centre, Cambridge, 1980.
[35] Benedek, P. (1980): *Steady-state Flowsheeting of Chemical Plants.* Elsevier, Amsterdam.
[36] Benedek, P., Pallai, I., Váczi, P. and Halász, L. (1980): FORSIM-A Flowsheet simulation program package of third generation. CHEMPLANT'80 Symp., Hévíz (Hungary), 3—5 Sept. 1980. *EFCE Publ. Ser. 10.*
[37] Rose, R. H. (1983): *EURECHA-UNICORN Flowsheeting program.* User manual, Version 4.0, ETH Zurich.
[38] Gallier, P. W. et al. (1980): ASPEN. Advanced capabilities for modelling and simulation of industrial processes. In: Squires, R. G. and Reklaitis, G. V. (eds) *Computer Applications to Chemical Engineering.* ACS Symp. Ser. 124.
[39] Perkins, J. D. (1979): *Efficient solution of design problems using a sequential-modular flowsheeting program.* CACE'79, EFCE Symp., Montreux (Switzerland), 1979.
[40] Broyden, C. G. (1971): *Math. Comp.*, **25** (114), 285.
[41] Froment, N. M. (1981): *SPEED-UP User's Manual.* Dept. of Chem. Eng. and Chem. Techn., IC. of Science and Technology, University of London.
[42] Evans, L. B., Boston, J. F. et al. (1979): ASPEN: *An advanced system for process engineering.* CACE'79, EFCE Symp., Montreux (Switzerland).
[43] Gorzinski, E. W. and Hutchinson, H. P. (1978): Towards a quasilinear process simulator — I. Fundamental ideas. *Comp. Chem. Eng.*, **2**, 189.
[44] Cunningham, J. R. and Westerberg, A. W. (1972): The GENDER Routines. *Proceedings of the Sixth Annual Princeton Conference on Information Sciences and Systems*, Princeton Univ., New Jersey.
[45] Babcock, P. D. (1982): *Thermodynamic data base design and structures for the* 1980*'s.* CHEM-COMP'82 Symp. Antwerp., Belgium, 25—27. May.
[46] Duff, I. S. (1979): MA28: a set of FORTRAN subroutines for sparse asymmetric linear equations. R8730. *AERE Harwell Report*, H. M. S. O. London 1979.
[47] Locke, M. H. and Westerberg, A. W.: The ASCEND—II system. Presented at *AICHE National Meeting*, Los Angeles (Nov. 1982).
[48] Duff, I. S. (1977): A survey of sparse matrix research. *Proc. IEEE*, **65** (1977), 500.
[49] Duff, I. S. (1982): A survey of sparse matrix software. *R*10512, *AERE Harwell Report*, 1982.
[50] Boston, M. J., Brit, H. I. (1978): A radically different formulation of the single stage flash problem. *Comp. Chem. Eng.* **2**, 109.
[51] Tímár, L., Simon, F., Csermely, Z. et al. (1980): *MASTEP — Sequential and simultaneous modular algorithm in a flowsheeting program.* CHEMPLANT'80 Hévíz (Hungary) 3—5 Sept. 1980. EFCE Publ. Ser., 10.
[52] Shacham, Kechat, E.: *The fraction separated method I. — For calculation of process flowsheets.* Dept. of Chem. Eng. TECHNION, Haifa, Israel, Report CE 73/74.
[53] Rippin, D. W. T. (1983): *Comp. Chem. Eng.*, **7** (2), 65.
[54] Rosen, E. M. (1962): *Chem. Eng. Prog.*, **58** (10), 69.

## 3.5 Literature suggested for further reading

### 3.5.1 Flowsheeting systems and related topics

Biegler, L. T. and Hughes, R. R. (1982): *AIChE J.*, **28** (6), 994.
Billingsley, D. S. (1984): *AIChE J.*, **30** (4), 686.
Black, C. (1982): *CEP 78,* (Dec.), 53.
Boston, M. J. (1980): Inside-out algorithms for multicomponent separation process calculations. In: Squires, R. G. and Reklaitis, G. V. (eds) *Computer Applications to Chemical Engineering.* ACS Symp. Ser., 124.
Bush, M. J., Pulido, J. and Johnson, A. I. (1978): Svrcek, W. I.: *Comp. Chem. Eng.*, **2,** 161.
Chien, H. H. Y. (1984): *Comp. Chem. Eng.*, **8** (1), 61.
Cho, Y. S. and Joseph, B. (1983): *AIChE J.*, **29** (2), 261.
Doering, F. J. and Gaddy, F. L. (1980): *Comp. Chem. Eng.*, **4** (2), 133.
Evans, L. B., Joseph, B. and Seider, W. D. (1977): *AIChE J.*, **23** (5), 658.
Fonyó, Z., Nishimura, H. and Yamashita, Y. (1983): *AIChE J.*, **29** (4), 538.
Hlavacek, V. (1977): *Comp. Chem. Eng.*, **1,** 75.
Lucia, A. and Machietto, S. (1983): *AIChE J.*, **29** (5), 705.
Mahalec, V., Kluzik, H. and Evans, L. B. (1979): *Comp. Chem. Eng.* **3,** 373.
Mathias, P. M., Boston, J. F. and Watanasiri, S. (1984): *AIChE J.*, **30** (2), 182.
Motard, R. L., Shacham, M. and Rosen, E. M. (1975): *AIChE J.*, **21,** 417.
Neville, J. M. and Seider, W. D. (1980): *Comp. Chem. Eng.*, **4** (1), 49.
Pierucci, S., Troiani, F. and Ranzi, E. (1982): *Comp. Chem. Eng.*, **6** (1), 19.
Prausnitz, J., Eckert, C. A., Orye, R. V. and O'Connell, J. P. (1980): *Computer Calculations for Multicomponent Vapor–liquid and Liquid–liquid Equilibria.* Prentice-Hall, Englewood Cliffs.
Russel, R. A. (1980): *Comp. Chem. Eng.*, **4** (3), 167.
Sargent, R. W. H. (1979): *Comp. Chem. Eng.*, **3** (1), 17.
Shacham, M. and Cutlip, M. B. (1982): *Comp. Chem. Eng.*, **6** (3), 209.
Simon, F., Siklós, J., Tímár, L. and Bácskai, S. (1980): An interactive system approach for estimation of physical and thermodynamic properties. CHEMPLANT'80 Symp., Hévíz (Hungary), 3—5 Sept. 1980. *EFCE Publ. Ser.* 10.
Singh, S. P. and Carnahan, B. (1981): *Comp. Chem. Eng.*, **5** (4), 263.
Thomson, D. W. (1982): *Comp. Chem. Eng.*, **6** (1), 27.
Thumbynayagam, R. K. M., Branch, S. J. and Winter, P. (1980): CONCEPT MK IV: A new approach to process flowsheeting. CHEMPLANT'80 Symp., Hévíz (Hungary), 3—5 Sept. 1980. *EFCE Publ. Ser.*, 10.
Tímár, L., Simon, F., Csermely, Z. et al. (1984): *Comp. Chem. Eng.*, **8** (3/4), 185.
Trevino-Losano, R. A., Kisala, T. P. and Boston, J. F. (1984): *Comp. Chem. Eng.*, **8** (2), 105.
Vasek, V. and Klemes, J. (1983): *Comp. Chem. Eng.*, 7 (3), 175.
Weisenfelder, A. J. and Olson, R. E. (1980): *CEP*, **76** (1), 40.
Westman, K. R., Lucia, A. and Miller, D. C. (1984): *Comp. Chem. Eng.*, **8** (3/4), 219.

### 3.5.2 Structure analysis and iteration procedures

Billingsley, D. S. (1984): *AIChE J.*, **30** (4), 686.
Crowe, C. M. (1984): *Comp. Chem. Eng.*, **8** (1), 35.
Gallun, S. E. and Holland, C. D. (1980): *Comp. Chem. Eng.*, **4** (2), 93.

Gros, H., Kaialuoto, S. and Mattson, I. (1977): Some new aspects on partitioning and tearing in steady-state process simulation. In: *Computer Applications in the Analysis of Chemical Data and Plants*. Science Press, Princeton.
Kechat, E. and Shacham, M. (1973): *Process Technology International*, **18** (1/2), 35.
Mah, R. S. H. (1983): *Comp. Chem. Eng.*, 7 (4), 239.
Mah, R. S. H. (1982): *CEP*, **78** (July), 84.
Mah, R. S. H. and Lin, T. D. (1980): *Comp. Chem. Eng.*, **4** (1), 75.
Motard, R. L. and Westerberg, A. W. (1981): *AIChE J.*, **27** (5), 725.
Murthy, C. L. N. and Husain, A. (1983): *Comp. Chem. Eng.*, 7 (2), 129.
Norman, R. L. (1965): *Am. Inst. of Chem. Engs.*, **11**, 450.
Orbach, O. and Crowe, C. M. (1971): *Can. J. Chem. Eng.*, **49**, 503.
Paloschi, G. (1982): *The numerical solution of nonlinear equations representing chemical processes*. PhD Dissertation. Imperial College of Science and Technology, University of London.
Pho, T. K. and Lapidus, L. (1973): *AIChE J.*, **19** (6), 1170.
Shacham, M. (1984): *RIChE J.*, **30** (1), 92.
Shacham, M. and Mah, R. S. H. (1978): *Comp. Chem. Eng.*, **2**, 64.
Stadtherr, M. A. and Hilton, C. M. (1982): *Comp. Chem. Eng.*, **6** (2), 115.
Tierman, J. C. (1977): *Comm. ACM.*, **13** (12), 722.
Westerberg, W. A. and Berna, Th. J. (1978): *Comp. Chem. Eng.*, **2** (1), 61.

# 4 ECONOMIC ANALYSIS OF CHEMICAL ENGINEERING SYSTEMS

by

F. TÁTRAI and P. SOLTÉSZ

## 4.1 Profitability analysis

*Introduction*

The word profitability is used as a general term to indicate the amount of profit that can be obtained from a given situation [3]. Profitability, therefore is, a common denominator to all business activities. The objective of the analysis is usually the maximization of the profit gained from any capital investment.

Capital investment for new equipment or a new process constitutes a current expenditure in the hope of obtaining a more than commensurate return of funds at a later time. All investment evaluation methods appraise the commercial attractiveness of estimated cash inflows as compared with estimated cash outflows.

Before any capital investment is made in a project or enterprise, the decision maker seeks answers to some important questions:

— whether the project is worth implementing at all;

— whether or not it would be more attractive to invest the money in another form of enterprise;

— if there are alternative processes, which of the possible process alternatives is economically the most attractive.

In order to answer these questions it is necessary to know how much profit can be obtained. This is the goal of the economic evaluation or economic analysis of a proposed investment.

Why do we use the profit as a measure of the worth of a project?

First of all, money is a common and suitable unit for the measurement and comparison of different things. Second, there is a real shortage of capital in the world, therefore capital can be regarded as a scarce resource. There is competition among alternative uses for this scarce resource, and the decision maker is the judge in this competition. He needs some objective measure to compare the rival investment alternatives, bearing in mind that he has to make the most of a scarce resource. The only way to do this is to accept the most profitable investment. Without proper investment it would no longer be possible to remain competitive.

Normally there are some "noneconomic" factors, which should also be taken into account before taking the final decision. Sometimes these factors are irrefutable, and override any economic considerations. These may involve health hazards, pollution, etc. Such noneconomic factors are not considered in this chapter.

To judge the profitability of a proposed investment one needs some insight into the future. We have to estimate some future cash flows into and out of the firm. This is a process that is associated with some uncertainty and risk. The question is whether the income will be greater than the investment and, if so, whether the difference will be large enough to justify the risk associated with the investment. The greater the difference, the higher the acceptable risk. The decision maker has to be aware of this uncertainty and risk. In order to be able to do this we need to know something about the probability of our estimates. For a detailed discussion of this subject see Rose [2].

When the engineer or manager has to decide whether a capital project for a new plant or equipment is worthwhile, he must choose from a bewildering array of analytical methods. In other words, he has to choose an appropriate economic design criterion. Profitability analysis methods fall into three major groups [10]:

— Return on investment, in which the profit after deduction of all costs is compared with the investment in the project.
— Payout time methods, which calculate the number of years between the start of commercial operation of the project and the time when total income from the project equals all expenditures, including plant investment.
— Project value or worth, which is a sum of money that represents the potential value of the entire project to the company.

There are naturally a number of variations in each of these groups, but only a few of them are widely used [6].

Before examining the principles leading to economic design criteria, it is necessary to explain the main concepts and terms used in profitability studies.

### 4.1.1 Definition of terms

In this section we define the terms used in profitability studies and introduce the notation which will be used in the remainder of the text. Furthermore, we shall examine the relationship of the different cost elements and profitability terms.

To establish a consistent notation, we will use the following conventions [14]:

— annual costs and incomes will be referred to by the capital letter A.
— capital expenditure (capital cost, working capital, etc.) will be referred to by the capital letter C,
— the future worth of money will be referred to by the capital letter F,
— the present worth of money will be referred to by the capital letter P,
— fractional interest rates of different kinds are denoted by i,
— t is used for the tax rate,
— the scrap value or salvage value will be referred to by the capital letter S.

This convention, introduced by Holland, Watson and Wilkinson [14] will be used throughout this chapter.

*Annual costs and cash flows*

Let As denote the revenue from the annual sales of the products, ATE the annual expense of production and sale. ACI, the annual cash income, is the difference between these two quantities:

$$ACI = As - ATE \tag{4.1}$$

Note that

(a) ATE contains no annual provision for plant depreciation;

(b) although, in a typical year of operation, the sales income, As, is the only source of revenue for the firm, there are two other possible sources of income:

— scrap income,
— working capital income.

These latter two can be realized only at the end of a project's life.

Subtracting the annual amount of tax, AIt from the annual cash income, we derive a new quantity, the annual net cash income

$$ANCI = ACI - AIt \tag{4.2}$$

The annual amount of tax can be calculated from the tax rate, t, and the taxable income (ACI–AD–AA), where AD is the annual depreciation, AA is the annual amount of any other tax allowances, AD and AA are book transaction values; they do not involve real cash flows.

$$AIt = (ACI - AD - AA) \times t \tag{4.3}$$

The value of t is determined by the tax authorities.

The net annual cash flow after tax, ANCF, is calculated according to the following expression:

$$ANCF = ANCI - ATC \tag{4.4}$$

where ATC is the annual capital charge, which is not necessarily zero after the first year of operation.

The above equations are illustrated by *Fig. 4.1.*

The net annual cash flow after tax is the basis of the net present worth and discounted cash flow rate of return calculations (see later in this chapter).

The total annual expenditure, ATE is often divided into two terms, the annual general expenditure, AGE, and the annual manufacturing cost or expense:

$$ATE = AGE + AME \tag{4.5}$$

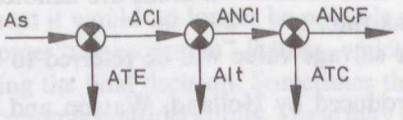



Fig. 4.1. Interdependence of annual cash flows

AGE, the annual general expenditure contains the following items:

administration,
sales,
shipping of product,
advertising and marketing,
technical service,
research and development,
finance.

The term AGE can be divided into two parts: the annual fixed general expenses, AFGE and the annual variable general expenses, AVGE:

$$AGE = AFGE + AVGE \tag{4.6}$$

Similarly, for the manufacturing costs:

$$AME = AFME + AVME \tag{4.7}$$

Another possible classification of the manufacturing expenses is the differentiation between direct and indirect manufacturing costs:

$$AME = ADME + AIME \tag{4.8}$$

Direct manufacturing costs include the following expenses (payments directly proportional to production rate):

— raw material (including delivery),
— catalysts, solvents, fine chemicals, additives, etc.,
— operating labour,
— utilities costs,
— operating maintenance and repairs, operating supplies,
— royalties, licence fees, patents.

Indirect manufacturing costs include expenses incurred as a result of, but not directly from the production process, and those that are a function of capital investment. These expenses are:

— payroll overhead,
— control laboratory,
— general plant overhead,
— packaging,
— storage facilities,
— property taxes,
— rent, insurance, taxes.

Some authors differentiate between a so-called 'labour term', proportional to the direct operating labour costs, and a 'fixed investment term', which is regarded as being proportional to the fixed capital costs (see below) [1].

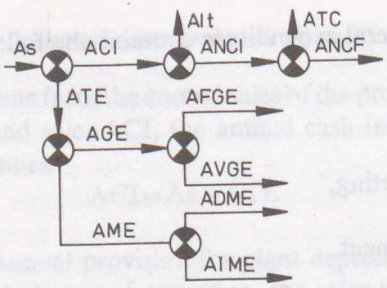



Fig. 4.2. Interdependence of annual cash flows

Our directed graph representing the relation of cash flows is now extended *(Fig. 4.2)*.

*Profits*

There are various definitions of profit in the literature. Profit can be calculated before tax is deducted, or on an after-tax basis. Accountants use terms of gross annual profit, AGP, and net annual profit, ANP, both calculated on a pre-tax basis:

$$AGP = As - AME - ABD \tag{4.9}$$

where ABD is the so-called balance sheet annual depreciation charge, which can differ from AD, the annual depreciation used for tax purposes [see Eq. (4.3)].

The net annual profit, ANP, can be calculated from the gross annual profit:

$$ANP = AGP - AGE \tag{4.10}$$

Combining Eq.(4.10) with Eqs (4.9) and (4.5), ANP may alternatively be written as

$$ANP = ACI - ABD \tag{4.11}$$

The net annual profit after tax, ANNP, is defined by the relation

$$ANNP = ANCI - ABD \tag{4.12}$$

where ANCI is the same as that defined in Eq.(4.2).

Our graph representation of the various cost elements involved in profit calculations is slightly different from that used for cash flow calculations *(Fig. 4.3)*.

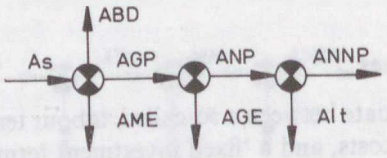



Fig. 4.3. Interdependence of terms used in profit calculations

The profit terms are used by accountants mainly in the construction of balance sheets.

Note that in profit calculations the balance sheet depreciation value, ABD, is subtracted from the cash income (therefore the profit terms always differ from the cash flow terms), but there is no allowance for the actual year's capital expenditure in profit calculations:

$$\text{ANNP} = \text{ANCF} + \text{ATC} - \text{ABD} \tag{4.13}$$

If there was no capital expenditure in the actual year of accountancy, the net annual profit after tax is the net annual cash flow less the book value of annual depreciation:

$$\text{ANNP} = \text{ANCF} - \text{ABD} \tag{4.14}$$

*Elements of capital cost calculation*

Capital cost (or investment, or capital expenditure) is defined as expenditure on factors of production in order to create physical assets which are themselves used to produce goods or other physical assets [59]. The capital cost is a fixed amount of money to be tied up and risked in a project or enterprise. This general definition becomes simpler in the process industries. Capital represents the investment in building and operating the process plant in order to manufacture a specific product [15]. This risked capital is the sum of (sometimes) several years' expenditure. (Note its difference from annual costs, cash flows and profits.)

The total capital cost, CTC, of a project consists of the

— fixed capital cost, CFC, the
— working capital cost, CWC, and the
— cost of land and other non-depreciable costs, CL:

$$\text{CTC} = \text{CFC} + \text{CWC} + \text{CL} \tag{4.15}$$

Some authors differentiate between fixed capital costs or investment in process equipment within the processing area (or site limits), and the investment in auxiliary services generally located outside the processing area. The former is called fixed investment in a narrow sense, the latter is called auxiliary investment cost [1]. Typical auxiliary investments are such items as

— steam generation facilities,
— fuel stations,
— fire protection systems,
— transformers,
— power stations, etc.

Such items serve the process under examination, and possibly other processes within a large process plant, as well. The subsystems served by these auxiliary

services pay a portion of the cost. However, in this chapter we regard the auxiliary equipment costs as a part of the fixed capital cost.

The fixed capital cost includes the following items [3, 15, 11]:

— major process equipment (i. e. reactors, tanks, pumps, filters, distillation columns, etc.),
— installation cost of major process equipment,
— process piping,
— insulation,
— instrumentation,
— auxiliary equipment (see above),
— outside lines (i. e. piping external to buildings, supports and posts for overhead piping, electric feeders from power stations),
— land and site improvements,
— buildings and structures,
— engineering and construction (design and engineering fees plus supervision of plant erection),
— contractors' fees.

Working capital, CWC, is the capital tied up on behalf of the project in the form of ready cash to meet operating expenses, inventories of raw materials, and products. Working capital is completely recoverable at any time (in theory), therefore no tax allowance is made for depreciation. The same holds for the cost of land. For this reason some authors regard the cost of land occupied as an element of the working capital.

The working capital cost includes the following items [15]:

— raw materials for plant startup,
— raw material, intermediate and finished product inventories,
— cost of handling and transportation of materials to and from stores,
— cost of inventory control, warehouse, associated insurance, security arrangements, etc.,
— money to carry accounts receivable (e. g. credit extended to customers, the value of product delivered, but not paid for), less accounts payable (value of raw materials received, but not yet paid for),
— money to meet payrolls needed for startup,
— readily available cash for emergencies,
— any additional cash required to operate the process or business.

A commonly used simplification is the calculation of the working capital from the value of raw material and product held at normal stock levels, plus the value of product delivered but not paid for, minus the value of raw materials received, but not yet paid for.

Working capital may vary from a very small fraction of the total capital invested into the project to almost the whole of it, depending on the process and industry. As a rule of thumb, the average working capital is of the order of 10 to 20% of the

fixed capital investment in the chemical process industries. Since it involves saleable material and expected payments, working capital is a relatively safe investment as compared to the fixed investment costs.

*Depreciation*

As we have seen above the annual amount of tax is calculated from the tax rate, t, and the taxable income, (ACI − AD − AA). The term AD in this equation is called depreciation. What is the meaning of this term? Why do we use it?

Originally, depreciation was defined as 'a reasonable allowance for the exhaustion, wear, and tear of property used in the trade or business, including a reasonable allowance for obsolence' [3].

Depreciation is a fraction of the fixed capital involved. Over the project life (called the 'useful life') the sum of the yearly depreciations must be equal to the capital invested [2].

Depreciation can be regarded as

— a tax allowance,
— a cost of operation,
— a means of creating a fund to finance plant replacement,
— a means of measuring the decrease in value with age [15].

In the first context the annual depreciation charge simply reduces the annual amount of tax.

In its second meaning, depreciation is considered to be a manufacturing cost in the same sense as raw material cost, which has to find its reflection in the value of the product.

The problem is that it is difficult to estimate the depreciation cost per unit of product. In this context, depreciation is employed as a means of distributing the original investment over the useful life of the project.

Third, depreciation can be regarded as a means of creating a fund to replace the plant. Replacement does not necessarily mean the replacement of the depreciated and sometimes obsolete process with a new, identical one. The management is free to invest the available capital in the most promising project, and therefore depreciation is regarded as a portion of the retained profit to be used in the financing of new capital projects.

In the fourth context, depreciation is a measure of the falling value of a plant or item of equipment. This decrease in value is a consequence of the fact that with the passage of time the remaining useful life (the number of years for which income will be earned) decreases. At the end of the useful life the equipment has only a salvage or scrap value, S.

Let us denote the useful life by s.

If the original fixed investment was CFC, the fund that has to be accumulated in s years is equal to (CFC − S).

There are several techniques for the calculation of depreciation, and we discuss four of them here:

— straight-line depreciation,
— the declining balance method,
— the sum-of-years-digits method, and
— the sinking found method.

The simplest one is straight-line depreciation. According to this method, the annual amount of depreciation, AD is expressed by the formula:

$$AD = (CFC - S)/s \tag{4.16}$$

The book value, B′, at the end of a particular year, s′, is given by:

$$B' = CFC - s'AD \tag{4.17}$$

The principal significance of a particular depreciation is for tax purposes, because annual depreciation is subtracted from the annual profit before tax is paid. The basis for depreciation in any particular case is a matter for agreement between tax authorities and the company [15], in accordance with the tax laws.

The sraight-line depreciation ensures depreciation of equal amounts of the value due to depreciation for each year of the project's life. For the investor it would be more beneficial to depreciate the equipment or plant by a greater amount during the early years of the project's life than during the later years. The effect of such types of depreciation methods is equivalent of a tax allowance in the early years of a project's life at the expense of the tax to be paid later, and it enhances the security of investments. There are two methods of this type: the declining balance and the sum-of-years-digits method.

According to the declining balance method, the book value at the end of the first year is

$$B' = CFC(1 - r)$$

where r is a fraction to be agreed to by the tax authorities.

The book value at the end of a particular year, s′, is defined by the expression:

$$B' = CFC(1 - r)^{s'} \tag{4.18}$$

If we choose the value of r to be twice the reciprocal of the useful life, i. e. $r = 2/s$, then the method is called the double-declining-balance method.

Although the declining-balance method allows equipment or process to be depreciated by a greater amount during the early years of the project, it does not ensure that the plant is depreciated to a zero value at the end of the useful project life.

The sum-of-years-digits method also allows equipment or process to be depreciated by a greater amount during the early years of the project than during the later years, but it does ensure that the plant is depreciated to a scrap value at the end of the project's useful life.

According to this rule, the annual amount of depreciation for a particular year, s′, is given by:

$$AD' = \frac{(s - s' + 1)}{(1 + 2 + 3 + \ldots s)} \times (CFC - S) \tag{4.19}$$

or

$$AD' = \frac{(2(s - s' + 1))}{s(s + 1)} \times (CFC - S) \tag{4.20}$$

The book value at the end of a particular year s′ can be given by:

$$B' = \frac{2(1 + 2 + \ldots + (s - s')) \times (CFC - S)}{s(s + 1)} + S \tag{4.21}$$

The fourth method discussed is the sinking fund method (rarely used in the rapidly changing circumstances of the modern chemical industry).

According to this method, the depreciation allowance is smaller in the early years of the project than in the later years, and consequently it has an effect on the security of investors opposite to the two methods described above.

In this method, the annual amount of depreciation is the same in each year of the useful life of the project. The series of equal amounts of depreciation is regarded as being invested at a fractional interest rate, i, and the investment is made at the end of each year over the service life of s years. The fund accumulated in this way builds up a future sum equal to CFC − S.

The annual amount of depreciation can be given by the following expression:

$$AD = (CFC - S) \times ff \tag{4.22}$$

where ff is the future worth factor (see later in this chapter under the heading 'Time value of money'), given by the relation:

$$ff = \frac{1}{(1 + i)^{s-1}} \tag{4.23}$$

To illustrate the effect of the different depreciation methods, let us calculate the fall in book value for an equipment characterized by the following data:

| | | |
|---|---|---|
| Fixed capital cost, CFC | = | \$ 230,000 |
| Scrap value, S | = | \$ 30,000 |
| Useful life in years, s | = | 10 |
| Fractional interest, i | = | 0.1 |

For this example we calculated the book values according to the four different depreciation methods. The results are summarized in *Table 4.1*.

Table 4.1. Book values of the same equipment item calculated with different depreciation methods

| Year | Depreciation method | | | |
|---|---|---|---|---|
| | Straight line | Sinking fund | Double declining | Sum of years digits |
| 1 | 210.0 | 217.45 | 148.00 | 193.64 |
| 2 | 190.0 | 203.64 | 147.20 | 160.91 |
| 3 | 170.0 | 188.46 | 117.76 | 131.81 |
| 4 | 150.0 | 171.76 | 94.21 | 106.36 |
| 5 | 130.0 | 153.39 | 75.37 | 84.55 |
| 6 | 110.0 | 137.17 | 60.29 | 66.36 |
| 7 | 90.0 | 110.94 | 48.23 | 51.82 |
| 8 | 70.0 | 86.49 | 38.58 | 40.91 |
| 9 | 50.0 | 59.59 | 30.87 | 33.64 |
| 10 | 30.0 | 30.00 | 24.69 | 30.00 |

The results in *Table 4.1* are shown graphically in *Fig. 4.4.*

Some of the depreciation methods discussed above are not permitted by the tax rules of a given country: it is therefore always necessary to consult the local tax authorities before calculating depreciation.

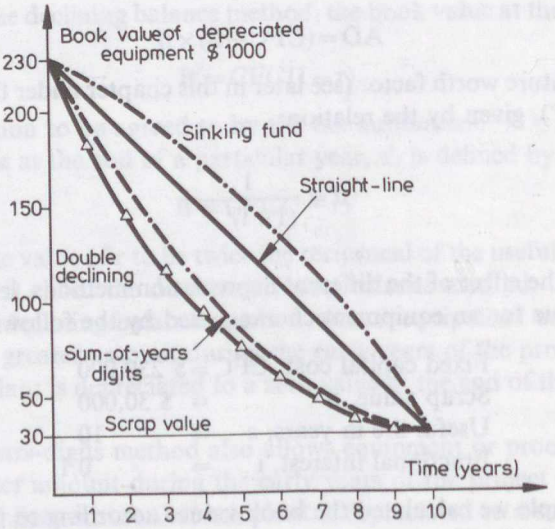



Fig. 4.4. Influence of depreciation method on book values

### 4.1.2 Measures of profitability

*Traditional measures of profitability*

Holland [16] gives a critical survey of some of the traditional methods of assessing profitability, in particular the rate of return and payout time methods.

*Rate of return*

Rate of return is defined in its most general form by the expression [3, 10]:

$$\text{Rate of return (\%)} = \frac{\text{Annual profit}}{\text{Invested capital}} \times 100$$

There are a number of variations of this simple method, depending on the precise meaning of the terms 'annual profit' and 'invested capital'.

Weinberger [10] presents a simple sample calculation of return on investment:

**Example 4.1**

The following data apply:

| | | |
|---|---|---|
| Investment required | | |
| New plant investment | $300,000 | |
| Working capital | $200,000 | |
| Income and expense statement | (per year) | |
| Sales income | | $750,000 |
| Manufacturing costs | $400,000 | |
| Sales + distribution costs | $100,000 | |
| General overhead costs | $150,000 | |
| Total costs | $650,000 | |
| Net annual cash income (pre-tax) | | $100,000 |
| Net annual cash income (after 52% tax) | | $48,000 |

$$\text{Rate of return on investment} = \frac{\$48{,}000 \times 100(\%)}{\$300{,}000 + \$200{,}000} = 9.6\%/\text{yr}$$

It is interesting to note that Weinberger used the rate of return defined above (with profit in the numerator), and he used cash flow terms in this example without any comment.

Some textbooks and papers tend to use 'annual saving' (e.g. Ref. [12]), others use different types of annuities. These facts indicate the necessity for using more precise definitions. In order to overcome this difficulty we will use the terms introduced in Section 4.2.

We shall discuss here the rate of return calculations based on four different types of cash flow and profit:

— the net annual profit before tax, ANP,
— the net annual profit after tax, ANNP,
— the annual cash income before tax, ACI,
— the annual cash income after tax, ANCI.

The investment capital term in the above equation can be one of the following:

— original investment,
— depreciated investment,
— average investment.

Combining all kinds of 'profit' with all kinds of invested capital, 12 different variations can be derived.

Instead of discussing all of the 12 methods in detail, we wish simply to show the effect of the choice of profit terms and capital cost terms.

To begin with the profit terms, we summarize their definitions, as introduced in Section 4.1.1.

The annual cash income was given as the difference between the revenue from annual sales, As, and the total annual expense, ATE:

$$ACI = As - ATE \qquad \text{[see Eq.(4.1)]}$$

The net annual cash income was:

$$ANCI = ACI - AIt \qquad \text{[see Eq.(4.2)]}$$

where AIt denoted the income tax.

The net annual profit before tax, ANP, was related to the above quantities by the relation

$$ANP = ACI - ABD \qquad \text{[see Eq.(4.11)]}$$

where ABD referred to the balance sheet annual depreciation.

A similar formula was derived for the net annual profit after tax, ANNP:

$$ANNP = ANCI - ABD \qquad \text{[see Eq.(4.12)]}$$

The total capital cost, CTC, of a project is the sum of three terms:

$$CTC = CFC + CWC + CL \qquad \text{[see Eq.(4.15)]}$$

where CFC, CWC and CL have the same meaning as before.

To demonstrate the effect of using different profit terms, we always use the original total invested capital as denominator.

First, the fractional interest rate of return using the net annual profit before tax is:

$$i = ANP/CTC \qquad (4.24)$$

or using Eq.(4.11):

$$i = ACI/CTC - ABD/CTC \qquad (4.25)$$

In the second case, using the net annual profit after tax, ANNP, the expression takes the form:

$$i = ANNP/CTC \tag{4.26}$$

or using Eq.(4.12):

$$i = ANCI/CTC - ABD/CTC \tag{4.27}$$

If we use the annual cash income before tax instead of profit, we get the simple expression:

$$i = ACI/CTC \tag{4.28}$$

In the last case, with the annual cash income after tax:

$$i = ANCI/CTC \tag{4.29}$$

The main disadvantage of using Eqs (4.24), (4.25), (4.26) and (4.27) is that the balance sheet annual depreciation, ABD is assessed arbitrarily. Its value has a considerable effect on the fractional interest rate of return, leading to difficulties in comparing different companies and especially companies from different countries. These difficulties may be overcome by using Eq.(4.28) or (4.29), and it is quite simple to deduce the particular value of depreciation, if necessary.

It is possible to generalize the above equations if we use average annual cash flows and profits over the length of useful life of the project instead of using values for a particular year.

To illustrate the effect of choice of capital investment term, we shall derive expressions for the cases when ANNP and ANCI were used as profit terms. Similar expressions can be developed using ANP and ACI values.

The expressions using the original capital investment are Eqs (4.26), (4.27) and (4.29).

If we calculate rate of return on depreciated investment, CFC has to be replaced by B′, the book value of the fixed capital investment at the end of a particular year, s′:

$$i = \frac{ANNP}{B' + CWC + CL} \tag{4.30}$$

This equation can be written in terms of Eq.(4.12) as:

$$i = \frac{ANCI - ABD}{B' + CWC + CL} \tag{4.31}$$

Using annual cash income after tax on depreciated total capital investment, the equation takes the form:

$$i = \frac{ANCI}{B' + CWC + CL} \tag{4.32}$$

It is important to note that the value of B′ depends strongly on the type of depreciation method used.

If we calculate the fractional interest rate of return on average investment, CTC has to be replaced by the average value of the capital investment, which can be expressed as:

$$(DFC+S+2CWC+2CL)/2 \tag{4.33}$$

since, at the beginning of the service life CTC = CFC + CWC + CL, and at the end of the useful life CTC = S + CWC + CL.

Using Eq. (4.35), we obtain the following expressions:

$$i=\frac{2ANNP}{CFC+S+2CWC+2CL} \tag{4.34}$$

This equation can be written in terms of Eq. (4.12) as:

$$i=\frac{2(ANCI-ABD)}{CFC+S+2CWC+2CL} \tag{4.35}$$

Holland *et al.* [16] notes that if $i$ is chosen on the basis of the average values of ANNP and ANCI over the length of the project, an average value of the working capital, CWC, must be used.

In the case when we use annual cash income after tax on depreciated total capital investment, the equation takes the form:

$$i=\frac{2ANCI}{CFC+S+2CWC+2CL} \tag{4.36}$$

*Payout period*

Another traditional measure of profitability is the payout period. Payout period, or payout time, is defined as the minimum length of time theoretically necessary to recover the original capital investment in the form of cash flow to the project based on total cash income minus all costs except depreciation [3].

Generally, for this method, the original capital investment means only the original, nondepreciable, fixed capital investment, and interest effects are neglected.

The cash income in this case is defined in the same way as in Section 4.1.1:

$$ANCF=ANCI-ATC \qquad \text{[see Eq. (4.4)]}$$

Since the net annual cash flow after tax, ANCF, is related to the net annual profit after tax, ANNP (if there is no capital expenditure in the actual year of accountancy), by the expression:

$$ANCF=ANNP+ABD \qquad \text{[see Eq. (4.14)]}$$

the payout period can also be formulated in profit terms.

The two definitions are:

$$\text{Payout time} = \frac{\text{depreciable fixed capital}}{\text{av. profit/yr} + \text{av. depr/yr}}$$

This definition, involving yearly average values for profit and depreciation, is used by Peters and Timmerhaus [3] and Weinberger [10].

According to our terminology, the payback period, PBP, is the time required for the cumulative net cash flow, ANCF, taken from the startup of the plant to become equal to the depreciable fixed-capital investment (CFC − S):

$$\sum_{s'=0}^{s'=\text{PBP}} \text{ANCF} = \text{CFC} - \text{S} \tag{4.37}$$

This definition is used by Holland *et al.* [16], Rose [2] and Herron [12]. This definition is not a real measure of profitability, but rather one of the time taken for the cash flows to recoup the original depreciable fixed capital investment. According to this definition, the cash flows after the break-even point have no effect on the PBP. However, the profitability is affected by these cash flows.

Therefore, PBP values can be misleading, especially if alternatives having considerably different useful lives are compared.

This method is used quite often, however. It has the advantage of being simple to understand and easy to use. Its value indicates the risk involved in the project; the longer the payout period, the greater the risk, especially for smaller firms with liquidity problems. Equipment investments with a short payout period are attractive, since unforeseen events occurring beyond the PBP cannot result in net loss [12].

Other definitions of payout period take the time value of money (see below) into consideration. These are designated as payout period including interest. These methods increase the payout period above that found with no interest charge, and give preference to investments that earn most of their profits during the early years of their useful life.

An example will help to place the methods discussed above in their true perspective.

**Example 4.2**

We will calculate different types of rate of return and the payback period for the project in Example 4.1.

(a) net annual cash income rate of return on:
- — original investment,
- — depreciated investment using straight-line depreciation,
- — average investment.

(b) the payback period,

(c) net annual profit (after tax) rate of return.

(d) To show the effect of depreciation method, we calculate the net annual cash income rate of return of depreciated investment using:

— straight-line,
— sinking fund,
— double declining, and
— sum of years digits depreciation methods.

The following data apply:

| | | |
|---|---|---|
| Fixed capital investment, CFC | : | $ 230,000 |
| Scrap value, S | : | $ 30,000 |
| Working capital, CWC | : | $ 30,000 |
| Cost of land, CL | : | $ 10,000 |

The cash flow data are summarized in *Table 4.2.*

Ad (a): Rate of return from annual net cash income, ANCI on original investment.

The original investment is calculated from Eq. (4.15):

$$CTC = \$\ 230{,}000 + \$\ 30{,}000 + \$\ 10{,}000 = \$\ 270{,}000$$

The fractional interest rate of return i, is defined by Eq. (4.29) only for positive cash flows. Therefore, we have to take the first year when the first positive ANCI

Table 4.2. Cash flow data percentage interest rate of return values for Example 4.1

| Year | ANCF $ 1000 | ANCI $ 1000 | i (ANCI on CTC) % |
|---|---|---|---|
| 1 | −10 | | |
| 2 | −20 | | |
| 3 | −30 | | |
| 4 | −230 | | |
| 5 | 20 | 20 | 7.4 |
| 6 | 30 | 30 | 11.1 |
| 7 | 40 | 40 | 14.8 |
| 8 | 50 | 50 | 18.5 |
| 9 | 55 | 55 | 20.4 |
| 10 | 60 | 60 | 22.2 |
| 11 | 40 | 40 | 14.8 |
| 12 | 30 | 30 | 11.1 |
| 13 | 20 | 20 | 7.4 |
| 14 | 90 | 20 | 7.4 |
| Av. | | | 13.5 |

Table 4.3. Percentage interest rate of return values on total, depreciated and average capital investment

| Year | Rate of return (%), ANCI on | | |
|---|---|---|---|
| | original total investment | depreciated investment (straight-line) | average investment |
| 5 | 7.4 | 8.0 | 11.8 |
| 6 | 11.1 | 13.0 | 17.6 |
| 7 | 14.8 | 19.0 | 23.5 |
| 8 | 18.5 | 26.3 | 29.4 |
| 9 | 20.4 | 32.4 | 32.4 |
| 10 | 22.2 | 40.0 | 35.3 |
| 11 | 14.8 | 30.8 | 23.5 |
| 12 | 11.1 | 27.8 | 17.6 |
| 13 | 7.4 | 22.2 | 11.8 |
| 14 | 7.4 | 28.6 | 11.8 |
| Av. | 13.5 | 24.8 | 21.5 |

value occurs (this is actually the first year of operation, year 5). Note that ANCI = ANCF + ATC, but in this case annual capital expenditure is zero throughout the useful life except for the last year, when working capital, cost of land and scrap value are recovered, causing a negative ATC value:

$$\text{ATC} = -\$\,30{,}000 - \$\,30{,}000 - \$\,10{,}000 = -\$\,70{,}000$$

From Eq. (4.29), the net annual cash income rate of return on the original investment for the fifth year is:

$$i = \$\,20{,}000/\$\,270{,}000 \times 100 = 7.4\%/\text{yr}$$

Values listed in *Table 4.2* were calculated in a similar way.

To calculate the rate of return values on the depreciated investment using the straight-line depreciation method, we can use the depreciated book values listed in the second column of *Table 4.1*.

Since the first year with positive net cash income is year 5, we can substitute the data for this year in combination with the first depreciated value into Eq. (4.32):

$$i = \frac{(\$\,20{,}000) \times 100}{\$\,210{,}000 + \$\,30{,}000 + \$\,10{,}000} = 8.0/\text{yr}$$

The subsequent values in the third column of *Table 4.3* were calculated in the same way.

We now proceed to calculate the net annual cash income rate of return on the

average investment using Eq. (4.36). For the first year with positive net annual cash flow:

$$i = \frac{2 \times (\$\ 20{,}000) \times 100}{\$\ 230{,}000 + \$\ 30{,}000 + \$\ 60{,}000 + \$\ 20{,}000} = 11.8\%/\text{yr}$$

The subsequent values in the fourth column of *Table 4.3* were calculated in the same way.

Comparing the values in the different columns, we can see that the numbers in the second column are always smaller than those in columns 3 and 4, i. e. the rate of return on the original investment is always smaller than on the depreciated or on the average investment.

On the other hand, comparing the trend of the values calculated with the same method, we can conclude that they have a maximum at the time of the maximum net annual cash income. The rate of return values on depreciated investment remain steadily high after the maximum. This is a consequence of the continually decreasing denominator (the depreciated value of the capital investment) of the equation used.

Ad (b): The calculation of the payback period is straightforward.

The difference between the original fixed capital and the salvage value is:

$$CFC - S = \$\ 230{,}000 - \$\ 30{,}000 = \$\ 200{,}000$$

Starting from the beginning of year 5 (the first year of operation), we listed the individual and accumulated cash flows in the following table:

| Year | s′ | ANCF (i) | Sum ANCF (i) |
|---|---|---|---|
| 5 | 1 | 20.000 | 20.000 |
| 6 | 2 | 30.000 | 50.000 |
| 7 | 3 | 40.000 | 90.000 |
| 8 | 4 | 50.000 | 140.000 |
| 9 | 5 | 55.000 | 195.000 |
| 10 | 6 | 60.000 | 255.000 |

From this Table we can conclude that the PBP is 5 years and one month. (Due to the uncertainty of future cash flow estimation, 5 years is a good estimate.)

Ad (c): The net annual profit (after tax) rate of return.

We will calculate the average value using the ABD value calculated by the straight-line depreciation method. The annual balance-sheet depreciation was:

$$(CFC - S)/10 = \$\ 20{,}000$$

According to Eq. (4.27), the average net annual after-tax profit rate of return is the difference between two terms: the average annual cash income rate of return and the fractional depreciation rate.

Table 4.4. Percentage interest rate of return values on depreciated investment with different depreciation methods

| Rate of return ANCI on depreciated investment | | | | |
|---|---|---|---|---|
| Year | Depreciation method | | | |
| | straight line | sinking fund | double declining | sum of years digits |
| 5 | 8.0 | 7.8 | 8.9 | 8.6 |
| 6 | 13.0 | 12.3 | 16.0 | 14.9 |
| 7 | 19.0 | 17.5 | 25.0 | 23.3 |
| 8 | 26.3 | 23.6 | 37.3 | 34.2 |
| 9 | 32.4 | 28.4 | 47.7 | 44.2 |
| 10 | 40.0 | 34.7 | 59.8 | 56.4 |
| 11 | 30.8 | 26.5 | 45.4 | 43.6 |
| 12 | 27.3 | 23.7 | 38.2 | 37.1 |
| 13 | 22.2 | 20.0 | 28.2 | 27.2 |
| 14 | 28.6 | 28.6 | 28.6 | 30.9 |
| Av. | 24.8 | 22.3 | 33.5 | 32.0 |

The latter is:

$$ABD/CTC = \$\ 20{,}000/270{,}000 = 0.074\ (7.4\%)$$

Consequently,

$$i = (13.5 - 7.4) = 6.1\%$$

Ad (d): To show the effect of the depreciation method used, we calculated the net annual cash income rate of return of depreciated investment using straight-line, sinking fund, double declining, and sum of years digits depreciation methods.

The calculations were based on the book-values of the depreciated investment (see *Table 4.1*), and the net annual cash flows listed in *Table 4.2*.

The results are shown in *Table 4.4*.

## Modern methods of assessing profitability

Most modern methods take the time value of money into account. We thus have to introduce this term.

### *Time value of money*

It is quite simple to understand that, if there is a chance of having $ 1000 now, or the same amount of money in two years, it is worth having it now. Here are some reasons why:

— if we put it in a bank immediately, it would earn interest;
— money has a tendency to lose its value due to inflation.

On the other hand, money can be loaned or borrowed. There is always a risk that it may not be returned. In this approach interest is the inducement offered to make the risk acceptable. When money is borrowed, interest is paid; when money is loaned, interest is received. The longer the period for which the money is loaned, the greater the total amount of interest paid. We define the present worth, $P$, of a sum of money, $F$, to be spent (or received) at a future time as the amount that would have to be set aside, at a specified interest rate $i$, to yield that future sum, $F$, when the specified time arrives. The future worth of the money, $F$, is therefore always greater than its present worth, $P$ [17].

Generally, interest is compounded, either in discrete time intervals or continuously.

The most common method is discrete compounding. According to this method, the future worth, $F$, and the present worth, $P$ are related by:

$$F = P(1+i)^n \tag{4.38}$$

where $i$ is the fractional interest rate, and $n$ is the number of periods for which interest is compounded.

Introducing a future worth factor, $ff$, with the definition:

$$ff = (1+i)^n \tag{4.39}$$

Eq. (4.38) can be simplified to the form:

$$F = P \times ff \tag{4.40}$$

Conversely,

$$P = F/(1+i)^n \tag{4.41}$$

or

$$P = F \times fd \tag{4.42}$$

where $fd$, the discount factor, is defined by:

$$fd = 1/ff = (1+i)^{(-n)} \tag{4.43}$$

In modern accounting practice interest is compounded annually.

Continuous discounting means that the interest is calculated and added continually to the principal. If we consider continuous discounting, the relation of present worth and future worth is slightly different.

At any time $x$, the total sum is the original sum of money (the principal) plus the interest payments. The rate of increase in time $x$ to $x+dx$ is proportional of the total sum, $F$, and the nominal interest rate, $i'$:

$$dF = F \times i' \, dx \tag{4.44}$$

The integration of Eq. (4.44) from year 0 to year $n$, for an initial value $P$ and a final value, $F$ after $n$ years, gives:

$$F = P \exp(i' n) \tag{4.45}$$

Equation (4.45) means that the future worth factor, *ff*, and the discount factor, *fd* have to be defined in this case as:

$$ff = \exp(i' n) \tag{4.46}$$

$$fd = \exp(-i' n) \tag{4.47}$$

and Eqs (4.40) and (4.42) still hold true.

It is important to note that the growth of the original principal is faster, if continuous compounding is used, as compared to discrete annual compounding with the same annual interest rate. However, the difference is significant only when applied to very large sums of money.

We will always use annual compounding periods in the calculation of interest.

If we have to gather a future sum of money, $F$, and we have a series of annual cash incomes, $A$, each year, invested (or put into a bank) at a fractional interest compounded annually, the relation of these terms can be derived from the theory of geometric series. The resulting equation is:

$$F = A \frac{((1+i)^{n-1})}{i} \tag{4.48}$$

or

$$F = A/faf \tag{4.49}$$

where *faf*, the annuity future worth factor is:

$$faf = i/(1+i)^{n-1} \tag{4.50}$$

If we have to calculate the present worth of $F$, we can use Eq. (4.42):

$$P = F \times fd = A \times fd/faf \tag{4.51}$$

The term *faf*/*fd* is called the annuity present worth factor, *fap* [17].

From the above equations, the annuity, $A$, required to build up a future sum of money, $F$, with a present worth $P$ is:

$$A = F \times faf \tag{4.52}$$

or

$$A = P \times fap \tag{4.53}$$

The modern methods for assessing profitability have some common features:

— most of them take the time value of money into account by using discount factors;

— they use annual discounted or undiscounted cash flows rather than profit terms;

— they try to express the cash flows during the whole life of the project rather than using yearly average values.

The modern profitability assessment methods presented in this chapter are as follows:

— net present value or worth (NPV),
— discounted cash flow rate or return (DCFRR),
— discounted break-even point (DBEP),
— equivalent maximum investment period (EMIP),
— interest recovery period (IRP),
— capital rate-of-return ratio (CRR).

*Net present value or worth (NPV)*

To define the net present value, we have to introduce the term annual discounted cash flow, ADCF, with the following definition:

$$\mathrm{ADCF} = \mathrm{ANCF} \times fd \tag{4.54}$$

Accordingly, we discount the net annual cash flow after tax, ANCF with a discount factor, $fd$, which is a function of the interest, $i$, and the number of years. If annual interest compounding is used, Eq. (4.54) takes the form:

$$\mathrm{ADCF}(n) = \mathrm{ANCF}(n)/(1+i)^n \tag{4.55}$$

The sum of the annual discounted cash flows over $s$ years, ADCF, is called the net present value, NPV

$$\mathrm{NPV} = \sum_{j=0}^{s} \mathrm{ADCF}(j) \tag{4.56}$$

If the calculated value is negative, the project is certainly not attractive. If its value is near zero, then the project has earned an interest rate $i$ on the expenditure involved. If the result is positive, this is the present day value of the money earned by the project after having paid interest at a fractional rate $i$ on the money invested in the project.

The sum in the above equation begins with $j=0$. However, in practice, there is some flexibility in defining the 'present'. It may be the time at which the calculations are being made. In this case the summation of all cash flows starts at year 0. Some authors define the present as the moment when production starts, which is definitely some years later than the time of calculating the NPV of the project. The summation might start at $j=-3$ for example. The starting point should be defined to include all cash flows yet to occur in connection with the project, but cash already spent is excluded [2]. NPV is a sum of money. In engineering terminology, it is an extensive property. It is additive, which is an advantage when a company is considering a package of projects, and different combinations of the subprojects have to be evaluated. There are simple calculator and computer programs in the literature to assist in the comprehensive analysis of alternative projects with the help of the NPV method, such as those published by Neal [53].

*Discounted cash flow rate of return*

If we substitute Eq. (4.55) into Eq. (4.56), we get the expression:

$$\mathrm{NPV} = \sum_{j=0}^{n} \frac{\mathrm{ANCF}(j)}{(1+i)^j} \tag{4.57}$$

A particularly interesting solution to Eq. (4.57) is the value of $i$ which makes NPV zero, i. e. the root of the equation $\mathrm{NPV} = 0$. This value of $i$ is known as the discounted cash flow rate of return, DCFRR. Equation (4.57) can be solved either graphically or by any root-finding numerical method. This is the real interest rate earned by the investment throughout its useful life. A DCFRR of 12% means that 12%/yr will be earned on the investment in addition to generating a sufficient amount of money to repay the original investment. It lays emphasis on the profitability per unit investment (in the terminology familiar to chemical engineers, an 'intensive property' of the investment), and is considered as an index of good management [2]. Other names of DCFRR are: Internal rate of return (IRR), true rate of return, profitability index, investor's rate of return, interest rate of return or, in economic circles, simply the yield. At this point in our discussion a simple question may occur to the reader: has DCFRR anything to do with the simple rate of return? In other words, is there any relationship between them? Yes, there is. This problem was solved by Lindsley [42] for periodic annual cash flows. He developed a procedure to calculate DCFRR if we know the process useful life and the simple rate of return.

There are other simple calculator and computer programs in the literature which help in the comprehensive analysis of alternative projects with the help of the DCFRR method, e. g. those published by Neal [54] or Wild [38]. Holland [18] notes that NPV and DCRFF computations do not consider depreciation as a separate factor, it is regarded only as an allowance for the reduction of income tax.

Before introducing the next profitability assessment method, let us look at a typical net annual discounted cash flow curve over the useful life of project *(Fig. 4.5)*.

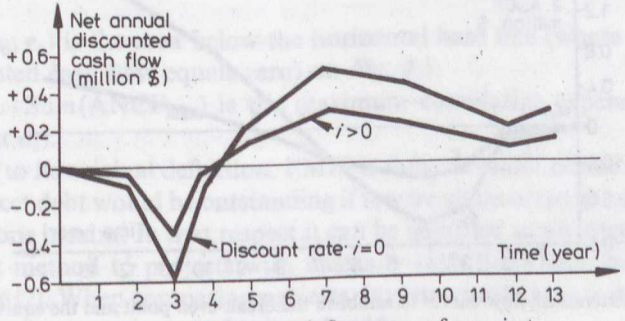



Fig. 4.5. Typical cash flow curve of a project

There are some typical stages through which each project passes. In the early stages of research and development the expenditure is relatively small. Once the detailed design is in progress, the rate of cash outflow starts to increase. This outflow of money accelerates further when items of equipment are purchased and construction work is in progress. There is a point in the time scale, the plant startup, when cash starts to flow in, although even after startup some additional capital expense may arise. During the steady operation of the plant cash flows in steadily.

The cash flow curve discussed above can be transformed by accumulating the cash flows and plotting the cumulated values on the $y$ axis. *Figure 4.6* shows typical cumulative discounted cash flow curves of a project (taken from Ref. [18]).

*Discounted break-even point*

*Table 4.5* lists process data — cash flows and discounted cash flows — for a typical project.

The data in *Table 4.5* are plotted in *Fig. 4.6.*

For the undiscounted case the appropriate cash flow curve is $a_0 b_0 c_0 d_0 e_0 f_0$. If the interest on capital (often referred to as cost of capital) is 10%, the appropriate cash flow curve is *abcdef*. The curve crosses the $x$ axis (the base line, where the cumulated discounted cash flow is zero) at point $e$. Up to this point, the capital is at risk. At this point the process has paid back its capital investment cost and produced the same return as if the same amount of principal had been put in a bank at a compound interest rate $i$. This point is called the discounted break-even point, DBEP. Beyond the break-even point the capital is no longer at risk, and the positive cumulative discounted cash flow above the base line indicates an excess return on the invested capital as compared to being invested on a fixed 10% interest rate.

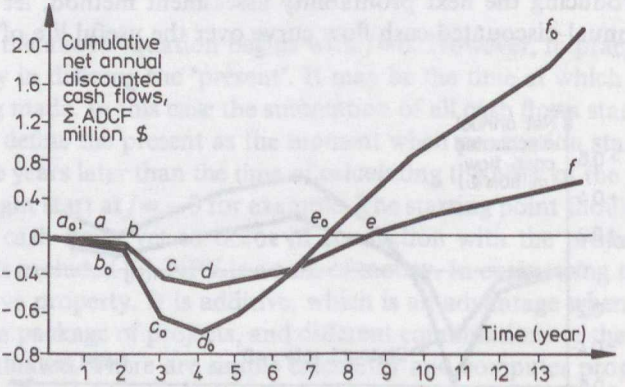



Fig. 4.6. Cummulative cash flow curves to establish the break-even point and the equivalent maximum interest period

Table 4.5. Cash flows and discounted cash flows for a typical project [18]

| Year | Undiscounted | | Discounted at 10% | | |
|---|---|---|---|---|---|
| | ANCF $ 1000 | Sum (ANCF) $ 1000 | *fd* | ADCF $ 1000 | Sum (ADCF) $ 1000 |
| 0 | − 10 | − 10 | 1.0 | − 10 | − 10 |
| 1 | − 30 | − 40 | 0.909 | − 27.27 | − 37.27 |
| 2 | − 60 | −100 | 0.826 | − 49.59 | − 86.86 |
| 3 | −750 | −850 | 0.751 | −563.48 | −650.34 |
| 4 | −150 | −1000 | 0.683 | −102.45 | −752.79 |
| 5 | +200 | −800 | 0.621 | +124.18 | −628.61 |
| 6 | +300 | −500 | 0.564 | +169.34 | −459.27 |
| 7 | +400 | −100 | 0.513 | +205.26 | −254.01 |
| 8 | +400 | +300 | 0.466 | +186.60 | − 67.42 |
| 9 | +360 | +660 | 0.424 | +152.68 | + 85.27 |
| 10 | +320 | +980 | 0.386 | +123.37 | +200.65 |
| 11 | +280 | +1260 | 0.350 | + 98.14 | +306.78 |
| 12 | +240 | +1500 | 0.319 | + 76.47 | +383.25 |
| 13 | +240 | +1740 | 0.290 | + 69.52 | +452.77 |
| 14 | +400 | +2140 | 0.263 | +105.33 | +558.10 |

Some authors suggest that a single number could never be enough to assess profitability. The shape of the cumulative cash flow and cumulative discounted cash flow curves, both before and after the break-even point, has to be taken into account.

*Equivalent maximum investment period (EMIP)*

Allen and Page [35] have introduced a quantity to account for the shape of the undiscounted cumulative cash flow curve up to the break-even point:

$$\text{EMIP} = \frac{\text{Area}(a_0, e_0)}{-\text{Sum}(\text{ANCF}_{\max})} \text{ (years)} \tag{4.58}$$

where area$(a_0, e_0)$ is the area below the horizontal base line (where the cumulated undiscounted cash flow equals zero) on *Fig. 4.5.*

The term $-\text{Sum}(\text{ANCF}_{\max})$ is the maximum cumulative expenditure on the project (point $d_0$).

According to its original definition, EMIP is the equivalent period during which the total project debt would be outstanding if it were all incurred at one instant and all repaid at one instant. In that respect it can be regarded as an attempt to extend the pay-back method to projects with unsteady cash flows with more than one capital outlay [2]. When comparing projects, investments having a shorter equivalent maximum investment period are preferable.

Holland [18] came to the conclusion that if the required capital investment were to be increased, it would be necessary to increase the rate of return after startup for the EMIP to remain the same. Also, if the development and construction stages are prolonged, i.e. the period with negative cash balance is longer, then this can be balanced by an increased rate of return. This fact underlines the role of time in project development and construction.

*Interest recovery period (IRP)*

The same mode of thought which resulted in the introduction of EMIP can be applied to the cumulative cash flow curve beyond the break-even point. The time period during which the area above the horizontal base line (where the cumulated discounted cash flow is zero) is equal to the area below it is called the interest recovery period (IRP). We can illustrate this idea on the cumulative cash flow curve of *Fig. 4.5*, where Area$(a_0e_0)$ has to be equal to Area$(e_0f_0)$. The time coordinate of $f_0-e_0$ is called interest recovery period (IRP).

*Capital rate-of-return ratio (CRR)*

The NPV value is the answer to the question of how much money can be generated by the investment. The DCFRR value seeks an answer to the question: how effectively is the committed capital used? However, these questions are academic if the company in question is unable to raise the money to undertake the project. Therefore, the maximum net outlay during the project's life is a question of equal importance. The capital rate-of-return ratio takes this factor into account. It is defined as the net present value, NPV, divided by the maximum cumulative expenditure or maximum net outlay, $-\text{Sum}(\text{ACF}_{\max})$ [18]:

$$\text{CRR} = \frac{\text{NPV}}{-[\Sigma(\text{ANCF})]_{\max}} \tag{4.59}$$

According to Holland [18], the CRR method is being increasingly used. However, his definition of CRR is somewhat strange, because NPV is a sum of discounted annual cash flows, while the maximum net outlay is calculated from undiscounted cash flows. A similar quantity, called the Benefit/Cost Ratio is used by Rose [2], with the discounted capital involved (over the project life) in the denominator. In this latter definition nominator and denominator are both discounted with the same discount rate.

### 4.1.3 Special aspects of profitability calculations

*Uncertainty and risk*

No profitability measure using discounted or undiscounted cash flows, e.g. discounted cash flow rates of return (DCFRR) or net present value (NPV) for future investments can ever be calculated accurately. These values are only predictions, because cash flow data for future projects are subject to uncertainty [21]. Sales, costs, prices and inflation all may vary considerably from estimated values.

Therefore, when stating predicted values of DCFRR and NPV for future investments, it is also necessary to give a correct measure of confidence in the predictions.

In addition to the technical risk in a new project, there is always the risk of choosing the wrong cash flow data. The criterion of a good decision is that it remains acceptable even if the estimates prove to be in error [2]. Statistical and probability techniques are used in the profitability analyses to account for risk.

Throughout the discussion in this chapter uncertainty and risk have been used synonymously. In a more exact treatment of the subject, risk is used to describe situations for which there are probability data. Past data can be used to predict the chance of certain things. In contrast, uncertainty refers to situations where there are no suitable past data on which to base an estimate of the chance of an occurrence. The sales of a new product, for instance, would be regarded as uncertain [2].

Rose discussed the problem of engineering investment decisions under uncertainty in full detail. We will discuss this problem in less detail, based mainly on the papers of Holland *et al.* [21–24].

The application of statistical methods to the assessment of uncertainty means that, knowing the distribution of some primary variables of the profitability calculations (e.g. the distribution of the sales volume, sales prices, capital investment, etc.), we may apply rules to predict the distribution of the profitability measures, such as DCFRR or NPV.

In this context 'knowing the distribution' means that we have some predicted values of the variable and know the type of probability distribution which apply to the predicted values. In some cases we know only the range of the variable and we assume that the variable follows some known probability distribution.

The most often used probability distribution is the Gaussian or normal distribution. Even when little is known about the true probability distribution of a variable, estimates of profitability can be made if the variables are assumed to be normally distributed. This assumption is often used as an approximation when it is known that the distribution is not normal, because only the normal (and Poisson) distributions are amenable to relatively simple mathematical manipulations [21]. The normal distribution has two parameters: the mean, $\mu$ and the standard deviation, $\sigma$.

The equation for the standardized Gaussian distribution is:

$$f(z)=\frac{\exp(-z^2/2)}{\sqrt{2\pi}} \tag{4.60}$$

where $z$ is the standard score with a mean value of zero and a standard deviation equal to unity. The standard score corresponding to a value of $x$ is:

$$z=(x-\mu)/\sigma \tag{4.61}$$

The shape of the curve of the Gaussian distribution is single-peaked and symmetrical about $z=0$ ($x=\mu$), and assymptotic to $f(z)=0$ at $z=\pm$infinity.

Equation (4.60) can be integrated between limits to determine the probability that the random value lies between the selected limits. Extensive tables of $f(x)$ and the associated integral are available in the literature. Most sources quote the original work of Fisher and Yates [61].

Table 4.6. Selected values from a standardized normal probability table

| Standard score | Ordinate of normal curve | Area under right-hand side of curve |
|---|---|---|
| 0.000 | 0.3989 | 0.00000 |
| 0.100 | 0.3970 | 0.03983 |
| 0.200 | 0.3910 | 0.07926 |
| 0.253 | 0.3864 | 0.10000 |
| 0.400 | 0.3683 | 0.15542 |
| 0.524 | 0.3478 | 0.20000 |
| 0.800 | 0.2897 | 0.28814 |
| 0.842 | 0.2798 | 0.30000 |
| 1.000 | 0.2420 | 0.34134 |
| 1.200 | 0.1942 | 0.38493 |
| 1.282 | 0.1953 | 0.40000 |
| 1.645 | 0.1032 | 0.45000 |
| 1.960 | | 0.47500 |
| 2.000 | 0.0540 | 0.47725 |
| 2.576 | 0.0145 | 0.49500 |
| 3.0000 | 0.0044 | 0.49865 |
| 4.000 | 0.0001 | 0.49997 |

*Table 4.6* contains an abbreviated version of the standardized normal probability distribution table.

To illustrate the application of statistical methods to the estimates of profitability, we shall use an example in which the economic parameters conform to the normal (Gaussian) probability distribution. The procedure is the same in cases when other (known) probability distributions apply.

The way in which these types of calculations are performed may be summarized as follows (using the NPV and DCFRR calculations to illustrate the procedure):

1. Collect any information available about primary variables which can be easily estimated. Information about sales revenue, annual total expenses and estimated fixed capital cost is required.
2. Develop equations for annual cash flows, ANCF, and annual discounted cash flows, ADCF, retaining the estimated primary variables in explicit form in the equations.
3. Calculate mean values and variances of the primary variables.
4. Develop equations to calculate the mean value and variance of the profitability measure and calculate them.
5. Knowing the two parameters of the normal probability distribution, calculate and plot the probability distribution of the profitability measure.

**Example 4.3**

Determine the probability distribution of NPV and DCFRR values for the project characterized by the following data and estimated values assuming that all variables have uncertainties following the Gaussian distribution [22]:

Ad (1): Data and estimates

| | |
|---|---|
| Sales volume, R: | 13,000–17,000t/yr |
| Sales price, cs: | \$295–\$305/t |
| Cost of raw materials, cr: | \$135/t |

Estimates of fixed capital investment, CFC from five independent sources:

\$3.85 million, \$4.2 million, \$3.7 million, \$4.25 million, \$4.5 million

| | |
|---|---|
| Working capital, CWC: | 0.12CFC |
| Labour requirement: | 2 men/3 shift |
| Labour rate, cl: | \$5.2/h |

Annual processing costs follow the equation:

$$AP = 0.09CFC + 13\,500cl \times N$$

| | |
|---|---|
| Annual general expenses: | \$450,000/yr |
| Plant service life: | 5 years |
| Depreciation: | straight-line |
| Cost of capital | 10%/yr |
| Tax allowance, AA | 0 |
| Income tax rate, $t$ | 0.45 |
| Salvage value, S | \$0 |

Ad (2): We consider the sales volume, sales price and the fixed capital investment as primary variables and develop equations for the net annual cash flow, ANCF and net annual discounted cash flow, ADCF:

We know from Eq.(4.4) that

$$ANCF = ANCI - ATC$$

Substituing Eq.(4.2) into Eq.(4.4) we have:

$$ANCF = ACI - AIt - ATC$$

Using Eq.(4.1) to calculate As and Eq.(4.3) to calculate income tax, AIt, we get the expression:

$$ANCF = ACI - (ACI - AD - AA)t - ATC$$

From Eq.(4.5):

$$ATE = AGE + AME$$

and for AME we have the expression:

$$AME = AP + R \times cr, \text{ therefore}$$

$$ATE = AGE + AP + R \times cr,$$

and using the equation proposed for the annual processing costs:

$$ATE = AGE + R \times cr + 0.09CFC + 13{,}500cl \times N \tag{4.62}$$

The depreciated values are calculated using the relation:

$$ADCF = ANCF \times fd \tag{4.63}$$

where *fd* is the discount factor, $1/(1+i)^n$.

Substituting the known data into the equations:

$$AP = 0.9CFC + 13{,}500(5.20) \times (2) \text{ results}$$

$$AP = 0.09CFC + 140{,}400$$

Substituting this result, the value of the annual general expenses, \$450,000/yr and the cost of raw materials, \$135/t into Eq.(4.62) gives:

$$ATE = 590{,}400 + 0.09CFC + 135R, \$/yr$$

The total capital cost using Eq.(4.15):

$$CTC = CFC + CWC + CL = CFC + .12CFC + 0$$

since the cost of land is zero.

The annual depreciation, AD, using the straight-line depreciation method, 5 years' useful life and zero scrap value:

$$AD = (CFC - 0)/5 = 0.2CFC, \$/yr$$

The sales income will be calculated using the relation $As = R \times cs$; therefore, using the expression for ATE developed above:

$$ACI = R(cs - 135) - 590{,}000 - 0{,}09CFC \tag{4.64}$$

and

$$AIt = 0.45(R(cs - 135) - 590{,}000 - 0.09CFC) - 0.45 \times AD, \$/yr$$

As the final equation for the net annual cash income we get the following expression:

$$ANCF = 0.55R(cs - 135) - 324{,}720 - 0.0595CFC + 45AD - ATC \quad (4.65)$$

Depreciation, AD, and annual capital expenditure are changing during the life of the project.

For Year 0, ACI = 0, AIt = 0, AD = 0 and ATC = CTC = \$1.12CFC/yr, and Eq.(4.65) takes the form:

$$ANCF(0) = 1.12CFC, \$/yr$$

For Years 1, 2, 3, 4 AD = 0.2CFC and ATC = 0 \$/yr. Therefore:

$$ANCF(I) = 0.55R(cs - 135) - 324{,}720 + 0.0405CFC, \$/yr \; i = 1, 2, 3, 4$$

For the final year of the project, AD = 0.2CFC and the working capital, CWC is recovered, therefore ATC = −0.12CFC. Using these values:

$$ANCF(5) = 0.55R(cs - 135) - 324{,}720 + 0.1605CFC, \$/yr$$

These are the expressions developed for the net annual cash flow, ANCF, with the parameters R, cs and CFC in the equations.

By substituting these results into Eq.(4.63) similar equations can be developed for the net annual discounted cash flow, ADCF. They are summarized in *Table 4.7*.

Ad (3): The mean values of the variables are:

$$R = (13{,}000 + 17{,}000)/2 = 15{,}000 \text{ t/yr}$$

$$cs = (295 + 305)/2 = 300 \text{ \$/t}$$

$$CFC = (4.2 + 3.85 + 4.5 + 3.7 + 4.25)/5 = \$4.1 \text{ million}$$

The standard deviations of the variables are:

Table 4.7. Equations for ADCF in Example 4.3

| Year | Expression for ADCF |
|---|---|
| 0 | $-1.12CFC$ |
| 1 to 4 | $(Y + 0.0405CFC) \sum_{n=1}^{n=4} \frac{1}{(1+i)^n}$ |
| 5 | $(Y + 0.0405\,CFC) \frac{1}{(1+i)^5}$ |
| where | $Y = 0.55R\,(cs - 135) - 324,\,720$ |

CFC, using the mean value of \$4.1 million and the definition of variance:

$$\sigma \cong S = \left[\frac{\Sigma(X_i - X)2}{n-1}\right]^{0.5}$$

$$\sigma(\text{CFC}) = \$322{,}100$$

For sales volume and sales price we have only the range of the estimated values. Therefore, we have to calculate their variance based on the fact that they are normally distributed around their mean value.

The cumulative probability of a normally distributed variable lying within four standard deviations of the mean is 0.49997 (see *Table 4.6*). Therefore, there is more than 99.99% certainty that the variable will fall within $\pm 4\sigma$ from the mean. In practice $\sigma$ may be taken as one eighth of the range.

$$\sigma(R) = (17{,}000 - 13{,}000)/8 = 500 \text{ t/yr}$$

$$\sigma(cs) = (305 - 295)/8 = \$1.25/\text{t}$$

Ad (4): Summing up the net annual discounted cash flow values in *Table 4.7*, we have an expression for the net present value of the project:

$$\text{NPV} = 2.0849\,\text{R}(\text{cs} - 135) - 0.89196\,\text{CFC} - 1{,}230{,}944 \qquad (4.66)$$

Substituting the mean values of the variables into Eq.(4.66), we have the mean value of the NPV:

$$\overline{\text{NPV}} = \$272{,}150$$

We follow the same procedure when calculating the variance of NPV.

Since NPV is a function of the three variables R, cs and CFC, we need some simple rules to enable us to calculate the variance of a function based on the variances of its variables. These simple rules are:

*Rule No. 1*

If a variable, $y$ is a function of several variables $x_1$, $x_2$, etc., we can write:

$$y = f(x_1, x_2, \ldots, x_n) \qquad (4.67)$$

If the variables $x_1, x_2, \ldots, x_n$, are independent of each other, and follow the normal probability distribution, then the variance of $y$:

$$\sigma_y^2 = \left(\frac{\partial y}{\partial x_1}\right)^2 \sigma_{x_1}^2 + \left(\frac{\partial y}{\partial x_2}\right)^2 \sigma_{x_2}^2 + \ldots + \left(\frac{\partial y}{\partial x_n}\right)^2 \cdot \sigma_{x_n}^2 \qquad (4.68)$$

where $\sigma_y^2$ is the variance, or the square of the standard deviation, of variable $y$, and $\sigma_{x_n}^2$ is the variance of $x_n$.

*Rule No. 2*

If the above function is linear, i.e.:

$$y = a_1 x_1 + a_2 x_2 + \ldots\ldots a_n x_n \qquad (4.69)$$

where all $a$'s are constants, Eq.(4.68) takes the following form:

$$\sigma_y^2 = a_1\sigma_{x_1}^2 + a_2\sigma_{x_2}^2 + \ldots + a_n\sigma_{x_n}^2 \tag{4.70}$$

*Rule No. 3*

If the function is a product of the independent variables, i.e.:

$$y = x_1 \times x_2 \tag{4.71}$$

then Eq.(4.68) can be written in the form:

$$\frac{\sigma_y^2}{y^2} = \frac{\sigma_{x_1}^2}{x_1^2} + \frac{\sigma_{x_2}^2}{x_2^2} \tag{4.72}$$

In Eq.(4.66) there are two variable terms:

$$2.0849\,\mathrm{R}(\mathrm{cs}-135) \quad \text{and} \quad -0.89196\,\mathrm{CFC}$$

The first term contains a product of two independent variables, R and cs, and the variance of the first term can therefore be calculated using Eq.(4.72) as:

$$\sigma_{\mathrm{R}\,(\mathrm{cs}-135)}^2 = (\overline{\mathrm{cs}}-135)^2\sigma_{\mathrm{R}}^2 + \overline{\mathrm{R}}^2\sigma^2(\mathrm{cs}-135)$$

and substituting the mean values and variances calculated in the previous step,

$$\sigma_{\mathrm{R}\,(\mathrm{cs}-135)}^2 = (165)^2\,(500)^2 + (15{,}000)^2\,(1.25)^2$$

$$\sigma_{\mathrm{R}\,(\mathrm{cs}-135)} = \$84{,}604/\mathrm{yr}$$

The equation for NPV in this example is linear in terms of R(cs − 135) and CFC with constant parameters, $a_1 = 2.0849$ and $a_2 = -0.89196$, therefore Eq.(4.71) can be applied:

$$\sigma_{\mathrm{NPV}}^2 = (2.0849)^2\,(84{,}604)^2 + (-0.89196)^2\,(322{,}100)^2$$

$$\sigma_{\mathrm{NPV}} = \$337{,}128$$

To calculate the mean and variance of discounted cash flow rate of return, DCFRR, we rearranged the equation for NPV to give the form:

$$\mathrm{NPV} = \mathrm{Y}\frac{(1+i)^{n+1}-1}{i(1+i)^n} - 1 + \mathrm{CFC}\sum_{k=0}^{n}\frac{K}{(1+i)^k} \tag{4.73}$$

where $n = 5$; $k = -1.12$ for year 0, 0.0405 for years 1 to 4, and 0.1605 for year 5, and

$$\mathrm{Y} = 0.55\mathrm{R}(\mathrm{cs}-135) - 324{,}720$$

The discounted cash flow rate of return, DCFRR, is the interest rate $i$ which makes NPV zero. This interest rate was determined for the mean value of the independent variables by an iterative procedure, and was found to be 12.19%.

To determine the variance of DCFRR, we first substitute NPV = 0 into Eq.(4.73) and rearrange it as follows:

$$\frac{\sum_{k=0}^{n} K/(1+i)^k}{\frac{(1+i)^{(n+1)}-1}{i(1+i)^n}-1} = \frac{0.55\mathrm{R}(\mathrm{cs}-135)-324{,}720}{-\mathrm{CFC}}$$

After differentiating both sides of the above equation, and substituting the corresponding numerical values ($n=5$, $i=\mathrm{DCFRR}=12.19\%$, and mean values of the independent variables) into it, we have the linear relationship:

$$\mathrm{d}i = \left(\frac{-0.680296}{4{,}100{,}000}\right)\mathrm{d}[\mathrm{R}(\mathrm{cs}-135)] + \left(\frac{0.312703}{4{,}100{,}000}\right)\mathrm{d}\mathrm{CFC}$$

where $i=\mathrm{DCFRR}$.

Therefore, the variance of DCFRR is related to the variance of R(cs − 135) and the variance of CFC by Eq.(4.70):

$$\sigma^2_{\mathrm{DCFRR}} = \left[\frac{-0.680296}{4{,}100{,}000}\right]^2 \sigma^2_{\mathrm{R(cs-135)}} + \left(\frac{0.312703}{4{,}100{,}000}\right)^2 \sigma^2_{\mathrm{CFC}}$$

Substituting the values of the variances of R(cs − 135) and CFC, the above equation yields the standard deviation:

$$\sigma_{\mathrm{DCFRR}} = 2.828\%$$

We can summarize the calculated results as:

| | Mean | Standard deviation |
|---|---|---|
| NPV | \$ 272,150 | \$ 337,128 |
| DCFRR | 12.19% | 2.828% |

Ad (5): As the last step, we can calculate the probability of attaining a specified NPV (or DCFRR), based on the assumption that both are normally distributed about their mean values. The calculation is illustrated on the example of NPV.

For NPV the standard score can be written as:

$$z = (\mathrm{NPV} - \overline{\mathrm{NPV}})/\sigma_{\mathrm{NPV}}$$

After substituting the appropriate values and rearranging, we get the expression:

$$\mathrm{NPV} = 337{,}128z + 172{,}150 \quad (\$)$$

Table 4.8. Cumulative probability of obtaining a given NPV and DCFRR or less

| Cumulative probability | Standard score | NPV $ | DCFRR %/yr |
|---|---|---|---|
| 0 | −4.000 | −1,076,362 | 0.88 |
| 2 | −2.054 | − 420,311 | 6.38 |
| 2.25 | −2.000 | − 402,106 | 6.65 |
| 10 | −1.282 | − 160,048 | 8.56 |
| 20 | −0.842 | − 11,712 | 9.81 |
| 30 | −0.524 | + 95,495 | 10.71 |
| 40 | −0.253 | + 186,857 | 11.47 |
| 50 | 0.000 | + 272,150 | 12.90 |
| 60 | 0.253 | + 357,443 | 12.91 |
| 70 | 0.524 | + 448,805 | 13.67 |
| 80 | 0.842 | + 556,012 | 14.57 |
| 90 | 1.282 | + 704,348 | 25.82 |
| 98 | 2.054 | + 964,611 | 18.00 |
| 99.998 | 4.000 | +1,620,662 | 23.50 |

We can now use *Table 4.8* to calculate the cumulative probability of a given NPV for year 5.

For example, if we choose $z=-2$, this means that we look for the probability of attaining an NPV value which is less than the mean value minus two times the standard deviation. The NPV value for $z=-2$ is

$$\text{NPV} = \$-402{,}106$$

and the cumulative probability of an NPV value less then −402,106 in the fifth year is about 2.25%.

Also, since the probability curve is symmetric, there is a $(100-2.25)=97.75\%$ probability that a $z$ value of less than 2 will be obtained. That means a net present value of $946,406. Other values calculated in the same way are shown in *Table 4.8* and depicted in *Fig. 4.7* for both profitability measures, NPV and DCFRR.

### Sensitivity analysis

Sensitivity analysis is a term familiar to systems engineers. Its aim is to determine to which input variables the response of the system is most sensitive. Sensitivity analysis is applied in cost engineering, because profitability is estimated on the basis of predicted costs and cash flows which are subject to error. In this case the system is the project under consideration, the input variables are the predicted parameters of the cash-flow calculations, and the response is the profitability of the project, as measured by some profitability assessment method. For example, some

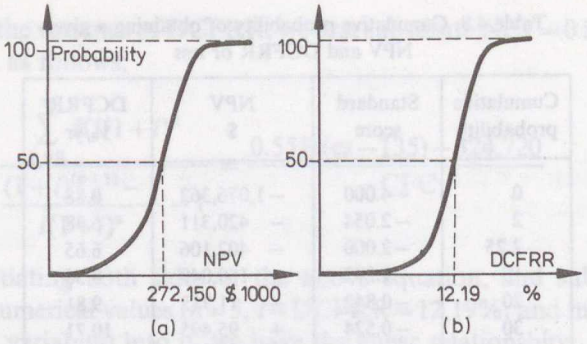



Fig. 4.7. Profitability of a given NPV of less (a) and of a given DCFRR of less (b)

of the possible questions that may be answered by sensitivity analysis are the following:

— if the sales revenue is 10% lower than predicted, what is the reduction in the calculated net present value, NPV, and in the discounted cash flow rate of return?

— if the fixed capital expenditure is 15% higher than originally estimated (this is a very common and everyday situation), what is the effect of this cost escalation on the payback period, PB, or on the equivalent maximum investment period, EMIP?

These problems can be analysed using the basic cost equation which show the interrelation of cost and cash-flow elements. We will show some examples to illustrate the development of sensitivity analysis methods. These examples demonstrate:

(*i*) the effect of a decrease in the sales revenue on the NPV of the project;
(*ii*) the effect of an increase in total expenditure, ATE, on the rate of return on original investment, based on annual cash income before tax, ACI, and annual cash income after tax, ANCI; and
(*iii*) the effect of an increase in fixed-capital investment, CFC, on the rate of return on original investment, based on annual cash income before tax, ACI, and annual net cash income, ANCI.

*The effect of a decrease in sales revenue [19]*

We have learned from Eq. (4.1) that the annual cash income is the difference between the annual sales income, As, and the annual cost of production and sale, ATE:

$$\mathrm{ACI} = \mathrm{As} - \mathrm{ATE}$$

Therefore, if the revenue from annual sales falls bellow the predicted value As by an amount dAs, the annual cash income is reduced by the same amount:

$$dACI = dAs \tag{4.74}$$

In view of Eq. (4.3), the reduction in cash income also reduces the annual income tax, AIt:

$$dAIt = dACI \times t \tag{4.75}$$

where t is the tax rate.

Combining Eq. (4.74) with Eq. (4.75) we get:

$$dAIt = dAs \times t \tag{4.76}$$

The net annual cash flow after tax is related to the annual cash income and the income tax by:

$$ANCF = ACI - AIt - ATC$$

where ATC is the annual capital charge, which is not necessarily zero after the first year.

Therefore, the effect of a reduction in the annual sales revenue affects the net annual cash flow by:

$$dANCF = dACI - dAIt$$

which can be expressed as:

$$dANCF = dAs\ (1-t) \tag{4.77}$$

The annual discounted cash flow, ADCF, can be calculated using Eq. (4.54):

$$ADCF = ANCF \times fd$$

where $fd$ is the discount factor, $1/(1+i)^n$.

Therefore,

$$dADCF = dANCF \times fd \tag{4.78}$$

Combining Eq. (4.77) with Eq. (4.78), we get:

$$dADCF = dAs\ (1-t) \times fd \tag{4.79}$$

and since the net present value, NPV, is the cumulative annual discounted cash flow,

$$dNPV = \text{Sum}\ (dADCF) \tag{4.80}$$

Combining this result with Eq. (4.79) we have:

$$dNPV = \text{Sum}\ (dAs\ (1-t) \times fd) \tag{4.81}$$

*The effect of an increase in total expense*

We have learned from Eq. (4.29) that the rate of return (on original investment) based on the annual cash income before tax is the ratio of the annual cash income, ACI, and the original total invested capital, CTC:

$$i = \text{ACI}/\text{CTC}$$

Similarly, according to Eq. (4.30), the rate of return (on original investment) based on the annual net cash income (after tax) is the ratio of the annual net cash income, ANCI, and the original total invested capital, CTC:

$$i = \text{ANCI}/\text{CTC}$$

If the total expense increases with dATE, the annual cash income decreases by the same amount. Therefore

$$-\text{dACI} = \text{dATE} \tag{4.82}$$

The effect of a decrease in annual cash income by an amount dACI is a reduction in the annual income tax by dAIt (see Eq. (4.75)), where

$$\text{dAIt} = \text{dACI} \times \text{t} \tag{4.83}$$

and applying Eq. (4.82), we have:

$$\text{dAIt} = -\text{dATE} \times \text{t}$$

According to Eq. (4.2), the annual net cash income, ANCI, is the difference between the annual cash income, ACI, and the annual amount of tax, AIt:

$$\text{ANCI} = \text{ACI} - \text{AIt}$$

Therefore, the change in annual net cash income, dANCI, as a consequence of an increase in annual total expenses, dATE, is:

$$\text{dANCI} = \text{dACI} - \text{dACI} \times \text{t} = \text{dACI}\ (1-\text{t}), \tag{4.84}$$

and using Eq. (4.82), we obtain the following expression:

$$\text{dANCI} = -\text{dATE}\ (1-\text{t}) \tag{4.85}$$

The expressions developed to describe the effect of increase in annual total expenses on annual cash income, ACI and annual net cash income, ANCI can be substitued into Eqs (4.29) and (4.30) to express the effect of increase of total expenses on the rate of return. The change in rate of return (based on cash income before tax), d$i$ is:

$$\text{d}i = \text{dACI}/\text{CTC} = -\text{dATE}/\text{CTC} \tag{4.86}$$

and similarly, for the rate of return based on net annual cash income:

$$\text{d}i = \text{dANCI}/\text{CTC} = -\text{dATE}(1-\text{t})/\text{CTC} \tag{4.87}$$

The negative sign means that the increase in total annual expense results in a decrease in the rate of return ratio. This reduction is less on an after-tax basis since the income tax is smaller.

*The effect of an increase in fixed-capital investment*

We learned from Eq.(4.15) that the total capital cost of a project is the sum of three terms: the fixed capital cost, CFC, the working capital demand, CWC and the cost of land and other depreciable costs, CL. Therefore, the primary effect of any increase in fixed-capital cost is an increase of the same amount in the total capital cost:

$$\mathrm{dCTC} = \mathrm{dCFC} \tag{4.88}$$

The increased total capital cost is equal to CTC + dCFC. Supposing that ACI, i.e. the cash income, is the same, the value of the rate of return with the increased fixed capital cost can be described by the expression:

$$i' = \mathrm{ACI}/(\mathrm{CTC} + \mathrm{dCFC}) \tag{4.89}$$

The rate of return decreases from $i$ to $i'$ by an amount d$i$. Therefore

$$\mathrm{d}i = i' - i = \mathrm{ACI}/(\mathrm{CTC} + \mathrm{dCFC}) - \mathrm{ACI}/\mathrm{CTC} \tag{4.90}$$

Rearranging and simplifying Eq.(4.90), we obtain the expression

$$\mathrm{d}i = \frac{-\mathrm{ACI} \times \mathrm{dCFC}}{(\mathrm{CTC} + \mathrm{dCFC}) \times \mathrm{CTC}} \tag{4.91}$$

If we use net annual cash income the picture is slightly more complicated. Since Eq.(4.2) still holds, we have to calculate the effect of increased fixed capital investment on the annual income tax.

For the sake of simplicity, we suppose that the depreciation allowance for tax purposes is the same as the book-value depreciation, and it is calculated by the straight-line depreciation method:

$$\mathrm{AD} = (\mathrm{CFC} - \mathrm{S})/\mathrm{s}$$

where S is the scrap value, and
s is the service life of the project.

The depreciation for the increased fixed capital investment, AD′, is:

$$\mathrm{AD}' = (\mathrm{CFC} + \mathrm{dCFC} - \mathrm{S})/\mathrm{s} \tag{4.92}$$

and therefore

$$\mathrm{dAD} = \mathrm{AD}' - \mathrm{AD} = \mathrm{dCFC}/\mathrm{s} \tag{4.93}$$

This increase in depreciation reduces the income tax payable by an amount equal to dAD × t. If the original value of the rate of return based on net annual cash income is denoted by $i$ and the new value by $i'$, the change in rate of return, d$i$, can

be described by the following expression:

$$di = i' - i = \frac{ANCI + dCFC/s \times t}{CTC + dCFC} - \frac{ANCI}{CTC} \tag{4.94}$$

This expression cannot be simplified very much. The term dCFC appears in both the numerator and denominator, but in the numerator with a multiplier, t/s. In the special case when t/s = ANCI/CTC, the change of CFC has no effect on the value of the rate of return on the original total investment based on net annual cash income.

*Inflation*

Inflation is the tendency for the buying power of money to decrease. If the rate of inflation for the coming years could be forecast with certainty, then these values should be incorporated in profitability calculations. However, this is not the case. An engineer can hardly predict that the price of a raw material such as oil will increase (or decrease) by 200–300% in a few months. But such rapid and unexpected changes in the cost data result in misleading, erroneous decisions. The only way to overcome this difficulty is to perform sensitivity analysis on the basis of inflation.

There are at least two basic methods by which inflation may be accounted for [3]:

— In preliminary studies it is often enough to incorporate inflation using an 'effective discount factor', which is defined for year $n$ as

$$fd' = (1 + ic)(1 + ii)^{-n} \tag{4.95}$$

where $ic$ is the cost of capital and $ii$ is the rate of inflation.

When this method is justified, $ic$ and $ii$ are both small compared to unity, so that their products and powers may be neglected. In this special case the effective discount factor is simply the sum of $ii$ and $ic$. This method is recommended for initial use only. The most attractive investment proposals have to be analysed by more detailed methods which take the different and varying inflation rates in each component of the selling price into account.

— In detailed economic studies more comprehensive methods are recommended, which allow for different inflation rates in each component of costs and prices.

## 4.2 Capital cost estimation

Capital cost is a fixed amount of money which will be tied up and risked in a project or enterprise. In the process industries it represents the investment needed to build and operate the process plant in order to manufacture a specific product [15].

The total capital cost, CTC, of a project consists of:

— fixed capital cost, DFC,
— working capital cost, CWC; and
— cost of land and other non-depreciable costs, CL, according to Eq.(4.15):

$$CTC = CFC + CWC + CL$$

The term 'capital cost estimation' is often used in a narrower sense: i.e. the estimation of fixed capital cost only.

The cost of land and other non-depreciable costs is a different matter, because cost of land may vary considerably depending on a variety of specific factors. There is no generally accepted method for the estimation of this item.

### 4.2.1 Estimation of fixed capital investment

Fixed capital investment is the amount of money tied up in depreciable facilities. The purpose of the estimate and the time and money spent on it, and consequently its accuracy, are closely correlated. Weinberger classified the methods of fixed capital cost estimation into three main categories [5]:

— Ratio methods, based on production capacity or sales volume;
— flowsheet-plus-factor methods, based on process flowsheets; and
— equipment-list-plus-factor methods based on equipment lists together with flowsheets.

The American Association of Cost Engineers has listed five types of capital cost estimates [25]:

| Type | Probable accuracy |
|---|---|
| Order of magnitude (ratio estimate) | +40% |
| Study (factored estimate) | +25% |
| Preliminary (budget authorization estimate) | +12% |
| Definite (project control estimate) | +6% |
| Detailed (firm estimate) | +3% |

However, independent of the type and accuracy of the method used, there are two factors in capital cost estimating to be considered, even if capital costs of similar plants are compared:

— time (date) of estimation; and
— location of the plant.

In this context similarity means that both plants produce the same product, by the same process, from the same (or very similar) raw materials.

The effect of time is taken into account by means of different cost indices. Cost indices express the relative cost of an item or of a whole plant in terms of the cost at a particular base time. Many different cost indices have been regularly published in the literature. The most useful ones in chemical engineering cost estimation are the following:

— Chemical Engineering Plant Cost Index;
— Marshall and Stevens Equipment Cost Index;
— Engineering News Construction and Building Index;
— Nelson Refinery Construction Cost Index.

The CE Plant Cost Index and the M&S Equipment Cost Index are compiled in the periodical Chemical Engineering on the page facing the inside back cover of each issue. The Nelson Refinery Construction Index appears in the Oil & Gas Journal, and the ENR Construction and Building Indices appear in Engineering News-Record. The above indices are calculated from various price indices published by the Bureau of Labor Statistics of the U.S. Department of Labor, which cover the prices of many products, the hourly earnings of workers in many branches of industry, salary trends for engineers, draftsmen, clerical and administrative staff, etc. [25]. (The original composition of the CE Plant Cost Index and the explanation of how it was put together appeared in 1963 [52]. Since then several papers have analysed the trends in CE Plant Cost Index (see refs [47–51]). The calculation method was revised in 1982 [55]).

*Figure 4.8* shows two graphs representing the change of the CE Plant Cost Index and the M&S Equipment Cost Index up to 1985.

An easy way to recalculate plant investment figures from a given base year to any other date (CFC(2)) is to multiply the fixed capital cost given for the base year (CFC(1)) by the ratio of the cost indices:

$$\mathrm{CFC}(2) = \mathrm{CFC}(1) \times \frac{\text{Cost Index}(2)}{\text{Cost Index}(1)} \tag{4.96}$$

There is no straightforward matter to assess the effect of location on capital investment costs, although geographical variations in capital costs can be quantified by means of a factor that is the ratio of capital costs between two locations. As one of the major chemical and petrochemical centers of the world, the U.S. Gulf Coast is often the reference base for plant construction costs.

The significant factors in creating differences between similar plants in different locations are [44]:

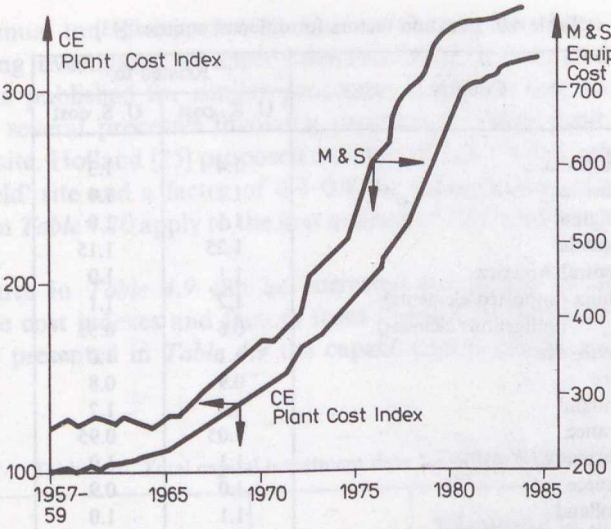



Fig. 4.8. Indexes to measure cost escalation in the chemical and process industries: (a) CE Plant Cost Index, (b) M & S Equipment Cost Index

— Influences of climate on foundations, building and equipment design;
— Labour rates and productivity;
— Indirect construction costs;
— Material and equipment costs, including the effect of tariffs, sales taxes, and rates of currency exchange;
— The extent to which material and equipment is indigenous or imported;
— Economic environment, including such factors as availability of labour, and general market conditions;
— Special factors, such as extensive overtime, travelling and hiring expenses, infrastructure of the specific area, etc.

Miller [44] has provided a method of estimating construction costs based on the seven factors listed above. He presents a detailed procedure for estimating a location factor for a plant at a new site, based on data for an analogous plant at another site. His method applies only to the U.K., Canada and U.S. Gulf Coast, with the location factors for these three sites being 1.0, 1.3 and 1.15, respectively.

Bridgewater [43] has revised Miller's method and extended the use of the location factor to several other countries. Location factors for chemical plants of similar functions (extracted from Bridgewater [43]) are listed in *Table 4.9*.

The application of location factors may also involve changing the time base of costs by use of a cost index. The cost index must be relevant to the country of location and the exchange rate has to be the one in effect at the data of conversion. Bridgewater came to the conclusion that location factors were independent of time if both inflation and parity exchange rates were considered.

Table 4.9. Location factors for different coutries [43]

| Country | Related to | |
|---|---|---|
| | U. K. cost | U. S. cost |
| Australia | 1.4 | 1.3 |
| Austria | 1.1 | 1.0 |
| Belgium | 1.1 | 1.0 |
| Canada | 1.25 | 1.15 |
| Central America | 1.1 | 1.0 |
| China (imported elements) | 1.2 | 1.1 |
| (indigenous element) | 0.6 | 0.55 |
| Denmark | 1.1 | 1.0 |
| Eire | 0.9 | 0.8 |
| Finland | 1.3 | 1.2 |
| France | 1.05 | 0.95 |
| Germany (West) | 1.1 | 1.0 |
| Greece | 1.0 | 0.9 |
| Holland | 1.1 | 1.0 |
| India (imported elements) | 2.0 | 1.8 |
| (indigenous element) | 0.7 | 0.65 |
| Italy, Japan | 1.0 | 0.9 |
| Malaysia | 0.9 | 0.8 |
| Middle East | 1.2 | 1.1 |
| New Zealand | 1.4 | 1.3 |
| North Africa (imported elements) | 1.2 | 1.1 |
| (indigenous element) | 0.8 | 0.75 |
| Norway | 1.2 | 1.1 |
| Portugal | 0.8 | 0.75 |
| South Africa | 1.25 | 1.15 |
| South America (North) | 1.5 | 1.35 |
| South America (South) | 2.5 | 2.25 |
| Spain (imported elements) | 1.3 | 1.2 |
| (indigenous element) | 0.8 | 0.75 |
| Sweden | 1.2 | 1.1 |
| Switzerland | 1.2 | 1.1 |
| Turkey | 1.1 | 1.0 |
| Yugoslavia | 1.0 | 0.9 |

*Notes*:

1. Increase a factor by 10% for each 1,000 miles, or part of it, that the new plant location is distant from major manufacturing or import centre, or both.
2. Investment incentives have been ignored.
3. Date of compilation: early 1978.

### Ratio methods for rapid estimates

One of the earliest methods for quick estimation of capital costs is based on a ratio of dollars invested per annual ton of capacity, CTC/R [3, 5, 25]. (Hence the name ratio method.) Published ratios cover a wide range — from about

$5–$500/annual ton of capacity. Even when no ratio has been reported for the process being investigated, an approximate ratio can be estimated by analogy from known data published for similar processes. *Table 4.9* lists some cost/capacity figures for several processes involving production plants built on a previously developed site. Holland [25] proposed a factor of 1.3–1.4 for units constructed on a 'green-field' site and a factor of 0.7–0.8 for enlargements of an existing plant. Cost data in *Table 4.10* apply to the first quarter of 1967 and for average conditions in the U.S.

The figures in *Table 4.9* can be corrected for time and location using the appropriate cost indexes and factors from *Table 4.8*. For plant capacities other than those presented in *Table 4.9* the capacity ratio exponents, $a$, may be used

Table 4.10. Total capital investment data for different processes [25]

| Product | Plant capacity T/year | Total capital cost $ | Capital cost per annual ton, $/t |
|---|---|---|---|
| Acetaldehyde (1) | 50,000 | 3,500,000 | 70 |
| Acetylene (2) | 75,000 | 9,000,000 | 127 |
| Ammonia | 500,000 | 16,000,000 | 32 |
| Ammonium phosphate | 250,000 | 2,500,000 | 10 |
| Ammonium sulphate | 140,000 | 1,200,000 | 9 |
| Butadiene (3) | 100,000 | 50,000,000 | 500 |
| Butadiene (4) | 200,000 | 70,000,000 | 350 |
| Cyclohexane | 100,000 | 800,000 | 8 |
| Ethyl alcohol (1) | 75,000 | 3,750,000 | 50 |
| Ethylene (5) | 300,000 | 15,000,000 | 50 |
| Ethylene dichloride | 25,000 | 3,200,000 | 127 |
| Ethylene oxide (1) | 100,000 | 9,000,000 | 90 |
| Glycerol | 35,000 | 5,000,000 | 157 |
| Hydrogen | 60,000 | 6,500,000 | 108 |
| Methanol (2) | 210,000 | 9,000,000 | 43 |
| Methyl chloride (6) | 10,000 | 500,000 | 50 |
| Oxygen | 150,000 | 2,250,000 | 15 |
| Phosphoric acid (as anhydride) | 100,000 | 2,400,000 | 24 |
| Polyisoprene | 30,000 | 5,000,000 | 320 |
| Sulphuric acid (contract process) | 280,000 | 2,100,000 | 8 |

Raw materials: (1) ethylene, (2) natural gas, (3) butane, (4) butylene, (5) refinery gases, (6) methanol

Table 4.11. Capacity ratio exponent

| Process | Exponent | Source |
|---|---|---|
| Acetylene | 0.75 | 25 |
| Ammonia | 0.74 | 25 |
| Ammonium nitrate | 0.54 | 25 |
| Benzene | 0.61 | 25 |
| Butadiene | 0.59 | 25 |
| Caustic soda | 0.35 | 25 |
| Chlorine | 0.35 | 25 |
| Complete refinery | 0.85 | 3 |
| Catalytic cracking | 0.78 | 3 |
| Coking only | 0.72 | 3 |
| Catalytic polymerization | 0.70 | 3 |
| Coke-oven gas separation | 0.82 | 3 |
| Ethanol | 0.60 | 25 |
| Ethylene | 0.58 | 25 |
| Ethylene oxide | 0.79 | 25 |
| Chamber sulphuric acid | | 3 |
| Contact sulphuric acid | 0.64–0.66 | 3 |
| Formaldehyde | 0.55 | 25 |
| Hydrogen cyanide | 0.71 | 25 |
| Isopropyl alcohol | 0.60 | 25 |
| Methanol | 0.83 | 25 |
| Nitric acid | 0.56 | 25 |
| Oxygen | 0.64 | 25 |
| Phosphoric acid | 0.58 | 25 |
| Refrigeration plant, $NH_3$ | 0.68 | 3 |
| Polyethylene (low pressure) | 0.67 | 25 |
| Solvent dewaxing | 0.76 | 3 |
| Styrene | 0.68 | 25 |
| Sulphuric acid (contact) | 0.62 | 25 |
| Urea | 0.59 | 25 |
| Water-gas plant | 0.81 | 3 |
| Water-gas shift conversion | 0.69 | 3 |
| Water treating plant | 0.91 | 3 |

within a twofold ratio of sizes:

$$\mathrm{CFC}(2) = \mathrm{CFC}(1) \times \left(\frac{\mathrm{Capacity}(2)}{\mathrm{Capacity}(1)}\right)^{a} \tag{4.97}$$

The capacity exponent has been found to be, on average, between 0.6 to 0.7 for many processes. In some papers the above equation is called the 'six-tenth rule' (see *Table 4.11*). In order to derive more accurate estimates by the six-tenth rule, the subdivision of the process plant into various process units is recommended.

The other ratio used in preliminary capital cost estimation is the Capital Ratio, CR, which is defined as the ratio of the fixed capital investment (for construction

Table 4.12. Typical capital ratio values

| Industry | Capital ratios | |
|---|---|---|
| | Lynn [4] (1939–52) | Holland [25] (1958) |
| Petroleum | 2.51 | 3.08 |
| Inorganics, heavy | 2.06 | 2.24 |
| Chemicals, general | 1.25 | 2.02 |
| Resins & plastics | 0.77 | 1.90 |
| Glass | 0.97 | 1.46 |
| Rubber | 0.75 | 1.04 |
| Pharmaceuticals | 0.53 | 0.92 |
| Soap & detergents | 0.48 | 0.69 |
| Foodstuffs, processed | 0.63 | 0.66 |

on a green-field site) to the annual sales income:

$$CR = CFC/As \tag{4.98}$$

Fixed capital in this case covers money invested in land, production facilities, warehousing, R&D costs, and any auxiliary facilities necessary to support the process. Capital ratio data were published in Refs [4, 5, 25, etc.]. Some authors refer to the same quantity as 'turnover ratio'.

*Table 4.12* lists capital ratios for some branches of the chemical industries. The Table shows that capital ratios are not independent of time, they have a tendency to rise.

### *Factored estimates*

Ratio methods do not require any design activity prior to estimating the capital cost and they can therefore be regarded as pre-design estimation methods. In contrast, factored estimates need some knowledge about the main equipment items, the main sizes and technological parameters of equipment, and the delivered cost of these items. The total delivered cost of equipment is the initial point of the factored estimate. In cases when only the purchase cost (free on board, 'fob') of equipment is known, this may be multiplied by a factor of 1.03 to give approximate delivered cost. The main steps of preparing a capital cost estimation, based on design, are listed by Holland [25] as:

1. Initial idea for the process and specification of the size and type of operation.
2. Collection of physical and chemical data from the literature or from laboratory experiments.
3. Preparation of a preliminary equipment flowsheet incorporating the required unit operations, and showing the main items of equipment.

4. Heat and mass balancing.

5. Specification of temperatures and pressures at selected points on the equipment flowsheet.

6. Design calculations to size the main items of equipment such as reactors, distillation columns, pumps, compressors, etc.

7. Preparation of a coded list of major items of process equipment such as storage tanks, reactors, compressors, columns, circulating pumps, heat exchangers, together with details such as size, capacity, materials of construction, pressures and temperatures.

8. Collection of cost data from literature, company records and quotations.

9. Estimation of the total delivered cost of all the items of major process equipment listed under point 7.

The first generally accepted method of converting the total delivered equipment cost to capital cost was that proposed by Lang (in Chilton [4]). After analysing [14] different processes, and grouping them into three classes as

— solids processing,
— mixed solids-fluid processing, and
— fluid processing,

he calculated the average factors relating the fixed capital cost to the delivered equipment cost in each category.

He found that for the equation

$$\text{CFC} = f \times \text{Sum CEQ}(i) \tag{4.99}$$

where CEQ($i$) is the delivered cost of a major item of process equipment, $f$ is the so-called Lang factor; the following factors apply:

$f = 3.1$ for solids processing
$f = 3.63$ for mixed solids-fluid processing
and
$f = 4.74$ for fluid processing.

Note that in his original paper Lang described the way to develop these factor values as the products of some other factors, and he recommended the use of self-developed factors, based on own company records.

Several attempts have been made to refine the Lang factor method and to break down the Lang factor into a series of other factors. In *Fig. 4.9* we have tried to construct one possible 'family tree' of the factored estimates without laying any claiming to completeness.

Chilton [4] developed correlations relating the main sizes of equipment to their installed costs. From these correlations the total equipment cost could be calculated (base year = 1949). He expressed the fixed capital cost as

$$\text{CFC} = F_1 \times F_2 \times \text{Sum CEQ}(i) \tag{4.100}$$

where $F_1 = 1 + f_1 + f_2 + f_3 + f_4 + f_5$
$F_2 = 1 + f_6 + f_7 + f_8$

$F_1$ relates the total physical cost of the plant to the total installed equipment cost, $F_2$ relates the total fixed capital cost to the total physical cost.

The factors $f$ account for the following items:

| Factor | Item | Average values |
|---|---|---|
| $f_1$ | Process piping | 0.1 −0.3 |
| $f_2$ | Instrumentation | 0.05−0.1 |
| $f_3$ | Manufacturing buildings | 0.2 −0.6 |
| $f_4$ | Auxiliary facilities | 0.05−0.25 |
| $f_5$ | Outside lines | 0.05−0.15 |
| $f_6$ | Engineering & construction | 0.2 −0.5 |
| $f_7$ | Contingencies | 0.2 −0.3 |
| $f_8$ | Size | 0.05−0.15 |

Miller [7] based his technique on the fact that there are three significant items causing variation in factor estimates:

— size of equipment;
— materials of construction;
— operating pressures.

If the size of the equipment increases the overall factor becomes smaller.

If the equipment is fabricated from more expensive materials such as stainless steel, glass-lined vessels, etc., the factor again becomes smaller.

If the operating pressure increases, it is also known that the overall factor will decrease.

To some extent, all these effects can be taken into account by one number, the 'average unit cost' of the process equipment. Average unit cost, CEav, of the process equipment is defined as the ratio of the total cost of process equipment, Sum CEQ($i$), to the number of equipment items, $n$. (The average unit cost is based on the main plant items only, and does not include miscellaneous unlisted items.)

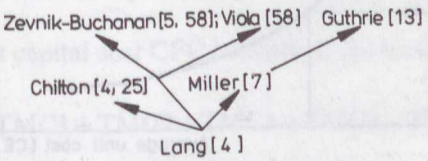



Fig. 4.9. Family tree of factored estimates

*Figure 4.10* illustrates how this characteristic measure can be used: each of the above mentioned three items increases the average unit cost, and the overall factor decreases as the average unit cost increases. Miller presented ranges of values of different factors to account for the effect of field erection, equipment foundations and supports, piping, insulation, all electrical services, instrumentation, buildings, etc. as a function of the average unit cost. These values apply strictly to the battery limit plant. Since there are actually four areas involved in plant estimates:

| | |
|---|---|
| Battery limit | (B/L) |
| Storage and handling | (S/H) |
| Utilities | (U) |
| Services | (SE), |

he corrected the B/L estimate with additional factors to give a complete B/L estimate. The last three items are often referred to in the literature as off-site facilities.

His estimation algorithm can be formulated as follows:

$$\text{CEav} = \text{Sum}\,(\text{CEQ}(i))/n \tag{4.101}$$

$$\text{B/L} = (1 + f_1 + f_2 + f_3 + f_4 + f_5 + \ldots + f_{10}) \times \text{Sum}\,(\text{CEQ}(i))$$

$$\text{S/H} = f_{11} \times \text{B/L}$$

$$\text{U} = f_{12} \times \text{B/L}$$

$$\text{SE} = f_{13} \times (\text{B/L} + \text{S/H} + \text{U})$$

$$\text{CFC} = \text{B/L} + \text{S/H} + \text{U} + \text{SE}$$

and all $f$ values are functions of the average unit cost, CEav.

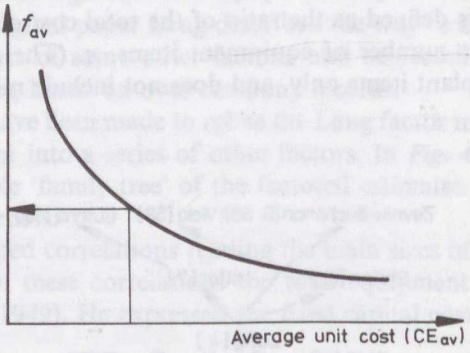



Fig. 4.10. The overall factor as a function of the average unit cost

### The modular technique (Guthrie)

One of the equipment-list-plus-factor methods for estimating the fixed capital cost (CFC) is that of Guthrie [13]. The method is discussed in detail in this section. For its implementation see the CAPCOS Manual [63].

Guthrie [13] collected and analysed feedback data of 42 process plant projects during the elaboration of his capital cost estimation method. All these plants were erected in the U.S.A. (Gulf Coast) in mid-1968. Hence, in applying the method, any differences in plant location have to be taken into account.

We shall survey Guthrie's method. Naturally, our text contains a number of citations from the original report, since we have neither the right nor the grounds to alter definitions or formulae. The only principal difference is that we do not consider that indirect costs as an individual module. Indirect costs are always estimated as a percentage of some 'direct' costs. There is no reason for collecting them into a separate module. This causes a slight modification in the meaning of the total module cost: it also contains the indirect costs.

All possible capital cost elements are grouped into five modules, each module representing cost elements having similar characteristics. The modules are as follows:

1. chemical processing;
2. solids handling;
3. site development;
4. industrial buildings;
5. off-site facilities.

The most important module is the first one, and it is therefore split up into seven equipment groups:

1. furnaces;
2. heat exchangers;
3. vertical pressure vessels;
4. horizontal pressure vessels;
5. pumps and drivers;
6. compression;
7. on-site tankage.

The estimates may be made independently for each module. The estimate levels may thus differ for the modules: e.g. one can mix accurate estimates for the main module (usually for the chemical processing) and rough estimates for the other modules.

The estimated project capital cost CFC consists of the total module costs of the five modules:

$$\mathrm{CFC} = \mathrm{TMC1} + \mathrm{TMC2} + \mathrm{TMC3} + \mathrm{TMC4} + \mathrm{TMC5} \qquad (4.102)$$

where TMC$j$ denotes the estimate of total module cost for module $j$.

TMC$j$ is assembled from six primary cost elements, some of them assembled from secondary elements; some secondary estimates of the indirect cost are collected from tertiary costs. Thus, for each module, a three-level hierarchy of cost elements may be built up. These 'pyramids' have similar characteristics but differ in particular data.

*Level 0: total module cost*

TMC — The total module cost is the sum of all primary cost elements described later:

$$\text{TMC} = E + m + L + I + C + F \tag{4.103}$$

*Level 1: primary cost elements*

$E$ — total equipment free on board (fob) cost. This value yields the base for every further estimate: all other cost elements are related to it. No standards are reported for $E$, and we therefore need an 'external' estimate.

The role of the chemical processing module is crucial, and it is therefore worth refining the total equipment fob cost, i.e. determining its details for each main equipment group. Let $Ei$ denote the total equipment fob cost of main equipment group $i$ ($i = 1, \ldots 7$) expressed as a percentage of $E$. A normal 'equipment mix' (i.e. the percentage distribution of $E$) is shown in *Table 4.13.*

$m$ — direct field material cost, i.e. cost of auxiliary materials needed to install all the equipment *(Table 4.14)*.

This cost is related to $E$, and its normal value depends on the module. In the following table normal values, $Mi$, are given for each main equipment group of the chemical processing module expressed as a percentage of the total equipment fob cost $Ei$ of the group.

Table 4.13. Standard equipment mix for the chemical processing module

| Equipment group | $Ei$ |
|---|---|
| 1. Furnaces | 14 |
| 2. Heat exchangers | 18 |
| 3. Vertical process vessels | 15 |
| 4. Horizontal process vessels | 8 |
| 5. Pumps and drivers | 7 |
| 6. Compression | 30 |
| 7. On-site tankage | 8 |
| Total equipment cost (e) | 100 |

Table 1.14. Direct field material costs for the chemical processing module

| Equipment group | % |
|---|---|
| 1. Furnaces | 34 |
| 2. Heat exchangers | 71 |
| 3. Vertical process vessels | 103 |
| 4. Horizontal process vessels | 63 |
| 5. Pumps and drivers | 72 |
| 6. Compression | 58 |
| 7. On-site tankage | 20 |

Note that if a standard equipment mix may be used in the first module, then the expected value of direct material costs $M1$ for them is 62% of the total fob cost $E1$.

A 'norm' value of $M2$ (solids handling) is also reported. It is 27% of $E2$. For the last three modules there is no 'norm'. In individual cases, however, we shall also estimate $m$ for items in these modules.

$L$ — direct field labour cost. This represents the labour costs required to erect and set up all the equipment belonging to the module. It does not contain indirect costs such as construction overheads. 'Norms' are reported, expressing $L$ as a percentage of $E$ for the first and second module:

| | |
|---|---|
| Chemical processing module | 58% |
| Solids handling module | 34% |

$I$ — indirect costs. The estimation of indirect costs is independent of the module type. Its 'norm' value is 34% of direct costs M&L of the module (see below).

The sum of primary costs, as defined so far, is called the bare module cost (BMC). The remaining two primary cost elements are both related to BMC; thus we have to calculate this term:

$$\text{BMC} = E + m + L + I \tag{4.104}$$

$C$— contingency. It is an allowance for costs that cannot be estimated directly but that will certainly occur. According to Guthrie its 'norm' value is 15% of the bare module cost BMC.

$F$— contractor's fee. Its 'norm' value is 3% of the bare module cost BMC.

*Level 2: secondary cost elements*

Now we can turn to the second level of our pyramid. Naturally, the list of secondary cost elements differs for different primary costs. Some secondary costs are assembled from tertiary cost elements.

Table 4.15. Normal percentage distribution of direct material costs for the chemical processing module

| | Furnace | Exchanger | Vessel (v) | Vessel (h) | Pump driver | Compressor | Tank |
|---|---|---|---|---|---|---|---|
| Piping | 53 | 63 | 58 | 63 | 42 | 35 | 55 |
| Concrete | 29 | 8 | 9 | 10 | 6 | 21 | 40 |
| Steel | 0 | 4 | 8 | 0 | 0 | 0 | 0 |
| Instruments | 12 | 14 | 11 | 10 | 4 | 14 | 0 |
| Electrical | 6 | 3 | 5 | 8 | 43 | 25 | 0 |
| Insulation | 0 | 7 | 8 | 8 | 4 | 4 | 0 |
| Paint | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| $m$ | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

$m$ — direct field material cost $m$ is assembled from the following secondary cost elements:

— piping;
— concrete foundation;
— minor steel structures such as ladder platforms, supports, etc.;
— insulation;
— paint;
— electrical materials required for the equipment.

In *Table 4.15* we list the normal percentage distribution of $mi$ ($i=1, \ldots, 7$).

The normal percentage distribution for the solids handling module is shown in *Table 4.16.*

Table 4.16. Normal percentage distribution of direct material costs for the solids handling

| Module | % |
|---|---|
| Piping | 6 |
| Concrete | 31 |
| Steel | 2 |
| Instruments | 0 |
| Electrical | 59 |
| Insulation | 0 |
| Paint | 2 |
| Total | 100 |

Table 4.17. Normal percentage distribution of direct labour costs ($L$) for the chemical processing module

| Equipment group | Erection % | Setting % |
|---|---|---|
| Furnaces | 100 | 0 |
| Exchangers | 87 | 13 |
| Vessels (v) | 85 | 15 |
| Vessels (h) | 87 | 13 |
| Pumps | 86 | 14 |
| Compression | 82 | 18 |
| Tanks | 94 | 6 |

All these standards are based on carbon steel and represent mid–1968 costs on a U. S. (Gulf Coast) job-site.

$L$— direct field labour cost $L$ is assembled from two secondary cost elements:

— material erection;
— equipment setting.

'Norm' values for these secondary costs are given only for the chemical processing module and are listed in *Table 4.17*.

$I$ — the secondary cost elements of the indirect cost ($I$) are the following:

— FIT: freight, insurances, taxes, import duties.
— construction overhead. This is related to direct field labour costs ($L$) and its reported average is approximately 67.4%.
— engineering costs. Total engineering cost is related to the direct material cost and its average value is 13.6%. This value was observed for chemical processing plants. It is recommended to adjust the project engineering cost by multiplying it by a factor Fpt if project characteristics are different. 'Norm' values for Fpt are shown in *Table 4.18*.

Table 4.18. Project type factors for engineering costs

| Project type | Fpt |
|---|---|
| Chemical processing | 1.0 |
| Chemical complex | 1.4 |
| Solids/fluid processing | 0.8 |
| Solids handling | 0.6 |
| Buildings only | 0.4 |

*Level 3: tertiary cost elements*

Two of the secondary elements of the indirect costs may be broken up as follows:

— Construction overhead contains the costs of seven tertiary cost elements
  — *Fringe benefits.* Employer's contribution to union funds for health and welfare, vacations, holidays, sick leave, retirement. It also contains travel and subsistence costs.
  — *Labour burden.* Employer's mandatory contribution to some insurance associations, e. g. social security, uneployment insurance.
  — *Field supervision.* Salaries, fringe benefits and overhead items for supervisory and field personnel.
  — *Temporary facilities.* Buildings, roadways vehicle parks, work areas, scaffolding, rigging, utilities, fencing and miscellaneous.
  — *Construction equipment.* Equipment rental and handling, plus freight to and from the job-site.
  — *Small tools.* Expendable construction tools, usually valued under $200.
  — *Miscellaneous field costs.* Job clean-up, watchmen, minor equipment repairs, medical services, welder tests, consumable supplies, warehousing, vendor services, job insurance.

In *Table 4.19* we show the 'norm' percentage distribution of construction overhead costs:

— Engineering cost can be assembled from six tertiary cost elements *(Table 4.20)*.

Finally, we list some other costs which can be derived from the costs described thus far, but which are sometimes used as base for comparison.

$M$ — some relations are defined using direct material costs $M$:

$$M = E + m \tag{4.105}$$

$M\&L$ — direct field material and labour costs. Some values are related to this term.

$$M\&L = M + L = E + m + L \tag{4.106}$$

Table 4.19. Percentage distribution of construction overhead costs

| | |
|---|---|
| Fringe benefits | 14.8 |
| Labour burden | 22.4 |
| Field supervision | 17.8 |
| Temporary facilities | 8.9 |
| Construction equipment | 14.8 |
| Small tools | 3.6 |
| Miscellaneous | 17.7 |

Table 4.20. Percentage distribution of engineering costs

| | |
|---|---|
| Project engineering | 14.3 |
| Process engineering | 4.5 |
| Design and drafting | 26.8 |
| Procurement | 0.8 |
| Office indirect overhead | 50.9 |

There are certain key relationships between the major cost elements described so far. These relationships are expressed in ratios, and they measure some important values such as labour productivity, etc. Their common name, which will be used in what follows, is 'factor'. Primary factors yield important checkpoints. Differences in plant location, labour productivity, etc. may be taken into account by using modified factors, if their proper value is known. The adjective 'primary' is used because we wish to distinguish these factors from those occurring in estimating individual equipment costs; primary factors measure relationships between certain cost elements and main equipment parameters such as type of construction material, design type, design pressure, etc. We now list the primary factors and their normal (mid–1968, U. S. Gulf Coast) values.

1. Material factor: ratio of direct material costs ($m$) and total fob equipment costs ($E$).
2. Labour factor: ratio of direct field labour costs ($L$) and total fob equipment costs ($E$).
3. $L/M$ ratio: this factor measures labour productivity. It may vary with locations and special activities. Note that labour factor and $L/M$ ratio is based on a composite rate of \$4.80/h.
4. Direct cost factor: ratio of all direct costs ($E+m+L$) to total fob equipment cost ($E$).
5. Indirect cost factor: ratio of module indirect costs ($I$) to all direct costs ($E+m+L$).
6. Bare module factor: ratio of bare module cost BMC ($=E+m+L+I$) to total equipment fob cost ($E$).
7. Total module factor: ratio of total module cost TMC to total fob equipment cost ($E$).

In *Table 4.21* we show the observed range and 'norm' values of these factors for the chemical processing module.

For the solids handling module average primary factors are reported *(Table 4.22)*.

If one has an up-to-date value for any of these factors one can modify the estimation method by simply calculating the appropriate cost elements using the factor. In this case the secondary and further cost elements can be estimated

Table 4.21. 'Norm' values of the primary factors for chemical processing module

| Factor | Ratio of | Range | | Norm |
|---|---|---|---|---|
| | | from | to | |
| Material | $m/E$ | 0.42 | 0.75 | 0.62 |
| Labour | $L/E$ | 0.54 | 0.66 | 0.58 |
| $L/M$ ratio | $L/M$ | 0.32 | 0.40 | 0.36 |
| Direct cost | $ML/E$ | 1.80 | 2.60 | 2.20 |
| Indirect cost | $I/ML$ | 0.32 | 0.45 | 0.34 |
| Bare module | BMC/$E$ | 2.38 | 3.64 | 2.95 |
| Total module | TMC/$E$ | | | 3.48 |

without any change in the relationships between the same cost details. The distribution tables in the previous section may be used for this purpose.

For example, suppose that we have an estimated total fob equipment cost of \$100,000 in the chemical processing module. We want to estimate all further costs together with their details. Suppose that we have no more information, i. e. Guthrie's 'norm' values will be used everywhere. Then we proceed as follows (all costs are expressed in multiples of \$1000):

$$E = 100$$

The next table may be calculated column by column. To calculate the first and second columns use *Table 4.12* (standard equipment mix) and *Table 4.13*. Multiplying the fob cost column (first) by 0.58 and 0.34 we get the third and fourth columns. In the last row and last column we find the sums of the columns and rows, respectively.

| Equipment group | $E$ | $M$ | $L$ | $I$ | BMC |
|---|---|---|---|---|---|
| Furnaces | 14.0 | 4.8 | 8.1 | 4.8 | 31.7 |
| Exchangers | 18.0 | 12.8 | 10.4 | 6.1 | 47.3 |
| Vessels (v) | 15.0 | 15.5 | 8.7 | 5.1 | 44.3 |
| Vessels (h) | 8.0 | 5.0 | 4.6 | 2.7 | 20.3 |
| Pumps & drivers | 7.0 | 5.0 | 4.1 | 2.4 | 18.5 |
| Compression | 30.0 | 17.4 | 17.4 | 10.2 | 75.0 |
| Tankage | 8.0 | 1.6 | 4.6 | 2.7 | 16.9 |
| Total | 100.0 | 62.1 | 57.9 | 34.0 | 254.0 |

Now, the lower right-hand position of our Table shows the bare module cost, BMC. We have to multiply it by 1.18 to obtain an estimate of the total module cost:

$$\text{TMC} = 1.18 \times 254 = 299.7$$

Table 4.22. 'Norm' value of the primary factors for the solids handling module

| Factor | Ratio of | Norm |
|---|---|---|
| Material | $m/E$ | 0.27 |
| Labour | $L/E$ | 0.34 |
| $L/M$ ratio | $L/M$ | 0.27 |
| Direct cost | $ML/E$ | 1.61 |
| Indirect cost | $I/E$ | 0.29 |
| Bare module | BMC/$E$ | 2.08 |
| Total module | TMC/$E$ | 2.46 |

If we need a more detailed view of the direct material cost we can use *Table 4.14*. For instance, from the Table above the expected direct material cost $M3$ of vertical pressure vessels is \$15,500. By the third column of *Table 4.14* we estimate the following distribution of $M3$:

| | |
|---|---|
| Piping | $0.58 \times 15.5 = 9.0$ |
| Concrete | $0.09 \times 15.5 = 1.4$ |
| Steel | $0.08 \times 15.5 = 1.2$ |
| Instruments | $0.11 \times 15.5 = 1.7$ |
| Electrical | $0.05 \times 15.5 = 0.8$ |
| Insulation | $0.08 \times 15.5 = 1.2$ |
| Paint | $0.01 \times 15.5 = 0.2$ |
| Total | 15.5 |

Now suppose that we have a firm value of 0.52 for the $L/M$ ratio. We can take this value into account by simply recalculating those columns of our table which are affected by the $L/M$ ratio, i. e. the third and fifth columns. We thus obtain the modified Table shown below:

| Equipment group | $E$ | $M$ | $L$ | $I$ | BMC |
|---|---|---|---|---|---|
| Furnaces | 14.0 | 4.8 | 9.8 | 4.8 | 33.4 |
| Exchangers | 18.0 | 12.8 | 16.0 | 6.1 | 52.9 |
| Vessels (v) | 15.0 | 15.5 | 15.9 | 5.1 | 51.5 |
| Vessels (h) | 8.0 | 5.0 | 6.8 | 2.7 | 22.5 |
| Pumps drivers | 7.0 | 5.0 | 6.2 | 2.4 | 20.6 |
| Compression | 30.0 | 17.4 | 24.6 | 10.2 | 82.2 |
| Tankage | 8.0 | 1.6 | 5.0 | 2.7 | 17.3 |
| Total | 100.0 | 62.1 | 84.3 | 34.0 | 280.4 |

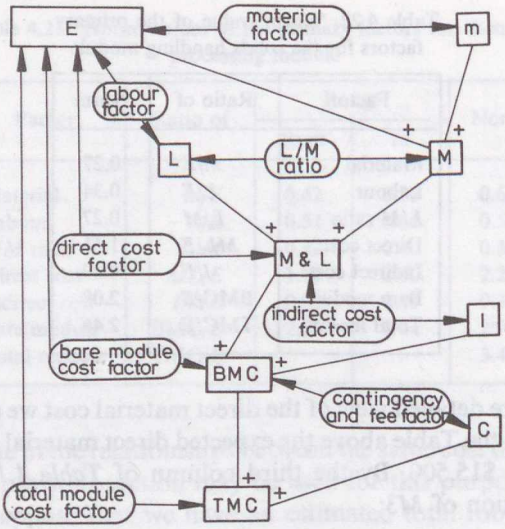



Fig. 4.11. Module costs and factors

and $TMC = 1.18 \times 280.4 = 330.9$

In *Fig. 4.11* we summarize all the main module costs and the factors between them.

*Cost estimation using main equipment size parameters*

Guthrie has also developed a long list of different equipment occurring in chemical process plants. For each item of equipment he determined its main size parameters as they essentially affect the cost. Finally, using correlation analysis, he gave the appropriate formulae to calculate the costs from the main size parameters.

Naturally, in chemical process plants the chemical processing module represents the most important module and essentially determines the total cost of the project. Therefore, Guthrie focused his attention on this module. Cost estimation is the most detailed in this module and the number of parameters considered is relatively high. In the other modules only one or two parameters are used. There is also a validity interval given for each parameter. The cost of an item of equipment having some parameters outside these limits can also be estimated by this method, but the expected accuracy is not very high.

For each equipment belonging to the same equipment group (e. g. for all heat exchangers, for all furnaces, etc.) we first calculate the base cost BC. This is not the price of the unit concerned but the cost of a standard unit of the same size as the one in question. Usually only one main parameter is considered here, and it is supposed that all other parameters take their standard values. This price is not actualized, so it is the expected mid–1968, U. S. Gulf Coast price.

When all base costs of an equipment group have been determined, we sum them up and determine the base dollar amount for this group. There are five ranges distinguished in the original report:

A. –$200.000
B. $200.000–$400.000
C. $400.000–$600.000
D. $600.000–$800.000
E. $800.000

Then we determine (using the appropriate tables) the ratios of the cost details and the base cost. These ratios depend on the group magnitude. In general, the higher the magnitude, the lower the ratio of the additional costs and the base cost. These ratios are the following:

| | | |
|---|---|---|
| Piping | /BC | Direct material |
| Concrete | /BC | |
| Steel | /BC | |
| Instruments | /BC | |
| Electrical | /BC | |
| Insulation | /BC | |
| Paint | /BC | |

| | | |
|---|---|---|
| Material erection | /BC | Direct field labour |
| Equipment setting | /BC | |
| Freight, insurance, taxes | /BC | Indirect costs |
| Other indirect costs | /BC | |

Now we are ready to proceed with our estimation in this equipment group. We have to repeat the following steps for each item of equipment in the group.

First, we calculate the fob cost $E$ of the actual equipment in the base year. This is calculated by multiplying the base cost by the appropriate factor or combination of factors. Each main parameter not considered in the calculation of the base cost gives a factor for this purpose. The formula for calculating $E$ from the base cost BC and the factors varies with the equipment groups.

On multiplying the equipment base cost by the appropriate ratios (see above), we obtain the additional cost elements (material, labour and indirect). Their sum gives the bare cost of the equipment.

The bare group cost is simply the sum of the equipment bare costs belonging to the group.

Let us illustrate this procedure by an example:

*Estimation of process furnaces*

Base year: 1968

| Parameter | Unit | Limits | Default | Factor |
|---|---|---|---|---|
| Absorbed heat duty | 1E6 Btu/h | 10–1000 | 100 | – |
| Furnace desing type | – | 1–3 | 1 | fd |
| | 1 – process heater | | | 1.00 |
| | 2 – pyrolysis | | | 1.10 |
| | 3 – reformer | | | 1.35 |
| Radiant tube material | – | 1–3 | 1 | fm |
| | 1 – carbon steel | | | 0.00 |
| | 2 – chrome/molybdenum | | | 0.35 |
| | 3 – stainless | | | 0.75 |
| Desing pressure | psi | 10–3000 | 100 | fp |
| | up to 500 | | | 0.00 |
| | 1000 | | | 0.10 |
| | 1500 | | | 0.15 |
| | 2000 | | | 0.25 |
| | 2500 | | | 0.40 |
| | 3000 | | | 0.60 |

The absorbed heat duty AHD is taken into account in calculating the base cost, and therefore there is no factor associated with it. The base cost is the mid-1968, U.S. Gulf Coast fob cost of a process heater made of carbon steel and with a design pressure of 500 psi. The formula for BC is as follows (expressed in $1000):

$$BC = 6.76 \times AHD^{0.793}$$

The formula for $E$ which provides the fob cost in the base year, 1968, but with the actual type, material and design pressure is

$$E = BC \times (fd + fm + fp)$$

Finally, the appropriate ratios, expressed as percentages of BC, are the following:

| Magnitude $ 1000 | –200 | –400 | –600 | –800 | –1000 |
|---|---|---|---|---|---|
| Piping | 18.5 | 18.0 | 17.7 | 17.5 | 17.5 |
| Concrete | 10.3 | 10.0 | 9.9 | 9.78 | 9.7 |
| Steel | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Instruments | 4.1 | 4.0 | 3.9 | 3.9 | 3.9 |
| Electrical | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Insulation | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Paint | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Material erection | 30.5 | 29.2 | 29.2 | 28.9 | 28.7 |
| Equipment setup | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| FIT | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Other indirects | 61.2 | 55.6 | 53.7 | 53.5 | 51.7 |

**Example:** Predict the 1972 costs of three similar process heaters, each of them designed to liberate 95 million Btu/h at 1400 psi supposing that stainless steel tubes are required. Calculate using 8%/yr escalation for materials and labour costs and 9%/yr for indirects. The calculations proceeds as follows (all costs in \$1000):

$$BC = 3 \times (6.76 \times 95^{0.793}) = 3 \times 250.2 = 750.6$$

Magnitude: 4 (600–800)

For each furnace:

$$fd = 1.00$$
$$fm = 0.75$$
$$fp = 0.15$$
$$E = 250.2 \times (1.00 + 0.75 + 0.15) = 475.4$$

From the fourth column of the table above, for each furnace:

$$\text{material factor} = 17.5 + 9.8 + 0.0 + 3.9 + 2.0 + 0.0 + 0.0 = 33.2$$
$$\text{labour factor} = 28.9 + 0.0 = 28.9$$
$$\text{indirect factor} = 0.0 + 53.5 = 53.5$$

Then, for the module:

$$E = 3 \times 475.4 \times 1.08^4 = 1940.3$$
$$m = 3 \times (250.2 \times 0.332) \times 1.08^4 = 339.0$$
$$L = 3 \times (250.2 \times 0.289) \times 1.08^4 = 395.1$$
$$I = 3 \times (250.2 \times 0.535) \times 1.09^4 = 566.8$$
$$BMC = 1940.3 + 339.0 + 395.1 + 566.8 = 3241.2$$
$$TMC = 1.18 \times 3241.2 = 3824.6$$

Thus, the expected total cost of the three furnaces in 1972 is \$3,824,600.

### 4.2.2 Working capital estimation

Working capital is the resource which enables a company to conduct its day-to-day business [30]. There are two types of working capital used in financial analysis: gross working capital and net working capital.

The former is defined as the total current assets, while the latter is the current assets minus the current liabilities. The component parts of the total current assets are:

— cash for monthly payments of operating expenses;
— accounts receivable; and
— inventories: raw-materials inventory,
work in progress inventory,
finished product inventory.

The current liabilities comprise:

— accounts payable;
— notes payable;
— bank loans;
— accruals payable; and
— taxes payable.

The best source of information for use in working capital estimates is data from similar or identical projects. Most companies have extensive records of their operation, so that quick and reliable estimates of working capital can be obtained. On the other hand, company standards often prescribe the amount of working capital to be allocated to a project. It is pointless to perform an evaluation on a basis that is not consistent with the standards of the company [5].

Company standards may be expressed as percentage of annual sales, or as a combination of three terms:

— $L$ months' supply of raw materials (at delivered cost);
— $M$ months' supply of products, at full manufacturing costs; and
— $N$ months' sales at full cost of sales.

In some cases the company standards may need some modification, since some raw materials or products may be:

— seasonal (needing a high inventory);
— shipped in very large volume, such as a tanker (also high inventory);
— very slow in terms of processing time required, such as aging or curing, e.g. in the tobacco industry (high inventory);

— arriving from remote locations, e.g. overseas (high inventory);
— produced and used in the same location (low inventory);
— supplied by pipeline (generally the lowest inventory).

Management normally aims at keeping inventories at the lowest possible level, consistent with customer satisfaction and efficient plant operation. Excessive inventories are unproductive, have no rate of return, and tie up money which could be invested in a more profitable business activity. Excessively high inventory can create a shortage of cash. In the less serious cases this means that accounts payable cannot be paid in time, and valuable cash discounts may be lost. In the worst case it may lead to liquidity problems and even to bankruptcy.

If there are no company standards, the estimator can apply some rules of thumb. For exploratory estimates a three-month sales period is suggested [5]. For more detailed evaluations, the following ranges may be applied (subject to change depending on the specific characteristics of the process):

— Raw materials (with delivery and handling) 1/2 to 2 months.
— Sales costs (sales income less pretax profit and depreciation) 1/2 to 4 months.

The working capital demand of a certain project can be related to the total capital investment. This ratio varies with different types of projects and with different companies. Most chemical companies use an initial working capital of about 10–20% of the total capital investment [3]. It is worth noting that working capital — unlike fixed capital investment — varies with production rate, but its estimate is usually based on steady operating conditions.

## 4.3 Manufacturing cost estimation

In Section 4.1.1 we discussed the annual cash income before tax, ACI, defined as the difference between the revenue from annual sales, As, and the total annual cost, ATE, required to produce and sell the product (excluding depreciation). ATE, the total annual cost of production and sales, was divided into two terms [see Eq.(4.5)] the annual manufacturing cost, AME, and the annual general expense or cost, AGE.

According to Peters and Timmerhaus [3], manufacturing costs are also known as operating or production costs. Weinberger [9] drew a distinction between manufacturing costs and operating costs, assuming that manufacturing costs included all costs associated with the manufacture of a product and preparing it for shipment, while operating costs included manufacturing costs and other cost elements such as costs of sales and distribution, indirect costs of company administration, company staff services and other items of company overhead costs. According to Weinberger's interpretation, manufacturing costs are synonymous with factory costs. He included the annual depreciation charge in the manufacturing costs, which is in conflict with our terminology.

Rudd and Watson [1] defined manufacturing costs so as to absorb all costs of production and sales of the product (excluding depreciation). Happel and Jordan [63], in calculating operating costs, excluded raw materials costs, and listed the annual depreciation charge among the most important items to be accounted for in calculating manufacturing costs.

The only conclusion that can be drawn from the above discussion is that care must be taken in comparing operating and manufacturing costs drawn from different sources. We will use the term manufacturing costs to cover the following items:

— direct manufacturing costs;
— raw materials costs (including delivery);
— catalysts, solvents, fine chemicals, additives, etc.;
— operating labour costs;
— utilities costs;
— operating maintenance and repairs, operating supplies; and
— royalties, patents, licence fees.

As well as:

— indirect manufacturing costs (those that are a function of capital investment);
— payroll overhead;
— quality control laboratory;
— general plant overhead;
— packaging;
— storage;
— property taxes; and
— rent, insurance, other kinds of taxes (except for income tax).

AGE, i.e. general expense, accounts for the following items:

— administration;
— sales and shipping;
— advertising and marketing;
— technical service;
— research and development; and
— finance.

We avoid using the term 'operating cost'.

This classification of the total annual expenses of production and sales is very close to that of Peters and Timmerhaus [3]. Their book is widely used and often cited in the chemical engineering literature. The only difference is that depreciation is not considered as a component part of indirect manufacturing expenses.

Why do we calculate these costs on an annual basis rather than on a daily basis or unit-of-product basis? Among the several reasons the most important ones are:

— Manufacturing and trading accounts are based on an accountancy period of one year. To compare the estimated manufacturing costs of a new product with any other products of the same company (or with the same product of another company) the most convenient way is to follow this rule.
— Seasonal variations can be smoothed out.
— Infrequently occurring but large expenses, such operation at less than full capacity and plant on-stream time, can easily be considered.

*Direct manufacturing costs*

(Raw materials, catalysts, solvents, fine chemicals, additives, etc.)

This is one of the major cost components in calculating manufacturing costs in the chemical and petrochemical industries, so the estimator should pay careful attention to this item. The quantity of raw materials and auxiliary materials (catalysts, solvents) has to be calculated from process material and energy balances. There are several computer programs to aid material and energy balancing in the chemical and petrochemical industries, primarily for the steady state of

continuous processes ('Flowsheeting programs', see Chapter 3). Once material consumption per unit of product or per unit of time has been established, we need some price data. The sources of price data can be one of the following:

— price quotations from suppliers;
— published market prices;
— in case of materials available from sources within the company or sold to markets controlled by the producing company (often referred to as captive materials [64]), data can be extracted from company sources, in consultation with the company's accountants.

Price quotations from potential suppliers are always preferred to published market prices. The prices of materials purchased on the open market are published regularly in journals such as Chemical and Engineering News, or The Oil, Paint and Drug Reporter. There are some factors to be considered when applying published market prices:

— freight and transportation costs must be included, and should be based on the form in which the materials are to be purchased in the final plant. Bulk shipments are cheaper than shipments in smaller containers, but require greater storage capacity and inventory [3];
— when competition exists among different suppliers, considerable discounts may be negotiated;
— if materials are purchased on the basis of long term contracts, the prices are often different from market prices (usually lower). The estimator should consult company officials about the existence of such contracts [63].

In the chemical industries the ratio of raw materials costs to total annual costs of production and sale is in the range of 0.1 to 0.5.

*Operating labour*

We have to estimate the number of operators required and the wages to be paid.

Prediction of the operating labour requirement of a project is a rather difficult task, unless the estimator has had extensive experience of operating similar plants, or has the assistance of some company employment records. There is some literature information on the operating labour requirement of several chemical processes, e.g. in Ref. [4]. In the absence of these sources, the estimator has to study the engineering flowsheet and employ some rule-of-thumb. One of the most widely used rules is that published by Wessel as early as 1952. His method has been recommended by Rudd and Watson [1], Peters and Timmerhaus [3], Holland *et al.* [26] and Happel and Jordan [63].

Wessel presented his operating labour requirement data for the U.S. chemical industry as a plot of man-hours per ton per processing step *vs.* plant capacity in tons per day. In this method, a process step is defined as any unit operation, unit

process, or combination thereof, which takes place in one or more units of integrated equipment on a repetitive cycle or continuously; e.g. reaction, distillation, drying, filtration, etc. There are three curves on the plot, representing three levels of automation and complexity: one for multiple small units for increasing capacity or completely batch processes (curve $A$), one for average conditions (curve $B$) and one for large equipment, highly automated, or fluid processing only (curve $C$).

Holland *et al.* [26] published a regression equation to represent the curves developed by Wessel:

$$\log Y = 0.783109X + 1.252 + B \tag{4.107}$$

where $Y$ is the operating labour (in man-hours per ton per processing step),
$X$ is plant capacity, tons/day;
$B$ is a constant representing the level of complexity and automation:

$$\begin{aligned} B &= \ \ 0.132 \text{ for curve } A \\ &= \ \ 0 \quad\ \ \text{for curve } B \\ &= -0.167 \text{ for curve } C \end{aligned}$$

The relationship between labour requirements and production rate is not linear. When plant capacities are scaled up or down, a 0.2 to 0.25 power of the capacity ratio is suggested [3].

The hourly wage rate can be set equal to the wage rate actually paid in the company, or can be extracted from literature sources. Hourly wage rates of operating labour at different locations and for different industries are published reguarly by the U.S. Bureau of Labor in the 'Monthly Labor Review'. For chemical processes, operating labour is usually the second largest cost item in manufacturing costs, but it seldom exceeds 15% of the total production costs. The importance of the correct estimation of operating labour costs is emphasised by the fact that some other manufacturing cost items are often related to the direct operating labour costs. To begin with, the supervision costs for operating labour are from 10% to 25% of the direct operating labour costs [26]. The greater the complexity of operation the higher is the percentage of the cost of supervision. In preliminary estimates, 15% is a fairly good approximation.

*Utilities costs*

This item includes the costs of fuel, steam, electricity, compressed air, process water, cooling water, refrigeration, and some special facilities supplied to the process from outside sources. The consumption of utilities has to be calculated from process material and energy balances. There are several computer programs to aid material and energy balancing in the chemical and petrochemical industries, primarily for steady state material and energy balances of continuous processes ('Flowsheeting programs', see Chapter 3). Once utilities consumption per unit of product or per unit of time has been established, we need some price data. The

sources of price data can be one of the following:

— company records of unit costs of utilities;
— for preliminary estimation, utilities may be priced as if they were purchased from an outside company.

Happel and Jordan [63] note that the rapid change in present day energy costs makes it impossible to fix a precise value on the cost of any utility.

There is no generally accepted short cut method for estimating utilities costs. When they are relatively small, an intelligent guess based on known costs for similar processes is often satisfactory. Published data for the utilities consumption per unit of production for various processes may also be used.

*Maintenance, repairs and operating supplies*

Maintenance and repairs costs may amount to a considerable sum in chemical plants, if the process is to be kept in good operating condition. These costs involve the cost of labour, materials, and supervision. Approximately half of the maintenance cost is for materials and half for labour and supervision [26].

Maintenance and repairs costs are generally related to the capital costs of the equipment and all sources agree on an average value of 6% [3, 26, 63]. The percentage may vary from 2 to 20% depending on the severity of the process. This is well illustrated by the following table.

| Type of operation | Maintenance cost as % of fixed capital investment |
|---|---|
| Simple process, mild operating conditions | 2–6 |
| Average process | 4–9 |
| Complicated process severe operating conditions (high pressure, temperature, highly corrosive process) | 7–11 |

Happel and Jordan [63] emphasize that maintenance costs vary with plant age. Maintenance demand is usually higher during the startup of a new plant, when minor changes are being made in equipment; when the operators are unfamiliar with the process, and so on. During the steady operation of the plant these costs are lower, but towards the end of the useful life of the plant maintenance costs will probably rise again.

Operating supplies are often related to the total costs of maintenance and repairs, and amount to about 15% thereof. This item covers the costs of lubricants, test chemicals, janitor supplies, charts and similar supplies, which cannot be classified as raw materials.

*Royalties, patents, licence fees*

Royalties are not generally considered in exploratory-type estimates since, at this stage, a patent search has not yet been made, except when the product itself is patented. For preliminary estimates 1–5% of sales price may be assumed in the absence of specific knowledge [9, 26].

*Indirect manufacturing costs*

The most reliable sources of indirect costs are company records of similar or identical processes.

*Payroll overhead*

This item accounts for such costs as state unemployment and disability payments, vacations, sick leaves, pensions, group insurance, accident and liability insurance, etc. These are the so-called 'fringe benefits' attached to the wage bill. This item is thus related to the direct labour costs, and amounts to 20–35% thereof [63].

*Control laboratory*

Chemical and petrochemical plants generally have laboratories for chemical analysis, testing of materials, and process improvement. The operating cost of such laboratories is about 10–25% of the direct operating labour costs [26, 63].

*General plant overhead*

This item includes the costs of medical and hospital services, safety services, restaurant and recreational facilities, effluent disposal and warehousing. Plant overhead costs may vary between 50 to 150% of the direct operating labour costs. Holland [26] notes that the larger the plant, the lower the plant overhead per unit of production.

*Packaging*

Packaging costs depend on the nature of the product, as well as its use, value and the unit size of the packaged product. It can be as high as one-third of the selling price for cosmetics and pharmaceuticals [26], and can be as low as 0.5 cents/lb for products shipped in railroad tank-cars [63].

*Storage*

This item accounts for the costs of storerooms of every kind: for tools, spare parts such as pumps, valves, piping, for protective garments, laboratory glassware, etc. This cost is generally related to the direct operating labour costs and may amount to 20% thereof [63].

*Property taxes*

Every plant must pay taxes to the local municipality. In the absence of specific data, this item may be taken as 2% of the fixed capital cost of the plant. Property taxes generally depend on location.

*Rent, insurance, other kinds of taxes*

Rent must be paid on land and buildings, and its yearly figure may amount to 8–10% of the value of the rented property.

Every chemical plant must have some insurance as protection against loss of property and income as a result of fire or other destructive events. Location and the hazardous nature of the materials both influence the cost of insurance. This item may normally amount to 1% of the fixed capital cost of the plant.

*General expenses*

These expenses cannot be charged directly to manufacturing costs, but it is necessary to include them in the complete economic analysis. Unfortunately, there is no single useful figure to account for this item.

*Administration*

This item contains the expenses associated with top management and administration (salaries and wages for administrators, secretaries, accountants, typists, etc., costs for office supplies and equipment, outside communications, administrative buildings). In the absence of specific information from company records or company standards, this cost may be taken as 40–60% of the operating labour [3].

*Sales shipping, advertising and marketing, technical sales service*

This item involves the costs of selling the product (expenses of sales and marketing offices, salaries, commissions and travelling expenses for salesmen, shipping expenses, cost of containers, technical sales service and advertising).

These costs are very sensitive to the nature of the product, to company policy, to the market position of the company, to the relative novelty of the product, etc. A probable range for these expenses is about 2 to 20% of the total product costs.

The higher figure applies to new products or to products sold in small quantities, whereas the lower one applies to well-known, high-volume products, such as bulk chemicals [3].

*Research and development costs*

To remain competitive in the market for chemical products, there is a constant need for new methods and products in the chemical industries. The costs incurred as a result of this research work typically include salaries and wages of all personnel engaged in this type of work, costs of materials and supplies, expenses of equipment and machinery involved and direct overhead costs. A useful figure for this item in the chemical industries is 2–5% of the sales income [3].

*Finance*

Interest must be paid on the borrowed capital. Its value is fixed at the time the capital is borrowed. In 1968 in the U.S.A. it was about 4 to 7% [3], in the eighties it was as high as 12–14%. If the investment is supplied from existing company funds, it is debatable whether interest should be charged as a cost. For income-tax statements definitely not, but for economic evaluation of different process alternatives, this is highly recommended.

*Rapid manufacturing cost estimates*

The manufacturing cost of a product can be regarded as the sum of two main components: materials costs, and conversion (or processing) cost. Berry [61] proposed a method for estimating the conversion costs based on the number of processing steps necessary to convert the raw materials to the desired product. The number of processing steps is a measure of complexity. The definition of a processing step is the same as was introduced in estimating the operating labour requirement by the Wessel method, so in this sense it is an extension of the Wessel's technique.

He introduced the following relation:

$$G = g \times P / C^{1-n} \tag{4.108}$$

where $G$ is the conversion cost (\$/lb of product),
$g$ is an empirical constant,
$P$ is the number of processing steps,
$C$ is the capacity (lb/y),
$n$ is the scaling factor (usually 0.3–0.6).

$g$ and $n$ are to be obtained by correlating values for a group of processes at various capacities.

The above equation shows that conversion costs decrease with increasing capacity, and are proportional to the number of processing steps. In using this method,

Berry notes that when adding processing steps together, the scale of product throughput must be roughly the same throughout the process. If production capacity is not within ±30%, as in the separate manufacture of two intermediates that are combined to make a final product, the costs of intermediate products must be calculated and a proportional part of that unit cost must be introduced, according to the amount of intermediate used.

Holland [26] proposed the 'good old' concept of breaking down the manufacturing costs into three parts (see also Ref.[1]):

$$\text{AME} = \text{FC} + \text{FL} + \text{FQ} = a \times \text{CFC} + b \times cl \times L + (u + m) \times R \qquad (4.109)$$

where FC is the fixed investment term,
FL is the labour term,
FQ is the production rate term,
CFC is the fixed capital investment,
$L$ is the operating labour requirement (man-h/yr),
$cl$ is hourly wage rate, \$/h,
$R$ is the production rate, t/yr,
$a$ is the investment related factor expressing the ratio of the investment related costs to the fixed investment costs,
$b$ is the labour related factor expressing the ratio of the labour related costs to the direct labour costs,
$u$ is the utility cost per unit of production,
$m$ is the materials cost per unit of production (this value should absorb the cost of raw materials, chemicals, catalysts, solvents, and some other costs proportional to production rate).

Typical investment related terms are:

the maintenance and repairs costs,
property taxes,
insurance,
safety expenses (fire protection, police, etc.),
general services (laboratory, roads, etc.),
administration services (office, accounting, legal, etc.).

Labour related costs are:

direct operating labour,
supervision,
payroll overhead, etc.

Production rate related costs are:
raw materials costs,
utilities costs,
some part of the maintenance costs,

chemicals, catalysts, solvents,
warehouse and shipping, quality control,
routine analysis,
royalties and fees.

Typical values for $a$ and $b$ are 0.09 and 2.25, respectively, whereas $m$ and $u$ have to be estimated separately.

## 4.4 Capital cost estimation by CAPCOS

### 4.4.1 The program CAPCOS

CAPCOS [63] is a computer program, written in FORTRAN 77, for the estimation of fixed capital costs in process plants using Guthrie's method. The user must have a full list of equipment required for the plant, together with its main size parameters. The yearly material, labour and indirect cost escalators from the base year (1968) must also be given. The output is in the form of two tables. The first one contains the estimated fob cost, material, labour and indirect cost for each cost item entered. The second table gives further details of the additional costs for the equipment belonging to the chemical processing module.

The program is able to work in interactive mode (in which the program interrogates the user about the input data) as well as in batch mode (in which an input file containing all input data has to be created by a line editor or a previous interactive run).

In interactive mode there is an option to save all input data in a file. This file has the format required for the batch mode, hence modified cost estimates can be made by making the appropriate alterations to this file and running it through the program in batch mode once again.

In interactive mode, at each program level, various help texts are available.

The program uses an easy to read keyword-driven input language in both interactive and batch modes.

The user is free to define the equipment parameters (pressure, temperature, length, absorbed heat duty, etc.) in any commonly used unit (e.g. SI, Imperial, CGS, etc.). The units can be redefined for each new item of equipment in the same calculation.

To meet the needs of industry, no estimation data is built into the program, but it is kept in separate files. For each equipment group the appropritate estimation method (including the list of necessary size parameters, their limits, defaults and units, the estimation functions, as well as the help texts belonging to the group, etc.) is described in a language developed especially for this purpose. The program CAPROS acquires the actual method by reading in the appropriate file. This feature confers the capability of altering the cost data and estimation methods

without having to change the program. The data description language is based on keywords. It allows any estimation method to be included in CAPCOS, even though its presentation may differ completely from the original Guthrie presentation. The creation of new methods as well as updating the old ones is recommended only for advanced users, since it requires a thorough understanding of the syntax and logic of the estimation method description language.

For further details see the CAPCOS Program Manual [63].

### 4.4.2 A complex example to illustrate capital cost estimation by CAPCOS

To illustrate capital cost estimation by CAPCOS we use a set of 'portfolio-problems' published by Powers [65]. The flowsheets were generated by a process synthesis program, called AIDES.

Process synthesis is a means of generating a set of plausible process flowsheets. These flowsheets must be examined in more detail by economic evaluation methods in order to select the process flowsheets upon which detailed engineering studies are to be conducted.

AIDES accepts chemical processing problems in a general format and creates a set of process flowsheets. The program asks the user for problem input data, for required physical and chemical property information, and then generates and evaluates various process alternatives. It interacts with the designer when making decisions, and maintains a record for each process alternative. Detailed description of the program and examples of the methods employed are available in the literature [63]. Powers [65] used this program to generate process flowsheets for the production of monochlorodecane by the direct chlorination of *n*-decane.

With the help of the AIDES program, Powers synthesized over 100 process flowsheets. From this set of (plausible) flowsheets he selected eight for further evaluation (partial optimization, sensitivity testing for the fractional conversion of monochlorodecane and the price of chlorine). We have used some of the flowsheets generated by Powers as a set of 'portfolio-problems' to test the CAPCOS program. Since CAPCOS needs some information about the specific equipment parameters, and Powers gave none of them, a special interactive process simulation and design program, PPDC (Preliminary Process Design and Costing, Ref. [66]) was used to derive the necessary design information for capital cost estimation.

General description of the sample problem:

There are 5 'main' chemical components in the technology:

*n*-decane (DEC)
monochlorodecane (MCD)
dichlorodecane (DEC)
chlorine (C12)
hydrogen chloride (HCl)

The following reactions are involved:

$$\underset{(DEC)}{n\text{-decane}} + \underset{(Cl_2)}{Cl_2} \rightarrow \underset{(MCD)}{\text{monochloro-decane}} + \underset{(HCl)}{HCl}$$

$$\underset{(MCD)}{\text{monochloro-decane}} + \underset{(Cl_2)}{Cl_2} \rightarrow \underset{(DCD)}{\text{dichloro-decane}} + \underset{(HCl)}{HCl}$$

The alternative processing schemes (see *Figs 4.12* and *4.13*) differ from each other in:

— species allocation, and
— separation schemes applied to separate the reaction products.

Flowsheet 1 *(Fig. 4.12)* represents the situation in which neither unreacted chlorine, nor unreacted decane is recycled. Any dichlorodecane formed in the second reaction remains in the monochlorodecane product.

Flowsheet 2 *(Fig. 4.12)* is the same as flowsheet 1, except that unreacted decane is recycled as a head-product of the distillation column.

Flowsheet 3 *(Fig. 4.12)* represents a situation in which both unreacted chlorine and unreacted decane are recycled. Unreacted chlorine is absorbed by the fresh

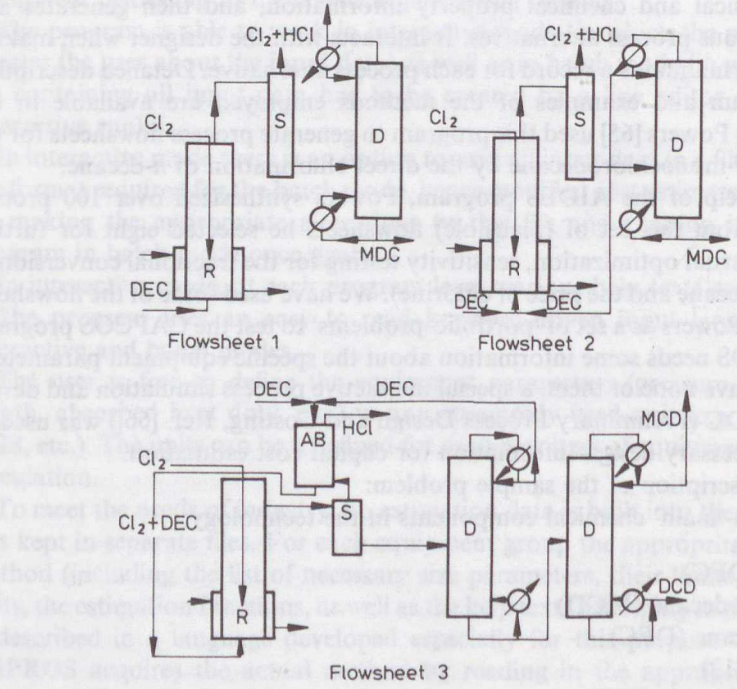
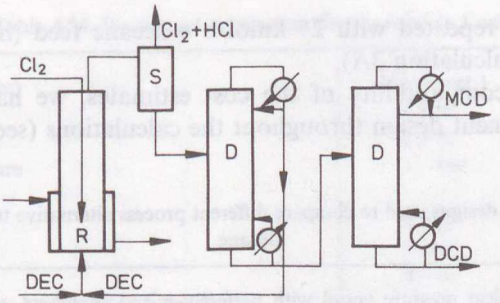
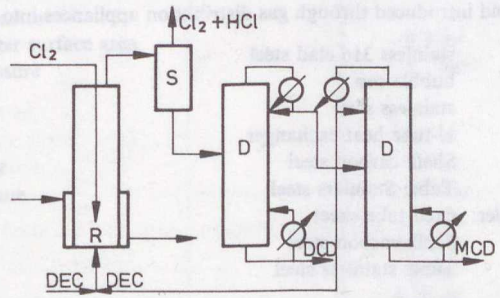



Fig. 4.12. Alternative processing schemes to produce monochlorodecane




Flowsheet 4




Flowsheet 5

Fig. 4.13. Alternative processing schemes to produce monochlorodecane

*n*-decane feed and the recycled head product (containing unreacted decane) from the first distillation column. Monochlorodecane is the head product of the second column, while the dichlorodecane by-product is driven off as the bottom product of the same column.

Flowsheet 4 and Flowsheet 5 *(Fig. 4.13)* represent a situation in which unreacted chlorine is discarded, while unreacted *n*-decane is recycled.

Flowsheets 4 and 5 differ only in the sequence and role of the two distillation columns. According to Flowsheet 4 unreacted decane is separated as head product in the first column, in contrast to Flowsheet 5, according to which it is recycled as the head of the second column. In Flowsheet 4 monochlorodecane is obtained as head product, while in the scheme depicted in Flowsheet 5 MCD is the bottom product of the second distillation column. A comparison of these two alternatives implies the comparison of two different separation sequences.

The question is whether it is better to apply the monochlorodecane–decane separation first, or whether is it better to transfer the monochlorodecane–dichlorodecane separation to the first place.

The flowsheeting and design calculations were made by program PPDC, supposing 50 kmol/h *n*-decane feed and approximately 2 : 1 *n*-decane-chlorine molar ratio. To study the effect of plant capacity on the capital required, the calculations for

Flowsheet 3 were repeated with 25 kmol/h *n*-decane feed (half of the original production rate, calculation 3A).

To ensure the comparability of the cost estimates, we have used the same ('standard') equipment design throughout the calculations (see *Table 4.23*).

Table 4.23. Equipment designs used to compare different process alternative to produce monochlorodecane

Reactor type: vertical pressure vessel with perfectly mixed chambers, separated by trays, equipped with preheater and interstage cooler. The gas/liquid reaction takes place in the chambers, at 120 °C and 1 bar. The *n*-decane stream flows downwards, the fresh chlorine feed is divided and introduced through gas distribution appliances into the chambers [61].

| | |
|---|---|
| Material: | stainless 316 clad steel |
| Tray type: | bubble cap |
| Material: | stainless steel |
| Preheater: | U-tube heat exchanger |
| Material: | Shell: carbon steel |
| | Tube: Stainless steel |
| Interstage cooler: | fixed tube sheet |
| Material: | Shell: carbon steel |
| | Tube: stainless steel |

Separator type: tray column (vertical pressure vessel, separated by trays)

| | |
|---|---|
| Material: | stainless 316 clad steel |
| Tray type: | valve |
| Material: | stainless steel |

Distillation towers: Tray columns with preheater, reflux condenser and reboiler

| | |
|---|---|
| Material: | stainless 316 clad steel |
| Tray type: | bubble cap |
| Material: | stainless steel |
| Preheater: | U-tube heat exchanger |
| Material: | Shell: carbon steel |
| | Tube: Stainless steel |
| Reflux condenser: | fixed tube sheet |
| Material: | Shell: carbon steel |
| | Tube: Stainless steel |
| Reboiler: | reboiler type heat exchanger |
| Material: | stainless steel |

Absorbtion train: tray column (vertical pressure vessel, separated by trays) with cooler to cool the *n*-decane solvent

| | |
|---|---|
| Material: | stainless 316 clad steel |
| Tray type: | bubble cap |
| Material: | stainless steel |
| Cooler: | fixed tube sheet |
| Material: | Shell: carbon steel |
| Tube: | stainless steel |

Table 4.24. Equipment parameters for Flowsheets 1 and 2

| | Flowsheet 1 | Flowsheet 2 |
|---|---|---|
| 1. Reactor (2 pieces) | | |
| Vessel height | 18.2 m | 18.2 m |
| diameter | 2.3 m | 2.3 m |
| Nominal pressure | 1 bar | 1 bar |
| Tray stack | | |
| height | 18.2 m | 18.2 m |
| diameter | 2.3 m | 2.3 m |
| Tray spacing | 0.4 m | 0.4 m |
| Interstage cooler built inside the reactor | | |
| heat exchanger surface area | 50.1 $m^2$ | 117.3 $m^2$ |
| nominal pressure | 1 bar | 1 bar |
| Preheater to reactor | | |
| heat exchanger surface area | 9.3 $m^2$ | 10.9 $m^2$ |
| nominal pressure | 1 bar | 1 bar |
| 2. Separator | | |
| Vessel height | 12.2 m | 12.4 m |
| diameter | 2.4 m | 2.5 m |
| Nominal pressure | 1 bar | 1 bar |
| Tray stack | | |
| height | 12.2 m | 12.4 m |
| diameter | 2.4 m | 2.5 m |
| Tray spacing | 0.76 m | 0.76 m |
| 3. Distillation column | | |
| Vessel height | 7.9 m | 12.0 m |
| diameter | 1.3 m | 2.0 m |
| Nominal pressure | 1 bar | 1 bar |
| Tray stack | | |
| height | 7.9 m | 12.0 m |
| diameter | 1.3 m | 2.0 m |
| Tray spacing | 0.3 m | 0.3 m |
| Preheater | | |
| heat exchanger surface area | 9.6 $m^2$ | 14.3 $m^2$ |
| nominal pressure | 500 psi | 500 psi |
| Reflux condenser | | |
| heat exchanger surface area | 10.2 $m^2$ | 23.6 $m^2$ |
| nominal pressure | 1 bar | 1 bar |
| Reboiler | | |
| heat exchanger surface area | 23.5 $m^2$ | 56.0 $m^2$ |
| nominal pressure | 800 psi | 800 psi |

*Tables 4.24* to *4.26* summarize the main sizes of equipment items (calculated by PPDC [66]) for the different process alternatives.

*Tables 4.27* to *4.32* summarize the capital cost calculation details and the various cost components for the flowsheet alternatives described above.

Table 4.25. Equipment parameters for Flowsheets 3 and 3A

| | Flowsheet 3 | Flowsheet 3A |
|---|---|---|
| 1. Reactor | | |
| Vessel height | 18.5 m | 29.4 m |
| | (2 pieces) | (1 piece) |
| diameter | 2.3 m | 1.8 m |
| Nominal pressure | 1 bar | 1 bar |
| Tray stack | | |
| height | 18.5 m | 29.4 m |
| diameter | 2.3 m | 1.8 m |
| Tray spacing | 0.4 m | 0.4 m |
| Interstage cooler built inside the reactor | | |
| heat exchanger surface area | 117.3 $m^2$ | 58 $m^2$ |
| nominal pressure | 1 bar | 1 bar |
| Preheater to reactor | | |
| heat exchanger surface area | 10.9 $m^2$ | 10.4 $m^2$ |
| nominal pressure | 1 bar | 1 bar |
| 2. Separator | | |
| Vessel height | 12.6 m | 10 m |
| diameter | 2.5 m | 2 m |
| Nominal pressure | 1 bar | 1 bar |
| Tray stack | | |
| height | 12.6 m | 10 m |
| diameter | 2.5 m | 2 m |
| Tray spacing | 0.76 m | 0.76 m |
| 3. Distillation column D1 | | |
| Vessel height | 12.0 m | 8.5 m |
| diameter | 2.0 m | 1.4 m |
| Nominal pressure | 1 bar | 1 bar |
| Tray stack | | |
| height | 12.0 m | 8.5 m |
| diameter | 2.0 m | 1.4 m |
| Tray spacing | 0.3 m | 0.3 m |
| Preheater | | |
| heat exchanger surface area | 14.3 $m^2$ | 10.2 $m^2$ |
| nominal pressure | 500 psi | 500 psi |
| Reflux condenser | | |
| heat exchanger surface area | 23.6 $m^2$ | 11.8 $m^2$ |
| nominal pressure | 1 bar | 1 bar |
| Reboiler | | |
| heat exchanger surface area | 56.0 $m^2$ | 27.60 $m^2$ |
| nominal pressure | 800 psi | 800 psi |
| 4. Distillation column D2 | | |
| Vessel height | 16.2 m | 11.5 m |
| diameter | 2.3 m | 1.6 m |
| Nominal pressure | 1 bar | 1 bar |
| Tray stack | | |
| height | 16.2 m | 11.5 m |
| diameter | 2.3 m | 1.6 m |
| Tray spacing | 0.3 m | 0.3 m |

Table 4.25 (continued)

| | | | |
|---|---|---|---|
| Reflux condenser | | | |
| heat exchanger surface area | 31.5 $m^2$ | | 15.8 $m^2$ |
| nominal pressure | 1 bar | | 1 bar |
| Rebolier | | | |
| heat exchanger surface area | 32.8 $m^2$ | | 16.4 $m^2$ |
| nominal pressure | 800 psi | | 800 psi |
| 5. Absorbtion train (5 towers in series) | | | |
| Vessel height | 2.0 m | each | 2.0 m |
| diameter | 0.6 m | | 0.5 m |
| Nominal pressure | 1 bar | | 1 bar |
| Tray stack | | | |
| height | 2.0 m | | 2.0 m |
| diameter | 0.6 m | | 0.5 m |
| Tray spacing | 0.4 m | | 0.4 m |
| Cooler to absorber | | | |
| heat exchanger surface area | 103.9 $m^2$ | | 51.9 $m^2$ |
| nominal pressure | 1 bar | | 1 bar |

Table 4.26. Equipment parameters for Flowsheets 4 and 5

| | Flowsheet 4 | Flowsheet 5 |
|---|---|---|
| 1. Reactor (2 pieces in series) | | |
| Vessel height | 18.2 m | 18.2 m |
| diameter | 2.3 m | 2.3 m |
| Nominal pressure | 1 bar | 1 bar |
| Tray stack | | |
| height | 18.2 m | 18.2 m |
| diameter | 2.3 m | 2.3 m |
| Tray spacing | 0.4 m | 0.4 m |
| Interstage cooler built inside the reactor | | |
| heat exchanger surface area | 117.3 $m^2$ | 117.3 $m^2$ |
| nominal pressure | 1 bar | 1 bar |
| Preheater to reactor | | |
| heat exchanger surface area | 10.9 $m^2$ | 20.9 $m^2$ |
| nominal pressure | 1 bar | 1 bar |
| 2. Separator | | |
| Vessel height | 12.4 m | 12.4 m |
| diameter | 2.5 m | 2.5 m |
| Nominal pressure | 1 bar | 1 bar |
| Tray stack | | |
| height | 12.4 m | 12.4 m |
| diameter | 2.5 m | 2.5 m |
| Tray spacing | 0.76 m | 0.76 m |

Table 4.26 (continued)

| | | |
|---|---|---|
| 3. Distillation column D1 | | |
| Vessel height | 12.0 m | 17.1 m |
| diameter | 2.0 m | 2.9 m |
| Nominal pressure | 1 bar | 1 bar |
| Tray stack | | |
| height | 12.0 m | 17.1 m |
| diameter | 2.0 m | 2.9 m |
| Tray spacing | 0.3 m | 0.3 m |
| Preheater | | |
| heat exchanger surface area | 14.3 $m^2$ | 14.3 $m^2$ |
| nominal pressure | 500 psi | 500 psi |
| Reflux condenser | | |
| heat exchanger surace area | 23.6 $m^2$ | 53 $m^2$ |
| nominal pressure | 1 bar | 1 bar |
| Reboiler | | |
| heat exchanger surface area | 56.0 $m^2$ | 67.0 $m^2$ |
| nominal pressure | 800 psi | 800 psi |
| 4. Distillation column D2 | | |
| Vessel height | 16.2 m | 15.3 m |
| diameter | 2.3 m | 1.8 m |
| Nominal pressure | 1 bar | 1 bar |
| Tray stack | | |
| height | 16.2 m | 15.3 m |
| diameter | 2.3 m | 1.8 m |
| Tray spacing | 0.3 m | 0.3 m |
| Reflux condenser | | |
| heat exchanger surface area | 31.5 $m^2$ | 41.4 $m^2$ |
| nominal pressure | 1 bar | 1 bar |
| Reboiler | | |
| heat exchanger surface area | 32.8 $m^2$ | 83.7 $m^2$ |
| nominal pressure | 800 psi | 800 psi |

*Table 4.27* is a complete output list with all of the results for Flowsheet 1. The subsequent *Tables* (*4.28* to *4.32*) contain only a summary of the main cost items.

*Table 4.33* gives a comparison of the different flowsheets.

Interpretation of the results:

(a) Flowsheet 1 is not very attractive: the capital cost (BMC)/annual capacity ratio is the highest, product purity is low.

(b) If (and only if) market specifications of the MCD product can be met by the low purity product, Flowsheet 2 is recommended. The recirculation of the unreacted decane is worthwhile.

(c) The absorption of chlorine can be justified if chlorine prices are extremly high or if environmental protection standards are very strict (Flowsheets 3 and 3A).

Table 4.27. Capital cost calculation results for Flowsheet 1 (Costs in $ 1000)

Free on board (FOB) Direct material (*M*)
Labour burden (*L*) Indirect (*I*)

| No. | Item name | *N* | FOB | *M* | *L* | *I* | Sum |
|---|---|---|---|---|---|---|---|
| 1 | Reactor | 2 | 111.3 | 51.4 | 49.1 | 59.4 | 271.2 |
| 2 | Separator | 1 | 58.2 | 26.8 | 25.6 | 31.0 | 141.7 |
| 3 | Distillation tower | 1 | 16.8 | 7.7 | 7.4 | 8.9 | 40.8 |
| 4 | Reactor tray stack | 2 | 70.2 | 0.0 | 0.0 | 0.0 | 70.2 |
| 5 | Separator tray stack | 1 | 14.9 | 0.0 | 0.0 | 0.0 | 14.9 |
| 6 | Distillation tray stack | 1 | 7.4 | 0.0 | 0.0 | 0.0 | 7.4 |
| 7 | Preheater | 1 | 3.0 | 1.4 | 1.3 | 1.9 | 7.6 |
| 8 | Interstage cooler | 1 | 10.4 | 4.1 | 3.6 | 5.5 | 23.6 |
| 9 | Distil. preheater | 1 | 5.0 | 1.5 | 1.3 | 1.9 | 9.6 |
| 10 | Condenser | 1 | 3.8 | 1.5 | 1.3 | 2.0 | 8.6 |
| 11 | Reboiler | 1 | 12.1 | 2.6 | 2.3 | 3.4 | 20.3 |
| | | | 313.0 | 97.0 | 91.9 | 114.0 | 615.9 |

Direct material cost details:

| No. | Piping | Concrete | Steel | Instr. | Electr. | Insul. | Paint | Sum |
|---|---|---|---|---|---|---|---|---|
| 1 | 29.7 | 4.9 | 4.0 | 5.7 | 2.5 | 4.0 | 0.6 | 51.4 |
| 2 | 15.5 | 2.6 | 2.1 | 3.0 | 1.3 | 2.1 | 0.3 | 26.8 |
| 3 | 4.5 | 0.7 | 0.6 | 0.9 | 0.4 | 0.6 | 0.1 | 7.7 |
| 7 | 0.9 | 0.1 | 0.1 | 0.2 | 0.0 | 0.1 | 0.0 | 1.4 |
| 8 | 2.6 | 0.3 | 0.2 | 0.6 | 0.1 | 0.3 | 0.0 | 4.1 |
| 9 | 0.9 | 0.1 | 0.1 | 0.2 | 0.0 | 0.1 | 0.0 | 1.5 |
| 10 | 1.0 | 0.1 | 0.1 | 0.2 | 0.0 | 0.1 | 0.0 | 1.5 |
| 11 | 1.6 | 0.2 | 0.1 | 0.4 | 0.1 | 0.2 | 0.0 | 2.6 |
| | 56.7 | 9.0 | 7.1 | 11.1 | 4.4 | 7.4 | 1.2 | 97.0 |

Labour cost details

| No. | Material erection | Equipment setting | Sum |
|---|---|---|---|
| 1 | 41.6 | 7.5 | 49.1 |
| 2 | 21.7 | 3.9 | 25.6 |
| 3 | 6.3 | 1.1 | 7.4 |
| 7 | 1.1 | 0.2 | 1.3 |
| 8 | 3.2 | 0.4 | 3.6 |
| 9 | 1.1 | 0.2 | 1.3 |
| 10 | 1.2 | 0.2 | 1.3 |
| 11 | 2.0 | 0.3 | 2.3 |
| | 78.1 | 13.8 | 91.9 |

Indirect cost details
(freight, insurance, taxes: fit)

| No. | Fit | Indirect cost | Sum |
|---|---|---|---|
| 1 | 4.8 | 55.4 | 59.4 |
| 2 | 2.1 | 29.0 | 31.0 |
| 3 | 0.6 | 8.3 | 8.9 |
| 7 | 0.2 | 1.7 | 1.9 |
| 8 | 0.5 | 5.0 | 5.5 |
| 9 | 0.2 | 1.8 | 1.9 |
| 10 | 0.2 | 1.8 | 2.0 |
| 10 | 0.3 | 3.1 | 3.4 |
| | 8.9 | 106.1 | 114.0 |

Table 4.28. Capital cost calculation results for Flowsheet 2

Monochlorodecane plant 2
Cost in 1000 $:
Free on board (FOB) Direct material (*M*)
Labour burden (*L*) Indirect (*I*)

| No. | Item name | *N* | FOB | *M* | *L* | *I* | Sum |
|---|---|---|---|---|---|---|---|
| 1 | Reactor | 2 | 111.3 | 51.4 | 49.1 | 59.4 | 271.1 |
| 2 | Separator | 1 | 60.7 | 28.0 | 26.8 | 32.4 | 147.8 |
| 3 | Distillation tower | 1 | 35.7 | 16.5 | 15.7 | 19.0 | 86.9 |
| 4 | Reactor tray stack | 2 | 70.2 | 0.0 | 0.0 | 0.0 | 70.2 |
| 5 | Separator tray | 1 | 20.2 | 0.0 | 0.0 | 0.0 | 20.2 |
| 6 | Distillation tray stack | 1 | 20.2 | 0.0 | 0.0 | 0.0 | 20.2 |
| 7 | Preheater | 1 | 3.3 | 1.6 | 1.4 | 2.1 | 8.4 |
| 8 | Interstage cooler | 1 | 22.1 | 7.0 | 6.2 | 9.3 | 44.7 |
| 9 | Distil. preheater | 1 | 6.4 | 1.9 | 1.6 | 2.5 | 12.4 |
| 10 | Condenser | 1 | 6.4 | 2.6 | 2.3 | 3.4 | 14.6 |
| 11 | Reboiler | 1 | 26.4 | 4.4 | 3.9 | 5.9 | 40.6 |
| | | | 382.9 | 113.2 | 106.9 | 133.9 | 737.0 |

Table 4.29. Capital cost calculation results for Flowsheet 3

Monochlorodecane plant 3
Costs in 1000 $:
Free on board (FOB) Direct material (*M*)
Labour burden (*L*) Indirect (*I*)

| No. | Item name | *N* | FOB | *M* | *L* | *I* | Sum |
|---|---|---|---|---|---|---|---|
| 1 | Reactor | 2 | 143.6 | 66.2 | 63.3 | 76.6 | 349.8 |
| 2 | Separator | 1 | 60.7 | 28.0 | 26.8 | 32.4 | 147.8 |
| 3 | Distillation tower D1 | 1 | 35.7 | 16.5 | 15.7 | 19.0 | 86.9 |
| 4 | Distillation tower D2 | 1 | 55.7 | 25.7 | 24.5 | 29.7 | 135.5 |
| 5 | Absorbtion train | 5 | 12.5 | 5.8 | 5.5 | 6.7 | 30.5 |
| 6 | Reactor tray stack | 2 | 71.4 | 0.0 | 0.0 | 0.0 | 71.4 |
| 7 | Separator tray stack | 1 | 20.2 | 0.0 | 0.0 | 0.0 | 20.2 |
| 8 | Distil D1 tray stack | 1 | 20.2 | 0.0 | 0.0 | 0.0 | 20.2 |
| 9 | Distil D2 tray stack | 1 | 35.9 | 0.0 | 0.0 | 0.0 | 35.9 |
| 10 | Absorber tray stack | 5 | 2.7 | 0.0 | 0.0 | 0.0 | 2.7 |
| 11 | Preheater | 1 | 3.3 | 1.6 | 1.4 | 2.1 | 8.4 |
| 12 | Interstage cooler | 1 | 22.1 | 7.0 | 6.2 | 9.3 | 44.7 |
| 13 | Distil preheater | 1 | 6.4 | 1.9 | 1.6 | 2.5 | 12.4 |
| 14 | Distil. D1 condenser | 1 | 6.4 | 2.6 | 2.3 | 3.4 | 14.6 |
| 15 | Distil. D2 condenser | 1 | 7.7 | 3.1 | 2.7 | 4.1 | 17.5 |
| 16 | Distil. D1 reboiler | 1 | 26.4 | 4.4 | 3.9 | 5.9 | 40.6 |
| 17 | Distil. D2 reboiler | 1 | 14.9 | 3.2 | 2.8 | 4.2 | 25.0 |
| 18 | Cooler to absorber | 1 | 20.5 | 6.5 | 5.7 | 8.6 | 41.4 |
| | | | 566.2 | 172.3 | 162.5 | 204.4 | 1105.4 |

Table 4.30. Capital cost calculation results for Flowsheet 3A

Note: Flowsheet 3A is the same as Flowsheet 3 but with half of the original throughput

Monochlorodecane plant 3A
Costs in 1000 $
Free on board (FOB) Direct material (*M*)
Labour burden (*L*) Indirect (*I*)

| No. | Item name | *N* | FOB | *M* | *L* | *I* | Sum |
|---|---|---|---|---|---|---|---|
| 1 | Reactor | 1 | 65.6 | 30.2 | 28.9 | 35.0 | 159.7 |
| 2 | Separator | 1 | 35.7 | 16.5 | 15.7 | 19.0 | 86.9 |
| 3 | Distillation tower D1 | 1 | 18.1 | 8.4 | 8.0 | 9.7 | 44.1 |
| 4 | Distillation tower D2 | 1 | 28.3 | 13.0 | 12.5 | 15.1 | 68.9 |
| 5 | Absorbtion train | 5 | 10.3 | 4.8 | 4.6 | 5.5 | 25.2 |
| 6 | Reactor tray stack | 1 | 33.1 | 0.0 | 0.0 | 0.0 | 33.1 |
| 7 | Separator tray stack | 1 | 9.2 | 0.0 | 0.0 | 0.0 | 9.2 |
| 8 | Distil D1 tray stack | 1 | 7.9 | 0.0 | 0.0 | 0.0 | 7.9 |
| 9 | Distil D2 tray stack | 1 | 14.4 | 0.0 | 0.0 | 0.0 | 14.4 |
| 10 | Absorber tray stack | 5 | 2.7 | 0.0 | 0.0 | 0.0 | 2.7 |
| 11 | Preheater | 1 | 3.2 | 1.5 | 1.4 | 2.0 | 8.2 |
| 12 | Interstage cooler | 1 | 11.4 | 4.5 | 4.0 | 6.0 | 25.8 |
| 13 | Distil. preheater | 1 | 5.2 | 1.5 | 1.3 | 2.0 | 10.0 |
| 14 | Distil D1 condenser | 1 | 4.1 | 1.7 | 1.5 | 2.2 | 9.4 |
| 15 | Distil D2 condenser | 1 | 5.0 | 2.0 | 1.9 | 2.6 | 11.3 |
| 16 | Distil D1 reboiler | 1 | 13.4 | 2.8 | 2.5 | 3.7 | 22.5 |
| 17 | Distil D2 reboiler | 1 | 9.7 | 2.0 | 1.8 | 2.7 | 16.2 |
| 18 | Cooler to absorber | 1 | 10.6 | 4.2 | 3.7 | 5.6 | 24.1 |
| | | | 287.7 | 93.1 | 87.5 | 111.1 | 579.5 |

Table 4.31. Capital cost calculation results for Flowsheet 4

Monochlorodecane plant 4
Costs in 1000 $:
Free on board (FOB) Direct material (*M*)
Labour burden (*L*) Indirect (*I*)

| No. | Item name | *N* | FOB | *M* | *L* | *I* | Sum |
|---|---|---|---|---|---|---|---|
| 1 | Reactor | 2 | 111.3 | 51.4 | 49.1 | 59.4 | 271.1 |
| 2 | Separator | 1 | 60.7 | 28.0 | 26.8 | 32.4 | 147.8 |
| 3 | Distillation tower D1 | 1 | 35.7 | 16.5 | 15.7 | 19.0 | 86.9 |
| 4 | Distillation tower D2 | 1 | 55.7 | 25.7 | 24.5 | 29.7 | 135.5 |
| 5 | Reactor tray stack | 2 | 69.4 | 0.0 | 0.0 | 0.0 | 69.4 |
| 6 | Separator tray stack | 1 | 19.5 | 0.0 | 0.0 | 0.0 | 19.5 |
| 7 | Distil D1 stack | 1 | 20.2 | 0.0 | 0.0 | 0.0 | 20.2 |
| 8 | Distil D2 stack | 1 | 36.4 | 0.0 | 0.0 | 0.0 | 36.4 |
| 9 | Preheater | 1 | 3.3 | 1.6 | 1.4 | 2.1 | 8.4 |

Table 4.31 (continued)

| No. | Item name | *N* | FOB | *M* | *L* | *I* | Sum |
|---|---|---|---|---|---|---|---|
| 10 | Interstage cooler | 1 | 22.1 | 7.0 | 6.2 | 9.3 | 44.7 |
| 11 | Distil preheater | 1 | 6.4 | 1.9 | 1.6 | 2.5 | 12.4 |
| 12 | Distil D1 condenser | 1 | 6.4 | 2.6 | 2.3 | 3.4 | 14.6 |
| 13 | Distil D2 condenser | 1 | 7.7 | 3.1 | 2.7 | 4.1 | 17.5 |
| 14 | Distil D1 reboiler | 1 | 26.4 | 4.4 | 3.9 | 5.9 | 40.6 |
| 15 | Distil D2 reboiler | 1 | 14.9 | 3.2 | 2.8 | 4.2 | 25.0 |
| | | | 496.1 | 145.1 | 137.0 | 171.8 | 950.0 |

(d) The decrease in capacity slightly increases the capital cost/capacity ratio (compare Flowsheet 3 with 3A).

(e) The distillation sequence represented by Flowsheet 4 is much better than that of Flowsheet 5. This is no surprise: the unreacted decane in Flowsheet 4 is passed through the first column only, which results in smaller column sizes and, consequently, lower capital cost.

Table 4.32. Capital cost calculation results for Flowsheet 5

Monochlorodecane plant 5
Costs in 1000 $:
Free on board (FOB) Direct material (*M*)
Labour burden (*L*) Indirect (*I*)

| No. | Item name | *N* | FOB | *M* | *L* | *I* | Sum |
|---|---|---|---|---|---|---|---|
| 1 | Reactor | 2 | 111.3 | 51.4 | 49.1 | 59.4 | 271.1 |
| 2 | Separator | 1 | 60.7 | 28.0 | 26.0 | 32.4 | 147.8 |
| 3 | Distillation tower D1 | 1 | 70.8 | 32.7 | 31.2 | 37.8 | 172.5 |
| 4 | Distillation tower D2 | 1 | 43.1 | 19.9 | 19.0 | 23.0 | 105.0 |
| 5 | Reactor tray stack | 2 | 69.4 | 0.0 | 0.0 | 0.0 | 69.4 |
| 6 | Separator tray stack | 1 | 19.5 | 0.0 | 0.0 | 0.0 | 19.5 |
| 7 | Distil D1 tray stack | 1 | 60.8 | 0.0 | 0.0 | 0.0 | 60.8 |
| 8 | Distil D2 tray stack | 1 | 20.0 | 0.0 | 0.0 | 0.0 | 20.0 |
| 9 | Preheater | 1 | 3.3 | 1.6 | 1.4 | 2.1 | 8.4 |
| 10 | Interstage cooler | 1 | 22.1 | 7.0 | 6.2 | 9.3 | 44.7 |
| 11 | Distil. preheater | 1 | 6.4 | 1.9 | 1.6 | 2.5 | 12.4 |
| 12 | Distil D1 condenser | 1 | 13.4 | 4.3 | 3.8 | 5.7 | 27.1 |
| 13 | Distil D2 condenser | 1 | 9.1 | 3.6 | 3.2 | 4.8 | 20.8 |
| 14 | Distil D1 reboiler | 1 | 29.6 | 4.9 | 4.4 | 6.6 | 45.4 |
| 15 | Distil D2 reboiler | 1 | 34.0 | 5.7 | 5.0 | 7.5 | 52.2 |
| | | | 573.7 | 160.9 | 151.7 | 191.0 | 1077.2 |

Table 4.33. MCD production, product purity and capital cost data summary

| Flowsheet | 1 | 2 | 3 | 3A | 4 | 5 |
|---|---|---|---|---|---|---|
| Capacity | | | | | | |
| kmol/h MCD | 17.9 | 42.9 | 41.1 | 20.6 | 41.1 | 41.1 |
| t/h | 3.81 | 9.04 | 7.38 | 3.69 | 7.38 | 7.38 |
| t/yr | 32,015 | 75,967 | 62,020 | 31,010 | 62,020 | 62,020 |
| Product composition | | | | | | |
| mol% DEC | 1.4 | 1.3 | 1.7 | 1.7 | 1.7 | 1.6 |
| MCD | 84.9 | 84.8 | 98.0 | 98.0 | 98.0 | 98.0 |
| DCD | 13.7 | 13.9 | 0.3 | 0.3 | 0.3 | 0.3 |
| Capital cost (on bare module cost level) | | | | | | |
| BMC/FOB | 616 | 737 | 1105 | 579 | 950 | 1077 |
| ($/t/yr) | 19.2 | 9.7 | 17.8 | 18.0 | 15.3 | 17.4 |

Notes
Abbrevations: DEC *n*-decane
MCD monochlorodecane
DCD dichlorodecane
Input streams: 50 kmol/h *n*-decane
25 kmol/h chlorine
Onstream time: 8400 h/yr
Flowsheet and desing calculations: PPDC [65]
Capital cost estimation: CAPCOS [62]
Equipment details: see *Tables 4.23–4.25*
Flowsheets: see *Figs 4.12–4.13*
Capital cost details: see *Tables 4.26–4.31*

## 4.5 References

[1] Rudd, D. F. and Watson, C. C. (1968): *Strategy of Process Engineering*. Wiley, New York.
[2] Rose, L. M. (1958): *Engineering Investment Decisions. Planning under Uncertainty*. Elsevier, Amsterdam.
[3] Peters, M. S. and Timmerhaus, K. D. (1968): *Plant Design and Economics for Chemical Engineers*. (2nd ed.) McGraw-Hill, New York.
[4] Chilton, C. H. (1960): *Cost Engineering in the Process Industries*. McGraw-Hill, New York.
[5] Weinberger, A. J. (1963): *Chem. Eng.*, November 25, 113.
[6] Earley, W. R. (1964): *Chem. Eng.*, March 30, 111.
[7] Miller, C. A. (1965): *Chem. Eng.*, September 13, 226.
[8] Souders, M. (1966): *Chem. Eng. Progr.*, **62** (3), 79.
[9] Weinberger, A. J. (1963): *Chem. Eng.*, December 23, 81.
[10] Weinberger, A. J. (1964): *Chem. Eng.*, March 30, 91.
[11] Westbrook, G. T. and Aris, R. (1961): *Ind. Eng. Chem.*, **53** (3), 181.
[12] Herron, D. P. (1967): *Chem. Eng.*, January 30, 125.

[13] Guthrie, K. M. (1969): *Chem. Eng.*, March 24, 119.
[14] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1973): *Chem. Eng.*, June 25, 103.
[15] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1973): *Chem. Eng.*, July 23, 118.
[16] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1973): *Chem. Eng.*, August 20, 139.
[17] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1973): *Chem. Eng.*, September 17, 123.
[18] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1973): *Chem. Eng.*, October 1, 80.
[19] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1973): *Chem. Eng.*, October 29, 115.
[20] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1973): *Chem. Eng.*, November 26, 83.
[21] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1973): *Chem. Eng.*, December 24, 61.
[22] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1974): *Chem. Eng.*, January 7, 105.
[23] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1974): *Chem. Eng.*, February 4, 73.
[24] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1974): *Chem. Eng.*, March 4, 119.
[25] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1974): *Chem. Eng.*, April 1, 71.
[26] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1974): *Chem. Eng.*, April 15, 91.
[27] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1974): *Chem. Eng.*, May 13, 105.
[28] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1974): *Chem. Eng.*, June 10, 83.
[29] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1974): *Chem. Eng.*, July 8, 93.
[30] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1974): *Chem. Eng.*, August 5, 101.
[31] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1974): *Chem. Eng.*, September 2, 62.
[32] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1974): *Chem. Eng.*, September 16, 119.
[33] Holland, F. A., Watson, F. A. and Wilkinson, J. K. (1974): *Chem. Eng.*, October 28, 107.
[34] Agarval, J. C. and Klumpar, I. V. (1975): *Chem. Eng.*, September 29, 66.
[35] Allen, D. H. and Page, R. C. (1975): *Chem. Eng.*, March 3, 142.
[36] Wild, N. H. (1976): *Chem. Eng.*, April 12, 153.
[37] Hoerner, G. M., Jr. (1976): *Chem. Eng.*, May 24, 141.
[38] Wild, N. H. (1977): *Chem. Eng.*, May 9, 137.
[39] Holland, F. A. and Watson, F. A. (1977): *Chem. Eng.*, February 14, 87.
[40] Holland, F. A. and Watson, F. A. (1977): *Chem. Eng.*, March 14, 133.
[41] Pikulik, A. and Diaz, H. E. (1977): *Chem. Eng.*, October 10, 106.
[42] Lindsley, J. (1979): *Chem. Eng.*, May 21, 201.
[43] Bridgewater, A. C. (1979): *Chem. Eng.*, November 5, 119.
[44] Miller, C. A. (1979): *Chem. Eng.*, July 2, 89.
[45] Kirkpatrick, D. M. (1979): *Chem. Eng.*, August 27, 103.
[46] Horowitz, B. A. (1980): *Chem. Eng.*, May 19, 169.
[47] Kohn, P. M. (1978): *Chem. Eng.*, May 8, 189.
[48] Ricci, L. J. (1975): *Chem. Eng.*, April 28, 117.
[49] Thorsen, D. P. (1972): *Chem. Eng.*, November 13, 168.
[50] Norden, R. B. (1969): *Chem. Eng.*, May 5, 134.
[51] Chilton, C. H. (1966): *Chem. Eng.*, April 25, 185.
[52] Arnold, T. A., Jr. and Chilton, C. H. (1963): *Chem. Eng.*, February 18, 132.
[53] Neal, G. W. (1982): *Chem. Eng.*, September 6, 131.
[54] Neal, G. W. (1982): *Chem. Eng.*, December 27, 73.
Neal, G. W. (1982): *Chem. Eng.*, November 1, 107.
[55] Matley, J. (1982): *Chem. Eng.*, April 19, 153.
[56] Harckney, J. W. (1983): *Chem. Eng.*, April 18, 99.
[57] Rubin, A. G. (1984): *Chem. Eng.*, October 15, 129.
[58] Martin, K. L. and Barna, B. H. (1982): *Chem. Eng.*, July 23, 73.
[59] Ward, Th. J. (1982): *Chem. Eng.*, September 17, 121.
[60] Horner, C. F. and Liebster, L. M. (1980): *Pocket Dictionary of Business Terms*. Newness Books, Hamlyn Publishing Group, Feltham, Middlesex.

[61] Fisher, R. H. and Yates, F. (1957): *Statistical Tables for Biological, Agricultural and Medical Research.* (5th ed.) Oliver and Boyd, London.
[62] BRD Offenlegungschrift 2,063,601 (1970).
[63] CAPCOS-Process Plant Capital Cost Estimation. User Manual, CSS–06. CHEMENG Software and Services. 69. Eastern Road, Selly Park, Birmingham, B29 7JX, England.
[64] Siirola, J. J., Powers, G. J. and Rudd, D. F. (1971): *AIChE J.*, **17** (3), 677.
[65] Powers, G. J. (1972): *Chem. Eng. Progr.*, **68** (8), August, 88.
[66] Tátrai, F. (1977): *Selection of optimal process flowsheets by computer methods.* (In Hungarian.) PhD. Thesis, Technical University of Budapest.

# 5 PRINCIPLES OF CHEMICAL PROCESS CONTROL

by

H. HAJDÚ and P. MIZSEY

## 5.1 Introduction

*The task of process control*

Modern, technically advanced chemical industry cannot work without process control.

Process control is applied in order to

(i) maintain the value of process variables eliminating external disturbances, and thereby ensuring stable operation;

(ii) vary the value of process variables as a function of a measurement or time, or according to an optimizing function;

(iii) satisfy various constraints, e.g. prevent flooding in an extraction column.

The so-called control system, which can achieve these tasks, consists of well-designed equipment (the control hardware: controllers, control valves, valve motors, measuring devices, computers, etc.), as well as trained designers and operators.

*The methods of control*

The outlet temperature of the steam-heated heat exchanger in *Fig. 5.1* is to be held constant in spite of variations in feed rate and feed temperature. This can be accomplished in two ways:

(*a*) measuring the outlet temperature and manipulating steam rate *(feedback)*;
(*b*) measuring feed temperature and manipulating steam rate *(feedforward)*.

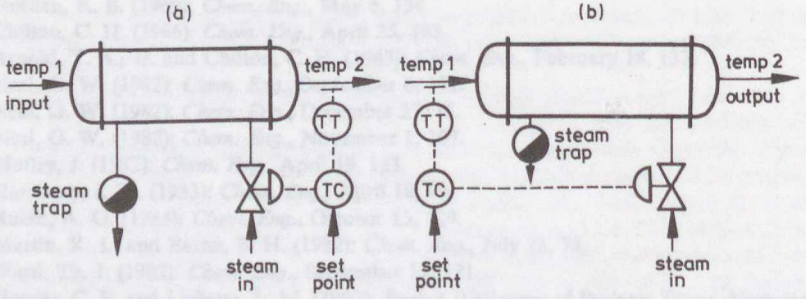



Fig. 5.1. Control of a heat exchanger: (*a*) feedback, (*b*) feedforward

In case (*a*) the measured value depends on the manipulation, in case (*b*) it does not. This is the basic difference between feedback and feedforward control.

Feedback control is able to eliminate the effects of changes in both feed temperature and feed rate, or any other disturbance (e.g. variation in steam pressure).

Feedforward control can eliminate only the effect of the disturbance on the measurement to which it applies; in some instances it does not even do that. In our heat exchanger example of *Fig. 5.1*, if the inlet temperature rises to a value for which the steam valve has to be closed down, the return of the outlet temperature to specification is not guaranteed. This is the main reason why feedforward is always used in combination with feedback.

Feedback has the disadvantage of requiring an offset in the control variable before control action. This implies the other main disadvantage of feedback: in case of erroneous application, feedback control may cause instability of the process controlled.

### *Stability*

The definition of stability is known from mechanics. A process is *stable* if after a pulse-like disturbance it returns within some time to its original state. A step-like disturbance may force the stable system into a new steady state (or back to the original steady state, see *Fig. 5.2*), but the output of an unstable process will

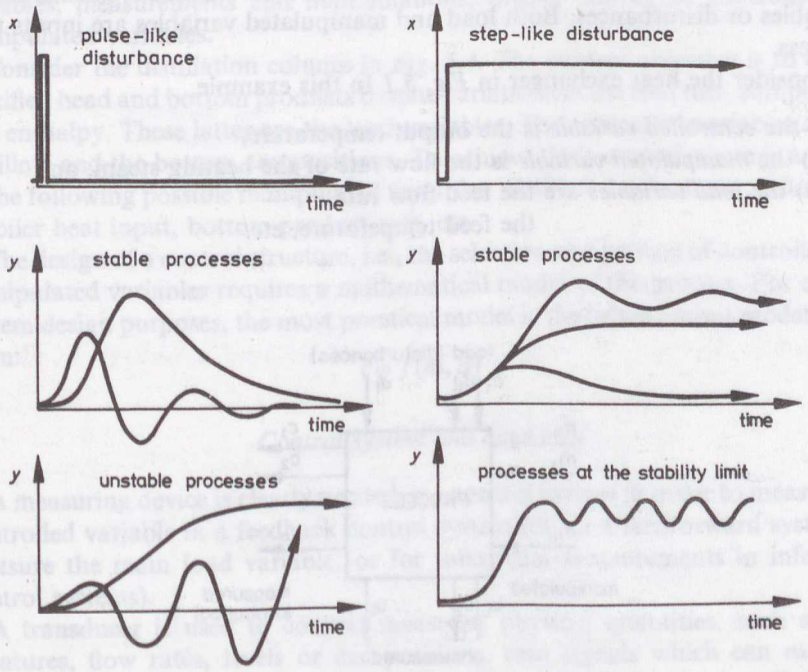



Fig. 5.2. Stable and unstable process responses

increase (or decrease) without limit (or until it reaches a natural limit, e.g. liquid level in a vessel will rise until it overflows). Some unstable processes show periodic instability: after a disturbance their output oscillates with increasing amplitude. At the stability limit the amplitude of the oscillations is constant.

Chemical processes are usually stable, but some unstable ones do exist. For example, irreversible exothermic reactions can be unstable and a reactor in which such a reaction takes place can only be operated in the steady state by the use of control.

*The controlled process and its variables*

A process has *input and output variables*: the *outputs are dependent* on the inputs, and on the characteristics of the process.

The general structure of a process is shown in *Fig. 5.3.*

The output variable to be controlled is the controlled variable. In general, it ought to be measured in order to control it; consider the output temperature of the heat exchanger in *Fig. 5.1*, for instance. However, it may occur that direct measurement of a controlled variable is impossible, e.g. the internal vapour velocity in a distillation column. In such cases, the variable is computed from inferential measurements.

The control variable is maintained at its desired value by changing a manipulated variable, thus eliminating the effect of uncontrolled variations in external load variables or disturbances. Both load and manipulated variables are inputs to the process.

Consider the heat exchanger in *Fig. 5.1* In this example,

(*i*) the *controlled variable* is the output temperature,
(*ii*) the *manipulated variable* is the flow rate of the heating steam, and
(*iii*) the *load variables* are the feed flow rate,
the feed temperature, etc.

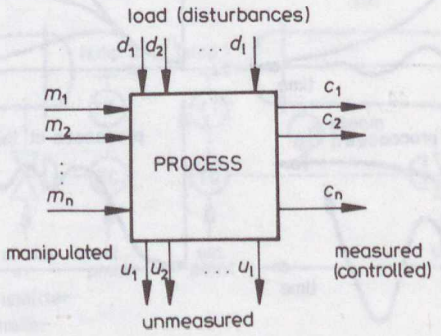



Fig. 5.3. Process and input/output variables

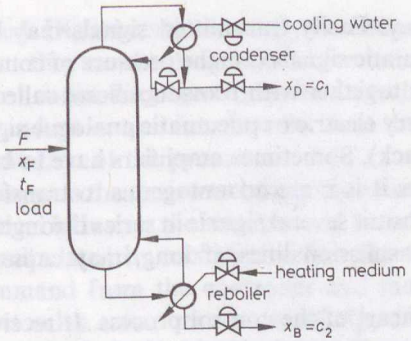



Fig. 5.4. Distillation column with possible manipulated variables

A system with one controlled output and one manipulated input is a Single Input, Single Output (SISO) system.

Many chemical processes have several controlled outputs and several possible manipulated inputs. A typical example is a distillation column. These Multiple Input, Multiple Output (MIMO) systems are a challenge to chemical engineers. They must consider the problems of: degrees of freedom (generally one controlled variable needs one manipulated variable); selection of the appropriate controlled variables, measurements and manipulations; proper pairing of controlled and manipulated variables.

Consider the distillation column in *Fig. 5.4*. The control objective is to obtain specified head and bottom products despite variations in the feed rate, composition and enthalpy. These latter are the load variables. The controlled variables are the distillate and the bottom compositions. To achieve their control, there is a choice of the following possible manipulated variables: distillate product rate, reflux rate, reboiler heat input, bottom product rate, etc.

The design of a control structure, i.e., the selection and pairing of controlled and manipulated variables requires a mathematical model of the process. For control system design purposes, the most practical model is the input-output model of the form:

$$c = f(m, d)$$

*Control system and hardware*

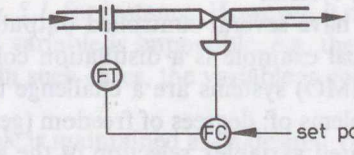
A measuring device is clearly needed in a control system in order to measure the controlled variable in a feedback control system (or, in a feedforward system, to measure the main load variable, or for inferential measurements in inferential control systems).

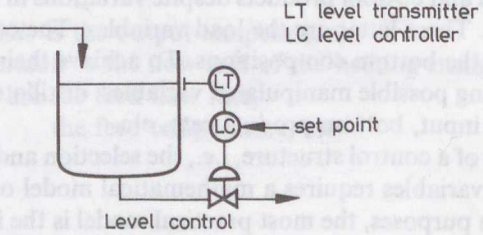
A transducer is used to convert measured physical quantities, such as temperatures, flow rates, levels or compositions, into signals which can easily be transmitted from the process to the controller, which may be installed some

hundreds of meters away. Easily transmitted signals may be electric current or voltage signals, or pneumatic signals (i.e. the pressure of compressed air or liquid). Measuring devices built together with transducers are called transmitters.

Transmission lines carry electric or pneumatic analogue signals from the process to the controller (and back). Sometimes amplifiers have to be installed into transmission lines. Sometimes it is more advantageous to transform analogue signals into digital ones and transmit several signals in series through a single transmission conduit. Pneumatic transmission lines, if long, may cause a considerable time delay.

The controller is the heart of the control process. It receives the control signal from the measuring device, compares it with its derived value, the set point, and according to the value and velocity of the error, gives a command signal to the process, telling it what manipulation is to be undertaken. The controller can be regarded as an analogue electric or pneumatic computer, on which simple control



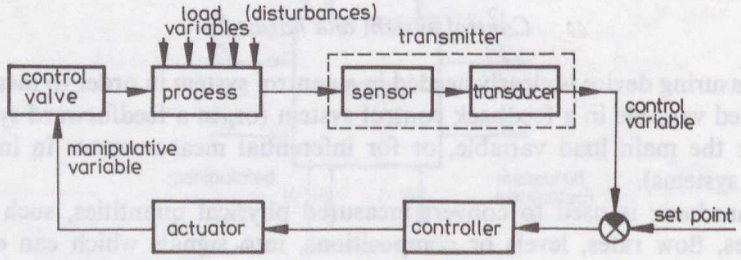




Fig. 5.5. Flow and level control loops and block diagram

laws can be implemented. Digital computers may also be used as controllers; they have the advantage that they can be programmed for more sophisticated control laws, and are capable of controlling more processes simultaneously. Digital controllers need an input and an output interface between the transmission line and the analogue instrumentation of the process.

The final control element is the device accomplishing the command given by the controller. In the chemical industry, the manipulated variable is most frequently a flow rate, which is manipulated by a control valve. The actuator of the control valve receives the command from the controller and moves the valve into the required position. Most valve actuators are pneumatic, since they are cheap and have a simple construction. Variable-speed pumps and compressors are also used as final control elements, as well as relay switches for on-off control.

A flow control and a level control system and the block diagram of a feedback control system are shown in *Fig. 5.5.*

## 5.2 Control dynamics

Feedback control is the basic regulating mechanism in control technology. However, negative feedback loops are capable of continuous oscillation and even instability — on the other hand, they may be sluggish and fail to control the controlled variable. It is thus of primary importance to understand the mechanism of feedback, and to examine its dynamics.

### 5.2.1 Basic concepts of feedback control

*Negative feedback*

Feedback may be negative or positive.

Positive feedback means that the controller works in the same direction as the disturbance. For example, in a heat exchanger where the output temperature is controlled *(Fig. 5.6)*, if the temperature is changing due to some disturbance, the

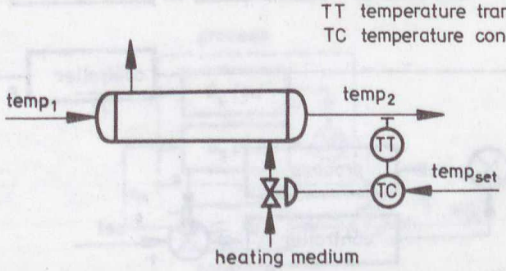



Fig. 5.6. Control of a heat exchanger

positive feedback controller will manipulate the flow of the heating medium so as to increase the temperature when it is above the set point and decrease when it is below. This feature is not conducive to regulation and causes instability.

Negative feedback works in the opposite sense. When the temperature, the controlled variable, is above the set point, i.e. it is higher than the desired value, the controller will reduce the heat input by manipulating some variable (called the manipulated variable). Of course if the temperature is below the set point it will increase the heat input. In short, negative feedback works always in the sense opposite to the measured deviation.

### *The block diagram*

*Figure 5.7* shows the block diagram and the simplified block diagram of a negative feedback loop, which may, for instance, be the heat exchanger control in *Fig. 5.6*. In the simplified block diagram the actuator, the control valve, the controlled process and the transmitter are all lumped together into the block labelled 'process'. Any classical simple control loop can be divided in this manner into two parts, the process and the controller, on the basis of the fact that the parameters of the controller are adjustable, but those of the process are not.

The mathematical treatment of networks similar to that represented in *Fig. 5.7*, and even of more complicated ones, is convenient if the signal transfer of all elements in the network can be represented by a simple linear input-output relationship of the form

$$dy = \underline{G}(\omega)dx \tag{5.1}$$

where $y$ is the system output signal,

$x$ is the system input signal,

$\underline{G}(\omega)$ is a function characterizing the signal transfer of the system. This is generally a function of the frequency of the input signal.

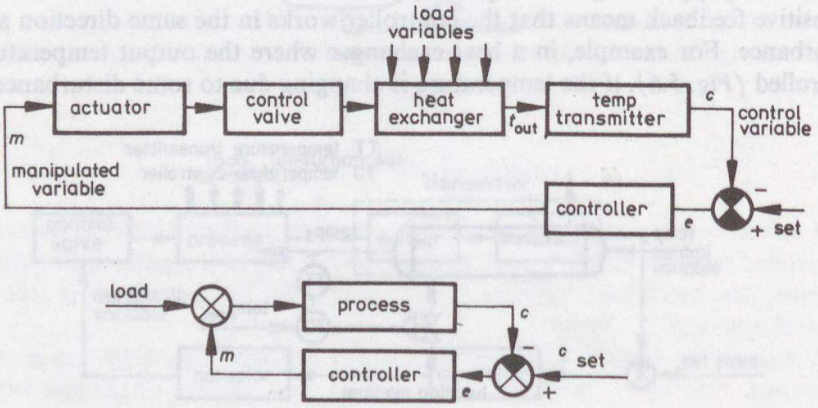



Fig. 5.7. Simplifying the block diagram

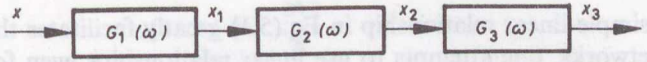



Fig. 5.8. Elements in series

In a linear network all linear transformations are valid. Thus the resulting frequency function of elements in series *(Fig. 5.8)* is the following:

$$dx_3 = G_3(\omega)dx_2 = G_3(\omega)G_2(\omega)dx_1 = G_3(\omega)G_2(\omega)G_1(\omega)dx$$

$$G(\omega) = \prod G_i(\omega) \tag{5.2}$$

The resulting frequency function of elements in parallel *(Fig. 5.9)* is:

$$dx_2 = G_1(\omega)dx_1 + G_2(\omega)dx_1 = [G_1(\omega) + G_2(\omega)]dx_1$$

$$G(\omega) = \sum G_i(\omega) \tag{5.3}$$

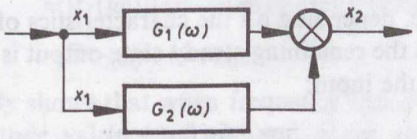



Fig. 5.9. Elements in parallel

The frequency function of a simple negative feedback control loop is derived as follows *(Fig. 5.10)*:

$$dx_S = G_L(\omega)dx_L + G_S(\omega)dx_m =$$

$$= G_L(\omega)dx_L + G_S(\omega)G_C(\omega)dx_{S,\,set} - G_S(\omega)G_C(\omega)dx_S$$

whence

$$dx_S = \frac{G_L(\omega)}{1 + G_S(\omega)G_C(\omega)}dx_L + \frac{G_S(\omega)G_C(\omega)}{1 + G_S(\omega)G_C(\omega)}dx_{S,\,set}. \tag{5.4}$$

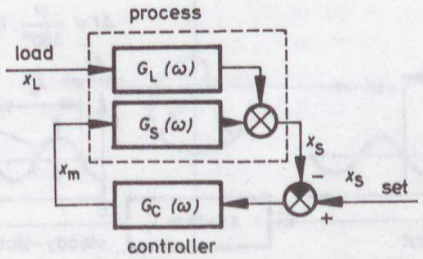



Fig. 5.10. Negative feedback control loop

Since the simple linear relationship in Eq.(5.1) greatly facilitates the treatment of control networks, one attempts to use linear relationships even for nonlinear systems, whenever possible. The technique of linearizing nonlinear input-output models will be presented in Section 5.3.3. It should be mentioned that most chemical engineering systems are nonlinear; some of them are suitable for linearization. Exceptions are, for instance, chemical reactors, where linearization may produce erroneous results.

*The frequency function*

The frequency response of a system is obtained experimentally by forcing the system, using a sinusoidal input (see *Fig. 5.11*):

$$x = a \sin \omega t \tag{5.5}$$

where $a$ is the amplitude of the sine wave (–), and
$\omega$ is the frequency of the sine wave (radians/s).

After a certain time, depending on the characteristics of the forced system, the transients die out and the remaining steady-state output is a sine wave having the same frequency $\omega$ as the input:

$$y = b \sin (\omega t + \varphi) \tag{5.6}$$

but with a different amplitude $b$, and shifted by phase angle $\varphi$ with respect to the input wave. The amplitude ratio and the phase angle are the components of $G(\omega)$, the frequency function vector:

the magnitude: $$|G(\omega)| = \frac{b}{a}(\omega) \tag{5.7}$$

the phase angle: $$\measuredangle G(\omega) = \varphi(\omega) \tag{5.8}$$

both depending on the frequency of the forcing function *(Fig. 5.11)*.

The frequency function vector can be represented graphically on a polar plot. $G(\omega)$ is conveniently expressed mathematically as a complex number, represented

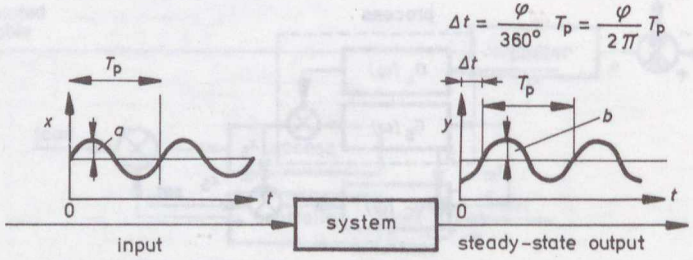



Fig. 5.11. Determination of the frequency response function

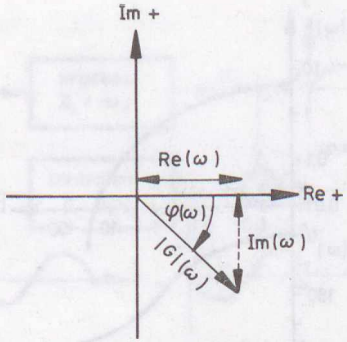



Fig. 5.12. Representation of the frequency function vector in the complex plane

in the complex plane *(Fig. 5.12)*:

$$G(i\omega) = |G(i\omega)|e^{i\varphi(i\omega)} =$$
$$= |G(i\omega)|[\cos\varphi(i\omega) + i\sin\varphi(i\omega)] = \tag{5.9}$$
$$= \mathrm{Re}_{G(i\omega)} + i\,\mathrm{Im}_{G(i\omega)}$$

This equation clearly shows that when frequency functions are in series (as in *Fig. 5.8*), then amplitude values multiply and phase shifts add. For parallel frequency functions *(Fig. 5.9)*, addition of the frequency functions is accomplished algebraically using the canonical form of the complex numbers (and graphically on a polar plot).

The most frequently used graphical representations of frequency functions over a broad frequency range are the Nyquist plot and the Bode plot. The Nyquist plot is a polar plot in which the end points of the individual vectors $G(\omega)$ are connected with a continuous line beginning at $\omega = 0$ and ending at $\omega = \infty$ *(Fig. 5.13)*. It has the drawback that frequencies $\omega$ are not indicated. The Bode plot consists of two

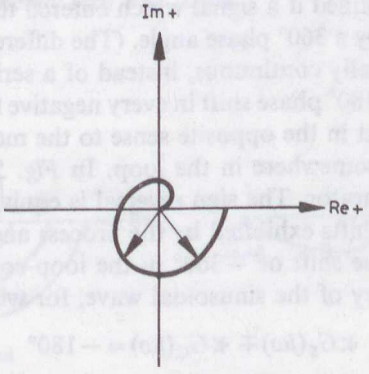



Fig. 5.13. Nyquist diagram

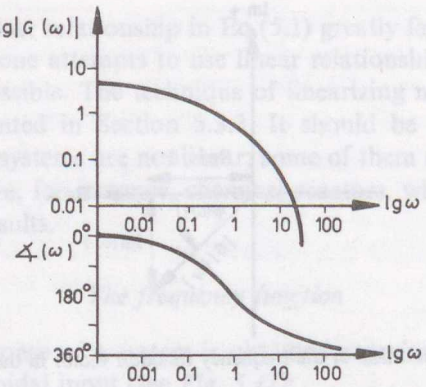



Fig. 5.14. Bode diagram

diagrams, one of which shows the logarithm of the amplitude ratio, and the other the phase lag (linear) both as a function of the logarithm of the frequency of the driving sine wave *(Fig. 5.14)*. The logarithmic scale of the amplitude diagram and the linear scale of the phase diagram are chosen to allow the multiplication of frequency functions in series by simple addition of distances in the diagram.

*Oscillation in a feedback loop*

Oscillation and resonance are characteristic of negative feedback loops. Resonance may be detrimental. A common example is the breakdown of a bridge caused by the rhythmical marching of a troop of soldiers.

The mechanism of oscillation is similar to a ball's bouncing. If a ball is being repeatedly struck (*i*) with the same force, (*ii*) at its highest position, it will continue to bounce to the same height (amplitude) at constant period. Should the force diminish, the amplitude will diminish, too. Should the ball be struck before its highest position, the period will be shortened.

In a feedback control loop a signal entering the loop circulates, as shown in *Fig. 5.15*. Oscillation is maintained if a signal which entered the loop returns with the same amplitude, shifted by a 360° phase angle. (The difference from ball bouncing is that the signal is generally continuous, instead of a series of pulses.)

There is an inherent $-180°$ phase shift in every negative feedback loop. Since the controller must always act in the opposite sense to the measured deviation, there must be a sign reversal somewhere in the loop. In *Fig. 5.15* the sign reversal is represented by the comparator. The sign reversal is equivalent to a $-180°$ phase shift. Thus, if the phase shifts exhibited by the process and the controller sum up to $-180°$, there is a phase shift of $-360°$ in the loop-condition (*ii*) for uniform oscillations. The frequency of the sinusoidal wave, for which

$$\measuredangle G_S(i\omega) + \measuredangle G_C(i\omega) = -180° \tag{5.10}$$

is the *critical frequency* of the loop.

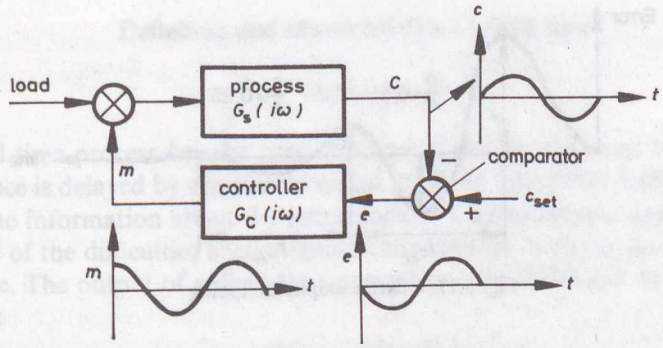



Fig. 5.15. Sign reversal in the negative feedback loop

Uniform oscillations are maintained in the control loop if the force exerted by the controller is the same after every cycle, that is, if the product of the amplitude ratios at the critical frequency is unity:

$$|G_C(i\omega_{cr})|\,|G_S(i\omega_{cr})| = 1 \tag{5.11}$$

Equations (5.10) and (5.11) can be combined into one equation, giving the representation of the vector in the complex plane:

$$G_C(i\omega_{cr})\,G_S(i\omega_{cr}) = -1 \tag{5.12}$$

This equation is a simple formulation of the Nyquist stability criterion. More exactly: for a stable loop, following the Nyquist curve of the open loop frequency function $(G_C(i\omega)\,G_S(i\omega))$ from $\omega = 0$ in the direction $\omega \to \infty$, point $(-1)$ must lie at the left. If point $(-1)$ is at the right, the loop is unstable. If the Nyquist curve traverses point $(-1)$, the loop is at the stability limit, exhibiting uniform oscillations at the critical frequency (see *Fig. 5.16*).

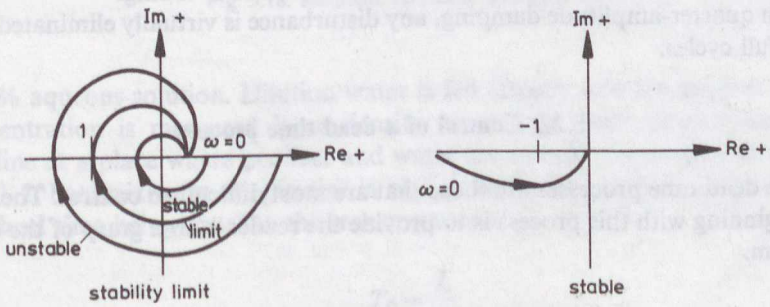



Fig. 5.16. The Nyquist stability criterion

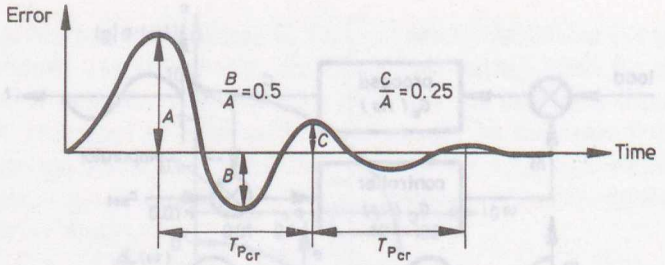



Fig. 5.17. Quarter-amplitude tuning

It is important to note that the Nyquist stability criterion uses the *open loop* frequency function to predict *closed loop stability!* Of course, it applies only to negative feedback loops.

How does a sine wave with the critical frequency enter the closed loop? Any pulse, impulse, or step signal may be decomposed into a sum of sine waves with different amplitudes, phases and frequencies. The sine components which have a frequency different from the critical frequency of the loop die out (because for these frequencies the loop gain is less than unity, or because for these frequencies the loop phase shift is different from $-360°$), and the loop will cycle with the critical frequency after a transient period. The forcing signal may enter the closed loop at any place, e.g. at the set point.

*Damping*

Damping of the oscillation (which has the critical frequency) is obtained by proper tuning of the controller. The controller is adjustable, whereas the parameters of the process are fixed.

Experience has shown that a gain product equal to 0.5 or less is generally advantageous. Then the amplitude of oscillation is attenuated by one-half each half cycle, and by one-quarter each full cycle. In *Fig. 5.17* $A/B$ is 0.5, and $A/C$, the so-called decay ratio, is 0.25. This is the so-called 'quarter-amplitude damping', often used in industry as a rule of thumb for controller tuning.

With quarter-amplitude damping, any disturbance is virtually eliminated within three full cycles.

### 5.2.2 Control of a dead time process

Pure dead time processes are those that are most difficult to control. The reason for beginning with this process is to provide the reader with a grasp of the control problem.

## Definition and characteristics of dead time

### *Definition*

A dead time process has the special property that its response to any kind of disturbance is delayed by an interval called the dead time *(Fig. 5.18)*. During this interval no information about the occurrence of a disturbance is available. This is the cause of the difficulties encountered in the control of processes dominated by dead time. The output of a dead time process may be described by the following equation:

$$y[t]=A_S x[t-T_D] \tag{5.13}$$

that is, the output, $y$ at time instant $t$ is proportional to the input at time instant $[t-T_D]$. Proportionality factor $A_S$ is the so-called steady-state gain of the process, its dimension is defined by those of $y$ and $x$ and by Eq.(5.13).

Dead time in the chemical industries is generally linked with transportation. As an example, consider a liquid product which is to be diluted with water to obtain

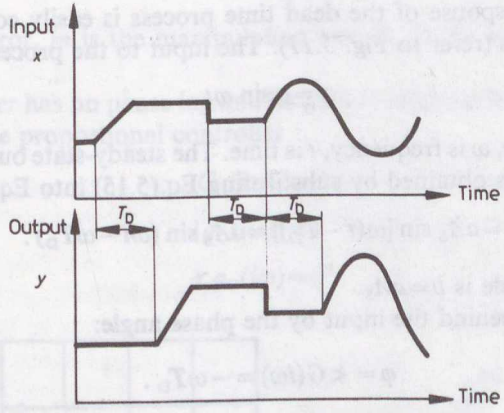



Fig. 5.18. Response of a dead time process

a 50% aqueous solution. Dilution water is fed directly into the product pipeline; concentration is measured by a density transmitter built downstream in the pipeline at a place where product and water are already thoroughly mixed *(Fig. 5.19)*. If backmixing in the pipeline is negligible, this is a pure dead time process, the dead time being equal to the transportation lag

$$T_D=\frac{L}{v} \tag{5.14}$$

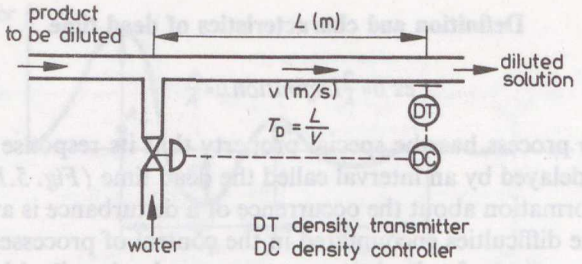



Fig. 5.19. Mixing in a pipeline is a dead-time process

The same phenomenon arises in tubular reactors, blending processes, and also in heat transfer processes; in fact, everywhere where molecules have to move through a distance in order to exert an effect. A typical dead time process is a conveyor belt transporting granulated solids. Sample pipelines to process analysers are also sources of dead time.

*Frequency function: gain and phase shift*

The frequency response of the dead time process is easily computed from its equation, as follows (refer to *Fig. 5.11*). The input to the process is a sine wave:

$$x = a \sin \omega t \tag{5.15}$$

where $a$ is amplitude, $\omega$ is frequency, $t$ is time. The steady-state output is also a sine wave. Its equation is obtained by substituting Eq.(5.15) into Eq.(5.13):

$$y = aA_S \sin [\omega(t - T_D)] = aA_S \sin (\omega t - \omega T_D) . \tag{5.16}$$

The output amplitude is $b = aA_S$.

The output lags behind the input by the phase angle:

$$\varphi = \measuredangle G(i\omega) = -\omega T_D . \tag{5.17}$$

The gain or magnitude of vector $G(i\omega)$ is the ratio of the amplitudes of the output and input sine waves:

$$|G(i\omega)| = \frac{b}{a} = \frac{aA_S}{a} = A_S \tag{5.18}$$

Thus, for a dead time process, $\measuredangle G(i\omega)$ depends on the frequency of the input, while $|G(i\omega)|$ does not.

We define the gain of the process as a product of the steady state and the dynamic gains:

$$|G(i\omega)| = A_S |g(i\omega)| \tag{5.19}$$

This is the definition of the dynamic gain, which will be found useful later. For the dead time process,

$$|g(i\omega)| = \frac{|G(i\omega)|}{A_S} = \frac{A_S}{A_S} = 1$$

We define the normalized frequency function vector of a process as

$$\underline{g}(i\omega) = \frac{\underline{G}(i\omega)}{A} \tag{5.20}$$

Of course, $\measuredangle g(i\omega) = \measuredangle G(i\omega)$ for all processes. $A_S$ is a scalar quantity.

The normalized frequency function vector of a dead time process is shown in *Fig. 5.20* in the form of Nyquist and Bode plots, using coordinates that are independent of the parameters of the individual system.

*Proportional control of dead time*

A proportional controller (or *P*-controller) is a simple device with an output that is proportional to the error at every time instant

$$m = A_C e \tag{5.21}$$

where $e$ is the error, $m$ is the manipulation and $A_C$ is the proportional gain of the controller.

The *P*-controller has no phase lag and its gain is constant for all input frequencies. Thus, for the proportional controller

$$|G_C(i\omega)| = A_C$$

$$|g_C(i\omega)| = 1$$

$$\measuredangle g_C(i\omega) = 0° .$$

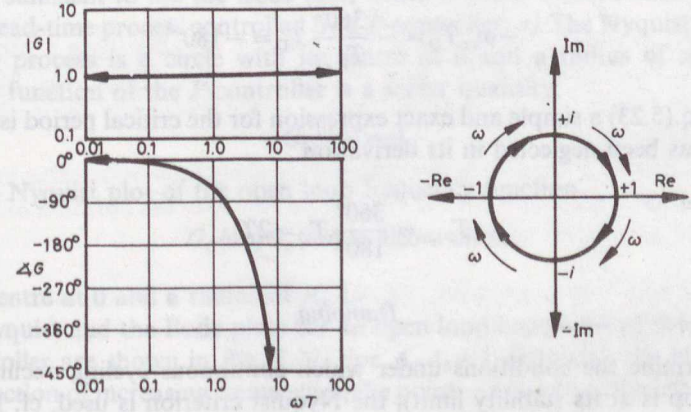



Fig. 5.20. Nyquist and Bode plots of dead time processes

*Figure 5.21* illustrates a simplified block diagram of a dead time process controlled by a *P*-controller. The phase lags are as follows:

process $$\varphi_S = -\omega T_D$$

$$\left(\text{radians if } \omega \frac{\text{rad}}{\text{s}}, \text{ degrees if } \omega \frac{\text{deg}}{\text{s}}\right)$$

controller $$\varphi_C = 0$$

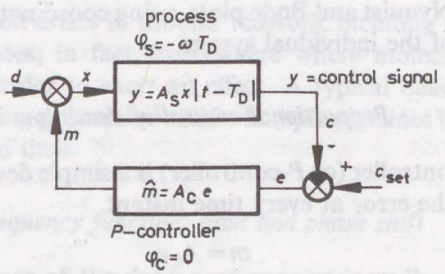



Fig. 5.21. Proportional control of dead time

*Critical frequency*

For a sine wave having the critical frequency, the phase lags of the process and the controller sum to $-180°$, which causes the sine wave to return to its starting point with $-360°$, i.e. zero phase shift:

$$\varphi_S + \varphi_C = -\omega_{cr} T_D + 0 = -180° \tag{5.22}$$

$$-\omega_{cr} T_D = -\frac{360°}{T_{P,cr}} T_D = -180° \tag{5.23}$$

From Eq.(5.23) a simple and exact expression for the critical period is obtained. Nothing has been neglected in its derivation.

$$T_{P,cr} = \frac{360°}{180°} T_D = 2T_D \tag{5.24}$$

*Damping*

To determine the conditions under which continuous cycling occurs (i.e. the control loop is at its stability limit), the Nyquist criterion is used, cf. Eqs (5.10) —(5.12). Thus, the stability limit is attained if the gain of the proportional control-

ler is set so that it satisfies the following equation:

$$|G_C(i\omega_{cr})|\,|G_S(i\omega_{cr})| = A_C A_S = 1 \tag{5.25}$$

This means that, if a signal composed of sinusoidal waves of a broad frequency spectrum (e.g. a pulse signal) enters the control loop in *Fig. 5.21* at any point, i.e. either as a disturbance, or as a set point change, then after a time necessary for the sine waves with other frequencies to die out, all signals in the loop, i.e. $c$, $e$ and $m$ will cycle with constant amplitude and a period equal to $2T_D$.

If the gain of the proportional controller is set so that

$$A_C A_S > 1$$

then the amplitudes of the successive cycles (with period $2T_D$) will be expanding, and the system is unstable.

As a rule of thumb, controllers are tuned for quarter-amplitude damping in order to obtain stable control which restores the steady-state within three full cycles *(Fig. 5.17)*.

For this damping a *P*-controller controlling a dead time process is tuned to produce a gain product equal to 0.5:

$$A_C A_S = 0.5\,.$$

With this tuning, the period of the signals in the control loop is essentially the critical one, $T_{P,cr} = 2T_D$ (it is, in fact, slightly longer), and the steady state is attained after a time of $3T_{P,cr} = 6T_D$.

*Open-loop frequency function*

To check the stability of the closed control loop by the Nyquist criterion, it is advisable to draw the Nyquist plot of the open-loop frequency function. In many cases it is sufficient to use the Bode plot, which is easier to construct.

For a dead-time process controlled by a *P*-controller: *a)* The Nyquist plot of the dead-time process is a circle with its centre at 0 and a radius of $A_S$. *b)* The frequency function of the *P*-controller is a scalar quantity:

$$G_C(i\omega) = A_C$$

Thus, the Nyquist plot of the open loop frequency function

$$G_C(i\omega)G_S(i\omega) \text{ is also a circle,}$$

with its centre at 0 and a radius of $A_S A_C$.

The Nyquist and the Bode plots for an open loop composed of dead time and a *P*-controller are shown in *Fig. 5.22*. For $A_C A_S < 1$, following the Nyquist plot in the direction of increasing frequencies, the point $-1$ is seen at the left, indicating that the loop is stable.

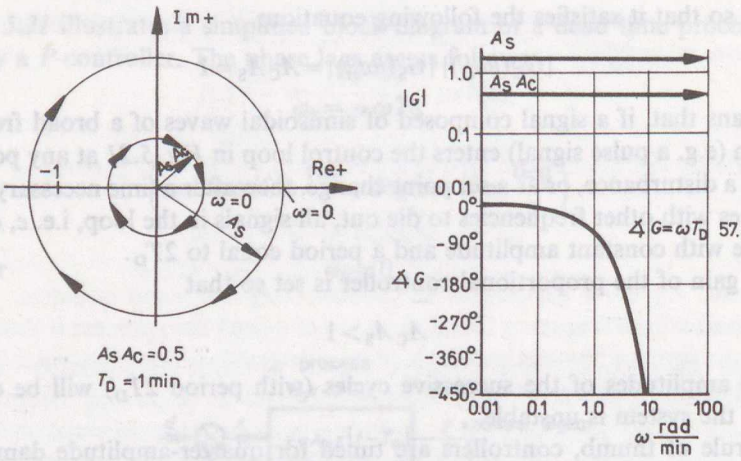



Fig. 5.22. Open loop frequency function for *P*-control of dead time process

### *Steady-state offset*

A control system is expected to keep the controlled variable at its set point, i.e. the error signal is kept at a value of zero (cf. *Fig. 5.21*). A load disturbance or a set point change will cause a transient error, but the controller is expected to eliminate it by appropriate manipulation: in the case of a load disturbance, it must return the control signal to its previous steady-state value; in the case of a set point change it must transduce the control signal to its new command value.

The *P*-controller has the deficiency of being unable to accomplish the above task: it generally works with a steady-state offset. This follows from its control law represented by Eq.(5.21):

$$\hat{m} = A_C e \tag{5.26}$$

where $\hat{m}$ means the deviation of the manipulated variable from a pre-set value $m_0$ for zero error and normal operative conditions: $c_{\text{set},0}$ and $d_0$. In order to balance a deviation from $c_{\text{set},0}$ or $d_0$ a value $m$ different from $m_0$ is needed, but this is produced only if $e \neq 0$ in the steady state.

The steady-state offset can be calculated from the steady-state gains in the loop. This calculation is valid only for stable systems. (Refer to the notation of *Fig. 5.20*.) The steady-state output of the process is

$$y = \hat{c} = A_S x = A_S \hat{m} + A_S \hat{d} \tag{5.27}$$

where $\hat{c} = c - c_0 = c - c_{\text{set},0}$

$\hat{m} = m - m_0$

$\hat{d} = d - d_0$

Combine Eq.(5.26) and $e=\hat{c}_{set}-\hat{c}$ with Eq.(5.27):

$$\hat{c}=A_S A_C(\hat{c}_{set}-\hat{c})+A_S\hat{d} \tag{5.28}$$

Rearrange to give

$$\hat{c}=\frac{A_S A_C}{1+A_S A_C}\hat{c}_{set}+\frac{A_S}{1+A_S A_C}\hat{d} \tag{5.29}$$

**Example 5.2.1**

A *P*-controller controlling a dead time process with $A_S=0.3$ is set for quarter amplitude damping, i.e. $A_S A_C=0.5$. What is the steady-state change of the controlled variable after a unity change

(*a*) in the set point, $\hat{c}_{set}=1$

(*b*) in the load variable, $\hat{d}=1$?

(*a*) For quarter amplitude damping

$$\frac{A_S A_C}{1+A_S A_C}=\frac{0.5}{1+0.5} \quad \text{thus} \quad \hat{c}=\frac{0.5}{1.5}1=0.333;$$

(*b*)

$$\hat{c}=\frac{A_S\hat{d}}{1+0.5}=\frac{A_S\hat{d}}{1.5}=\frac{0.3}{1.5}=0.2$$

Without control, a unit change in $d$ would cause a change in $c$, the controlled variable of

$$A_S\hat{d}=0.3\cdot 1=0.3$$

Thus, proportional control leaves $\frac{0.2}{0.3}=\frac{1}{1+A_S A_C}=0.66$, i.e., 66% of the offset without control!

To eliminate steady-state error, $A_S A_C=\infty$ would be needed, but this is prohibited by the stability requirement.

### Integral control of dead time

*Integral controller, definition*

The *P*-controller is unable to eliminate offset, so another control mode is needed. The (integral) *I*-controller is a device producing a steady state only in the case of zero offset:

$$\frac{dm}{dt}=\frac{1}{I}e \tag{5.30}$$

$\left(\text{thus, } \frac{dm}{dt} = 0 \text{ if } e = 0\right)$, or, integrating Eq.(5.30):

$$\hat{m} = \frac{1}{I} \int_0^t e \, dt \tag{5.31}$$

Here $I$ is the tuning parameter of the $I$-controller. The output of the $I$-controller is proportional to the error integral. The response to a step input is shown in *Fig. 5.23*. It is shown that $I$ is the time necessary for the output to repeat the input: it is named *integral* or *reset* or *repeat* time.

Thus, the integral control mode eliminates offset in the steady state. There is, however, certainly a penalty for that.

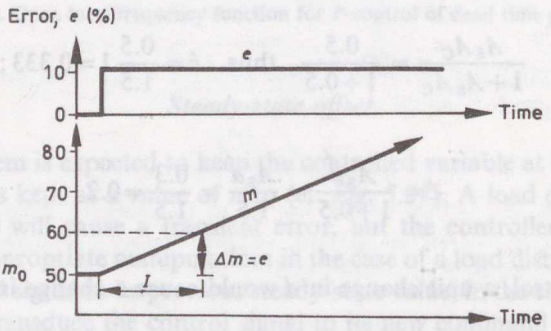



Fig. 5.23. Integral time of the $I$-controller

*I-controller, frequency function*

The gain and phase angle of the $I$-controller will be calculated before examining its effect on control loop behaviour. The frequency function is easily calculated using Eq.(5.31) (refer to *Fig. 5.11*). The input to the controller is the sine wave:

$$e = a \sin \omega t \tag{5.32}$$

Substitute this into Eq.(5.31)

$$\hat{m} = \frac{1}{I} \int a \sin \omega t \, dt = \frac{a}{I} (-\cos \omega t) \tag{5.33}$$

Since

$$-\cos \varphi = -\sin (90° - \varphi) = \sin (\varphi - 90°) \tag{5.34}$$

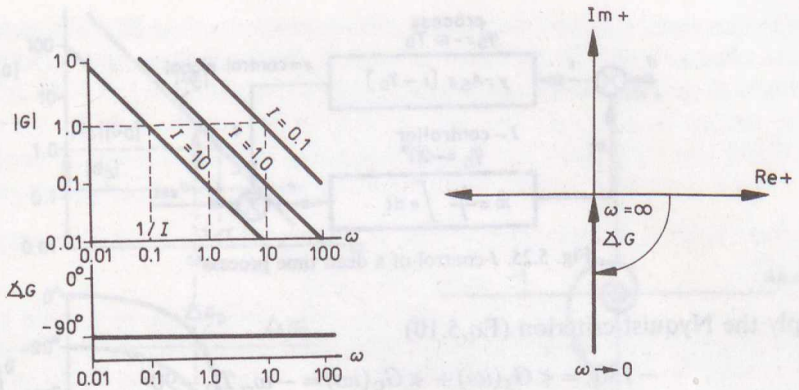



Fig. 5.24. Nyquist and Bode plot of the *I*-element

we obtain

$$\hat{m} = \frac{a}{I\omega} \sin(\omega t - 90°) \tag{5.35}$$

whence the magnitude of the frequency function vector, or gain, is

$$|G(i\omega)| = |g(i\omega)| = \left(\frac{a}{I\omega}\right) \Big/ (a) = \frac{1}{I\omega} \tag{5.36}$$

It is inversely proportional to the frequency of the input: when $\omega \to 0$, $|G| \to \infty$; when $\omega \to \infty$, $|G| \to 0$.

The phase angle of the frequency function vector is clearly

$$\measuredangle G(i\omega) = \measuredangle g(i\omega) = -90° \tag{5.37}$$

independent of the frequency of the input. Thus, all the frequency function vectors of the *I*-controller lie along the negative imaginary axis of the complex plane. The Nyquist and the Bode plots of the frequency function are shown in *Fig. 5.24.*

Recall that for the dead time process it is the magnitude of the frequency function which is independent; the phase lag is proportional to frequency.

*The control loop with dead time process and I-controller*

The simplified block diagram of the control loop is shown in *Fig. 5.25.*

In order to determine the *critical frequency* of the loop, the phase shifts are needed. Process [Eq.(5.17) with $\omega$ in degree/s]:

$$\measuredangle G_S(i\omega) = -\omega T_D$$

*I*-controller [Eq.(5.37)]:

$$\measuredangle G_C(i\omega) = -90°$$

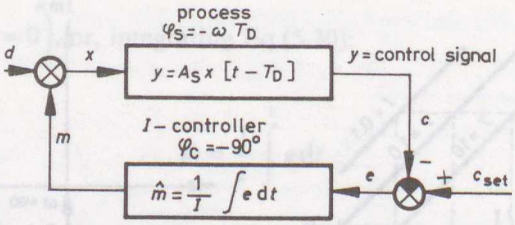



Fig. 5.25. *I*-control of a dead time process

Apply the Nyquist criterion (Eq.5.10)

$$-180° = \measuredangle G_S(i\omega) + \measuredangle G_C(i\omega) = -\omega_{cr} T_D - 90° \tag{5.38}$$

Whence:

$$\omega_{cr} = \frac{90°}{T_D}\,\frac{\text{degree}}{\text{s}} = \frac{\pi}{2T_D}\,\frac{\text{radian}}{\text{s}} \tag{5.39}$$

The critical period:

$$T_{P,cr} = \frac{360°}{\omega_{cr}} = 4T_D \tag{5.40}$$

The penalty for eliminating offset is found in Eq.(5.40): the critical period of the control loop with an *I*-controller is double that found with a *P*-controller, and so is the time of a transient. This result is quite general for any process and control loop: an *I*-controller slows control down.

Damping in the control loop with an *I*-controller is obtained by the proper setting of the sole tuning parameter *I*.

At the stability limit [cf. Eqs (5.11), (5.18) and (5.36)]

$$1 = |G_S(i\omega_{cr})|\,|G_C(i\omega_{cr})| = A_S \frac{1}{I\omega_{cr}} = \frac{2T_D A_S}{\pi I} \tag{5.41}$$

whence

$$I_{cr} = \frac{2}{\pi} T_D A_S \tag{5.42}$$

If quarter-amplitude damping is desired, then

$$0.5 = |G_S(i\omega_{cr})|\,|G_C(i\omega_{cr})|$$

and

$$I_{(0.5)} = \frac{4}{\pi} T_D A_S \tag{5.43}$$

In order to damp oscillations, the reset time has to be augmented. The minimum value of *I* is the critical one given by Eq.(5.42).

The choice of the reset time does not influence the critical frequency, which depends only on the dead time of the process.

The open loop frequency function for the integral control of dead time is plotted in the Bode diagram *(Fig. 5.26)* using the following equations:

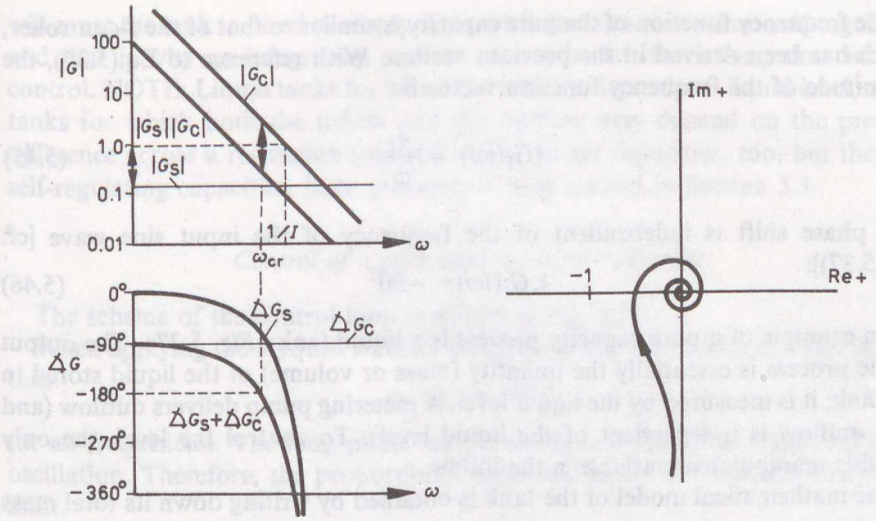



Fig. 5.26. Open loop Nyquist and Bode plots for *I*-control of a dead time process

gain: $$|G_S|\,|G_C| = A_S \frac{1}{I\omega}$$

phase shift: $$\measuredangle G_S + \measuredangle G_C = -T_D\omega \times 57.3° - 90°$$

$\omega$ in radians/s

Both gain and phase shift are monotonic functions of frequency, so stability can also be checked from the Bode plot. The Nyquist plot is constructed using the gain and phase values from the Bode plot. The point $-1$ is seen at the left, and thus the control loop is stable.

### 5.2.3 Control of capacity processes

By contrast with dead-time processes, capacity processes are easy to control. The input-output mathematical model of a pure capacity process is similar to that of the *I*-controller:

$$y = A_I \int x\,dt \tag{5.44}$$

where $x$ is input, $y$ is output, both being deviations from the steady-state values,

$$A_I \text{ is gain, its dimension being } \frac{[y]}{[x] \times \text{time}}$$

Pure capacity is thus an integrating element, as is the *I*-controller, for which $A_I = \frac{1}{I}$ [cf. Eq.(5.31)].

The frequency function of the pure capacity is similar to that of the *I*-controller, which has been derived in the previous section. With reference to Eq.(5.36), the magnitude of the frequency function vector is

$$|G_I(i\omega)| = \frac{A_I}{\omega} \tag{5.45}$$

The phase shift is independent of the frequency of the input sine wave [cf. Eq.(5.37)]:

$$\measuredangle G_I(i\omega) = -90° \tag{5.46}$$

An example of a pure capacity process is a liquid tank *(Fig. 5.27)*. The output of the process is essentially the quantity (mass or volume) of the liquid stored in the tank; it is measured by the liquid level. A metering pump delivers outflow (and thus outflow is independent of the liquid level). To control the level, the only possible manipulative variable is the inflow.

The mathematical model of the tank is obtained by writing down its total mass balance:

$$A\frac{dH}{dt} = W_{in} - W_{out} \tag{5.47}$$

whence the change in level is:

$$\hat{H} = \frac{1}{A}\int_0^t (W_{in} - W_{out})\,dt \tag{5.48}$$

Equation (5.48) is identical to Eq.(5.44), where $y = \hat{H}$, $A_I = \frac{1}{A}$, $x = (W_{in} - W_{out})$. Thus this tank is an integrating element, the frequency function of which is shown in *Fig. 5.24*, with a step response similar to that shown in *Fig. 5.23*.

The step response of the integrating process indicates that any small change in the flows, which implies a non-steady-state:

$$W_{in} - W_{out} \neq 0 \tag{5.49}$$

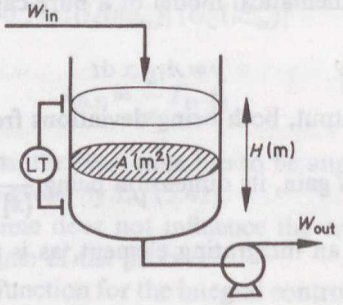



Fig. 5.27. Pure capacity element: liquid tank with fixed outflow

will cause the tank to flood or empty. This attribute is known as non-self-regulation and that is why an integrating process cannot be left for a long time without control. NOTE: Liquid tanks for which the outflow depends on liquid level, or gas tanks for which both the inflow and the outflow may depend on the pressure difference across a restriction (pipeline or valve) are capacities, too, but they are self-regulating capacities. Such processes will be treated in Section 5.3.

*Control of a pure capacity by P-controller*

The scheme of the control loop is shown in *Fig. 5.28.*

When applying the Nyquist stability criterion to this control loop, it can be seen that

$$\varphi_S + \varphi_C = -90°$$

for all frequencies. The loop phase lag never reaches 180°, thus there will be no oscillation. Therefore, the proportional band can be set without any theoretical limit.

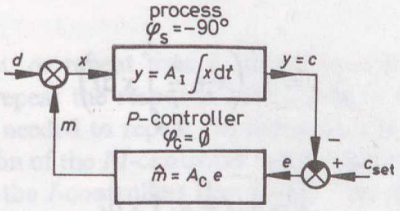



Fig. 5.28. *I*-element controlled by *P*-controller

In consequence, pure capacity processes can be controlled by a simple *P*-controller set to a high $A_C$ proportional gain (narrow proportional band), with no oscillation or steady-state offset.

An *I*-controller cannot be used to control an *I*-process because both the *I*-controller and the pure capacity process exhibit a constant phase shift of −90° for all frequencies *(Fig. 5.29)*:

$$\varphi_S = \varphi_C = -90°$$

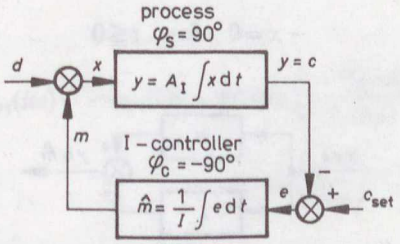



Fig. 5.29. *I*-element controlled by *I*-controller

thus the loop phase shift sums to $-180°$

$$\varphi_S + \varphi_C = -180°$$

independently of the frequency and of controller tuning.

### 5.2.4 The *PI*-controller

Due to the fact that the *I*-controller may produce unstable control if applied to a process with a free *I*-element, it is almost always combined with a proportional controller.

The ideal *PI*-controller consists of a *P*- and an *I*-controller in parallel *(Fig. 5.30)*. This combines the best features of the *P*- and *I*-controllers: the proportional offset is eliminated with little loss of response speed (compared to an *I*-controller) and stable operation can be achieved with almost any process. This controller combination is the one that is most frequently used. It is described by the following equation:

$$y = A_P \left( x + \frac{1}{I} \int_0^t x \, dt \right) \tag{5.50}$$

or

$$y = A_P x + A_I \int_0^t x \, dt \tag{5.51}$$

where $y$ is output (for a controller, the manipulation),
$x$ is input (for a controller, the error),
$A_P$ is the static gain of the *P*-controller,
$I$ is the integral time of *PI*-controller,
$A_I = \frac{A_P}{I}$ is the static gain of the *I*-controller, and $t$ is time.

The step response of the *PI*-controller is obtained from Eq. (5.50) if the input is constant:

$$x = 0 \quad \text{for} \quad t \leqq 0$$

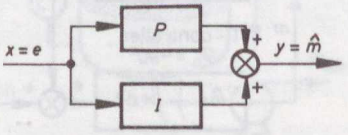



Fig. 5.30. Scheme of *PI*-controller

$$x = a \quad \text{for} \quad t > 0$$

$$y = A_P\left(a + \frac{at}{I}\right). \tag{5.52}$$

For $t = I$, $y = A_P(2a)$ (see *Fig. 5.31*).

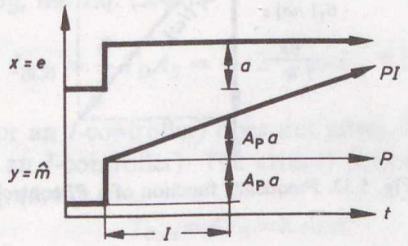



Fig. 5.31. Step response of a *PI*-controller

The integral, or reset, or repeat time $I_I$ of a *PI*-controller is the time necessary for the *I* part to repeat the response of the *P*-part. (The repeat time of an *I*-controller is the time needed to repeat the input; cf. *Fig. 5.23*.)

The frequency function of the *PI*-controller can be obtained as the vectorial sum of those of the *P*- and the *I*-controllers [Eq. (5.3)]. This is best represented on a polar diagram *(Fig. 5.32.)*.

From *Fig. 5.32* it is evident that the magnitude of the vector $G_{PI}(i\omega)$ is:

$$|G_{PI}(i\omega)| = \sqrt{|G_P(i\omega)|^2 + |G_I(i\omega)|^2}$$

$$= \sqrt{A_P^2 + \left(\frac{A_P}{I\omega}\right)^2}$$

$$= A_P\sqrt{1 + \left(\frac{1}{I}\right)^2} \tag{5.53}$$

The phase shift

$$\measuredangle G_{PI}(i\omega) = -\varphi_{PI}(i\omega) = \arctan\left(\frac{-G_I(i\omega)}{T_P(i\omega)}\right)$$

$$= \arctan\left(\frac{-1}{I\omega}\right) \tag{5.54}$$

Both the magnitude and the phase shift of a *PI*-controller depend on frequency.

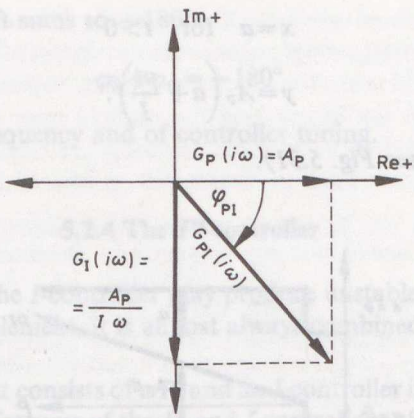



Fig. 5.32. Frequency function of a *PI*-controller

For

$$I\omega \to 0 \quad |G_{PI}| = \frac{A_P}{I} \to \infty \quad -\varphi_{PI} \to -90°$$

$$I\omega \to \infty \quad |G_{PI}| \to A_P \quad -\varphi_{PI} \to 0°$$

Thus for low frequencies (near the steady state) the controller behaves as an *I*-controller, eliminating steady-state offset; and for high frequencies it behaves as a *P*-controller with very small phase lag — thus if properly set it will only slightly increase the phase lag of the control loop, the critical period and the settling time.

The phase lag of the *PI*-controller depends only on the parameter *I*. *I* should be set so that the phase lag of the controller is smaller than 10° at the critical frequency of the process.

**Example 5.2.2**

Consider a dead-time process with the parameters $A_S = 2$, $T_D = 2$ min. The task is to control it (*a*) with a *P*-controller; (*b*) with an *I*-controller; (*c*) with a *PI*-controller. The controller is to be tuned to quarter-amplitude damping. Offset and settling time are to be estimated.

**Solution**

(*a*) Tune the *P*-controller to quarter-amplitude damping:

$$A_C A_S = A_C \cdot 2 = 0.5$$

$$A_C = 0.25$$

Offset is calculated using Eq. (5.29):

$$\frac{\hat{c}}{\hat{c}_{set}} = \frac{A_S A_C}{1 + A_S A_C} = \frac{0.5}{1.5} = 0.333$$

$$\frac{\hat{c}}{d} = \frac{A_S}{1 + A_S A_C} = \frac{2}{1.5} = 1.333$$

The settling time is $3T_{P,cr}$. Use Eq. (5.24):

$$T_{set} = 3T_{P,cr} = 3 \cdot 2 \cdot T_D = 3 \cdot 2 \cdot 2 \text{ min} = 12 \text{ min}$$

(*b*) With an *I*-controller there is no offset. To tune the controller for quarter-amplitude damping, use Eq. (5.43):

$$I_{(0.5)} = \frac{4}{\pi} T_D A_S = \frac{4}{\pi} \cdot 2 \cdot 2 \text{ min} = 5.1 \text{ min}$$

The value of $I$ (for an *I*-controller) does not affect the phase shift (which is always $-90°$ for an *I*-controller). The critical period of the control loop is obtained from Eq. (5.40):

$$T_{P,cr} = 4T_D = 8 \text{ min},$$

$$T_{set} = 3T_{P,cr} = 24 \text{ min}$$

(*c*) With a *PI*-controller, there is no offset. The critical frequency, where the dead-time process alone exhibits $-180°$ phase shift

$$\omega_{cr} T_D = \pi \rightarrow \omega_{cr} = \frac{\pi}{T_D} = \frac{\text{rad}}{\text{min}}$$

This is the frequency at which the *PI*-controller should produce a $-10°$ phase shift. Use Eq. (5.54):

$$-\varphi_{PI} = \arctan\left(\frac{-1}{I\omega}\right) = -10°,$$

$$I = 3.61 \text{ min}$$

The critical frequency of the control loop is where

$$\varphi_S + \varphi_C = -T_D\omega + \arctan\left(\frac{-1}{I\omega}\right) = -\pi$$

By trial and error, this is

$$\omega_{cr} = 1.40 \frac{\text{rad}}{\text{min}},$$

whence

$$T_{P,cr} = \frac{2\pi}{\omega_{cr}} = 4{,}49 \text{ min},$$

and the settling time is $3T_{P,cr} = 13.5$ min, which is slightly more than the 12-min value found for *P*-control. For quarter-amplitude damping $A_P$ remains to be calculated.

$$|G_S||G_C| = 0.5,$$

where [Eq. (5.53)]

$$|G_C| = A_P \sqrt{1 + \left(\frac{1}{I\omega}\right)^2}$$

$$= A_P \sqrt{1 + \left(\frac{1}{3.61 \cdot 1.40}\right)^2} = A_P \cdot 1.019$$

whence

$$A_P = \frac{0.5}{2 \cdot 1.019} = 0.245,$$

i. e. slightly smaller than 0.25, the value found for the *P*-controller.

### 5.2.5 *PD*- and *PID*-controllers

Dead-time processes are well controlled with *PI*-controllers as shown in the previous sections. In the case of multicapacity processes the critical period may be very slow and it would be advantageous to speed up control. In such cases control involves derivative action.

The input-output equation of the ideal derivative element is:

$$y = D \frac{dx}{dt} \tag{5.55}$$

where $y$ is output, $x$ is input, $D$ is derivative time constant, and $t$ is time.

Of course, Eq. (5.55) cannot be realized because the derivative of the input at time instant $t$ cannot be known; it can be estimated only from previous measurements. So a real derivative element always contains a small time delay.

Moreover, for an ideal step input which would occur in zero time and for which $\frac{dx}{dt} = \infty$, the ideal derivative element should produce infinite output in zero time delay, which is also impossible.

The frequency function of the *D*-element is obtained by solving Eq. (5.55) for $x = a \sin \omega t$:

$$y = D\omega \cos\omega t = aD\omega \sin(\omega t + 90°)$$

whence $|G_D(i\omega)| = D\omega$ dependent on $\omega$

$\measuredangle G_D(i\omega) = +90°$ independent from $\omega$

The frequency function of the *D*-element in the complex plane in a series of vectors lying along the positive imaginary axis; its Nyquist and Bode plots are shown in *Fig. 5.33*.

The ideal *D*-element would thus reduce the loop lag by 90°; a real *D*-element can produce a phase shift of +60°.

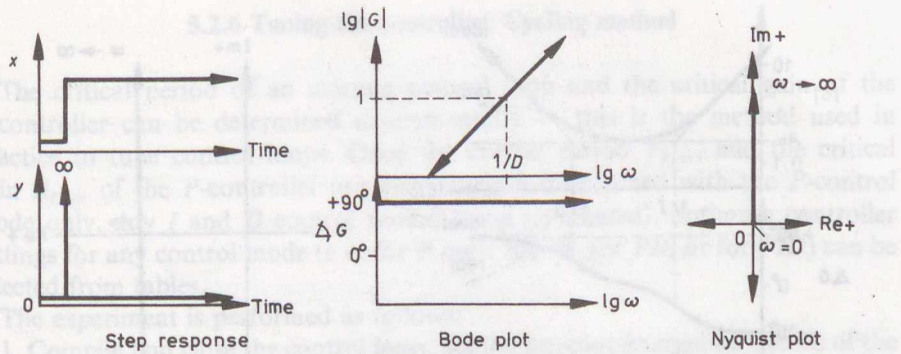



Fig. 5.33. Plots of a *D*-element

The derivative element cannot be applied in a controller as the only control mode, since it responds only to a change in its input, and a constant input has no effect at all on its output.

The *D*-element is therefore used only in combination with a *P*-controller (*PD*-controller) or with a *PI*-controller (*PID*-controller). Since the *PD*-controller does not contain the *I* operating mode, it cannot eliminate steady-state offset completely and is used only for special applications.

The *PID*-controller is the most general controller. The ideal *PID*-controller consists of a *P*, an *I* and a *D* element in parallel *(Fig. 5.34)*, its input-output equation is:

$$y = A_P x + A_I \int_0^t x\,dt + A_D \frac{dx}{dt} = A_P \left( x + \frac{1}{I} \int_0^t x\,dt + D \frac{dx}{dt} \right) \tag{5.56}$$

where $y = m - m_0$ is the output or change in the manipulated variable,

$m_0$ is the output of the controller in the steady state, when the error, its integral and its derivative are zero,

$x = e$ is the input of the controller: the error signal, $A_P$, $I$, $D$ are tuning parameters of the controller, and

$t$ is time.

The Nyquist plot of the *PID*-controller is obtained by adding the Nyquist plots

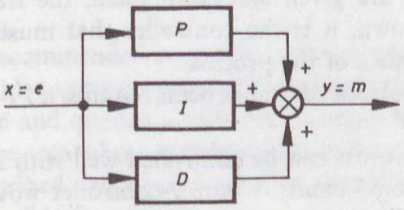



Fig. 5.34. Scheme of a *PID*-controller

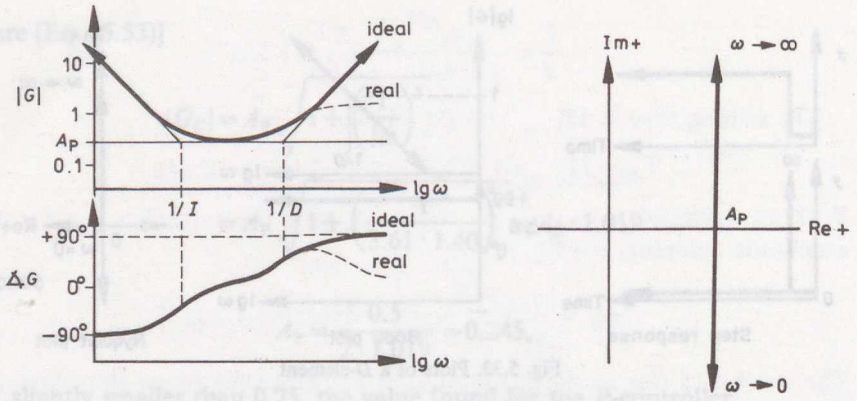



Fig. 5.35. Bode and Nyquist plots of a *PID*-controller

of the *P*, *I* and *D* elements (polar plot). Both the Nyquist and the Bode diagrams are shown in *Fig. 5.35*. It can be seen that

(*i*) for low frequencies, $\omega < \frac{1}{I}$, the *PID*-controller behaves like an *I*-controller, and may cause instability in a control loop by increasing both phase lag and dynamic gain;

(*ii*) for high frequencies $\omega > \frac{1}{D}$ the *PID*-controller behaves like a *D*-element and may cause instability in a control loop by increasing its dynamic gain (although it decreases phase lag);

(*iii*) thus parameters *I* and *D* should be tuned so that $\frac{1}{I} < \omega_{cr} < \frac{1}{D}$, where $\omega_{cr}$ is the critical frequency of the process along (or with a *P*-controller).

*Conclusion.* Negative feedback control loops are stable if their frequency function satisfies the Nyquist criterion:

$$\measuredangle G_S(i\omega_{cr}) + \measuredangle G_C(i\omega_{cr}) = -180°$$

$$|G_S(i\omega_{cr})|\,|G_C(i\omega_{cr})| < 1$$

Since the parameters are given and, with them, the frequency function of the process, $G_S(i\omega)$ is known, it is the controller that must be tuned in consideration of the characteristics of the process.

The proper control of a dead time process requires a *PI*-controller; the minimum settling time is $6T_D$.

A single-capacity process can be controlled well with a simple *P*-controller set to a narrow proportional band. A pure *I*-controller would cause instability and thus must not be applied.

### 5.2.6 Tuning the controller. Cycling method

The critical period of an existing control loop and the critical gain of the *P*-controller can be determined experimentally — this is the method used in practice to tune control loops. Once the critical period $T_{P,cr}$, and the critical gain $A_{C,cr}$ of the *P*-controller are known (both determined with the *P*-control mode only, any *I* and *D* control modes being eliminated), optimum controller settings for any control mode (i. e. for *P* only, for *PI*, for *PD*, or for *PID*) can be selected from tables.

The experiment is performed as follows:

1. Compile and close the control loop. Set the setpoint (command signal) of the controller.

2. Eliminate *I* and *D* control modes, possibly by setting *I* to its maximum and *D* to its minimum possible value.

3. Set $A_C$ to its minimum value. (Most analogue controllers are scaled as $P=\frac{1}{A_C}$.)

4. Apply a small driving signal to one of the inputs of the control loop (this may be one of the load inputs as well; but it is most conveniently applied to the set point). Be sure to apply a driving signal containing a broad frequency spectrum (it must also contain the unknown critical frequency, e. g. a pulse or a stop).

5. Observe the output of the control loop (in most cases the controlled variable is the output, but the manipulated variable may also be used). If no oscillation occurs, increase the controller gain $A_C$ and repeat step 4 until uniform oscillation results. If the amplitude of the oscillation increases, the controller gain $A_C$ must be decreased.

6. The controller gain $A_C$ causing constant amplitude oscillation of the control loop output signal is the critical one: $A_{C,cr}$, and the period of one oscillation is the critical period $I_{P,cr}$.

7. Use *Table 5.1* to calculate the optimum controller settings. Here $K_{cr}$ is the critical and $K$ the recommended loop gain. By definition

$$K=A_C A_S$$

and, since the process gain $A_S$ is constant,

$$\frac{K}{K_{cr}}=\frac{A_C}{A_{C,cr}}$$

*Table 5.2* gives the recommended controller settings given by Ziegler and Nichols, which are applicable to processes consisting of dead time and single capacity. It gives stable operation and quarter amplitude damping for most control loops. Controller tuning will be treated more thoroughly in Section 5.3.8.

The method just described can also be used for simulated control loops which may be tested using the *TACS* program.

Table 5.1. Optimum controller settings

Process parameters by cycling method: $A_{C,cr}$, $T_{P,cr}$
Zeigler—Nichols settings

| Controller | $A_C$ | $I$ | $D$ |
|---|---|---|---|
| $P$ | $0.5\ A_{C,cr}$ | | |
| $PI$ | $0.45\ A_{C,cr}$ | $T_{P,cr}/1.2$ | |
| $PID$ | $0.6\ A_{C,cr}$ | $T_{P,cr}/2$ | $T_{P,cr}/8$ |

Table 5.2. Process parameters from process reaction curve: $A_S$, $T_D$, $T$

| Controller | Ziegler—Nichols | Cohen—Coon |
|---|---|---|
| $P$ | $A_S A_C = K = \frac{T}{T_D}$ | $K = \frac{T}{T_D} + \frac{1}{3}$ |
| $PI$ | $k = 0.9\frac{T}{T_D}$<br>$I = \frac{T_D}{0.3}$ | $K = 0.9\left(\frac{T}{T_D} + \frac{1}{11}\right)$<br>$\frac{I}{T} = \frac{3.33(T_D/T)\,(1+11(T_D/T))}{1+2.2(T_D/T)}$ |
| $PID$ | $K = 1.2(T/T_D)$<br>$I = 2T_D$<br>$D = 0.5T_D$ | $K = 1.35(T/T_D)+0.27$<br>$\frac{I}{T} = \frac{2.5(T_D/T)\,(1+5(T_D/T))}{1+0.6(T_D/T)}$<br>$\frac{D}{T} = \frac{0.37(T_D/T)}{1+0.2(T_D/T)}$ |

## 5.3 Linear control theory

Before tackling more complex control problems it is necessary to survey the mathematical methods of linear control theory. The fundamentals of classical linear control theory are very important in practice, notwithstanding the fact that most chemical engineering systems are nonlinear. The mathematical models of most nonlinear systems may, however, be linearized around their operating points, in order to use the powerful linear techniques.

### 5.3.1 Classification of processes

There are several ways of classifying processes and their dynamics.

1. Mathematical model
   (a) Linear: if all equations and all function in the equations are linear.
   (b) Nonlinear: if the statement above is not true. In this case the principle of superposition does not apply. That is, the response of the system to several independent inputs cannot be determined independently and then added.
2. Independent variables
   (a) Lumped-parameter system: described by an ordinary differential equation; the dependent variable is a function of time only; for example, a perfectly mixed tank reactor.
   (b) Distributed-parameter system: described by a partial differential equation; the dependent variable is a function of time and of spatial variables; an example is the tubular reactor.
3. Order:

An $n$th order linear, lumped-parameter, single input single output process can be described by an $n$th order linear, ordinary differential equation of the form:

$$a_n \frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \ldots + a_1 \frac{dy}{dt} + a_0 a = \\ = b_m \frac{d^m x}{dt^m} + b_{m-1} \frac{d^{m-1} x}{dt^{m-1}} + \ldots + b_1 \frac{dx}{dt} + b_0 x \qquad (5.57)$$

where $x$ is input,
$y$ is output,
$t$ is time, and
$a$, $b$ are constants.

In real systems $n$ is greater than $m$. Inequality $m > n$ would mean that the response for a step input must be infinite in zero time. This cannot be realized in a finite system (cf. the $D$-element, Section 5.2.5).

Multiple input—multiple output systems are characterized by a system of differential equations.

In control technology it is convenient to write Eq.(5.57) in the following form:

$$T_n^n \frac{d^n y}{dt^n} + T_{n-1}^{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \ldots + T_1 \frac{dy}{dt} + y = \\ = A_S \left[ \tau_m^m \frac{d^m x}{dt^m} + \tau_{m-1}^{m-1} \frac{d^{m-1} x}{dt^{m-1}} + \ldots + \tau_1 \frac{dx}{dt} + x \right] \qquad (5.58)$$

where $T_i = \sqrt[i]{\frac{a_i}{a_0}}$; $\tau_j = \sqrt[j]{\frac{b_j}{b_0}}$ are time constants, and

$$A_S = \frac{b_0}{a_0} = \frac{y(t=\infty)}{x(t=\infty)} \text{ is steady-state gain.}$$

Very important special cases are:

the first-order element:

$$T\frac{dy}{dt} + y = A_S x \tag{5.59}$$

the second-order element:

$$T_2^2 \frac{d^2y}{dt^2} + T_1 \frac{dy}{dt} + y = A_S x \tag{5.60}$$

A dead-time element (defined in Section 5.2.2) can be approximated only in this representation by $n$th order differential equation where $n \to \infty$.

### 5.3.2 Typical input signals

Linear processes can be characterized by their responses to typical input signals (besides being characterized by their differential equations).

Typical input signals are also applied for the experimental determination of the dynamic behaviour of an unknown process (identification).

The typical input signals are the following:

(*i*) step: a change in the input signal value during zero time;

(*ii*) pulse: a step upwards and, after time $\Delta t$, the same step downwards;

(*iii*) impulse: a pulse in time $\Delta t \to 0$;

(*iv*) ramp: a change in the input signal with constant speed; and

(*v*) a sinusoidal signal.

These deterministic signals are shown and defined in *Fig. 5.36*.

The sinusoidal input signal is used to obtain the frequency response that has already been treated in Section 5.2.1 (see *Fig. 5.11*). The sinusoidal input is to be repeated several times with the same frequency and amplitude, until a steady-state response is obtained; the steady state is indicated by at least three constant amplitude and phase sinusoidal outputs while step, impulse and ramp inputs may be applied only once or twice. Moreover, the experiment has to be carried out with different frequencies over a broad frequency range to obtain a meaningful frequency function, thus it is lengthy and expensive. However, in the presence of noise, it is advantageous that a single frequency output is to be measured.

Identification of an unknown process can also be performed using stochastic signals, such as white noise or plant noise.

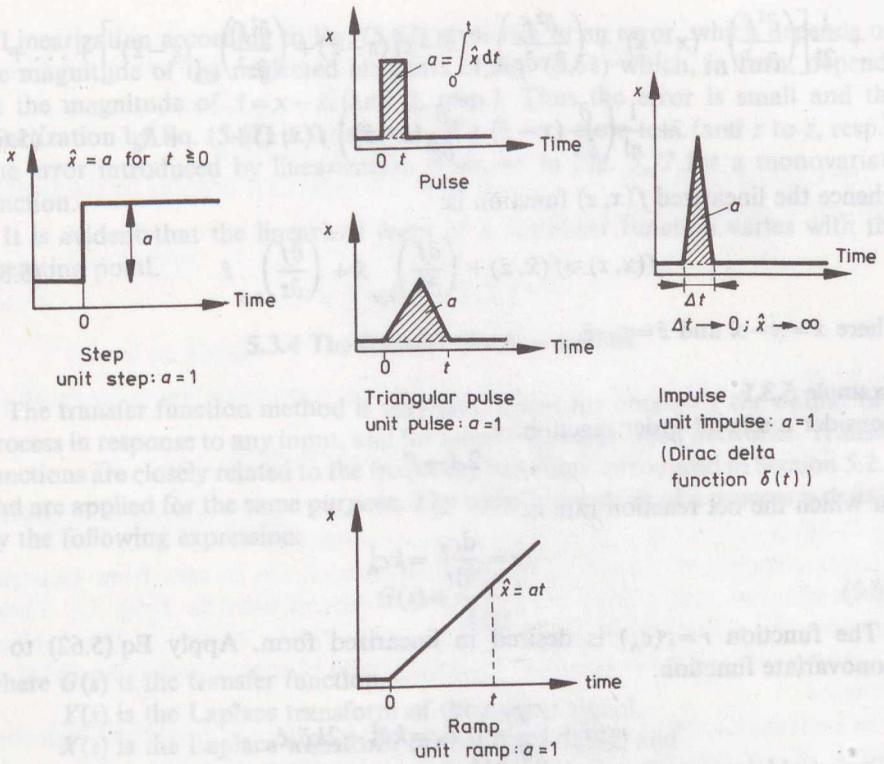



Fig. 5.36. Deterministic input signals

### 5.3.3 Linearization

Mathematical models of nonlinear processes are often linearized in the vicinity of an operating point, so that they may be tackled by linear mathematical techniques, e.g. the frequency function method.

Linearization of a nonlinear function is accomplished by expanding the nonlinear function in Taylor series and then truncating the series, leaving only the first derivative(s). The nonlinear function has to be differentiable at the operating point. Thus the nonlinear function is simply substituted by its tangent.

Consider a nonlinear function, $f(x, z)$, of two variables. $\bar{x}$ and $\bar{z}$ are the values of the variables at the operating point, at the steady state. The Taylor series of this function about point $(\bar{x}, \bar{z})$ is:

$$f(x,z)=f(\bar{x},\bar{z})+\left(\frac{\partial f}{\partial x}\right)_{\bar{x},\bar{z}}(x-\bar{x})+\left(\frac{\partial f}{\partial z}\right)_{\bar{x},\bar{z}}(z-\bar{z})+$$

$$+\frac{1}{2!}\left[\left(\frac{\partial^2 f}{\partial x^2}\right)_{\bar{x},\bar{z}}(x-\bar{x})^2+\left(\frac{\partial^2 f}{\partial x\,\partial z}\right)_{\bar{x},\bar{z}}(x-\bar{x})(z-\bar{z})+\left(\frac{\partial^2 f}{\partial z^2}\right)_{\bar{x},\bar{z}}(z-\bar{z})^2\right]+\ldots+$$

$$+\ldots+\frac{1}{n!}\left(\frac{\partial}{\partial x}(x-\bar{x})+\frac{\partial}{\partial z}(z-\bar{z})\right)^n f(\bar{x},\bar{z})+\ldots+R_n \qquad (5.61)$$

whence the linearized $f(x, z)$ function is:

$$f(x,z)\approx f(\bar{x},\bar{z})+\left(\frac{\partial f}{\partial x}\right)_{\bar{x},\bar{z}}\hat{x}+\left(\frac{\partial f}{\partial z}\right)_{\bar{x},\bar{z}}\hat{z} \qquad (5.62)$$

where $\hat{x}=x-\bar{x}$ and $\hat{z}=z-\bar{z}$

**Example 5.3.1**
Consider a second-order reaction:

$$2A\rightarrow P$$

for which the net reaction rate is:

$$r=\frac{\mathrm{d}c_\mathrm{P}}{\mathrm{d}t}=kc_\mathrm{A}^2$$

The function $r=r(c_\mathrm{A})$ is desired in linearized form. Apply Eq.(5.62) to a monovariate function:

$$r\approx\bar{r}+\left(\frac{\partial r}{\partial c_\mathrm{A}}\right)_{\bar{c}_\mathrm{A}}\hat{c}_\mathrm{A}=k\bar{c}_\mathrm{A}^2+2k\bar{c}_\mathrm{A}\hat{c}_\mathrm{A}$$

**Example 5.3.2**
The input of component $A$ to a continuous reactor is $A=Wc_\mathrm{A}$, where both the volume rate $W$ and its concentration $c_\mathrm{A}$ are variables. Obtain the function $A=A(W, c_\mathrm{A})$ in linearized form. Apply Eq.(5.62):

$$A\approx\bar{A}+\left(\frac{\partial A}{\partial W}\right)_{\bar{W},\bar{c}_\mathrm{A}}\hat{W}+\left(\frac{\partial A}{\partial c_\mathrm{A}}\right)_{\bar{W},\bar{c}_\mathrm{A}}\hat{c}_\mathrm{A}=\bar{W}\bar{c}_\mathrm{A}+\bar{c}_\mathrm{A}\hat{W}+\bar{W}\hat{c}_\mathrm{A}$$

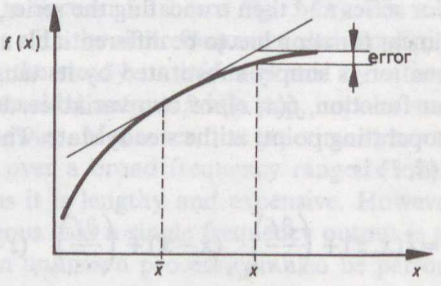



Fig. 5.37. Error of linearization

Linearization according to Eq. (5.62) gives rise to an error, which depends on the magnitude of the neglected elements of Eq. (5.61) which, in turn, depends on the magnitude of $\hat{x}=x-\bar{x}$ (and $\hat{z}$, resp.). Thus the error is small and the linearization by Eq. (5.62) is satisfactory if $x$ is very close to $\bar{x}$ (and $z$ to $\bar{z}$, resp.). The error introduced by linearization is shown in *Fig. 5.37* for a monovariate function.

It is evident that the linearized form of a nonlinear function varies with the operating point.

### 5.3.4 The transfer function method

The transfer function method is very convenient for obtaining the output of a process in response to any input, and for handling complicated networks. Transfer functions are closely related to the frequency functions introduced in Section 5.2.1, and are applied for the same purpose. The transfer function of a process is defined by the following expression:

$$G(s)=\frac{Y(s)}{X(s)} \tag{5.63}$$

where $G(s)$ is the transfer function,

$Y(s)$ is the Laplace transform of the output signal,

$X(s)$ is the Laplace transform of the input signal, and

$s$ is a complex variable, the so-called Laplace transform variable.

Thus to use the transfer function method, Laplace transformation has to be performed.

*Laplace transformation*

Transform operations are used to transform difficult problems into more tractable forms. Laplace transformation is a method of transforming linear differential equations into algebraic equations.

**Example 5.3.3**

Transform the differential equation of the *PID*-controller:

$$m(t)=m(0)+A_P e(t)+A_I \int_0^t e(t)\,dt+A_D \frac{de(t)}{dt}$$

The Laplace transform is:

$$M(s)=A_P E(s)+A_I \frac{E(s)}{s}+A_D s E(s)$$

if

$$\int_{t=0} e\,dt = 0 \quad \text{and} \quad \left(\frac{de}{dt}\right)_0 = 0$$

The Laplace transform of a time domain function $f(t)$ is denoted by the symbol $F(s)$ and is defined as follows:

$$F(s) = \mathcal{L}[f(t)] = \int_0^\infty f(t)\, e^{-st} dt \tag{5.64}$$

where $\mathcal{L}[f(t)]$ is the symbol that indicates Laplace transformation of the function $f(t)$ in brackets,
$s = a + i\omega$ is a complex variable, the Laplace variable.

Laplace transformation is applicable to linear functions for which the integral in Eq. (5.64) is convergent. This is not a serious restriction.

Laplace transformation is a linear operation, as it can be seen from Example 5.3.3 which has been solved using the basic theorems listed in *Table 5.3*. These theorems are derived from the definition in Eq. (5.64), as well as the Laplace transforms of some functions most frequently used in control technique, listed in *Table 5.4*.

To perform Laplace transformation of a time function or the inverse transformation of a Laplace transform into a time function it is not necessary to resort to the definition given in Eq. (5.64); it is generally sufficient simply to look up the function in *Table 5.3* and/or *5.4*. The technique is illustrated by some examples.

**Example 5.3.4.** *First-order lag. Transfer function*

A first-order element is characterized by the first-order differential equation, Eq. (5.59):

$$T\frac{dy(t)}{dt} + y(t) = A_S x(t) \tag{5.65}$$

where $y(0) = 0$, and $\left(\frac{dy}{dt}\right)_0 = 0$ (steady state) and
$x(0) = 0$.

Having defined the functions and initial conditions, we can proceed to the transformation. Applying the linearity theorem, each term in Eq. (5.65) can be transformed individually:

$$\mathcal{L}\left[T\frac{dy(t)}{dt}\right] + \mathcal{L}[y(t)] = \mathcal{L}[A_S x(t)] \tag{5.66}$$

Table 5.3. Laplace transform theorems

1. Linearity theorem

$$\mathcal{L}[Kf(t)] = K\,\mathcal{L}[f(t)] = KF(s) \qquad (K = \text{constant})$$

$$\mathcal{L}[f_1(t) \pm f_2(t)] = F_1(s) \pm F_2(s)$$

2. Real differentiation theorem

First derivative

$$\mathcal{L}\left[\frac{\mathrm{d}}{\mathrm{d}t} f(t)\right] = sF(s) - f(0)$$

General $n$th derivative

$$\mathcal{L}\left[\frac{\mathrm{d}^n}{\mathrm{d}t^n} f(t)\right] =$$

$$= s^n F(s) - s^{n-1} f(0) - s^{n-2} \frac{\mathrm{d}}{\mathrm{d}t} f(0) - \ldots - s^2 \frac{\mathrm{d}^{n-2}}{\mathrm{d}t^{n-2}} f(0) - s \frac{\mathrm{d}^{n-1}}{\mathrm{d}t^{n-1}} f(0)$$

3. Real integration theorem

$$\mathcal{L}\left[\int f(t)\,\mathrm{d}t\right] = \frac{F(s)}{s}$$

$$\mathcal{L}\left[\overset{1}{\int}\overset{2}{\int}\ldots\overset{n}{\int} f(t)\,\mathrm{d}t^n\right] = \frac{1}{s^n} F(s)$$

4. Initial value theorem

$$f(0) = \lim_{s \to \infty} sF(s)$$

5. Final value theorem

$$f(\infty) = \lim_{s \to 0} sF(s)$$

6. Time-shift theorem

$$\mathcal{L}[f(t-a)] = e^{-at}\,\mathcal{L}[f(t)] = e^{-nt} \cdot F(s)$$

The transform of the first term is, using Theorems 1 and 2:

$$\mathcal{L}\left[T\frac{\mathrm{d}y(t)}{\mathrm{d}t}\right] = T\mathcal{L}\left[\frac{\mathrm{d}y(t)}{\mathrm{d}t}\right] = T(sY(s) - y(0))$$

$$= TsY(s)$$

The transform of the second term is by definition

$$\mathcal{L}[y(t)] = Y(s)$$

The transform of the third term is by definition

$$\mathcal{L}[x(t)] = X(s), \qquad \text{whence} \qquad \mathcal{L}[A_S x(t)] = A_S X(s)$$

The entire Laplace domain equation is

$$TsY(s)+Y(s)=A_S X(s) \tag{5.67}$$

which can be rearranged to obtain the transfer function defined by Eq. (5.63):

$$\frac{Y(s)}{X(s)}=G(s)=\frac{A_S}{Ts+1} \tag{5.68}$$

The transfer function of a process can be used to obtain its response to any input:

$$Y(s)=G(s)\,X(s) \tag{5.69}$$

Table 5.4. Laplace transform pairs

| $F(s)$ | $f(t)$ |
|---|---|
| $1$ | $\delta(t)$ (impulse) |
| $\frac{1}{s}$ | $u(t)$ (step) |
| $\frac{1}{s^n}$ $(n=1, 2, \ldots)$ | $\frac{t^{n-1}}{(n-1)!}$ |
| $\frac{1}{s\pm a}$ | $e^{\mp at}$ |
| $\frac{1}{s(s\pm a)}$ | $\frac{1}{\pm a}(1-e^{\mp at})$ |
| $\frac{s}{s^2+a^2}$ | $\cos at$ |
| $\frac{a}{s^2+a^2}$ | $\sin at$ |
| $\frac{s}{s^2-a^2}$ | $\mathrm{ch}\ at$ |
| $\frac{a}{s^2-a^2}$ | $\mathrm{sh}\ at$ |
| $\frac{1}{(s+a)^2+b^2}$ | $\frac{1}{b}e^{-at}\sin bt$ |
| $\frac{s+a}{(s+a)^2+b^2}$ | $e^{-at}\cos bt$ |
| $\frac{1}{(s+a)^n}$ | $\frac{1}{(n-1)!}t^{n-1}e^{-at}$ |
| $\frac{ab}{(s+a)(s+b)}$ | $\frac{1}{b-a}(e^{-at}-e^{-bt})$ |

To obtain the impulse response, the Laplace transform of the unit impulse is to be substituted for $X(s)$ in Eq. (5.69). From *Table 5.4* it is found that

$$\text{for} \quad f(t)=1\cdot\delta(t) \qquad F(s)=1$$

Thus the Laplace transform of the impulse response of any process is

$$Y(s)=G(s) \tag{5.70}$$

in other words, the transfer function of a process is obtained by the Laplace transformation of its impulse response.

**Example 5.3.5** *First-oder lag. Impulse response*

The Laplace transform of the impulse response of the first-oder lag is obtained by substituting Eq. (5.68) into Eq. (5.69):

$$Y(s)=\frac{A_S}{Ts+1}=\frac{\frac{A_S}{T}}{s+\frac{1}{T}}=\frac{A_S}{T}\left(\frac{1}{s+\frac{1}{T}}\right) \tag{5.71}$$

The time domain function of Eq.(5.68) is found in *Table 5.4* as

$$y(t)=\frac{A_S}{T}e^{-\frac{t}{T}} \tag{5.72}$$

The impulse response of a first-oder lag is shown in *Fig. 5.38*.

From Eq.(5.72):

$$y(0)=\frac{A_S}{T}\left(\frac{dy}{dt}\right)_0=-\frac{A_S}{T^2}$$

$$y(T)=\frac{A_S}{T}\cdot 0.368$$

$$y(\infty)\to 0$$

**Example 5.3.6.** *First-order lag. Step response*

The Laplace transform of the step response of the first-order element can be found in two ways:

1. by multiplying its transfer function by the Laplace transform of the unit step input;
2. by substituting $x(t)=1\cdot u(t)$, the expression for unit step input, into its differential equation given in Eq. (5.65).

The first method has been illustrated: to obtain the impulse response. Now use the second way.

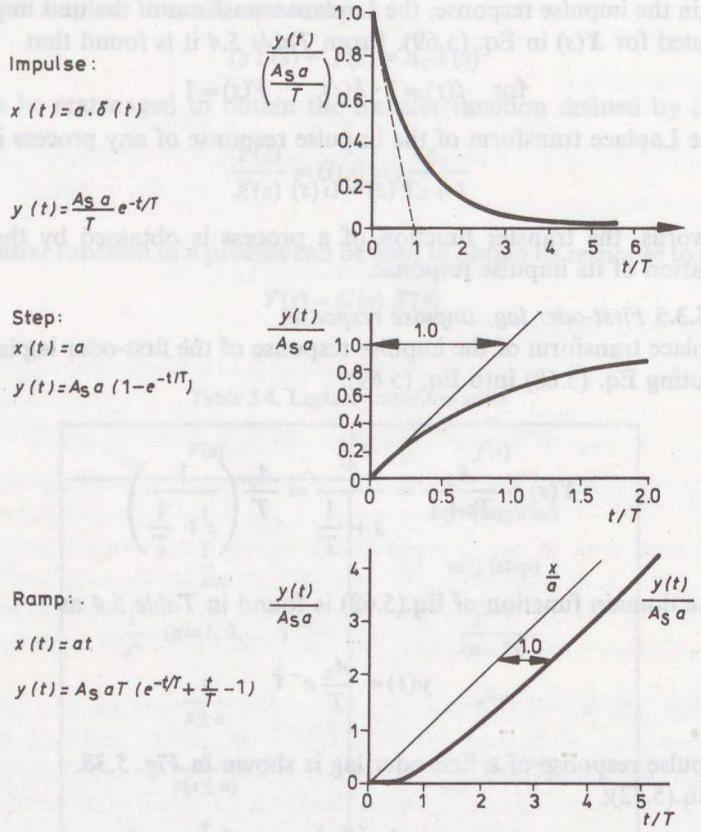



Fig. 5.38. Response functions of a first-order lag

Equation (5.65) is to be transformed, with $x(t)=1\cdot u(t)$, unit step input occurring at $t=0$.

The Laplace transforms of terms 1 and 2 were found in Example 5.3.4. Using *Table 5.4*:

$$\mathcal{L}[A_S u(t)] = A_S \frac{1}{s}.$$

The Laplace transform of Eq. (5.65) for $x(t)=u(t)$ is obtained as:

$$TsY(s) + Y(s) = A_S \frac{1}{s} \tag{5.73}$$

from which $Y(s)$ is expressed as:

$$Y(s) = \frac{A_S}{T} \frac{1}{s\left(s+\frac{1}{T}\right)} \tag{5.74}$$

This can be found in *Table 5.4*:

$$y(t)=A_S\left(1-e^{-\frac{t}{T}}\right) \tag{5.75}$$

The values of the step response function are:

$$y(0)=A_S(1-1)=0 \qquad \left(\frac{dy}{dt}\right)_0=\frac{A_S}{T}=\frac{y(\infty)}{T}$$

$$y(T)=A_S(1-e^{-1})=A_S\cdot 0.632$$

$$y(\infty)\to A_S$$

The function described by Eq. (5.75) is shown in *Fig. 5.38*.

*Inversion of Laplace transforms by expansion in partial fractions*

The inverse transforms of functions not listed in *Table 5.4* can be obtained by several methods.

Inversion of a Laplace transform is denoted by

$$\mathcal{L}^{-1}[F(s)]=f(t) \tag{5.76}$$

The inverse transform can be obtained by contour integration in the complex plane:

$$f(t)=\frac{1}{2\pi i}\int_{a-i\omega}^{a+i\omega} e^{st}F(s)\,ds \tag{5.77}$$

Other methods for inversion are the convolution method and expansion in partial fractions. The partial fraction method is presented here. It consists in converting the function to be inverted into the form of a sum of simpler functions, the inverse transforms of which can be found in the transform tables:

$$F(s)=F_1(s)+F_2(s)+\ldots+F_n(s) \tag{5.78}$$

The Laplace transform, for which we require the time inverse generally has the form

$$F(s)=\frac{P(s)}{Q(s)} \tag{5.79}$$

which is obtained by the Laplace transformation of the general ordinary differential equation, Eq.(5.57). Here $P(s)$ and $Q(s)$ are polynomials in $s$ such that the order of the numerator is less than that of the denominator (if not, then it has to be converted into such a form by division).

Next the roots of polynomial $Q(s)$ are to be obtained. $Q(s)$ is factored into its roots:

$$Q(s)=(s-s_1)(s-s_2)\ldots(s-s_n)=0 \tag{5.80}$$

On the basis of Eq.(5.80) the function $F(s)$ can be expanded in terms of the roots of $Q(s)$:

$$F(s)=\frac{P(s)}{Q(s)}=\frac{C_1}{s-s_1}+\frac{C_2}{s-s_2}+\ldots+\frac{C_n}{s-s_n} \tag{5.81}$$

The inverse transform of Eq.(5.81) is simply (cf. *Table 5.4*)

$$f(t)=C_1e^{s_1t}+C_2e^{s_2t}+\ldots+C_ne^{s_nt} \tag{5.82}$$

The constants $C_1$, $C_2$ ... $C_n$ can be determined by making use of the fact that Eq.(5.81) is an identity that is valid for every value of $s$ or by applying the following relation for distinct roots:

$$C_i=\lim_{s\to s_i}[(s-s_i)F(s)], \qquad i=1\ldots n \tag{5.83}$$

Equation (5.83) can be checked by applying it to the right-hand-side of the identity [Eq.(5.81)]. Approaching the limit $s\to s_i$, all terms containing $C_j$ (where $j\neq i$) become zero, except the term containing $C_i$, which is $C_i\cdot 1$.

If some roots of $Q(s)$ are repeated, the procedure is slightly different. Consider, for instance, that one of the roots of $Q(s)$ is repeated twice, thus $Q(s)$ is factored into the following equation [$Q(s)$ is of order $(n+1)$, it has $n$ roots]:

$$Q(s)=(s-s_1)^2(s-s_2)\ldots(s-s_n) \tag{5.84}$$

In this case $F(s)$ is to be expanded as follows:

$$F(s)=\frac{P(s)}{Q(s)}=\frac{C_1}{(s-s_1)^2}+\frac{C_2}{s-s_1}+\frac{C_3}{s-s_2}+\ldots+\frac{C_{n+1}}{s-s_n} \tag{5.85}$$

Constants $C_1$ and $C_2$ are determined as:

$$C_1=\lim_{s\to s_1}[(s-s_1)^2F(s)] \tag{5.86}$$

$$C_2=\lim_{s\to s_1}\left[\frac{d}{ds}(s-s_1)^2F(s)\right] \tag{5.87}$$

**Example 5.3.7.** *First-order lag. Ramp response*

The unit ramp function is $x(t)=1\cdot t$ *(Fig. 5.36)*. Its Laplace transform is $X(s)=\frac{1}{s^2}$.

The Laplace transform of the ramp response of a first-order lag is obtained from its transfer function

$$Y(s)=G(s)X(s)=\left(\frac{A_S}{Ts+1}\right)\left(\frac{1}{s^2}\right) \tag{5.88}$$

Equation (5.88) is to be rearranged so that the coefficient of the highest power of $s$ in the denominator is 1:

$$Y(s)=\frac{A_S}{T}\left(\frac{1}{s+\frac{1}{T}}\right)\left(\frac{1}{s^2}\right) \tag{5.89}$$

It is not necessary to write $Q(s)$ in the polynomial form, since it is already in the required factored form. It is evident that $Q(s)$ has three roots:

$$s_1=-\frac{1}{T}, \qquad s_2=0, \qquad s_3=0$$

Applying Eq.(5.85), $Y(s)$ is expanded in partial fractions:

$$Y(s)=\frac{A_S}{T}\left(\frac{C_1}{s+\frac{1}{T}}+\frac{C_2}{s^2}+\frac{C_3}{s}\right) \tag{5.90}$$

The transform of Eq.(5.90) is obtained by using *Table 5.3*:

$$y(t)=\frac{A_S}{T}\left(C_1 e^{-\frac{t}{T}}+C_2 t+C_3\right) \tag{5.91}$$

$C_1$ is determined by applying Eq.(5.89):

$$C_1=\lim_{s\to-\frac{1}{T}}\frac{s+\frac{1}{T}}{\left(s+\frac{1}{T}\right)s^2}=T^2 \tag{5.92}$$

$C_2$ is determined by applying Eq.(5.86) to Eq.(5.89):

$$C_2=\lim_{s\to 0}\frac{s^2}{\left(s+\frac{1}{T}\right)s^2}=T \tag{5.93}$$

$C_3$ is determined by applying Eq.(5.87):

$$C_3=\lim_{s\to 0}\left[\frac{\mathrm{d}}{\mathrm{d}s}\frac{s^2}{\left(s+\frac{1}{T}\right)s^2}\right]=\lim_{s\to 0}\left[\frac{\mathrm{d}}{\mathrm{d}s}\frac{1}{\left(s+\frac{1}{T}\right)}\right]=$$

$$=\lim_{s\to 0}-\frac{1}{\left(s+\frac{1}{T}\right)^2}=-T^2 \tag{5.94}$$

$C_1$, $C_2$ and $C_3$ are substituted into Eq.(5.91) to obtain the required time re-

sponse:

$$y(t)=\frac{A_S}{T}\left(T^2e^{-\frac{t}{T}}+T\cdot t-T^2\right)=A_S\left[Te^{-\frac{t}{T}}+(t-T)\right] \tag{5.96}$$

The response of a *P*-element with a gain $A_S$ to a unit ramp input $1\cdot t$ is $y(t)=$ $=A_St$. This can be compared to the response of the first order element, which is

$$y(t)\cong A_S(t-T) \qquad \text{for} \quad t>3T \tag{5.97}$$

since $e^{-3}=0.05$, $e^{-4}=0.018$, the first term in Eq.(5.96) dies out with time.

The ramp response is shown in *Fig. 5.38*. It can be seen from both the figure and from Eq.(5.97) that the ramp response of a first-order process lags by $T$, its time constant, behind that of a proportional element.

*Block diagrams*

The transfer function representation has the convenience that the output of a process is obtained by simply multiplying its transfer function by the input: $Y(s)=G(s)X(s)$, cf. Eq.(5.63). Owing to this feature it is very suitable for the treatment of networks. Moreover, since the Laplace transformation is linear, addition and subtraction of time domain signals also holds for their Laplace transforms.

A similar quality of the frequency functions has already been utilized in Section 5.2.1.

In the frequency function representation the elements were represented by their frequency functions. In the transfer function representation the elements are represented by their transfer functions, which are the blocks, and the signals are represented by their Laplace transforms, which are the arrows.

The overall frequency functions were derived for elements in series (Eq.5.2), in parallel (Eq.5.3) and for a simple negative control loop (Eq.5.4). Analogous equations are obtained for the Laplace transforms of the signals and the transfer functions of the elements.

Elements in series:

$$G(s)=G_1(s)\cdot G_2(s)\cdot G_3(s)\cdot \ \ldots\ \cdot G_n(s) \tag{5.98}$$

in parallel:

$$G(s)=G_1(s)+G_2(s)+\ \ldots\ +G_n(s) \tag{5.99}$$

The transfer functions for a simple negative feedback control loop, with a block diagram shown in *Fig. 5.39* can be derived as

$$W_D(s)=\frac{G(s)}{D(s)}=\frac{G_D(s)}{1+G_C(s)G_s(s)} \tag{5.100}$$

$$W(s)=\frac{G(s)}{R(s)}=\frac{G_C(s)G_s(s)}{1+G_C(s)G_s(s)} \tag{5.101}$$

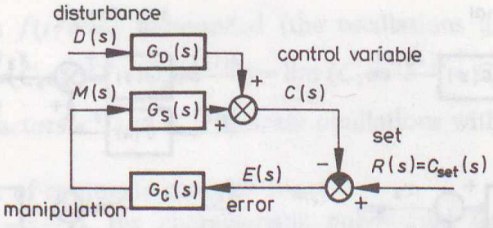



Fig. 5.39. Negative feedback loop. To the derivation of its transfer function

The block diagram of a positive feedback loop is shown in *Fig. 5.40*. The transfer function for this loop is derived as:

$$W(s)=\frac{Y(s)}{X(s)}=\frac{G(s)}{1-G(s)H(s)} \tag{5.102}$$

To facilitate calculations with the block diagram linear transformations are possible. Some of these are shown in *Fig. 5.41*.

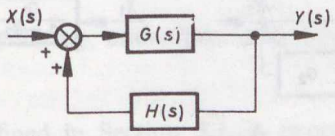



Fig. 5.40. Positive feedback loop

*Stability analysis*

Recall Eq. (5.82) for the time inverse of a Laplace transform:

$$f(t)=C_1e^{s_1t}+C_2e^{s_2t}+\ldots+C_ne^{s_nt} \tag{5.103}$$

Here, $s_1$, $s_2$,... $s_n$ are complex numbers obtained as the roots of $Q(s)$, the characteristic polynomial [cf. Eq. (5.80)], and the roots are the characteristics of the time function $f(t)$, which may be unbounded, bounded or oscillating.

*Case 1.* All characteristic roots are real negative numbers,
e. g. $s_1=-k_1$, $s_2=-k_2$, etc.
Then $\lim_{t\to\infty} f(t)=0$, $f(t)$ is bounded, because $\lim_{t\to\infty} e^{-kt}=0$

*Case 2.* One of the characteristic roots is a positive real number, $s_i=+k$.
Then $\lim_{t\to\infty} f(t)=\infty$, unbounded, because $\lim_{t\to\infty} e^{+kt}=\infty$

*Case 3.* There is a complex pair among the characteristic roots; they always appear as conjugate pairs.

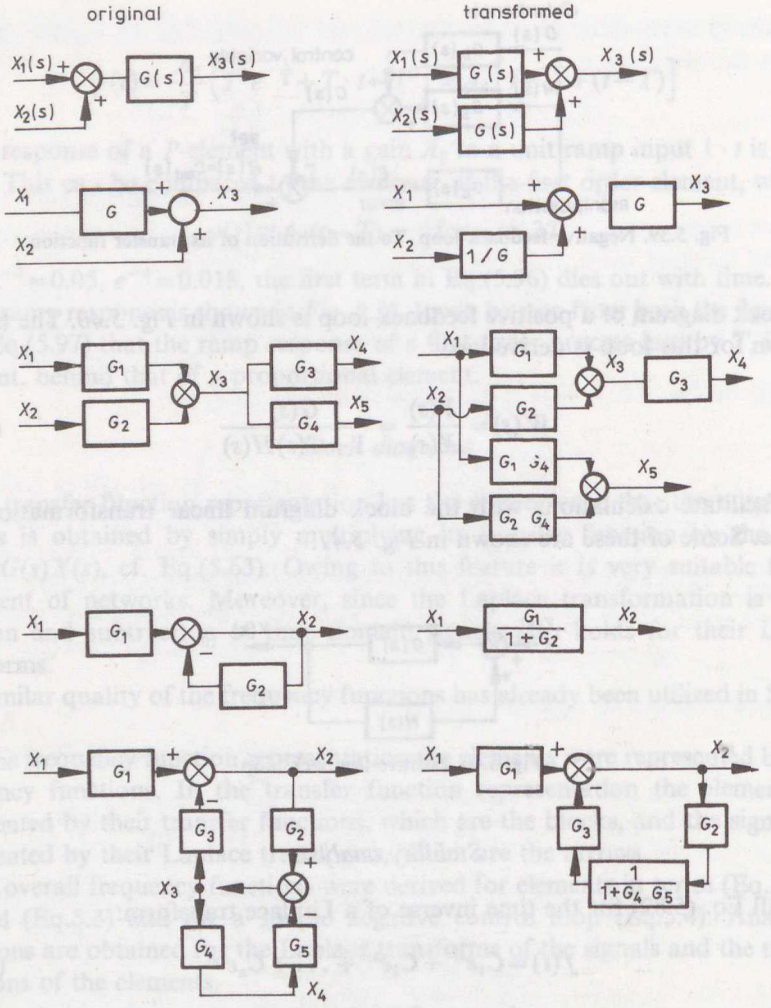



Fig. 5.41. Linear transformations of the block diagram

*Case 3a.* $s_k = -a+ib$, $s_m = -a-ib$, i. e. the real parts of the conjugate complex roots are negative.

Then $\lim_{t\to\infty} f(t)=0$, bounded, with damped oscillation, because

$\lim_{t\to\infty} (C_k e^{-a+ib} + C_m e^{-a-ib}) = 0$

and the factors $e^{ib}$ and $e^{-ib}$ indicate oscillations with frequency $b$.

*Case 3b.* $s_k = +a+ib$, $s_m = +a-ib$, i. e. the real parts of the conjugate complex roots are positive.

Then $\lim_{t\to\infty} f(t)=\infty$, unbounded (the oscillations are expanding), because $\lim_{t\to\infty}(C_k e^{+a+ib}+C_m e^{+a-ib})=\lim_{t\to\infty}(C_k e^{+a}e^{-ib}+C_m e^{+a}C_m e^{-ib})=\infty$, and the factors $e^{ib}$ and $e^{-ib}$ indicate oscillations with frequency $b$.

*Note*: For a pair of conjugate complex roots, $a_k=-a+ib$, $s_m=-a-ib$, the characteristic polynomial $Q(s)$ is conveniently factored as $(s-s_k)(s-s_m)=(s+a-ib)(s+a+ib)=s^2+2as+a^2+b^2=(s+a)^2+b^2$.

In the expansion in partial fractions this will give the two terms:

$$\frac{C_k b+C_m(s+a)}{(b+a)^2+b^2}$$

for which *Table 5.4* gives

$$\mathcal{L}^{-1}\left[\frac{C_k b}{(s+a)^2+b^2}\right]+\mathcal{L}^{-1}\left[\frac{C_m(s+a)}{(s+a)^2+b^2}\right]=$$
$$=C_k e^{-at}\sin bt+C_m e^{-at}\cos bt$$

*Case 4.* There is an imaginary pair among the characteristic roots: $s_k=ib$, $s_m=-ib$.
Then $f(t)$ is oscillating, undamped with a constant amplitude and frequency $b$.

Stability has been defined in Section 5.1. A process is stable when, after a pulse-like disturbance, it returns to its original state within some time. This is equivalent to the statement that the limit of the impulse response of a stable process is zero, for $t\to\infty$.

For a unit impulse, $f(t)=1\cdot\delta(t)$, $F(s)=1$.

The Laplace transform of the impulse response is:

$$Y(s)=G(s)\,X(s)=G(s)$$

Thus for a stable process:

$$\lim_{t\to\infty}(\mathcal{L}^{-1}[G(s)])=0 \qquad (5.104)$$

The stability definition expressed by Eq. (5.104) suggests the stability criterion that, for a stable linear process, the real parts of all characteristic roots of the transfer function must be smaller than 0. If the characteristic roots of the transfer function are plotted in the complex plane, all roots must lie in the left half-plane *(Fig. 5.42)*.

Unfortunately, the stability criterion based on the examination of the characteristic roots is of only limited use.

The roots of the denominator of the transfer function must be known, which may cause problems for a higher-order process. For a control loop, even if the

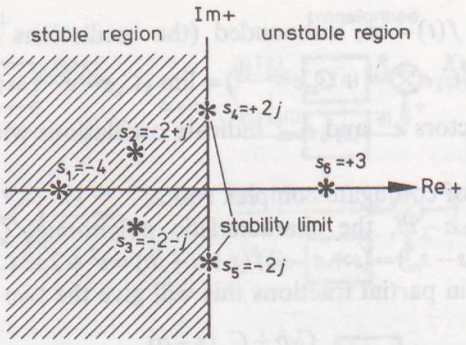



Fig. 5.42. Stable region of characteristic roots

roots of the open loop transfer function $G_L(s)=G_C(s)\,G_S(s)$ are known from the model of the process [refer to Eqs (5.100) and (5.101)], and even if the open loop is stable, the characteristic roots of the closed loop transfer function, i.e. the roots of the characteristic equation

$$1+G_C(s)\,G_S(s)=0 \tag{5.105}$$

are unknown, and the closed loop may be unstable. Equation (5.105) cannot be solved to find its roots if dead time is involved, because a dead time process can be approximated by a differential equation of infinite order.

*Routh–Hurwitz stability criterion*

The Routh–Hurwitz stability criterion is used to check the stability of a system for which the characteristic equation can be written and is of finite order, without solving the equation to find the roots. The procedure can also be applied to control loops, where the characteristic equation is:

$$Q(s)=1+G_C(s)\,G_S(s)=0 \tag{5.106}$$

Expand the characteristic equation into the following form:

$$Q(s)=a_n s^n+a_{n-1}s^{n-1}+\ldots+a_1 s+a_0=0 \tag{5.107}$$

Let $a_n$ be positive. If it is negative, both sides of Eq. (5.107) should be multiplied by $-1$.

*First step*: If any of the coefficients $a_n$, $a_{n-1}\ldots a_0$ is negative, there is at least one characteristic root which has a positive real part, and the system is unstable. No further analysis is needed.

*Second step*: If every coefficient is positive, then the following array (Routh array) should be constructed:

$$\begin{array}{lllll} 1 & a_n & a_{n-2} & a_{n-4} & \cdots \\ 2 & a_{n-1} & a_{n-3} & a_{n-5} & \cdots \\ 3 & b_1 & b_2 & b_3 & \cdots \\ 4 & c_1 & c_2 & c_3 & \cdots \\ 5 & d_1 & d_2 & d_3 & \cdots \\ \vdots & & & & \\ n-1 & w_1 & w_2 & w_3 & \end{array} \tag{5.108}$$

where the coefficients $c$, $d$, . . . , etc. are calculated as

$$b_1 = \frac{a_{n-1}a_{n-2} - a_n a_{n-3}}{a_{n-1}}$$

$$b_2 = \frac{a_{n-1}a_{n-4} - a_n a_{n-5}}{a_{n-1}}$$

$$b_3 = \frac{a_{n-1}a_{n-6} - a_n a_{n-7}}{a_{n-1}}$$

$$\vdots$$

$$c_1 = \frac{b_1 a_{n-3} - a_{n-1}b_2}{b_1}$$

$$c_2 = \frac{b_1 a_{n-5} - a_{n-1}b_3}{b_1}$$

$$\vdots$$

$$d_1 = \frac{c_1 b_2 - b_1 c_2}{c_1}$$

$$d_2 = \frac{c_1 b_3 - b_1 c_3}{c_1}$$

$$\vdots$$

$$\text{etc.} \tag{5.109}$$

Examine the elements of the first column of the array in (Eq. 5.108):

$$a_n,\ n_{n-1},\ b_1,\ c_1,\ d_1,\ \ldots,\ w_1$$

A system is stable if all elements in the first column of the Routh array are positive.

**Example 5.3.8**

Check the stability of the process having the following transfer function:

$$G(s) = \frac{A}{1 - Ts} = -\frac{\frac{A}{T}}{s - \frac{1}{T}}$$

The characteristic equation of this process is:

$$Q(s)=s-\frac{1}{T}=0$$

The characteristic root is evident: $s_1=+1/T$. It is positive, thus the process is unstable.

Check the stability of this process by the Routh–Hurwitz criterion. $a_n=1$, i. e. it is positive, $a_{n-1}=-1/T$, i. e. negative. Thus, the process is found to be unstable by this method, too. The impulse response of this process is obtained as:

$$y(t)=\mathcal{L}^{-1}[G(s)\,X(s)]=\mathcal{L}^{-1}\left[\frac{\frac{A}{T}}{s-\frac{1}{T}}\cdot 1\right]$$

$$y(t)=-\frac{A}{T}e^{+\frac{t}{T}}$$

$y\to\infty$ when $t\to\infty$. Thus, the process is unstable.

Stabilize this unstable process using proportional control. Denote the proportional gain of the *P*-controller by $A_C$. The block diagram of the control loop is shown in *Fig. 5.43*. Eqs (5.100) and (5.101):

$$Q(s)=1+G_C(s)\,G_S(s)=1+\left(\frac{A}{1-Ts}\right)A_C=0$$

which can be arranged to

$$Q(s)=1-Ts+AA_C=-Ts+(1+AA_C)=0$$

The characteristic root is

$$s_1=+\frac{1+AA_C}{T}$$

$s_1$ is negative, and the control loop is stable if

$$A_C<0 \qquad \text{and} \qquad |A_C|>\frac{1}{A}$$

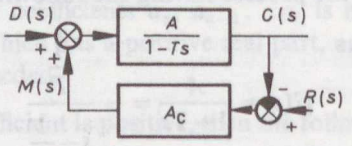



Fig. 5.43. To Example 5.3.8

### 5.3.5 The frequency function method

The analysis of frequency response is the third alternative method (besides the differential equation method and the transfer function method) for describing and examining the dynamics of linear systems. All these methods contain essentially the same information.

The frequency function method has already been mentioned and used in Section 5.2. It has the advantage of being directly applicable to dead time processes and to distributed parameter processes, and of being capable of utilizing direct experimental results. Frequency functions can be depicted as Nyquist and Bode plots; the resulting Bode plot of elements in series can easily be constructed by the simple addition of distances in the diagram. The stability of a negative feedback loop can easily be checked by the use of the Nyquist stability criterion (Section 5.2.1).

*Determination of the frequency function*

The frequency function of a process can be determined in three ways.

1. Experimentally, as already described in Section 5.2.1 and shown in *Fig. 5.11.* The process is driven by a continous sinusoidal signal until, after some time, a sine wave of the same frequency can be detected at the output. The result of the experiment is
   (*a*) the amplitude ratio of the output to the input sine wave, this is the magnitude of the frequency function vector $|G(i\omega)|$;
   (*b*) the phase shift of the output with respect to the input, which is the phase angle argument of the frequency function, $\measuredangle G(i\omega)$.

   The aim of the investigation is to obtain the frequency function over a broad frequency range, so it is a lengthy and expensive process (in the case of chemical processes, which are slow), but in the case of noisy signals it is advantageous that only one single frequency signal is to be detected.

   The frequency function can also be computed from responses to input signals containing a broad spectrum of frequencies, e. g. from impulse or step responses, or responses to white noise with the aid of Fourier analysis.
2. From the differential equation of the process, which can be solved substituting $x(t) = a \sin \omega t$, in simple cases, like in Section 5.2 for dead time, interval and differential elements.
3. From the transfer function, making the formal substitution $s = i\omega$.

*Frequency function from transfer function*

The procedure is illustrated by the example of the first-order process.

**Example 5.3.9.** *First-order lag. Frequency function*

The differential equation of a first-order element is given by Eq.(5.59):

$$T\frac{dy}{dt}+y=A_S x$$

By Laplace transformation:

$$TsY(s)+Y(s)=A_S X(s)$$

This can be rearranged to obtain the transfer function:

$$\frac{Y(s)}{X(s)}\equiv G(s)=\frac{A_S}{Ts+1}$$

Substitute $s=i\omega$

$$G(i\omega)=A_S\frac{1}{Ti\omega+1}\frac{(-iT\omega+1)}{(-iT\omega+1)}=A_S\left(\frac{1}{1+T^2\omega^2}-i\frac{T}{1+T^2\omega^2}\right) \quad (5.110)$$

This has the form $G(i\omega)=\mathrm{Te}_G+i\,\mathrm{Im}_G$ [cf. Eq.(5.9), *Fig. 5.12*]. The magnitude and the phase are obtained as:

$$|G(i\omega)|=\sqrt{\mathrm{Re}_G^2+\mathrm{Im}_G^2}=$$

$$=A_S\sqrt{\left(\frac{1}{1+T^2\omega^2}\right)^2+\left(\frac{T\omega}{1+T^2\omega^2}\right)^2}= \quad (5.111)$$

$$=A_S\frac{1}{\sqrt{1+T^2\omega^2}}$$

$$\measuredangle G(i\omega)=\arctan\frac{\mathrm{Im}_G}{\mathrm{Re}_G}=\arctan(-\omega T). \quad (5.112)$$

The frequency function can be tabulated for different frequencies, e.g.

| $\omega$ | $\omega T$ | $\lvert G(i\omega)\rvert$ | $\measuredangle G(i\omega)$ |
|---|---|---|---|
| 0 | 0 | $A_S$ | 0° |
| $\frac{1}{T}$ | 1 | $A_S\frac{1}{\sqrt{2}}$ | −45° |
| $\to\infty$ | $\to\infty$ | $A_S\frac{1}{T\omega}$ | $\to$ −90° |

For small frequencies ($\omega<1/T$) (e. g. close to the steady state) the first-order element resembles a *P* element; for high frequencies ($\omega>1/T$) (e. g. at the start of

a step or a pulse driving signal) it resembles to an $I$ element. The Bode and the Nyquist diagrams of the first-order element are shown in *Fig. 5.44*. The amplitude characteristic in the Bode diagram is 'broken' at frequency $1/T$, called also the 'corner frequency' or 'breakpoint frequency'.

The general method for obtaining the frequency function from a transfer function can now be summarized:

1. $s = i\omega$ is substituted into the transfer function.
2. The expression obtained is arranged in the form:

$$G(i\omega) = \mathrm{Re}_G(i\omega) + i\,\mathrm{Im}_G(i\omega) \tag{5.113}$$

3. The magnitude of the frequency function is:

$$|G(i\omega)| = \sqrt{[\mathrm{Re}_G(i\omega)]^2 + [\mathrm{Im}_G(i\omega)]^2} \tag{5.114}$$

4. The phase angle of the frequency function is:

$$\sphericalangle G(i\omega) = \arctan \frac{\mathrm{Im}_G(i\omega)}{\mathrm{Re}_G(i\omega)} \tag{5.115}$$

5. The frequency function is evaluated for different frequencies.

It still remains to prove that a simple substitution of $s = i\omega$ into the transfer function of an arbitrary $n$th order system yields the same frequency function as that obtained experimentally by driving the system with a continuous sine wave.

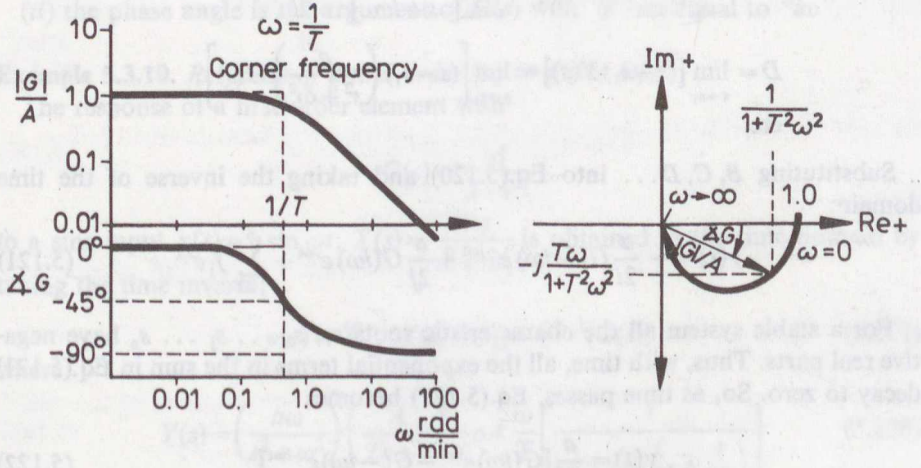



Fig. 5.44. Bode and Nyquist plots of first-order lag

The transfer function is:

$$G(s)=\frac{Y(s)}{X(s)} \tag{5.116}$$

where both $Y(s)$ and $X(s)$ are reciprocals of polynomials in $s$; $Y(s)$ are of $n$th and $X(s)$ are of $m$th order where $m<n$, and both polynomials can be factored into their roots:

$$G(s)=\frac{Y(s)}{X(s)}=\frac{(s-p_1)(s-p_2)\ldots(s-p_m)}{(s-s_1)(s-s_2)\ldots(s-s_n)} \tag{5.117}$$

The input is a sine wave the Laplace transform of which is, from *Table 5.4*,

$$X(a)=a\left(\frac{\omega}{s^2+\omega^2}\right) \tag{5.118}$$

$$Y(s)=X(s)G(s)=\left[\frac{a\omega}{s^2+\omega^2}\right]\left[\frac{(s-p_1)(s-p_2)\ldots(s-p_m)}{(s-s_1)(s-s_2)\ldots(s-s_n)}\right] \tag{5.119}$$

Expansion in partial fractions gives

$$Y(s)=\frac{aG(s)}{s^2+\omega^2}=\frac{B}{s+i\omega}+\frac{C}{s-i\omega}+\frac{D}{s-s_1}+\ldots+\frac{Q}{s-s_n} \tag{5.120}$$

Determine $B, C, D, \ldots Q$, using Eq.(5.83) in Section 5.3.4, and making use of the fact that Eqs (5.119) and (5.120) are identical equations for all values of $s$:

$$B=\lim_{s\to -i\omega}[(s+i\omega)Y(s)]=\lim_{s\to i\omega}\left[\frac{\omega aG(s)}{s-i\omega}\right]=-\frac{a}{2i}G(-i\omega)$$

$$C=\lim_{s\to i\omega}[(s-i\omega)Y(s)]=\lim_{s\to i\omega}\left[\frac{\omega aG(s)}{s+i\omega}\right]=\frac{a}{2i}G(i\omega)$$

$$D=\lim_{s\to s_1}[(s-s_1)Y(s)]=\lim_{s\to s_1}\left[(s-s_1)\left(\frac{a}{s^2+\omega^2}\right)G(s)\right],$$

etc.

Substituting $B, C, D\ldots$ into Eq.(5.120) and taking the inverse of the time domain:

$$y(t)=-\frac{a}{2i}G(-i\omega)e^{-i\omega t}+\frac{a}{2j}G(i\omega)e^{i\omega t}+\sum_{j=1}^{n}f_j e^{s_j t} \tag{5.121}$$

For a stable system all the characteristic roots $s_1, s_2 \ldots s_j \ldots s_n$ have negative real parts. Thus, with time, all the exponential terms in the sum in Eq.(5.121) decay to zero. So, as time passes, Eq.(5.121) becomes

$$y(t)=\frac{a}{2i}[G(i\omega)e^{i\omega t}-G(-i\omega)e^{-i\omega t}] \tag{5.122}$$

*Note*: $G(i\omega)$ and $G(-i\omega)$ are conjugate complex functions, which are characterized by equal real parts and imaginary parts that are also equal but opposite in sense. Represented in polar coordinates, the magnitudes are equal, and the arguments are equal too, but of opposite signs.

Using the Euler equations:

$$\begin{aligned} e^{ix} &= \cos x + i \sin x \\ e^{-ix} &= \cos x + i \sin x \end{aligned} \tag{5.123}$$

we obtain

$$\begin{aligned} \sin x &= \frac{1}{2i}(e^{ix} - e^{-ix}) \\ \cos x &= \frac{1}{2i}(e^{ix} + e^{-ix}) \end{aligned} \tag{5.124}$$

Write $G(i\omega)$ and $G(-i\omega)$ in the polar form of complex numbers:

$$G(i\omega) = |G(i\omega)|\, e^{i \measuredangle G(i\omega)}$$

$$G(-i\omega) = |G(-i\omega)|\, e^{i \measuredangle G(-i\omega)} = |G(i\omega)|\, e^{-i \measuredangle G(i\omega)}$$

Using these forms:

$$y(t) = a|G(i\omega)| \frac{e^{i[\omega t + \measuredangle G(i\omega)]} - e^{-i[\omega t + \measuredangle G(i\omega)]}}{2i} \tag{5.125}$$

we obtain the result:

$$y(t) = a[|G(i\omega)| \sin(\omega t) + \arg G(i\omega)] \tag{5.126}$$

Thus it has been proved that

(*i*) the amplitude ratio is the absolute value of $G(s)$ with "$s$" set equal to "$i\omega$";
(*ii*) the phase angle is the argument of $G(s)$ with "$s$" set equal to "$i\omega$".

**Example 5.3.10.** *Response of a first-order lag to a sine input*

The response of a first-order element with

$$G(s) = \frac{A}{Ts + 1}$$

to a sine input $x(t) = a \sin \omega t$; $X(s) = \dfrac{a\omega}{s^2 + \omega^2}$ is obtained in the time domain by taking the time inverse

$$y(t) = \mathcal{L}^{-1}[X(s)G(s)] = \mathcal{L}^{-1}[Y(s)] \tag{5.127}$$

where

$$Y(s) = \left(\frac{a\omega}{s^2 + \omega^2}\right)\left(\frac{A}{Ts + 1}\right) = aA\frac{\omega}{T}\left[\frac{1}{(s^2 + \omega^2)\left(s + \frac{1}{T}\right)}\right] \tag{5.128}$$

The roots of the characteristic equation

$$Q(s)=(s^2+\omega^2)\left(s+\frac{1}{T}\right)=0$$

are evidently $+i\omega$, $-i\omega$ and $-1/T$. However, to reduce complex algebraic work, it is more practical to factor $Y(s)$ as

$$Y(s)=aA\frac{\omega}{T}\left(\frac{B\omega+Cs}{s^2+\omega^2}+\frac{D}{s+\frac{1}{T}}\right) \tag{5.129}$$

which, transformed into the time domain, gives

$$y(t)=aA\frac{\omega}{T}\left(B\sin\omega t+C\cos\omega t+De^{-\frac{t}{T}}\right)$$

$$=aA\frac{\omega}{T}\left(\sqrt{B^2+C^2}\sin\left(\omega t+\arctan\frac{C}{B}\right)+De^{-\frac{t}{T}}\right) \tag{5.130}$$

The coefficients $B$ and $C$ are evaluated from

$$B\omega+Ci\omega=\lim_{s\to i\omega}\frac{(s^2+\omega^2)}{(s^2+\omega^2)\left(s+\frac{1}{T}\right)}=\lim_{s\to i\omega}\frac{1}{\left(s+\frac{1}{T}\right)}=$$

$$=\left(\frac{1}{i\omega+\frac{1}{T}}\right)\left(\frac{\frac{1}{T}-i\omega}{\frac{1}{T}-i\omega}\right)=\frac{T-i\omega T^2}{1+\omega^2T^2} \tag{5.131}$$

whence

$$B=\frac{\frac{\omega}{T}}{1+\omega^2T^2};\qquad C=-\frac{T^2}{1+\omega^2T^2} \tag{5.132}$$

$$\arctan\frac{C}{B}=\arctan(-\omega T)=\varphi \tag{5.133}$$

and

$$aA\frac{\omega}{T}\sqrt{B^2+C^2}=aA\sqrt{\frac{1+\omega^2T^2}{(1+\omega^2T^2)^2}}=aA\frac{1}{\sqrt{1+\omega^2T^2}} \tag{5.134}$$

Coefficient $D$ is evaluated using Eq.(5.83):

$$D=\lim_{s\to-1/T}\frac{s+\frac{1}{T}}{(s^2+\omega^2)\left(s+\frac{1}{T}\right)}=\frac{1}{\frac{1}{T^2}+\omega^2}=\frac{T^2}{1+\omega^2T^2} \tag{5.135}$$

Substitute Eqs (5.133), (5.134) and (5.135) into Eq.(5.130):

$$y(t)=aA\left[\left(\frac{1}{\sqrt{1+\omega^2T^2}}\right)\sin\left(\omega t-\arctan\left(-\omega T\right)\right)+\left(\frac{\omega T}{1+\omega^2T^2}\right)e^{-t/T}\right] \quad (5.136)$$

Compare Eq.(5.136) with the result of Example 5.3.9 given in Eqs (5.111) and (5.112). After a time longer than $3T$ the transient decays to zero and the steady-state part of the response remains: this is a sine wave with the same frequency $\omega$ as the input. Its amplitude is $\frac{A}{\sqrt{1+\omega^2T^2}}$ times that of the input. Its phase lag with respect to the input is

$$\varphi=\arctan\left(-\omega T\right)$$

In the case of higher-order elements the transient part is a sum of functions similar to the transient term in Eq.(5.136). The time necessary for the transients to decay is longer than $3T_{max}$ where $T_{max}$ is the greatest time constant. Since chemical engineering processes may have time constants of the order of several hours, the experimental determination of the frequency function by sinusoidal driving is rather expensive and is rarely applied.

### 5.3.6 Simple linear systems

*Table 5.5* contains the differential equations and the transfer functions of the most important systems. Some of them, such as *P*, *I*, *D*, *PI* and *PID* elements, have been treated already in Section 5.2 and are not repeated here.

*Figure 5.45* represents the impulse, step and ramp responses of these systems.

*Figure 5.46* shows the frequency functions represented as Nyquist and Bode plots.

*First-order element*

A first-order element is described by a first-order ordinary differential equation.

$$T\frac{dy}{dt}+y=Ax \quad (5.137)$$

Its transfer function was derived in Example 5.3.4 as Eq.(5.68)

$$G(s)=\frac{A_S}{Ts+1} \quad (5.138)$$

The unit impulse response [Example 5.3.5; Eq.(5.72)]:

$$y(t)=\frac{A_S}{T}e^{-\frac{t}{T}}$$

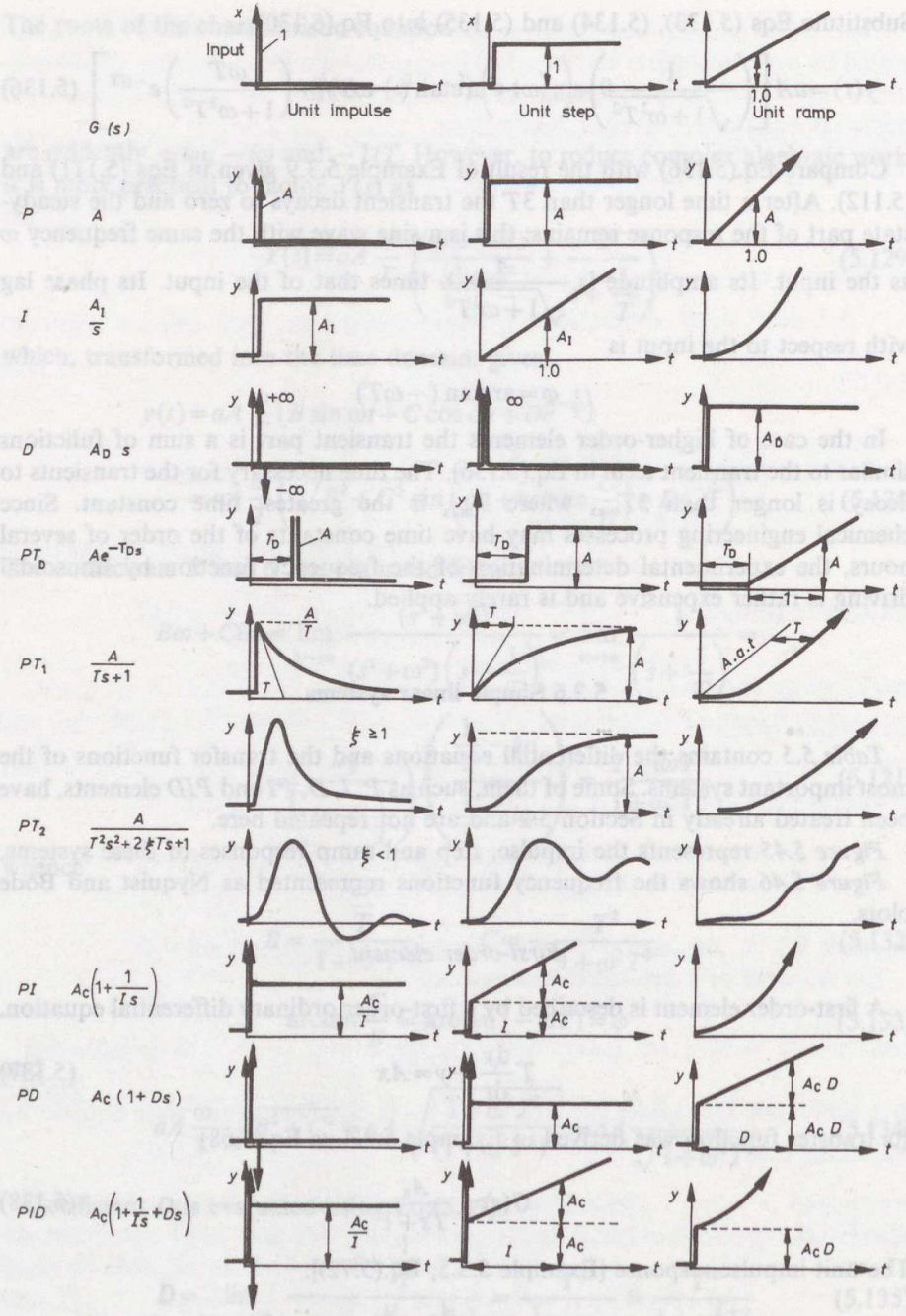



Fig. 5.45. Response functions of linear systems

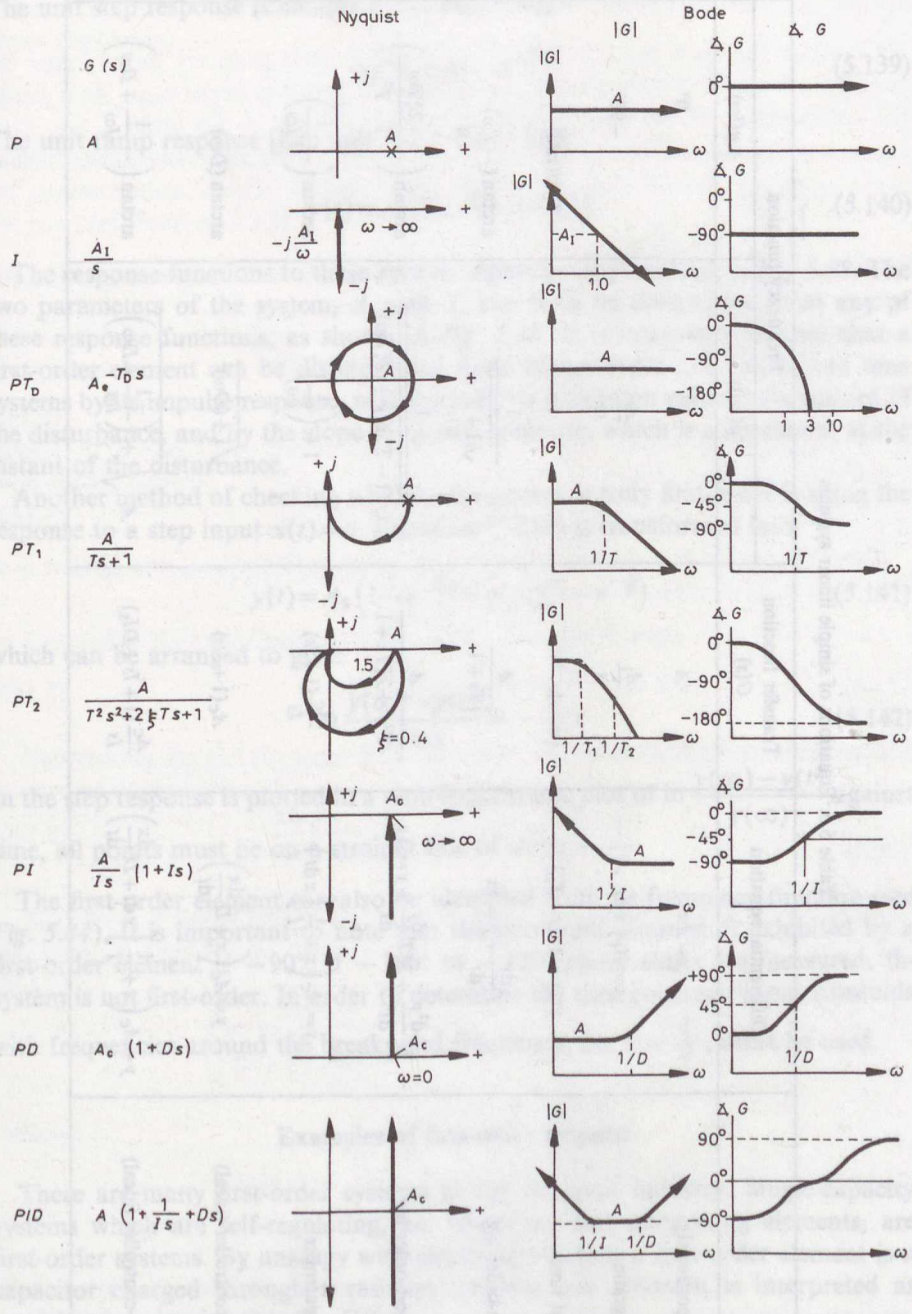



Fig. 5.46. Frequency functions of linear systems

Table 5.5. Equations of simple linear systems

| | Differential equation | Transfer function $G(s)$ | Frequency function | |
|---|---|---|---|---|
| | | | $\lvert g(i\omega)\rvert$ | $\measuredangle g(i\omega)$ |
| Proportional | $y = Ax$ | $A$ | $1$ | $0°$ |
| Integrating | $v = A_I \int x\,dt$ | $\frac{A_I}{s}$ | $\frac{A_I}{\omega}$ | $-90°$ |
| Dead time | $y(t) = Ax(t - T_D)$ | $Ae^{-T_D s}$ | $1$ | $-T_D\omega$(rad) |
| First-order lag | $T\frac{dy}{dt} + y = Ax$ | $\frac{A}{Ts+1}$ | $\frac{1}{\sqrt{1+T^2\omega^2}}$ | $\arctan(-T\omega)$ |
| Second-order lag | $T^2\frac{d^2y}{dt^2} + 2\xi T\frac{dy}{dt} + y = Ax$ | $\frac{A}{T^2s^2 + 2\xi Ts + 1}$ | $\frac{1}{\sqrt{(1-T^2\omega^2)^2 + (2\xi T\omega)^2}}$ | $\arctan\left(-\frac{2\xi T\omega}{1-T^2\omega^2}\right)$ |
| *PI*-controller | $y = A_C\left(x + \frac{1}{I}\int x\,dt\right)$ | $\frac{A_C}{Is}(1+Is)$ | $1 + \left(\frac{1}{I\omega}\right)^2$ | $\arctan\left(-\frac{1}{I\omega}\right)$ |
| *PD*-controller (ideal) | $y = A_C\left(x + D\frac{dx}{dt}\right)$ | $A_C(1+Ds)$ | $\sqrt{1+(D\omega)^2}$ | $\arctan(D\omega)$ |
| *PID*-controller (ideal) | $y = A_C\left(x + \frac{1}{I}\int x\,dt + D\frac{dx}{dt}\right)$ | $\frac{A_C}{Is}(1 + Is + DI_s)$ | $A_C\sqrt{1+\left(\frac{-1}{I\omega} + D\omega\right)^2}$ | $\arctan\left(-\frac{1}{I\omega} + D\omega\right)$ |

The unit step response [Example 5.3.6; Eq.(5.75)]:

$$y(t)=A_S\left(1-e^{-\frac{t}{T}}\right) \tag{5.139}$$

The unit ramp response [Example 5.3.7; Eq.(5.96)]:

$$y(t)=A_S\left[Te^{-\frac{t}{T}}+(t-T)\right] \tag{5.140}$$

The response functions to these specific inputs are represented in *Fig. 5.38*. The two parameters of the system, $A_S$ and $T$, can both be determined from any of these response functions, as shown in *Fig. 5.38*. It is important to note that a first-order element can be distinguished from higher-order and from dead time systems by its impulse response, which attains its maximum value at the instant of the disturbance, and by the slope of its step response, which is a maximum at the instant of the disturbance.

Another method of checking whether the system is truly first-order is using the response to a step input $x(t)=a$. Equation (5.139) is transformed into:

$$y(t)=A_S\left(1-e^{-\frac{t}{T}}\right)=y(\infty)\left(1-e^{-\frac{t}{T}}\right) \tag{5.141}$$

which can be arranged to give:

$$\ln\frac{y(\infty)-y(t)}{y(\infty)}=-\frac{1}{T}t \tag{5.142}$$

In the step response is plotted in a semi-logarithmic plot of $\ln\dfrac{y(\infty)-y(t)}{y(\infty)}$ against time, all points must lie on a straight line of slope $-\dfrac{1}{T}$.

The first-order element can also be identified from its frequency function (see *Fig. 5.44*). It is important to note that the minimum phase shift exhibited by a first-order element is $-90°$. If $-100°$ or $-120°$ phase shifts are measured, the system is not first-order. In order to determine the time constant, input sinusoids with frequencies around the breakpoint frequency, i.e. $\omega=\dfrac{1}{T}$, must be used.

### Examples of first-order systems

There are many first-order systems in the chemical industry. Single-capacity systems which are self-regulating, i.e. which are not integrating elements, are first-order systems. By analogy with electrical practice, a first-order element is a capacitor charged through a resistor, and its time constant is interpreted as capacitance × resistance, $T=CR$.

*Continuous stirred tank reactor (CSTR)*

Consider the tank reactor shown in *Fig. 5.47*. It is perfectly mixed, thus the concentration in it is uniform, independent of the position in the tank. As a result, the composition of the output stream is equal to that in the tank: $c_0 = c$. First order, irreversible, isothermal reaction takes place in the tank with reaction rate $k$:

$$A \rightarrow P$$

The volume holdup in the tank, $V$, is constant. There is no density change, thus the mass balance may be written in terms of volume rates:

$$\text{input} = \text{output} + \text{accumulation}$$

$$W_i \, dt = W_0 \, dt + dV \tag{5.143}$$

However, $V = \text{const}$, thus

$$W_i = W_0 = W \tag{5.144}$$

Suppose that $W$ is constant in time. The input variable is $c_i$, the output is $c$. The component balance is:

$$\text{input} = \text{output} + \text{consumed} + \text{accumulated}$$

$$W c_i \, dt = W c \, dt + V k c \, dt + V \, dc \tag{5.145}$$

The component balance can be arranged to give:

$$\left(\frac{W}{W + Vk}\right) c_i = c + \left(\frac{V}{W + Vk}\right) \frac{dc}{dt} \tag{5.146}$$

This is the differential equation of a first-order element, with

$$A_S = \frac{W}{W + Vk}; \qquad T = \frac{V}{W + Vk} \tag{5.147}$$

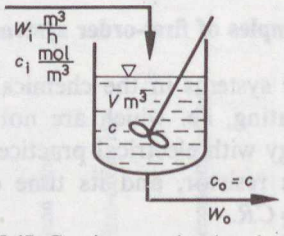



Fig. 5.47. Continuous stirred tank reactor

where $V$ is capacitance and $1/(W+Vk)$ is resistance. In the steady-state $\frac{dc}{dt}=0$, and from Eq.(5.146):

$$c_A = \frac{W}{W+Vk} c_{A_i} \tag{5.148}$$

If the reaction in the *CSTR* is not first-order, but the reaction rate is a function of concentration, $r=f(c_A)$, then linearization and perturbation variables are to be used (cf. Section 5.3.3):

$$\hat{r} = \hat{c}\left(\frac{\partial r}{\partial c}\right)_{\bar{c}} \tag{5.149}$$

where $\bar{c}$ is the concentration at the steady state. The steady-state component balance is:

$$W\bar{c}_i = W\bar{c} + V\bar{r}\,, \tag{5.150}$$

where $\bar{c}_i$, $\bar{c}$ and $\bar{r}$ are the steady-state values. The component balance in the non-steady-state is:

$$\text{input} = \text{output} + \text{consumed} + \text{accumulated} \tag{5.151}$$

$$Wc_i\,dt = Wc\,dt + Vr\,dt + V\,dc$$

Rearrange Eq.(5.151):

$$Wc_i = Wc + Vr + V\frac{dc}{dt} \tag{5.152}$$

Subtracting Eq.(5.150) from Eq.(5.152) we obtain the differential equation in perturbation variables ($\hat{c}_i = c_i - \bar{c}_i$, $\hat{r} = r - \bar{r}$):

$$W\hat{c}_i = W\hat{c} + V\hat{r} + V\frac{d\hat{c}}{dt} \tag{5.153}$$

Substitute $\hat{r}$ from Eq.(5.149) and arrange into the form of the differential equation of a first-order element:

$$\left(\frac{W}{W+V\left(\frac{\partial r}{\partial c}\right)_{\bar{c}}}\right)\hat{c}_i = \hat{c} + \left(\frac{V}{W+V\left(\frac{\partial r}{\partial c}\right)_{\bar{c}}}\right)\frac{d\hat{c}}{dt} \tag{5.154}$$

whence

$$A_S = \frac{W}{W+V\left(\frac{\partial r}{\partial c}\right)_{\bar{c}}}; \qquad T = \frac{V}{W+V\left(\frac{\partial r}{\partial c}\right)_{\bar{c}}}$$

Parameters $A_S$ and $T$ both depend on the operation point.

Perfectly mixed tanks are generally first-order elements. It is conceivable that a perfectly mixed blending tank, where no reaction occurs, where the output is the

concentration of the blend, and where the input variables are the feed rate and the feed concentration, is a first-order lag. A perfectly mixed continuous heater (or cooler) with continuous inflow and outflow is a first-order lag, too: the output is the temperature of the outflow, the input variables are the temperature and the flow rate of the inflow, as well as the heat input. Other examples of first-order elements are:

— thermometer bulb,
— liquid tank discharging through a valve,
— gas tank.

Here the differential equation of the liquid tank will be derived. The derivation for the case of a gas tank is quite similar.

*Self-regulating liquid tank*

The liquid level system presented in Section 5.2.3 and *Fig. 5.27* is an integrating process, because its outflow is fixed by a metering pump: the outflow is independent of the liquid level in the tank.

The liquid tank in *Fig. 5.48* discharges through a valve, which represents a resistance. The flow through the valve depends on the valve opening, which is included in the valve coefficient $k_v$ ($k_v$ changes with the valve opening), and on the pressure drop $(p_1 - p_2)$ across the valve:

$$W_c = k_v \sqrt{\frac{p_1 - p_2}{\varrho}} \tag{5.155}$$

The dependence of the outflow rate on the liquid level is the strongest when the whole pressure drop $(p_1 - p_2)$ is covered by the liquid head in the tank:

$$p_1 - p_2 = \varrho g H . \tag{5.156}$$

Using Eq.(5.156), we obtain the outflow as a (nonlinear) function of the level:

$$W_0 = k_v \sqrt{\frac{\varrho g H}{\varrho}} = k \sqrt{H} \tag{5.157}$$

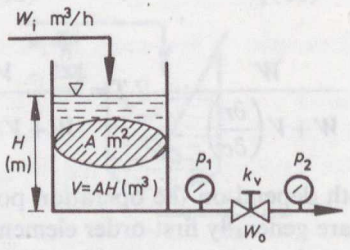



Fig. 5.48. Self-regulating liquid tank

Linearization around the steady state ($\bar{W}_0$, $\bar{H}$) is needed:

$$W_0 = \bar{W}_0 + \hat{H}\left(\frac{\partial W_0}{\partial H}\right)_{\bar{H}} . \tag{5.158}$$

The steady-state mass balance of the tank is:

$$\bar{W}_i = \bar{W}_0 \qquad \text{and} \qquad \bar{H} = \left(\frac{\bar{W}_0}{k}\right)^2 \tag{5.159}$$

thus, in the steady state, the level is determined by the inflow. The mass balance in the non-steady-state is:

$$\text{input} = \text{output} + \text{accumulation}$$

$$W_i\,\mathrm{d}t = W_0\,\mathrm{d}t \; + A\,\mathrm{d}H . \tag{5.160}$$

Substituting Eqs (5.159) and (5.158) into Eq.(5.160) and rearranging, the differential equation of a first-order lag is obtained with perturbation variables:

$$\frac{1}{\left(\frac{\partial W_0}{\partial H}\right)_{\bar{H}}}\hat{W}_i = \hat{H} + \frac{1}{\left(\frac{\partial W_0}{\partial H}\right)_{\bar{H}}}\frac{\mathrm{d}\hat{H}}{\mathrm{d}t} \tag{5.161}$$

The time constant is

$$T = \frac{A}{\left(\frac{\partial W_0}{\partial H}\right)_{\bar{H}}}$$

where $A$ is capacitance, and

$\left(\frac{\partial H}{\partial W_0}\right)_{\bar{H}}$ is resistance.

*Note 1*: In the case of pressurized tanks, when the pressure loss through the restriction is covered partly by the liquid head in the tank and partly by the pressure above the liquid (or by a centrifugal pump), the total effective pressure drop can be expressed in terms of the liquid head, i.e. $p_1 - p_2 = B\varrho g H$ is substituted into Eq.(5.156), and the final result is the same.

*Note 2*: Observe that in this nonlinear process both the steady-state gain and the time constant depend on the operating point ($\bar{W}$, $\bar{H}$). Compare this with the result obtained for the nonlinear *CSTR*, Eq.(5.154). However, the dynamic gains for high frequencies are independent of the operating point [cf. Eq.(5.111) and subsequent equations]:

$$|G(i\omega)| = \frac{A_S}{\sqrt{1+T^2\omega^2}}$$

for $\omega \to \infty$ $|G(i\omega)| \dfrac{A_S}{T\omega}$

For the *CSTR* $|G| \rightarrow \frac{V}{W}\frac{1}{\omega}$

For the liquid level tank $|G| \rightarrow \frac{1}{A}\frac{1}{\omega}$

This phenomenon is frequently encountered in chemical engineering processes.

### Second-order systems

A second-order system is modelled by a second-order differential equation:

$$T^2\frac{d^2y}{dt^2}+2\xi T\frac{dy}{dt}+y=Ax \tag{5.162}$$

where $y$ is output, $x$ is input, $T$ is time constant, $A$ is process gain, and $\xi$ is damping factor.

Second-order systems can be classified into three categories according to the value of the damping factor.

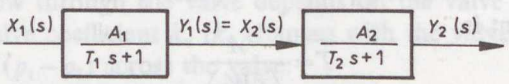



Fig. 5.49. Two first-order lags in series

*1. Overdamped system*

In this case the damping factor is greater than unity. This happens whenever two first-order systems are in series and their time constants are not equal. For example: two *CSTR*s in series, a thermometer in a heat exchanger, buffer tank and *CSTR*, and so on. *Figure 5.49* represents such a case.

The transfer function is:

$$\frac{Y_2(s)}{X_1(s)}=\left(\frac{A_1}{T_1s+1}\right)\left(\frac{A_2}{T_2s+1}\right)=\frac{A_1A_2}{T_1T_2s^2+(T_1+T_2)s+1}=\frac{A}{T^2s^2+2\xi Ts+1} \tag{5.163}$$

Comparing Eq.(5.163) with Eq.(5.162), the parameters are:

$$A=A_1A_2$$

$$T=\sqrt{T_1T_2} \tag{5.164}$$

$$\xi=\frac{1}{\sqrt{T_1T_2}}\left(\frac{T_1+T_2}{2}\right) \tag{5.165}$$

If $T_1 \neq T_2$ then $\xi < 1$.

The responses to typical unit inputs are *(Fig. 5.45)*:

unit impulse $$y(t)=\frac{A}{T_1-T_2}\left(e^{-\frac{t}{T_1}}-e^{-\frac{t}{T_2}}\right) \tag{5.166}$$

The slope of the impulse response is a maximum at the instant of the disturbance $A/(T_1T_2)$. The impulse response has a maximum at time $t_{max}=\sqrt{T_1T_2}$.

unit step $$y(t)=A\left[1-\frac{1}{T_1-T_2}\left(T_1e^{-\frac{t}{T_1}}-T_2e^{-\frac{t}{T_2}}\right)\right] \tag{5.167}$$

$$\left(\frac{dy}{dt}\right)_{t=0}=0$$

unit ramp $$y(t)=A\left[\frac{1}{T_1-T_2}\left(T_1^2e^{-\frac{t}{T_1}}-T_2^2e^{-\frac{t}{T_2}}\right)+\left(t-(T_1+T_2)\right)\right]$$

$$\text{for}\quad t>3T_1\quad\text{and}\quad t>3T_2,\quad y(t)=A\left(t-(T_1+T_2)\right) \tag{5.168}$$

*2. Critically damped system*

This is the same as Case 1, presented in *Fig. 5.49*, with the transfer function in Eq.(5.163), but $T_1=T_2$. Thus the model parameters are

$$T=\sqrt{T_1T_2}=T_1=T_2 \tag{5.169}$$

$$\xi=\frac{1}{\sqrt{T_1T_2}}\left(\frac{T_1+T_2}{2}\right)=1 \tag{5.170}$$

The system is called critically damped, because it does not exhibit oscillations in response to the typical inputs but attains the steady state faster than over-damped systems with the same $T$ parameter. The responses to typical inputs are *(Fig. 5.45)*:

unit impulse $$y(t)=\frac{A}{T^2}te^{-\frac{t}{T}} \tag{5.171}$$

$$\text{with}\quad\left(\frac{dy}{dt}\right)_0=\frac{A}{T^2}\quad\text{and}\quad t_{max}=t$$

unit step $$y(t)=A\left[1-e^{-\frac{t}{T}}\left(1+\frac{t}{T}\right)\right] \tag{5.172}$$

with $$\left(\frac{dy}{dt}\right)_0=0$$

unit ramp $$y(t)=A\left[e^{-\frac{t}{T}}(t+2T)+(t-2T)\right] \tag{5.173}$$

for $t>3T$ $\quad y(t)=A(t-2T)$

*3. Underdamped system*

The damping factor is less than unity, but always greater than zero. Second-order underdamped systems cannot be constructed from two first-order elements in series; they are second-order elements in nature, which is due to inertial effects and internal interactions. Examples of second-order underdamped systems are: a *U*-tube manometer; a mass suspended from a spring, etc.

The underdamped character manifests itself in the impulse or step response by overshooting the final value and reaching it after damped oscillations. The frequency of the damped oscillations is

$$\omega=\frac{1}{T}\sqrt{1-\xi^2} \tag{5.174}$$

whence the meaning of parameter $T$: $1/T$ would be the oscillation frequency if the system were undamped, i.e. for $\xi=0$.

The responses to typical inputs are presented in *Fig. 5.45*.

All second-order systems (independent of the value of the damping factor) can be distinguished from first-order systems by examining the initial slopes of the impulse and step responses. The initial slope of the impulse response of a first-order system is infinite, while that of a second-order system is a maximum (for higher-order systems it is zero). The initial slope of the step response of a first-order system is a maximum, that of second- (or higher-) order systems is zero.

The transfer function of second-order systems is obtained by the Laplace transformation of the differential equation, Eq.(5.162):

$$G(s)=\frac{Y(s)}{X(s)}=\frac{A}{T^2s^2\xi Ts+1}=\frac{A/T^2}{s^2\dfrac{2\xi}{T}s+\dfrac{1}{T^2}} \tag{5.175}$$

The magnitude and phase angle of the frequency function are obtained by substituting $s=i\omega$ into Eq.(5.175) and following the procedure described in Section 5.3.5:

$$|G(i\omega)|=\frac{A}{\sqrt{(1-\omega^2T^2)^2+(2\xi\omega T)^2}} \tag{5.176}$$

$$\sphericalangle G(i\omega) = \arctan\left(-\frac{2\xi\omega T}{1-\omega^2 T^2}\right) \tag{5.177}$$

Of course the response to a sine input with frequency $\omega$ is, in the case of underdamped systems, another sine wave with frequency $\omega$, but the frequency function shown in *Fig. 5.46* reveals its underdamped character. There exists a resonance frequency $\omega = \frac{1}{T}$, where the dynamic gain of underdamped systems is several times the static gain. Important values of the frequency function are:

$$\omega \to 0 : |G| \to A \qquad \sphericalangle G \to 0° \tag{5.178}$$

$$T = 1 : |G| = \frac{A}{2\xi} \qquad \sphericalangle G = -90° \tag{5.179}$$

$$\omega \to \infty : |G| = \frac{A}{\omega^2 T^2} \qquad \sphericalangle G \to -180° \tag{5.180}$$

$$\omega \to \infty : \frac{\mathrm{d}\log|G|}{\mathrm{d}\log|\omega|} \to -2, \tag{5.181}$$

which represent the slope of the Bode plot.

Thus second-order systems may readily be identified from their frequency functions, which should be determined around the frequency $\omega = 1/T$.

It is important to note that the minimum phase of second-order elements is $2 \cdot (-90°) = -180°$.

### Higher-order systems

Elements of order $n$ are described by $n$th order ordinary differential equations, such as Eq.(5.57); their transfer functions are the reciprocals of $n$th order polynomials in $s$.

They usually consist of first- (and second-) order elements in series, in other words, they are multicapacity systems. The capacities of a multicapacity system can be interacting or noninteracting (isolated). In noninteracting multicapacity systems the signals are transferred in one direction only; the conditions in the $m$th capacity do not depend on the conditions in the $(m+1)$th capacity. A cascade of *CSTR*s is an example of a noninteracting multicapacity system. In such systems the time constants of the individual capacities can be calculated independently from their mathematical models (or $T_i = C_i R_i$).

In interacting multicapacity systems the signals are transferred in both directions, as in a series of interconnected gas tanks, where a pressure change at both ends of the system would cause the pressure to change in all tanks *(Fig. 5.50)*. Other examples of interacting multicapacity systems are: multitray distillation

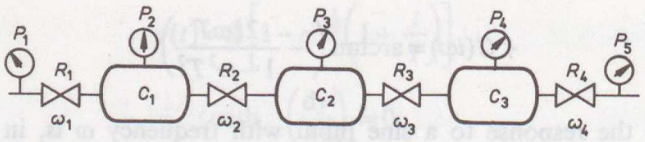



Fig. 5.50. Interacting gas tanks

columns, countercurrent extractors and many other chemical engineering processes. Consider the series of three interacting gas tanks presented in *Fig. 5.50.*

The input variables are $p_1$ and $p_5$. In the steady state

$$\bar{w}_1 = \bar{w}_2 = \bar{w}_3 = \bar{w}_4 = \frac{\bar{p}_1 - \bar{p}_2}{R_1} = \frac{\bar{p}_2 - \bar{p}_3}{R_2} = \frac{\bar{p}_3 - \bar{p}_4}{R_3} = \frac{\bar{p}_4 - \bar{p}_5}{R_4}$$

In the non-steady-state

$$\text{input} = \text{output} + \text{accumulation}$$

$$\frac{\hat{p}_1 - \hat{p}_2}{R_1} = \frac{\hat{p}_2 - \hat{p}_3}{R_2} + C_1 \frac{\mathrm{d}\hat{p}_2}{\mathrm{d}t}$$

$$\frac{\hat{p}_2 - \hat{p}_3}{R_2} = \frac{\hat{p}_3 - \hat{p}_4}{R_3} + C_2 \frac{\mathrm{d}\hat{p}_3}{\mathrm{d}t}$$

$$\frac{\hat{p}_3 - \hat{p}_4}{R_3} = \frac{\hat{p}_4 - \hat{p}_5}{R_4} + C_3 \frac{\mathrm{d}\hat{p}_4}{\mathrm{d}t}$$

This third-order system is described by a mathematical model consisting of three first-order differential equations which, however, are not independent of each other. If interaction is ignored, the equation for the first tank can be arranged as:

$$\left(\frac{1}{\frac{1}{R_1} + \frac{1}{R_2}}\right)\hat{p}_1 + \left(\frac{1}{\frac{1}{R_2} + \frac{1}{R_3}}\right)\hat{p}_3 = \hat{p}_2 + \left(\frac{1}{\frac{1}{R_1} + \frac{1}{R_2}}\right) \cdot \frac{\mathrm{d}\hat{p}_2}{\mathrm{d}t}$$

whence the time constant for the first tank can be expressed as:

$$T_1 = \frac{C_1}{\frac{1}{R_1} + \frac{1}{R_2}}; \quad \text{and} \quad T_i = \frac{C_i}{\frac{1}{R_i} + \frac{1}{R_{i+1}}}$$

If interaction is not ignored, the mathematical model can be solved to obtain three time constants $T_{i,\,\mathrm{eff}} \neq T_i$. The effect of interaction is to increase the greatest individual time constant and at the same time to reduce the smaller ones. Interaction will make the process more sluggish (it increases the sum of the effective time constants) but easier to control.

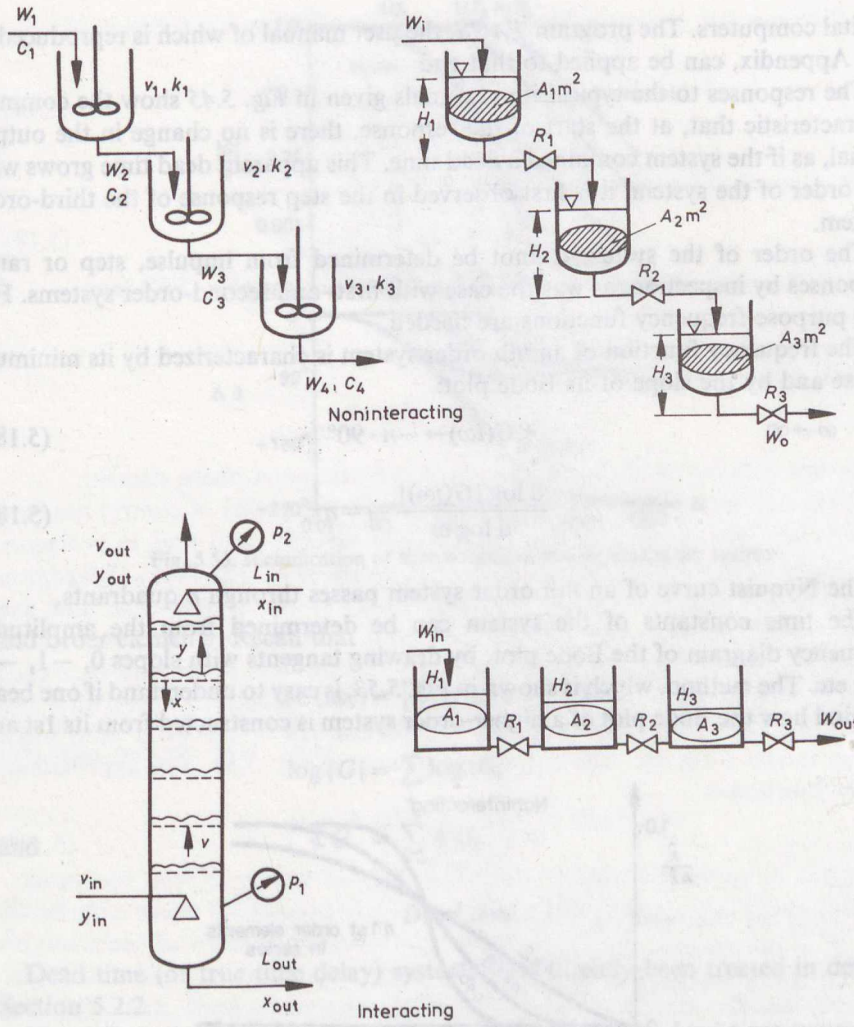



Fig. 5.51. Examples of interacting and noninteracting multicapacity systems

Some examples of interacting and noninteracting systems are shown in *Fig. 5.51*. The step responses are compared in *Fig. 5.52*.

Interacting or not, *n*-capacity systems are modelled by *n*th order differential equations and *n*th order transfer functions.

The time responses of higher-order systems may be calculated by inverse transformation of the Laplace transforms, but this needs considerable algebraic work. Typical responses in analytical form up to fourth-order may be found in handbooks. For higher-order systems numerical integration is easier, using analogue or

digital computers. The program *TACS*, the user manual of which is reproduced in the Appendix, can be applied to that end.

The responses to the typical input signals given in *Fig. 5.45* show the common characteristic that, at the start of the response, there is no change in the output signal, as if the system contained a dead time. This apparent dead time grows with the order of the system; it is first observed in the step response of the third-order system.

The order of the system cannot be determined from impulse, step or ramp responses by inspection, as was the case with first- and second-order systems. For this purpose frequency functions are needed.

The frequency function of an $n$th order system is characterized by its minimum phase and by the slope of its Bode plot:

for $\omega \to \infty$ $$\measuredangle G(i\omega) \to -n \cdot 90° \tag{5.182}$$

$$\frac{\mathrm{d} \log |G(i\omega)|}{\mathrm{d} \log \omega} \to -n \tag{5.183}$$

The Nyquist curve of an $n$th order system passes through $n$ quadrants.

The time constants of the system can be determined from the amplitude-frequency diagram of the Bode plot, by drawing tangents with slopes 0, $-1$, $-2$, $-3$, etc. The method, which is shown in *Fig. 5.53*, is easy to understand if one bears in mind how the Bode plot of a higher-order system is constructed from its 1st and

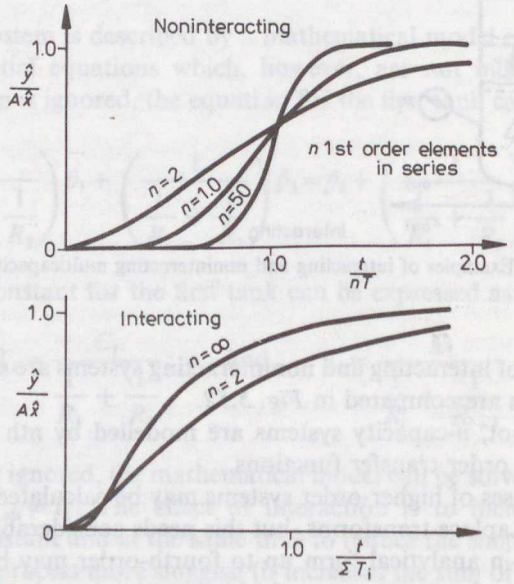



Fig. 5.52. Step responses of interacting and of noninteracting systems

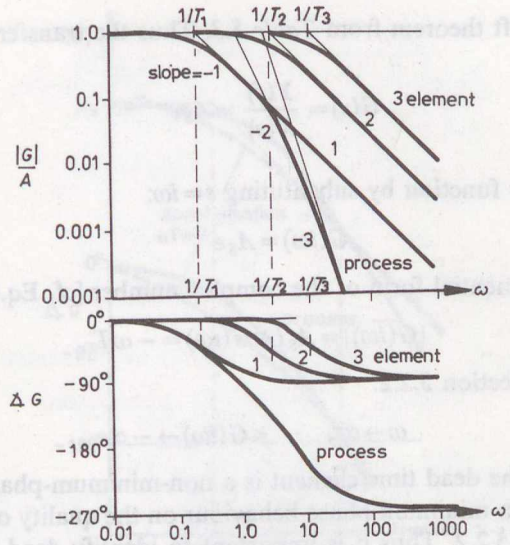



Fig. 5.53. Identification of time constants of a multicapacity system

2nd order elements. Recall that

$$|G(i\omega)| = \prod_i |G(i\omega)|\,, \qquad \text{thus}$$

$$\log|G| = \sum_i \log|G_i|\,,$$

and

$$\measuredangle G = \sum \measuredangle G_i\,.$$

*Dead time*

Dead time (or true time delay) systems have already been treated in detail in Section 5.2.2.

The differential equation method cannot be applied to dead time systems because they correspond to infinite-order systems.

The transfer function and the frequency function methods are applicable, however. The transfer function is obtained directly from the equation

$$y[t] = A_S x[t - T_D] \tag{5.184}$$

where $y[t]$ is output at moment $t$,

$A_S$ is steady-state gain of the process, and

$x[t-T_D]$ is input at moment $(t-T_D)$.

The Laplace transform is

$$Y(s) = A_S X \mathcal{L}[x(t-T_D)] = A_S X(s) e^{-sT_D} \tag{5.185}$$

using the time shift theorem from *Table 5.3*. Thus the transfer function is:

$$G(s)=\frac{Y(s)}{X(s)}=A_S e^{-sT_D} \tag{5.186}$$

and the frequency function by substituting $s=i\omega$:

$$G(i\omega)=A_S e^{-i\omega T_D} \tag{5.187}$$

Recall the exponential form of the complex number [cf. Eq.(5.9)] thus

$$|G(i\omega)|=A_S \quad i\not\propto G(i\omega)=-\omega T_D$$

as was found in Section 5.2.2.

When $\omega\to\infty$, $\quad \not\propto G(i\omega)\to-\infty$

This means that the dead time element is a non-minimum-phase element.

The effect of non-minimum-phase behaviour on the quality of control has been shown in Section 5.2.2. Thus it is important to identify dead time in processes.

Pure dead time is easily identified from typical responses *(Fig. 5.45)*. If dead time is in series with capacities (single or multicapacity systems), pure dead time and apparent dead time in a step response are hard to distinguish from each other.

For purposes of controller setting, capacity systems with or without pure dead time are generally approximated by a dead time–single capacity or a dead time–two capacity model. The determination of the parameters of the dead time–single capacity model from the step response is shown in *Fig. 5.54*. The approximate transfer function is

$$G(s)=\frac{A_S e^{-T_D s}}{T_1 s+1} \tag{5.188}$$

where $T_D$ is the effective dead time,
$T_1$ is the effective time constant, and
$A_S$ is the steady-state process gain.

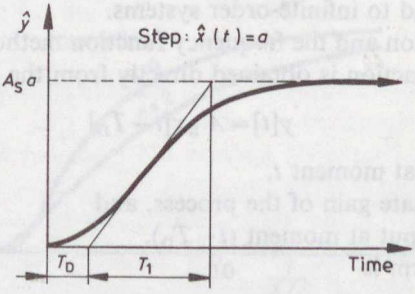



Fig. 5.54. Approximation of multicapacity step response

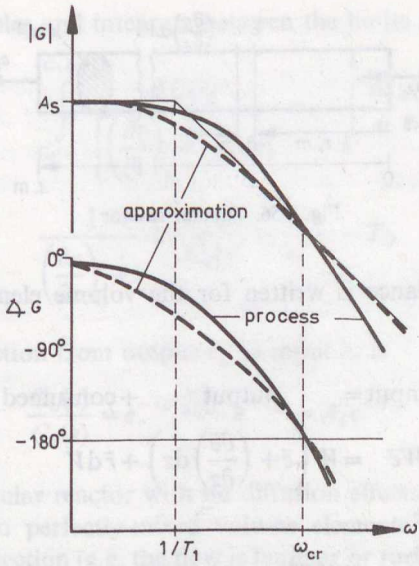



Fig. 5.55. Approximation of multicapacity process by first-order lag with dead time, Bode plot

The frequency function of the approximate model should fit that of the original process in the vicinity of the critical frequency, since it is used to determine controller settings *(Fig. 5.55)*.

*Distributed parameter systems*

So far we have discussed lumped parameter systems consisting of completely mixed units where the process variable is a function of time only. These processes are described by ordinary differential equations.

In distributed parameter systems the process variable is a function both of time and of position. Such systems are described by partial differential equations. The mathematical treatment of such systems is a little different. Applying Laplace transformation, ordinary differential equations are obtained, the solution of which gives the transfer functions directly.

The procedure is shown using a tubular reactor *(Fig. 5.56)* as an example. This is a unidimensional problem with axial symmetry. We suppose the following conditions: plug flow, no density change, no diffusion effects, no thermal effects (isothermal reaction).

The reaction rate $r = -\frac{1}{V}\frac{dc}{dt}$ is dependent on concentration, but an average value of $\left(\frac{\partial r}{\partial c}\right)_{t=0}$ can be used over the length of the reactor.

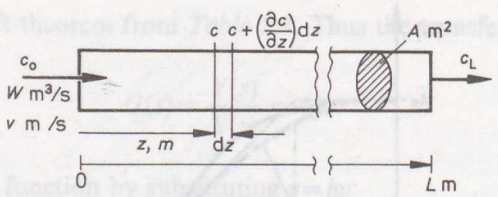



Fig. 5.56. Tubular reactor

The component balance is written for the volume element $dV = A\,dz$ of the reactor of lenght $L$.

In the steady state

$$\text{input} = \text{output} + \text{consumed}$$

$$W\bar{c} = W\left(\bar{c} + \left(\frac{\partial c}{\partial z}\right)dz\right) + \bar{r}\,dV \tag{5.189}$$

In the non-steady state

$$\text{input} = \text{output} + \text{consumed} + \text{accumulated}$$

$$Wc = W\left(c + \left(\frac{\partial c}{\partial z}\right)dz\right) + r\,dV + V\frac{\partial c}{\partial t} \tag{5.190}$$

The perturbation variables are defined as:

$$\hat{c}(z, t) = c(z, t) - \bar{c}(z, 0) \tag{5.191}$$

$$\hat{r}(z, t) = r(z, t) - \bar{r}(z, 0) \tag{5.192}$$

$\hat{r}(z, t)$ is function of $\hat{c}$

$$\hat{r}(z, t) = \left(\frac{\partial r}{\partial c}\right)_0 \hat{c}(z, t)\,. \tag{5.193}$$

By subtracting Eq.(5.189) from Eq.(5.190) and substituting Eq.(5.193) and $dV = A\,dz$, the linearized partial differential equation in perturbation variables is obtained:

$$0 = W\frac{\partial c}{\partial z}dz + A\left(\frac{\partial r}{\partial c}\right)_0 \hat{c}\,dz + A\frac{\partial \hat{c}}{\partial t}dz \tag{5.194}$$

giving:

$$-v\frac{\partial \hat{c}}{\partial z} = \left(\frac{\partial r}{\partial c}\right)_0 \hat{c} + \frac{\partial \hat{c}}{\partial t} \tag{5.195}$$

Apply Laplace transformation:

$$-v\frac{dC(s)}{dz} = \left(\frac{\partial r}{\partial c}\right)_0 C(s) + sC(s) \tag{5.196}$$

Separate the variables and integrate between the limits $z=0$ and $z=L$:

$$\int_{C_0(s)}^{C_L(s)} \frac{dC(s)}{\left[\left(\frac{\partial r}{\partial c}\right)_0 + s\right] C(s)} = -\int_0^L \frac{dz}{v} \tag{5.197}$$

$$\frac{1}{\left(\frac{\partial r}{\partial c}\right)_0 + s} \ln \frac{C_L(s)}{C_0(s)} = \frac{L}{v} = -T_D \tag{5.198}$$

then the transfer function from output $c_L$ to input $c_0$ is

$$\frac{C_L(s)}{C_0(s)} = e^{-T_D\left(\frac{\partial r}{\partial c}\right)_0} \cdot e^{-T_D s} = A_S e^{-T_D s} \tag{5.199}$$

Thus a plug flow tubular reactor with no diffusion effects corresponds to a pure time delay, i.e. $n \to \infty$ perfectly-mixed volume elements. If the velocity is not uniform in the cross-section (e.g. the flow is laminar or turbulent, diffusion and/or back-mixing exist), the tubular reactor corresponds to a multicapacity system, the order of which depends on the mixing conditions.

The transfer and frequency functions of distributed parameter systems may also be approximated by first-order dead time models [Eq.(5.188)], or second-order dead time models.

### 5.3.7 Control loop behaviour

#### Open loop transfer and frequency function

The open loop frequency function is important for checking the stability of the closed control loop. The method has been introduced in Section 5.2.1 and was applied in Section 5.2.2 for dead time process and in Section 5.2.3 for pure capacity or integrating process. It is based on the simplified Nyquist stability criterion [see Eq.(5.12)]:

$$L(i\omega) = G_C(i\omega) G_S(i\omega) = -1 \tag{5.200}$$

or

$$\measuredangle L(i\omega) = \measuredangle G_C(i\omega) + \measuredangle G_S(i\omega) = -180° \tag{5.201}$$

and

$$\begin{aligned} |L(i\omega)| = 1 &= |G_C(i\omega)|\,|G_S(i\omega)| = \\ &= A_C |g_C(i\omega)|\, A_S |g_S(i\omega)| = \\ &= K |g_C(i\omega)|\,|g_S(i\omega)| \end{aligned} \tag{5.202}$$

where $L(i\omega)$ is the open loop frequency function,
$\omega = \omega_{cr}$ is the critical frequency, if Eq.(5.201) is satisfied,
$K = A_C A_S$ is the loop gain,
$K = K_{max}$ and $A_C = A_{C,max}$ is the critical gain, if Eq.(5.202) is satisfied.

*Open loop frequency function with P-controller*

The *P*-controller (i.e. the ideal *P*-controller) has no phase lag at any frequency, thus it does not modify the phase-frequency function of the controlled process. This characteristic offers the possibility of determining the critical frequency of the process alone (called natural or ultimate frequency), i.e. the frequency at which

$$\measuredangle G_S(i\omega_0) = -180° \tag{5.203}$$

where $\omega_0$ is the natural frequency.

For processes with known parameters, the frequency function of the process is calculated, plotted on a Bode diagram, and the natural frequency is simply read off.

Continuous cycling of the signals in a control loop occurs if Eq.(5.200) is satisfied:

$$A_C A_S |g(i\omega_0)| = 1 \tag{5.204}$$

i.e.

$$A_{C,\max} = \frac{1}{A_S |g_S(i\omega_0)|} \quad \text{or} \quad K_{\max} = \frac{1}{|g_S i\omega_0|} \tag{5.205}$$

$A_{C,\max}$ is read simply from the Bode plot of the process.

For industrial processes with unknown parameters $A_{C,\max}$ and $\omega_0$ are determined experimentally by trial. The loop is closed through a *P*-controller and the gain of the controller is changed until continuous cycling of the process output (or the manipulated variable) occurs in response to a step input. The frequency of the output signal is $\omega_0$, the natural frequency of the process, the gain set at the *P*-controller is $A_{C,\max}$, the ultimate gain. This method has been described in detail in Section 5.2.6.

The method fails for some processes: even with the controller set to its highest proportional gain, continuous cycling cannot be achieved. The reason for this is explained below.

The pure capacity, or integrating process with the transfer function

$$G_S(s) = \frac{A_S}{s} \tag{5.206}$$

has a phase lag of $\measuredangle G_S = -90°$, independent of the frequency of the input signal. As has already been discussed in Section 5.2.3, this process can be controlled safely by a *P*-controller set to the largest possible gain, since the phase shift cannot cross $-180°$. The first-order lag process with the transfer function

$$G_S(s) = \frac{A_S}{Ts+1} \tag{5.207}$$

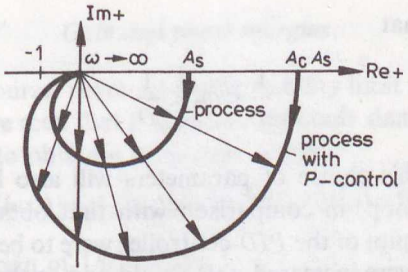



Fig. 5.57. Second-order process with *P*-controller is always stable

exhibits a phase lag (cf. *Table 5.3*)

$$\measuredangle G_S(i\omega) = \arctan(-\omega T)$$

For $\omega \to \infty$, the minimum phase $\measuredangle G_S \to -90°$, which is the same situation as was encountered with the pure capacity process.

A second-order lag process with the transfer function

$$G_S(s) = \frac{A_S}{T^2 s^2 + 2\xi T s + 1} \tag{5.208}$$

exhibits a phase lag (see Section 5.3.6)

$$\measuredangle G_S(i\omega) = \arctan\left(-\frac{2\xi T}{1-\omega^2 T^2}\right)$$

which has a minimum for $\omega \to \infty$: $\measuredangle G_S \to -2 \cdot 90°$. At the same time the dynamic gain of the process approaches zero:

$$\text{for} \quad \omega \to \infty \qquad |G_S| \to \frac{A_S}{\omega^2 T^2}$$

The situation is best visualized in the Nyquist diagram, *Fig. 5.57*. It is evident that a second-order process may be safely controlled by a *P*-controller set to the highest possible gain; oscillation will always be damped.

Processes with dead time or processes higher than second-order can be brought to continuous cycling with a *P*-controller.

### *Open loop frequency function with PID-controller*

Any combination of controllers other than proportional ones will substantially modify the open loop frequency function as compared with that of the process.

In *Fig. 5.58* the Bode plot of a process is shown, together with that of a *PID*-controller. It is evident that the critical frequency of the open loop will be closest to the natural frequency of the process if the *I* and *D* parameters of the

controller are set so that

$$\frac{1}{I} < \omega_0 < \frac{1}{D} \tag{5.209}$$

At the same time this choice of parameters will also have less effect on the dynamic gain of the loop, in comparison with that obtained with a simple *P*-controller. If the Bode plot of the *PID*-controller were to be moved towards higher frequencies, i.e. if $1/I$ were increased, this would add to the phase lag of the open loop, decreasing the critical frequency. This implies a slower control loop. Simultaneously, the frequency-dependent gain of the *I*-controller part would add to the dynamic gain of the process, eventually causing instability with the proportional gain set to a level that may safely be used with *P*-control only.

When the Bode plot of the *PID*-controller is moved towards smaller frequencies, i.e. when $1/D$ is diminished, the *D*-part of the controller is effective in diminishing the phase lag of the open loop around $\omega_0$, and thus it speeds up control. At the same time frequency dependent gain of the *D*-controller part adds to the process gain, and this can eventually cancel the effect of the phase lead. This depends, of course, on the slope of the amplitude-frequency function of the process.

The above considerations are also readily applicable to *PI*- and *PD*-controller combinations.

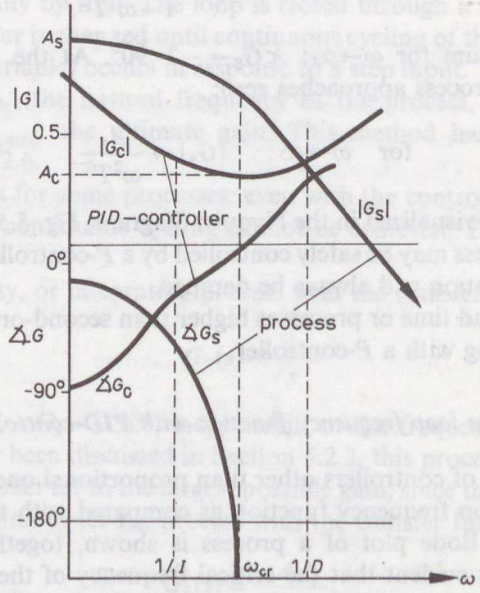



Fig. 5.58. Bode plot of process with *PID*-controller

*Gain and phase margins*

Controllers are, of course, never set to the stability limit defined by Eq.(5.200). In Section 5.2.1 we have seen that if quarter-amplitude damping is aimed for, the controller must be set to obtain a loop gain of 0.5:

$$|L(i\omega_{cr})| = A_S A_C |g_S(i\omega_{cr})|\,|g_C(i\omega_{cr})| = 0.5$$

This means a gain margin of $1/0.5 = 2.0$.

The definition of the gain margin is

$$\text{gain margin} = \frac{1}{|L(i\omega_{cr})|} \tag{5.210}$$

and it expresses the distance from the stability limit.

A similar value is needed for the phase lag of the open loop. This is found in the phase margin. Denote by $\varphi(L=1)$ the open loop phase lag at the frequency where $|L(i\omega)| = 1$, that is, at the gain crossover frequency. Then

$$\text{phase margin} = 180° - \varphi(L=1) \tag{5.211}$$

The phase margin is the phase shift which can be added to the phase shift of the open control loop to arrive at the stability limit of the closed control loop (at the gain crossover frequency).

The gain and the phase margins are easily read from the Bode and Nyquist plots of the open loop *(Fig. 5.59)*.

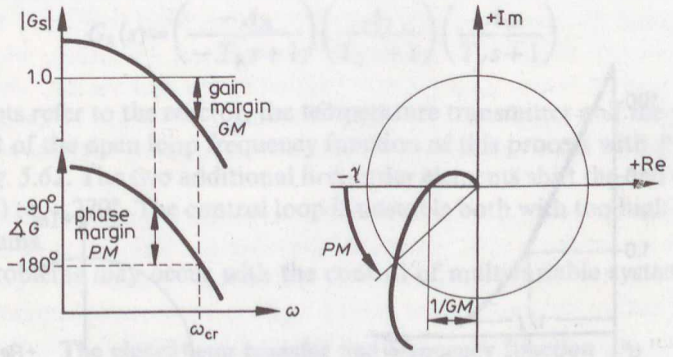



Fig. 5.59. Gain and phase margins

Both the gain and the phase margins account not only for the damping, but also for uncertainties in the knowledge of the process parameters. It is important to draw attention to the fact that most chemical engineering processes are nonlinear: their gain and also their time constants (including dead time) might be dependent

on the operating point. Gain variations in the range of 1:10 are also possible. The gain and phase margins (i.e. the controller parameters) must be set in consideration of these facts.

*The Nyquist stability criterion*

Hitherto the simplified Nyquist stability criterion formulated by Eq.(5.200) was used, which is readily applicable to monotonic open loop frequency functions. The rigorous Nyquist criterion is used to solve more complex problems.

The simplest example for which the rigorous Nyquist criterion has to be applied is the pure capacity process controlled by a *PI*-controller. The Bode and Nyquist plots of the open loop are shown in *Fig. 5.60.* For $\omega \to 0$, the loop phase lag is $-180°$ and the loop gain is infinite: there are two integrators in the loop. With increasing frequency the phase lag decreases towards $-90°$ and the loop gain decreases towards zero. The stability of this control loop can be checked by drawing the Nyquist plot. Following the plot from $\omega = 0$ towards $\omega = \infty$, the point $(-1, 0_{-j})$ is seen to be at the left, and thus the control loop is stable.

Conditional stability may occur in the case of chemical reactors in which exothermic reaction occurs. Some of these reactors are unstable without control. Continuous stirred tank reactors are first-order lag processes; at their unstable operating point both the steady-state gain and the time constant are negative. The transfer function is:

$$G(s) = \frac{-A_S}{-Ts+1} = \frac{A_S}{T_S - 1}$$

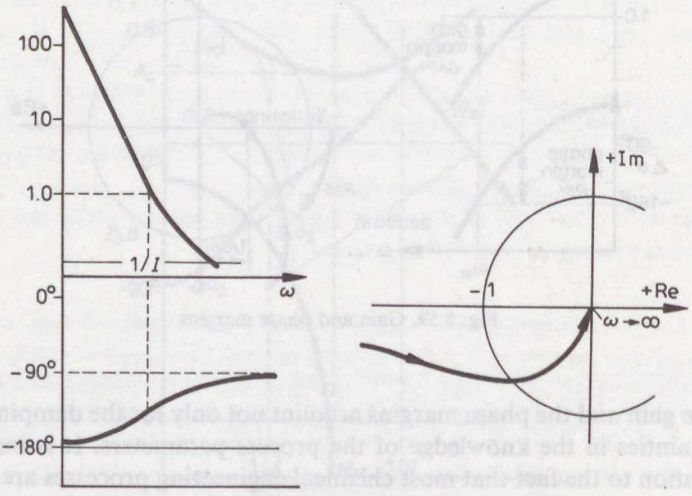



Fig. 5.60. *I* process controlled by *PI*-controller

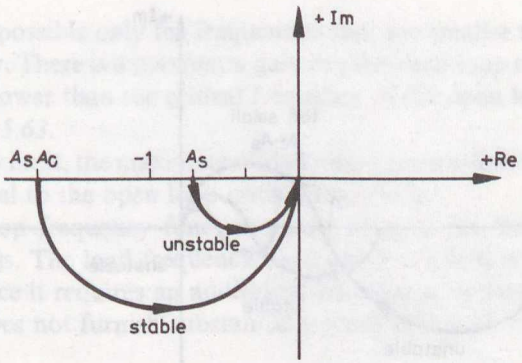



Fig. 5.61. Stabilizing an unstable reactor

This process was dealt with in Section 5.3.4, Example 5.3.8. The Nyquist plot of such a frequency function is shown in *Fig. 5.61*. It begins at point $(-A_S, 0)$ and ends at (0, 0). For a stable control loop $|A_S A_C| > 1$ is to be set, then the point $(-1, 0)$ is seen to be at the left. It is unusual that the controller gain must be increased to ensure stable operation.

When there are other dynamic elements in the control loop besides the reactor, the controller gain may have a high and a low limit for stable operation. Suppose that there are two first-order elements (e.g. the temperature sensor and the actuator of the control valve) both having smaller time constants than that of the reactor (of course, in absolute value). The transfer function of the process is:

$$G_S(s) = \left(\frac{-A_R}{-T_R s + 1}\right)\left(\frac{A_T}{T_T s + 1}\right)\left(\frac{A_V}{T_V s + 1}\right)$$

The subscripts refer to the reactor, the temperature transmitter and the valve. The Nyquist plot of the open loop frequency function of this process with *P*-control is shown in *Fig. 5.62*. The two additional first-order elements shift the end of the plot by $2 \cdot (-90°)$ to $-270°$. The control loop is unstable both with too high or too low controller gains.

Similar problems may occur with the control of multivariable systems.

### The closed loop transfer and frequency function

The open loop transfer and frequency functions are obtained by multiplying the respective functions of all dynamic elements in the control loop. The open loop functions are not to be confused with those of the closed loop.

The frequency function of a negative feedback control loop is given in Eq.(5.4). Since every control loop has at least two inputs the set point (or command variable) and one load variable, there are at least two closed loop frequency functions:

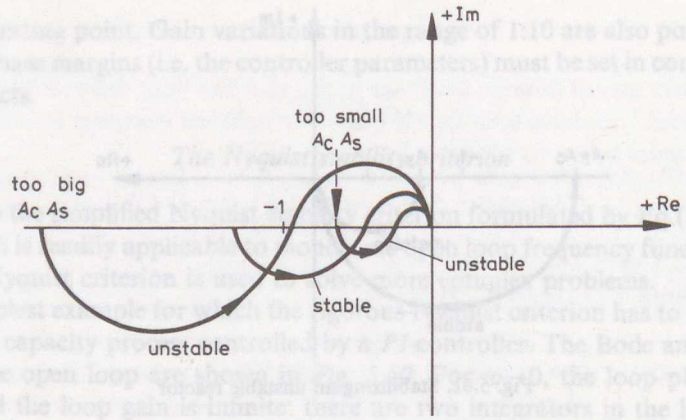



Fig. 5.62. Stability of control loop with unstable reactor and two small lags

for the set point:
$$W(i\omega)=\frac{G_S(i\omega)G_C(i\omega)}{1+G_S(i\omega)G_C(i\omega)} \tag{5.212}$$

for the load:
$$W_d(i\omega)=\frac{G_d(i\omega)}{1+G_S(i\omega)G_C(i\omega)} \tag{5.213}$$

The closed loop frequency functions are denoted by $W$ to distinguish them from the corresponding open loop functions. All closed loop frequency functions of the same control loop have the same denominator: $1+L(i\omega)$, where the frequency dependent vector $L(i\omega)$ is to be added to the unit vector, which has zero phase angle. The closed loop frequency functions can easily be calculated using complex algebra, if the frequency functions of the process and the controller are known.

The closed loop frequency functions can be determined experimentally by driving the closed loop by a sine signal (varying the frequency) at the set point or at the load input, respectively, and recording the control signal, but this method is rather tedious and is rarely applied to chemical engineering processes.

The closed loop frequency functions, of course, are characteristic of the behaviour of the control loop. Since we require that the control loop causes the control signal to follow the set point signal exactly, with no phase shift at all frequencies, the required frequency function is

$$|W(i\omega)|=1, \qquad \measuredangle W(i\omega)=0^\circ \qquad \text{for all } \omega.$$

In reality this is possible only for frequencies that are smaller than the open loop critical frequency. There is a maximum gain at the closed loop resonant frequency, which is a little lower than the critical frequency of the open loop. This situation is shown in *Fig. 5.63.*

At the stability limit, the maximum closed loop gain is infinite and the resonance frequency is equal to the open loop critical frequency.

The closed loop frequency functions may be used for the determination of controller settings. The load frequency function, Eq.(5.213) is less often used for this purpose, since it requires an additional knowledge of the frequency function $G_d(i\omega)$, but it does not furnish substantially more information.

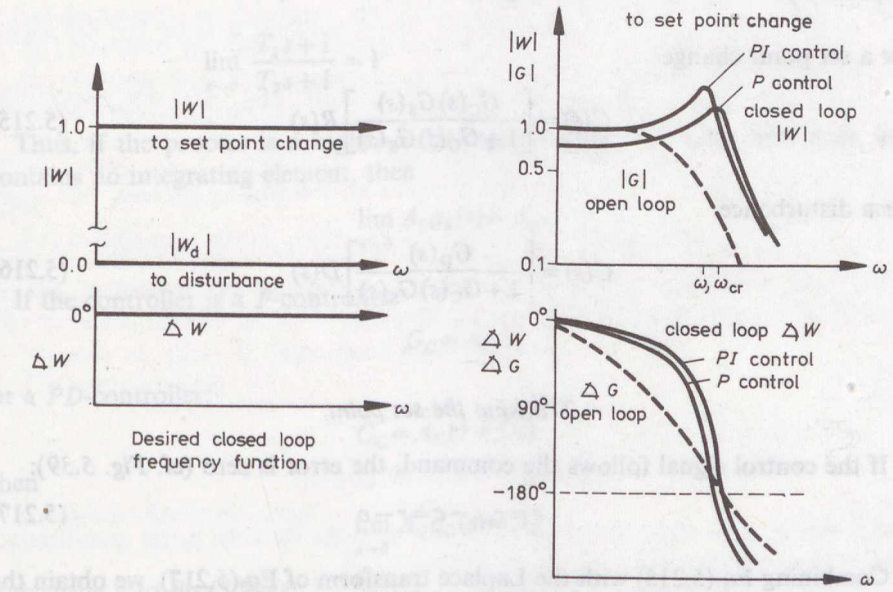



Fig. 5.63. Bode plots of closed control loop frequency functions

### Steady-state offset. Final value theorem

We have seen in Section 5.2.2 that, if a dead time process is controlled by a *P*-controller, the control loop is unable to restore the control signal to its original value after a step-like disturbance of the load variable, and it is also unable to make the control signal exactly follow a change in the set point. In other words, the error signal is different from zero, not only during a transient, but also in the steady state. This steady-state offset forced us to incorporate integral action in the controller. However, we did not examine the problem of steady-state error for the general case.

The final steady-state value of any linear time function can be determined using the final value theorem, Item 5 in *Table 5.3*, if the Laplace transform of the function is known:

$$f(\infty) = \lim_{s \to 0} sF(s) \tag{5.214}$$

It is important to check the stability of the function, before applying the final value theorem, because it gives no information about stability.

The Laplace transform of $c$, the control variable for the closed negative feedback control loop, is calculated using Eqs (5.100) and (5.101) in Section 5.3.4, respectively:

for a set point change

$$C(s) = \left[\frac{G_C(s)G_S(s)}{1 + G_C(s)G_S(s)}\right] R(s) \tag{5.215}$$

for a disturbance

$$C(s) = \left[\frac{G_D(s)}{1 + G_C(s)G_S(s)}\right] D(s) \tag{5.216}$$

*Tracking the set point*

If the control signal follows the command, the error is zero (cf. *Fig. 5.39*):

$$e = c_{set} - c = r - c \tag{5.217}$$

Combining Eq.(5.215) with the Laplace transform of Eq.(5.217), we obtain the Laplace transform of the error function:

$$E(s) = R(s) - \left(\frac{G_C(s)G_S(s)}{1 + G_C(s)G_S(s)}\right) R(s)$$

$$= \left(\frac{1}{1 + G_C(s)G_S(s)}\right) R(s) \tag{5.218}$$

Suppose a unit step in the set point: $R(s) = \frac{1}{s}$, then the final value of the error is [Eq.(5.214)]

$$e(\infty) = \lim_{s \to 0} s \left(\frac{G_C(s)G_S(s)}{1 + G_C(s)G_S(s)}\right) \frac{1}{s} = \lim_{s \to 0} \frac{G_C(s)G_S(s)}{1 + G_C(s)G_S(s)} \tag{5.219}$$

To obtain $e(\infty)=0$, the product $G_C(s)G_S(s)$ must go to infinity if $s$ goes to zero. This means that either the controller or the process must contain at least one integrator (the transfer function of which is $A/s$, cf. *Table 5.2*).

In the absence of an integrator, the controlled variable follows the set point with some error. The magnitude of the error is calculated from Eq.(5.219). To facilitate the calculation, consider that

$$\lim_{s\to 0} e^{-T_D s}=1$$

$$\lim_{s\to 0} \frac{1}{(T_1 s+1)^n (T_2 s+1)(T^2 s^2+2\xi T s+1)}=1$$

$$\lim_{s\to 0} \frac{T_1 s+1}{T_2 s+1}=1$$

Thus, if the process is a higher-order lag, possibly also with dead time, but contains no integrating element, then

$$\lim_{s\to 0} A_S g_S(s)=A_S$$

If the controller is a *P*-controller

$$G_C=A_C$$

or a *PD*-controller:

$$G_C=A_C(1+Ds)$$

then

$$\lim_{s\to 0} A_C g_C(s)=A_C$$

Substitute into Eq.(5.219):

$$e(\infty)=\lim_{s\to 0} s\left(\frac{1}{1+A_C g_C(s) A_S g_S(s)}\right)\frac{1}{s}=\frac{1}{1+A_C A_S} \tag{5.220}$$

Equation (5.220) indicates that a *P*- or a *PD*-controller may be used to control such processes, for which a small steady-state error can be tolerated and the process stability allows the use of a high-proportional gain. Recall from Eq.(5.205) that

$$(A_C A_S)_{max}=\frac{1}{g_C(i\omega_{cr}) g_S(i\omega_{cr})} \tag{5.221}$$

For the reader's orientation we may say that to a unit step a steady-state error $e(\infty)=0.05$, i.e. 5% of the step point change, is generally acceptable.

*Disturbance rejection*

The controller is expected to restore the controlled variable to the same steady-state value as it had before the disturbance: $\hat{c}(\infty)=0$.

Using Eq.(5.216), in response to a unit step disturbance: $D(s)=\frac{1}{s}$, we obtain

$$\hat{c}(\infty)=\lim_{s\to 0} s\left[\frac{G_d(s)}{1+G_C(s)G_S(s)}\right]\frac{1}{s}=\lim_{s\to 0}\frac{G_d(s)}{1+G_C(s)G_S(s)} \tag{5.222}$$

The following situations are possible:

1. $G_C(s)$ implies a *P*- or *PD*-controller. $G_S(s)$ contains a term $1/s$, but $G_d(s)$ does not. Then $\hat{c}(\infty)=0$.

2. $G_C(s)$ implies a *P*- or *PD*-controller. Both $G_D(s)$ and $G_S(s)$ are integrating processes containing a term $1/s$. Then $\hat{c}(\infty)=A_d/A_C A_S$.

3. $G_C(s)$ implies a *PI*- or *PID*-controller, containing a $1/s$ term. Then $\hat{c}(\infty)=0$, except in the case when $G_d(s)$ means an integrating process, but $G_S(s)$ not.

Thus the steady-state error for both command signal changes and disturbances is generally eliminated by using an integrating mode in the controller, and is diminished by the use of high loop gain, $K=A_C A_S$.

### 5.3.8 Controller tuning

*Control performance criteria*

In order to determine optimum controller settings we need some quantitative criteria to compare different control systems.

The performance criteria are not only used to determine the *P*, *I*, *D* parameters of the controller. Hitherto we have supposed that the parameters of the controlled process are given. In reality, even in the case of an existing industrial process, it is possible to choose different types of sensors and final control elements, to decide their placement, and thus, to influence the parameters of the process. As an example: the product composition of a distillation column may be controlled by (*a*) measuring the boiling temperature continuously at an intermediate tray; (*b*) measuring the composition intermittently by a process analyser from ($b\hat{a}$) a continuous vapour sample taken from the column head or ($b\hat{b}$) a continuous liquid sample taken from the reflux accumulator. In case (*b*) there is a dead time incorporated in the intermittent analysis, the effect of which can be evaluated by use of the performance criteria.

The desired performance of a closed-loop system can be specified in the time domain. The traditional test input signal is a step change in the set point. The

performance criteria based upon the response of the control signal are shown in *Fig. 5.64:*

(*i*) overshoot: $A/B$. This can be estimated from the phase margin of the open loop frequency function, or from the maximum, $M$, of the closed loop frequency function to the set point as follows:

| Phase margin | $M$ | Overshoot |
|---|---|---|
| $\leqq 30°$ | $\geqq 1.5$ | $M$–0.1 |
| 30°–40° | 1.5–1.3 | $M$ |
| 60° | 1.0 | $M$ |

(*ii*) decay ratio or damping: $A/C$. This can be estimated from the gain margin;

(*iii*) rise time: the time necessary to first reach the new steady-state value;

(*iv*) response time: the time it takes for the amplitude of the oscillations to decay to some fraction (generally $\pm 5\%$) of the final change of the control signal;

(*v*) period of oscillations, $T_p$. For a well-tuned control loop the response time is about 3 times the period. The period is estimated as the critical period of the open loop.

(*vi*) steady-state offset; estimated by the final value theorem.

The integral performance criteria are based upon the error $e = c_{set} - c$ during the response time *(Fig. 5.64)*. They may be calculated by computer for control loops with no offset or determined by experiment.

The error integral: (5.223)

$$EI = \int_0^\infty e\,dt$$

has the disadvantage of giving zero at the stability limit;

Integrated absolute error: (5.224)

$$IAE = \int_0^\infty |e|\,dt$$

given the total area at both sides of zero error;

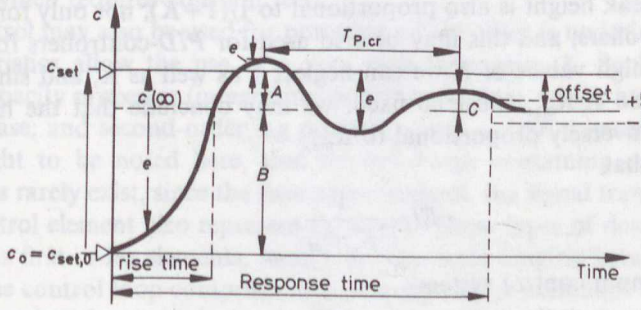



Fig. 5.64. Performance criteria based on closed loop step response

Integrated square error: (5.225)

$$ISE = \int_0^\infty e^2 \, dt$$ confers a penalty on overshoot;

Integrated time-weighted absolute error: (5.226)

$$ITAE = \int_0 |e| t \, dt$$ implies a penalty at long response times.

The optimum controller settings determined by the various criteria are different. Which criterion is to be used depends on the process.

*Quick estimation of the error integral*

This method can be used for a quick choice between different control systems. The best control system is that for which:

$$K_{cr}\omega_{cr} = \max \tag{5.227}$$

This quick criterion is based upon the step response, *Fig. 5.64* and the error integral, Eq.(5.223). It is supposed that the decay ratios in the control systems to be compared are equal and thus the oscillations are similar, and only the error areas under the first peaks need to be considered.

The area of the first peak is proportional to the product of its height and its length. The length of a peak is half the oscillation period; thus it is inversely proportional to the oscillation frequency, which, in turn, can be roughly estimated as the critical frequency. As a consequence, the error integral is inversely proportional to the critical frequency. It is important to note that when comparing controller settings, the phase shift of the controller is to be included in the calculation of the critical frequency.

The height of the first peak depends only on $K = A_S A_C$, the open loop gain product, which has been demonstrated by experience. Similarly to the steady-state offset (with a *P*-controller only), which is proportional to $1/(1 + A_S A_C)$; cf. Eq.(5.220), the peak height is also proportional to $1/(1 + K)$, not only for *P*-, but also for *PI*-controllers; and this may be also used for *PID*-controllers for rough estimation. For high values of $K$ we can neglect 1 as well as $K$, and since for a certain decay ratio $K/K_{max}$ must be fixed, we may conclude that the height of the first peak is inversely proportional to $K_{max}$.

Thus we find that

$$EI \sim \frac{1}{K_{cr}\omega_{cr}}$$

and for the optimum control system

$$EI = \min, \qquad K_{cr}\omega_{cr} = \max$$

**Controller selection**

To make a decision upon the control mode to be selected for a given process, let us summarize the features of the various control modes:

Proportional mode:

(*i*) produces no phase lag, thus it does not slow down the response of the process;
(*ii*) is unable to eliminate offset, except in the case of an integrating process.

Integral mode:

(*i*) eliminates offset,
(*ii*) exhibits phase lag (max. $-90°$) as well as frequency dependent gain, thus it has a retarding and a destabilizing effect.

Derivative mode:

(*i*) unable to eliminate error as unique control mode;
(*ii*) produces phase lead (max. $+60°$ for real controllers), thus speeds up control and has a stabilizing effect.

A three-mode *PID*-controller seems to be the best since it confers all kinds of benefits. But, at the same time, it presents a more complex tuning problem, which is hard to overcome, particularly with the nonlinear processes frequently encountered in chemical engineering. Recall that the *PID* parameters ought to be tuned to have $\frac{1}{I} < \omega_0 < \frac{1}{D}$, which is hard to obtain if $\omega_0$ changes with the operating point.

*P-controller*

This is the cheapest device and is the simplest to tune. It is used for control problems where offset is allowed or even desirable. An example of the latter is the level control in a buffer tank used to average out inflow fluctuations before a process which requires constant feed.

*P*-control may also be used for processes where offset is undesirable, but whose characteristics allow the use of a very high loop gain, $K$. Such processes are: single-capacity processes (integrating or first-order lag), which have a $-90°$ minimum phase; and second-order lag processes, with $-180°$ minimum phase.

It ought to be noted here, that control loops containing only one or two capacities rarely exist, since the sensing instrument, the signal transformer and the final control element also represent time lags. These types of device may be considered as first-order elements, their time constants ranging between 0.2 to 30 s. But, if the control loop comprises one capacity with a dominant time constant at least one order of magnitude greater than the second greatest, then the process may

be treated as a single capacity process and it may be controlled by a *P*-controller set to high loop gain, *K*. Such processes are: liquid tanks, stirred tank reactors, gas pressure controllers.

**Example 5.3.11.** *Control of a process with dominant time constant.*

Consider a liquid vessel having a time constant of 30 min. The time constant of the level transmitter, the pneumatic signal transmission line and the pneumatic control valve actuator sum together to 1 min. Although the transmitter and the valve actuator are first-order elements, the sum of the smaller time constants may be treated as pure dead time. This is an approximation, but it is on the safe side, since the dead time is less advantageous to control than a multicapacity system.

The natural period of the process is calculated as that of a single capacity, dead time process. Suppose the dominant time constant exhibits a $-90°$ phase lag at the natural period, then (cf. *Table 5.3*)

$$-\omega_0 T_D - 90° = -180°, \qquad \omega_0 = \frac{90°}{T_D}$$

$$T_{P,0} = \frac{2\pi}{\omega_0} = 4T_D = 4 \text{ min}$$

Verify: $$\varphi(T = 30 \text{ min}) = \arctan(-\omega_0 T) =$$

$$= \arctan\left(-\frac{2\pi}{4} \cdot 30\right) = -98.6°$$

At $T_{P,0} = 4$ min period, the dynamic gain is:

for the dead time: 1

for the first-order lag:

$$|g(i\omega)| = \frac{1}{\sqrt{1 + T^2\omega_0^2}} = \frac{1}{\sqrt{1 + \left(30\,\frac{2\pi}{4}\right)^2}} = 0.021$$

The maximum loop gain is

$$K_{max} = \frac{1}{|g_S(i\omega_0)|} = \frac{1}{0.021} = 47.6$$

To obtain quarter amplitude damping, the loop gain is set to

$$K = 23.8$$

The error to a unit step change in the set point is

$$e(\infty) = \frac{1}{1 + K} = 0.04, \text{ acceptable}$$

*PI-controller*

This is used when offset is not tolerable and the process has no dominating time constant. A *PI*-controller may also be used for processes dominated by dead time. It has the disadvantage of rendering the control more sluggish due to the time lag of the *I*-controller.

*PID-controller*

The *D* action produces a phase lag of about 60° (in a real controller). This almost cancels the phase lag of one capacity (the minimum phase is −90°). Thus it may decrease loop phase lag and increase the critical frequency in the case of processes consisting of three or four capacities, and speed up control substantially. *D* action is of no help in the case of processes dominated by dead time, because a dead time element corresponds to an infinite capacity system, of which only one is cancelled by *D* action.

*D* action cannot be applied to processes with substantial measurement noise, e.g. flow control.

**Controller tuning**

Several tuning tables giving controller settings are offered in handbooks. These are based upon some performance criteria applied to an appropriate process model. The simplest process model is a first-order lag with dead time, with the transfer function

$$G(s) = \frac{A_S e^{-T_D s}}{Ts+1} \tag{5.228}$$

The Ziegler—Nichols and the Cohen—Coon tuning tables are based upon this model and give controller settings for quarter-amplitude damping. The Cohen—Coon settings were determined using the *ISE* criterion.

The parameters of the process model are determined from the process reaction curve, as shown in *Fig. 5.65*. The process reaction curve is the open loop response

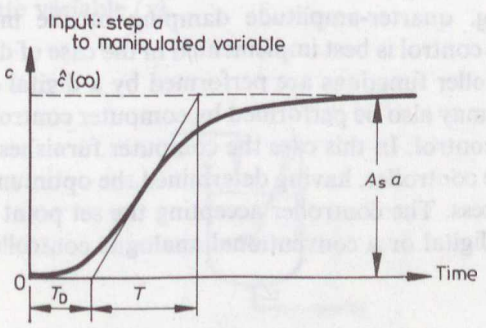



Fig. 5.65. Process reaction curve

of the controlled variable to a unit step input in the manipulated variable. All elements of the control loop must be included in the response, except the controller.

The determination of the parameters can be seen in *Fig. 5.65*. A tangent is drawn to the inflection point of the process reaction curve. The model dead time is determined as the time at which this tangent intersects the base line. The time at which the tangent intersects the 100% response line is the sum of the model time constants, $T+T_D$. The model gain factor, $A_S$, is, of course, the 100% response to unit input.

There are some difficulties in taking the process reaction curve. It may take a long time to obtain 100% response. It is difficult to draw the tangent line accurately. Thus the controller settings so obtained will be inaccurate, too.

The model parameters, critical period $T_{P,cr}$ and critical gain $A_{C,cr}$, may be determined more exactly. The experimental determination of these parameters involves the cycling method described in Section 5.2.6. They may also be determined from the open loop frequency function of the process, cf. Section 5.3.7. The controller settings based on these parameters are also given in *Table 5.1*.

### *Adaptive and optimal control*

Most chemical engineering systems are nonlinear and if their operating point is changed, their gain factors and time constants change too. This may cause instability in a well-tuned control loop, and make it necessary to re-tune control loops frequently.

This difficulty may be overcome by adaptive control. Adaptive control means automatic adaption of the controller settings to the varying process parameters. Its simplest form is gain adaptive control, when only the controller gain is varied. Since process parameter changes are mostly connected with changes in the flow rate through the process, gain adaptive control may be implemented by measuring the position of the control valve which manipulates the flow rate, and feeding the correlation between valve position and controller gain into a microprocessor controller or into a computer, which adjusts the controller settings.

Self-tuning controllers are able to set all their tuning parameters by the use of some criterion, e.g. quarter-amplitude damping or the integral performance criterion. Adaptive control is best implemented in the case of direct digital control, i.e. when the controller functions are performed by a digital computer.

Optimal control may also be performed by computer control, but in this case by using supervisory control. In this case the computer furnishes the set point (command signal) to the controller, having determined the optimum operating point of the controlled process. The controller accepting the set point and controlling the process may be a digital or a conventional analogue controller.

## 5.4 The state space method

### 5.4.1 The state of a system

The state of any system means the past, present, and the future of the system. The state space method gives all this information. It is not limited to linear systems: it is capable of describing nonlinear and variable parameter systems, too. Using the state space method it is possible to investigate the system in the time domain, since it depends on the past history of the system. The classical linear methods are capable of describing linear systems especially in the frequency domain, and they neglect initial conditions.

The state space method uses the so-called state variables and state equations to determine the state and the dynamic behaviour of the system. The state variables are not always the output variables of the system and they cannot always be measured or observed. The set of the state variables $x_1(t)$, $x_2(t)$, ..., $x_n(t)$ must satisfy the following two conditions:

1. At any entirely randomly selected instant of time ($t=t_0$) the state variables $[x_1(t_0), x_2(t_0), \ldots, x_n(t_0)]$ describe the initial condition of the system $[x_1(t_0), x_2(t_0), \ldots, x_n(t_0)]$.
2. The inputs of the system have to be specified for all $t \geqq t_0$, then the state variables completely determine the behaviour, the state of the system at any time $t$.

**Example 5.4.1**

Determine the state variable of a perfectly mixed tank reactor using the first-order reaction equation and its state equation *(Fig. 5.66)*.

The well-known differential equation of such a system is:

$$T\frac{dc_1}{dt}+c_1=A_S c_0 \tag{5.229}$$

The concentration in the reactor ($c_1$) completely specifies the state of the system, so it is the state variable ($x$).

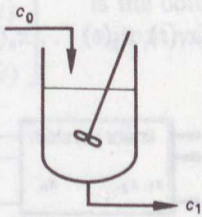



Fig. 5.66. State variables of a multivariable system

Laplace transform and solve the equation ($c_i$ cannot be neglected at $t=0$, $c_i(0)$ is a constant):

$$A_S c_0(s) = c_i(s) + T[c_i(s)s - c_i(0)]. \tag{5.230}$$

Rearranging:

$$c_i(s) = \frac{A_S c_0(s)}{1+Ts} + \frac{Tc_i(0)}{1+Ts} \tag{5.231}$$

Inverse transform Eq.(5.231) in the case of step disturbance of $c_0$ at $t \geq 0$:

$$c_i(t) = A_S c_0\left(1 - e^{-\frac{t}{T}}\right) + c_i(0)e^{-\frac{t}{T}} \tag{5.232}$$

It is clear that if $c_i(0)$ and the input at $t \geq 0$ are known, the state of the system is determined. Thus $c_i$ satisfies the basic requirements for a state variable. Rearranging Eq.(5.229) one obtains:

$$\frac{dc_i}{dt} = -\frac{1}{T}c_i + \frac{A_S}{T}c_0\,, \tag{5.233}$$

which is the state equation of a first-order element. It is usual to denote the state variable by $x$, the input variable by $u$, and the output variable by $y$. Generally, the state equation is:

$$\frac{dx(t)}{dt} = ax(t) + bu(t) \tag{5.234}$$

and the so-called output equation is:

$$y(t) = cx(t) + du(t) \tag{5.235}$$

In Example 5.4.1 the output equation is simply:

$$y(t) = x(t) \tag{5.236}$$

### 5.4.2 State equations of multivariable systems

Consider a linear multivariable system with $k$ inputs and $m$ outputs (see *Fig. 5.67*).

The state variables are $x_1, x_2 \ldots x_n$. The system dynamics are described by $n$ first-order differential equations, the state equations being similar to Eq.(5.234)

$$\frac{dx_i(t)}{dt} = f[x_1(t), x_2(t) \ldots x_n(t), u_1, u_2 \ldots u_k] \tag{5.237}$$

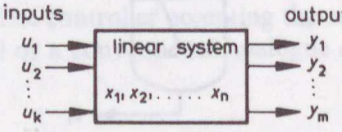



Fig. 5.67. To Example 5.4.1

or

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{n} a_{i,j}x_j(t) + \sum_{v=1}^{k} b_{i,j}u_v(t) \tag{5.238}$$

The output equations are:

$$y_p(t) = \sum_{j=1}^{n} c_{p,j}x_j(t) + \sum_{v=1}^{m} d_{p,v}u_v(t) \quad \text{where} \quad \begin{matrix} p=1,2\ldots m \\ v=1,2\ldots k \end{matrix} \tag{5.239}$$

It is much more convenient to use the matrix form of representation:

$$\frac{d\underline{x}(t)}{dt} = \mathbf{A}\underline{x}(t) + \mathbf{B}\underline{u}(t) = f[x(t), u(t)] \tag{5.240}$$

or

$$\dot{x} = \mathbf{A}\underline{x}(t) + \mathbf{B}\underline{u}(t) \tag{5.241}$$

The output equation is

$$\underline{y}(t) = \mathbf{C}\underline{x}(t) + \mathbf{D}\underline{u}(t) \tag{5.242}$$

where

$$\underline{\dot{x}} = \begin{bmatrix} \frac{dx_1(t)}{dt} \\ \frac{dx_2(t)}{dt} \\ \vdots \\ \frac{dx_n(t)}{dt} \end{bmatrix} \qquad \underline{x} = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix}$$

is the column matrix of state variables

$$\underline{u} = \begin{bmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_k(t) \end{bmatrix}$$

is the column matrix of input variables

$$\underline{y} = \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_m(t) \end{bmatrix}$$

is the column matrix of output variables

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & & & \\ \vdots & & & \\ a_{n1} & \ldots\ldots & & a_{nn} \end{bmatrix}$$

is an $n \times n$ square matrix, the state matrix

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1k} \\ b_{21} & & & \\ \vdots & & & \\ b_{n1} & \dots\dots & & b_{nk} \end{bmatrix}$$

is an $n \times k$ matrix, the input matrix

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & & & \\ \vdots & & & \\ c_{m1} & \dots\dots & & c_{mn} \end{bmatrix}$$

is an $m \times n$ matrix, the output matrix

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1k} \\ d_{21} & & & \\ \vdots & & & \\ d_{m1} & \dots\dots & & d_{mk} \end{bmatrix}$$

is an $m \times k$ matrix, the transmission

### 5.4.3 State equations of higher-order elements

Consider an $n$th order linear system with single input and single output. Its general differential equation is:

$$\frac{d^n y(t)}{dt^n} + a_1 \frac{d^{n-1} y(t)}{dt^{n-1}} + a_2 \frac{d^{n-2} y(t)}{dt^{n-2}} + \dots + a_{n-1} \frac{dy(t)}{dt} + a_n y(t) = u(t) \tag{5.243}$$

The state variables are:

$$\begin{aligned} x_1(t) &= y(t) \\ x_2(t) &= \frac{dy(t)}{dt} \\ &\vdots \\ x_n(t) &= \frac{d^{n-1} y(t)}{dt} \end{aligned} \tag{5.244}$$

and the state equations:

$$\begin{aligned} \frac{dx_1(t)}{dt} &= x_2(t) \\ \frac{dx_2(t)}{dt} &= x_3(t) \\ &\vdots \\ \frac{dx_{n-1}(t)}{dt} &= x_n(t) \end{aligned} \tag{5.245}$$

$$\frac{dx_n(t)}{dt} = -a_n x_1(t) - a_{n-1} x_2(t) - \ldots - a_2 x_{n-1}(t) - a_1 x_n(t) + u(t)$$

The output equation is very simple:

$$y(t) = x_1(t) \tag{5.246}$$

The matrix representation of Eqs (5.244)—(5.245) is:

$$\begin{bmatrix} \frac{dx_1(t)}{dt} \\ \frac{dx_2(t)}{dt} \\ \vdots \\ \frac{dx_n(t)}{dt} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -a_n & -a_{n-1} & \ldots\ldots & & -a_1 \end{bmatrix} \cdot \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \cdot u(t) \tag{5.247}$$

and the output matrices are:

$$\mathbf{C} = [1 \quad 0 \quad 0 \quad \ldots \quad 0]$$

and

$$\mathbf{D} = \mathbf{0} \text{ (null matrix)}$$

because the output equation is a simple scalar equation [Eq.(5.247)].

Real systems are often both higher-order and multivariable. The state space method offers the possibility to reduce the dynamic mathematical model of such real systems to a set of first-order differential equations. This is the method applied by *TACS* (Teaching Aid for Control Studies).

**Example 5.4.2**

Determine the state variables and state equations of a general second-order element.

The differential equation is:

$$T^2 \frac{d^2 y(t)}{dt^2} + 2\xi T \frac{dy(t)}{dt} + y(t) = u(t) \tag{5.248}$$

The state variables:

$$\begin{aligned} x_1(t) &= y(t) \\ x_2(t) &= \frac{dy(t)}{dt} \end{aligned} \tag{5.249}$$

and the state equations:

$$\frac{dx_1(t)}{dt} = x_2(t)$$
$$\frac{dx_2(t)}{dt} = -\frac{1}{T^2}x_1(t) - \frac{2\xi}{T}x_2(t) + \frac{1}{T^2}u(t) \tag{5.250}$$

In matrix form

$$\begin{bmatrix} \frac{dx_1(t)}{dt} \\ \frac{dx_2(t)}{dt} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{1}{T^2} & -\frac{2\xi}{T} \end{bmatrix} \cdot \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{T^2} \end{bmatrix} \cdot u(t) \tag{5.251}$$

The output equation is:

$$y(t) = x_1(t) \tag{5.252}$$

**Example 5.4.3**

Consider a mixing unit *(Fig. 5.68)* and determine its state variables and state equations. $c_1$ and $c_2$ are constant. The other variables are time-dependent. Flows can change.

This is a multivariable (2) system. The total mass balance:

$$\frac{dV}{dt} = W_1 + W_2 - W_3 \tag{5.253}$$

The component balance:

$$\frac{d(Vc_3)}{dt} = W_1c_1 + W_2c_2 - W_3c_3 \tag{5.254}$$

$$V\frac{dc_3}{dt} + c_3(t)\frac{dV}{dt} = W_1c_1 + W_2c_2 - W_3c_3 \tag{5.255}$$

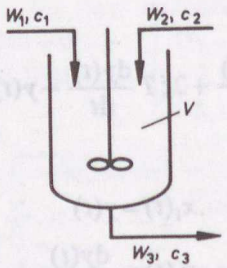



Fig. 5.68. To Example 5.4.3

Substituting the total mass balance into the component balance we obtain:

$$V\frac{dc_3}{dt} = W_1c_1 + W_2c_2 - W_3c_3 - [c_3(W_1 + W_2 - W_3)] \tag{5.256}$$

$$V\frac{dc_3}{dt} = W_1(c_1 - c_3) + W_2(c_2 - c_3) \tag{5.257}$$

The state variables are: $V$ and $c_3$

The state equations are:

$$\frac{dV}{dt} = W_1 + W_2 - W_3 \tag{5.258}$$

$$\frac{dc_3(t)}{dt} = -\left(\frac{W_1(t)}{V} + \frac{W_2(t)}{V}\right)c_3(t) + \frac{c_1}{V}W_1(t) + \frac{c_2}{V}W_2(t) \tag{5.259}$$

This is a variable parameter equation.

The state equations in matrix form are:

$$\begin{bmatrix} \frac{dc_3(t)}{dt} \\ \frac{dV(t)}{dt} \end{bmatrix} = \begin{bmatrix} -\left(\frac{W_1}{V} + \frac{W_2}{V}\right) & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} c_3(t) \\ V(t) \end{bmatrix} + \begin{bmatrix} \frac{c_1}{V} & \frac{c_2}{V} & 0 \\ 1 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} W_1(t) \\ W_2(t) \\ W_3(t) \end{bmatrix} \tag{5.260}$$

The state and input matrices are also functions. They are not constant.

The output equation is simply

$$y(t) = c_3(t) \tag{5.261}$$

This example can be solved in a different way, so that the state and input matrices should be constant. In this case linearization and perturbation variables have to be used. The linearization, of course, brings its usual error with it, but it is much easier to solve.

The linearized state equations are:

$$\begin{bmatrix} \frac{d\hat{c}_3}{dt} \\ \frac{d\hat{V}}{dt} \end{bmatrix} = \begin{bmatrix} \frac{-\bar{W}_3}{\bar{V}} & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \hat{c}_3 \\ \hat{V} \end{bmatrix} + \begin{bmatrix} \frac{c_1 - \bar{c}_2}{\bar{V}} & \frac{c_2 - \bar{c}_3}{\bar{V}} & 0 \\ 1 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} \hat{W}_1 \\ \hat{W}_2 \\ \hat{W}_3 \end{bmatrix} \tag{5.262}$$

### 5.4.4 The solution of the state equations

The solution of the state equations

$$\dot{\underline{x}} = \mathbf{A}\underline{x} + \mathbf{B}\underline{u}$$

can be performed by the Laplace transformation method or by a classical method, or some numerical method.

The most convenient way of solving the state equations is to use a computer. There are many suitable numerical methods, such as the trapezoid, Euler, or Runge–Kutta, etc. methods. The solution of the state equation is the dynamic behaviour of the modelled system in the time domain.

## 5.5 Notation

| | |
|---|---|
| $a, b, c, \ldots$ | constants |
| $A, B, C, \ldots$ | |
| $A$ | gain factor |
| $A$ | area |
| $c$ | controlled variable |
| $C(s)$ | its Laplace transform |
| $c$ | concentration |
| $d$ | disturbance, load variable |
| $D(s)$ | its Laplace transform |
| $D$ | derivative time constant |
| $e$ | error |
| $E(s)$ | its Laplace transform |
| $g(s) = \frac{G(s)}{A}$; $g(i\omega)$ | normalized transfer or frequency function |
| $G(s)$, $G(I\omega)$ | transfer or frequency function |
| $H$ | height |
| $i = \sqrt{-1}$ | |
| $I$ | integral time constant |
| $K$ | loop gain |
| $L(s)$ | open loop transfer function |
| $m$ | manipulated variable |
| $p$ | pressure |
| $\mathscr{L}$ | Laplace operator |
| $t$ | time |
| $T$ | time constant |
| $T_D$ | dead time |
| $T_P$ | period of oscillation |

| | |
|---|---|
| $V$ | volume |
| $W$ | flow rate |
| $W(s)$, $W(i\omega)$ | closed loop transfer or frequency function |
| $x$ | input |
| $X(s)$ | its Laplace transform |
| $y$ | output |
| $Y(s)$ | its Laplace transform |
| $\varphi$ | phase angle |
| $\omega$ | frequency |
| $\omega_0$ | natural frequency |

*Subscripts:*

| | |
|---|---|
| C | refers to the controller |
| cr | critical |
| S | refers to the process |

*Superscripts:*

| | |
|---|---|
| $\bar{}$ | steady state |
| $\hat{}$ | deviation |

## 5.6 Literature

[1] Coughanowr, D. R. and Koppell, L. B. (1965): *Process Systems Analysis and Control*. McGraw-Hill, New York.

[2] Luyben, W. L. (1973): *Process Modeling, Simulation, and Control for Chemical Engineers*. McGraw-Hill, New York.

[3] Stephanopoulos, G. (1984): *Chemical Process Control*. Prentice-Hall, New York.

[4] Shinskey, F. G. (1979): *Process Control Systems*. (2nd ed.) McGraw-Hill, New York.

[5] Smith, C. L. (1972): *Digital Computer Process Control*. Intact Educational Publishers, Pa.

[6] Lipták, G. B. and Venczel, K. (eds) (1982): *Instrument Engineers Handbook*. Chilton, Pa.

# 6 MEASURING SUBSYSTEMS FOR THE PROCESS CONTROL OF CHEMICAL ENGINEERING SYSTEMS

by

G. E. VERESS

Due to the rapid development of analytical chemistry and an increase in the number of novel methods, the specialist faces an increasingly difficult situation in selecting the most appropriate method for a given chemical measuring task.

At the same time, the increasing trend towards the automation of the methods of analytical chemistry makes it necessary to develop and unify the basic principles of automatic analysis.

In order to solve the above problems some pioneering steps have been published in the analytical chemistry literature concerning the unified treatment of these methods, and a new branch of science, 'chemometrics' (chemical metrology) has been born.

In this chapter, the fundamentals of the unified, system theoretical treatment of analytical chemical methods will be summarized (Section 6.1), followed by an analysis of the most important problems of signal processing (Section 6.2), signal interpretation (Section 6.3) and the characterization of chemical measuring methods (Section 6.4).

## 6.1 Chemical measuring systems

The intention is to develop a unified treatment and characterization of chemical measuring systems, and thus Section 6.1.1 deals with the definition of chemical measuring systems and Section 6.1.2 with the general questions of their mathematical modelling.

### 6.1.1 Definition of chemical measuring systems

*Concept of analytical chemical system*

Analytical chemical systems are purpose-built systems which assign relevant chemical analytical information to the substance to be analyzed, according to the purpose of the system, i. e. they are systems that produce chemical information. As purpose-built systems, the production of information is carried out in order to reach a certain aim, and thus the systems comprise object and control subsystems.

Analytical chemical systems usually form subsystems of economic systems, or

they are economic systems in their own sight and thus their purpose is economical or they have a related, equivalent 'natural' (technical) aim. Consequently, they should be characterized from an economical viewpoint.

It follows from this definition that any qualitative or quantitative analytical chemical method or procedure may be regarded as an analytical chemical system.

In a physical sense, an analytical chemical system generally consists of the substance to be analyzed, the materials and equipment necessary for the investigation, the experimental measuring instrument, the person carrying out the analysis and, finally, the analytical protocol, i.e. instructions concerning the investigation. In the case of complete automation, the person and the protocol are replaced by a computer and its program.

As with all systems, including analytical chemical systems, the foundation, operation, storage during nonoperational periods, development and dismounting can be defined as basic modes of operation. Operation consists of running and maintaining the system, the latter consisting of maintenance and repair. Of these, we will be concerned with the running mode only, which will be termed, according to everyday usage, the operation.

A fundamental characteristic of the operation of analytical chemical systems is the time necessary to 'produce' the chemical information, i.e. the operation time of the system. Accordingly, the operation of analytical chemical systems as a function of time can be characterized by an input of a 'flow' of material to be analyzed and an output of a 'flow' of chemical information.

Depending on whether the analyzed substances or the information produced are input or output in 'batch' or 'continuous' manner, respectively, the analytical system is called a batch (discrete) or continuous system, respectively.

Depending on whether the analyzed substance is input in a closed or open chain, systems exit with on-line or off-line substance flow. Similarly, systems with on-line and off-line information flow may be distinguished.

There are three levels of hierarchy of analytical chemical systems. One may discuss the instrumental level, which may be a system consisting of a single instrument including the controller of the instrument (of course, which may be a human or a computer), represented by a simple or compound instrument or an instrument system; the laboratory level works as a subsystem (including the room, equipment and the chief of the laboratory; and one may define the institutional level, which works as an independent system. This section deals with the instrumental level only.

*Principle of operation of analytical chemical systems*

The principle of operation of analytical chemical systems is that the chemical information on the substance to be analyzed, which generally cannot be determined directly, is determined on the basis of measurable, observable physical, biological, or other analytical information.

This indirect determination is, of course, possible only if the relationships between the analytical and the chemical information are known, i.e. if we can perform an analytical chemical correlation.

The principle of analytical chemical correlation and thus of chemical analysis is that the substance to be analyzed is divided into common constituents for which there is a mutually unambiguous or at least unambiguous connection between the chemical and analytical information defined on the basis of the chemical and analytical properties of the constituents.

Consequently, the operation of analytical chemical systems consists in principle in defining the concept of constituent, i.e. the smallest analyzed unit of the sample on the basis of the chemical information or chemical property to be determined, and then assigning an analytical property to this constituent, for which there is an unambiguous relationship between the analytical information defined by this property and the chemical information to be determined.

Since the principle of analytical chemical systems is that chemical information is measured indirectly, through the determination of analytical information, analytical chemical systems consist of two main parts, the analytical investigation subsystem, which produces analytical information, and the analytical chemical correlation subsystem, which assigns chemical information to analytical information (see *Fig. 6.1*).

The investigation subsystem assigns to the sample, on the basis of the measurable, determinable analytical properties, a statement on the qualitative and/or quantitative analytical composition of the sample, that is the analytical information.

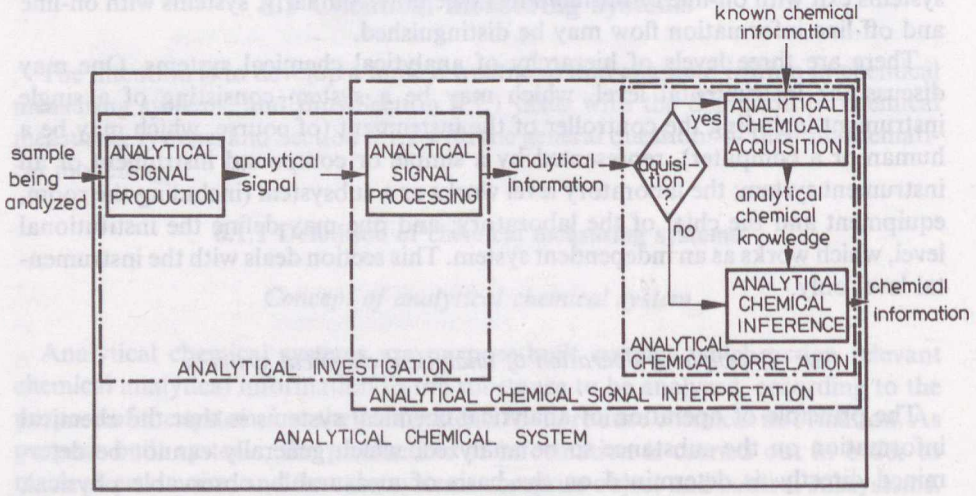



Fig. 6.1. Structure of analytical chemical systems

The correlation subsystem, on the basis of the correlations between the analytical and chemical properties of the samples, draws conclusions on the chemical information pertaining to the sample on the basis of the analytical information.

*Analytical chemical correlation*

Analytical chemical correlation consists of two main parts: analytical chemical data acquisition and analytical chemical inference.

During analytical chemical data acquisition, or calibration, analytical chemical knowledge can be derived from prior knowledge, and analytical information may be determined from calibration samples using known chemical information. Using this analytical chemical knowledge during analytical chemical inference, chemical information pertaining to the sample is assigned to the analytical information obtained from the sample (see *Fig. 6.1*).

Two types of correlation can be distinguished: nonsimultaneous (*a priori*) correlation, in which the knowledge is acquired first and then an inference is made on the basis of this knowledge, and simultaneous, in which acquisition of knowledge and inference take place simultaneously. This definition, therefore, covers not only the classical 'calibration' measurements but also other methods (e.g. the method of standard additions).

*Analytical chemical sign process*

From the perspective of semiotics, the science of signs, analytical chemical systems may be regarded as a controlled sign process, i.e. semiosis. The basis of their operation is the sign process in which the sign producing unit, the analytical sign producer, assigns a sign to the sample to which the interpreter unit of the process, the analytical chemical sign interpreter, assigns the 'meaning' of the sign, i.e. the chemical information pertaining to the sample. Thus, an analytical chemical system consists of analytical sign producer and analytical chemical sign interpreter units (see *Fig. 6.1*).

The sign produced by the analytical sign producer may be any kind of sign, taken in a semiotic sense, which thus may include the colour of a colour reaction in a classical chemical qualitative analysis, or a spectrum, etc.

As with any sign process, it must be stressed here that the analytical chemical system is determined jointly by the chain comprising sample — sign producer — sign — sign interpreter — chemical information, and thus the component parts may not be discussed independently.

Therefore, the choice of the sign of an analytical chemical system is never determined by the sign producer or the sign itself, but rather by the complete analytical chemical system which also include the sign interpreter unit. Thus, for example, the result of a viscosity measurement, a viscosity value, can be regarded as a sign of an analytical chemical system only if there exists a sign interpreter

which can be used to draw an inference on the chemical quality of the sample from the value of its viscosity. On the other hand, a thermoanalytical measurement, if its aim is not the determination of chemical information but of a physical property (e.g. melting point) cannot be regarded as an analytical chemical system.

*Analytical sign production*

The analytical sign production unit consists of two main parts, the sample preparation and analytical investigation processes.

The part of the material to be investigated on which the investigation is in fact carried out is called sample. If the material is sampled at several spatially different, then points the spatial distribution of the chemical constitution of the substance can be determined. In most cases the sample is a finite spatial part of the substance, and thus the chemical constitution determined on the basis of the sample is an average constitution of the substance for the given spatial element. Of course, one must carefully ensure, e.g. by homogenizing of the sample, that the sample is representative of the material.

A similar line of thought can be used to define the chemical constitution of material changing in time, which may refer to a given point of time or to a time interval, yielding a varying, dynamic constitution or a time average of the constitution.

The process of sample preparation transforms the substance to be investigated into a sample which is able to produce a sign by means of the sign production process. Sample preparation consists of a pretreatment stage, and a sampling of the pretreated subject.

The most important part of analytical sign production and of the entire analytical chemical system is the analytical investigation process, which assigns to the sample obtained by sample preparation, a sign that is interpretable by the sign interpreter unit. The analytical investigation unit consists of analytical treatment, detection and reading out.

*Analytical chemical sign interpretation*

The operation of the analytical sign interpreter unit, which assigns to the sign produced by the analytical sign production unit the analytical chemical 'meaning' of the sign, i.e. the chemical information pertaining to the sample, can be divided into analytical sign processing and analytical chemical correlation. Analytical sign processing assigns the statement pertaining to the so-called analytical properties (relevant from the viewpoint of analytical chemistry), i.e. the 'analytical information' to the sign. Signal processing methods are discussed in Section 6.2.

Analytical correlation relates the chemical information pertaining to the given analytical information to the chemical information pertaining to the sample.

Consequently, sign interpretation is a joint process of sign processing, acquisi-

tion and inference. In the case of samples with known chemical information, by means of the analytical information produced by sign processing, knowledge is produced in calibration (learning), i.e. a correlation is produced between chemical and analytical information. The analytical information of the sample with unknown chemical information is then converted into chemical information by inference, using analytical chemical knowledge. Therefore, sign processing is closely connected with the process of correlation.

The analytical chemical interpretation of the sign may be different, depending on the type of chemical information required and the level of sign interpretation. Signal interpretation may be simple — e.g. 'fingerprint'-like identification — but it may also be a complex process model. The best known sign interpretation methods are reviewed in Section 6.3.

### 6.1.2 Mathematical modelling of chemical measuring systems

The principle of the operation of analytical chemical systems is that during the process of analytical chemical acquisition (calibration), analytical chemical knowledge is produced, i.e. the chemical information input vs. analytical information output model of analytical investigations is achieved, and during analytical chemical inference, in the possession of the model of analytical investigation, chemical information is produced from analytical information.

Several models of an analytical investigation are possible. Verbal–logical models are used by expert systems, symbolic models by syntactic pattern recognition systems, algebraic models by algebraic–logical systems, etc., but most of the analytical chemical methods are based on numerical methods.

The next two sections deal with the logical modelling and measure theoretical modelling of analytical chemical systems.

*The logical model of analytical chemical systems*

Let us denote chemical information relevant to the sample by $S^c$ and analytical information by $S^a$. The analytical investigation corresponds to the logical implication

$$\forall m(S^c m \rightarrow S^a m)$$

i.e. to the statement that the sample with chemical information $S^c$ possesses analytical information $S^a$. Consequently, from a given item of chemical information a given unit of analytical information always follows, but for the occurrence of a given item of analytical information the given chemical information is not a necessary condition. Our knowledge may therefore be presented as an array of implications.

Analytical chemical knowledge can be acquired by analytical chemical learning, by means of determining the analytical information of samples with known chemi-

cal information. The rule of inference used in acquisition is 'modus ponens' (assertion mode). If $m_c$ denotes a calibration sample, $S^c m_c$ denotes our previous knowledge of the chemical information of the calibration sample and $S^a m_c$ denotes the analytical information obtained on the calibration sample. This mode of inference can then be written as

$$\{\forall m_c (S^c m_c \rightarrow S^a m_c); \quad S^c m_c\} \Rightarrow S^a m_c$$

This rule states the necessity of the existence of analytical information.

From a logical point of view, analytical inference may be of several types. If in the analytical investigation the absence of any analytical property is observed, the analytical inference is, from the logical point of view, the inference rule of 'modus tollens' (discarding mode), and thus the proof is

$$\{\forall m_c (S^c m_c \rightarrow S^a m_c); \quad \overline{S^a} m_c\} \Rightarrow \overline{S^c} m_c$$

This rule states that, from the absence of the given analytical information, the absence of the given chemical information follows with certainty. Consequently, in analytical chemistry in general chemical information pertaining to the absence of given constituents which follow from the absence of given analytical information can be obtained by a logically correct procedure.

Conclusions on the presence of chemical information may be drawn only if our analytical chemical knowledge can be converted into equivalences rather than implications, i.e.

$$\forall m_c (S^c m_c \leftrightarrow S^a m_c)$$

For a mutually unambiguous assignment it is necessary that only one piece of analytical information belongs to any piece of chemical information that gives rise to this analytical information.

In order to be able to present analytical chemical knowledge in the form of equivalence, it is necessary that the set of all possible chemical information is finite and that calibration is carried out for all of its elements. This condition is fulfilled, for instance, if the given constituent is an atom, then its chemical properties may assume a finite set of values only.

If the analytical chemical knowledge can be expressed as a set of equivalences, analytical chemical inferences are logically correct, proving inferences:

$$\{\forall m_c (S^c m_c \leftrightarrow S^a m_c); \quad S^a m_c\} \Rightarrow S^c m_c$$

However, most of the analytical chemical inferences are, from the perspective of formal logic, not proving, but credible (plausible) statements:

$$\{\forall m_c (S^c m_c \rightarrow S^a m_c); \quad S^a m_c\} \overset{?}{\Rightarrow} S^c m_c$$

i.e. the inference $S^c m_c$ is only plausible, not certain. Therefore, the credibility of analytical and chemical information should be checked by appropriate methods (Section 6.4.5).

*A measure theoretical model for analytical chemical systems*

A measure theoretical model based on such ideas as element, property of element, component and quantity of component can be used for the description of an analytical investigation.

The material to be investigated is modelled by a set that is called the material. The smallest investigated unit of the material is modelled as an element, and thus the material may be seen to be a set of elements.

The property of an element as the subject of investigation is modelled by a function defined on the elements, which is called the 'property of element' function. The property of an element can be regarded as an intensive property of the material.

A component may be modelled by defining it as a set of elements corresponding to a value of a 'component defining' property.

The set of components is the power set of the material. The amount of material and of the components is modelled as a measure on the component set, and it is called the 'quantity of component'. This latter can be regarded as an extensive property of the material. The distribution of materials and of material quantity can be defined based on the idea of property defining component.

This measure theoretical model can be used for describing the operation of analytical chemical systems. In these systems two types of properties are assigned to the elements investigated: the observed analytical properties and the inferred chemical properties. Based on them, observed analytical components and inferred chemical components may be defined. Similarly, two types of quantities, the observed analytical and the inferred chemical quantities are assigned to the material or to the components.

If the analytical investigation is a linear, time-invariant system then as a conclusion of the measure theoretical model, it may be expressed that the fact that the analytical distribution of material (quantity) is a mixed distribution of the analytical distributions of the chemical components (quantities), weighted by the chemical distribution of material (quantity).

The analytical investigation means the determination of the analytical distribution of material (quantity), the acquisition is concerned with the determination of the analytical distribution of the chemical components, and inference is the determination of the chemical distribution of material (quantity) using the mixed distribution model of material (quantity) determined by the investigation.

In the case of qualitative analysis, inference is the determination of the chemical distribution of material by the resolution of the analytical distribution of the material and, in the case of quantitative analysis, inference means the determination of the chemical distribution of material quantity by the resolution of the analytical distribution of the material quantity.

## 6.2 Signal processing methods

### 6.2.1 Signal transformations

*Basic principles*

In the processing of measured data, a basic role is played by transformations and operations which assist in the estimation of relevant information from a set of noisy signals.

The concepts of signal analysis and signal processing cannot be defined precisely, but they include, in general, all transformations and operations that can be used to distinguish between useful information and noise.

The major advance of signal processing can be dated to the last decade, although a number of its fundamentals had been known long before, primarily in the field of electronics. The literature and nomenclature of the subject are far from unambiguous, different authors using different terms for the same concepts.

One school of the various signal analysis techniques is based on the theory of stochastic processes, but these techniques of analysis are also widely applied when the properties assumed to be characteristic of stochastic processes cannot be proved to exist for the process observed.

Depending on the domain in which the signals are interpreted (i.e. on the choice of independent variable), two basic types of signal can be distinguished: wave-form (time domain) and spectrum (frequency domain). (In a broader sense, the correlogram can be regarded as a special case of the time domain, in which the independent variable is the difference [in time] between the two signals to be compared.)

In principle, an infinite number of signal transformations and operations can be conceived, and thus the number of signal processing methods is, in principle, very large. However, the number of methods applied in practice is not too great. In this chapter we shall attempt to discuss these methods in a systematic way.

The most important operations that are performed on signals are convolution and correlation, whereas, of the transformations that are applied to the signals. Fourier transformation, derivation and integration play a significant role.

On the basis of transformations carried out on a given signal and on the basis of operations performed between signals, smoothed, sharpened (or resolved), derivative and integrated forms can be distinguished, in addition to the original signal.

No sharp distinction is drawn between discrete and continuous signals and transformations or operations. From the viewpoint of computer technology, the discrete form is of significance, and in practice calculations are performed using the discrete forms.

From the aspects of discrete transformations and operations, digital filters play a leading role. Linear digital filters are represented by the relationship

$$y_j = \sum_k a_k x_{j-k} + \sum_i b_i y_{j-i}$$

where $y$ is the filtered and $x$ is the non-filtered signal, and $a$ and $b$ are the filter coefficients. It can be shown that convolution (i.e. smoothing and sharpening), correlation, differentiation, integration and discrete Fourier transformation may all be represented by digital filters, differing only in the filter coefficients.

The applications of the methods of signal analysis in measurement technique are very widespread. One of the main fields of application is signal analysis in the frequency domain and on the basis of correlograms, another main field is the improvement signal-to-noise ($S/N$) ratio by smoothing the signal. A further important application is resolution enhancement by sharpening (where resolution is reduced by the instrument function), and the accentuation of certain curve features by differentiation or integration.

*Signal transformations*

Modern chemical measurement methods and evaluation procedures involve an ever increasing panoply of mathematical techniques and of transformations, in particular.

The concept of mathematical transformations is very broad and it is difficult to define unambiguously. By transformation we mean such a projection between sets as generally assigns to geometric patterns, functions or sets of numbers another pattern, function, number or set of numbers, or vector.

Mathematical transformations may be classified in different ways. For example, on the basis of the connected sets of a transformation one can distinguish geometric transformations, which assign geometric sets to geometric sets; function transformations, which assign functions to functions; vector–vector transformations which assign sets of numbers to sets of numbers; function–vector transformations, assigning vectors to functions, etc.

Classification may also be done on the basis of the purpose of the transformation. Accordingly, there are feature extraction transformations, in which to a data set which is in itself difficult to interpret (measured points, curve, etc.) generally fewer or more easily interpretable pieces of data are assigned, in a given manner. Such methods are, for instance, factor analysis, the Karhunen–Loeve transformation, the method of moments, or the curve fitting techniques, in which the appropriately chosen parameters of the fitted curve type are adjusted to the experimental points. Differentiation can also be regarded as feature extraction transformation when, for instance, a curve is replaced by its derivative, which reflects the changes in the process much more clearly.

Coordinate transformation is involved when, instead of the results of the measurement themselves, a function (logarithm, exponential, etc.) of these data is plotted for the sake of better comprehension. Moreover, coordinate transformation is applied in certain cases to simplify further calculations, e.g. when an exponential calibration curve is brought into linear form by calculating the logarithms.

Transformation may also be a part of a measurement technique, when the feature in question is a transform of the signal actually measured. An example is Fourier transform infrared spectroscopy, when the spectrum is obtained by Fourier transformation of the directly measured interferogram.

Of course, classification according to goals is not unambiguous. Fourier transformation may also be used to simplify calculations, e.g. in the solution of a differential equation, and it may be classified as a feature extraction transformation process when it is applied to the determination of characteristic frequencies.

Series expansions according to various sets of functions can also be regarded as transformations. In such cases a set of functions $\{\varphi_1, \ldots, \varphi_k, \ldots\}$ is given. Expansion of function $f(t)$ into this set of function is defined as the determination of coefficients $a_1, a_2, \ldots, a_k, \ldots$ for which the equation

$$y(x) = \sum_{k=1}^{\infty} a_k \varphi_k(x)$$

holds, or, if the equation cannot be satisfied, the expression

$$\int_{-\infty}^{\infty} \left[ y(x) - \sum_{i=1}^{\infty} a_i \varphi_i(x) \right]^2 \mathrm{d}x$$

is a minimum. Of course, other conditions can also be conceived, for instance, the condition that at a given point, the substitution values of function $y(x)$ and the approximating function $\sum_{i=1}^{\infty} a_i \varphi_i(x)$ and the values of their derivatives should be equal (Taylor series). In several cases the function $y(x)$ is approximated, rather than the infinite series, by a finite slice, $\sum_{k=1}^{N} a_k \varphi_k(t)$, of the given function. It may also happen that the function set $\{\varphi\}_{k=1}^{\infty}$ is uncountable, i.e. it cannot be characterized by natural numbers $i = 1, 2, \ldots$, but only by a continuously varying parameter, e.g. $\{\varphi_v\}$, $v \in (-\infty, \infty)$. In this case the above sum is replaced by an integral, and numbers $a_k$, $k = 1, 2, \ldots, N$ by function $a(v)$, and then

$$y(x) = \int_{-\infty}^{\infty} a(v) \varphi_v(x) \mathrm{d}v$$

Set $\{a_k\}$, $k = 1, 2, \ldots, N$ or function $a(v)$, can be termed the expansion or transform, respectively, of function $y(x)$ into system $\varphi$. Expansion is therefore a function–vector or function–function transformation. Depending on the system of functions $\{\varphi_k\}$, various transformations exist.

Expansions or transformations into orthogonal sets of functions are particularly convenient, since in this case the determination of the individual coefficients $a_i$ or functions $a(v)$ can be performed independently. Two functions, $\varphi_i(x)$ and $\varphi_j(x)$, are called orthogonal over interval $D$ if

$$\int_D \varphi_i(x) \varphi_j(x) \mathrm{d}x = 0$$

If for a function

$$\int_D \varphi_k^2(x)\mathrm{d}x = 1$$

$\varphi_k(x)$ is called orthogonal. If the elements of system $\{\varphi_k\}$ are normalized and pairwise orthogonal, they form an orthonormal set. In this case the values of $a_k$, as mentioned, can be evaluated independently, by means of the equation

$$a_k = \int_D y(x)\varphi_k(x)\mathrm{d}x$$

or the equation

$$a(v) = \int_D y(x)\varphi_v(x)\mathrm{d}x$$

If set $\{\varphi_k\}$ is a trigonometric system, i.e.

$$\{\varphi_k\} = \{a, \cos 2\pi kt, \sin 2\pi kt\}_{k=1}^{\infty}$$

the operation is called Fourier expansion or Fourier transformation. The latter transformation is, of course, not the only possibility for orthogonal transformation.

For example, set $\{\varphi_k\}$ may also be an orthogonal polynomial system, in which case the transformation is an orthogonal polynomial transformation. If $\{\varphi_k\}$ is a polynomial system but it is nonorthogonal, the operation is called polynomial expansion. A practical drawback of such a transformation is that the coefficients $a_k$ cannot be determined independently, and thus it leads to numerically less convenient, less tractable and more unstable methods.

Recently, an increasingly wide application has been found for transformations according to orthonormal systems in which the elements of the system are the sign functions, i.e. step functions assuming the values of $+1$ and $-1$ or $c$, $-c$ and 0 only (where $c$ is constant). There are several full orthogonal systems consisting of such step functions. In practical applications, the most well known of them are the Walsh, Handamard and Haar systems. Transformations, which bear the same names, are performed on the basis of these functions.

The transformation which brings the transformed pattern, function or vector back into the original form is called the inverse transformation. However, in many instances, the inverse transformation does not exist. For instance, feature extraction transformation have generally no inverse, thus one cannot determine the calibration points from the calibration curve. It also happens that the inverse transformation exists but it is not unique. For example, an original curve can be restored from its derivative by the inverse transformation, integration, but only with the uncertainty of an additive constant.

*Feasibility checking*

The essence of feasibility checking (feasibility test) is to decide whether the deviation of a given point of a set of measured signal points from its neighbours can be regarded as a random or an informative change in the signal or whether it is due to a (presumably exceptional) error. The aim of feasibility checking is thus the rejection of data which do not belong to the 'useful' signal, arising from an outlier due to some major disturbance.

The basis of feasibility checking is the assumption that a signal set which is subject to random errors is 'monotonous', i.e. the difference between a given point and the ones preceding and following it may exceed a given value only in the case of outliers. In the following discussion we assume an equidistant series of signals.

Feasibility checking has several variants, depending on the number of points in the vicinity of the point to be investigated and on the way these points are taken into account.

The simplest, most common check of the feasibility of a point is based on one preceding point. One may here investigate simply the absolute value of the difference, i.e. if point $y_{j-1}$ is feasible (in this case the point $y_1$ is generally regarded as feasible), then point $y_j$ is feasible if

$$|y_{j-1}-y_j| \leq a$$

where $a$ is a constant, depending on the problem, and point $y_j$ is an 'outlier' if

$$|y_{j-1}-y_j| > a$$

In this case the simplest correction is

$$y_j \equiv y_{j-1}$$

Of course, it is assumed here that the signal $y(x)$ is sampled at a sufficient frequency, i.e. the distance $x_{j-1}- x_j$ is much less than the peak width of a feature of the function.

Instead of the above absolute differences it is sometimes more reasonable to use relative differences. Point $y_j$ is then feasible if

$$\left|\frac{y_{j-1}-y_j}{y_{j-1}}\right| \leq a$$

otherwise it is an outlier. The correction may be the same as in the previous case.

More sophisticated but more reliable feasibility tests are those in which several neighbours are taken into consideration. The simplest form of this test is when point $y_j$ is accepted if

$$\left|\frac{y_{j-1}+y_{j+1}}{2}-y_j\right| \leq a$$

otherwise it is rejected. The outlier is then replaced by the correction

$$y_j \equiv \frac{y_{j-1}+y_{j+1}}{2}$$

In the general case the feasibility of several points is tested simultaneously by taking account not of one but of several preceding and following points. Since, however, the aim of feasibility checking is the removal of outliers, the elimination of large discrepancies the application of quite simple methods is generally sufficient in practice.

### 6.2.2 Fourier transformation

*Concept of Fourier transformation*

Fourier transformation, which has been known for a long time and has had a number of theoretical and practical applications, has become very important in the last decade in chemical measuring techniques.

Fourier transformation is a function–function transformation, $\mathcal{F}[.]$, which assigns function $F(v)$ to a function $f(t)$ according to the following equation:

$$\mathcal{F}[f(t)] = F(v) = \int_{-\infty}^{\infty} f(t)\, e^{-2\pi i v t}\, dt$$

where $i = \sqrt{-1}$ is the imaginary unit.

The definition of the inverse Fourier transformation, $\mathcal{F}^{-1}[.]$, is the equation

$$\mathcal{F}^{-1}[F(v)] = f(t) = \int_{-\infty}^{\infty} F(v)\, e^{2\pi i v t}\, dv$$

It should be noted that functions $f(t)$ and $F(v)$ may also be complex. Utilizing the fact that

$$e^{-2\pi i v t} = \cos(2\pi v t) - i \sin(2\pi v t)$$

the above equations can be written in the forms:

$$F(v) = \int_{-\infty}^{\infty} f(t) \cos(2\pi v t)\, dt - i \int_{-\infty}^{\infty} f(t) \sin(2\pi v t)\, dt$$

and

$$f(t) = \int_{-\infty}^{\infty} F(v) \cos(2\pi v t)\, dv + i \int_{-\infty}^{\infty} F(v) \sin(2\pi v t)\, dv$$

When $f(t)$ is a real function, the above equations imply that the real part of the Fourier transform $F(v)$ of function $f(t)$ is

$$\mathrm{Re}(v) = \int_{-\infty}^{\infty} f(t) \cos(2\pi v t)\, dt$$

and the imaginary part is

$$\mathrm{Im}(\nu) = -\int_{-\infty}^{\infty} f(t) \sin(2\pi\nu t)\,\mathrm{d}t$$

Consequently,

$$F(\nu) = \mathrm{Re}(\nu) + i\,\mathrm{Im}(\nu)$$

It can be seen from the above equation that, if $f(t)$ is an even function, then $\mathrm{Im}(\nu) = 0$, and if it is odd, $\mathrm{Re}(\nu) = 0$. Therefore $\mathrm{Re}(\nu)$ may be called even and $\mathrm{Im}(\nu)$ called odd Fourier transform. The same holds for the inverse Fourier transformation.

*The physical interpretation of Fourier transformation*

One of the reasons for the wide application of Fourier transformation is the relationship of physical importance according to which, if $f(t)$ is a function of a time-intensity character, then the Fourier transform $F(\nu)$ of $f(t)$ is a frequency-amplitude type function, i.e. $F(\nu)$ characterizes the spectrum of $f(t)$. Function $F(\nu)$ gives the resolution of $f(t)$ into sine and cosine functions. It can be seen that $f(t)$ can be given as a superimposition of sine and cosine functions of frequency $\nu$ where $\nu \in (-\infty, \infty)$, taken with different weights, since integration can be regarded as infinitely fine summation. The weights of components of frequency $\nu$ in the sum (integral) give $F(\nu)$, i.e. just the Fourier transform of function $f(t)$. It should be noted that argument $t$ of function $f(t)$ may also be other than time, e.g. it may represent distance. The physical dimension of variable $\nu$ is always the reciprocal of the dimension of $t$.

According to the above, $f(t)$ is called a time domain function, whereas $F(\nu)$ is a frequency domain function, i.e. Fourier transformation is a transformation between time and frequency domains. According to a more exact description of complex Fourier transformation, the amplitude of frequency $\nu$ in the Fourier spectrum of function $f(t)$ is

$$A_f(\nu) = |F(\nu)| = \mathrm{sqr}\left(\mathrm{Re}^2(\nu) + \mathrm{Im}^2(\nu)\right)$$

and its phase is

$$\Theta(\nu) = \mathrm{atan}\left(\mathrm{Im}(\nu)/\mathrm{Re}(\nu)\right)$$

Fourier transforms are also defined mathematically for negative frequencies, therefore the actual amplitude is a sum of the values corresponding to the appropriate positive and negative frequencies. The Fourier spectrum of a real function is symmetrical.

*Some important properties of Fourier transformation*

Some properties of Fourier transformation that are of practical importance are as follows:

(a) $\mathcal{F}$ and $\mathcal{F}^{-1}$ are linear transformations, therefore the principle of superimposition holds, i.e.

$$\mathcal{F}[cf(t)] = c\,\mathcal{F}[f(t)]$$

$$\mathcal{F}[f_1(t)+f_2(t)] = \mathcal{F}[f_1(t)] + \mathcal{F}[f_2(t)]$$

(b) Parseval (energy conservation) theorem:

$$\int_{-\infty}^{\infty} f^2(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} |F(v)|^2\,\mathrm{d}v$$

(c) The amplitude spectrum:

$$A_f(v) = \operatorname{sqr}\left(\mathrm{Re}^2(v) + \mathrm{Im}^2(v)\right)$$

is invariant with respect to time shifts in $f(t)$, i.e. to the phase.

(d)

$$\mathcal{F}^{-1}(F) = \overline{\mathcal{F}(\bar{F})}$$

where the vinculum denotes complex conjugation. (Expression of inverse transformation with direct transformation.)

(e)

$$\mathcal{F}[f_1 * f_2] = \mathcal{F}[f_1] \cdot \mathcal{F}[f_2]$$

where $*$ denotes convolution operation, i.e.

$$f_1(t) * f_2(t) = \int_{-\infty}^{\infty} f_1(t-\tau) f_2(\tau)\,\mathrm{d}\tau$$

This property implies that simple multiplication in the frequency domain corresponds to a convolution in the time domain and vice versa; to a convolution in the frequency domain there is a corresponding multiplication in the time domain, i.e.

$$\mathcal{F}[f_1] * \mathcal{F}[f_2] = \mathcal{F}[f_1 \cdot f_2]$$

(f) Differentiation, integration:

$$\mathcal{F}\left[\frac{\mathrm{d}^n f(t)}{\mathrm{d}t^n}\right] = (i2\pi v)^n \mathcal{F}[f(t)]$$

multiplication in the frequency domain corresponds to differentiation in the time domain, whereas

$$\mathcal{F}\left[\int f(t)\,\mathrm{d}t\right] = (i2\pi v)^{-1} \mathcal{F}[f(t)]$$

i.e. integration in the time domain corresponds to division in the frequency domain.

*Practical techniques of Fourier transformation*

Fourier transformation relates to functions that are continuous in the region $(-\infty, \infty)$, and the transformed function has the same property. In practical cases, however, one has only discrete values sampled at intervals $\Delta t$ in a finite $[0, T]$, and the resulting transform also consists of discrete values in a limited region only. It is obvious that this difference will cause discrepancies, and the resulting transformed values will differ from the theoretical values. Following on from the above, these discrepancies arise for two reasons.

One of the reasons is that the region of integration is, rather than $(-\infty, \infty)$, a finite region $[0, T]$ or $[-T, T]$. This is equivalent to a situation in which the $f(t)$ function defined over the true region $(-\infty, \infty)$ is regarded as having a zero value outside the region $[0, T]$, i.e. it is multiplied by a window function $w_{0,T}(t)$ which has a value of unity in the region $[0, T]$ and zero everywhere else. This multiplication in the time domain corresponds to a convolution in the frequency domain, which causes broadening of the spectral lines and the appearance of side lobes on the lines. This effect is called leakage. The extent of leakage (or line broadening) can be decreased by increasing the sampling region ($T$) or by the application of special window functions.

It is important to note that no distortion occurs in the resulting discrete frequencies if the function to be transformed is periodic and the sampled region coincides with the period. Consequently, $T$ should be selected carefully.

It should be noted that $T$, owing to the band broadening effect, also determines resolution. In this regard we should state that the condition of the distinction between frequencies $\nu_1$ and $\nu_2$ is that

$$T \geq 1/(\nu_1 \nu_2)$$

The other reason for the distortions is the discrete sampling at intervals $\Delta t$. This sampling of function $f(t)$ corresponds to a multiplication of $f(t)$ by a sampling function, which is transformed into the frequency domain as a convolution of $F(\nu)$ with the Fourier transform of the sampling function. This is equivalent to a summation of the shifted $F(\nu)$ functions, where the size of the shift is $1/\Delta t$. Consequently, if the spectrum of the original function contains frequencies higher than $1/(2\Delta t)$, a distortion, called folding, occurs in the case of discrete Fourier transformation: the amplitudes of frequencies greater than $1/(2\Delta t)$ are folded over those of the frequencies less than $1/(2\Delta t)$. This distortion is also known as aliasing.

The distortion can be reduced by decreasing the sampling interval, i.e. by increasing the number of data points. It can be completely eliminated if the sampling frequency is at least twice the highest frequency in the spectrum of the function, i.e. elimination is possible only with band limited functions. In the case of such a sampling, the resulting discrete Fourier transformed values correctly represented the full function $f(t)$. This is Shannon's sampling theorem. To avoid

this sampling problem, in some cases high frequency filtering is performed on the function before Fourier transformation.

In the case of a sampling interval of $\Delta t$ over a region of $[0, T]$, the recognizable frequencies are the fundamental frequency $1/T$ and its overtones $2/T$, $3/T$, ..., $1/(2\Delta t)=J/(2T)$, where $J$ is the number of sample points. The correct choice of $T$ is also very significant for this reason. $J$, $\Delta t$ and $T$ are connected by the equations $T=J\Delta t$ or $1/T=1/(J\Delta t)$.

Utilizing the properties of pulse function $\delta(t)$, the following relationship of the discrete Fourier transformation can be set up for the discrete points of function $f(t)$ sampled at intervals $\Delta t$ in the region $[0, T]$:

$$F(\nu)=\Delta t \sum_{j=0}^{J-1} \int_{0}^{T} f(t)\delta(t-j\Delta t)\cdot e^{-2\pi i\nu t}\,dt=\Delta t \sum_{j=0}^{J-1} f_j \cdot e^{-2\pi i\nu_j \Delta t}$$

where

$$f_j=f(j\Delta t)$$

i.e.

$$F(\nu_i)=\Delta t \sum_{j=0}^{J-1} f_j \exp(-i2\pi\nu_i t_j)$$

or, in a different form,

$$F_i=\Delta t \sum_{j=0}^{J-1} f_j(\cos(2\pi\nu_i t_j)-i\sin(2\pi\nu_i t_j))$$

where $f_j$ is the $j$th sampled data point ($j=0, 1, \ldots J-1$),

$F_i$ is the transformed value corresponding to the $i$th frequency, $\nu_i$, where $\nu_i=i/T$ ($i=0, \pm 1, \ldots, \pm J/2$),

$t_j$ is the time of $j$th sampling; $t_j=j\Delta t$, and

$J$ is the number of points and $T=J\Delta t$.

The discrete inverse Fourier transform is obtained on the basis of similar considerations:

$$f_j=(1/\Delta t)\sum_{i=0}^{J-1} F_i \exp(+i2\pi ij)$$

or

$$f_j=(1/\Delta t)\sum_{i=0}^{J-1} F_i[(\cos(2\pi ij)+i\sin(2\pi ij))]$$

which is clearly identical to the Fourier series of function $f$.

### 6.2.3 Walsh and Haar transformations

*Walsh functions*

The Walsh functions represent an ordered set of square waves. All functions may assume the values of $+1$ and $-1$, and they form an orthogonal set of functions. They are defined over a time range of length $T$. They are characterized by two

arguments, period t, which is generally given in a form normalized to $T$, i.e. as $t/T$ (thus periods vary over the range [0, 1]), and order n. The Walsh functions are denoted as $\mathrm{WAL}(n,t)$, and for most purposes they are ordered according to the ascending number of zero crossings.

One of the expressions of the functions for $N=2$ terms is

$$\mathrm{WAL}(n,t)=\mathrm{WAL}(n_{p-1}, n_{p-2}, \ldots, n_0; t_{p-1}, t_{p-2}, \ldots, t_0)= \\ =\prod_{r=0}^{p-1}(-1)^{n_{p-1-r}\cdot(t_r+t_{r+1})}$$

where $n$ and $t$ are expressed as binary numbers and $n_{p-1}, n_{p-2}, \ldots, n_0$ and $t_{p-1}, t_{p-2}, \ldots, t_0$ are their binary digits.

It can be shown that these functions are orthogonal, i.e.

$$\sum_{t=0}^{N-1}\mathrm{WAL}(m,t)\cdot\mathrm{WAL}(n,t)=\begin{cases}N & \text{if } n=m\\ 0 & \text{if } n<>m\end{cases}$$

and if the functions are all divided by $N$, they form an orthonormal system.

Introducing the following notation for the classification of functions on the basis of even or odd wave symmetry

$$\mathrm{WAL}(2k,t)=\mathrm{CAL}(k,t)$$

$$\mathrm{WAL}(2k-1,t)=\mathrm{SAL}(k,t)$$

we may then define two further Walsh series, which resemble the cosine and sine series.

The Walsh functions are symmetric about the centre of the region. If the set of the independent variables is defined as $-1/2<t<+1/2$, the functions have either mirror symmetry (CAL functions) or central symmetry (SAL functions) corresponding to the symmetry relationship

$$\mathrm{WAL}(n,t)=\mathrm{WAL}(t,n)$$

This has the advantage that transformation and its inverse can be performed by the same mathematical operation, which simplifies applications.

The order of Walsh functions discussed above is called a sequency order (Walsh order, Walsh–Harmuth–Kaczmarz order).

The concept of sequence can be defined as one half of the average number of zero crossings occurring in unit time, which is thus independent of frequency. Consequently, sequency is a special case of frequency, which is applied to sinusoidal waves. Over the region $-1/2<t<+1/2$ the number of zero crossings is $2k$, thus in the SAL or CAL arrangement $k$ means sequency.

Another frequently used order of Walsh functions is the natural order (normal, diadic, binary or Paley order). This order was first applied by Paley, the notation

of function arranged in this order is PAL($n$, $t$). These functions can also be defined as the eigenfunctions of the logical differential operator.

The natural order is applied mainly in theoretical mathematics and in image transmission, whereas the sequency order is applied primarily in telecommunication and signal processing techniques related to spectral analysis and filtering.

The relationship between WAL and PAL series can be expressed as

$$\mathrm{PAL}(g(k), t)=\mathrm{WAL}(k, t)$$

If order $k$ is expressed as a binary number

$$k=(k_p, k_{p-1}, \ldots k_1, k_0)_2$$

then $g(k)$ can be obtained by the modulo-2 summation

$$g_i(k)=k_i \oplus k_{i+1}$$

For example, if $k=5_{10}=0101_2=k_3, k_2, k_1, k_0$, then

$$g_1(k)=1 \oplus 0=1$$

$$g_2(k)=0 \oplus 1=1$$

$$g_3(k)=1 \oplus 0=1$$

thus $g(k)=111_2=7_{10}$, and therefore PAL(7, $t$)=WAL(5, $t$).

*Walsh transformation*

All functions $f(t)$ which can be integrated over the open interval (0, 1) can be represented by a Walsh series defined on the open interval (0, 1):

$$f(t)=a_0\mathrm{WAL}(0, t)+a_1\mathrm{WAL}(\mathrm{i}, t)+\ldots$$

i.e.

$$f(t)=\sum_{k=1}^{\infty} a_k\mathrm{WAL}(k, t)$$

where the coefficients are given by equation

$$a_k=\int_0^1 f(t)\mathrm{WAL}(k, t)\mathrm{d}t$$

and

$$\mathrm{WAL}(0, t)=\mathrm{i}$$

Thus the following pair of transformations can be defined:

$$f(t)=\sum_{k=0}^{\infty} F(k)\mathrm{WAL}(k, t)$$

and

$$F(k)=\int_0^1 f(t)\mathrm{WAL}(k, t)\mathrm{d}t$$

This definition is applicable to continuous, limited functions where the time is $0<t<1$. In numerical transformation $N$ equidistant points which are on the continuous function must be selected from the region [0, 1]. $N$ should be a power of two, $N=2^P$. Then integration may be replaced by summation. Applying the trapezoid rule for the $N$ sample points $x_i$, the pair of discrete Walsh transformations can be given as

$$X_n=(1/N)\sum_{i=0}^{N-1} x_i \mathrm{WAL}(n, i) \qquad n=0, 1, 2, \ldots N-1$$

$$x_i=\sum_{n=0}^{N-1} X_n \mathrm{WAL}(n, i) \qquad i=0, 1, 2, \ldots N-1$$

Similar transformations can be obtained by applying CAL and SAL functions on the series $x_i$:

$$X_C(k)=(1/N)\sum_{i=0}^{N-1} x_i \mathrm{CAL}(k, i)$$

and

$$X_S(k)=(1/N)\sum_{i=0}^{N-1} x_i \mathrm{SAL}(k, i)$$

The transformations are linear and it thus follows from $x_i \leftrightarrow X_n$ and $y_i \leftrightarrow Y_n$ that $ax_i+by_i \leftrightarrow aX_n+bY_n$, where $a$ and $b$ are real constants and $\leftrightarrow$ represents the transform pair.

The transformation is also symmetrical. $\mathrm{WAL}(n, i)$ is symmetrical about the centre of sequency if $n$ is even and antisymmetric if $n$ is odd $[i=0, 1, 2, \ldots (N-1)]$. Thus, if the set of $x_i$ is axially symmetric about the centre, the transformation uses Walsh coefficients of even order only (CAL function), and if it is centrally symmetric, only the odd order coefficients are finite (SAL function).

Walsh transformation can be performed by a simple matrix multiplication. Since the matrices of a Walsh transformation are symmetric, transformation and its inverse are identical except for a factor of $1/N$, in contrast to Fourier transformation.

*Haar transformation*

The Haar functions, too, form a complete, orthonormal system of functions consisting of square waves. They have a number of important properties; thus, for example, they are capable of approximating a given function by a small number of terms with high precision. They can assume three values: 0, $A$ and $-A$, where $A$ is a function of $\sqrt{2}$. They can be defined over the region $0 \leq t \leq 1$. Their notation is $\mathrm{HAR}(n, t)$, where $n$ is related to the number of zero crossings similar to the Walsh functions. For a full definition, information must also be given on the amplitude of the function. A definition may be, for instance,

$$\mathrm{HAR}\,(2^p+n,t)=\begin{cases}\mathrm{sqr}_{(2^p)} & \text{if } n/2^p\le t<\left(n+\frac{1}{2}\right)\Big/2^p\\ -\mathrm{sqr}_{(2^p)} & \text{if } \left(n+\frac{1}{2}\right)\Big/2^p\le t\le(n+1)/2^p\\ 0 & \text{otherwise}\end{cases}$$

where $p=1, 2, \ldots$ and $n=0, 1, \ldots 2^{p-1}$.

The functions are orthonormal:

$$\int_0^1 \mathrm{HAR}\,(m,t)\,\mathrm{HAR}\,(n,t)\,\mathrm{d}t=\begin{cases}1 & \text{if } m=n\\ 0 & \text{if } m<>n\end{cases}$$

Haar functions converge better than Walsh functions. The Walsh expansion of some continuous functions may be divergent in certain points; this may not occur with Haar expansions. It should be noted that a relationship between Haar and Walsh functions can be derived.

A given continuous function $f(t)$ can be expanded into Haar functions in the region $0<t<1$ by means of the equation

$$f(t)=\sum_{n=0} C_n\mathrm{HAR}(n,t)$$

where

$$C_n=\int_0^1 f(t)\mathrm{HAR}(n,t)\mathrm{d}t$$

From these equations, the discrete Haar transformation and its inverse can be written as follows

$$X_n=(1/N)\sum_{i=0}^{N-1} x_i\mathrm{HAR}(n,i/N)$$

$$x_i=\sum_{n=0}^{N-1} X_n\mathrm{HAR}(n,i/N)$$

where $i, n=0, 1, \ldots N-1$.

The above equations can be expressed in matrix form as

$$X=(1/N)\mathbf{H}x$$

$$x=\mathbf{H}^{-1}X$$

where $\mathbf{H}$ and $\mathbf{H}^{-1}$ are the matrices of direct and inverse Haar transformations.

The matrices are not symmetrical, unlike those of Walsh transformation, and thus separate transformation operations are necessary for the forward and inverse transformations.

*Applications of Walsh and Haar transformations*

Walsh and Haar transformations are applied in several fields: in spectral resolution, frequency filtering, telecommunication and in various branches of signal processing, such as pattern recognition, image formation, speech analysis and processing, analysis of diagnostic signals, etc. The fields of application in chemical data processing, however, have only partially been clarified.

A theoretically interesting and thus extensively studied area is the possibility of the synthesis of various waveforms by means of Walsh and Fourier transformations, and the comparison of these transformations. According to these investigations, in the case of continuous functions, for waveform synthesis with a given accuracy, approximately twice as many Walsh terms are required as Fourier terms. This disadvantage is, however, balanced by the significant reduction of calculation times, and thus, if rapidity is required, the application of Walsh functions may be more advantageous.

In the synthesis of square waves or other noncontinuous signals, Walsh transformation may be decisively more advantageous than Fourier transformation. This was the reason why Walsh transformation was first suggested for the spectral analysis of noncontinuous and transient waveforms.

### 6.2.4 Convolution and correlation

*Convolution*

The operation called convolution and denoted by $*$, involving functions $y_1(x)$ and $y_2(x)$, is defined as

$$y(x)=y_1(x)*y_2(x)=\int_{-\infty}^{\infty} y_1(\tau)y_2(x-\tau)\mathrm{d}\tau$$

According to this definition, function $y(x)$ is the convolution of functions $y_1(x)$ and $y_2(x)$.

On the basis of the above equation, the convolution operation consists of the following steps:

1. Take the mirror image of $y_2(\tau)$ about the ordinate axis.
2. Shift $y_2(-\tau)$ by the amount $x$.
3. Multiply the shifted function $y_2(x-\tau)$ by $y_1(\tau)$.
4. Integrate the area under the product $y_1(\tau)y_2(x-\tau)$.

The area under the product $y_1(\tau)y_2(x-\tau)$ is the value of the convolute at point $x$.

Convolution is a symmetrical operation, i.e.

$$\int_{-\infty}^{\infty} y_1(\tau)y_2(x-\tau)\mathrm{d}\tau=\int_{-\infty}^{\infty} y_1(x-\tau)y_2(\tau)\mathrm{d}\tau$$

The discrete form of convolution is

$$y(j)=\sum_{k=0}^{K-1} y_1(k)y_2(j-k)$$

*Smoothing*

The aim of smoothing is the reduction of the effect of random noise on the signal, i.e. the suppression of noise. It is based on a principle according to which it is assumed that the random noise superimposed on a measured signal is alternating in sign, and thus, by taking into account several points simultaneously, its effect can be 'averaged'.

The simplest form of smoothing is boxcar integration, in which a certain number of the elements of data set $y_1, \ldots, y_J$ are added, thereby producing a set of points which are 'looser' than the original, but smoother at the same time. With this method the value of the $j$th smoothed point is

$$\tilde{y}_j=1/(2m+1)\sum_{k=-m}^{m} y_{j-k} \qquad j=1, \ldots, J$$

i.e. instead of the $J$ original points $y_1, \ldots, y_J$, only the points $y_m$, $y_{m+(2m+1)}$, $y_{m+2(2m+1)}, \ldots, y_{J(2m+1)}$ i.e. only $J/(2m+1)$ points remain.

The above procedure can also be carried out by summing not only about every $(2m+1)$st point but about all points. The simple moving average is a smoothing technique in which the smoothed value of the $j$th point can be obtained from the expression

$$\tilde{y}_j=1/(2m+1)\sum_{k=-m}^{m} y_{j-k}$$

but in this case $j=m, \ldots, J-m$ includes all points, i.e. $J-2m$ smoothed points are obtained.

Thus, in the simple moving average algorithm the values of $2m+1$ data points are added, divided by the number of points, and the result is assigned to the middle member of the series of points as its smoothed value. In the next step one point is dropped from one end of the series and another is taken at the other end, and the procedure is repeated while there are still any data points remaining.

This method is very simple and often satisfactory, but it has some disadvantages. It does not preserve the original shape of the signal since it smooths the sharp jumps. Therefore in those types of measurement in which sharp changes of signal can be expected, e.g. in liquid chromatography, the peaks are distorted and flattened they take the shape of the smoothing function. This reduces the resolution of the signal, thus the use of the simple moving average is not recommended here.

By using a weighted moving average the disadvantages of the simple average can be significantly reduced. In this case the central and directly adjacent points can be taken into account with higher weights than the points on the (mowing) wings. The

weighted moving average is calculated by the expression

$$\tilde{y}_j = \frac{\sum_{k=-m}^{m} c_k y_{j-k}}{\sum_{k=-m}^{m} c_k} = \sum_{k=-m}^{m} a_k y_{j-k}$$

where the values $c_k$ or $a_k$ are the appropriate weights. The smoothed value of data point $j$ is therefore a function of $2m+1$ $y_j$ values and the applied coefficients $c_k$, i.e. it is the weighted average of $2m+1$ points.

Smoothing, or noise filtering, can also be performed by polynomial fitting. The polynomial that best fits the given set of data points can always be found. The order of the polynomial and the number of data points to be taken into account for smoothing should be decided in advance. When these numbers are fixed, the coefficients of the polynomial chosen can be determined by the method of least squares. With these coefficients the squares of the differences between the data points and the ordinates calculated from the polynomial at the same abscissa have the smallest sum.

The smoothed value at a given point is the value of the polynomial at the given abscissa. The algorithm is continued in the same way as in the case of the weighted moving average method. Consequently, the smoothed value is an appropriate linear combination of the data points considered, where the coefficients are independent of the values of the basis points, depending only on the number of these points and the order of the polynomial. Determining these coefficients, they will be the weights $a_k$ in equation

$$\tilde{y}_j = \sum_{k=-m}^{m} a_k y_{j-k}$$

i.e. smoothing by polynomial fitting is a special case of the weighted moving average method. An advantage of polynomial fitting is that it influences signal profile and intensity only to a minor extent.

It can be seen that all smoothing methods correspond to a special convolution, in which the smoothed function is a convolution of the original function and the weighting function.

*Auto- and cross-correlation functions*

In addition to convolution, the operation of correlation has an important role, too. Correlation between functions $y_1(x)$ and $y_2(x)$ can be defined as

$$y(x) = \int_{-\infty}^{\infty} y_1(\tau) y_2(x+\tau)\, d\tau$$

or

$$y(j) = \sum_{k=0}^{K-1} y_1(k) y_2(j+k)$$

The operation of correlation consists of the following steps.

1. Shift $y_2(\tau)$ by the amount $x$
2. Multiply the shifted function $y_2(x+\tau)$ by $y_1(\tau)$
3. Integrate the area under the product $y_2(x+\tau)y_1(\tau)$ The area under the product $y_2(x+\tau)y_1(\tau)$ is the value of the correlation at point $x$.

The similarity between correlation and convolution is obvious. For the special case when either $y_1(x)$ or $y_2(x)$ is an even function, convolution and correlation are equivalent.

If $y_1(x)$ and $y_2(x)$ are different, $y(x)$ is the cross-correlation of these functions, if they are the same, $y(x)$ is called the autocorrelation function of $y_1(x)$. The autocorrelation function is even, and has a maximum at $x=0$. The function shows the periodicity or correlations within the signal, i.e. it shows how close the relationship is between the values of the function at points $x$ and $x+\tau$.

### 6.2.5 Curve fitting

*Curve fitting by the method of least squares*

In chemical measurements a task that frequently occurs is the mathematical modelling of a set of observed, measured data, the determination of the type and parameters of a function that properly fits the data points. This is the problem, for instance, in modelling signals, for example in the approximation of chromatographic peaks or calibration curves.

Curve fitting and parameter estimation can be discussed at a high level of mathematical sophistication see Chapter 2. However, in this section only such problems are discussed as do not require the assumptions of mathematical statistics.

Curve fitting by the method of least squares means that the measured signal $y(x)$ is approximated by a function $f(x, p)$ so that parameter $p$ of the function is chosen to minimize the squares of differences, $S(\underline{p})$, which can be defined as

$$S(\underline{p}) = \int_{x_L}^{x_U} w(x)\,(y(x)-f(x,p))^2\,dx$$

in the continuous case, or

$$S(p) = \sum_{j=1}^{J} w_j(y_j - f(x_j, p))^2$$

in the discrete case, where $w(x)$ or $w_j$, respectively, is a weight function.

The determination of the $k$ components $p_1, p_2, \ldots p_k$ of parameter vector $p$ on the basis of the least squares criterion is a minimization problem with $k$ parameters. For an arbitrary function $f(x, p)$ this problem can generally by solved by numerical methods only. Such methods are, for instance, the gradient method,

the Nelder–Mead (modified simplex), Newton–Raphson, Powell methods, etc. the choice of which depends on the type of the function to be fitted.

In some cases one may start from the fact that to satisfy the condition

$$\min S(\underline{p})$$

the derivative of $S(\underline{p})$ with respect to $p$ should be the 0 vector, i.e. the partial derivatives of $S(\underline{p})$ with respect to the $p_k$ parameters are zero:

$$\partial S(\underline{p})/\partial \underline{p} = \underline{0}$$

This system of conditions leads to a system of equations of $k$ unknowns that is generally nonlinear in $p$, the solution of which gives an estimate $\hat{p}$ of the required parameter vector $\underline{p}$. In fortunate cases the system is easy to solve, but it often requires the application of numerical approximations, e.g. the Gauss–Newton method.

*The estimation of the parameters of linear functions*

The solution of the system of equations $\partial S(p)/\partial p = 0$ becomes particularly simple when the function $f(x, p)$ is linear in parameters $p$, i.e.

$$f(x, \underline{p}) = \underline{p}^T \underline{g}(x)$$

or, alternately,

$$f(x, \underline{p}) = p_1 g_1(x) + p_2 g_2(x) + \ldots + p_k g_k(x)$$

It is evident that linearity in $x$ is not required, since the values of $x$ are known. It is easy to see that by means of substitutions $x_1 = g_1(x)$, $x_2 = g_2(x)$, ..., $x_k = g_k(x)$ this equation transforms into

$$Y = \underline{p}^T \underline{x}$$

These equations have no constant term. This, however, involves no restriction, since if an unknown constant $c$ were included, then by introducing $x_{k+1} = 1$ and $p_{k+1} = c$, and inserting the term $p_{k+1} \mathbf{x}_{k+1}$ the above form could be obtained once again.

If, therefore, function $f(\underline{p}, x)$ is linear in $\underline{p}$, the above set of equations is simplified into the following linear system of equations with $k$ unknowns:

$$\sum_{i=1}^{N} (y_i - p^T x_i) \underline{x}_i = 0$$

i.e.

$$\sum_{i=1}^{N} (\underline{x}_i \underline{x}_i^T) p = \sum_{i=1}^{N} y_i \underline{x}_i$$

where $x_i^T$ is the transpose of $x_i$, i.e. it is a row vector.

A more convenient form can be obtained by introducing **X**, the observation matrix:

$$\mathbf{X}=\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & & & \\ x_{N1} & x_{N2} & \dots & x_{Nn} \end{bmatrix}$$

where $x_{ij}$ is the $i$th occurrence of the $j$th independent variable $\underline{x}_j$, i.e. the $i$th row of $\mathbf{X}$ is the vector denoted by $\underline{x}_i$ in the above discussion. Let us denote the measured vector $(y_1, y_2, \dots y_N)$ by $y$, thus

$$\underline{y}=\mathbf{X}\underline{p}$$

and then the form of the least squares criterion is

$$\min_{p} (\underline{y}-\mathbf{X}\underline{p})^{\mathrm{T}} (\underline{y}-\mathbf{X}\underline{p})$$

from which, by differentiation with respect to $p$, the system of equations

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\underline{\hat{p}}=\mathbf{X}^{\mathrm{T}}\underline{y}$$

can be obtained, which is equivalent to the above, scalar equation. The form of this system of equations in the traditional notation is

$$\begin{aligned}
&\sum_i x_{i1}x_{i1}p_1 + \sum_i x_{i1}x_{i2}p_2 + \dots + \sum_i x_{i1}x_{in}p_n = \sum_i x_{i1}y_i \\
&\vdots \\
&\sum_i x_{i2}x_{i1}p_1 + \sum_i x_{i2}x_{i2}p_2 + \dots + \sum_i x_{i2}x_{in}p_n = \sum_i x_{i2}y_i \\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\
&\sum_i x_{iN}x_{i1}p_1 + \sum_i x_{iN}x_{i2}p_2 + \dots + \sum_i x_{iN}x_{in}p_n = \sum_i x_{iN}y_i
\end{aligned}$$

This is usually called a Eulerian system of equations.

From these equations the unknown set of parameters can be calculated by

$$\underline{\hat{p}}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\underline{y}$$

The solution of this matrix equation becomes extremely simple when $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is a diagonal matrix. In this case it is not necessary to carry out matrix inversion, since the inverse is obtained simply by replacing the diagonal elements by their reciprocals. The matrix is diagonal if the columns of $\mathbf{X}$ are pairwise orthogonal, i.e. if their scalar products are 0. The setting up of such orthogonal experimental conditions is the aim of experiment design.

It is a further advantage of orthogonality that in this case the components of $\underline{p}$ can be determined independently, since

$$\hat{p}_i=(1/d_i)\mathbf{X}^{\mathrm{T}}\underline{y}$$

where $d_i$ is the $i$th element of the main diagonal of the diagonal matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. It

is advantageous for the stability of the method if the diagonal elements are the same, or at least not widely divergent.

In contrast, the more distant experimental set-up from the orthogonal one, i.e. the more correlated (dependent) the columns of **X**, the less stable is the solution of the above set of equations. This is the case of an ill-defined, 'ill-conditioned' system of equations, when the rounding errors and other inaccuracies of numerical calculation cause very large errors. The detection of ill-conditioning is very important in estimating the reliability of the results obtained, see Chapter 2.

*Investigation of the goodness of estimation*

After determining an estimate $\underline{\hat{p}}$ for parameter vector $\underline{p}$, the following, partially interdependent demands may occur:

(a) characterization of the goodness of fit;
(b) testing the correctness of the assumed function;
(c) characterization of the reliability of the resulting parameter set $\underline{p}$.

The answer to these questions obviously plays a key role; it may also occur that the aim of fitting was precisely the investigation of these questions.

The investigation of the goodness of parameter estimation is generally based on statistical tests. However, the methods which will be discussed below do not involve statistical assumptions, investigations or probability theory.

The goodness of fit can be characterized in a number of ways. A good measure is, for instance the value of the error squares which is to be minimized at $\underline{\hat{p}}$, i.e. the value of $S(\underline{\hat{p}})$. After the estimation of parameters, this is already easy to calculate, since $f(x, \underline{\hat{p}})$ $x$ and $\hat{p}$ are known.

The value of function $f$ with parameter $\underline{\hat{p}}$ at a given value of $x$ is called the fitted value of function $y(x)$ at $x$, i.e.

$$\hat{y}(x) = f(x, \underline{\hat{p}})$$

The difference between the measured and fitted values is called the residue or the residual error, and denoted by function $r$:

$$r(x) = y(x) - \hat{y}(x)$$

or

$$r(j) = y(x_j) - \hat{y}(x_j)$$

With the above notation,

$$S(\underline{\hat{p}}) = \int_{x_L}^{x_U} r^2(x)\,dx$$

or

$$S(\underline{\hat{p}}) = \sum_{j=1}^{N} r_j^2$$

i.e. $S(\hat{p})$ is the sum of residual squares, in other terms the 'residual error'.

To check the correctness of fit one can utilize the fact that, because of the properties of minimum of function $S(\underline{p})$, the sum of residuals is zero.

In the discrete case, a method for the investigation of the applicability of the equation which does not require parallel measured data, is based on the test of the signs of residuals, since in the case when the model is correct, the signs + and − must vary randomly. The randomness of signs can be investigated by means of appropriate statistical tests.

A graphical plot of the measured signal and the fitted curve may also be of help in the judgement of the correctness of a model. If the residuals are also plotted as a function of $x$, it can be seen whether the variation of residuals is random, which supports the correctness of the model, or is a function of $x$, which suggests that the model leads to systematic errors and distortions.

*Fitting by orthogonal polynomials*

The system of functions

$$\{\varphi_i(x)\}_0^m$$

defined over a finite set of points $x_j$ is called orthogonal for a given non-negative weight $\varrho(x)$ if

$$\sum_{j=1}^{J} \varphi_i(x_j)\varphi_l(x_j)\varrho(x_j) = \begin{cases} 0 & \text{if } i<>1 \\ k_l & \text{if } i=1 \end{cases}$$

If functions $\varphi_i(x)$ are polynomials for which the highest power of $x$ is $i$, then the system $\{\varphi_i(x)\}_{i=0}^n$ is called an orthogonal polynomial system.

Polynomials orthogonal with respect to a given weight can be calculated by means of the Gram-Schmidt orthogonalization procedure from the non-negative integer powers of $x$, or in explicit forms by using the Rodrigues equations, or by recursive procedures. Some orthogonal polynomial systems that are often used in practice are given below.

Let the given region be $[-1, +1]$ and take $\varrho(x)=(1-x)^\alpha(1+x)^\beta$, the 'almost everywhere positive' weight function, where $\alpha>-1$ and $\beta>-1$. The polynomials defined in this manner satisfy the orthogonality conditions discussed; they are called Jacobian polynomials. The cases of such polynomials that occur in practice are

| | |
|---|---|
| $\alpha=\beta=0$ | (Legendre polynomials) |
| $\alpha=\beta=1/2$ | (Chebyshev polynomials of the second kind) |
| $\alpha=\beta=1$ | (Gegenbauer polynomials) |

The Jacobian polynomials can be written in explicit form by the following Rodrigues equation:

$$(1-x)^\alpha(1+x)^\beta P_n^{(\alpha,\beta)}(x) = \frac{(-1)^n}{2^n n!}\left(\frac{d}{dx}\right)^n\left((1-x)^{n+\alpha}(1+x)^{n+\beta}\right)$$

They can be generated by the following recursive equation:

$$2n(n+\alpha+\beta)(2n+\alpha+\beta-2)P_n^{(\alpha,\beta)}(x)=$$
$$=(2n+\alpha+\beta-1)\cdot[(2n+\alpha+\beta)(2n+\alpha+\beta-2)\cdot x-\alpha^2-\beta^2]P_{n-1}^{(\alpha,\beta)}(x)-$$
$$-2(n+\alpha-1)(n+\beta-1)(2n+\alpha+\beta)P_{n-2}^{(\alpha,\beta)}(x)$$

for $n=2, 3, \ldots$

$$P_0^{(\alpha,\beta)}(x)\equiv 1$$

and

$$P_1^{(\alpha,\beta)}(x)\equiv\frac{1}{2}(\alpha+\beta+2)x+\frac{1}{2}(\alpha-\beta)$$

The polynomials generated by the weight function $\exp(-x)x^{\alpha}$ $(\alpha>-1)$ over the region $[0, +\infty)$ are called Laguerre polynomials, whereas those produced by weight function $\exp(-x^2)$ over $[-\infty, +\infty)$ are the Hermite polynomials.

The fitting of these polynomials is also carried out by the method of least squares. If the system of polynomials is $\{\varphi_i(x)\}_{i=0}^{n}$, and the experimental data are $y_j$, then the minimum of function

$$S(p)=\sum_{j=1}^{J}\left(y_j-p_j\sum_{i=0}^{k}\varphi_j(x_i)\right)^2$$

is to be determined with respect to $p_1, \ldots p_k$. The above minimum problem is solved by the methods discussed in the previous section.

### 6.2.6 Factor analysis

*The model of factor analysis*

Let us assume that the components of the measured analytical signal, as a vector, are not independent, and their values depend, linearly, on various factors. In these cases factor analysis can be applied to reveal the linear relationship between the signal, i.e. the analytical information, and the factors.

Factor analysis is a statistical method which can be used to determine the number of 'factors' underlying the measured properties under the given conditions of accuracy, i.e. the number of relevant properties, called factors, which account for the changes in the measured data.

The observed or measured $n$ properties of the system under study are characterized by the $n$-dimensional property vector $\xi=(\xi_1, \xi_2, \ldots, \xi_n)^T$, which can be regarded as a random variable. The components of $\xi$ are the various properties (e.g. the discrete points of a spectrum or chromatogram). Let us now perform measurement of $\xi$ on a large number of objects (sample, substance, etc.). The result

of the $i$th measurement is denoted by the vector

$$y_i = (y_{1i}, y_{2i}, \ldots, y_{ni})^T, \qquad i = 1, 2, \ldots N$$

They will form a data matrix

$$\mathbf{Y}[n \times N] = \begin{bmatrix} y_{11}, \ldots y_{1N} \\ \vdots \qquad \vdots \\ y_{n1}, \ldots y_{nN} \end{bmatrix}$$

(the square bracket indicates the dimension of the matrix), the $i$th row of which refers to the values of property $i$ measured over $N$ objects, whereas column $j$ refers to the measured values of properties $\xi_1, \xi_2, \ldots, \xi_n$ of object $j$. The measured value $y_{ij}$ depends on one hand on the object and on the other hand on the experimental error.

The observed properties $\xi_1, \xi_2, \ldots, \xi_n$ are generally not independent. Our aim is just to determine the correlations between them. The following discussion deals with the linear factor model, which is restricted to linear correlations only. The restriction of linearity is in general not necessary, but this is the simplest and most frequently discussed case. The task is, from the mathematical point of view, to find $r$ common factors

$$\underline{\varphi}[r] = (\varphi_1, \varphi_2 \ldots \varphi_r)$$

and $n$ so-called individual factors

$$\underline{\varepsilon}[n] = (\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n)$$

Moreover an arbitrary matrix $\mathbf{A}[n \times r]$ and a diagonal matrix $\mathbf{B}[n \times n]$, by means of which the properties studied can be expressed in the form

$$\underline{\xi} = \mathbf{A}\underline{\varphi} + \mathbf{B}\underline{\varepsilon}$$

possibly under the condition that $r \ll n$.

Furthermore, the conditions

$$c(\varphi_i, \varphi_j) = \delta_{ij} \qquad i, j = 1, 2, \ldots r$$

$$c(\varepsilon_i, \varepsilon_j) = \delta_{ij} \qquad i, j = 1, 2, \ldots n$$

and

$$c(\varphi_i, \varepsilon_j) = 0, \qquad i = 1, 2, \ldots r; \quad j = 1, 2, \ldots n$$

must be fulfilled, where $c(\,,\,)$ means the covariance of the two variables and $\delta_{ij}$ is 1 if $i = j$ and 0 otherwise.

The values of the coefficients of individual factors should be as low as possible, i.e. the sum of $b_{ij}^2$ should be a minimum, where $b_{ij}$ is the element of matrix $\mathbf{B}$. Since $\mathbf{B}$ is a diagonal matrix, $b_{ij} = 0$ if $i \neq j$.

The aim of factor analysis is, therefore, to find the smallest possible number of non-correlated factors $\varphi_1, \varphi_2, \ldots, \varphi_r$ of unit variance under the condition that

$$\underline{\xi} \approx \mathbf{A}\underline{\varphi}$$

hold, i.e. the appropriate linear combinations of common factors $\varphi$ should describe the investigated properties $\xi$ with sufficient accuracy.

In the following discussions, when we use the term 'factor' we mean common factors. The minimum number, $r$, of factors necessary for the above description is in several cases a characteristic value, e.g. it may characterize the number of components in a substance. Matrix

$$\mathbf{A}[n \times r] = \begin{pmatrix} a_{11}, \ldots, a_{1r} \\ \vdots \\ a_{n1}, \ldots, a_{nr} \end{pmatrix}$$

is called the matrix of factor weights or the weight matrix, and its elements $a_{ij}$ are called factor weights. Similarly, the diagonal matrix $\mathbf{B}[n \times n]$ is called the weight matrix of individual factors, the elements $b_{ii}$ are called individual factor weights. The value of $b_{ii}$ gives the weight of individual factor $\varepsilon_i$, i.e. it shows the error of the description of property $\xi_i$ by the common factors.

Weights $a_{ij}$, which play an important role in the analysis, can be estimated on the basis of data matrix $\mathbf{Y}$. The values of factors belonging to the $N$ measurements can be found in the matrix

$$\mathbf{F}[r \times N] = \begin{bmatrix} f_{11}, \ldots f_{1N} \\ \vdots \\ f_{r1}, \ldots f_{rN} \end{bmatrix}$$

and those of individual factors in

$$\mathbf{E}[n \times N] = \begin{bmatrix} e_{11}, \ldots, e_{1N} \\ \vdots \quad\quad \vdots \\ e_{n1}, \ldots, e_{nN} \end{bmatrix}$$

These matrices are connected by the basic equation

$$\mathbf{Y} = \mathbf{AF} + \mathbf{BE}$$

i.e.

$$y_{i,j} = \sum_{k=1}^{r} a_{ik} f_{kj} + b_{ii} e_{ij} \qquad i = 1, 2, \ldots n; \quad j = 1, 2, \ldots N$$

Let us denote the covariance matrix of properties calculated from the measured $\mathbf{Y}$ values by $\mathbf{C}_y$. Since the factors are noncorrelated and have unit variance, one obtains on the basis of the above equation

$$\mathbf{C}_y = \mathbf{AA}^T + \mathbf{B}^T\mathbf{B}$$

where $\mathbf{A}^T$ and $\mathbf{B}^T$ denote the transposes of $\mathbf{A}$ and $\mathbf{B}$, respectively. Since $\mathbf{B}$ is diagonal, $\mathbf{B}^T\mathbf{B}$ is also diagonal, and thus the off-diagonal terms of $\mathbf{C}_y$ and $\mathbf{AA}^T$ matrices are identical. If, therefore, weight matrix $\mathbf{A}$ is known, the covariances between the original variables can be reconstructed.

For the observed standard deviations $s_i^2 = c(y_i, y_i)$ one may write:

$$s_i^2 = \sum_{j=1}^{r} a_{ij}^2 + b_{ii}^2 \qquad i = 1, 2, \ldots, n$$

The extent of the reproducibility of variances by means of factors can therefore be characterized by the sum

$$\sum_{i=1}^{n} b_{ii}^2 \qquad i = 1, 2, \ldots n$$

which should be as small as possible. Further characteristic values are the so-called communalities,

$$h_i = \sum_{j=1}^{r} a_{ij}^2 / S_i^2 \qquad i = 1, 2, \ldots, n$$

which are close to 1 in the case of a good approximation. If $r = n$, then $h_i = 1$ and

$$\sum_{i=1}^{n} b_{ii}^2 = 0$$

can be fulfilled. Otherwise $0 \leq h_i \leq 1$ and

$$\sum_{i=1}^{n} b_{ii}^2 \geq 0$$

The basic steps of factor analysis are the determination of the minimum required number of factors, $r$, the matrix **A** of factor weights and, in certain cases, the values of factors from the data matrix **Y**.

Since the conditions set for the factor system are purely statistical conditions, the factors obtained can generally not be identified with any physical or chemical properties. Therefore, a further aim of factor analysis may be to facilitate the interpretation of factors by means of their transformation or rotation.

*The methods of factor analysis*

There are several methods for the determination of the necessary number of factors, $r$ and the weight matrix **A**. The most widely known of these methods are the method based on maximum likelihood estimation, the centroid method and the principal component method.

The method based on maximum likelihood estimation is statistically the best one, since it gives the best undistorted estimates for the factor weights, and it is independent of the choice of the units of properties. However, a disadvantage in computer applications is its slow convergence.

The centroid method is statistically inferior, but it is very simple and can be implemented on small computers and even on calculators.

The most frequently applied method for the calculation of the weight matrix is the principal component method, which has been known for a long time in the literature of physics and mathematics, sometimes under the name of the eigenvalue problem.

The model of the method of principal components is as follows. We look for the linear transform $\eta[n]=(\eta_1, \eta_2, \ldots \eta_n)^T$ of property vector $\xi[n]$:

$$\underline{\eta}=\mathbf{D}\underline{\xi}$$

where the transformation matrix is $\mathbf{D}[n \times n]$, for which the components of $\eta$ are uncorrelated and the variance of $\eta_1$ is a maximum, moreover, in the same sequence, the variances of $\eta_2, \eta_3, \ldots, \eta_n$ are also maximum and the components remain noncorrelated. The elements $d_{ij}$ of matrix $\mathbf{D}$ should thus be determined so that the following relations are satisfied:

$$\max_{d_{11}, d_{12}, \ldots d_{1n}} c(\eta_1, \eta_1), \qquad \eta_1 = \sum_{j=1}^{n} d_{1j}\xi_j$$

and

$$\max_{d_{i1}, d_{i2}, \ldots d_{in}} c(\eta_i, \eta_i) \qquad c(\eta_i, \eta_k)=0, \qquad k<i,$$

$$\eta_i = \sum_{j=1}^{n} d_{ij}\xi_j, \qquad i=2, 3, \ldots, n$$

For the sake of uniqueness, it is required that

$$|d_i|^2 = \sum_{j=1}^{n} d_{ij}=1 \qquad i=1, 2, \ldots, n$$

The vector $\underline{\eta}$ which satisfies these conditions is called the principal component vector corresponding to property vector $\underline{\xi}$, and the coordinates $\eta_1, \eta_2, \ldots, \eta_n$ are principal components.

Once matrix $\mathbf{D}$ is known, the values of the principal component vector corresponding to the observed $\underline{y}_i=(y_{i,1}, y_{i2}, \ldots, y_{in})^T$ values of property vector $\underline{\xi}$ can be calculated. These data, denoted by $x_i=(x_{i1}, x_{i2}, \ldots, x_{in})^T$ can be determinated from equation

$$x_i=\mathbf{D}v_i, \qquad i=1, 2, \ldots, n$$

The elements of matrix $\mathbf{D}$ are estimated from data matrix $\mathbf{y}$. Let us denote the covariance matrix of $\xi$ by $\mathbf{C}_y$ and that of $\eta$ by $\mathbf{C}_x$, calculated from observations $\mathbf{Y}$ and $\mathbf{X}$, respectively. If the components of $\xi$ are random variables with mean 0 and variance 1, then $\mathbf{C}_y$ is identical to the correlation matrix, and, because of the relationship $\eta=\mathbf{D}\xi$, one way write

$$\mathbf{C}_x=\mathbf{D}\mathbf{Y}\mathbf{Y}^T\mathbf{D}^T=\mathbf{D}\mathbf{C}_y\mathbf{D}^T$$

Since the components of $\eta$ are supposed to be noncorrelated, $\mathbf{C}_x$ is diagonal. On the basis of the above conditions, the elements of $\mathbf{D}$ may readily be estimated.

It can be shown that the $i$th row of matrix $\mathbf{D}$, i.e.

$$\underline{d}_i^T = (d_{i1}, d_{i2}, \ldots, d_{in})$$

is the unit eigenvector belonging to the $i$th highest eigenvalue, $\lambda_i$, of covariance matrix $\mathbf{C}_y$, which should be orthogonal to the other eigenvectors. Moreover, the standard deviation of $\eta_i$ is just $\lambda_i$, and thus for the diagonal elements of diagonal matrix $\mathbf{C}_y$ it can be shown that $\lambda_i \geq \lambda_2 \geq \ldots \geq \lambda_n$.

Accordingly, the determination of $\mathbf{D}$ is reduced to an eigenvalue problem, for which the necessary computer programs are available.

Since condition $|\underline{d}_i| = 1$ was fixed, and since the orthogonality of eigenvectors $\underline{d}_i$ is a condition which can be met uniquely for the various eigenvalues, it follows that $\mathbf{D}$ is an orthonormal matrix, i.e. its rows are mutually orthogonal vectors of unit length, and therefore $\mathbf{D}^{-1} = \mathbf{D}^T$, i.e. the inverse of $\mathbf{D}$ is identical to its transpose. Therefore,

$$\mathbf{C}_y = \mathbf{D}^T \mathbf{C}_x \mathbf{D} = \sum_{i=1}^{n} \lambda_i d_i d_i^T = \sum_{i=1}^{n} \lambda_i \mathbf{D}_i$$

where $\mathbf{D}_i = d_i^T d_i$ is an $n \times n$ matrix. This procedure breaks covariance matrix $\mathbf{C}_y$ down into the of $n$ matrices. Each term of the sum corresponds to a principal component, since $d_i$ is the weight vector necessary for producing the $i$th principal component and $\lambda_i$ is the variance.

Consequently, noncorrelated principal components $\eta_i$ and their weights 'explain' the correlations between components $\xi_i$. In the above sum matrices $\mathbf{D}_i$ have weights $\lambda_i$, thus $\lambda_i$ can be taken as being characteristic of the contribution on $\eta_i$ to the reproduction of $\mathbf{C}_y$.

Since $\Sigma \lambda_i = \Sigma s_i^2$, in order to characterize how the first $r$ components reproduce $\mathbf{C}_y$, one can use the ratio

$$p_r = \sum_{i=1}^{r} \lambda_i \Big/ \sum_{i=1}^{n} s_i^2$$

for which $0 \leq p_r \leq$ i. If the maximum number of linearly independent components $\xi_i$ is $k$, which can also be termed as the degree of freedom of the system, then $p_r = 1$ if $r \geq k$ and $p_r < 1$ if $r < k$. Consequently, in this case the first $k$ principal components, $\eta_1, \eta_2, \ldots, \eta_k$ fully characterize the covariance matrix $\mathbf{C}_y$.

Since, rather than $\xi$ only the measured data matrix, $\mathbf{Y}$, is known, the estimates of $\mathbf{C}_y$ and $\mathbf{D}$ contain errors, and thus it is probable that, in cases when $i < n$ $p_i < 1$ will be obtained. The degree of freedom can therefore be estimated only with a given uncertainty. Usually a threshold, $1 - \delta$ is set up, and $r$ is determined so that $p_{r-1} < 1 - \delta \leq p_r$. In this case the $n$ properties $\xi_1, \xi_2, \ldots, \xi_n$ can be characterized 'sufficiently well' by the first $r$ principal components.

The role of principal components in factor analysis is as follows. One possible way of describing the $r$ factors, assumed to have unit variance and to be noncor-

related, is

$$\varphi_i = \eta_i / \sqrt{\lambda_i}, \qquad i = 1, 2, \ldots, r$$

and their estimated values are

$$f_i = y_i / \sqrt{\lambda_i}, \qquad i = 1, 2, \ldots, r$$

The factor weights can be estimated by the equation

$$a_{ij} = d_{ij} \sqrt{\lambda_i}, \qquad i = 1, 2, \ldots, r$$

As has already been mentioned, the system of factors is not unambiguous, in contrast to the principal components, which are unambiguous in the case of different eigenvalues.

It is an important relationship that the correlation of $i$th property and $j$th factor is $a_{ij}$. This may play a decisive role in the practical interpretation of factors, since it shows which properties are in a stronger and which in a weaker linear relationship with the factors. Therefore, in factor analysis, the determination of the weight matrix is the most important task, since this can be used to draw conclusions concerning the number and properties of unknown factors that may be hidden in the background.

The determination of the matrix of factor values **F** is not always necessary but, if it is, it can be calculated, e.g. by regression.

Consequently, the determination of factors by the method of principal components is relatively easy and readily programmable. It should be noted, however, that the condition that the individual factors should not be correlated with the factors is generally not met. This, in turn, implies that the off-diagonal elements of $\mathbf{C}_y$ and the $\mathbf{A}^T\mathbf{A}$ product may also be different.

### *The physical meaning of factors, rotation of factors*

We have seen how factor weights and the values of factors can be determined from data matrix **Y**. There remains the most important procedure for the chemist, the determination of the physico-chemical meaning of the factors. As we have mentioned, factors have been treated so far only on the basis of mathematical definitions, which completely ignore the specific properties of the system being studied. The interpretation of factors may often be facilitated by the rotation of factors.

In the rotation of factors the fact is utilized that factors are not unambiguous, and the system of factors obtained can be rotated arbitrarily in the subspace determined by it without violating the relationships that have been set up. One may therefore attempt to perform rotations which bring the factors into positions that are easier to interpret by chemists or engineers.

Factor rotation, which is in fact an appropriately chosen linear coordinate transformation is of various types. One may distinguish between orthogonal and

nonorthogonal, so called (target) rotations. In the former rotation the factors retain their orthogonality during rotation, i.e. their noncorrelated character, whereas in the latter this condition is omitted and one attempts to rotate the factors into vectors corresponding to some parameters of physical significance.

Orthogonal rotation can be expressed mathematically by a multiplication with a so-called orthogonal rotation matrix. Of the several orthogonal rotations known we may mention, as an example, the most widespread 'varimax' rotation, in which the objective is that in the weight matrix of the rotated factors the values in the individual columns, which show the correlation between the original properties and the factors, should be close as possible to $\pm 1$ or 0. In this case the rotated factors are practically independent of some properties, and very strongly correlated with other properties, making interpretation straightforward for the engineer.

In target rotation orthogonality is not demanded; but the factors, or some of them, are rotated by means of the appropriate choice of the rotation matrix into vectors having a previously fixed position. Since the rotated vectors remain in all cases within the subspace determined by the factors, rotation into a given vector is feasible only if this vector is also in the subspace of the factors. If not, the usual process is to rotate it into a vector of the subspace that is closest to the given vector, which is the projection of the given vector in the subspace.

Target rotation is particularly suitable in chemical engineering when one has to decide of a given physical quantity whether or not it is included in the factor space, i.e. whether it can be regarded as a factor that influences the system, being in some relation with the properties studied. If so, one of the factors is rotated by target rotation to make it coincide with the property in question.

Target rotation can be carried out as follows. Data matrix $\mathbf{X}$ can be estimated by means of the factor matrix and the weight matrix in the form of

$$\mathbf{X} \approx \mathbf{AF}$$

where in the $i$th row of factor matrix $\mathbf{F}[r \times N]$ the values of $f_{i1}, f_{i2}, \ldots, f_{iN}$ belonging to the $N$ studied objects of the $i$th factor can be found. In a geometric sense this is one of the axes of the $r$ dimensional subspace of the $N$ dimensional space, the factor space. One of these axes is to be rotated within the factor space in a way that it coincides with a given accuracy with vector $\underline{z}$ having components $z_1, z_2, \ldots z_N$ measured on the objects corresponding to the physical quantity of interest. This can be done if and only if physical quantity $\psi$ is within the $r$ dimensional subspace of the factors.

The target rotation of factor matrix $\mathbf{F}$ is equivalent to a multiplication by matrix $\mathbf{T}[r \times r]$, i.e. the rotated matrix

$$\bar{F}[r \times n] = \begin{pmatrix} \bar{f}_{11}, \bar{f}_{12}, & \cdots & \bar{f}_{1n} \\ \bar{f}_{21} & & \vdots \\ \vdots & & \\ \bar{f}_{n1}, & \cdots & \bar{f}_{nn} \end{pmatrix}$$

can be written in the form

$$\bar{\mathbf{F}} = \mathbf{TF}$$

Since in target rotation orthogonality is generally not required, the only assumption that should be made concerning **T** is that it is not singular.

For the rotation of the $i$th factor, the elements $t_{i1}, t_{i2}, \ldots, t_{ir}$ of the $i$th row of matrix **T** should be determined so that the sum of squares

$$S = \sum_{k=1}^{N} (z_k - \bar{f}_{ik})^2$$

should be a minimum, where

$$\bar{f}_{ik} = \sum_{j=1}^{r} t_{ij} f_{jk}, \qquad k = 1, 2, \ldots, N$$

It follows from the minimality condition that the $t$ values should satisfy the linear system of equations

$$\frac{\partial S}{\partial t_{ij}} = 0, \qquad j = 1, 2, \ldots, r$$

The solution of this system readily yields the values of $t_{ij}$.

If the minimum obtained for $S$ can be practically regarded as practically 0, then $\psi$ is in the factor space and the rotated vector can be identified with property $\psi$. If the minimum of $S$ is significantly higher than 0, it may be assumed that $\psi$ is outside the factor space and cannot be characterized sufficiently by the factors.

The above procedure can also be applied if some, at most $n-r$, values of $z_k$ are unknown. Then the terms containing the unknown $z_k$-s are omitted from the sum defining $S$. The procedure is invariably performable in this case, too, and $t_{ij}$ and thus the values $\bar{f}_{ik}$ can be determined. The values of $\bar{f}_{ik}$ may be taken as estimates for the missing values $z_k$.

In the process of target rotation, matrix **A** of factor weights is transformed according to the equation $\bar{\mathbf{A}} = \mathbf{AT}^{-1}$, and thus, after rotation, the equation

$$\mathbf{X} \approx \mathbf{AF} = \bar{\mathbf{A}}\bar{\mathbf{F}}$$

still holds.

### 6.2.7 Processing of density function-like signals

A significant fraction of the signals produced by analytical chemical measuring systems have a 'density function' character, such as spectra or chromatograms, for instance, or a 'distribution function' character, which is the integral of the former, such as thermogravimetric curves produced by chemical decomposition processes. Of course, within these groups several types can be distinguished. The main common characteristics of these groups will be reviewed below.

The processing and interpretation of density (and distribution) function-like signals may be performed in various ways, starting with fingerprint-like identification and ending with qualitative and quantitative analysis on the basis of chro-

matograms and spectra. This section deals with the processing of density function-like signals for the qualitative identification and quantitative determination of the constituents, of a sample.

*Density function-like signals*

In the foregoing discussion density function-like signals have been defined as measured signals $y(x)$ (where $x$ may represent time, wavenumber, etc.) of which it is assumed that they represent the algebraic sum of the signals $f_k(x)$ of the individual components, where $k$ is the index of the component, the baseline $b(x)$ and a random noise $n(x)$, i.e.

$$y(x)=\sum_{k=1}^{K} f_k(x)+b(x)+n(x)$$

As to the signals of components, it is generally assumed that they are identical in type, differing only in parameters, and it is often assumed, too, that they are of the form

$$f_k(x)=q_k f\left(\frac{x-\mu_k}{\sigma_k}\right)$$

where function $f(..)$ is a so-called 'unimodal' function which equally describes the signals of all components and has a density function character, i.e.

$$f(-\infty)=f(+\infty)=0 \quad \text{and} \quad \int_{-\infty}^{+\infty} f(\tau)\,d\tau=1$$

Furthermore, $\mu_k$ is a parameter characteristic of the quality of the component, whereas $q_k$ is proportional to the quantity of the component. With this assumption, the aim of quantitative analysis is the qualitative identification of the components on the basis of parameters $\mu_k$, and their quantitative determination on the basis of parameters $q_k$.

Baseline $b(x)$ is a stochastic signal the change of which is slower and smaller than that of the signals of the components, and much slower than the change of random noise, but similar in magnitude. Owing to this low rate of change, the stochastic character of the baseline is neglected, treating it as a deterministic signal.

Random noise $n(x)$ is a stochastic process that changes much faster than the signals of the components. Its magnitude is usually smaller than that of the useful signal, and it is assumed to be a 'white noise' with an expected value of zero and a normal distribution.

In the foregoing discussion, no distinction has been made between the continuous signal $y(x)$ and its discretized values $y(x_1), \ldots, y(x_N)$, or, using a simpler notation, $y_1, \ldots, y_N$, which can be expressed in vector form as

$$\underline{Y}=\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

*Processing of density function-like signals*

Like the interpretation of all analytical signals, the interpretation of density function-like signals also consists of two main steps, signal processing, i.e. the production of analytical information, and analytical chemical correlation. In the interpretation of signals of such a type, signal processing is called primary data processing and analytical chemical correlation is termed as secondary data processing. Signal processing will be discussed immediately below, whereas signal interpretation will be reviewed in Section 6.3.

The first main task of the interpretation of density function-like signals for the purposes of qualitative and quantitative analysis is signal processing, i.e. the determination of analytical information from the measured signal (chromatogram, spectrum, etc.). This analytical information includes qualitative (retention time, characteristic wavenumber, etc.) and quantitative (peak area, peak height, etc.) parameters.

Signal processing consists of two main steps, the determination of the signals of components and the characterization of the signals, i.e. the peaks of components.

The determination of the signals of components can be divided into the following partly interrelated steps:

(i) Noise filtering: suppression of the effect of random noise.

(ii) Baseline correction: estimation of baseline and subtraction from the measured signal.

(iii) Peak resolution: resolution of overlapping component signals into individual component signals.

The characterization of individual component signals (peaks) means the generation of qualitative and quantitative parameters, i.e. the analytical information, from the component signals.

The flow chart of signal processing can be seen in *Fig. 6.2.*

Since noise filtering has already been discussed above (Section 6.2.4), the next sections deal with baseline correction, resolution of overlapping peaks and peak characterization.

*Baseline correction*

One of the most difficult tasks of the processing of density function-like signals is appropriate baseline correction. Baseline correction consists of three main steps:

1. selection of the sections of signal that can be regarded as baseline;
2. estimation, by interpolation, of the missing baseline sections;
3. correction for the estimated baseline, i.e. its subtraction from the measured signal.

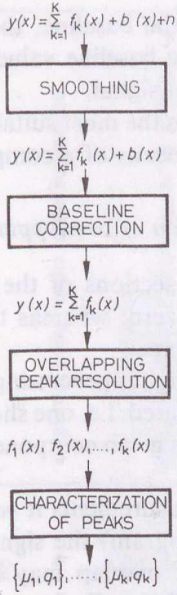



Fig. 6.2. Processing of density function-like signals

To determine which sections (points) of the smoothed signal containing acceptable (feasible) data can be regarded as baseline, or by contrast, which can be regarded as component peaks, is in many cases a very hard task; in a number of instances the answer is not even unique.

Most of the methods are based on the determination of the extrema (minima or maxima) and inflection points of the smoothed signal by the formation of differences or by differentiation. It is evident that maximum locations of the signal correspond generally to the peaks of components or their superimposition, whereas the minima are usually baseline points or valley points between overlapping peaks.

After deciding, by means of some appropriate method, which sections of the smoothed curve can be regarded as baseline sections, the second step of baseline correction, baseline estimation, involves the appropriate interpolation or, occasionally, extrapolation between (or beyond) the known sections.

The simplest case of baseline interpolation is when the last point of a baseline section and the first point of the next baseline section are connected by a straight line.

The essence of baseline interpolation by means of enveloping straight lines is the determination of the enveloping straight lines of known baseline sections. The slopes of these lines may not exceed a present value. If this criterion is satisfied, the enveloping straight lines so determined can be regarded as baseline. The boundaries of interpolation are the common points of the baseline and the signal. Thus, the interpolation interval also uniquely determines the baseline itself.

Once one is in possession of the baseline, the third and last step of baseline correction is easy to perform: the baseline values should be subtracted from the smoothed values of the analytical signal.

In up-to-date analytical systems the most suitable method of baseline correction can be selected interactively by means of a computer display.

*Resolution of overlapping peaks*

After baseline correction, the sections of the signal which had been baseline sections are identically equal to zero, whereas the nonzero sections contain the peak(s) of one or more components.

Therefore, in the next step of signal processing the uniqueness (multiplicity) of the given peaks should be investigated, i.e. one should determine whether the given peak or peak system is a result of a given component peak or whether it arises from the peaks of several components.

During analytical chemical measurements it very often occurs that in the measured signal (spectrum, chromatogram) the signals of two or more components merge or overlap. The extent of overlap may be so great that the number of components cannot be determined visually from the measured set of points. Thus, the goal of peak resolution is the determination of the number, location and, if possible, the area of component peaks from the measured set of points as the only source of data.

For the resolution of overlapping peaks, a great number of quite different methods has been developed. Which of them is to be applied depends on the nature of the curves, the extent of their overlap and the computer background available.

For the resolution of density function-like signals, a large number of techniques may be applied ranging from simple geometric methods to the deconvolution technique applying Fourier transformation.

The diversity of methods makes it difficult to compare them, but we hope that the data shown in *Tables 6.1* and *6.2* will help the reader in selecting the appropriate method.

The tables summarize the characteristic properties and application conditions of these methods by taking into account some aspects that are commonly encountered in practice, with the aim of making the choice between them easier for the solution of a given task. As a summary of the tables we may state that the choice of the appropriate method depends primarily on the type of information that one wishes to acquire by means of resolution. For example, if the aim is the detection of a contaminant or a trace component, as well as other constituents, only statistical methods may be considered. An important factor is a knowledge of the number of components. It is also relevant whether the parameters of the components should be known and, if so, how accurate they should be. Moreover, if necessary, is it possible to make them more accurate by other methods? By attempting to

Table 6.1. Conditions of the applicability of decomposition methods

| The principle of the method | Method applied | Should curve type be assumed? | Should number of components be known? | Is it applicable to strongly overlapping peaks? | Information obtained after decomposition |
|---|---|---|---|---|---|
| Geometric methods | perpendicular tangent triangle | no | yes | no | area below component curves |
| Curve fitting | LKN method | | yes (no) | | parameters of components |
| | cut | | no | no | |
| | short-cut | yes | yes | | |
| | use of standard curves | no | yes | yes | weights of components in the suspension |
| Moments | | yes | yes | yes | values of parameters |
| Statistical methods | Bayes method | yes | yes | yes | presence of minor components in addition to main peak |
| | cross correlation | | | | |
| | factor analysis | no | no | yes | number and quantity of components |
| Narrowing of component curves | depends on type of curves | yes | no | yes | number of components + possibility for determination of parameters |
| Deconvolution | Fourier-transformation pseudo-deconvolution | no | no | yes | number of components + possibility for determination of parameters |
| | series expansion | | | | |

Table 6.2. Characterization of decomposition methods

| Main principles | Practical method | Applicable to continuous processing of curve? | Precision | Reproducibility | Computing effort | Note |
|---|---|---|---|---|---|---|
| Geometric | perpendicular tangent triangle | no | very large error | poor | minimum | very simple |
| Curve fitting | LKN method | | depends on initial guess | depends on initial guess | great | good initial guess required |
| | cut | | medium | medium | great | |
| | short-cut | no | poor | poor | little | |
| | use of standard curves | | generally very large error | poor | medium | applicable only if the maximum locations well reproducible |
| Moments | | yes | medium | medium | little | |
| Statistical methods | Bayes method | no | depends on problem | medium | little | applicable to detection of small quantities |
| | cross correlation factor analysis | | | | medium | |
| Narrowing of component curves | depends on type of curves | no | good | good | great | |
| Deconvolution | Fourier transformation pseudo-deconvolution | no | good | good | great | advantage that weight function of system can be given numerically |
| | series expansion | yes | | | little | |

answer these questions, most of the methods can be eliminated *a priori*. One should also investigate whether one may assume any particular curve shape for the mathematical description of the measured curve, and what the restrictions of doing so may be.

Of course, the application of any method is limited by the extent of overlap between the peaks. The computer facilities at hand may also impose limitations, although this factor tends to diminish in significance.

*Characterization of component peaks*

The last step of the processing of density function-like signals is the determination of analytical information from the resolved, individual components, i.e. the determination of parameters $\mu_k$ characteristic of quality and $q_k$ characteristic of quantity from functions (peaks) $f_k(x)$.

Evidently, the determination of parameters $\mu_k$ and $q_k$ may be carried out in several different ways, and the choice of the appropriate method also depends on the knowledge acquired so far. When, for example, the peak shapes are known and curve resolution is carried out by means of curve fitting, this process also yields the desired parameters.

The simplest way to characterize the peaks is to assume that the qualitative parameter $\mu_k$ is nearly identical to the location of the peak's maximum and thus only the latter is to be determined, and that the quantitative parameter $q_k$ is equal to the area below the peak, and thus that it can be obtained by simple integration.

Of the several methods applicable for the description of peaks, the method of moments will now be discussed in detail. Moments are characteristic of the shapes of peaks, and such reasonable meanings can be attached to them as area, width, skew, etc.

From the values $y_i$, $1=1, 2, \ldots, N$ of a density function-like curve measured at equidistant points $x_i$, the moments of the peak can be determined by the following expressions:

— the $s$th (ordinary) moment:

$$M(s)=\sum_{i=1}^{N} x_i^s\, y_i, \qquad s=0, 1, \ldots$$

— the $s$th relative moment:

$$RM(s)=M(s)/M(0), \qquad s=0, 1, \ldots$$

— the $s$th central moment:

$$C(s)=\sum_{i=1}^{N} (x_i-RM(1))^s\, y_i \qquad s=0, 1, \ldots$$

— the $s$th relative central moment:

$$RC(s)=C(s)/M(0), \qquad s=0, 1, \ldots$$

It should be noted that the above formulae give the so-called discrete moments calculated from the sampled values of the given curve. These equations are the approximations of the integral definitions applying to continuous curves, which can be written as follows:

$$M_c(s)=\int_{x_l}^{x_u} x^s y(x)\,dx$$

$$RM_c(s)=M_c(s)/M_c(0)$$

$$C_c(s)=\int_{x_l}^{x_u} (x-RM_c(1))^s\, y(x)\,dx$$

$$RC_c(s)=C_c(s)/M_c(0)$$

where $x_l$ and $x_u$ are the abscissae corresponding to the beginning and end of the peak, respectively.

If $y(x)$ is a density function, taken in a statistical sense, i.e. if

$$M(0)=\int_{-\infty}^{+\infty} y(x)\,dx=1$$

then the relative moments are identical to the ordinary moments. The relative moments mean moments as they are related to unit area.

The most important moments and their meanings will now be reviewed.

(a) M(0), the 0th ordinary moment, is the area of the curve below the peak between values $x_l$ and $x_u$, yielding parameter $q_k$ of quantitative analysis.

(b) RM(1), the first relative moment, is the expected value of the distribution characterized by the curve shape considered as a distribution function, i.e. in a visual sense the location of the centre of gravity of the peak. In the case of symmetric peaks, this is coincident with the location of the peak maximum. If the relative first moment is to the left or to the right of the maximum, this means that the peak is tailing in the appropriate direction, which may indicate that the peak is a superimposition of several components.

(c) RC(2), the second relative central moment, is the dispersion or width height of the distribution characterized by the curve considered as a density function, i.e. it is characteristic of the width of the peak relative to unit area. Peaks wider than average may again call attention to overlapping peak components.

(d) Skewness, calculated from the equation

$$\text{skewness}=RC(3)/RC(2)**(3/2)$$

is characteristic of the asymmetry of the peak. If the peak is symmetrical, its value is 0, if it is negative, the peak is skewed to the left, if positive it is skewed to the

right. From this parameter similar conclusions can be drawn as from the relation of peak maximum and expected value.

(e) The ratio

$$\text{excess} = \text{RC}(4)/\text{RC}(2)**2 - 3$$

is characteristic of the 'flatness' of the curve. In the case of a Gaussian-type curve, excess is zero, for 'sharper' curves it is negative, and for 'flatter' ones it is positive. High excess values indicate, once again, the presence of overlapping peaks.

In the formula for excess, variance is in the denominator, and thus excess is defined as a width related to a normal distribution of the same variance, therefore the two quantities have different meanings. A positive value of excess indicates a plateau-like peak.

It can be seen from the above that the moments may give valuable information for the quantitative and qualitative evaluation of analytical peaks.

## 6.3 Signal interpretation methods

The ultimate aim of the signal processing methods described in Section 6.2 is the interpretation of analytical signals. Therefore, this section will review the signal interpretation methods used in analytical chemistry.

As has been mentioned in Section 6.1, analytical chemical signal interpretation consists of analytical signal processing and analytical chemical correlation. Analytical chemical correlation reveals and utilizes the relationships between analytical and chemical information, meaning, on the one hand, modelling of the analytical investigation (calibration) and, on the other hand, the application of this model (analytical chemical inference).

Several signal interpretation methods have been introduced and are widely used in chemical analysis. For qualitative analysis, algebraic–logical, pattern recognition, library search methods will be demonstrated (Section 6.3.1). Of the several types of quantitative analytical methods, the ones based on linear models will be discussed (Section 6.3.2).

### 6.3.1 Qualitative analysis by supervised classification

*Principle of pattern recognition*

The purpose of pattern recognition is to classify objects, i.e. materials or phenomena, into classes, categories or groups. In chemical metrology, the objects to be classified are always the materials involved in the analysis.

At the outset of the application of pattern recognition, the concept of the class of objects, i.e. of materials to be examined, must be defined. The classes are

interpreted in terms of the property which defines the class represented by the object. In chemical metrology, such a class-defining property is either a chemical characteristic, e.g. the presence or absence of a chemical group, or a practical property, e.g. the manufacturer of a sheet of paper, the origin of an archaeological find, the quality of a product, etc.

In general, the property which defines a class is a qualitative property which cannot be measured directly, whereas in some cases it is measurable, but the measurement would be very time consuming or expensive.

In addition to the class-defining property, the object may exhibit several other properties, which may be much easier to measure or observe. If there exists a relationship between easily measurable properties and the hard to measure class-defining ones, it is worth performing the classification in an indirect way. The principle of pattern recognition is to carry out the classification of objects according to difficult to measure or nonmeasurable properties in an indirect manner, by determining the more readily measurable or determinable properties connected with the nonmeasurable class-defining property.

For classification, a set of characteristic quantities-measured properties-called the pattern vector representing the 'pattern' of the object can be assigned to the objects. The set of properties constitutes the pattern space. If the properties can be expressed quantitatively, each object is a point in the pattern space defined by the values of the quantitative properties. Into this space one can introduce metric (i.e. the concept of 'distance') or at least a similarly measure.

Two objects in the pattern space are considered to be close to one another if their distance is small and far apart if their distance is large. On this analogy, two objects are regarded as being similar if the similarity measure of their points is high, and dissimilar if this value is low.

It is evident that indirect classification by means of pattern recognition is based on the assumption that a relationship does exist between the pattern of the object and its class defining property, i.e. the pattern of the object is expected to be characteristic of the class of the object. Pattern recognition methods assume that objects that are similar according to the class defining property (i.e. belonging to the same class) are close to one another in pattern space, whereas objects belonging to different classes are relatively far apart in pattern space.

This assumption implies that the pattern points of objects belonging to different classes are separated in the pattern space, and thus the pattern space can be divided into subspaces corresponding to the classes. On the basis of this assumption, classification by means of pattern recognition is performed according to the following rule: an object belongs to the class corresponding to the pattern subspace in which the pattern point of the object can be found.

*Feature selection*

In every pattern recognition method, classification and recognition are preceded by feature selection, i.e. the determination of the most important properties, the features of the pattern (see *Fig. 6.3*). Feature selection includes all transformations performed in pattern space which facilitate the distinguishing of classes, i.e. classification. Feature selection is to be carried out if the observed or measured pattern values of the objects are found to be non-independent or the information necessary for the classification is hidden or, possibly, if the pattern values contain redundant or even disturbing pieces of information. Thus, feature selection aims at transforming the pattern space into a feature space containing the minimum amount of information necessary for classification, and eliminating all redundant and disturbing pieces of information so as to make the recognition of patterns easier. The important properties remaining after feature selection are called features.

The methods available for solving the feature selection problem can be classified as problem oriented and mathematical feature selection methods.

A number of methods can be found for solving problem oriented feature selection, based on our knowledge of the given classification problem. For example, if one wants to classify compounds on the basis of their infrared spectra, according to the presence or absence of an aromatic ring, a possible problem oriented feature selection is to disregard all parts of the spectra that are unconnected with the vibrations of an aromatic ring.

Several mathematical methods of feature selection are also available. They can be divided into two main categories, depending on whether the components of the pattern vector are examined separately or simultaneously.

The component-by-component examination of the pattern vector can be achieved by transformations or by the elimination of irrelevant components. Component-by-component transformations of pattern vectors either eliminate the different weights of the components due to their different scalings by mapping them into the regions (0, 1), (−1, +1), etc., or aim at the better or linear separation of classes by logarithmic, quadratic, etc. transformations.

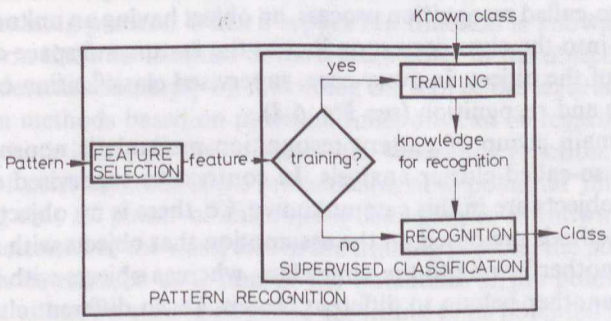



Fig. 6.3. Pattern recognition by supervised classification

A number of mathematical methods exist for the ordering and selection of the relevant components of the pattern vector. The basic principle of these procedures is the application of indices which show the importance and effectiveness of the individual components of the pattern vector according to the classification. In general, the given component of the pattern vector can be regarded as being important if its standard deviation is low enough within the classes, and, simultaneously, if the means of the various classes are sufficiently different. One index that expresses this relative importance is the widely used Fisher ratio. As can be seen, this concept is analogous to the resolution used in the case of overlapping peaks.

The simultaneous transformations of the components of the pattern vector aims primarily at the elimination of the dependence and the redundancy of components, and thereby at decreasing the number of dimensions.

Such transformations, for instance, the orthogonal transformations: Fourier, Haar and Walsh transformations, principal component and factor analyses, etc. Orthogonal transformations preserve distance in Euclidean space, and therefore they cannot be used to improve the separation of classes. They are, however, quite suitable for dimension reduction, since several types of functions can be approximated well in 'finite' dimensional space by orthogonal transformations. The latter can be regarded as weighted sums, therefore they generate less noisy signals from noisy ones, thus improving the reliability of the classification.

*Types of classification methods*

There is a wide choice of pattern recognition classification methods. These methods can be divided into two different types according to their function. The first group includes supervised classification, or learning with a teacher, or pattern recognition proper. In this group, classes are given beforehand, since a set of objects of known classes is available. With the use of this so-called training set, the location of classes in feature space is determined by finding the subspace of the feature space into which the object points of known classification fall. This procedure is called learning or training. Using the knowledge thus acquired in the course of the so-called recognition process, an object having an unknown classification is placed into the class corresponding to the feature subspace containing the feature point of the object. In every case, supervised classification consists of two steps: training and recognition (see *Fig. 6.3*).

The other main group of pattern recognition methods is nonsupervised classification, the so-called cluster analysis. In contrast to supervised classifiers, the classes of the objects are in this case unknown, i.e. there is no object of a specified class. These methods are based on the assumption that objects with feature points close to one another belong to the same class, whereas objects with feature points far from one another belong to different classes, i.e. to different clusters.

Thus, the basic idea of cluster analysis methods is the distance or similarity

measure, and their common characteristic is the determination of the 'homogeneous' groups of the points of objects in the feature space. A group of points of the objects is considered to be homogeneous if their inter-class distance is smaller than their intra-class distance.

From the user's point of view, an important distinction should be made between cluster analysis and supervised classification. The former should be used if there is no *a priori* knowledge available about the classes. Nevertheless, the clusters obtained by cluster analysis may have no physical meaning, they are rather a stochastic sorting of objects, since clusters depend on the measuring method, distance measure, feature selection and the cluster analysis method itself. However, if clusters obtained by cluster analysis have real physical meaning, these clusters as classes are suitable for classification by supervised learning methods.

*Supervised classification methods*

There are differences among various supervised classification methods, i.e. among the mathematical models which describe, in the feature space, the location and separation of feature points of objects belonging to the various classes. In chemical metrology, the characteristic points, the separating functions, the potential function and the closed region methods are widely applied.

In the case of classification methods based on characteristic points, training means merely the recording of the characteristic points of the objects of known classes, and recognition is the determination of the classes of objects having the closest characteristic points to the unclassified object. The class of the object is given by the majority class of the neighbouring points around the characteristic points of the object of unknown class (weighting points, K-nearest neighbour, K-NN method).

The library search methods using data banks are special types of characteristic point classifiers, a special type of the 1-NN method.

The basic idea of the separating or discriminant function methods is to determine separating functions which separate the feature space into subspaces belonging to the classes on the basis of the feature points of objects of known class in the course of the training process. When a separation function is known, recognition is easy to perform since the location of the feature point of the object, i.e. its class, can usually be determined simply by examining the sign of the separating function.

Classification methods based on potential functions can be regarded as a combination of methods based on the point and the separating function, respectively. The principle of these methods is the introduction of a 'potential' into the feature space, meaning that all points of the objects belonging to a known class have a potential characteristic of the class, and in the training process, the potential of the feature space is determined as a sum of the potentials of the points. An object belonging to an unknown class is classified according to its potential in the feature space.

A special class of supervised classification methods of great importance are *closed region methods*. Important examples of this method are the SIMCA and Prima methods. The SIMCA method (Soft Independent Modelling of Class Analogy, or Statistical Isolinear Multi-Category Analysis) is based on the examination of the location of the object points in the principal component space of the feature space. In this space, the boxes surrounding the points of the objects belonging to the same classes are determined in the training process, whereas recognition is carried out by determining the class of the box in which the point of the object of the unknown class falls, or else the object is regarded as an 'outlier' if its point is outside the boxes.

*Supervised classification as a signal interpretation method*

Analytical chemical methods can be regarded as systems which produce chemical information by the interpretation of measured, observed analytical signals. This signal interpretation means the processing of the analytical signal to generate analytical information and the inference of chemical information from analytical information. On comparing this signal interpretation process shown in *Fig. 6.1* with the pattern recognition process shown in *Fig. 6.3* it can be seen that supervised pattern recognition methods may be regarded as special types of analytical signal interpretation, because in this case the qualitative chemical composition of the sample is the property defining the classes, the analytical signal is the pattern, data processing is an equivalent of feature selection, analytical information is the feature, analytical chemical correlation is the classification, calibration is the training and the analytical chemical inference is identical to recognition. Consequently, pattern recognition can be used as a signal interpretation method in the case of qualitative analysis if there are some measured properties relating to the chemical property to be examined and if they are suitable for indirect classification.

The interpretation of mass, infrared, NMR and X-ray spectra as well as qualitative analysis by means of polarograms, voltammograms, etc., are based on this principle. Similarly, it is possible to distinguish between *cis*- and *trans*-configurations of unsaturated carbonyl compounds from their ultraviolet and infrared spectra, or to characterize the acidity of chemical elements from their valence, melting point, atomic radius, electronegativity, etc.

There are several other types of signal interpretation methods for qualitative analysis. Pattern recognition is similar to the algebraic–logic methods which are also based on the data processing of qualitative proper characterized quantitatively. The advantage of pattern recognition over these other methods lies in the simplicity of the former. Unfortunately, there are disadvantages, too: pattern recognition methods are suitable for determining only a limited number of components, whereas the algebraic–logic methods can be used to determine a set of functional groups simultaneously.

*Supervised classification using analytical chemical data*

Another main group of the application of supervised classification methods in analytical chemistry includes classification problems in which the quantitative characterization of the sample is performed by a certain analytical chemical system, i.e. the pattern is chemical or analytical information. It must be mentioned here that since there is an unambiguous relationship in quantitative analysis between the analytical and chemical properties, it is often immaterial which type of information is applied. Examples of such applications are the classification of archaeological artifacts, corpi delictorum and environmental pollutants according to their origin on the basis of their chemical composition, classification of products by quality or manufacturer, and classification of materials by their biological activity on the basis of their chemical properties.

### 6.3.2 Quantitative analysis by linear analytical investigation

Most of the quantitative analytical methods are subject to the assumption that the analytical investigation can be described by means of a linear model, i.e. the analytical information obtained by signal processing is a linear function of chemical information.

In the following discussion, according to the type of analytical information, quantitative analysis based on discrete, continuous or 'disperse' types of analytical information will be distinguished (disperse information is obtained from density function-like signals).

#### Linear discrete multicomponent systems

In the practice of analytical chemistry it may often be assumed that both analytical and chemical information are discrete or may be discretised, i.e. can be modelled as vectors, and that analytical information is a linear function of chemical information, i.e. a linear model is a good approximation for the analytical investigation.

*The modelling of linear, discrete, multicomponent systems*

It is assumed that the system which yields the analytical information is linear, i.e. there is a linear relationship between the amounts of components, i.e. the chemical information, and the determined analytical information:

$$\underline{y} = \mathbf{A}\underline{x}$$

where $\underline{x}[n]$ represents the amounts of components, $\underline{y}[m]$ is the analytical information, i.e. the values of the measured or calculated quantitative analytical property

and $\mathbf{A}[m \times n]$ is the so called calibration matrix, which describes the linear relationship between the former. It is assumed that $m \geq n$, i.e. at least as many pieces of analytical information are determined as the number of components.

Writing out the vectors and matrix in full, we have

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} a_{11} & \ldots\ldots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \ldots\ldots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

where $x_j$ is the amount of the $j$th component ($j=1, 2, \ldots, n$), $y_i$ is the value of the $i$th piece of analytical information ($i=1, 2, \ldots, m$) and $a_{ij}$ is the specific value, sensitivity of the $i$th piece of analytical information with respect to the $j$th component.

As has already been mentioned, analytical chemical correlation consists of calibration and analytical chemical inference. During calibration, the unknown matrix $\mathbf{A}$ of the above equation is estimated, then the inverse of this matrix is calculated and used for analytical chemical inference.

*Calibration*

Calibration measurements should be performed to estimate of calibration matrix $\mathbf{A}$ used by the model. Let us assume that $N$ such measurements have been carried out. Denote the amounts of components present in the calibration mixtures by matrix $\mathbf{X}[n \times N]$ and the values of analytical information obtained in the calibration measurements by matrix $\mathbf{Y}[m \times N]$. Then we have the approximate relationship

$$\mathbf{Y} \approx \mathbf{A} \times \mathbf{X}$$

or, writing the matrices in full,

$$\begin{pmatrix} y_{11} & \ldots & y_{1N} \\ \vdots & & \vdots \\ y_{m1} & \ldots & y_{mN} \end{pmatrix} \approx \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix} \begin{pmatrix} x_{11} & \ldots & x_{1N} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{nN} \end{pmatrix}$$

where $x_{jk}$ is the amount of component $j$ in the $k$th calibration mixture ($j=1, \ldots, n$, $k=1, \ldots, N$) and $y_{ik}$ is the value of the $i$th analytical property for the $k$th calibration mixture ($i=1, \ldots, m$, $k=1, \ldots, N$).

The elements of $\mathbf{A}$ are estimated by means of the method of least squares, i.e. so that for all analytical properties $y_i$ the following condition is satisfied:

$$\min_{a_{i1}, \ldots, a_{in}} \sum_{k=1}^{N} \left( y_{ik} - \sum_{j=1}^{n} a_{ij} x_{jk} \right)^2 \qquad i=1, \ldots, m$$

The $a_{ij}$ coefficients ($i=1, \ldots, m$, $j=1, \ldots, n$) can be determined from these criteria. The estimate $\hat{\mathbf{A}}$ for the approximation of calibration matrix $\mathbf{A}$ can be obtained from the equation

$$\hat{\mathbf{A}} = \mathbf{Y}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}$$

*Analytical chemical inference*

In analysis, it is not the model of the analytical chemical investigation but its inverse, i.e. equation

$$\underline{x} = \mathbf{B}\underline{y}$$

that is required, since our aim is to set up the relationship which can be used to determine the quantity of the components of a mixture on the basis of the measured analytical information. Analytical matrix **B** in the latter equation, which is the generalized inverse of calibration matrix **A**, can be estimated by the use of the estimated calibration matrix using the equation

$$\hat{\mathbf{B}} = (\hat{\mathbf{A}}^{\mathrm{T}}\,\hat{\mathbf{A}})^{-1}\,\hat{\mathbf{A}}^{\mathrm{T}}$$

Consequently, for analytical chemical inference we use the equation

$$\underline{x} = \hat{\mathbf{B}}\underline{y}$$

On the basis of analytical information $\underline{y}^*$, belonging to an unknown mixture $\underline{x}^*$, the estimate of the latter, i.e. the estimated quantities, $\underline{\hat{x}}^*$, of the components of the unknown mixture can be calculated from the equation

$$\underline{\hat{x}}^* = \hat{\mathbf{B}}\underline{y}^*$$

**Linear continuous multicomponent systems**

In certain fields of analytical chemistry a relatively new multicomponent quantitative analytical method, the determination of composition on the basis of library or standard curves has become widespread.

*The model of quantitative analysis on the basis of library curves*

The gist of this method is the assumption that the observed, measured analytical information, $y(z)$ of a mixture consisting of components $i = 1, \ldots, n$ is a superimposition of the signals of components, $y_i(z)$, i.e.

$$y(z) = \sum_{i=1}^{n} y_i(z)$$

Moreover, it is assumed that the pieces of analytical information $y_i(z)$ of the components are proportional to the amounts $q_i$ of the components, i.e.

$$y_i(z) = q_i \times f_i(z), \qquad i = 1, \ldots, n$$

where $f_i(z)$ is the 'normalized' analytical information related to unit quantities of the components. Taking this into account, we have

$$y(z) = \sum_{i=1}^{n} q_i f_i(z)$$

i.e. the analytical information is a sum of the individual items of analytical information of the components weighted by their amounts.

The normalized analytical information, $f_i(z)$ of component $i$ is often called the library curve or standard curve.

The essence of quantitative analysis on the basis of library curves is the above model of analytical investigation. In principle, this method is applicable in the case of all density function-like curves for which it may be assumed that analytical information is additive and the pieces of analytical information of the individual components are proportional to the quantities of the components. However, the practical applicability of the method depends on whether the value of the indenpendent variable, $z$, can be set accurately, and reproducibly, so that the analytical information is not shifted along $z$ axis. For this reason, the method is inapplicable in chromatography, but is applicable in X-ray diffraction, neutron activation analysis, gamma spectroscopy, fluorescence spectroscopy and NMR spectroscopy.

It is noted that the above principles from the basis of the so called simultaneous multicomponent rank annihilation (SMRA) method used for the processing of two dimensional signals (e.g. the excitation-emission matrix).

*Calibration*

The purpose of calibration is the determination of the normalized analytical information, i.e. the library curves of the components, which can be accomplished in various ways.

The simplest principle but one that is not always applicable is the determination, during calibration, of analytical information for known amounts of pure components only. Then, for the analytical information $y^k(z)$ obtained in the $k$th calibration measurement one has the approximation

$$y^k(z) \approx q_i f_i(z) \qquad k=1, \ldots, N; \quad i=1, \ldots, n, \quad N=n$$

and on the basis of the this equation the estimate for $f_i(z)$, the normalized analytical information of the components, can be determined from equation

$$\hat{f}_i(z) = (1/q_i)\, y^k(z)$$

*Analytical chemical inference*

Once one is in possession of the estimated analytical information on the components, $f_i(z)$, $i=1, \ldots, N$, the analytical chemical inference, i.e. the estimation of the unknown quantities $q_i^*$ $i=1, \ldots, n$ of the components, can be performed by means of the method of least squares, on the basis of the criterion

$$\min_{q_1^*, \ldots, q_n^*} \int_{z_l}^{z_u} \left[ y^*(z) - \sum_{i=1}^{n} q_i^* \hat{f}_i(z) \right]^2 dz$$

where $y^*(z)$ is the analytical information pertaining to the unknown mixture.

### Interpretation of density function-like signals

For the interpretation of density function-like signals for qualitative analysis, as discussed in Section 6.2.7, the analytical information comprises the values of quality parameters $\mu_1, \ldots, \mu_k$, characteristic of quality, and quantity parameters $q_1, \ldots, q_k$, proportional to quantity.

In the case of qualitative identification, the first step of analytical chemical correlation is calibration, in which the quality parameters of substances involved in the analysis are determined and stored, as analytical chemical knowlegde, in appropriate form (table, index system, etc.).

The second step of qualitative identification, analytical chemical inference, means the comparison of the values of measured quality parameters with the stored values. In general, a component with qualitative parameter $\mu_k$ is said to be the $n$th component if

$$\mu_k \in \mu_n \pm \Delta$$

where $\mu_n$ is the stored value of the $n$th component and $2\Delta$ is the so called window, which is an interval depending on the experimental error of the particular method.

Qualitative identification may then be followed by quantitative analysis. The essence of the latter is that as a first step, during calibration, that factor is determined which gives the value of quantitative parameter $q_k$ belonging to a known amount of component $k$, and then, once one is in possession of this calibration knowledge, during the phase of analytical chemical inference, the quantity of component $k$ is determined from the value of $q_k$ obtained in the analysis.

## 6.4 Performance characteristics of measuring systems

Analytical chemical systems are generally parts of certain high level systems; a process analysis instrument may, for example, be a subsystem of the process control unit of a chemical production system. It is therefore necessary to characterize analytical chemical systems according to the perspective of the process itself; i.e. to give the characteristics of the process analyser which are relevant from the aspects of the industrial process to be controlled.

Of course, the operation of analytical chemical systems may generally be characterized by numerous properties; they are referred to by some authors as performance characteristics, by others as figures of merit. According to Wilson, "The performance characteristics are a set of quantitative ... values for parameters of fundamental importance in assessing the suitability of the method for any given purpose."

The performance characteristics of analytical chemical systems allow the characterization of the methods of analytical chemistry in a unique, comparative manner,

including the specification of the aim of control appropriate from the viewpoints of the user and that of the restrictions, the comparison of various methods and the selection of the method useful for the given task.

### 6.4.1 Types of performance characteristics

There is no unique classification for the performance characteristics in analytical chemistry; in the following discussion functional, operational and evaluating performance characteristics will be distinguised.

*Functional performance characteristics*

Functional performance characteristics are a set of properties which characterize the function of the analytical chemical system as a system producing chemical information from the sample.

A functional characteristic of a method is, for example, whether the chemical information gives the spatial or the temporal distribution of the sample or spatial or temporal average value. Another functional characteristics of the method is the indentity, quality and number of constituents that are determinable. Specificity means in this respect that only the given component gives a signal, selectivity means that the determination of a given component is not perturbed by the presence of other components. An important characteristic is the limit of detection, which is the lowest amount of a given component that can be detected.

A functional charecteristic of an analytical method is the number of components that are determinable, more specifically the range and the lower and upper limits of qualitation. A basic characteristic is the type of the model of the analytical investigation, i.e. of the 'mapping' which assigns analytical information to chemical information. Sensitivity, another characteristic, is the change in analytical information caused by a unit change in the amount of the given component; in the case of linear relationships this is the slope of the calibration curve.

One may regard as a functional performance charecteristic the deviation between the estimated and known chemical information obtained with the calibration and control samples, respectively. The statistical estimates of these characteristics are SEC, SEP and classification abilities.

Performance characteristics may be defined not only for the analytical investigation and the complete analytical chemical system, but also for the analytical signal producing unit. Functional characteristics relevant to the latter are, e.g. baseline, band width, resolution, background, drift, etc.

*Operational performance characteristics*

One of the basic operational performance characteristics is the quality (matrix) and quantity of the samples used for calibration, control and analysis, and the method of sample pretreatment.

An important operational performance charateristic is the time required for calibration, control and analysis, including the time throughput of the analysis, i.e. the number of samples per unit time.

The resources necessary for calibration, control and analysis are also important operational performance characteristics, particularly the availability and quality of the operatives and the apparatus required.

Further important operational performance characteristics are the reliability, maintainability and reparability of the analytical chemical system, newed as a module serving the previous features. Software must be considered with regard to its error reporting ability. Finally, the ruggedness, steadiness and stability in time of the system are important.

The quality of operation is determined by the practicability and difficulty of the analytical chemical system. Clear and sufficient system documentation is the extent of control and automation of the system and, having regard to its expansion, its controllability and possibility of automation.

An important operational performance charactristic is safety. In certain cases users may demand security as well.

*Evaluating performance characteristics*

Analytical chemical systems may be evaluated from a very great number of aspects. Of these, statistical (Section 6.4.2), information theoretical (Section 6.4.3), decision theoretical (Section 6.4.4) and credibility characterization (Section 6.4.5) are of fundamental importance.

In addition to the evaluative performance characteristics listed above and to be discussed below, a method may obviously be evaluated according to various other points of view. An important aspect is the economic one, the aim of which is the estimation of the costs of calibration, control and analysis, the value of the chemical information produced, and thus the estimation of the profit gained through the analysis.

The theoretical reliability characterization of analytical chemical systems may also be of some importance: this relates to the estimation of the length of time for which the system will operate continuously without error or breakdown.

From the user's point of view an important performance characteristic is the relevance of the method, which characterizes the importance of the resulting analytical information for the user.

*Composition of performance characteristics*

In order to compare performance characteristics or to combine them into a single characteristic, it is first necessary to normalize them. There are several methods for normalization; one of the possible ways is to calculate

$$\eta = \frac{\eta_{\text{actual}} - \eta_{\text{worst}}}{\eta_{\text{best}} - \eta_{\text{worst}}}$$

The above normalized values always lie between 0 and 1:

$$0 <= \eta <= 1$$

and the characteristic is better the higher its value.

For a given task, more than one performance characteristics must usually be taken into account. The most frequent case is when the most important performance characteristic is regarded as the target of optimatization i.e. as the 'objective function', whereas the others are taken into account as 'constraints', presented in the form of limiting conditions. This is (or may be) equivalent to a treatment in which a certain function of several performance characteristics is regarded as a composite performance characteristic.

Let us assume that we have $n$ different, normalized performance characteristics. They may be combined to form any number of composite (normalized) performance characteristics. The best known composite performance characteristics are the following:

Arithmetic mean:

$$\eta = \frac{\Sigma \eta_i^0}{n}$$

Weighted arithmetic mean:

$$\eta_w = \frac{\Sigma w_i \eta_i^0}{n}$$

Product:

$$\eta_\pi = \Pi \eta_i^0$$

Weighted product:

$$\eta_{\pi w} = \Pi (\eta_i^0)^{w_i}$$

Geometric mean:

$$\overline{\eta_\pi} = \sqrt[n]{\Pi \eta_i^0}$$

Weighted geometric mean:

$$\overline{\eta_{\pi w}} = \sqrt[n]{\Pi (\eta_i^0)^{w_i}}$$

Minimum:

$$\eta_{min} = \min \{\eta_1^0, \ldots, \eta_n^0\}$$

The runs of arithmetic mean, product and minimum performance characteristics can be seen, for two — characteristic cases, in *Fig. 6.4.*

More than one performance characteristic may also be compounded by means of the so-called multiple criteria optimization process.

### 6.4.2 Statistical characteristics

Several properties of analytical chemical methods may be characterized by statistical methods. This section deals only with the statistical characteristics of single component quantitative analysis, disregarding those of qualitative analysis, classification, the features of multicomponent analysis, the description of the

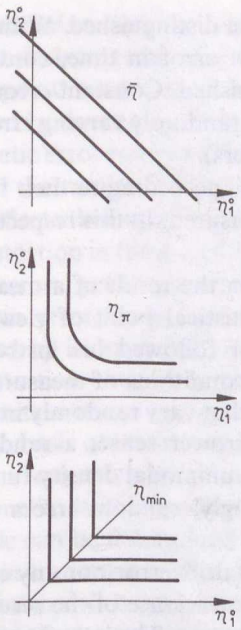



Fig. 6.4. Composite performance characteristics

analytical signals as a stochastic process, etc. Furthermore, we are not going to discuss such important particulars as limit of determination, reproducibility, repeatability, etc.

*The probability theoretical model of the result of measurement*

By result of measurement one generally means the value of a quantity to be measured as determined by the measurement. In the case of analytical chemical measuring systems, two kinds of measured data can be distinguished: analytical information and chemical information. The result of measurement is regarded as a random variable.

The true value is the theoretically obtainable quantity. It is regarded as being unknown in analysis and regarded as a known, constant value in the measurement of standard material.

Measurement error is the difference between the result of measurement and the true value, i.e.

$$\varepsilon^* = \xi - m^0$$

where $\varepsilon^*$ is the error, $\xi$ is the result of measurement and $m^0$ is the true value. Since the result of measurement is regarded as a random variable and the true value as a constant, the error of measurement is also a random variable.

Several types of error can be distinguished. Within one series of measurements, depending on the variation of error in time, constant, trend-like and randomly varying errors can be distinguished. Constant error is the systematic error, trend-like error is the drift error and randomly varying errors are the random and esoteric errors (extreme errors, blunders).

Errors may also be classified according to their functional relationship with the quantity (true value) to be measured. In this respect there are additive and proportional errors.

The errors superimposed on the result of a measurement may be caused by a number of factors. From statistical point of view, errors are treated under the assumption that the operator followed his instructions and the various other prescriptions concerning the conditions of measurement.

In a broad sense, all errors that vary randomly in time, including extreme errors, are random errors. In a narrower sense, a random error is an error varying randomly in time which has a unimodal density function with zero expected value and finite variance. Accordingly, random errors can be characterized by their variances (fluctuations).

If there is neither a varying drift error nor any extreme error superimposed on the result of measurement, the variance of the result of the measurement is equal to the variance of the random error. The latter is a measure of the precision of the measurement, provided that extreme and varying drift errors are absent.

The variance of random error is determined from the variance of the result of the measurement. The effects of random error can be reduced by averaging parallel measurements.

A type of randomly varying error that is scarcely to be defined is the extreme error. This error, also referred to as an outlier, a mistake, or a blunder, is a random error that does not necessarily have zero expected value, its value is generally significantly different from the true value, and it has a high variance.

By way of eliminating such errors, two types of extreme error can be distinguished. By means of feasibility checking one can filter out the results of measurement, containing random errors, which deviate by at least an order of magnitude from the true value. The statistical tests used for the filtering of outliers are able to eliminate results which fall within an order of magnitude of the true value, exceeding it only by some multiple of the variance.

The interpretation of extreme errors 'defined' by feasibility checks is more unanimous and justified than that of the such errors as may be spotted by outlier checks; from the statistical point of view the investigation of outliers is generally not justified.

The systematic error or distortion is a constant error, and thus cannot be regarded as a random variable. Instead, it is an unknown constant that has to be estimated, which produces a difference between the true value and the result of measurement. The systematic error appears in the expected value of the result of measurement, in the absence of extreme and drift errors, it can be interpreted as

the difference between the true value and the expected value of the result of measurement. If a sample with known true value is measured, and there is no extreme or drift error superimposed on the result of measurement, the effect of random error can be suppressed by averaging the results of several parallel measurements, and then the systematic error can be estimated as the difference between the true value and the mean of the results of measurement.

Whereas the random error reflects the imprecision of the result of measurement, the systematic error causes a distortion in the direction of expected value, i.e. a bias. The real error is a sum of the errors, yet their unambiguous common description is still impossible, since random error is characterized by its variance, whereas systematic error is characterized by its expected value.

Drift error is an error or distortion varying (slowly) in time, which is constant at a given instant, i.e. in this instant it can be regarded as a systematic error.

Drift may be observed on the basis of the investigation of the results of parallel measurements carried out over a longer stretch of time. Its detection requires time series analysis or control chart methods. Its presence is also detectable if true values are unknown, but its magnitude can be determined only on the basis of measurements on samples having known true values. Moreover, drift cannot be estimated on its own, but only in combination with systematic error. Its estimation is only slightly affected by the presence of extreme and random errors.

For chemical measurements, the additivity of the various types of error may often be assumed:

$$\varepsilon^* = \varepsilon + \varepsilon^s + \varepsilon^b + \varepsilon(t)$$

and thus the result of measurements has the form

$$\xi = m^0 + \varepsilon + \varepsilon^s + \varepsilon^b + \varepsilon(t)$$

where $\varepsilon$ is the random, $\varepsilon^s$ the systematic, $\varepsilon^b$ the esoteric and $\varepsilon(t)$ the drift error, i.e. the result of measurement is a sum of the true value and the random, systematic, extreme and drift errors.

The direct statistical characterization of this model for practical purposes is impossible. In order to move closer to a statistical characterization, two steps must be taken: the model should be simplified under certain conditions, and assumptions must be made on the distributions of random variables.

In the simplification it is usually justifiable to assume that if the measuring system is controlled, the system eliminates extreme errors, and regularly corrects drift and systematic errors by means of test measurements, and thus it is primarily the random error which should be characterized statistically.

Prior to making any assumption on distributions, it should be noted that since the calibration model produces a relationship between chemical and analytical information, no independent assumptions can be made on the distribution of two variables. In the practice of analytical chemical metrology, it is often correct to

assume that the random error of analytical information has a normal distribution. Under this assumption more powerful methods can be used for the statistical characterization of the results of measurements than if a normal distribution is not assumed.

In the following discussion, assuming the use of a controlled measuring system and the (approximate) normal distribution of random error, we shall deal with the hypothesis testing, interval estimation and comparison of the true value, precision and systematic error, taking account, too, of the statistical characterization of calibration lines.

## Hypothesis testing

This section deals with the hypothesis testing of the true value, systematic error and precision.

### *Hypothesis testing of true value*

If random errors are normally distributed, a correct statistical test of the hypothesis on the true value can be performed by means of the single-sample $t$ test, based on the so-called $t$ statistic.

First it is assumed that there is no systematic error in the measurement, i.e. the expected value of the result of measurement is the true value. This is expressed by the so called $H_0$ hypothesis:

$$H_0 : \mu = m^0$$

Concerning the true value an assumption is made; it is assumed that the composition of the (standard) sample is known, consequently all elements of the $t$ statistic are known, and thus

$$t = (\bar{m} - m^0)/s^* \cdot \sqrt{n}$$

assumes a certain calculated value. However, it is also known that for normally distributed random errors under the validity of the $H_0$ hypothesis the $t$ statistic follows Student's distribution. Therefore, hypothesis testing here means a comparison of the calculated value of the $t$ statistic derived according to the above equation with the tabulated value of Student's distribution belonging to a predetermined level of significance $1-\alpha$, and if the calculated value is lower than the tabulated one, the hypothesis is accepted, and rejected otherwise.

The acceptance of a hypothesis means, from the statistical viewpoint, that the given sample at the given level of significance does not contradict the hypothesis. However, this is not certainty, it is only a probability statement, and thus it may be false in two ways. It may happen that, on the basis of the given sample the hypothesis, although true, is rejected. This is called an error of the first kind. In contrast, if on the basis of the given sample, the hypothesis, although false, is accepted, an error of the second kind is committed.

The above consideration is easy to extend to the case when a constant systematic error, $\varepsilon^m$, of known magnitude is assumed. In this case the $t$ statistic has the form

$$t=(\bar{m}-(m^0+\varepsilon^m))/s^* \operatorname{sqr} n$$

and the subsequent treatment is the same as the above.

*Hypothesis testing of systematic error*

The testing of our hypothesis on systematic error is analogous to the hypothesis testing of true value.

If it is assumed that for normally distributed random errors the systematic error is $\varepsilon^m$, our hypothesis for the measurement of a sample with known $m^0$ value is

$$H_0: \mu=m^0+\varepsilon^m$$

This hypothesis can be tested by the $t$ statistic

$$t=(\bar{m}-(m^0+\varepsilon^m))/s^* \cdot \operatorname{sqr} n$$

as discussed above.

*Hypothesis testing of precision*

For normally distributed results of measurements, the hypothesis on the precision of the measurement can be tested by means of a $\chi^2$ test. In this case our hypothesis is

$$H_0: D^2[\xi]=\sigma^2$$

where $\sigma^2$ is the assumed variance. Now, to test the hypothesis, one calculates the statistic

$$\chi^2=n\times S^2/\sigma^2$$

If the hypothesis is correct, this statistic is of $\chi^2$ form with $n-1$ degrees of freedom. Therefore, the hypothesis is accepted if the value of the statistic is less than or equal to the tabulated value of the $\chi^2$ distribution, and is rejected if the calculated value is higher than the tabulated one.

**Interval estimations**

This section deals briefly with the fundamentals of the interval estimations of true value, systematic error and precision.

*Confidence interval of the true value*

The aim of the analytical chemist is the best possible estimation of the true value. If the result follows a normal distribution, then a confidence interval can be given which contains the true value in question with a given probability. The confidence interval of the true value is

$$P(\bar{m}-t_{n-1,\alpha}s^*/\sqrt{n}<=m^0<=\bar{m}+t_{n-1,\alpha}s^*/\sqrt{n})=1-\alpha$$

where $\bar{m}$ is the mean of the sample, $s^*$ is the standard deviation, $n$ is the number of parallel measurements, $m^0$ is the true value in question and $t$ is the value of Student's distribution. The equation expresses the fact that the statement in the brackets is true with a probability of $1-\alpha$.

Consequently, it is stated that if there is no systematic error, the true value falls into the interval $\bar{m}\pm t_{n-1,\alpha}\cdot s^*/\sqrt{n}$ with a probability of $1-\alpha$ where the value of $\alpha$ depends on the given analytical task.

In the calculation of the confidence interval, only the value of $t_{n-1,\alpha}$ requires some explanation. There is a theorem in statistics which states that if for a random variable with normal distribution the expected value is subtracted from the sample mean, divided by the standard deviation and multiplied by the square root of the number of samples, a random variable obeying the $n-1$ parameter Student's $t$ distribution is obtained. Therefore, in our case, the $t$ statistic:

$$t=(\bar{m}-m^0)/s^*\sqrt{n}$$

as random variable is an $n-1$ parameter Student's distribution. Thus, for any given pair of values of $n$ and $\alpha$ the appropriate value of $t_{n-1,\alpha}$ can be found in the table of Student's distribution and substituted into the above equation to obtain the required confidence interval.

As an illustration, the values of $t_{n-1,\alpha}$ are given for some of the values of $\alpha$ and $n=10$:

| $1-\alpha$ | 0.95 | 0.99 | 0.999 |
|---|---|---|---|
| $t_{n-1,\alpha}$ | 2.262 | 3.250 | 4.781 |

If systematic errors are also superimposed on the measured data, the interval estimate for the true value will be modified as follows

$$P(\bar{m}-\varepsilon^m-t_{n-1,\alpha}s^*/\sqrt{n}<=m^0<=\bar{m}-\varepsilon^m+t_{n-1,\alpha}s^*/\sqrt{n})=1-\alpha$$

where $\varepsilon^m$ is the systematic error.

So far we have assumed that, as is common in analytical chemistry, that the theoretical variance is unknown. If it were known, the confidence interval would

have been written in terms of $\sigma$ instead of $s$, and instead of the $t$ statistic the so-called $u$ statistic should have been used, which has a standardized normal distribution.

*Confidence interval of systematic error*

When a standard sample with an exactly known true value is measured, i.e. if it can be assumed that the true value of $m^0$ is accurately known, in an analogous way to the interval estimation of true value, a confidence interval of

$$P\left(\bar{m}-m^0-t_{n-1,\alpha}s^*/\sqrt{n}<=\varepsilon^m<=\bar{m}-m^0+t_{n-1,\alpha}s^*/\sqrt{n}\right)=1-\alpha$$

is obtained for the systematic error, $\varepsilon^m$.

*Confidence interval of the standard deviation of random error*

As has already been mentioned, if a normal distribution is assumed, the standard deviation is a 'good' estimate for the theoretical standard deviation, characteristic of the precision of the method. Similarly, on the basis of the standard deviation, assuming a normal distribution, an interval can be given within which the unknown theoretical standard deviation, i.e. the unknown square root of variance is contained with a probability of $1-\alpha$. One can derive that

$$P\left(\sqrt{n}s^*/\sqrt{c_2}<=\sigma<=\sqrt{n}s^*/\sqrt{c_1}\right)=1-\alpha$$

where constants $c_1$ and $c_2$ are the values of the $\chi^2$ distribution at $\alpha/2$ and $1-\alpha/2$, respectively, taken from the $\chi^2$ distribution table. The basis of this confidence interval estimation is the so-called $\chi^2$ statistic

$$\chi^2=ns^{*2}/\sigma^2$$

which, in the case of a sample having a normally distributed variable, obeys the $n-1$ parameter $\chi^2$ distribution.

Consequently, on the basis of the above equation, an asymmetric interval

$$K_1<=\sigma<=K_2$$

can be given which contains the standard deviation (precision) of the analytical measurements with a probability of $1-\alpha$. For the sake of unambiguity, when referring to such intervals, the values of $\alpha$ and $n$ should be given.

**Statistical comparisons**

This section deals with those methods of mathematical statistics that are applicable to the comparison of the precisions and systematic errors of measurement series (measuring methods), and of the true values of samples.

*Comparison of the precisions of series of measurement*

Let us assume that the $j$th series of measurements yielded the results

$$m_{j1}, m_{j2}, \ldots, m_{jn_j}$$

The task is to compare the precision and/or accuracy of this series to results

$$m_{k1}, m_{k2}, \ldots, m_{kn_k}$$

obtained in another series of measurements. A more general task is the comparison of the results of more than two series of measurements.

Obviously, a great number of such tasks can be found in analytical chemistry. When, for example, an analytical method is modified, the results of the original and modified measurements should be compared, the effects of the given experimental conditions may be studied, the temporal stability of a given instrument may be checked, etc.

The precisions, i.e. the standard deviations of random errors, can be compared, under the assumption of normal distribution, by the $F$ test for two series of measurements or by the so-called Bartlett test for more than two series.

The basis of the $F$ test is the $F$ statistic, which can be obtained from the equation

$$F = s_j^{*2}/s_k^{*2}$$

where subscripts $j$ and $k$ refer to the series to be compared. The ratio should be formed so that $F \geq 1$. Starting from the hypothesis

$$H_0 : \sigma_j = \sigma_k$$

if the hypothesis is true, then the $F$ statistic is a random variable obeying the $n_j - 1$, $n_k - 1$ parameter $F$ distribution. For hypothesis testing the value of $F$ is calculated, this is compared with the value of the $n_j - 1$, $n_k - 1$ parameter $F$ distribution corresponding to the significance level $\alpha$, and if the calculated value is lower than the tabulated one, the hypothesis is accepted, and rejected otherwise.

The discussion of Bartlett's test, which applicable to the comparison of more than two series, is omitted here: it can be found in statistical textbooks.

*Comparison of the systematic errors of series of measurement*

A frequently occurring problem in analytical chemistry is the comparison of the accuracy of certain series of measurements. If, for instance, it is known that one of the methods is free of systematic errors, the systematic errors or the accuracies of other methods can be estimated by comparison.

Depending on the conditions, there are various methods for the comparison of the systematic errors of measurement series. Assuming a normal distribution it is first of all further assumed that the theoretical variance is unknown (this is the most frequent case). This case can be divided into two sub-cases depending on whether

the (unknown) theoretical variance can be regarded as being the same or different in the two series.

In the former case, the two-sample $t$ test can be used. In this test the $t$ statistic has the form

$$t = \frac{\bar{m}_1 - \bar{m}_2}{\sqrt{(n_1-1)s_1^{*2} + (n_2-1)s_2^{*2}}} \sqrt{\frac{n_1 n_2 (n_1+n_2-2)}{n_1+n_2}}$$

The hypothesis is, then, that

$$H_0 : \varepsilon_1^{\mathrm{m}} = \varepsilon_2^{\mathrm{m}}$$

The run of the test is the same as of the single-sample test, the only difference being that the parameter of Student's distribution is $n_1+n_2-1$.

In analytical chemistry, the most frequent problem is a comparison of the accuracies of two normally distributed series of measurement with unknown and different theoretical variances. In this case the more complicated Welch test should be applied, which is based on the $t$ statistic

$$t = \frac{\bar{m}_1 - \bar{m}_2}{\sqrt{\frac{s_1^{*2}}{n_1} + \frac{s_2^{*2}}{n_2}}}$$

which, if the hypothesis

$$H_0 : \varepsilon_1^{\mathrm{m}} = \varepsilon_2^{\mathrm{m}}$$

is true, obeys Student's $t$ distribution approximately. The $t$ distribution has $f$ degrees of freedom, where the number of the degrees of freedom can be calculated from the equation

$$f = \frac{1}{\frac{c^2}{n_2-1} + \frac{(1-c)^2}{n_1-1}}$$

where constant $c$ can be determined from the sample according to the equation

$$c = \frac{s_2^{*2}}{n_2}\left(\frac{s_1^{*2}}{n_1} + \frac{s_2^{*2}}{n_2}\right)$$

The rest of the procedure of hypothesis testing is the same as in the single-sample $t$ test.

It should be noted that in the analytical chemical literature, the so-called two-sample $u$ test can also be found, which can be used to compare the systematic errors of two series of measurement if they are normally distributed and their theoretical variances are known. Since in our opinion this cannot be regarded as a case that occurs frequently, the test is not discussed here.

To compare the accuracies of more than two, normally distributed series of measurement with the same but unknown theoretical variances, the methods of variance analysis can be applied (with certain further restrictions).

*Comparison of true values*

An everyday task of analytical chemistry is the comparison of samples. The identity of two samples is determined in the same manner as systematic errors are compared.

The statistical determination of the identity of two substances may be performed by means of the two-sample $t$ test applied on measurements obeying normal distributions with the same but unknown variance. The procedure is the same as with systematic errors, assuming here the hypothesis

$$H_0 : m_1^0 = m_2^0$$

If the two samples are measured in different measuring systems, then, if the standard deviations are the same, the two-sample $t$ test is still applicable, whereas if the variances are different, the Welch test can be applied as was discussed in connection with the comparison of systematic errors.

The identity of two samples can also be determined if the two measuring systems have systematic errors, but in this case, with either the two-sample $t$ test or the Welch test, the sample means should be corrected for systematic errors before the test.

## The statistical characterization of the calibration line

The basis of single-component, quantitative chemical analysis is the assumption that there exists a relationship represented by a continuous function between the quantity of substance, $x$ (chemical information) and the analytical property obtained by the measurement, $y$ (analytical information). This function is called the calibration function, and its inverse is called the evaluation function or the inverse calibration function.

In the determination of calibration functions, the main problems arise from the fact that the measurement of both variables (quantity, $x$ and analytical property, $y$) is prone to experimental error. In most cases it can be assumed that the error in the measurement of quantity is negligible, and thus that $x$ is accurate, whereas the analytical variable $y$ contains an error. It is assumed that this error has a continuous, symmetrical distribution with zero expected value and finite variance, and it is also assumed that the measurements are independent.

*Estimation of calibration line*

It may often be assumed that the calibration function is linear, or at least that in a given interval it can be well approximated by a linear function. Then,

$$y = a + bx$$

is the theoretical calibration function, regarded as a true calibration function, in which parameters $a$ and $b$ are, according to the usual interpretation, the intercept and slope of the line, respectively.

Parameters $a$ and $b$ may be estimated on the basis of the data pairs $\{x_i, y_i\}$, $i=1, \ldots, N$ obtained from the calibration measurements by means of the method of least squares, i.e. through the condition

$$\min_{a,b} \sum_{i=1}^{N} (y_i - \hat{a} - \hat{b}x_i)^2$$

Solving the normal equations

$$\sum_{i=1}^{N} (y_i - \hat{a} - \hat{b}x_i) = 0$$

and

$$\sum_{i=1}^{N} x_i (y_i - \hat{a} - \hat{b}x_i) = 0$$

one obtains

$$\hat{b} = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \hat{c}) \Big/ \sum_{i=1}^{N} (x_i - \bar{x})^2$$

and

$$\hat{a} = \hat{c} - \hat{b}\bar{x}$$

where

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

and

$$\hat{c} = \frac{\sum_{i=1}^{N} y_i}{N}$$

Thus, as an estimate for the calibration line we have the experimental calibration line

$$\hat{y} = \hat{a} + \hat{b}x$$

or, after transformation,

$$y = \hat{c} + \hat{b}(x - \bar{x})$$

Consequently, the least squares estimation of the parameters of the calibration line provides parameters $\hat{a}$ and $\hat{b}$ for which the sum of the squares of the vertical

(along axis $y$) $d_i$ distances between the points $\{x_i, y_i\}$, $i=1, \ldots, N$ and the line $\hat{a}+\hat{b}x$ is a minimum, where the sum is extended to all ($N$) measurements.

The experimental calibration line can be used to give an estimate for the analytical property belonging to a given concentration. Thus, the estimate $\hat{y}^*$ of $y^*$ belonging to a given $x^*$ is

$$\hat{y}^* = \hat{c} + \hat{b}(x^* - \bar{x})$$

The variance of this estimate is

$$D^2[\hat{y}^*(x^*)] = \sigma^2 \left[ 1/N + (x^* - \bar{x})^2 \Big/ \sum_{i=1}^{N} (x_i - \bar{x})^2 \right]$$

The confidence interval of the estimate $\hat{y}^*$ belonging to a given $x^*$ can be determined on the basis of the random variable

$$t_{y^*} = \frac{y^* - \hat{y}^*}{s^* \sqrt{1/N + (x^* - \bar{x})^2 \Big/ \sum_{i=1}^{N} (x_i^* - \bar{x})^2}}$$

which follows a Student distribution with $N-2$ degrees of freedom, by means of the relationship

$$\hat{y}^* - t_{1-\frac{\alpha}{2}} s^* \sqrt{\frac{1}{N} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}} \leq y^* \leq \hat{y}^* + t_{1-\frac{\alpha}{2}} s^* \sqrt{\frac{1}{N} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}}$$

This latter relationship, which represents the confidence limits of $y^*$ for any point $x^*$, may be given graphically as a hyperbola with asymptotes passing points $\bar{x}$, $\hat{c}$ calculated from the calibration points $\{x_i, y_i\}$, $i=1, \ldots, N$.

It can be seen from the above relationship that a reliable estimate can be obtained for a point $y^*$ belonging to given $x^*$ if the residual error is small, the number of calibration measurements is large, the calibration measurements are far from the 'point of gravity' $\bar{x}$, and the given point $x^*$ is close to the point of gravity, $\bar{x}$.

The aim of the determination of the calibration function is generally the construction of the analytical evaluation function, or inverse calibration function, and on the basis of the latter, the estimation of the concentration of the unknown sample from its known analytical signal.

The inverse calibration line is of the form

$$x = (y - \hat{c})/\hat{b} + \bar{x}$$

An equivalent expression is

$$x = (y - \hat{a})/\hat{b}$$

On the basis of the inverse calibration line, it is possible to estimate the unknown concentration, $x^*$ on the (also unknown) analytical information of which ($y^*$) the $M$ parallel measurements $y_1^*, \ldots, y_M^*$ were performed.

In this case the value of $\hat{x}^*$ used for the estimation of $x^*$ can be given as

$$\hat{x}^* = (\hat{y}^* - \hat{c})/\hat{b} + \bar{x}$$

where

$$\hat{y}^* = \left( \sum_{i=1}^{M} y_i^* \right) \Big/ M$$

It is assumed that the variance of the observations on $y^*$ is the same as that of the calibration measurements, thus

$$D^2[\hat{y}^*] = \sigma^2/M$$

For the difference of $\hat{y}^*$ and estimate $y^*$ one may write

$$E[\hat{y}^* - y^*] = 0$$

$$D[\hat{y}^* - y^*] = \sigma^2 [1/M + 1/N + (x^* - \bar{x})^2 / \sum_{i=1}^{N} (x_i - \bar{x})^2$$

The statistic

$$t = \frac{\hat{y}^* - y^*}{s^* \sqrt{\frac{1}{M} + \frac{1}{N} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}}}$$

has Student's distribution with $N$—2 parameters, which stems from the fact that for the estimation of $\sigma$ two of the $N$ independent measurements were constrained by the calculation of the slope and intercept.

Using this statistic, the following interval estimate can be given for the unknown concentration, $x^*$:

$$\hat{x}^* - t_{1-\frac{\alpha}{2}} \cdot \frac{s^*}{\hat{b}} \sqrt{\frac{1}{M} + \frac{1}{N} + \frac{(\hat{x}^* - \bar{x})^2}{\Sigma (x_i - \bar{x})^2}} \leq x^* \leq \hat{x}^* +$$

$$+ t_{1-\frac{\alpha}{2}} \frac{s^*}{\hat{b}} \sqrt{\frac{1}{M} + \frac{1}{N} + \frac{(\hat{x}^* - \bar{x})^2}{\Sigma (x_i - \bar{x})^2}}$$

where the estimate of $\hat{x}^*$ was obtained from the inverse of the calibration line. This relationship gives a good approximation for low values of $s^*/\hat{b}$. The exact expression for the confidence interval is

$$\bar{x}+\frac{\hat{y}^*-\hat{c}}{\hat{b}\gamma}-\frac{\sqrt{1-\gamma}}{\gamma}\sqrt{\left(\frac{1}{M}+\frac{1}{N}\right)\gamma\sum_{i=1}^{N}(x_i-\bar{x})^2+\frac{(\hat{y}^*-\hat{c})^2}{\hat{b}^2}}\leq x^*\leq$$

$$\leq\bar{x}-\frac{\hat{y}^*-\hat{c}}{\hat{b}\gamma}+\frac{\sqrt{1-\gamma}}{\gamma}\sqrt{\left(\frac{1}{M}+\frac{1}{N}\right)\gamma\sum_{i=1}^{N}(x_i-\bar{x})^2+\frac{(\hat{y}^*-\hat{c})^2}{\hat{b}^2}}$$

where

$$\gamma=1-\frac{t_{1-\frac{\alpha}{2}}\cdot s^{*2}}{\hat{b}^2\sum_{i=1}^{N}(x_i-\bar{x})^2}$$

In analytical chemistry, this confidence interval is extremely significant, since it can be used to estimate an unknown quantity of substance on the basis of the given calibration measurements and the parallel measurements on the sample of unknown composition.

### 6.4.3 Information theoretical characteristics

The information theoretical treatment of analytical chemical systems is of fundamental importance since it allows one numerically to characterize the amount of information produced by analytical chemical systems.

In this section, the fundamentals of information theory will first be reviewed and then the analytical chemical systems will be characterized by means of the theory.

#### Fundamentals of information theory

In a broad sense, information is applied in its everyday sense, meaning communication, knowledge, news, data, etc. Various aspects of information exist, one may discuss its content, truth, comprehensibility, value, the volume of information carried, etc.

Information theory (in the mathematical sense) deals only with the 'amount' of information, disregarding all other aspects. Information theory is now a huge field of science, which, on the basis of the interpretation of the amount of information, includes the theory of telecommunication systems as well as the theory of coding/decoding.

*Information, amount of information*

In information theory, information is the marking, selection or occurrence of one of a set of known, given possibilities, events or statements. Regarding a set $\mathscr{A}$ of events $A_1, \ldots, A_n$, where $A_i \in \mathscr{A}$ $(i=1, \ldots, n)$, information is defined as the selection of one of the elements of this set.

It is assumed below that the set of events, $\mathscr{A}$, forms a complete system of events. Let us assign to the complete system of events a discrete probability distribution, $p_1, \ldots, p_n$, where $p_i = P(A_i)$, $i = 1, \ldots, n$, for which

$$\sum_{i=1}^{n} p_i = 1$$

On the basis of the above concepts, in the case of a discrete probability distribution, the amount or unexpectedness of individual item of information $A_i$ is defined as

$$I^d(A_i) = \log(1/p_i) = -\log p_i$$

If the base of the logarithm in the formula is 2, the unit of the amount of individual information $I(A_i)$ is the bit; if its base 10, the unit is called a Hartley; and if its base e, the unit is called a nat.

For continuous probability distribution, information can be described, instead of the numbers $p_i$ characteristic of discrete distribution, by the distribution function $F(x)$ or, if it exists, the density function $f(x)$ of the distribution. Then, using the density function of continuous random variable, the amount of individual information can be given in a similar way as

$$I^c(x) = -\log f(x)$$

Since, however,

$$p(x) \approx f(x)dx$$

the amounts of discrete and continuous individual information cannot be regarded as identical.

On the basis of the above the areas of application of information theory can clearly be defined. Information theory is applicable only when the set of possibilities is given, is known a priori, and the possibilities may be observed in principle as many times as necessary, i.e. a random mass event corresponds to them, and thus the probability distribution of the set of possibilities can be estimated.

### *Entropy*

Whereas individual events are characterized by the amounts of individual information, a given system of events can be described by its entropy. The entropy or mean information content of a discrete system of events is the expected value of the individual amounts of information of the system:

$$H^d(\mathscr{A}) = \sum_{i=1}^{n} p_i I^d(A_i)$$

i.e.

$$H^d(\mathscr{A}) = \sum_{i=1}^{n} p_i \log(1/p_i) = -\log \sum_{i=1}^{n} p_i \log p_i$$

The entropy for continuous random variables is defined as

$$H^c(\xi) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx$$

This cannot be regarded as a generalization of the discrete case, since

$$p(x) \log p(x) \approx f(x) \log [f(x)]\, dx + f(x) \log [dx]\, dx$$

The properties of the entropies belonging to the two systems are shown in *Table 6.3*, from which the different properties of the two types are obvious.

The entropy of a system of events can be regarded as the extent of the uncertainty of the probability distribution belonging to the system.

We have been concerned so far with one dimensional random variables only. If variables $\xi$ and $\eta$ are regarded simultaneously, further kinds of entropy may be defined, the joint entropy $H(\xi, \eta)$ and the conditional entropy $H(\xi|\eta)$. The definitions are given in *Table 6.4*.

These entropies obey the following relationships:

$$H(\xi, \eta) = H(\eta, \xi)$$

$$H(\xi, \eta) = H(\xi|\eta) + H(\eta) = H(\eta|\xi) + H(\xi)$$

$$H(\xi) >= H(\xi|\eta)$$

$$H(\eta) >= H(\eta|\xi)$$

### *Transinformation*

A very important concept of information theory is transinformation, which can be regarded as a measure of interrelation between two event spaces.

The definition of individual transinformation is

$$I(x_i; y_j) = \log \left(p(x_i|y_j)/p(x_i)\right)$$

which can be rearranged as

$$I(x_i; y_j) = \log \left(1/p(x_i)\right) - \log \left(1/p(x_i|y_j)\right)$$

Table 6.3. The properties of the entropies of discrete and continuos distributions

| Property \ Distribution | Discrete | Continuous |
|---|---|---|
| set of values | $0 <= H$ | $0 <> H$ |
| finiteness | finite | may be infinite |
| effect of transf. | invariant | noninvariant |

Table 6.4. Definitions of joint and conditional entropies

| Entropy \ Distribution | Discrete | Continuous |
|---|---|---|
| $H(,)$ | $-\sum_i \sum_j p_{ij} \log p_{ij}$ | $-\int_{-\infty}^{\infty} \int f(x, y) \log f(x, y)\, dx\, dy$ |
| $H(\mid)$ | $-\sum_i \sum_j p_{ij} \log p_{i\mid j}$ | $-\int_{-\infty}^{\infty} \int f(x, y) \log f(x \mid y)\, dx\, dy$ |

i.e.

$$I(x_i; y_j) = I(x_i) - I(x_i \mid y_j)$$

Consequently, individual transinformation is the difference between the *a priori* individual information concerning event $x_i$ and the a posteriori individual information (with known $y_j$) of event $x_i$.

It is noted that an equivalent definition of transinformation is

$$I(x_i; y_j) = \log \left(p(x_i, y_j)/p(x_i)p(y_j)\right)$$

It follows from the definition that

$$I(x_i; x_i) = I(x_i)$$

i.e. the individual transinformation taken for the same variable gives the amount of individual information.

Mean transinformation is the expected value of the individual transinformations, i.e. it can be defined as

$$I^{d}(\xi; \eta) = \sum_i \sum_j p(x_i, y_j) \log \left(p(x_i, y_j)/p(x_i)p(y_j)\right)$$

or, for continuous distributions

$$I^{c}(\xi; \eta) = \int_{-\infty}^{+\infty} \int f(x, y) \log \left(f(x, y)/f(x)f(y)\right) dx\, dy$$

Mean transinformation defined in the above manner is a possible measure of the interdependence of two probability distributions, since it has the property that

$$I(\xi; \eta) = 0$$

if and only if the two variables are independent, whereas if $\xi$ is a unique function of $\eta$, i.e. if

$$\xi = g(\eta)$$

then

$$I(\xi;\eta)=H(\xi)$$

It follows from the definition of mean transinformation that

$$I(\xi;\eta)>=0$$

$$I(\xi;\eta) \quad =I(\eta;\xi)$$

and

$$I(\xi;\xi) \quad =H(\xi)$$

The differences between the properties of discrete and continuous mean transinformations can be seen from *Table 6.5*.

Table 6.5. Properties of the mean transinformation of discrete and continuous distributions

| Property \ Distribution | Discrete | Continuous |
|---|---|---|
| Set of values | $0<=I^d$ | $0<=I^c$ |
| Finiteness | finite | generally finite |
| Effect of transformation | invariant | invariant for linear transformation |

It can be seen on the basis of *Table 6.5*, that, in contrast to entropies, the differences between the two types of transinformation are less significant.

The connections between mean transinformation and the various types of entropy can be given by the following relationships:

$$I(\xi;\eta)=H(\xi)-H(\xi|\eta)$$

$$I(\xi;\eta)=H(\eta)-H(\eta|\xi)$$

$$I(\xi;\eta)=H(\xi)+H(\eta)-H(\xi;\eta)$$

In practice, mean transinformation is generally calculated by the first two equations, since the monovariate and conditional distributions are usually known, whereas the joint distribution necessary for the calculation of $H(\xi,\eta)$ can very seldom be estimated.

*Communication systems*

The most important application field of information theory is communication theory; several basic concepts of information theory have their origin in its application in this field.

The definition of communication systems is very general; any kind of communication, semiotic process, measurement or observation can be regarded as a communication system.

Let us denote the signals arising from the transmitter by random variable $\xi$. A given value of this variable, if discrete, is called letter, and a given (finite) string of these letters is called a word or a communication. If the successive letters are independent, the system has 'no memory', whereas if they are dependent, the system has a 'memory'. The received signals are denoted by random variable $\eta$. With this notation, a communication system can be characterized by the following entropies:

$H(\xi)$: mean information content pro letter of transmitter
$H(\eta)$: mean information content pro letter of receiver
$H(\xi, \eta)$: uncertainty of communication system
$H(\xi|\eta)$: uncertainty of transmitter under known received signal
$H(\eta|\xi)$: uncertainty of receiver under known transmitted signal

In addition to the above entropies, the system can also be characterized by mean transinformation, which can also be taken as the information transmitted by the system. Transinformation is indeed characteristic of a communication system since, for a noise-free channel

$$I(\xi; \eta) = H(\xi) = H(\eta) = H(\xi, \eta)$$

whereas for a channel with independent input and output

$$H(\xi; \eta) = 0$$

**Characterization of analytical chemical systems by means of information theory**

As was discussed above, analytical chemical investigations and their measurements can be regarded as analytical chemical systems in which analytical information is produced by the analytical investigation unit via processing observed or measured analytical signals, from which chemical information is obtained by inference.

The applications of information theory in analytical chemistry may be divided into two main categories. On the basis of information theory one may characterize, on the one hand, the amount of analytical and chemical information by means of entropy, and on the other hand the amount of information produced by analytical chemical systems can be given, by treating these systems as communication systems, by means of transinformation.

Of course, the precondition of the applicability of information theory is that the possible analytical and chemical statement have been defined in advance, and that the outcomes of these statements can be observed arbitrarily many times, in order to estimate the probability distribution of the system of statements.

If the probability distribution is unknown, a preferred strategy is to assume the highest disorder, i.e. the distribution which has the highest possible entropy under

the given conditions. Thus, if there is no other prior knowledge, one should assume for the case of (discrete) qualitative statements and for a random variable defined over a finite domain an even (rectangular) distribution, an exponential distribution for a variable defined over the nonnegative domain of the real numbers, and a normal distribution for a variable defined over the complete real domain.

*Amount of analytical and chemical information*

Information theory can be used to characterize the amount of analytical information produced by the analytical investigating system and that of the chemical information produced by analytical chemical correlation. This requires the amount of individual information for individual analysis and entropy for an analytical method. For this, the given systems of events should be defined properly and uniquely, for both analytical and chemical information.

In this respect, the 'fingerprint' type of analytical information, such as is incorporated, for instance, in data banks, has been the cause of debates and misunderstanding. Some authors regard the spectrum or chromatogram, when used as 'fingerprint', as a composite density function, and accordingly a calculated 'entropy' is formally assigned to it. We feel that this interpretation covers no properly defined system of events, and thus the calculation has no basis in information theory. In such cases analytical information should be defined rather on the basis of a system of statements in which the possibilities are the 'fingerprints' actually present and distinguishable in the data bank.

We do not agree with the information theoretical treatment of signals, either, since a 'signal' is not 'information', but only an 'information carrier', and thus in our opinion no information can be assigned to a signal.

It is also evident that the same 'signal', i.e. a chromatogram, may be the carrier of different types of analytical (and chemical) information, depending on the manner of data processing. At one extreme, it can be used as a 'fingerprint' for the qualitative characterization of a single mixture, while at the other extreme it also serves for the quantitative analysis of all components. In these two extreme cases the signals are identical, but the analytical information (and, consequently, the chemical information) will be entirely different in nature.

It should be noted that in either qualitative or quantitative analysis, arising from the applicability of information theory, one may discuss the choice on the basis of pre-set possibilities only, i.e. the possible statements must be known *a priori*. Accordingly, in the application of information theoretical models defined in this way the discovery of a new substance is impossible, 'identification' means here a choice from among a set of given, known substances. Thus, it may be stated, for instance, that the sample behaves like (similarly to) a given, already known substance, and if its behaviour is different from those already known (and considered in the particular problem), then the given system of statements and thus the information theoretical model should be modified.

*Analytical chemical systems as communication systems*

Communication systems transmit information, thus both their input and their output consist of information. In order to interpret analytical chemical systems or certain parts of them as communication systems, it is necessary to define input and output information appropriately.

The definition of information output is evident in all cases, since the output of an analytical investigation is analytical information and that of the complete analytical chemical system is chemical information. Problems may emerge, however, in the definition of the inputs of analytical chemical systems regarded as communication system. It is only the input to the analytical chemical correlation system (analytical information) that is information in the usual sense. In the other systems the basis of the definition of information input is that the transinformation of the communication system gives the amount of information produced by the measurement (observation); therefore, only that *a priori* chemical or analytical information is regarded as input to the analytical chemical system, which was at our disposal before (or without) the measurement (observation).

According to this definition, analytical chemical systems or certain parts of them may be regarded as communication systems. The complete analytical chemical system can be regarded as a communication system the input of which is the *a priori* chemical information concerning the substance and the output is the *a posteriori* chemical information concerning the substance after the measurement (observation). In principle, a complete analytical chemical system can also be regarded as a communication system of which the input is the *a priori* analytical information, this, however, has no practical significance.

Analytical investigation can be regarded as a communication system with an input of *a priori* knowledge on analytical (or chemical) information and an output of analytical information after measurement (observation).

Analytical chemical correlation can be regarded as a communication system with an input of analytical information after measurement (observation) and an output of chemical information after measurement (observation). Here, as was mentioned, no *a priori* information should be defined.

Of course, the inputs and outputs of analytical chemical systems (or subsystems) regarded as communication systems cannot be taken as being independent; neither in qualitative nor in quantitative analyses. It is therefore worth discussing the system model that describes the relationship between them.

In the case of classical qualitative analysis, the general expression of transinformation usually has a simpler form. If it can be assumed that before analysis there are $m^b$ possibilities of equal probability for the identity of the substance unknown to us, and after analysis, i.e. once we are in possession of analytical information, there remain $m^a$ equally probable possibilities for the identity, a special form of transinformation will be

$$I = \log (m^b/m^a)$$

Another group of qualitative analysis comprises the modern methods, usually based on data banks, which produce chemical information by means of logico–algebraical methods from the given analytical information. Such a method is, for example, the determination of functional groups (chemical information) from spectroscopic data (analytical information). With these methods, the inputs to the communication system are the possible characteristic features of the spectra, the outputs are the possible functional groups of the compounds, and thus the model of the communication system is the logical, spectral feature vs. functional group correlation matrix.

Assuming a certain probability distribution for the spectral features and for the functional groups, the joint probability distribution characteristic of the communication system can be determined on the basis of the logical correlation matrix and the given (marginal) distributions. On the basis of the probability distributions, the transinformation of the system can be determined from

$$I = \sum_i \sum_j p_{ij} \log (p_{ij}/p_i \cdot p_j)$$

where $p_i$, $p_j$ and $p_{ij}$ denote the probabilities of the occurrence of feature $i$, functional group $j$ and their joint occurrence, respectively.

Transinformation is an efficient measure of the information produced by the system, since if the groups and features are independent, $I=0$, whereas maximum transinformation is obtained in correlation if there is a one-to-one assignment between them, i.e. if in matrix $p_{ij}$ only the diagonal elements are non-zero.

In quantitative analysis, the relationship defined by the assumed model of the communication system determines the distribution of the dependent variable. Assuming the validity of the additive noise model, i.e. that the output is the sum of the input and noise, the density function $f_\eta(x)$ of output $\eta$ will be the convolution of the density functions $f_\xi(x)$ of the input, $\xi$, and $f_\varepsilon(x)$ of the error, $\varepsilon$:

$$f_\eta(x) = f_\xi(x) \times f_\varepsilon(x)$$

where " $\times$ " denotes convolution. Thus, for the additive noise model, mean transinformation can be calculated from

$$I(\xi; \eta) = H(\xi \times \varepsilon) - H(\varepsilon)$$

In principle, the above considerations apply to the information theoretical description of any communication system, and thus applies equally to the complete analytical chemical system or to its subsystems. However, in the case of analytical chemical systems, a sharp distinction should be made between analytical investigation, which produces analytical information, and the complete analytical system including correlation, which produces chemical information. Analytical information always contains less uncertainty than chemical information, since the latter contains the uncertainties of both calibration and analytical chemical correlation.

Therefore, the entropy of chemical information is generally higher than that of analytical information, and in the evaluation of chemical information the uncertainty of calibration should also be taken into account.

*Conclusion*

Analytical chemical systems, as information sources, can also be characterized by the methods of information theory. This, owing to the quantitation of the information produced, is very significant. The amounts of resulting analytical and chemical information can be characterized by the amounts of individual information and by mean information, called entropy. Analytical chemical systems and certain subsystems of them may also be regarded as communication system, and their information production may be characterized by their mean mutual information, called transinformation. A precondition of the information theoretical treatment of analytical chemical systems is that the operation of the system can be characterized statistically.

### 6.4.4 Decision theoretical characteristics

The decision theoretical characterization of analytical chemical methods is extremely important since the concept of loss enables one to link statistical and economical characteristics.

This section is concerned primarily with the fundamentals of statistical decision theory, and then with the decision theoretical treatment of analytical chemical methods.

**Basic concepts of statistical decision theory**

The parameter space describes the status of the system. A state parameter is a random variable, and its possible values span the parameter space. The state parameter may be a scalar or vector quantity, which can be characterized by its distribution function.

The uncertainty of the random variable which describes the state of the system can be decreased by means of information obtained by mathematical statistical observation. The statistical observation of the system yields the statistical sample, which is also a random variable, and its possible values span the sample space. The relationship between the state parameter and the sample can be described by a conditional distribution function.

The elements of the decision space are the possible actions or decisions. The relationship between decision space and sample space is given by the decision function, which assigns an element of the decision space to the elements of the sample space. The most important aim of statistical decision theory is the determination of the 'best' decision function, where 'best' relates to the consequence of decision.

The consequence of a decision can be given by the loss, more particularly by the real, nonnegative loss function. Loss depends on two variables, the state variable and the decision variable. Since the state variable is a random variable, the value of the loss function is also a random variable.

It follows from the definition of the decision function that in the case of fixed decision function the loss function depends on the state variable and the sample. Loss is generally defined as negative gain. Loss in general is regarded as an abstract quantity, in the case of economic problems it can be expressed in monetary units.

Depending on the given type of decision, various kinds of loss function can be defined. If the parameter and decision spaces assume discrete values, the loss function can be given as a 'loss matrix'. If the parameter and decision spaces are continuous, the loss function forms a multidimensional surface.

The risk function is the expected value of the loss function with respect to the sample. In the case of fixed decision function, risk is a function of the state parameter only. For the construction of the loss function, one should know the conditional distribution function of the sample according to the state parameter.

For different decision functions the loss function will also be different, i.e. with the application of different decision functions implies that different risk values will be obtained for the same parameter values, and thus the decision functions may be compared on the basis of the loss.

The optimum decision function can be selected on the basis of various criteria. The most widespread optimization criteria are discussed below.

If a decision function exists which, as compared to all other decision functions, has the lowest risk for any parameter value, this function is said to be a unanimously best decision function, and all other decision functions are not acceptable. Therefore, a decision function is acceptable only if there is no unanimously better decision function.

If there is no unanimously best decision function, one may introduce for the characterization of the decision functions the concept of expected risk, which is the expected value of the risk function for the given state parameter space and fixed decision function. On the basis of the expected risk, the decision function with the lowest mean risk can be regarded as an optimum. The minimum of the expected risk is called Bayesian risk.

Another method for selecting decision function is the minimax method. The basic principle is that a maximum of 'malevolence' is assumed from the opponent, i.e. the least favourable state parameter is always assumed. This means that the state parameter always has such a value for which the risk is maximum. In this case the optimum decision function is the one for which the maximum risk according to the parameter space is the lowest.

This minimax method provides the maximum security against large losses, but is not optimal in the sense that the method is based on the worst case instead of the one, that may be expected.

## Decision theoretical characterization of analytical chemical systems

During the opinion of analytical chemical systems five types of decision systems may be defined:

— decision concerning analytical information, obtained during signal processing on the basis of the analytical signal;
— decision concerning chemical information, obtained during analytical chemical correlation (inference) on the basis of analytical information;
— decision concerning chemical information, obtained during signal interpretation on the basis of the signal;
— decision concerning the status of the high-level system that applies the analytical chemical system, obtained by the user on the basis of chemical information;
— decision concerning the status of the high-level system that applies the analytical chemical system, obtained by the user on the basis of analytical information.

The various decision systems are summarized in *Table 6.6*.
These decisions will now be discussed on the basis of statistical decision theory.

### *Analytical signal processing as decision*

Analytical investigation, which assigns analytical information to the sample, can be characterized in decision theory as follows. The parameter space of the decision model is the set of analytical information, i.e. the quantitative or qualitative statements concerning the analytical properties of the sample. The aim of analytical investigation is the determination of analytical information, i.e. the actual values of some parameters. With this aim analytical signals are generated which carry information on the parameters; the possible analytical signals form the sample space.

Our decision, i.e. our statement concerning the analytical properties of the sample, is taken on the basis of the analytical signal obtained. The possible decisions, the estimated analytical information, span the decision space. According to this interpretation, state parameter space and decision space, representing the spaces of unknown and estimated analytical information, are identical, whereas the signal space, representing a sample space, is different.

The decision function of the decision theoretical model is the operation of signal processing. Analytical information is a function of the analytical signal, and analytical information can be acquired only if the prescription of signal processing are known.

The quality of analytical information may be characterized by the decision theoretical model. Analytical investigation is prone to errors, and the error of the

Table 6.6. Various decision systems

| System characterized by risk | Parameter space | Subsystem regarded as sampling | Sample space | Subsystem regarded as decision | Decision space | Decision function |
|---|---|---|---|---|---|---|
| Analytical subsystem | analytical information of sample | analytical sign production | analytical sign | analytical sign processing | estimated analytical information | signal processing |
| Analytical chemical system | chemical information of sample | investigating subsystem | analytical information | analytical chemical inference | estimated chemical information | inverse calibration function |
| Analytical chemical system | chemical information of sample | analytical sign production | analytical sign | analytical sign interpretation | estimated chemical information | signal interpretation |
| User system | user parameter of sample | analytical chemical system | chemical information | user decision | estimated user parameter of sample | special |
| User system | user parameter of sample | analytical chemical system | analytical information | analytical chemical inference and user decision | estimated user parameter of sample | special |

estimation of analytical information may be characterized by the loss function. The expected value of loss as a function of an analytical signal, i.e. the risk or expectable loss, is then a decision theoretical measure of the quality of an analytical investigation.

*Analytical chemical correlation as decision*

The full analytical system may be characterized in two ways from the aspects of decision theory. Analytical chemical correlation and analytical chemical signal interpretation may both be regarded as decisions.

Analytical chemical correlation, which assigns chemical information to analytical information, may be characterized in decision theory by the following model.

The parameter space of a decision model is a collection of chemical information, i.e. qualitative or quantitative statements on the chemical properties of the sample. The aim of the operation of the analytical chemical system is the determination of the actual values of parameters, which is chemical information. With this aim analytical information is generated during an analytical investigation, and this spans the sample space. Our decision or statement on the chemical properties of the sample is taken on the basis of the analytical information obtained. The possible decisions, again chemical information, form the decision space. In this interpretation state parameter space and decision space are identical, both representing chemical information, whereas the space of analytical information as sample space is different.

The decision function of the decision theoretical model corresponds to analytical chemical correlation. Chemical information is a function of analytical information, this function is determined during calibration.

The quality of an analytical chemical system may be characterized by its decision theoretical model. The operation of an analytical chemical system is prone to error, the error of the estimation of chemical information may be characterized by the loss function. The expected value of loss as a function of analytical information, i.e. the risk or expected loss is, then, a decision theoretical measure of the quality of an analytical chemical system.

*Analytical chemical signal interpretation as decision*

From the perspective of decision theory, analytical chemical systems may also be characterized by regarding not only analytical chemical correlation but the complete process of analytical chemical signal interpretation as a decision process.

The state parameter space of the decision model is in this case, too, the set of chemical information relevant to the chemical properties of the sample. The analytical signal is generated by the analytical signal production unit, and the signals span the sample space. Our decision, chemical information, is taken on the

basis of the analytical signal obtained. The possible decisions, chemical information, form the decision space. State parameter space and decision space are identical, both representing chemical information, sample space, the space of analytical signals, is different.

The decision function of the model corresponds to complete analytical chemical signal interpretation. Of this, signal processing was determined on the basis of prior knowledge, while correlation was determined on the basis of calibration.

The error of the operation of an analytical chemical system may be characterized by the loss function, and the quality of the operation by the expected value of loss, the risk.

It follows from the above characterization of analytical chemical systems that during their operation two types of decision should actually be taken: analytical information is estimated on the basis of an analytical signal and chemical information is estimated on the basis of analytical information.

*User decision on the basis of chemical information*

The decision of user systems applying analytical chemical systems may be based either on chemical or analytical information. This section deals with the former case.

The state parameter space of the decision model is, again, the collection of chemical information on the chemical properties of the sample. Chemical information is generated by the analytical chemical system, and this forms the sample space. On the basis of the chemical information obtained, user decision, or user information is obtained. This user information, the possible decisions of the user, form the decision space. State parameter space and sample space are identical, both representing chemical information, whereas decision space is different.

The decision function of the decision theoretical model corresponds to the decision of the user. In controlling tasks this is just the 'control law'.

The error of the operation of user systems may be characterized by the loss function, which depends on the actual chemical composition of the sample and on the decision of the user. The quality of the operation of the system may be characterized by the expected loss, i.e. the risk.

*User decision on the basis of analytical information*

The decisions of user systems applying analytical chemical systems may be based not only on chemical but also on analytical information. If it may be assumed that analytical information is a unique function of chemical information, analytical correlation will be an implicit part of the user decision. This is the basic principle of several control systems containing analytical chemical systems, and this justifies

the application of 'incomplete' analytical chemical systems (which contain no analytical chemical correlation).

The state parameter space of the decision model is, again, the set of chemical information relevant to the chemical properties of the sample. The analytical investigation unit generates analytical information, which forms the sample space. On the basis of the analytical information obtained, first chemical information is determined, and then user decision is made, and user information is generated. The set of possible user information spans the decision space. In this case the state parameter, sample and deeision spaces are all different.

The decision function of the decision theoretical model corresponds to the composition of analytical chemical correlation and user decision. In controlling tasks this is the 'control law'.

The error of the operation of user systems can be characterized by the loss function, and the quality of operation by the expected loss, i.e. the risk, defined similarly to the loss and risk of decisions on the basis of chemical information.

*Classification as statistical decision*

The supervised classification methods of pattern recognition can be used to classify objects characterized by property or pattern vectors into previously defined classes. The concepts of decision modelling can be applied to these classification problems as follows:

— the state parameter space contains the true classes of objects;
— the sample space contains the pattern vectors describing the objects;
— the decision space contains the classes estimated by the pattern recognition method.

The parameter space and the decision space are identical, containing classes, whereas the sample space is a different space of pattern vectors.

In the case of pattern recognition, the decision function is the classification rule which assigns an estimated class to all objects represented by pattern vectors. This function is determined by means of 'learning' on a training set. For example, in the case of binary separation, this function is the equation of a hyperplane, and the decision is that if the pattern vector of the object is on one side of the plane, the pattern belongs to one class, and if on the other side, the pattern belongs to the other class.

The loss function, defined over the state parameter and decision spaces, gives the loss which occurs when the true class of the object does not agree with the class given by pattern recognition. In the case of correct classification this loss is zero, otherwise it may assume different values depending on the effect of the various errors, i.e. on the seriousness of the consequences of misclassification. For example, if the structure of a molecule is to be elucidated on the basis of a spectrum, then

a false decision that the molecule does not contain a given structural unit involves a higher loss than the false decision that it contains the given unit. In reality it is the other way round, since in the first case the correct solution is excluded in this step, whereas in the second the extra structural units may be eliminated in later steps.

*Connections between decision theoretical and statistical characterization*

The operation of an analytical chemical system or one of its subsystem can be characterized statistically by the error of information estimation, more precisely the mean square error, standard deviation of the estimation, whereas in decision theory it can be characterized by loss, mean loss or risk.

Loss depends on the error of estimation, thus it is an 'analogous' concept, expressing the 'economic value' of estimation, whereas risk or mean loss depends on the standard deviation of estimation, being analogous to the latter, expressing its economic value.

'Individual' information generated by analytical chemical systems can be characterized statistically by error, in decision theory by the loss depending on it, whereas the 'average' operation of the system can be characterized statistically by standard deviation, and in decision theory by the risk depending on it.

*Connections between decision theoretical and information theoretical characterization*

The connection between decision theoretical and information theoretical treatments of analytical chemical systems depends on the following analogies. In decision theory, the goodness of a single decision is characterized by loss, in information theory the information production of a single measurement can be given by the individual joint information. The quality of a system or a method is characterized by risk in decision theory and by the mean joint information, i.e. transinformation, in information theory.

Thus, analytical chemical systems may be characterized by risk in decision theory or by transinformation in information theory. The 'perfect' operation of analytical chemical systems means on the one hand a deterministic decision, i.e. that its risk is zero, and on the other hand a perfect information communication, i.e. that transinformation is equal to the uncertainty before measurement. The 'completely wrong' operation of an analytical system would mean maximum risk in decision theory and zero transinformation in information theory. Risk and transinformation can thus be used equally for the characterization of analytical chemical systems.

*Decision theoretical description of analytical chemical systems. Summary*

During the operation of analytical chemical systems, decisions on analytical and chemical information are made on the basis of analytical signals, and on chemical information on the basis of analytical information. In the application of analytical chemical systems, user decisions may be based either on the signal or on analytical or chemical information. These decisions, based on the statistical characteristics of analytical chemical systems, may also be described by statistical decision theory, if the losses or their expected value, the risk, can be estimated.

The decision theoretical description of analytical chemical methods is important from the perspective of the uniform treatment of decisions and the introduction of the concepts of loss and risk.

### 6.4.5 Characteristics for credibility

Similarly to the description of analytical chemical systems by means of information theory, the characterization of their credibility is important since it enables one numerically to characterize the truth value and credibility of information produced by analytical chemical systems.

This section deals first with our proposal for the interpretation of credibility, and then with the characterization of analytical chemical systems.

#### Credibility of statements and conclusions

*The concept and measure of credibility*

Credibility of statements or of the conclusions drawn from statements is the truth value of the statement or conclusion. A statement or conclusion is true if it faithfully reflects reality.

However, the majority of statements refer not directly to reality but, indirectly, to the so-called knowledge background which contains relevant and true prior knowledge. This immediately involves the relativity of the truth of statements. It is assumed of background knowledge that it is a true, compound statement consisting of noncontradicting individual statements. This means that background knowledge is a sufficiently faithful description of reality. It must be noted, however, that background knowledge is not permanent, since it is generally expanding, and is sometimes modified.

For the characterization of credibility, the measure, more briefly the concept, of credibility is introduced. The credibility of a statement is its relative truth value. Credibility is not necessarily measurable.

Let us denote the credibility of statement $S$ relative to background $S_b$ by $C(S, S_b)$. The (extent or measure of) credibility is, then, the nonnegative, sigma-

additive set function, taken as measure in the mathematical sense, defined on set $S$ of statements.

Credibility has the following main properties.

If statement $S$ fully corresponds to background $S_b$, i.e. statement $S$ is certain on the basis of background $S_b$, one has

$$C(S, S_b) = 1$$

Similarly, if conclusion $S^*$ is proving, one may write

$$C(S^*, S_b) = 1$$

If statement $S$ contradicts background $S_b$, i.e. statement $S$ is positively false on the basis of background knowledge $S_b$, then

$$C(S, S_b) = 0$$

Consequently, the credibility of statements and conclusions may vary between 0 and 1, i.e.

$$0 <= C(S, S_b) <= 1$$

The credibilities of two statements, $S_i$ and $S_j$ may be compared only in the case of the same background. If the relation

$$C(S_i, S_b) > C(S_j, S_b)$$

holds, one may state that statement $S_i$ is more credible than statement $S_j$ (with respect to the same background, $S_b$).

Statement $S$ and its complementary, i.e. negation, $\bar{S}$ are related by the equation

$$C(S, S_b) + C(\bar{S}, S_b) = 1$$

This is called the complementing principle.

The assumption that the background is a fully true compound statement may be described by the equation

$$C(S_b, S_b) = 1$$

The expansion of background knowledge may affect the truth of a given statement. The inequality

$$C(S, S_b) < C(S, S_b \& S_{b1})$$

(where the & sign denotes conjunction, i.e. logical 'and') means that if statement $S_{b1}$ is proved, statement $S$ is more credible than in the possession of background $S_b$ alone.

A unique connection may be created between statements and the expansion of background knowledge. If $S_i$ and $S_j$ are given statements, the relationship

$$C(S_i \& S_j, S_b) = C(S_i, S_b) \times C(S_j, S_b \& S_i)$$

holds. This is the general multiplication principle.

*Interpretation of credibility*

In a broader sense, credibility is not identical to the concept of probability used in probability calculus, being a broader concept with more than one possibility of interpretation. For the conversion of the credibilities of statements and conclusions into numbers several suggestions have been made (probability logic, multivalued logic, etc.), however, even in mathematical logic, no unique methodology has been developed.

In the sense of probability theory, the concept of credibility can be interpreted as conditional probability:

$$C(S, S_b) = P(S \mid S_b)$$

i.e. credibility is the probability of the truth of statement $S$, i.e. the occurrence of event $S$, under the condition that background knowledge $S_b$ is true.

Since, however, it was assumed that $S_b$ is true, i.e.

$$P(S_b) = 1$$

the above conditional probability may be given as the unconditional probability

$$C(S, S_b) = P(S)$$

Thus, credibility is often denoted simply as probability.

We should note that the interpretation of credibility by means of probabilities is not new at all; it has sound foundations in probability theory known as probability logic.

The basic assumption is that inference, in natural language, is mostly a possibilistic rather than probabilistic concept:

$$C(S, S_b) = \text{Poss}\,[C(S, S_b) = 1]$$

The possibility distribution induced by arbitrary statement $S$ is equivalent to the membership function in fuzzy logic (fuzzy measure).

Unlike probability, the concept of possibility does not assume repeated experiments, being thus nonstatistical in character.

Probability theory and fuzzy logic have been presented as having quite distinct theoretical foundations for reasoning in an uncertain environment, but they have much in common.

*Credibility of analytical chemical inference*

If the premises of a syllogism are characterized by credibility, the credibility of the conclusion is a function of the credibility of the premises. There has been, however, no unique function discovered so far for the calculation of the credibility of conclusions. Of the several possible approaches, the ones based on probability logic, Bayes's theorem, fuzzy logic and reliability theory will be discussed here.

*Probability logic.* In this discussion inference will be slightly simplified, ignoring the universal quantifier together with the additional premise on its domain. Probability logic can be used to calculate the probability of a conclusion from the probabilities of the premises involved. The method is limited to cases when modus ponens could be used if the premises were not 'weakened' by uncertanty. For this case

$$\begin{array}{ll} S^a \rightarrow S^c & P(S^a \rightarrow S^c) = 1 - \varepsilon \\ \underline{S^a} & \underline{P(S^a) = 1 - \delta} \\ S^c & P(S^c \mid S^a) = 1 - (\varepsilon/(1-\delta)) \end{array}$$

where $\delta < \varepsilon$. The credibility of the conclusion is the conditional probability of $S^c$, provided that $S^a$ is true (present).

*Bayes's theorem.* The credibility of the conclusion, $P(S^c \mid S^a)$, is calculated from $P(S^a \mid S^c)$ using Bayes's theorem. Unfortunately, this theorem pays no regard to $P(S^a)$.

*Fuzzy logic.* Here, the credibility of the conclusion is calculated by the compositional modus ponens law. The name of this law indicates that it can be applied to cases where modus ponens would hold if the premises were certain. The syllogism of the inference is then

$$\begin{array}{ll} S^a \rightarrow S^c & P(S^a \rightarrow S) \\ \underline{S^a} & \underline{P(S^a)} \\ S^c & P(S^c) = \min\left[P(S^a \rightarrow S^c), P(S^a)\right] \end{array}$$

The credibility of the conclusion is the minimum credibility of the premises involved. If $S^c$ is inferred from a number of different rules, the credibility of $S^c$ is the maximum of these rules ('maxmin' rule).

*Reliability theory.* This theory can be used to calculate the credibility of the conclusion in inference by a simple multiplication rule

$$C(\text{conclusion}) = \prod_i C(\text{premise}_i) \cdot C(\text{inference})$$

where the credibility of the inference process is defined as the credibility of the conclusion if all the premises had a credibility of 1. Reliability theory is applied mainly to composite inferences.

### *Credibility of chemical information*

The credibility of analytical information depends on the quality (goodness) of the operation of analytical investigation, i.e. analytical signal production and signal processing; assuming once again, of course, the given background knowl-

edge. The credibility of analytical information depends on the credibility of sample identification, on the credibility of the conditions of an analytical investigation and on the credibility of signal processing methods.

The credibility of the known chemical information of calibration samples depends on the independent method used for the determination or checking of this information or else on the credibility of the certification given, e.g. by the company producing the standard.

The credibility of analytical chemical knowledge depends on one hand on the credibilities of the premises (i.e. analytical information obtained during calibration measurements and the known chemical information on the calibration samples, discussed above) and on the other hand on the credibility of the correlations used during analytical chemical acquisition.

The credibility of chemical information obtained by analytical chemical correlation depends on the credibility of analytical information utilized in analytical chemical correlation, the credibility of analytical chemical information and on the credibility of analytical chemical correlation.

If chemical information is determined in parallel by several independent analytical chemical methods, the credibility of chemical information can be calculated by the equation

$$C(S^c) = 1 - \prod_{i=1}^{n} (1 - C(S_i^c))$$

where $C(S_i^c)$ is the credibility of chemical information $S_i^c$ obtained by the $i$th method.

*Logical characterization of analytical chemical systems. Conclusion*

Since analytical chemical correlation is, in the formal logical sense, not proving but only credible, and since both analytical information obtained in analytical investigation and chemical information used for calibration are uncertain, chemical information produced by analytical chemical systems is not certain, only credible. It is therefore recommended to estimate the credibility of both information and correlation.

The idea of credibility is a basis for the real, practicable characterization of the applicability of chemical information produced by analytical chemical systems.

### 6.4.6 Selection of analytical chemical measuring techniques for quality control

The selection of an optimum analytical chemical technique for quality control is a very complex problem. To characterize the chemical measuring techniques, suitable performance characteristics must be defined, and a number of these characteristics must be investigated for the techniques applied; thus such as cost,

time, ruggedness, temporal stability, etc. must be taken into account. One of the most important characteristics is the credibility of the quality determined by the analytical technique selected. Based on this, and on other performance characteristics, an optimum analytical chemical technique for quality control can be selected by a multi-objective method.

Although the literature of quality control and performance characteristics is immense, still no method exists for the selection of analytical chemical techniques for quality control.

*Quality assurance systems*

Quality assurance is a hierarchical, multilevel system, which consists of qualification, quality control and quality assurance levels, as shown in *Fig. 6.5*.

The lowest level of a quality assurance system is the qualification level, which produces quality from the sample. This consists of investigation and correlation (see *Fig. 6.6*). The former generates observed qualifying information from the sample and the latter has two modes of operation, acquisition and inference. Acquisition produces knowledge for qualification from known quality and observed qualifying information. Inference produces the quality of the sample from qualifying knowledge and observed qualifying information.

Qualifying correlation, i.e. the correlation between the observed qualifying information and sample quality, depends not only on the relationship existing between them, but also on the method adopted for their mathematical description.

The middle level of a quality assurance system is the level of quality control. This consists of three operations: sampling, qualification and the application of quality

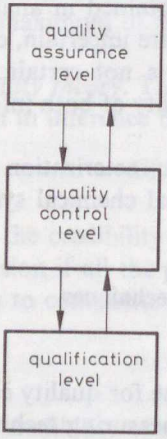



Fig. 6.5. Hierarchical structure of quality assurance system

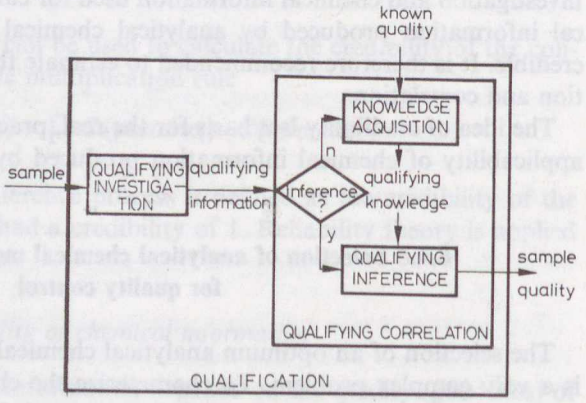



Fig. 6.6. Operations on qualification level

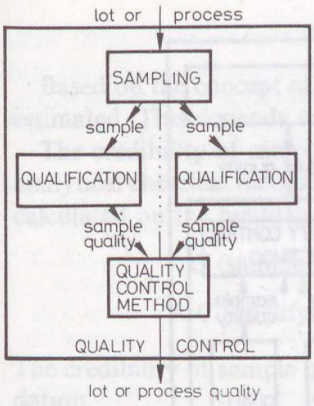



Fig. 6.7. Operations on quality control level

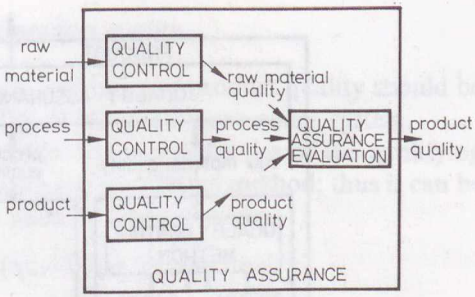



Fig. 6.8. Operations on quality assurance level

control methods, e.g. statistical quality control (see *Fig. 6.7*). Quality control uses the qualities of samples and produces qualitative statements on the quality of the batch or the process under investigation.

The highest level of quality assurance systems is the level of quality assurance. It consists of quality control and quality assurance evaluation, as shown in *Fig. 6.8*. Quality control contains the control of raw material, product and process. Based on the output of quality control, quality assurance evaluation takes decision on the quality of product taking into account the expected income and/or loss.

As a summary, the complete structure of a quality assurance system is shown in *Fig. 6.9*.

*Chemical measuring techniques for qualification*

One of the most important questions of qualification is the selection of qualifying properties produced by qualifying investigations using suitable analytical chemical techniques. Qualifying properties must meet two requirements: first, their measurement should be easy, cheap and suitable for automation and, second, the properties should be able to give a sufficient characterization of quality.

The automation of qualification is often hindered by the fact that the analytical chemical methods of qualification prescribed by standards cannot or can hardly be automated. There is so far no sound statistical and legal basis for the comparison of results obtained by officially prescribed standard methods and by their 'substitute' methods amenable to automation.

The analytical chemical technique selected must give information which has a good correlation with the defined quality, i.e. the quality of a product should be determinable from the observed qualifying information using qualifying correlation.

When selecting a suitable analytical chemical technique, the applied qualifying correlation method must also be taken into account. The performance characteris-

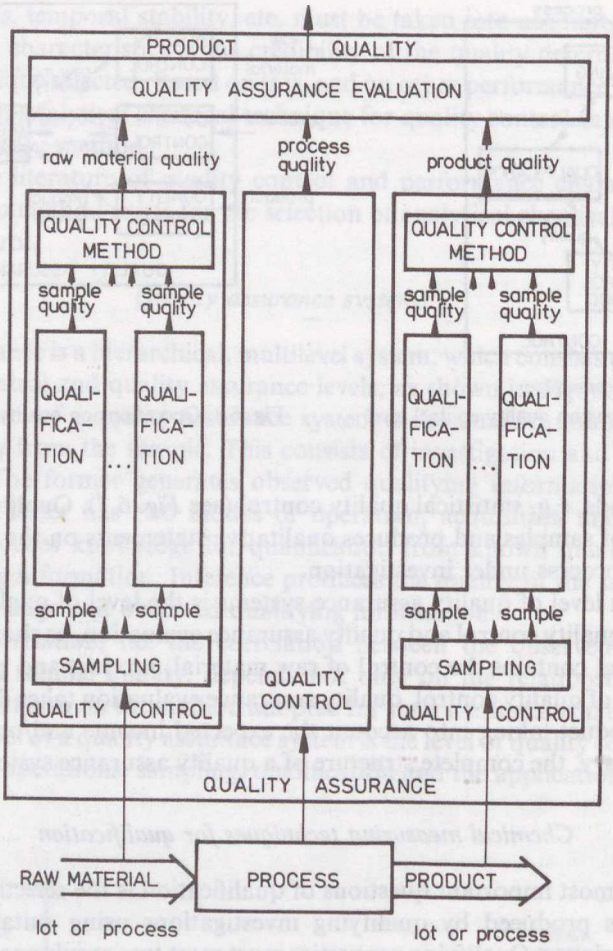



Fig. 6.9. Structure of quality assurance system

tics of the selected analytical chemical technique must meet the requirements of the correlation method applied for qualification. In this respect, supervised classification (pattern recognition) methods are very effective correlation methods for qualification.

Suitable performance characteristics must be determined for the characterization and comparison of analytical chemical techniques.

One of the most important performance characteristics of analytical chemical techniques is credibility, a logical evaluating performance characteristic. Credibility plays a distinguished role, since quality assurance requires the characterization of the credibility of statements on product quality.

*Credibility of product quality*

Based on the concept of credibility, the credibility of product quality should be estimated. This depends on the credibility of the quality assurance system.

The credibility of sample quality depends on the credibilities of the qualifying analytical chemical methods and the qualifying correlation method; thus it can be calculated on the qualification level by means of the equation

$$C\,(\text{sample quality}) = C\,(\text{qualifying information}) \times$$
$$\times\, C\,(\text{qualifying knowledge}) \times C\,(\text{qualifying inference})$$

The credibility of sample quality may be increased by quality assurance and validation.

The credibility of the quality of product and of material can be calculated on the quality control level by equation

$$C_{QC}\,(\text{product or raw material quality}) = C\,(\text{sampling}) \times$$
$$\times\, C\,(\text{sample quality}) \times C\,(\text{quality control method})$$

On the level of quality assurance, the credibility of the product can be calculated approximately by the relationship

$$C_{QA}\,(\text{product quality}) = 1 - [1 - C_{QC}\,(\text{product quality})] \times$$
$$\times\, [1 - C_{QC}\,(\text{raw material}) \times C_{QC}\,(\text{process quality})]$$

*The role of validation*

Validation is a systematic control of all parts of the total quality assurance system. The validation of qualification is necessary to increase the credibility of the quality of sample. The credibility of qualification knowledge can be increased by the validation of the standard materials, etc. applied. Up to now, no report has been found on the validation of qualifying inference, although this become a question of great importance in the application of computer algorithms.

On the level of quality control, the validation of sampling and of the methods of quality control is necessary. The former is a common though very difficult problem, whereas the latter has remained unsolved up to now.

On the level of quality assurance, the validation of quality assurance evaluation is necessary.

*Methods for selecting analytical chemical techniques*

Based on the credibility of quality and other performance characteristics of analytical chemical methods, the selection of the optimum technique for quality control can be solved by one of the multiobjective methods.

Using a multiobjective function for the selection of analytical chemical techniques for quality control, the performance characteristics should be selected and weighted first, then the analytical chemical methods should be evaluated and finally the order of preference of techniques should be determined by experts.

## 6.5 Literature

### 6.5.1 Chemical measuring systems

[1] Danzer, K., Than, E. and Molch, D. (1976): *Analytik-Systematischer Überblick*. Akad. Verlag, Leipzig.
[2] Bartels, H. A. (1977): *Chemische Analyse—Basiswissen*. Springer, Berlin.
[3] Kowalski, B. T. (ed.) (1977): Chemometrics: Theory and Application. *Am. Chem. Soc.*, Symposium Series, No. 52, Washington, D. C.
[4] De Voe, J. F. (ed.) (1977): Validation of the Measurement Process. *Am. Chem. Soc.*, Symposium Series, No. 63, Washington, D. C.
[5] Massart, D. L., Dijkstra, A. and Kaufman, L. (1978): *Evaluation and Optimization of Laboratory Methods and Analytical Procedures*. Elsevier, Amsterdam.
[6] Liteanu, C. and Rica, J. (1980): *Statistical Theory and Methodology of Trace Analysis*. Ellis Horwood Ltd, Chichester.
[7] Kateman, G. and Pijpers, F. W. (1981): *Quality Control in Analytical Chemistry*. Wiley, New York.
[8] Vandeginste, B. G. M. (1981): *Digital Simulation of Laboratory for Structural Analysis*.
[9] Sajevics, A. B. (1981): *Analytical Service as a System*. (In Russian.) Khimiya, Moscow.
[10] Hippe, Z. (ed.) (1981): *Data Processing in Chemistry*. Elsevier, Amsterdam.
[11] Heller, S. R. and Potenzone, R. Jr. (eds) (1983): Computer Applications in Chemistry. *Analytical Chemistry Symposia Series*, Vol. 15. Elsevier, Amsterdam.
[12] Kowalski, B. R. (ed.) (1984): Chemometrics—Mathematics and Statistics in Chemistry. *NATO-ASI-Series-C. 138*. D. Reidel Publ. Co., Dordrecht.
[13] Sharaf, M. A., Illman, D. L. and Kowalski, B. R. (1986): *Chemometrics*. Wiley, New York.
[14] Kowalski, B. R. (1980): Chemometrics. *Anal. Chem.*, **52,** 112R—122R.
[15] Frank, I. E. and Kowalski, B. R. (1982): Chemometrics. *Anal. Chem.*, **54,** 232R–243R.
[16] Delaney, M. F. (1984): Chemometrics. *Anal. Chem.*, **56,** 261R–277R.
[17] Bosman, Th. J. L. and Kateman, G. (1984): A survey of the chemometric publications of 1983. *Chemometric Newsletter*, No. 11, 3–30.
[18] Ramos, L. S., Beebe, K. R. and Carey, W. P. (1986): Chemometrics. *Anal. Chem.*, **58,** 294R–315R.
[19] Malissa, H.: *Z. Anal. Chem.*, **222,** 100–119 (1966), *Z. Anal. Chem.*, **238,** 81–89 (1978), *Z. Anal. Chem.*, **247,** 1–8 (1969), *Z. Anal. Chem.*, **256,** 7–14 (1971), *Z. Anal. Chem.*, **258,** 363–364 (1972), *Z. Anal. Chem.*, **271,** 97–100 (1974).
[20] Kienitz, H. and Kaiser, R. (1968): Datenverarbeitung in der analytischen Chemie. *Z. Anal. Chem.*, **237,** 241–263.
[21] Kaiser, H. (1970): *Anal. Chem.*, **42,**(2), 24A; *Anal. Chem.*, **42**(4), 26A.
[22] Kaiser, H. (1973): Grundlagen zur Beurteilung von Analysenverfahren. In Korte, F. (ed.): *Methodicum Chimicum*. Thieme, Stuttgart.
Gottschalk, G. (1971); *Z. Anal. Chem.*, **256,** 257–270. Kaiser, R. (1972): *Z. Anal. Chem.*, **261,** 1–10. Gottschalk, G.: *Z. Anal. Chem.*, **258,** 1–12 (1972), *Z. Anal. Chem.* **275,** 1–10 (1975), *Z. Anal. Chem.*, **278,** 1–12 (1976), *Z. Anal. Chem.*, **280,** 205–218 (1976), *Z. Anal. Chem.*, **282,** 1–15 (1976), *Z. Anal. Chem.*, **285,** 199–214 (1977).

### 6.5.2 Signal processing methods

[1] Box, G. E. P. and Jenkins, G. M. (1976): *Time Series Analysis*. Holden-Day, San Francisco.
[2] Malinowski, E. R. and Howery, D. G. (1980): *Factor Analysis in Chemistry*. Wiley, New York.
[3] Cadzow, J. A. and Van Landingham, H. F. (1980): *Signals, Systems and Transforms*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
[4] Kelly, P. C. and Harris, W. E. (1971): Estimation of chromatographic peaks with particular consideration of effects base-line noise. *Anal. Chem.*, **43**, 1770–1783.
[5] Kelly, P. C. and Harris, W. E. (1971): Application of Method of Maximum Posterior Probability to Estimation of Gas-Chromatographic Peak Parameters. *Anal. Chem.*, **43**(10), 1184–1195.
[6] Franzen, J. (1979): Theory of $GC^2$ — an unusual approach. *J. of HRC and CC*, **2**, 324–328.
[7] Helmer, J. C. (1980): A theory of measurement with applications to spectrometry. *Anal. Chim. Acta*, **122**, 27–40.
[8] Wijtvliet, J. J. M. (1972): *Ultimate Retention Time Accuracy*. Thesis. Technical University, Eindhoven.

### 6.5.3 Signal interpretation methods

[1] Martens, H. and Naes, T. (1984): Multivariate calibration I–II. *Trends in Analytical Chemistry*, **2**, 204–210, 266–271.
[2] Osten, D. W. and Kowalski, B. R. (1985): Background detection and correction in multicomponent analysis. *Anal. Chem.*, **57**, 908–917.
[3] Blackburn, J. A. (1965): Computer program for multicomponent spectrum analysis using least-squares method. *Anal. Chem.*, **37**, 1000–1003.
[4] Bezjak, A. (1975): Theory of a general method for phase analysis. *Anal. Chem.*, **47**, 790–793.
[5] Brown, S. D. (1986): The Kalman filter in analytical chemistry. *Anal. Chim. Acta*, **181**, 1–26.
[6] Rutan, S. C. (1987): Kalman filtering approaches for solving problems in analytical chemistry. *J. of Chemometrics*, **1**, 7-18.
[7] Sanchez, E. and Kowalski, B. R. (1986): Generalized rank annihilation factor analysis. *Anal. Chem.*, **58**, 496–499.
[8] Carey, W. P., Beebe, K. R., Sanchez, E. et al. (1986): Chemometric analysis of multisensor arrays. *Sensors and Actuators*, **9**, 223–234.

*Quality Control*

[1] Duncan, A. J. (1965): *Quality Control and Industrial Statistics*. R. D. Irwin Inc., Homewood, I11.
[2] Zeleny, M. (ed.) (1976): *Multiple Criteria Decision Making*. Springer, Berlin.

### 6.5.4 Characterization of chemical measuring systems

*Performance characteristics*

[1] Wilson, A. L. (1970): *Talanta*, **17**, 21.
[2] Kuss, H. M. and Wünsch, G. (1984): *Z. Anal. Chem.*, **319**. 492.

*Statistics*

[1] Gottschalk, G. (1962): *Statistik in der quantitativen chemischen Analyse*. F. Enke Verlag, Stuttgart.
[2] Nalimov, V. V. (1963): *The Application of Mathematical Statistics to Chemical Analysis*. Pergamon, London.

[3] Doerffel, K. (1966): *Statistik in der analytischen Chemie*. VEB Deutscher Verlag für Grundstoffindustrie, Leipzig.
[4] Eckschlager, K. (1965): *Fehler bei der chemischen Analyse*. Akad. Verlag, Leipzig.
[5] Eckschlager, K. (1969): Errors and Measurement. In: *Chemical Analysis*. Van Nostrand, London.

*Literature*

[1] Currie, L. A., Filliben, J. J. and De Voe, J. R. (1972): *Anal. Chem.*, **44**, 497R.
[2] Shoenfeld, P. S. and De Voe, J. R. (1976): *Anal. Chem.*, **48**, 403R.

*Information theory*

[1] Shannon, C. E. and Weaver, W. (1949): *The Mathematical Theory of Communications*. University of Illinois Press, Urbana, I11.
[2] Reza, F. M. (1961): *An Introduction to Information Theory*. McGraw-Hill, New York.
[3] Eckschlager, K. and Stepanek, V. (1979): *Information Theory as Applied to Chemical Analysis*. Wiley, New York.
[4] Cleij, P. and Dijkstra, A. (1979): *Z. Anal. Chem.*, **298**, 97.
[5] Liteanu, C. and Rica, I. (1979): Utilization of the amount of information in evaluation of analytical methods. *Anal. Chem.*, **51**(12), 1986.
[6] Frank, I., Veress, G. E. and Pungor, E. (1982): *Hungarian Scientific Instruments*, **54**, 1.
[7] Veress, G. E. and Pungor, E. (1983): Informationstheorie in der analytischen Chemie. *Österr. Chemie-Zeitschrift*, **84**(5), 127.

*Statistical decision theory*

[1] Wald, A. (1950): *Statistical Decision Functions*. Wiley, New York.
[2] Ferguson, T. (1967): *Mathematical Statistics. A Decision Theoretical Approach*. Academic Press, New York.
[3] Frank, I. E., Pungor, E. and Veress, G. E. (1981): *Anal. Chim. Acta*, **133**, 433 and 443.

*Credibility*

[1] Hársing, L. (1981): *The Logic of Scientific Reasoning*. (In Hungarian.) Akadémiai Kiadó, Budapest.
[1a] von Wright, G. H. (1951): *A Treatise on Induction and Probability*. London.
[2] Ruzsa, I. (1971): *Novel Results in Modal Logic*. (In Hungarian.) Akadémiai Kiadó, Budapest.
[3] Reschner, N. (1969): *Many-valued Logic*. MacGraw-Hill, New York.
[4] Iwin, A. A. (1975): *Grundlagen der Logik von Wertungen*. Akademie-Verlag, Berlin.
[5] Deans, D. R. (1971): *J. Chrom. Sci.*, **9**, 729.
[6] Kelly, P. C. (1972): *Anal. Chem.*, **44**(11), 28A.

# APPENDICES

*Appendix A2:* Appendix to Chapter 2
REPROCHE

*Appendix A4:* Appendix to Chapter 4
CAPCOS

*Appendix A5:* Appendix to Chapter 5
TACS

*Appendix A6:* Appendix to Chapter 6
STATCHAR

# APPENDIX A2

## Appendix to Chapter 2

### EURECHA

European Committee For Computers in Chemical Engineering Education

### REPROCHE

Regression Programs for Chemical Engineers

MANUAL

Program version 1
June 1985

Manual authors: S. VAJDA and P. VALKÓ

Laboratory for Chemical Cybernetics
Eötvös Loránd University

Budapest
Múzeum krt. 6—8.
Hungary
H-1055

# CONTENTS


A2—1 INTRODUCTION 495

A2—2 PROBLEMS AND METHODS 495
- A2—2.1 Estimation of parameters in algebraic models 495
- A2—2.2 Goodness of fit and interpretation of estimates 496
- A2—2.3 Solution of differential equations 496
- A2—2.4 Estimation of parameters in differential models by indirect least squares method 497
- A2—2.5 Estimation of parameters in differential models by direct integral least squares method 497
- A2—2.6 Estimation of parameters in dynamic models 498
- A2—2.7 Estimation from tube reactor data 498

A2—3 OVERVIEW OF PROGRAM OPTIONS 499
- A2—3.1 Capacity 500

A2—4 USER-SUPPLIED SUBROUTINES 500
- A2—4.1 Algebraic model 500
- A2—4.2 Differential model 501
- A2—4.3 Dynamic model 502
- A2—4.4 Tube reactor 504

A2—5 DATA INPUT 505
- A2—5.1 Problem name and type 506
- A2—5.2 Problem size and variables 506
- A2—5.2.1 Size and variables of algebraic model 506
- A2—5.2.2 Size and variables of differential model 506
- A2—5.3 Observations 507
- A2—5.3.1 Real data for algebraic model 507
- A2—5.3.2 Simulated data for algebraic model 507
- A2—5.3.3 Real data for differential model 508
- A2—5.3.4 Simulated data for differential model 508
- A2—5.4 Estimation control parameters 509
- A2—5.5 Plots and repeated estimation 510
- A2—5.6 Some notes on data input 510

A2—6 TEST EXAMPLES 512
- A2—6.1 Algebraic model 512
- A2—6.2 Differential model 515

A2—7 INSTRUCTIONS FOR LOADING THE PROGRAM 517

A2—8 REFERENCES 518

# A2—1 INTRODUCTION

REPROCHE is a regression program specifically designed for solving nonlinear parameter estimation problems in chemistry and chemical engineering. The program applies both to algebraic and differential models and requires user-supplied subroutines of very simple structure. Since neither data preparation nor interpretation of results demand a high degree of statistical training, it is suitable for student use. To study possible difficulties in solving nonlinear estimation problems, models can also be fitted to sets of simulated 'observations'.

REPROCHE is based on programs that involve efficient and robust numerical methods and that have been used many years both for teaching and research. These programs were, however, thoroughly revised and made compatible with each other. Particular attention was given to user-supplied subroutines and to the input, since these are the points where mistakes are made. Required programming efforts have been minimized using computer-developed derivatives in solving both differential equations and estimation problems. The input can be batch free-read or interactive, either with supplementary printer or with all results being presented on the terminal. Since the program is self-explanatory in interactive mode, its many options can be easily exploited.

# A2—2 PROBLEMS AND METHODS

## A2—2.1 Estimation of parameters in algebraic models

The model is given by

$$\underline{y} = f(\underline{x}, \underline{p}) \qquad \text{(A2—1)}$$

where $\underline{y} = (y_1, y_2, \ldots, y_{ny})^T$, $\underline{x} = (x_1, x_2, \ldots, x_{nx})^T$ and $\underline{p} = (p_1, p_2, \ldots, p_{np})^T$ represent vectors of dependent variables (response functions), indenpendent variables and parameters, respectively. The observed responses at the $i$th sample point are

$$\tilde{y}_{ij} = f_j(\underline{x}_i, \underline{p}) + \varepsilon_{ij}, \qquad j = 1, 2, \ldots, ny \qquad \text{(A2—2)}$$

where $\underline{x}_i = (x_{i1}, x_{i2}, \ldots, x_{i,n})^T$ and $\varepsilon_{ij}$ are independent variables and error in the $j$th response function, respectively, both at $i$th sample point. Performing $nm$ observations, we estimate the parameters $\underline{p}$ of Eq. (A2—1) from the sample

$$\{(x_{i1}, x_{i2}, \ldots, x_{i,nx}, \tilde{y}_{i1}, \tilde{y}_{i2}, \ldots, \tilde{y}_{i,ny}), \qquad i = 1, 2, \ldots, nm\}$$

by minimizing the objective function

$$Q(\underline{p}) = \sum_{i=1}^{nm} [y_i - f(x_i, p)]^T \, \mathbf{w}_i [y_i - f(x_i, p)] \qquad \text{(A2—3)}$$

where $\mathbf{w}_i$ is an $ny \times ny$ diagonal matrix of weight's $w_{ij}$, $j = 1, 2, \ldots, ny$, in the $i$th

sample point. The function is minimized by a Gauss–Newton–Marquardt procedure [1], involving normalized parameters and the cross-product matrix. Some further details are given in [2].

Control parameters of the estimation procedure are the maximum number of iterations and relative error tolerance on parameters. Starting estimates for parameters are also input data.

The estimation requires one user-supplied subroutine to calculate dependent variables $y_1, \ldots, y_{ny}$ in terms of independent variables $x_1, \ldots, x_{nx}$ and parameters $p_1, \ldots, p_{np}$ according to [1].

## A2—2.2 Goodness of fit and interpretation of estimates

Let $\underline{\hat{p}}$ denote the least squares estimate of the parameters obtained by minimizing (A2—3). The goodness of the fit is characterized by

— minimum of the objective function, $Q(\underline{\hat{p}})$,
— residual errors, $\hat{\varepsilon}_{ij} = \tilde{y}_{ij} - f_j(\underline{x}_i, \underline{p})$,
— residual variance,
— serial correlation coefficient of the residuals [1].

Interpretation of the estimates $\underline{\hat{p}}$ includes

— estimated parameters $\underline{\hat{p}}$,
— variances of estimates,
— confidence intervals at 95% level,
— correlation matrix of the parameters,
— principal component analysis of the normed information matrix for checking near-singularity and its possible sources. Details of this powerful method of analysis are given in [3].

## A2—2.3 Solution of differential equations

Kinetic differential equations of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\underline{c} = f(t, \underline{c}, \underline{p}), \qquad c(0) = \underline{c}^0 \qquad \text{(A2—4)}$$

where $\underline{c} = (c_1, \ldots c_{nc})^T$, $\underline{p} = (p_1, \ldots p_{np})^T$ are solved by a semi-implicit Runge–Kutta method [4]. Initial conditions are $\underline{c}(0) = \underline{c}^0$. The method enables one to solve highly stiff kinetic equations. Solutions are printed at the specified time points $t_1, t_2, \ldots t_{nm}$. Control parameters of the procedure are relative error tolerance on the solution and maximum number of steps taken by the differential equation solver between two neighbouring time points.

The solution requires one user-supplied subroutine to calculate the derivatives $dc_1/dt, \ldots dc_{nc}/dt$ in terms of time $t$, concentrations $c_1, \ldots, c_{nc}$ and parameters $p_1, \ldots, p_{np}$, as given in (A2-4).

## A2—2.4 Estimation of parameters in differential models by indirect least squares method

We consider the model (A2—4) and observe concentrations $c=(c_1, \ldots c_{nc})^T$ at time points $t_1, t_2, \ldots, t_{nm}$. If the observation in the $i$th sample point is

$$\underline{c}_i=(c_{i1}, c_{i2}, \ldots, c_{i,nc})^T, \quad \text{where} \quad \tilde{c}_{ij}=c_j(t_i)+\varepsilon_{ij}$$

then after performing $nm$ observations the data obtained are

$$\{(t, \tilde{c}_{i1}, \ldots \tilde{c}_{i,nc}), \qquad i=1, 2, \ldots nm\}$$

To estimate the parameter $\underline{p}$ in (A2—4) we integrate the differential equation by the procedure A2—2.3 and apply the estimation algorithm A2—2.1 to this numerical solution. The required partial derivatives with respect to the parameters are computed by solving sensitivity equations that correspond to (A2—4) [5]. Estimates are interpreted as in Section A2—2.2.

Initial conditions $\underline{c}^0$ for some species may also be considered as parameters to be estimated. It should be emphasized that using this method you may select some of the concentrations as observed variables and no measurement of all concentrations is required. In this case $ny<nc$ and you must specify $ny$ components in $\underline{c}$ as response functions.

To solve the estimation problem the single user-supplied subroutine required is the same as that described in Section A2—2.3.

## A2—2.5 Estimation of parameters in differential models by direct integral least squares method

In order to avoid the necessity of solving Eqs (A2—4) in each iteration step of the estimation procedure, differential equations are transformed into the equivalent set of integral equations

$$\underline{c}(t_i)=\underline{c}^0+\int_0^{t_i} \underline{f}(\tau, \underline{c}(\tau), \underline{p})\, d\tau \qquad \text{(A2—5)}$$

The basic idea of the direct integral method [6, 7] is to approximate the integrals in (A2—5) by cubic spline functions $\underline{s}_p^f$ interpolating the points $\underline{f}(t_i, \tilde{\underline{c}}_i, \underline{p})$, $i=1, 2, \ldots, nm$. Then the integrals are analytically evaluated and regarding (A2—5) as an algebraic model, the method presented in Section A2—2.1 applies. Spline

interpolation also has a great advantage in computing the required derivatives of $c$ with respect to the parameters [2].

The method requires all species in (A2—4) to be measured, but reduces computer time requirements by up to 100 times. It yields quite good estimates in case of sufficiently dense sample points and moderate measurement errors. In any case, however, application of the method is advantageous for obtaining good starting parameter values for the more involved indirect procedure discussed in Section A2—2.4 [2].

To solve the estimation problem you must supply the same subroutine as in Sections A2—2.3 and A2—2.4.

## A2—2.6 Estimation of parameters in dynamic models

Frequently the observed variables are not concentrations but some related quantities and we have the model

$$\frac{\mathrm{d}}{\mathrm{d}t}\underline{c}=\underline{f}(t,\underline{c},\underline{p}), \qquad \underline{c}(0)=\underline{c}^0$$
$$\underline{y}=\underline{h}(t,\underline{c},\underline{p}) \tag{A2—6}$$

where both $\underline{f}$ and $\underline{h}$ are known functions with unknown parameters. The indirect method given in Section A2—2.4 applies also to this more general case. The only difference is that you must supply two further subroutines for calculating the response functions $y_1, \ldots y_{ny}$, and the partial derivatives $\partial y_i/\partial p_j$, $i=1, 2, \ldots ny$, $j=1, 2, \ldots, np$ in terms of $t$, $c_1, \ldots c_{nc}$ and $p_1, \ldots p_{np}$ according to Eqs (A2—6). Further details will be given in Section A2—4.

## A2—2.7 Estimation from tube reactor data

Assume that Eq. (A2—4) describe the steady-state concentration profiles of a distributed parameter process, there $t$ is considered as a length coordinate (or catalyst weight). In experiments the length is usually fixed, i.e., $t=L$, while reactor load, input concentrations, and additional parameters (e.g., temperature) may be varied from one sample point to another. We can solve this estimation problem as an algebraic one, considering experimental conditions as independent variables $x_{i1}, x_{i2}, \ldots, x_{in}$ in the $i$th sample point and computing the response through the integration of Eqs (A2—4) from $t=0$ to $t=L$ with the corresponding constants. This latter operation is supported by a specific subroutine TUBE. The approach requires two user-supplied subroutines. The first generates the initial conditions (and reactor length, if variable) and calls the differential equation solver TUBE. The second subroutine computes the derivatives as described in Section A2—2.3.

## A2—3 OVERVIEW OF PROGRAM OPTIONS

Before detailed description of data input a summary of program options is given in *Table A2—3.1.*

Table A2—3.1. Overview of program options

| Select | Options | Remark |
|---|---|---|
| Model? | algebraic (1) or differential (2) | |
| Responses? | concentrations (1) or else (2) | only for differential model |
| Data? | real (1) or simulated (2) | |
| Data from file? | yes or no | only for real data |
| Are data points equidistant? | yes or no | only for simulation (a) |
| Additive noise? | no (0), absolute (1) or relative (2) | only for simulation |
| Do you wish parameter estimation? | yes or no | if no, you may change data ($b$, $c$) |
| Weights? | corresponding to the assumption of: | $i$-sample, $j$-response |
| | — identical variance (0) | $w_{ij}=1$ |
| | — identical relative variances (1) | $w_{ij}=1/y_{ij}^2$ |
| | — the same user specified variances in each sample point (2) | $w_{ij}=1/\sigma_j^2$ ($ny$ numbers) |
| | — arbitrary user-specified variances (3) | $w_{ij}$ arbitrary ($ny \times nm$ numbers) |
| Method? | direct integral (1) or indirect (2) | only for differential model |
| Plots? | no (0), residuals (1), fir (2) or off line (3) | off-line plot means creating a file |
| Repeat estimation? | yes or no | if yes, return to data specification |

*Notes to Table A2—3.1*

(a) In a repeated estimation you can use the previous data set by entering ⟨CR⟩. Notice that on some computers ⟨CR⟩ (carriage return) should be preceeded by at least one blank character.

(b) By entering ⟨CR⟩ when respecifying data and control parameters previous values are left unchanged.

(c) It is important to emphasize that you must select the parameters to be estimated while the others are kept at their initial values. In a differential model you may also select initial concentrations as unknown parameters. For an explanation of how to use this option, consider a model with the parameters $p_1$, $p_2$ and $p_3$ i.e., $npr=3$, where $npr$ denotes the number of the parameters, (known + un-

known). If you select indices 3 and 5 for the parameters to be estimated, then the actual number of unknown parameters in $np=2$ and the parameters are $p_3$ and $c_2^0$, since a vector with entries $(p_1, p_2, p_3, c_1^0, c_2^0, \ldots c_{nc}^0)$ is considered as an extended vector of potential unknown parameters and $c_2^0$ is the fifth entry in it. Parameters $p_1$, $p_2$, and initial conditions $c_1^0$, $c_3^0$, ..., $c_{nc}^0$ will be kept fixed during the estimation.

### A2—3.1 Capacity

The program is so dimensioned that it can handle

— 20 parameters (known + unknown) (*npr*),
— 20 independent variables (*nx*),
— 20 dependent variables (response functions) (*ny*),
— 20 concentration variables (*nc*),
— 500 sample points (*nm*),

Larger problems require re-dimensioning of the variables. Sizes of some auxiliary arrays in the program depend on several integers *npr*, *nx*, *ny*, *nc* and *nm*. If one of these arrays is not sufficiently large for the actual problem, prompts indicate the required re-dimensioning.

## A2—4 USER-SUPPLIED SUBROUTINES

Creation of user-supplied subroutines is shown by simple examples.

### A2—4.1 Algebraic model

Consider the single-response model

$$y = p_1 + \frac{x_1}{p_2 x_2 + p_3 x_3} \qquad \text{(A2—7)}$$

with independent variables $x_1$, $x_2$, and $x_3$, and parameters $p_1$, $p_2$ and $p_3$ [8]. This model is specified by the following subroutine.

```
SUBROUTINE FUNA (NP, NY, NX, P, X, Y)
DIMENSION X(NX), Y(NY), P(1)
Y(1)=P(1)+X(1)/(P(2)*X(2)+P(3)*X(3))
RETURN
END
```

The parameters of the subroutine are as follows:

*Input*:

NP — number of parameters (*np*),
NY — number of dependent variables (*ny*),
NX — number of independent variables (*nx*),
P — vector of parameters (*p*),
X — NX-vector of independent variables (*x*).

*Output*:

Y — NY-vector of dependent variables (*y*).

*Note*: Please do write P(1) in the DIMENSION statement instead of P(NP), since NP denotes the number of parameters to be actually estimated and this may be less than the total number *npr* of parameters (known + unknown).

## A2—4.2 Differential model

Consider the reactions [9]

$$\begin{aligned} &1.\ \ A+B\rightarrow C+D \\ &2.\ \ A+D\rightarrow 2X \end{aligned} \qquad \text{(A2—8)}$$

with reaction rates

$$r_1 = k\,[A][B], \quad r_2;\quad k_2\,[A][D] \qquad \text{(A2—9)}$$

Using identifications $P(1)=k_1$, $P(2)=k_2$,

$$C(1)=[A],\quad C(2)=[B],\quad C(3)=[C],\quad C(4)=[D]$$

and C(5)=[X], kinetic equations of the form (A2—4) are specified for this simple system by the following subroutine:

```
SUBROUTINE FUND (NP, NC, T, P, C, DC, X)
DIMENSION C(NC), DC(NC), P(1), X(1)
R1=P(1)*C(1)*C(2)
R2=P(2)*C(1)*C(4)
DC(1)=-R1-R2
DC(2)=-R1
DC(3)=R1
DC(4)=R1-R2
DC(5)=2*R2
RETURN
END
```

The parameters of the subroutine are as follows:

*Input*:

NP — number of parameters (*np*),
NC — number of concentration variables (*nc*),
(i.e., number of differential equations),

T — time ($t$),
P — vector of parameters ($P$),
C — NC-vector of concentration variables ($c$),
X — dummy vector used in tube reactor estimation only.
*Output*:
DC — NC-vector of time derivatives ($dc/dt$).
*Notes*:
(1) Please do write P(1) in the DIMENSION statement instead of P(NP).
(2) This subroutine is sufficient for estimating parameters in differential models if the observed response functions are selected concentrations. The same subroutine also applies also in the case of estimating some of the initial concentrations as additional parameters.

## A2—4.3 Dynamic model

Assume that the observed variables for the kinetic model discussed in Section A2—4.2 are

$$y_1 = p_3[\mathrm{A}] + p_4[\mathrm{B}]$$
$$y_2 = [\mathrm{D}] \qquad \text{(A2—10)}$$

i.e., $y_1$ is a linear combination (possibly with unknown contribution coefficients) of the concentrations [A] and [B]. The second response is [D]. In addition to the subroutine given in Section A2—4.2 you need an additional one to specify the model of the form (A2—6).

```
SUBROUTINE YGEN (NC, NY, NP, T, P, C, Y)
DIMENSION C(NC), Y(NY), P(1)
Y(1)=P(3)*C(1)+P(4)*C(2)
Y(2)=C(4)
RETURN
END
```

Parameters of the subroutine are as follows:
*Input*:
NC — number of concentration variables ($nc$),
NY — number of response functions ($ny$),
NP — number of parameters ($np$),
T — time ($t$),
P — vector of parameters ($p$),
C — NC-vector of concentration variables ($c$).
*Output*:
Y — NY-vector of response functions ($y$)
*Note*: (1) Please do write P(1) in DIMENSION statement instead of P(NP).

A further subroutine is needed to compute the derivatives $\partial y_i/\partial p_j$, $i=1, 2, \ldots, ny$; $j=1, 2, \ldots, np$, where $np$ is the actual number of parameters to be estimated. The NY × NP array of these derivatives will be denoted by DYDP. According to (A2—6), by the rule of differentiating composite functions

$$\frac{\partial y}{\partial p_j} = \frac{\partial h}{\partial c}\frac{\partial c}{\partial p_j} + \frac{\partial h}{\partial p_j} \qquad \text{(A2—11)}$$

Evaluating (A2—11) you can use the derivatives $\partial c_j/\partial p_j$ stored in the NX × NP array DCDP.

To show the most complex application of this option assume that $p_3$ in (A2—10) is a parameter to be estimated in addition to rate constants $p_1$ and $p_2$. Then exploiting (A2—11) the matrix DYDP is evaluated by the following subroutine:

```
      SUBROUTINE YJAC (NC, NY, NP, T, P, C, DCDP, DYSP)
      DIMENSION C(NC), P(1), DCDP(NC, NP), DCDY(NY, NP)
      DO 10 I=1, NY
      DO 10 J=1, NP
   10 DYDP(I, J)=0.
      DO 20 J=1, 2
      DYDP(1, J)=P(3)*DCDP(1, J)+P(4)*DCDP(2, J)
   20 DYDP(2, J)=DCDP(4, J)
      DYDP(1, 3)=C(1)
      RETURN
      END
```

Parameters of this subroutine are as follows:

*Input*:

NC — number of concentration variables ($nc$),
NY — number of response functions ($ny$),
NP — number of parameters to be estimated ($np$),
T — time ($t$),
P — vector of parameters ($p$),
C — NC-vector of concentration variables ($c$),
DCDP — NC × NP array of derivatives $\partial c_i/\partial p_j$, $i=1, 2, \ldots, nc$; $j=1, 2, \ldots, np$.

*Output*:

DYDP — NY × NP array of derivatives $\partial y_i/\partial p_j$, $i=1, 2, \ldots, ny$; $j=1, 2, \ldots, np$.

*Notes*:

(1) Please do write $P(1)$ in the DIMENSION statement instead of $P(NP)$.

(2) In contrast to the previous subroutines, this one strictly corresponds to the case in which the parameters to be estimated are $p_1$, $p_2$, and $p_3$. If you fix some of them (e.g., $p_2$), than YJAC must be appropriately changed to evaluate the $2 \times 2$ matrix of the derivatives with respect to $p_1$ and $p_3$ instead of computing a $2 \times 3$ matrix as in the present case.

## A2—4.4 Tube reactor

We estimate the Arrhenius parameters $k_1 = P(1)$, $k_2 = P(2)$, $E_1 = P(3)$, and $E_2 = P(4)$ of the consecutive reaction $A \rightarrow B \rightarrow C$ from steady-state experiments performed in a catalytic reactor at different weights of the catalyst, input concentrations of A, and temperatures. All concentrations are observed. As in Section A2—4.2, the differential equations are specified by the subroutine

```
      SUBROUTINE FUND (NP, NC, T, P, C, DC, X)
      DIMENSION C(NC), DC(NC), P(1), X(1)
      R=8.314
      R1=R(1) *EXP (-P(3)*(1./X(2)-1./300.)/R)*C(1)
      R2=P(2) *EXP (-P(4)*(1./X(2)-1./300.)/R)*C(2)
      DC(1)=-R1
      DC(2)=R1-R2
      SC(3)=R2
      RETURN
      END
```

where X(2) denotes the actual temperature. Integration is performed with respect to the catalyst weight. The estimation problem, however, is solved as an algebraic one with the independent variables

X(1) — catalyst weight
X(2) — temperature
X(3) — input concentration of *A*

and requires a user-supplied subroutine as follows.

```
      SUBROUTINE FUNA (NP, NY, NX, P, X, Y)
      DIMENSION X(NX), Y(NY), P(1)
      COMMON/AUX/AUX (12000)
      Z1=0.
      Z2=X(1)
      Y(1)=X(3)
      Y(2)=0.
      Y(3)=0.
      CALL TUBE (Z1, Z2, NP, NY, P, Y, X)
      WRITE (1, 101) Z1, Z2, (Y(I), I=1,3)
  101 FORMAT (5E12.4)
      RETURN
      END
```

The aim of this routine is the solution of the differential equations from Z1 to Z2. At each call the vector *X* contains the actual independent variables (i.e. the experimental conditions). The parameters of the TUBE subroutine are as follows:

*Input*:
Z1 — lower limit of the interval of integration,
Z2 — upper limit of the interval of integration,
NP — number of parameters,
NY — number of concentration variables,
(i.e., number of differential equations),
*P* — vector of parameters,
*Y* — initial concentrations (at Z1),
*X* — vector of independent variables,
(only transferred to the subroutine FUND).
*Output*:
Y — concentrations at Z2.

*Note*: The options of Section A2—4.3 cannot be used here and the differential equations should be written for observed variables. If some components of *Y* are unobserved, zero weights can be used to eliminate them from the objective function (see No. 6 in Section A2—5.4).

## A2—5 DATA INPUT

Thought the input data required for parameter estimation are almost independent of the model (algebraic or differential), data specifying the system and observations are slightly different. According to the specific problem studied, you must select different sections of input data as shown in *Fig. A2—1.*

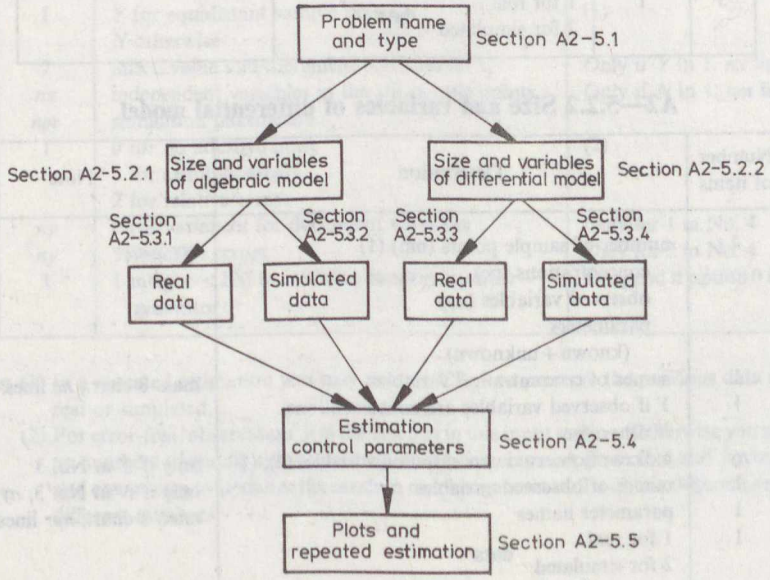



Fig. A2—1. Structure of data input

## A2–5.1 Problem name and type

| No. | Number of items | Description | Note |
|---|---|---|---|
| 1 | 1 | problem name | max. 12 char. |
| 2 | 1 | 1 for algebraic model<br>2 for differential model | |

## A2–5.2 Problem size and variables

### A2–5.2.1 Size and variables of algebraic model

| No. | Number of items | Description | Note |
|---|---|---|---|
| 1 | 4 | number of sample points (*nm*)<br>independent variables (*nx*)<br>dependent variables (*ny*)<br>parameters (known + unknown) | denoted by *npr* ≥ *np* |
| 2 | 1 | names of independent variables | max. 8 char., *nx* lines |
| 3 | 1 | names of dependent variables | max. 8 char., *ny* lines |
| 4 | 1 | parameter names | max. 8 char., *npr* lines |
| 5 | 1 | 1 for real data<br>2 for simulated data | |

### A2—5.2.2 Size and variables of differential model

| No. | Number of items | Description | Note |
|---|---|---|---|
| 1 | 4 | number of sample points (nm) (1)<br>concentrations (nc)<br>observed variables (ny)<br>parameters (known + unknown) | *npr* ≥ *np* |
| 2 | 1 | names of concentration variables | max. 8 char., *nc* lines |
| 3 | 1 | *Y* if observed variables are concentrations<br>*N* otherwise | |
| 4 | *ny* | indices of observed concentration variables (2), (3) | only if *Y* in No. 3 |
| 4′ | 1 | names of observed variables | only if *N* in No. 3, *ny* lines |
| 5 | 1 | parameter names | max. 8 char., *npr* lines |
| 6 | 1 | 1 for real data<br>2 for simulated data | |

*Notes*: (1) Initial time point ($t=0$) should also be considered as a sample point.
(2) 4 or 4′, depending on whether or not observed variables are concentrations.
(3) This item is dropped if $ny=nc$, i.e., all concentration variables are observed.

# A2—5.3 Observations

## A2—5.3.1 Real data for algebraic model

| No. | Number of items | Description | Note |
|---|---|---|---|
| 1 | 1 | <CR> or for included data otherwise file number (integer >0) | (1) |
| 2 | $nx+ny$ | independent and dependent variables in the $i$th sample | $nm$ lines (2) |
| 3 | 1 | $Y$ if parameter estimation is required<br>$N$ if you want to change data | N-return to No. 1. |

*Notes*: (1) Any number >0 defines the logical number of a file containing observations. This option is particulary useful in interactive mode.

(2) No. 2 is dropped if observations are read from a separate file specified in No. 1.

## A2—5.3.2 Simulated data for algebraic model

| No. | Number of items | Description | Note |
|---|---|---|---|
| 1 | 1 | $Y$ for equidistant sample points<br>$N$ otherwise | (1) |
| 2 | 2 | initial value and increment for variable $x_j$ | Only if $Y$ in 1. $nx$ lines |
| 2′ | $nx$ | independent variables in the $i$th sample points | Only if $N$ in 1. $nm$ lines |
| 3 | $npr$ | simulation parameters | |
| 4 | 1 | 0 for no additive noise<br>1 for absolute errors<br>2 for relative errors | (2) |
| 5 | $ny$ | error variances for dependent variables | Only for 1 in No. 4 |
| 5′ | $ny$ | % relative errors | Only for 2 in No. 4 |
| 6 | 1 | 1 integer <100 to initialize random number generator | Dropped if option 0 in No. 4 |

*Notes*: (1) In a repeated estimation you may press <CR> to preserve the previous data set, either real or simulated.

(2) For error-free 'observation' it is the last line in this input section. Otherwise you must enter $ny$ numbers to specify error variances or relative errors as given in 5 and 5′, respectively, and one integer to initialize the random number generator. To obtain different samples use different numbers.

## A2—5.3.3 Real data for differential model

| No. | Number of items | Description | Note |
|---|---|---|---|
| 1 | 1 | <CR> or 0 for included data otherwise file number (>0) | (1) |
| 2 | $ny+1$ | time and observed variables in the $i$th sample point | $nm$ lines (2) |
| 3 | 1 | *Y* if parameter estimation is required<br>*N* if you want to change data | N-return to No. 1 |

*Notes*: (1) Any number >0 defines the logical number of a file containing observations. This option is particulary useful in interactive mode.
(2) No. 2 is dropped if observations are read from a seperate file specified in No. 1.

## A2—5.3.4 Simulated data for differential model

| No. | Number of items | Description | Note |
|---|---|---|---|
| 1 | 1 | *Y* for equidistant sample points<br>*N* otherwise | (1) |
| 2 | 2 | initial time and step size | only if *Y* in No. 1 |
| 2′ | $nm$ | sample times (2) | only if *N* in No. 2 |
| 3 | $npr$ | simulation parameters | |
| 4 | $nc$ | initial concentrations for simulation | |
| 5 | 2 | number of steps and error tolerance for solving differential equations | suggested: 100, .001 |
| 6 | 1 | *Y* for printing concentrations<br>*N* no printing | |
| 7 | 1 | 0 for no additive noise<br>1 for absolute errors<br>2 for relative errors | (3) |
| 8 | $ny$ | error variances for dependent variables | only for 1 in No. 7 |
| 8′ | $ny$ | % relative errors | only for 2 in No. 7 |
| 9 | 1 | 1 integer <100 to initialize random number generator | |

*Notes*: (1) See (1) in Section A2—5.3.2.
(2) 2 or 2′ depending on whether or not sample points are equidistant. In 2′ enter at most 20 numbers in one line. If these 20 numbers occupy several lines, use * as the last character to show that a continuation line follows. If $nm>20$, use several lines.
(3) See (2) in Section A2—5.3.2.

## A2—5.4 Estimation control parameters

| No. | Number of items | Description | Note |
|---|---|---|---|
| 1 | 1 | number of parameters to be estimated (*np*) | (1) |
| 2 | *np* | indices of parameters to be actually estimated | (2) |
| 3 | *npr* | starting parameter estimates for known + unknown parameters | (3) |
| 4 | *nc* | initial concentrations | only for differential model |
| 5 | 2 | maximum number of iterations relative step threshold | |
| 6 | 1 | 0 for identical variances<br>1 for identical relative variances<br>2 for the same user-specified variances in each sample point<br>3 for arbitrary variances | weights corresponding to the selected assumption on (4) error variance |
| 7 | *ny* | standard errors of observed variables (5) | for option 2 in No. 6 |
| 7′ | *ny* | standard errors of observed variables in *i*th sample point (6) | for option 3 in No. 6 *nm* lines |
| 8 | 1 | 1 for direct integral method<br>2 for indirect method (7) | only for differential model |

*Notes*: (1) Recall that *npr* and *np* denote the number of parameters (known + unknown) and the number of parameters actually estimated. Since in differential models initial concentrations may also be estimated, in such cases we may have $np > npr$.

(2) Note that here we consider the vector $p_1, p_2, \ldots, p_{npr}, c_1^0, c_2^0, \ldots, c_{nc}^0$ and select its entries which are to be estimated.

(3) Parameters and initial concentrations not selected in No.2 are kept fixed at their values specified in No.3 and No.4, respectively.

(4) Let $i$ denote the sample point and $j$ the response function, i.e., the $j$th observed variable in $i$th sample point is $y_{ij}$. The different weighting options are as follows:

0: $w_{ij} = 1$
1: $w_{ij} = 1/Y_{ij}^2$ ($w_{ij} = 1.E\ 10$ if $y_{ij} \leq 1.E{-}5$)
2: $w_{ij} = 1/\sigma_j^2$, where $\sigma_j^2$ is the variance of the variable $y_j$ and must be specified in No.7 of the input.
3: $w_{ij} = 1/\sigma_{ij}^2$, where $\sigma_{ij}^2$ is the variance of the variable $y_j$ in the sample point $t_j$ and must be specified in No.7′ of the input.
Note that no further information for weighting is required when using options 0 or 1.

(5) If $\sigma_j < 0$, then $w_{ij} = 0$ for $i = 1, 2, \ldots, nm$, thus we omit the response $y_j$ from the sum of squares objective function.

(6) If $\sigma_{ij} < 0$, then $w_{ij} = 0$, thus we omit the observation $y_{ij}$ from the sum of squares objective function. It is advantageous if the $j$th response has not been measured in the $i$th sample point.

(7) After No. 8 the program performs parameter estimation and prints results. Further input data are needed only for creating plots and repeating the estimation. Since these options are

recommended to use in interactive mode, in batch mode input data are usually terminated by the following 2 additional cards:

0
N

which show that neither plots nor repeated estimation are required. Otherwise go to the next input section.

We describe these sections separately. It should be emphasized that the following information is necessary only if the program is to be run in batch mode, since it is self-explanatory in the interactive mode and its many options can easily be exploited.

## A2—5.5 Plots and repeated estimation

The options of this section enable you to obtain plots on the line-printer or display for a quick inspection of the residuals and the goodness of the fit. Creating such plots if self-explanatory in the interactive mode. You can also create a file for off-line plotting of some selected response functions computed as the least squares estimates of the parameters.

| No. | Number of items | Description | Note |
|---|---|---|---|
| 1 | 1 | 0 for no plots<br>1 for plotting residuals<br>2 for plotting the fit<br>3 for creating a file for off-line plotting (1) | select plot;<br>options 2 and 3<br>only for models with 1 independent variable |
| 2 | 2 | indices of independent and dependent variables | if $nx > 1$ |
| 2′ | 1 | index of dependent variable | if $nx = 1$ |
| 3 | 1 | *Y* if you change scales<br>*N* for automatic scaling minimum and maximum for independent and dependent variables | only if *Y* in No. 9 |
| 4 | 4 | | |
| 2″ | 2 | logical number of a file and number of response functions to be written on the file (*nf*) | only for option 3 in No. 1 |
| 3″ | *nf* | indices of response functions to be written on the file | |
| 4″ | 1 | number of desired sample points | only for option 3 in No. 1 |
| 5 | 1 | *Y* if you wish repeated estimation<br>*N* stop | if *Y*, return to No. 5 in *Table* A2—5.2.1 or 6 in *Table* A2—5.2.2, depending on model type |

*Notes*: (1) Assume that you select *nf* response functions to be written on the file in *n* sample points. The values of *nf*, the desired responses, and *n* are defined in No. 2″, 3″ and 4″. The created file has the following structure:

*n* equidistant values of the independent variable;
*n* values of the selected response 1;
⋮
*n* values of the selected response *nf*.

Format on this file is 7E11.5.

## A2—5.6 Some notes on data input

(a) The following points should be noted:
— Numerical data must be separated by a comma or blanks.
— All data are read as real. Decimal points are only necessary when the data include a fractional part.
— Up to 20 data items can be packed on one 80 character line. Continuation lines are indicated by a star (*)
— Carriage return (<CR>), returning zero characters on the line in recognized and used for input control or returning default values as indicated by the prompts.
— Break is $$$$, which at any point of the input enables one to return to an earlier point in the program and to re-enter erroneous data. Exceptions are questions of Yes/No type, where break is the single character $. Then $ has the same effect as $$$$ following other prompts.

(b) Using the program interactively (i.e. in MODE = 1 or MODE = 2, see Section A2—7, the prompts listed in Section A2—5 are preceeded by the question

SELECT INPUT MODE:

INTERACTIVE (0), BATCH (1), OR COMBINED (2)

If you enter 1 or 2, a further prompt appears:

ENTER NAME OF THE INPUT FILE

In this case a direct file will be opened on unit 8. Option 1 implies a virtually batch operation, reading all data from the input file as in MODE = 0, option 2 reads only the observations from the input file, while 0 does not require the file at all.

(c) The supplied version of the program operates in MODE = 2 and opens a direct file 'RESULTS' on unit 7 instead of printing results directly. For further information see Section A2—7.

# A2—6 TEST EXAMPLES

## A2—6.1 Algebraic model

We fit the model (A2—7) discussed in Section A2—4.1 to the data given by Bard [8] and consisting of the values of $x_1$, $x_2$, $x_3$ and $y$ in 15 sample points. Notice that following the original paper [8], variables $x_1$, $x_2$, and $x_3$ will be denoted by X, V and W, respectively. We read these data from a file having logical number 8 and we use identical weights. *Table A2—6.1.1* shows data input in interactive mode. Data input in batch mode is presented in *Table A2—6.1.2*, omitting lines for the residual plot.

Table A2—6.1.2. Data input in batch mode for algebraic model

```
BARD TEST
1
15  3  1  3
X
V
W
Y
p1
p2
p3
1
8
Y
3
1  2  3
1  1  1
10  0.001
0
0
N
```

This test example has been solved by many well-known parameter estimation methods. As shown in [8] and [10], minimization of the objective function required 5—30 iteration steps, depending on the method. We obtained the same results in 5 iterations, thus the algorithm used in REPROCHE is competitive with the best available ones. Notice that the residual plot clearly shows one outlying observation.

Table A2—6.1.1. Data input in interactive mode

```
**********************************************************************************
      REPROCHE — REGRESSION PROGRAMS FOR CHEMICAL ENGINEERING
**********************************************************************************

ENTER NAME OF THE INPUT FILE
BARCOM
ENTER PROBLEM NAME(MAXIMUM 12 CHARACTERS)
BARD TEST
SELECT MODEL: ALGEBRAIC (1) OR DIFFERENTIAL (2)?
1
ENTER 4 INTEGERS TO SPECIFY NUMBER OF SAMPLE POINTS
                             NUMBER OF INDEPENDENT VARIABLES
                             NUMBER OF DEPENDENT VARIABLES
                             NUMBER OF PARAMETERS (KNOW+UNKNOWN)
15  3  1  3

NAMES OF INDEPENDENT VARIABLES
ENTER NAME FOR VARIABLE  1
     X
ENTER NAME FOR VARIABLE  2
     V
ENTER NAME FOR VARIABLE  3
     W

NAMES OF DEPENDENT VARIABLES
ENTER NAME FOR VARIABLE  1
     Y

PARAMETER NAMES
ENTER NAME PARAMETER  1

     P1
ENTER NAME FOR PARAMETER  2

     P2
ENTER NAME FOR PARAMETER  3

     P3

REAL(1) OR SIMULATED(2) DATA?
1
IF DATA INPUT FROM FILE ENTER ITS LOGICAL NUMBER
8
```

Table A2—6.1.1 (continued)

TABLE OF OBSERVATIONS

| SAMPLE POINT | VARIABLES | | | |
|---|---|---|---|---|
| | X | V | W | Y |
| 1 | 0.1000E+01 | 0.1500E+02 | 0.1000E+01 | 0.1400E+00 |
| 2 | 0.2000E+01 | 0.1400E+02 | 0.2000E+01 | 0.1800E+00 |
| 3 | 0.3000E+01 | 0.1300E+02 | 0.3000E+01 | 0.2200E+00 |
| 4 | 0.4000E+01 | 0.1200E+02 | 0.4000E+01 | 0.2500E+00 |
| 5 | 0.5000E+01 | 0.1100E+02 | 0.5000E+01 | 0.2900E+00 |
| 6 | 0.6000E+01 | 0.1000E+02 | 0.6000E+01 | 0.3200E+00 |
| 7 | 0.7000E+01 | 0.9000E+01 | 0.7000E+01 | 0.3500E+00 |
| 8 | 0.8000E+01 | 0.8000E+01 | 0.8000E+01 | 0.3900E+00 |
| 9 | 0.9000E+01 | 0.7000E+01 | 0.7000E+01 | 0.3700E+00 |
| 10 | 0.1000E+02 | 0.6000E+01 | 0.6000E+01 | 0.5800E+00 |
| 11 | 0.1100E+02 | 0.5000E+01 | 0.5000E+01 | 0.7300E+00 |
| 12 | 0.1200E+02 | 0.4000E+01 | 0.4000E+01 | 0.9600E+00 |
| 13 | 0.1300E+02 | 0.3000E+01 | 0.3000E+01 | 0.1340E+01 |
| 14 | 0.1400E+02 | 0.2000E+01 | 0.2000E+01 | 0.2100E+01 |
| 15 | 0.1500E+02 | 0.1000E+01 | 0.1000E+01 | 0.4390E+01 |

DO YOU WISH PARAMETER ESTIMATION(Y/N)?

Y

ENTER NUMBER PARAMETERS TO BE ESTIMATED

3

ENTER 3 INTEGERS TO SPECIFY INDICES OF UNKNOWN PARAMETERS
1 2 3

ENTER 3 NUMBERS TO SPECIFY PARAMETER ESTIMATES
1 1 1

ENTER MAXIMUM OF NUMBER OF ITERATIONS AND RELATIVE STEP THRESHOLD

10 .001

SELECT WEIGHTS CORRESPONDING TO THE ASSUMPTION OF
IDENTICAL VARIANCES(0)
IDENTICAL RELATIVE VARIANCES(1)
THE SAME USER-SPECIFIED VARIANCES IN EACH SAMPLE POINT(2)
ARBITRARY USER-SPECIFIED VARIANCES(3)

0

The output is shown in Chapter 2 *(Fig. 2.2)*.

## A2—6.2 Differential model

We solve the test example given by Eakman [9] and discussed in Section A2—4.2. The set of 'bad' data is taken from [9] assuming that all concentrations are measured. Following the original paper [9], identical weights are used. We first apply the direct integral method and then improve estimates by applying the indirect method. Reading data from a separate file having logical number 8, input

Table A2—6.2.1 Data input batch mode for differential model: direct method followed by indirect one

```
EAKMAN BAD
1
11  5  5  2
A
B
C
D
X
Y
K1
K2
1
8
Y
2
1  2
.5  .5
1.5  1  0  0  0
10  0.00001
0
100  0001
1
0
Y
```

```
2                    Repeated estimation
<CR>                 by indirect method
Y
<CR>
<CR>
<CR>
<CR>
<CR>
2
O
N
```

Table A2—6.2.2. Eakman's data [10]

| Table of observations | | | | | |
|---|---|---|---|---|---|
| $t$ | A | B | C | D | X |
| 0.000E+0 | 0.1500E+1 | 0.1000E+0 | 0.0000E+0 | 0.0000E+0 | 0.0000E+0 |
| 0.100E+1 | 0.1115E+1 | 0.6198E+0 | 0.3111E+0 | 0.3033E+0 | 0.4920E−1 |
| 0.200E+1 | 0.9465E+0 | 0.5279E+0 | 0.5352E+0 | 0.4372E+0 | 0.1321E+0 |
| 0.300E+1 | 0.7022E+0 | 0.4256E+0 | 0.6145E+0 | 0.5058E+0 | 0.2048E+0 |
| 0.400E+1 | 0.6495E+0 | 0.2993E+0 | 0.7034E+0 | 0.5479E+0 | 0.2770E+0 |
| 0.500E+1 | 0.5133E+0 | 0.2468E+0 | 0.6494E+0 | 0.5916E+0 | 0.3294E+0 |
| 0.600E+1 | 0.5383E+0 | 0.1855E+0 | 0.7144E+0 | 0.5876E+0 | 0.4688E+0 |
| 0.700E+1 | 0.4501E+0 | 0.2103E+0 | 0.8014E+0 | 0.5738E+0 | 0.4837E+0 |
| 0.800E+1 | 0.3689E+0 | 0.1979E+0 | 0.8312E+0 | 0.5874E+0 | 0.4671E+0 |
| 0.900E+1 | 0.3961E+0 | 0.1489E+0 | 0.9286E+0 | 0.5725E+0 | 0.6203E+0 |
| 0.100E+2 | 0.3397E+0 | 0.1061E+0 | 0.8383E+0 | 0.5756E+0 | 0.6599E+0 |

Table A2—6.2.3. Direct integral estimates

| Parameter | Estimate | Stand. Dev. | Lower bound | Upper bound |
|---|---|---|---|---|
| K1 | 0.30826E+00 | 0.22679E−02 | 0.30373E+00 | 0.31280E+00 |
| K2 | 0.10744E+00 | 0.22709E−02 | 0.10290E+00 | 0.11198E+00 |

Table A2—6.2.4. Indirect estimates

| Parameter | Estimate | Stand. Dev. | Lower bound | Upper bound |
|---|---|---|---|---|
| K1 | 0.30415E+00 | 0.68620E−02 | 0.29043E+00 | 0.31788E+00 |
| K2 | 0.10421E+00 | 0.32615E−02 | 0.97691E−01 | 0.11074E+00 |

data in batch mode are shown in *Table A2—6.2.1. Table A2—6.2.2* gives the observations as contained in the file and printed by the program. The vector of initial concentrations (1.5, 1, 0, 0, 0) is known [10].

Though the data set taken from [8] is considered here as a 'real' one, in fact it has been obtained by Eakman by solving the kinetic equations at parameter values $k_1 = 0.3$ and $k_2 = 0.1$ and then adding relative errors. The direct integral method yields estimates which are very close to these true parameter values and requires 4 iteration steps. Results can be improved by 2 steps over the indirect method. The direct integral and the indirect estimates are shown in *Tables A2—6.2.3* and *A2—6.2.4*, respectively. Notice that the rate of convergence is much better than the one reported in [9].

## A2—7 INSTRUCTIONS FOR LOADING THE PROGRAM

Transfer to different computers should cause no problem since the coding is strictly 1966 ANSI FORTRAN. The only necessary change should be replacing the IBM-specific statement CALL CLEAR in the subroutine PLOAUX by an equivalent, or deleting it. The function of this statement is to clear the screen before displaying a plot.

On loading the program it must be decided if batch, interactive or interactive + printer mode is required (modes 0 to 2, respectively).

MODE 0 is straightforward batch operation requiring input and output files as units 5 and 6.

MODE 1 is interactive, requiring input and output as the same unit, unit 5.

MODE 2 allows for interactive input and operating dialogue, but with the results, being stored on a separate file for later disposal to a printer (or directly on the printer output file). This requires units 5 and 7 to be defined.

The mode is chosen by activating one of three lines in the BLOCK DATA segment:

```
DATA IR(5),  IW(6),  IPMT(6),  MODE(0)
DATA IR(5),  IW(5),  IPMT(5),  MODE(1)
DATA IR(5),  IW(7),  IPMT(5),  MODE(2)
```

IR, IW, IPMT are the unit identifiers of input, output and prompt peripheries, respectively. The first line belongs to batch mode, the second one to interactive mode and the third one to interactive + printer mode. From these three lines let the chosen one remain active and transform the other two into comments by writing a letter *C* in the first position of the lines. If you want to use the program in another mode, change the position of two letters *C* and recompile the BLOCK DATA segment.

The tape supplied contains 5 files as follows.

(1) Source file without user-supplied subroutines (approximately 2500 lines of FORTRAN code.

(2) User-supplied subroutine A2—4.1 to the test example A2—6.1.

(3) User-supplied subroutine A2—4.2 to the test example A2—6.2.

(4) Data input A2—6.1.2 to solve test example A2—6.1 in batch mode.

(5) Data input A2—6.2.2 to solve test example A2—6.2 in batch mode.

To estimate parameters in a particular algebraic or differential model, it is advantageous to change the appropriate FORTRAN lines in files (2) or (3), respectively, and then to compile and load it with the main part of the REPROCHE program. Note that user-supplied subroutines are very simple to prepare.

The program uses the IBM subroutine GGL (IX, X, N, IER) for generating a pseudo-random real number X, uniformly distributed between 0 and 1. It is advisable to replace this segment by one from the actual library. GGL is called in segment SNORM, IX is a starting value, $N=1$, and IER is an error flag which is not checked further in the program. The subroutine is needed only in simulation studies.

## A2—8 REFERENCES

[1] Bard, Y. (1974). *Nonlinear Parameter Estimation*. Academic Press, New York.

[2] Vajda, S., Valkó P. and Yermakova, A. (1990): Direct-indirect procedure for estimation of kinetic parameters. *Comp. et Chem. Eng.*

[3] Vajda, S., Valkó P. and Turányi, T. (1985): Principal component analysis of kinetic models. *Int. J. Chem. Kinet.*, **17**, 55.

[4] Gottwald, B. A. and Wanner, G. (1982): Comparison of numerical methods for stiff differential equations in biology and chemistry. *Simulation*, **17**, 169.

[5] Valkó, P. and Vajda, S. (1984): An extended ODE solver for sensitivity calculations. *Comp. et Chemistry*, **8**, 255.

[6] Himmelblau, D. M., Jones, C. R. and Bischoff, K. B. (1967): Determination of rate constants for complex kinetic models. *Ind. Eng. Chem. Fund.*, **6**, 539.

[7] Yermakova, A., Vajda, S. and Valkó, P. (1982): Direct integral method via spline-approximation for estimating rate constants. *Appl. Catal.*, **2**, 139.

[8] Bard, Y. (1970): Comparison of gradient methods for the solution of nonlinear parameter estimation problems. *SIAM J. Numer. Anal.*, **7**, 157.

[9] Eakman, J. M. (1969): Strategy for estimation of rate constants from isothermal reaction data. *Ind. Eng. Chem. Fund.*, **8**, 53.

[10] Garcia-Pena, J., Azen, S. P. and Bergman, R. N. (1983): On a modification of Marquardt's compromise: Rationale and applications. *Appl. Math. Corp.*, **12**, 1.

# APPENDIX A4

## Appendix to Chapter 4

# CAPCOS

Process Plant Capital Cost Estimation

MANUAL

Program Version 1
September 1986

Manual authors: F. TÁTRAI and P. SOLTÉSZ

Department of Chemical Technology
Technical University of Budapest

Budapest,
Budafoki út 42.
Hungary
H-1111

# CONTENTS


PREFACE 520
A4—1 INTRODUCTION 521
A4—2 PRINCIPLES AND DEFINITIONS 521
A4—2.1 Definition of terms used by the estimation 522
A4—2.2 Factors — relationships of cost elements 523
A4—2.3 The modules 524
A4—2.4 Estimating costs of equipment units 527
A4—2.5 Escalation 527
A4—3 INPUT SPECIFICATION TABLES 528
Index
A4—3.1 Chemical processing equipment 530
A4—3.2 Solids handling and special equipment 536
A4—3.3 Site development works 542
A4—3.4 Industrial buildings 545
A4—3.5 Chemical plant structures 546
A4—3.6 Offsite facilities 547
A4—4 PROGRAM USE 552
A4—4.1 Cost estimation with CAPCOS 552
A4—4.2 Updating the data files with CAPCOM 554
A4—4.2.1 Data file structure 555
A4—4.2.2 The ITEMS file 560
A4—5 EXAMPLES 560
A4—5.1 Example to demonstrate batch use 560
A4—5.2 Example to demonstrate the use of the data description language 562
A4—6 NOTES ON PROGRAM INSTALLATION 565


## PREFACE

CAPCOS is a program to estimate capital costs (fixed capital investment) in process plants. The first version proved to be useful in demonstrating the interdependency of cost elements, primarily for teaching purposes. To meet the needs of industry, some new features were built into the program:

— The user is free to define the equipment parameters (pressure, temperature, length, absorbed heat duty, etc.) in any commonly used unit (SI, British, CGS, etc.). The units can be redefined for each new equipment item in the same calculation.

— Industrial users who have their own cost files can use their own cost data and the cost estimation of the CAPCOS modules in the same calculation. The FOB (free on board) cost (see later) given by the user for any equipment item will simply owerwrite the estimated FOB cost of the item, and all cost elements will be calculated on the new (user defined) basis.

— Program and cost data (including the constants of cost estimation functions and the arithmetic form of these functions) are completely separate. This feature creates a new possibility: the possibility of altering the basic cost data and estimation functions without changing the program. Furthermore, it offers the possibility to create user defined equipment items, with the necessary cost modules, without any change in the program.

To faciliate all of these new possibilities a new data description and input language has been developed, and a completely new program was written.

## A4—1 INTRODUCTION

One of the critical phases in the early stage of a process plant project occurs when an economic evaluation must be prepared. Many process alternatives must be generated and analyzed before an economically viable project emerges. It is very important to use consistent estimating techniques so that alternatives can be compared on the same basis, and comparisons can also be made between projects. CAPCOS is a program to estimate capital costs (fixed capital investment) in a process plant, and to demonstrate the interdependency of cost elements for teaching purposes.

The program package CAPCOS was based on a 'modular technique' published by K. M. Guthrie (*Chem. Eng.*, **24**, March, 1969).

This method was developed based on feedback data from 42 process plant projects in the United States in 1968. The different cost elements can be updated using factorial relationships.

The program is suitable for cost studies on single equipment items and also total investment cost estimation of complete processes.

## A4—2 PRINCIPLES AND DEFINITIONS

The accurate and consistent estimation of capital costs is of great importance in engineering chemical plants. The preliminary cost estimation is indispensable for decisions concerning alternative solutions.

All cost estimating techniques have inherent inaccuracies. This calls for identical methods and identical basic data for calculations comparing alternative solutions. To fulfil the above task, a set of subroutines is needed to estimate all cost factors.

The method published by K. M. Guthrie (Capital Cost Estimating, *Chem. Eng.*, March 24, 1969) has been choosen as the basis of the method. This method was based on the analysis of 42 projects implemented in the United States in 1968.

## A4—2.1 Definition of terms used by the estimation

The following cost factors are involved:

| Identification | Description |
|---|---|
| E | *FOB* (free on board) *cost* of equipment, i.e. equipment price excluding transport costs. |
| M | *Auxiliary materials*. The costs of material which are needed up to commissioning.<br>Possible breakdown:<br>Piping<br>Concrete<br>Steel<br>Instruments<br>Electrical<br>Insulation<br>Paint |
| L | *Field labour costs*: The labour cost incurred during erection and commissioning<br>Possible breakdown:<br>Material erection<br>Commissioning |
| I | *Indirect costs*:<br>1. Import duties, taxes, subsidies, etc.<br>2. Construction overheads<br>3. Construction engineering |
| B | *Module cost*<br>B=E+M+L+I |

The sum of the module cost gives the total capital cost. Contingencies should be added to complete the estimate (this is not done by the program).

The contingency covers expenses which will definitely be incurred, but the detailed estimation of which is not feasible. The contingency is *not* meant to compensate for inaccuracies of the estimation, but to cover actual cost factors.

## A4—2.2 Factors — relationships of cost elements

Statistical relationships have been developed between the cost elements listed in Section A4—2.1. The ratios of the average values of cost elements can be used to estimate several elements based on a simple known element. This is generally the FOB cost.

The following ratios — referred to as factors — have been developed as part of the Guthrie method:

| Factor identification | Description | Meaning |
|---|---|---|
| MF | material factor | auxiliary material to FOB |
| LF | labour factor | field labour cost to FOB |
| DF | direct cost factor | direct cost to FOB |
| IF | indirect factor | indirect cost to FOB |
| BF | bare module factor | module cost to FOB |

The interrelationships of factors and cost elements are shown in *Fig. A4—1.* Cost elements are represented by rectangles. Arrows point from one cost element to another, and the factors on the arrows convert one element into another.

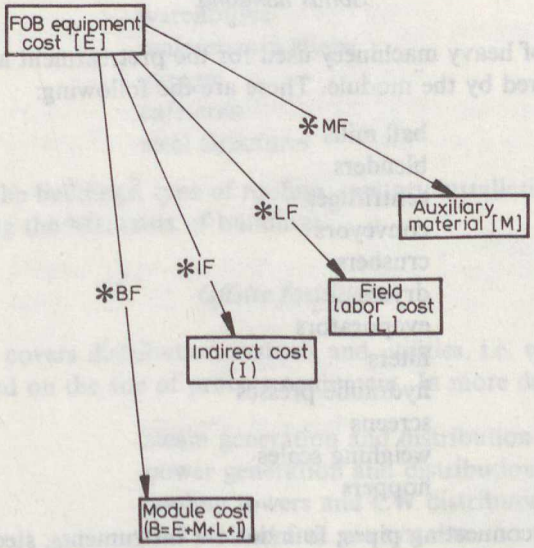



Fig. A4—1. Interrelation of factors and cost elements

## A4—2.3 The modules

All potential costs have been grouped into six modules, five of which are direct cost modules with an indirect one.

These modules are the following:

1. Chemical processing
2. Solids handling
3. Site development
4. Industrial buildings
5. Offsite facilities
6. Project indirect costs

*Chemical processing*

These represent the ML costs of the process equipment of the plant. The ML costs of interconnecting pipes, instruments, minor supports, platforms and ladders, concrete foundation, insulation and painting, electrical installations are also included, up to plant start-up. Cost estimating is the most detailed in this module, since chemical processing represents the most important module in a chemical project.

*Solids handling*

The MD costs of heavy machinery used for the pretreatment and transport of materials are covered by the module. These are the following:

ball mills
blenders
centrifuges
conveyors
crushers
dryers
evaporators
filters
hydraulic presses
screens
weighing scales
hoppers

The cost of interconnecting pipes, foundation, instruments, steel supports and electrical installations are included in the ML cost of each item of equipment.

Cost elements are elaborated in less detail than for chemical processing. Estimating is based on one or sometimes two parameters.

*Site development*

The costs of the following activities and equipment are included in this module:

land surveys and fees
dewatering and drainage
site clearing
excavation
grading
underground sewers
piling
roads, walkways and paving
parking lots
landscaping
fencing
fire protection facilities

*Industrial buildings*

These are the following:

administration offices
laboratories
medical
warehouses
maintenance shops
garages
cafeterias
steel structures

The sizes of the buildings, type of roofing, sanitary installations are considered when calculating the ML costs of buildings.

*Offsite facilities*

This module covers distribution systems and utilities, i.e. equipment which is generally located on the site of process equipment. In more details:

steam generation and distribution system
power generation and distribution system
cooling towers and CW distribution system
fuel oil and fuel gas systems
flowdown and flare
pollution control facilities
fireloops and hydrants

separators and ponds
yard lighting and communications
receiving, storage, shipping

Within the last item the following ML costs are grouped:

automation
railroads
docks and wharfs
tankage

*Project indirect costs*

Indirect costs are incurred by all equipment belonging to the above five modules. These indirect costs may be the following:

freight, insurance, taxes
sales, other taxes
freight, packing, insurance
ocean freight
marine insurance
import duties

Construction overhead

fringe benefits
labour burden
field supervision
temporary facilities
construction equipment
small tools
miscellaneous

Construction engineering costs

project engineering
process engineering
design and drafting
procurement
home office construction
office overhead

## A4—2.4 Estimating costs of equipment units

Each estimate for equipment is based on major parameters of the unit concerned. For chemical processing equipment multiple parameters are used, while only one or two are used for the others.

In the first step a so-called base cost is calculated. This is not the price of the unit being considered, but the FOB cost of a standard unit of the same size as the one in question. The actual value for the unit considered is calculated by multiplying the base cost by the following factors:

*Pressure factor*: used if the design pressure of the base unit differs from that of the actual one.

*Material factor*: the base units are for carbon steel, additional costs due to other construction materials can be taken into account through multiplying by this material factor.

Of course, the individual factors and the appropriate formulae vary with each type of equipment. They are defined in the input specification tables.

## A4—2.5 Escalation

The calculated costs to be brought up from 1968 to present-day costs. This is achieved with the following three of inflation factors:

*Material inflation factor* which brings material costs [$E+M$] up-to-date from 1968 base level

*Labour inflation factor* which brings labour costs [L] up-to-date from the 1968 base level

*Indirect costs inflation factor* which brings indirect costs [I] up-to-date from the 1968 base level.

# A4—3 INPUT SPECIFICATION TABLES

## Index


| *Chemical processing equipment* | *Classification number* |
|---|---|
| Process furnace | 1 |
| Direct fired heater | 2 |
| Heat exchanger | 3 |
| Double pipe heat exchanger | 4 |
| Air cooler | 5 |
| Pressure vessel, vertical | 6 |
| Pressure vessel, horizontal | 7 |
| Tray stack | 8 |
| Packing | 9 |
| Lining | 10 |
| Pump | 11 |
| Process gas compressor | 12 |

| *Solids handling and special equipment* | *Classification number* |
|---|---|
| Agitators | 14 |
| Air compressors | 15 |
| Air conditioners | 16 |
| Air dryers | 17 |
| Bagging machines | 18 |
| Blenders | 19 |
| Blowers and fans | 20 |
| Boilers (industrial) | 21 |
| Centrifuges | 22 |
| Conveyors | 23 |
| Cranes | 24 |
| Crushers | 25 |
| Crystallizers | 26 |
| Dryers | 27 |
| Ducting | 28 |
| Dust collectors | 29 |
| Ejectors | 30 |
| Elevators | 31 |
| Evaporators | 32 |
| Filters | 33 |

| | |
|---|---|
| Flakers | 34 |
| Generator sets (portable) | 35 |
| Hoppers | 36 |
| Hydraulic presses | 37 |
| Mills | 38 |
| Screens | 39 |
| Stacks | 40 |
| Tank heaters | 41 |
| Weight scales | 42 |

*Site developments works*

| | |
|---|---|
| Dewatering and drainage | 43, type 1—3 |
| Fencing | 43, type 4—10 |
| 3—18 | |
| Fire protection | 43, type 13—14 |
| Piling | 43, type 15—21 |
| Roads, walkways, paving | 43, type 22—26 |
| Sewer facilities | 43, type 27—30 |
| Site clearing, excavation | 43, type 31—34 |
| Foundation excavation | 43, type 35—37 |
| Trench excavation | 43, type 38—42 |
| 3—17, 18, 19 | |
| Miscellaneous materials | 43, type 43—46 |
| Industrial buildings | 44 |
| 3—20, 21, 22 | |
| Chemical plant structures | 45 |
| Offsite facilities | 46 |
| Air systems | 46, type 1—2 |
| Blowdown and flare | 46, type 3 |
| Cooling tower and cooling water | 46, type 4—6 |
| Fireloop and hydrants | 46, type 7 |
| Fuel systems | 46, type 8—9 |
| General water systems | 46, type 10—12 |
| Power generation distribution | 46, type 13—16 |
| Receiving, shipping, storage | 46, type 17—36 |
| 3—25, 26, 28, 29 | |
| Steam generation and distribution | 46, type 37—39 |
| Yard lighting and comm. | 46, type 40 |
| Yard transfer lines | 46, type 41 |
| Offsite storage | 47 |

## A4—3.1 Chemical processing equipment

Classification number 1
*Process furnace*
Equipment input data

| Secondary keywords | Interpretation | Acceptable range | Cases |
|---|---|---|---|
| dty | design type | 1–3 | 1—process heater<br>2—pyrolysis heater<br>3—reformer (without catalyst) |
| mty | radiant tube material | 1–3 | 1—carbon steel<br>2—chrome/moly<br>3—stainless |
| p | pressure | 0.689–206.84 bar | |
| ahd | absorbed heat duty | 632.88–63.288 kW | |

Classification number 2
*Direct fired heater*
Equipment input data

| Secondary keywords | Interpretation | Acceptable range | Cases |
|---|---|---|---|
| dty | design type | 1–3 | 1—cylindrical<br>2—dowtherm |
| mty | radiant tube material | 1–3 | 1—carbon steel<br>2—chrome/moly<br>3—stainless steel |
| p | pressure | 0.68947–241.3145 bar | — |
| ahd | absorbed heat duty | 63.288–6328.8 kW | |

Classification number 3
*Heat exchanger*
Equipment input data

| Secondary keywords | Interpretation | Acceptable range | Cases |
|---|---|---|---|
| dty | design type | 1–4 | 1—kettle or reboiler<br>2—floating head<br>3—U tube<br>4—fixed tube sheet |
| p | pressure | 0.68947–68.947 bar | — |
| a | heat transfer surface area | 9.29–929 $m^2$ | — |

Classification number 4
*Double pipe heat exchanger*
Equipment input data

| Secondary keywords | Interpretation | Acceptable range | Cases |
|---|---|---|---|
| mty | shell/tube material | 1–2 | 1—carbon steel/carbon steel<br>2—carbon steel/stainless steel |
| p | pressure | 0.68947–68.947 bar | — |
| a | calculated heat transfer surface area | 0.0929–9.29 $m^2$ | — |

Classification number 5
*Air cooler*
Equipment input data

| Secondary keywords | Interpretation | Acceptable range | Cases |
|---|---|---|---|
| a | area of air-cooled surface | 143.995–14399.5 $m^2$ | — |
| p | tube length | 0.68947–68.947 bar | — |
| l | tube length | 0.3048–6.096 m | — |
| mty | tube material code | 1–4 | 1—carbon steel<br>2—aluminium<br>3—stainless steel<br>4—monel |

Classification number 6
*Pressure vessel, vertical*
Equipment input data

| Secondary keywords | Interpretation | Acceptable range | Cases |
|---|---|---|---|
| p | pressure | 0.68947–68.947 bar | — |
| h | vessel height | 0.6096–30.48 m | — |
| d | vessel diameter | 0.3048–3.048 m | — |
| mty | pressure vessel material | 1–8 | 1—carbon steel, clad<br>2—stainless 316, clad<br>3—monel, clad<br>4—titanium, clad<br>5—carbon steel, solid<br>6—stainless 316, solid<br>7—monel, solid<br>8—titanium, solid |

Classification number 7
*Pressure vessel, horizontal*
Equipment input data

| Secondary keywords | Interpretation | Acceptable range | Cases |
|---|---|---|---|
| p | pressure | 0.68947–68.947 bar | — |
| h | vessel length | 0.6096–30.48 m | — |
| d | vessel diameter | 0.3048–3.048 m | — |
| mty | pressure vessel material | 1–8 | 1—carbon steel, clad<br>2—stainless 316, clad<br>3—monel, clad<br>4—titanium, clad<br>5—carbon steel, solid<br>6—stainless 316, solid<br>7—monel, solid<br>8—titanium, solid |

Classification number 8
*Tray stack*
Equipment input data

| Secondary keywords | Interpretation | Acceptable range | Cases |
|---|---|---|---|
| dty | tray type | 1–6 | 1—grid (no downcomer)<br>2—plate<br>3—sieve<br>4—trough or valve<br>5—bubble cap<br>6—Koch cascade |
| mty | tray material | 1–3 | 1—carbon steel<br>2—stainless steel<br>3—monel |
| sp | tray spacing | 0.245–0.762 m | — |
| d | tray diameter | 0.3048–3.048 m | — |
| h | tray stack height | 1.524–30.48 m | — |

Classification number 9
*Packing*
Equipment input data

| Secondary keywords | Interpretation | Acceptable range | Cases |
|---|---|---|---|
| typ | packing type | 1–7 | 1—Raschig rings<br>2—Berl saddles<br>3—activated carbon<br>4—alumina<br>5—coke<br>6—crushed limestone<br>7—silica gel |
| mty | material type of packing | The number of cases and the acceptable range depend on the type defined | |
| siz | size of packing | | |
| v | packed volume | 0.02832–212.4 $m^3$ | — |

Classification number 10
*Lining*
Equipment input data

| Secondary keywords | Interpretation | Acceptable range | Cases |
|---|---|---|---|
| typ | lining type | 1–6 | 1—acid brick<br>2—firebrick<br>3—rubber<br>4—refractory<br>5—gunite<br>6—chemical lead |
| a | lining surface area | 0–929 $m^2$ | — |
| th | thickness | 0.03–0.06 m in case 1<br>0.04–0.09 m in case 2<br>0 –0.003 m in case 3<br>0.01–0.05 m in case 4<br>0.01–0.05 m in case 5<br>Ø in case 6 | |
| M | total mass of lining | 0–6803.85 in case 6 only | |

Classification number 11
*Pump*
Equipment input data

| Secondary keywords | Interpretation | Acceptable range | Cases |
|---|---|---|---|
| typ | pump type | 1–4 | 1—centrifugal pump/motor<br>2—centrifugal pump/turbine<br>3—reciprocating pump/motor<br>4—reciprocating pump/steam |
| dip | differential pressure | 0.68947–68.947 bar | |
| sup | suction pressure | 0.68947–68.947 bar | |
| t | system pressure | 256.5–397.89 K | |
| mty | material type | 1–10 (for types 1-2)<br>1–4 (for types 3–4) | 1—cast iron<br>2—bronze<br>3—cast steel<br>4—stainless<br>5—carpenter 20<br>6—worthite<br>7—hastalloy C<br>8—monel<br>9—nickel<br>10—titanium |

Classification number 12
*Progress gas compressor*
Equipment input data

| Secondary keywords | Interpretation | Acceptable range | Cases |
|---|---|---|---|
| typ | gas compressor type | 1–5 | 1—centrifugal/motor<br>2—reciprocating/steam<br>3—centrifugal/turbine<br>4—reciprocating/motor<br>5—reciprocating/gas engine |
| bp | brake power | 22.065–2942 kW | |

## A4—3.2 Solids handling and special equipment

| Name | Classification number | Equipment input data | | | | |
|---|---|---|---|---|---|---|
| | | Secondary keywords | Interpretation | Acceptable range | | Cases |
| Agitators | 14 | typ | design type | 1–2 | | 1—propeller |
| | | *C* | power rating | 0.1–100,000 | (hp) | |
| Air compressor | 15 | *C* | volumetric flow rate | 0.1–100,000 | ($ft^3$/m) | |
| Air conditioners | 16 | typ | design type | 1–3 | | 1—window vent<br>2—floor-mounted<br>3—roof-top |
| | | *W* (only if typ = 3) | weight | 10–40 | (tons) | |
| Air dryers | 17 | *C* | capacity | 0.1–100,000 | ($ft^3$/min) | |
| Bagging machines | 18 | typ | design type | 1–2 | | 1—weight<br>2—volume |
| | | *C* | capacity | 1–100,000 | (bags/min) | |
| Blenders | 19 | *C* | capacity | 1–100,000 | ($ft^3$) | |
| Blowers and fans | 20 | *C* | capacity | 1–100,000 | ($ft^3$) | |
| Boilers (industrial) | 21 | *P* | pressure | 14–700 | psi | |

| | | | | | |
|---|---|---|---|---|---|
| Centrifuges | 22 | typ | design type | 1–4 | 1—horizontal basket<br>2—vertical basket<br>3—solid bowl (SS)<br>4—sharples (SS) |
| | | *D* (in cases 1, 2) | diameter | 1–1,000 in | |
| | | *C* (in cases 3, 4) | capacity | 1–1,000 hp | |
| Conveyors | 23 | typ | design type | 1–5 | 1—belt<br>2—bucket (height)<br>3—roller<br>4—screen<br>5—vibrating |
| | | *L* | length | 1–10,000 ft | |
| | | *W* (for typ 1, 3, 5) | width | 12–56 cm (typ = 5)<br>10–24 cm (typ = 3)<br>8–40 cm (typ = 5) | |
| | | *D* (for typ 4) | diameter | 4–20 cm | |
| | | *C* (for typ 2) | capacity | 10–150 tons/h | |
| Cranes | 24 | *SP* | span | 5–120 ft | |
| | | *C* | capacity | 1–1000 tons | |

## A4—3.2 (continued)

| Name | Classification number | Equipment input data | | | | |
|---|---|---|---|---|---|---|
| | | Secondary keywords | Interpretation | Acceptable range | | Cases |
| Crushers | 25 | typ | crusher type | 1–4 | | 1—cone<br>2—gyratory<br>3—jaw<br>4—pulverizers |
| | | *C* | capacity | 0.1–10,000<br>1–10,000 | (tons/h)<br>(1b/h) | |
| Crystallizers | 26 | typ | crystallizer type | 1–3 | | 1—growth<br>2—forced circulation<br>3—batch |
| | | *C* (for typ 1, 2) | capacity | 0.1–10,000 | (tons/day) | |
| | | *V* (for typ 3) | volume | 1–10,000 | gal | |
| Dryers | 27 | typ | dryer type | 1–3 | | 1—drum<br>2—pan<br>3—rotary vacuum |
| | | *A* | area | 1–10,000 | $ft^2$ | |
| Ductwork | 28 | *L* | length | 1–10,000 | ft | |
| | | typ | type of ductwork | 1–3 | | 1—aluminium<br>2—galvanized<br>3—stainless |

| | | | | | |
|---|---|---|---|---|---|
| Dust collectors | 29 | typ | type of dust collector | 1–3 | 1—cyclones<br>2—cloth filter<br>3—precipitators |
| | | *C* | capacity | 0.1–10,000 ($ft^3$/min) | |
| Ejectors | 30 | typ | ejector type | 1–3 | 1—1-stage<br>2—4-stage, barometric<br>3—5-stage, barometric |
| | | *C* | capacity | 1–10,000 (lb/h) | |
| | | *SU* | suction pressure | 0–30 torr (gauge)<br>2–25 torr (gauge)<br>0.2–2 torr (gauge) | |
| Elevators | 31 | *C* | capacity (for typ 2) | 1,000–15,000 lb | |
| | | *H* | height | 1–1,000 ft | |
| | | typ | elevator type | 1–2 | 1—freight<br>2—passenger |
| Evaporators | 32 | typ | evaporator type | 1–4 | 1—forced circulation<br>2—vertical tube<br>3—horizontal tube<br>4—jacketed vessel (glass-lined) |
| | | *A* (for typ 1—3) | area | 1–10,000 $ft^2$ | |
| | | *V* (for typ 4) | volume | 1–10,000 gal | |

## A4—3.2 (continued)

| Name | Classification number | Equipment input data | | | | |
|---|---|---|---|---|---|---|
| | | Secondary keywords | Interpretation | Acceptable range | | Cases |
| Filters | 33 | typ | filter type | 1–5 | | 1—plates and press<br>2—pressure leaf-wet<br>3—pressure leaf-dry<br>4—rotary drum<br>5—rotary disc |
| | | *A* | effective area | 1–100,000 | ($ft^2$) | |
| Flakers (drum) | 34 | *A* | effective area | 0.1–100,000 | ($ft^2$) | |
| Generator sets (portable) | 35 | *C* | capacity in kW | 1–120 | (kW) | |
| Hoppers | 36 | typ | hopper type | 1–2 | | 1—conical<br>2—silos |
| | | *C* | capacity | 1–100,000 | ($ft^3$) | |
| Hydraulic presses | 37 | *P* | pressure | 0–1200 | (psi) | |
| | | *A* | plate area | 1–3 | | |
| Mills | 38 | typ | mill type | 1-3 | | 1—ball<br>2—roller<br>3—hammer |
| | | *C* | capacity | 1–10,000 | (tons/h) | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Screens | 39 | typ | | 1–2 | | 1—vibrating single<br>2—vibrating double |
| | | *A* | surface area | 1–10,000 | $ft^2$ | |
| Stacks | 40 | *D* | diameter | 20–60 | in | |
| | | *H* | height | 1–100 | ft | |
| Tank heaters | 41 | typ | heater type | 1-2 | | 1—steam coil<br>2—immersion |
| | | *A* (for typ 1) | surface area | 1-10,000 | $ft^2$ | |
| | | *C* (for typ 2) | capacity | 1–10,000 | kW | |
| Weighing scales | 42 | typ | scale type | 1–3 | | 1—portable beam<br>2—portable dial<br>3—truck |
| | | *W* (for typ 3) | weight | 10–75 | tons | |

## A4—3.3 Site development works

| Name | Classification number | Equipment input data | | | |
|---|---|---|---|---|---|
| | | Secondary keywords | Interpretation | Acceptable range | Cases |
| Site development | 43 | typ | site development type | 1–46 | Dewatering and drainage:<br>1—pumping system (rented)<br>2—well-point dewatering system<br>3—drainage trench<br>Fencing:<br>4—complete fence, light<br>5—complete fence, heavy<br>6—chain, link<br>7—gate, light<br>8—gate, heavy<br>9—gate, chain<br>10—corner post<br>Fire protection:<br>11—pumps, fire-house and 2 fire-trucks<br>Piling:<br>13—untreated wood<br>14—creosoted wood<br>15—precast concrete<br>16—concrete cast in place<br>17—steel pipe, concrete filled<br>18—steel section<br>19—sheet piling, steel<br>20—sheet piling, wood<br>21—pile driver setup<br>Roads, walkways and pavings:<br>22—paving, reinforced, with subbase |

| | | | | | |
|---|---|---|---|---|---|
| Site development | 43 | typ | | 1–46 | 23—paving, asphalt top, on existing base<br>24—paving asphalt top, with subbase<br>25—gravel surface<br>26—parking lots, black-top surface<br>Sewer facilities:<br>27—asbestos cement pipe<br>28—concrete pipe, reinforced<br>29—vitrified clay piping<br>30—septic tank<br>Site clearing, excavation and grading:<br>31—machine cuts<br>32—clearing and grubbing<br>33—general grading<br>34—final levelling<br>Foundation excavation:<br>35—machine excavation<br>36—machine plus hand trim<br>37—hand work<br>Trench excavation:<br>38—machine<br>39—hand labour<br>40—trench shoring, sheeting<br>41—machine plus hand trim<br>42—hand labour only<br>Miscellaneous materials:<br>43—sand<br>44—gravel<br>45—dirt fill<br>46—crushed limestone |
| | | *tim* (for typ 1, 2)<br>*L* (for typ 3, 4, 5, 6)<br>*A* (for typ 12) | renting time (month)<br>length (ft)<br>area ($yd^2$) | 1–100<br>1–100,000<br>1–10,000 | |

## A4—3.3 (continued)

| Name | Classification number | Equipment input data | | | |
|---|---|---|---|---|---|
| | | Secondary keywords | Interpretation | Acceptable range | Cases |
| Site development (continued) | 43 | *L* (for typ 13–18) | length of piling (ft) | 1–100,000 | |
| | | *A* (for typ 19, 20) | area of sheet piling ($ft^2$) | 1–100,000 | |
| | | *A* (for typ 22–26) | area of roads ($yd^2$) | 1–10,000 | |
| | | *TH* (for typ 22) | top layer thickness (cm) | 2–10 | |
| | | *TH* (for typ 24) | top layer thickness (cm) | 1–5 | |
| | | *TH* (for typ 25) | top layer thickness (cm) | 1–10 | |
| | | *L* (for typ 27–29) | pipe length (ft) | 1–10,000 | |
| | | *D* (for typ 27–29) | pipe diameter (cm) | 10–50 | |
| | | *V* (for typ 31) | volume ($yd^3$) | 1–100,000 | |
| | | *A* (for typ 32–34) | site area ($yd^2$) | 1–100,000 | |
| | | *V* (for typ 35–37) | excavated volume ($yd^3$) | 1–100,000 | |
| | | *L* (for typ 38) | length (ft) | 1–100,000 | |
| | | *D* (for typ 38) | depth (ft) | 1–8 | |
| | | *V* (for typ 39) | excavated volume ($yd^3$) | 1–100,000 | |
| | | *A* (for typ 40) | area ($ft^2$) | 1–100,000 | |
| | | *V* (for typ 41–46) | volume ($yd^3$) | 1–100,000 | |

## A4—3.4 Industrial buildings

| Name | Classification number | Equipment input data | | | |
|---|---|---|---|---|---|
| | | Secondary keywords | Interpretation | Acceptable range | Cases |
| Industrial buildings | 44 | *C* | building characteristics | 1–3 | (1) low cost: prefabricated light-steel frame and roof, transit of metal sheet walls, concrete floor, minimum furnishing and accomodation inside<br>(2) average: prefabricated medium steel frame and trusses, wood or concrete walls, concrete floor and roof, average furnishing and accomodation inside<br>(3) high cost: custom-designed heavy-steel frame, heavy sidings, brick or concrete block walls, concrete floor and roof, expensive furnishing and accomodation inside |
| | | typ | building type | 1-9 | 1—administration office<br>2—cafeteria<br>3—compressor house with bridge crane<br>4—control house (equipped)<br>5—garage<br>6—maintenance shop<br>7—laboratory and medical<br>8—process building<br>9—warehouse |
| | | *HV* | heating and ventillating | 0–1 | 0—no heating and ventillating system<br>1—heating and ventillating |
| | | *PL* | plumbing | | 0—none<br>1—necessary |
| | | *FIR* | fire prevention equipment (alarms, extinguishers, sprinkling systems) | | 0—none<br>1—necessary |

## A4—3.5 Chemical plant structures

| Name | Classification number | Equipment input data | | | |
|---|---|---|---|---|---|
| | | Secondary keywords | Interpretation | Acceptable range | Cases |
| Chemical plant structures | 45 | *CHR* | structure characteristics | 1–3 | (1) light: not high, mostly access and platform areas, minor equipment supports and stairways, open construction<br>(2) medium: average height, designed for average elevated equipment supports, platform areas and stairways, open construction used for most of the structure<br>(3) heavy: high, designed for heavy equipment supports, crane or catalyst leading structures, some covered construction (metal sheetings with potential wind loads), outside freight elevator structure (but elevator not included) |
| | | *A* | plant area ($ft^2$) | 1–10,000 | |
| | | *H* | height (ft) | 1–50 (for CHR = 1)<br>1–200 (for CHR = 2)<br>1–300 (for CHR = 3) | |

## A4—3.6 Offsite facilities

| Name | Classification number | Equipment input data | | | |
|---|---|---|---|---|---|
| | | Secondary keywords | Interpretation | Acceptable range | Cases |
| Offsite facilities | 46 | typ | offsite facility type | 1–41 | Air systems:<br>1—instrument air (compressor facilities, air dryer, air receiver and distribution)<br>2—plant air (compressor facilities, air receiver and distribution<br>Blowdown and flare:<br>3—for general purposes (including flare lines, blowdown drum and disposal pit)<br>Cooling tower and cooling water distribution:<br>4—cooling tower<br>5—distribution system for general purposes<br>6—river intake installation<br>Fireloop and hydrants:<br>7—for general purposes<br>Fuel systems:<br>8—fuel oil (incl. pumps, storage, piping, controls and distribution)<br>9—fuel gas (incl. receiver, piping, controls and distribution) |

## A4—3.6 (continued)

| Name | Classification number | Equipment input data | | | |
|---|---|---|---|---|---|
| | | Secondary keywords | Interpretation | Acceptable range | Cases |
| Offsite facilities | 46 | typ | offsite facility type | 1–41 | General water system:<br>10–treated water, filtered and softened<br>11–distilled water<br>12–drinking and service water, general facilities<br>Power generation and distribution:<br>13–generating facilities<br>14–electrical distribution for general purposes<br>15–main transformer station<br>16–secondary transformer station<br>Receving, shipping, storage:<br>17–forklift truck<br>18–pallet truck<br>19–payloader<br>20–tank trailer<br>21–tractor<br>22–tractor shovel<br>23–automotive shipping facilities<br>24–docks and wharves, light construction<br>25–docks and wharves, medium construction<br>26–docks and wharves, heavy construction<br>27–dredging general operations<br>28–railroad, straight track |

| | | | | | |
|---|---|---|---|---|---|
| Offsite facilities | 46 | typ | offsite facility type | 1-41 | 29–turnout<br>30–bumper<br>31–blinker and gate<br>32–grade and ballast<br>33–locomotive (battery)<br>34–locomotive (diesel)<br>35–tank car<br>36–railroad shipping facilities<br>Steam generation and distribution:<br>37–packaged boiler unit<br>38–field erected boiler unit<br>39–steam distribution for general purposes<br>Yard lighting and communications:<br>40–for general purposes<br>Yard transfer lines and purposes:<br>41–for general purposes |
| | 46 | *C* (for typ 1–3) | cost ($) | 100–1,000,000 | |
| | | *C* (for typ 4) | capacity (gal/min) | 1000–50,000 | |
| | | *T* (for typ 4) | temperature difference (F) | 10–30 | |
| | | *C* (for typ 5, 6) | capacity (gal/min) | 1–10,000 | |
| | | *C* (for typ 7–9) | cost ($) | 100–1,000,000 | |
| | | *V* (for typ 10, 11) | volume (gal) | 0–100,000 | |
| | | *C* (for typ 12) | cost ($) | 100–1,000,000 | |
| | | *C* (for typ 13) | generatig capacity (kVA) | 500–20,000 | |
| | | *C* (typ 14) | capacity (kW) | 500–50,000 | |
| | | *C* (typ 15) | capacity (kVA) | 1,000–30,000 | |
| | | *C* (typ 16) | capacity (kVA) | 300–3,000 | |
| | | *C* (typ 17) | capacity (1b) | 2000–15,000 | |
| | | *PTY* (typ 18) | pallet truck type | 1–2 | 1–hydraulic<br>2–electric |

## A4–3.6 (continued)

| Name | Classification number | Equipment input data | | | |
|---|---|---|---|---|---|
| | | Secondary keywords | Interpretation | Acceptable range | Cases |
| | | *PLT* (typ 19) | payloader type | 1–4 | 1—small, gas engine<br>2—big, gas engine<br>3—small, diesel<br>4—big, diesel |
| | | *MTY* (typ 20) | material type | 1–3 | 1—carbon steel<br>2—aluminium<br>3—stainless |
| | | *TTY* (typ 21) | tractor type | 1–2 | 1—gasoline<br>2—diesel |
| | | *V* (for typ 22) | bucket volume ($yd^3$) | 1–5 | |
| | | *A* (for typ 24–26) | area ($ft^2$) | 1–10,000 | |
| | | *V* (for typ 27) | volume ($yd^3$) | 1–1,000 | |
| | | *L* (for typ 28) | length (ft) | 1–10,000 | |
| | | *L* (for typ 32) | length (ft) | 1–10,000 | |
| | | *W* (for typ 33) | weight (tons) | 5–15 | |
| | | *W* (for typ 34) | weight (tons) | 1–5 | |
| | | *C* (for typ 37) | steam capacity (1b/h) | 10,000–150,000 | |
| | | *P* (for typ 37) | steam pressure (psi) | 10–700 | |
| | | *T* (for typ 37) | superheat (F) | 0–400 | |
| | | *C* (for typ 38) | steam capacity (1b/h) | 150,000–300,000 | |
| | | *P* (for typ 38) | steam pressure (psi) | 300–1100 | |
| | | *T* (for typ 38) | superheat (F) | 0–400 | |
| | | *C* (for typ 39) | steam capacity (1b/h) | 1–10,000 | |
| | | *C* (for typ 40, 41) | cost ($) | 5000–100,000,000 | |

| | | | | | |
|---|---|---|---|---|---|
| Offsite storage | 47 | type | storage vessel type | 1–4 | 1—small capacity storage<br>2—high capacity storage<br>3—horizontal pressure storage vessel<br>4—spherical pressure storage vessel |
| | | *C* | tank capacity (gal) | 10–100,000 | |
| | | *RTY* | roof type | 1–2 | 1—light<br>2—conical |
| | | *MTY* | material type | | 1—carbon steel<br>2—aluminium<br>3—rubber lined<br>4—lead lined<br>5—stainless<br>6—glass lined |

if typ = 1

## A4—4 PROGRAM USE

The CAPCOS program enables the user to interrogate the individual data files for each type of equipment (files 1.D to 47.D), and to present the resulting costs summarized in tabular form. The CAPCOM program creates the individual cost data files interactively with the user, or batchwise using 1.DRS–47.DRS as the input.

CAPCOM is used to permanently update the cost files, and CAPCOS to obtain the cost estimates from these files. The use of these two programs is described in the following two sections.

### A4—4.1 Cost estimation with CAPCOS

On running CAPCOS, the user must first define the mode in which it is to operate, as requested by the first prompt:

mode 1 — batch (data file input results file output)
mode 2 — interactive (terminal input terminal output)
mode 3 — batch + printer (data file input
results to terminal plus
results to file for printer)

To use the batch modes, the input file must have been created previously either by an editor or by the input file saved from a previous interactive run. Interactive use gives the option of saving the input file for later runs. Modified cost estimates can therefore be made by making the appropriate alterations to this file with the normal editor before running it in batch mode through the program again.

In interactive mode the program then asks for a file name in which it will store this input data for further use. CAPCOS.INP is this default file name.

The user must give a title to his estimate (maximum 72 characters). In interactive mode the program asks for the title; in batch mode the first line of the input file is considered to be the title. The individual cost data and the conditions of the estimate can be described using the primary and secondary keywords of the CAPCOS input language. The sequence of the primary keywords is arbitrary. There are, however, four primary keywords which control all estimates later in the sequence. These are the following:

YEAR to determine the year of the estimate
ESC to give escalation from the base year and
BRITISH/SI to define the standard units of equipment data.

It is recommended that these conditions be defined first. The repetitive use of these keywords in the same estimate overwrites the previously defined conditions.

YEAR 1987

means that the actual year is 1987.

The program uses three escalation factors:

Material cost means yearly escalation factor from base cost year for material costs.

Labour cost means yearly escalation factor from base cost year for labour costs.

Indirect cost means yearly escalation factor from base cost year for indirect costs.

The values can be specified following the ESC keyword, e.g.

ESC 1.07, 1.08, 1.12

The standard (default) units in which the equipment data will be given can be fixed by the keywords

BRITISH
SI

Hence, when equipment information is given without units, this keyword will have defined the units that will be assumed.

The defaults for these keywords are the following:

YEAR 1968
ESC 1.0, 1.0, 1.0
SI

Typing in the estimate control keywords the run is now set up to receive the details of equipment items to be included in the cost estimate. Each item is entered by first entering the item identification number, and then the number of identical units

e.g. ITEM 1,2

The standard data for this item are then displayed, showing the data necessary to define the equipment, and the default values used in the data file. The user is then asked to enter a title to define his specific equipment item (in batch mode the next line is considered to be the title), then in the following lines the user can input the data specific to this unit. If any item data corresponds to the standard data, there is no need to re-enter it. Each data item is in a separate line, beginning with the secondary keyword in the first four positions of the line. Data requiring units (e.g. pressure, heat duty, dimensions, etc.) have the units defined in positions 5 to 12 (starting at position 5). If the units correspond to the default defined by BRITISH or SI at the beginning of the run, then the units can be omitted by leaving the positions 5 to 12 blank.

Errors can be corrected by re-entering the data item. The sequence is not fixed.

After all input data for the item have been defined, the keyword

OK

indicates that this data input session is complete, and the program expects a new item to be entered. When all data items have been entered, the keyword

END

closes the items and instructs the program to proceed to the estimating phase. Between the ITEM and OK keywords only secondary keywords are accepted. At this point the user is asked whether he requires the file to be saved with the input data. The cost estimate is presented in a number of tables.

During the specification of each item, it is possible to enter equipment cost data (FOB) which override the estimation procedure. This enables equipment for which separate costs have been obtained (e.g. from offers from equipment manufacturers) to be included in the estimate. Such data make all the equipment input data otherwise required redundant. To correctly allow for inflation involving the FOB value, this must be the year to which this cost applied. Default cost units are 1000$. Other cost units can be entered exactly as with the units for the other equipment data items — i.e. in positions 5 to 12, starting at position 5.

e.g. FOB 1000$ 24.3, 1984

The keyword SHOW, when used between ITEM and OK keywords displays the present information for that equipment item. It can therefore be used to check that the input data is correct before closing with OK and moving on to the next item.

*Note*: The keyword structure of the input data enables simple identification and modification of the input files, and simple correction during interactive data input. It also enables the HELP function to be called at any time by giving the keyword followed by HELP

e.g. YEARHELP

## A4—4.2 Updating the data files with CAPCOM

Each equipment class has its own data file (.D) used by CAPCOS to produce the cost estimate. These D files are in a special form which has been generated by running the CAPCOM program. CAPCOM uses an input file containing the cost information in a form understandable to the user, to generate the data file that is understandable to CAPCOS.

When run interactively, CAPCOM reads an editor-prepared input file (the name is optional) and prepares one file for CAPCOS (again the name is optional).

The CAPCOS file names are optional but must not exceed 5 characters.

Note that a CAPCOM file becomes accessable for CAPCOS only if it is contained in the ITEMS file.

In batch mode, a batch file is used (name optional) consisting of a list of input followed by output file names. CAPCOM then proceeds to use the input file named in the list to create the output file named as the following file in the list. Hence batch mode enables a series of these files to be handled in one run.

Runs are completed by returning CR when asked for the next input file name.

### A4—4.2.1 Data file structure

The data file or CAPCOM provides all the cost information for the whole item class. This involves sub-division of the class, algebraic cost correlations with significant equipment parameters, pressure and material allowances — which are often in tabular form – and the division of the indirect costs — equally segmented for different cost levels. Finally, texts need to be stored for equipment names, and helps, screens and units must be included.

CAPCOM achieves all this with its special language, based on keywords. This is completely flexible, thus allowing any cost estimation method to be included in CAPCOS even though its presentation may differ completely from the original Guthrie presentation.

The following section details each of these keywords. This language is, because of its flexibility, not easy to understand without considerable study. However, CAPCOM input files are available (.DSR), which enable a clearer picture to be obtained. Updates to Guthrie's information can be achieved by minor changes to the .DSR files. It is necessary to have a thorough understanding of the CAPCOM input language only for a complete restructuring of the costing information.

The principle of the language is to read all data into a known location $X(i)$ each time defining $(i)$. Calculate keywords allow new variables $X(ii)$ to be created as a function of the $X$ variables read in as data. Logical decisions are also possible. FOB, BC and field costs can be calculated and their $X$ locations specified.

All the data files have the same structure, and use the same 12 primary keywords. The sequence of these primary keywords is arbitrary.

The first line of the file is the name of the equipment (less than 72 characters) e.g. "direct fired heater".

The primary keywords are the following:

| | | |
|---|---|---|
| INPUT | FOR | ENDIF |
| CODE | BC | MLI |
| BASE | FIELD | |
| CONSTANT | IND | |
| CALCULATE | IF | |

The second line of the file contains the base year of the equipment cost. The keywords are identified by their first four characters.

The use of the primary keywords is as follows:

1 INPUT XXXX i, type, units, min, max, default, n

enables the definition of the significant equipment parameters, their units, range of validity and the typical default value. Also the appropriate texts for the help file are entered.

XXXX is a secondary keyword (created at this point by the user) defining the input data to follow (e.g. AHD for absorbed heat duty) secondary keywords consisting of 4 characters or fewer. The secondary keywords and the number of them for the different equipment items can be different. All numeric data begin after the 12th character.

*i* points out an x-vector element, where the data will be placed [X(I)] between 1 and 50

*type* is an integer, indicating the type of units in which the data are given. The type codes are:

0 indicates that it is an input without units (i.e. dimension)
1 indicates length
2 indicates area
3 indicates volume
4 indicates mass
5 indicates pressure
6 indicates temperature
7 indicates work (or enthalpy)
8 indicates power (e.g. KW)
9 indicates volumetric rate (e.g. $m^3/s$)
10 indicates mass rate (e.g. kg/h)
11 indicates time
12 indicates value

*units* is an integer between 1 and 10 which indicates a measurement unit used inside the cost estimation routine (internal units). Input data are converted to these units before the calculation starts.

Type and units as a pair of integer numbers are indicated in the following table. This table has 12 rows, one for each 'type' and 10 columns corresponding to the different internal units.

The contents of the table are the names of the measurement units

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| m | ft | cm | in | yd | | | | | |
| $m^2$ | $ft^2$ | $cm^2$ | $in^2$ | $yd^2$ | | | | | |
| $m^3$ | $ft^3$ | $cm^3$ | $in^3$ | $yd^3$ | | | | | |
| kg | lb | g | t | tons | | | | | |
| Pa | psi | bar | atm | at | torr | | | | |
| K | F | C | R | | | | | | |
| kJ | MBtu | erg | kWh | kcal | cal | hph | 1atm | mkp | |
| kW | MBtu/h | erg/s | W | kcal/s | cal/s | hp | Btu/s | kcal/h | mkp/s |
| $m^3/h$ | gal/min | l/h | $m^3/min$ | gal/h | $ft^3/min$ | $ft^3/h$ | | | |
| kg/h | lb/h | tons/day | tons/h | | | | | | |
| h | min | day | month | | | | | | |
| $ | 10$ | 100$ | 1000$ | 10,000$ | 100,000$ | | | | |

*min*, *max* are lower and upper bounds for the data specified by the secondary keyword.

*default* default value for the data specified by the secondary keyword

$n$ indicates that $n+1$ help lines of text will follow the data input lines.

2. CODE $i, j, n$

enables the program to assign one of a number of values for a specific parameter, depending on the integer value or code of another parameter

$i$ — indicates that the result will be placed in $x(i)$

$j$ — indicates that the evaluation of $x(i)$ is based on the integer value or code stored in $x(j)$

$n$ — number of cases (i.e. number of integer values of $(j)$)

the following lines give the $n$ parameter values

$$(V(1), \quad 1=1, 2, \ldots n).$$

As a result of the evaluation of the line the following lines give the $n$ parameter values:

$$x(1)=v(x(1))$$

3. CASE $i, j, n$

enables the program to assign one of a number of values for a parameter, depending on the interval in which the first parameter lies.

$i$ — indicates that the result will be placed in $x(i)$

$j$ — indicates that the evaluation of $x(i)$ is based on the value stored in $x(j)$

$n$ — number of intervals

the following lines give the $n$ parameter values

This keyword is always followed by $(2 \times n+1)$ numbers on the next lines. The first $n+1$ numbers are boundary values of the intervals $(h(i), i=0, \ldots, n)$. The remaining $n$ numbers are the values associated with each of the intervals $(v(i))$. As a result of the evaluation of the line beginning with the keyword CASE

$$x(i)=v(i) \qquad \text{if} \qquad h(i-1)=x(j)\, h(i)$$

4. CONSTANT $i$, value

enables a numerical constant to be entered into the correlations

$i$ indicates that "value" should be placed into $x(i)$

5. CALCULATE $i$

This keyword indicates that a formula will follow to calculate $x(i)$. The formula is interpreted according to the Polish notation, and should not exceed 72 characters. To explain this construction let us define a stack pointer, *sp*. At the beginning of the evaluation $sp=1$. If we write $xj$ in the formula, the value of $xj$ will be placed on the stack, and the stack pointer increased by 1. (It must be less than 20.)

The basic operations and their algebraic rules are the following:

* $s(sp-2)=s(sp-1)*(sp-2)$ : stack pointer moves $sp=sp-1$
\+ $s(sp-2)=s(sp-1)+(sp-2)$ : stack pointer moves $sp=sp-1$
\- $s(sp-2)=s(sp-1)-(sp-2)$ : stack pointer moves $sp=sp-1$
/ $s(sp-2)=s(sp-1)/(sp-2)$ : stack pointer moves $sp=sp-1$
E $s(sp-2)=s(sp-1)$ to power of $s(sp-2)$ : stack pointer moves $sp=sp-1$
L $s(sp-2)=$logarithm $s(sp-2)$ on the base of $s(sp-1)$ : stack pointer moves $sp=sp-1$

The value of $sp$ cannot exceed 20.

This concept is relatively simple and can be seen more clearly by inspecting some of the formulae in the DSR files.

6. FOB $i$

indicates that a formula will follow to calculate the FOB cost of the equipment item which will be placed in $x(i)$.

*Note*: If the user gives the FOB cost of the equipment item during the input procedure this value always overwrites the calculated FOB cost.

7. BC $i$

indicates that a formula will follow to calculate the BC cost of the equipment item which will be placed in $x(i)$.

*Note*: The definition of the formula to calculate the base cost of the equipment item is compulsory, and must be the inverse of the FOB formula.

The reasoning behind the CALCULATE-FOB-BC procedure is that factors are prepared under the CALCULATE keyword from which the BC and FOB can be determined. The formula following the FOB keyword calculates the FOB from these factors. The following BC keyword provides the formula to back-calculate the BC from the FOB. Hence, if the user overwrites the FOB cost, the BC is re-evaluated by this formula, since the BC is required in the evaluation of auxiliary materials, field labour and indirect costs.

8. FIELD $n$

enables the subdivision of field costs into the items listed below to be achieved. Since this sub-division is dependent on the magnitude of the base costs, the factors are given in a table with the columns of the table built up from the different base cost dollar ranges.

$n$ integer indicating the number of BC ranges

This line is followed by a table containing the detailed cost components for the specific equipment item. The table has 12 rows and $n$ columns:

row 1 : The maximum base cost of each BC range (in 1000 \$, $n$ values)

The following detailed cost elements expressed in % FOB are given for each of the BC ranges. The % values generally decrease with increasing BC dollar magnitude.

row 2: cost of piping
row 3: concrete
row 4: steel
row 5: instruments
row 6: electrical
row 7: insulation
row 8: paint
row 9: material erection
row 10: equipment setting
row 11: freight, insurance, taxes
row 12: indirect cost

*Note*: These costs are in % of FOB, but BC is required to determine the interval in which the factor is to be found.

9. MLI *i, j, k*

enables the auxiliary material, field labour and indirect costs to be located. For interpretation of these factors see Section A4—2 of this manual.

*i, j, k* are the integer pointers to the variables $x(i)$, $x(j)$, $x(k)$ in which the auxiliary material, field labour and indirect costs have been calculated and allocated. These are generally percentage values of FOB cost.

*Note*: Each equipment definition file should contain either a FIELD or an MLI keyword.

10. IF XX *i*, value

enables logical decisions to be made within the calculation of the variables

XX should be replaced by one of the following codes:

| | |
|---|---|
| EQ | is equal |
| NE | is not equal |
| LT | is less than |
| GT | is greater than |
| GE | is greater or equal |

*i* points to $x(i)$ which must be an input variable.

In evaluating the logical relationship, $x(i)$ is compared to 'value'. The following part of the equipment description file up to the next ENDIF keyword will be evaluated only if the relationship holds.

11. ENDIF

indicates the end of IF section. It must always be used in conjunction with IF.

Notes on generating/updating item description files:

(a) The keywords must be located at the beginning of the lines.

(b) Numbers can start anywhere after the 12th position of the line, and they can occupy spaces up to the 72nd position.

(c) Numbers must be separated by a space or a colon.

(d) Any line can be continued. *H* as the non-blank character of a line indicates that a continuation line will follow. The continuation line must not contain any keyword.

### A4—4.2.2 The ITEMS file

This file is used to tell CAPCOS which files are the actual description files. The file name 'ITEMS' must not be changed.

The first line of the ITEMS file contains the number of item description files '*n*'. The next *n* lines contain the filenames on their first 4 positions (therefore the item description file name must not exceed 4 characters), and the item name (5th—28th positions). The item name is optional. The possible item names are listed in interactive mode whenever ITEMHELP is entered.

To update the item description file we have to locate its name in the ITEMS file, then, to update the corresponding .DSR file and to compile it under the same name by CAPCOM.

In order to delete an item description file from CAPCOS we have to delete the corresponding line of the ITEMS file and to decrease the first number (the number of items descriptor).

Finally, a new item group can be added in the following way:

First set up the corresponding .DSR file using a normal editor, then translate it by CAPCOM into a file which has a name no longer than 4 characters, write it into the ITEMS file on a separate line (location) and increase the items descriptor number.

## A4—5 EXAMPLES

### A4—5.1 Example to demonstrate batch use

To use the program in the interactive modes (mode 2 and 4), there is no need for any special instructions, and therefore no need for any examples.

To use the program in batch mode, one has to create an input file with all the necessary input data.

This example demonstrates the input data requirements and input data file structure of the most frequently used equipment costing routines: the process equipment. The title is

Default test for process equipment

```
default test for process equipment
brit
item                    1, 1
process furnace
dty                     1
mty                     1
p       psi             100
ahd     mbtu/h          100
ok
item                    2, 1
direct fired heater
dty                     1
mty                     1
p       psi             100
ahd     mbtu/h          10
ok
item                    3, 1
heat exchanger
dty                     2
p       psi             100
a       ft2             1000
mty                     1
ok
item                    4, 1
double pipe
mty                     1
p       psi             100
a       ft2             10
ok
item                    5, 1
air cooler
a       ft2             15500
p       psi             100
l       ft              10
mty                     1
ok
item                    6, 1
vertical vessel
mty                     1
p       psi             40
h       ft              20
d       ft              4
ok
item                    7, 1
horizontal vessel
mty                     1
p       psi             40
h       ft              20
d       ft              4
ok
item                    8, 1
tray stack
dty                     1
mty                     1
sp      ft              1.66666667
d       ft              3
h       ft              30
dk
item                    9, 1
packing
typ                     1
mty                     1
siz     ft              .0322808399
v       ft3             1000
ok
item                    10, 1
lining
typ                     1
a       ft2             1000
th      ft              .0984251969
ok
```

The keyword >brit< indicates that the default values of all measurement units are the British ones.

Item 1, 1

indicates that a data set for item 1 (process furnace) will follow, and there is only one piece of that kind.
The user defined name of the equipment (for the sake of simplicity) is process furnace.

```
dty          1
```

indicates that the design type of the process furnace is process heater.

```
mty          1
```

indicates that the process heater is made of carbon steel.

```
p      psi       100
```

indicates that the nominal pressure of the furnace in psi units is 100.

```
ahd    mbtu/h    100
```

indicates that the absorbed heat duty in the specified heater is 100 mbtu/h.

```
ok
```

this statement indicates that the item specification is finished, a new equipment specification comes.

The interpretation of the remaining part of this example (with the help of the input specification tables in Section A4—3.1) is a straightforward task for anybody who is willing to learn the use of this program.

## A4—5.2 Example to demonstrate the use of the data description language

This example illustrates the creation of the item description files. If you would like to make changes in the cost modules, or to create new ones, this example helps to check your understanding of the syntactic rules explained in Section A4—4.
The example is the data description file for heat exchangers.

The file name is >heat exchanger<.

The base year is 1968.

The first input item is the design type (dty) which will be set to X(1), the measurement unit type is 0, the unit is 0, and the acceptable range of the input data is 1—4. The default value is 2, and 4+1 help lines should follow.

The next input data description comes after the help lines. The second input item is pressure (indicated with the mnemonic p), it should be set to X(5), measurement unit type is 5(pressure), and psi is specified as unit. The acceptable range is 10—1000 psi, and the default value is 100 psi. There is only one help line with the interpretation of the mnemonic p.

The next keyword, >case<, specifies 5 cases. The first pair of numbers 6, 5 indicates that the number in X(5) (pressure, in psi) will be evaluated, and the result will be stored in X(6). The next line contain $2 \times 5 + 1$ numbers. The first 5+1 are

the boundary values of the intervals within which the pressure data should fall. The remaining 5 numbers are the values associated with the intervals previously defined. These values are the possible values of X(6). The interpretation of the remaining two input parameter definitions is straightforward.

The keyword >calculate< indicates that the value of X(9) is calculated as a sum of X(03) and X(08).

The keyword >code< indicates that the value of the code, X(10), will be defined according to the value of X(9), and there are 45 code values.

The next five lines contain these 45 values.

The keyword >constant< indicates, that a constant value, 0.629 will be set to X(20).

The next new keyword is >FOB<. This line indicates that the FOB cost of this equipment will be placed to X(12). The next line indicates that the calculation of the FOB cost is as follows:

$$X(12) = (X(02) + X(06)) \times X(10) \times X(11)$$

The line with the keyword >bc< can be interpreted in the same way.

The line with the keyword >field< specifies 5 possible intervals for the base cost dollar magnitude. To interpret the numbers in the rows of the matrix see Section A4—4.

```
heat exchanger
1968
input   dty          1, 0, 0, 1, 4, 2, 4
heat exchanger design type
1 — kettle or reboiler
2 — floating head
3 — u tube
4 — fixed tube sheet
input   p            5, 5, 2, 10, 1000, 100, 0
pressure
case                 6, 5, 5
0, 150, 300, 400, 800, 1500, 0, 0.1, 0.25, 0.52, 0.55
code                 2, 1, 4
1.35, 1.00, 0.85, 0.80
input   a            7, 2, 2, 100, 10000, 1000, 0
calculated surface area
case                 8, 7, 5
0, 100, 500, 1000, 5000, 10000, 0, 9, 18, 27, 36
input   mty          3, 0, 0, 1, 9, 1, 9
shell/tube material
1 — carbon steel          / carbon steel
2 — carbon steel          / brass
```

```
3 — carbon steel          / monel
4 — carbon steel          / stainless steel
5 — stainless steel       / stainless steel
6 — carbon steel          / monel
7 — monel                 / monel
8 — carbon steel          / titanium
9 — titanium              / titanium
calculate                 9
x03  x08  +  =
code                      10, 9, 45
1.00, 1.05, 1.60, 1.54, 2.50, 2.00, 3.20, 4.10, 10.28 *
1.00, 1.10, 1.75, 1.78, 3.10, 2.30, 3.50, 5.20, 10.60 *
1.00, 1.15, 1.82, 2.25, 3.26, 2.50, 3.65, 6.15, 10.75 *
1.00, 1.30, 2.15, 2.81, 3.75, 3.10, 4.25, 8.95, 13.05 *
1.00, 1.52, 2.50, 3.52, 4.50, 3.75, 4.95, 11.10, 16.60
constant                  20, 0.629
constant                  21, 0.1102
calculate                 11
x21  x20  x07  e  *  =
fob                       12
x11  x10  x02  x06  +  *  *  =
bc                        13
x10  x02  x06  +  x12  /  /  =
field                     5
 200,     400,     600,     800,     1000
45.6,    45.1,    44.7,    44.4,    44.3
 5.1,     5.0,     5.0,     5.0,     5.0
 3.1,     3.0,     3.0,     3.0,     3.0
10.2,    10.1,    10.0,     9.9,     9.8
 2.0,     2.0,     2.0,     2.0,     2.0
 4.9,     4.8,     4.7,     4.7,     4.7
 0.5,     0.5,     0.5,     0.5,     0.5
55.4,    54.7,    54.2,    53.9,    53.8
 7.6,     6.5,     5.9,     5.5,     5.2
 8.0,     8.0,     8.0,     8.0,     8.0
86.7,    78.8,    75.9,    75.5,    73.0
end
```

## A4—6 NOTES ON PROGRAM INSTALLATION

The original program disk contains two FORTRAN source programs with their executable compilations and a number of data files.
The Microsoft FORTRAN 77 V3.10 compiler has been used.

Programs:

1. CAPCOS.FOR (compiled: CAPCOS.EXE)
Main program which organizes the capital cost estimation procedure.
2. CAPCOMP.FOR (compiled: CAPCOMP.EXE)
Compiler program for advanced users who intend to introduce new equipment definitions or to alter existing ones.

Data files:

1. ITEMS file
This file contains as its first record the number of different equipment description files (n). n succeeding records contain the filenames (4 characters) and the corresponding item group name (24 characters). The ITEMS file name must not be modified.
2. *. DSR files
These files contain the item group estimation descriptions written in CAPCOS language (see above).
3. ??. D files
These files contain the item group estimation method description in 'compiled' form, e. g. in a form produced by program CAPCOMP. Therefore ??. D files occur in the ITEMS file and not *. DSR files.

It is recommended to copy both EXE files and all ??.D files together with ITEMS file to a hard disk, if possible, since CAPCOS often uses disk files.

To load and run these programs simply enter (as usual in MS—DOS) CAPCOS or CAPCOMP, respectively.

# APPENDIX A5

## Appendix to Chapter 5

### TACS

A Teaching Aid for Control Studies on IBM PC
Program Version 3

### ETACS

An Extended Teaching Aid for Control Studies on IBM PC
Program Version 1

MANUAL

June 1987

Manual authors: S. FARKAS and P. VAJDA

Laboratory for Chemical Cybernetics
Eötvös Loránd University

Budapest
Múzeum krt. 6—8.
Hungary
H-1055

# CONTENTS


A5—1 INTRODUCTION 569
A5—1.1 Main features 569
A5—1.2 Dynamic units: principles 570
A5—1.3 Available dynamic units in TACS 570
A5—1.4 Additional units in ETACS 575
A5—1.5 Preparation of new model subroutines 581

A5—2 HOW TO USE THE PROGRAM 584
A5—2.1 Input-output options and general remarks 584
A5—2.2 Summary of input data 585
A5—2.3 Additional options in ETACS 585
A5—2.4 Example 585

A5—3 PROGRAM IMPLEMENTATION 596

A5—4 FILES ON THE DISKETTE 598

A5—5 REFERENCES 598

# A5—1 INTRODUCTION

TACS is general program for simulating the dynamic behaviour of control systems. Written mainly for undergraduate courses in process control, this program will enable students to understand the most common methods of control (open loop, feedback, feedforward); to study the 'standard' controller actions of on-off, proportional, reset, and derivative as well as selection of appropriate controller settings. Plots constructed by the program clearly show the dynamic behaviour of uncontrolled and controlled processes.

Simple program usage ensures a friendly environment for the students. A number of basic dynamic elements, enabling him to construct control loops can be defined simply by selecting the appropriate type indices, variable names and initial values. The elements can be organized into a control network of arbitrary complexity. Simulating the dynamic behaviour of the network, any output variables can be tabulated and plotted. There are possibilities for temporary inclusion of new units.

Together with any textbook written for undergraduate courses, the TACS program is also suitable for self-study and will bring physical meaning to the mathematical expressions defining control systems.

ETACS is an extended version of TACS. It provides subroutines to study realistic models and enables further analysis of dynamic behaviour. It needs, however, double the computer memory.

## A5—1.1 Main features

*TACS:*

A number of simple dynamic elements (e.g., integrator, differentiator, first- and second-order units) as well as further elements enabling one to construct control loops can be defined by selecting the appropriate type index, variable names and values.

The elements can be organized into a control network of arbitrary complexity.

Difference or differential equations produced by the network are solved within the specified time interval, with specified time step and input functions.

Any variable can be tabulated or plotted.

Further subroutines defining new units can be included in the system.

*Additional features of ETACS:*

More realistic dynamic models are provided for some typical chemical engineering processes.

Identification units can be used to approximate an input-output pair by a discrete time linear system.

Frequency response plots can be constructed for any input-output pair.

Phase diagrams and spatial distributions can be plotted for appropriate process units.

## A5—1.2 Dynamic units: principles

TACS and ETACS are based on the state-space approach. Each dynamic unit is described by the following vectors:

$U$ — vector of input variables; the components of $U$ are time and selected output variables of other units.

**X** — matrix of state variables; for linear models, presently in the system, zero initial state values are automatically assumed, except for some additional units in ETACS where the program requires their specification. If required, values of state variables are preserved after completed runs, thereby enabling one to continue the simulation.

$Y$ — vector of output variables; any output variable can be connected to another unit as its input variable, and it can be tabulated or plotted.

$P$ — vector of parameters that specify the unit

Each dynamic unit is implemented as a subroutine. At actual time $T$ of simulation the subroutine acts as follows: from $X(T)$ and $U(T)$ it computes $X(T+DT)$, and $Y(T+DT)$, where $DT$ denotes the time step of the simulation. In each subroutine these calculations are carried out independently, without formulating the whole set of difference or differential equations corresponding to the network.

Therefore, the program directly describes the dynamic behaviour of a discrete system with time step $DT$. However, by selecting a small step size, continuous systems can also be approximated adequately for educational purposes. Algorithms for linear units are based on analytic solutions, thereby reducing run time considerably.

The same approach is used in some units of ETACS that do not model specific processes but perform operations. For example, the input of an identification unit is the input–output pair to be identified, its state variables are the current parameter estimates, its output contains the final parameter vector of the linear model fitted to the input–output pair, whereas the parameters of the actual identification unit specify the identification procedure itself (e.g. number of iterations and termination conditions).

## A5—1.3 Available dynamic units in TACS

*Table A5—1* lists the available dynamic units and defines their input, state, and output variables and parameters. No knowledge of this table is required, however, since, when selecting a unit its specification is supported by approximate questions appearing on the display.

Table A5—1. Available dynamic units

| Unit name | Type index | Inputs | States (1) | Output | Parameters | Error code |
|---|---|---|---|---|---|---|
| Step function generator | 1 | 1—time | 1—not used | step function | 2—step time<br>—amplitude | — |
| Sine function generator | 2 | 1—time | 1—not used | sine function | 3—amplitude<br>—frequency (rad/s)<br>—initial phase (rad) | — |
| Pulse function generator | 3 | 1—time | 1—not used | pulse function | 2—pulse time<br>—amplitude | — |
| Ramp function generator | 4 | 1—time | 1—not used | ramp function | 2—initial time<br>—slope | — |
| 1-st user-written unit (2) | 5 | NI1— | NR1*NC1 | NO1— | NP1— | 50–59 |
| 2-nd user-written unit (2) | 6 | NI2— | NR2*NC2 | NO2— | NP2— | 60–69 |
| 3-rd user-written unit (2) | 7 | NI3— | NR3*NC3 | NO3— | NP3— | 70–79 |
| 4-th user-written unit (2) | 8 | NI4— | NR4*NC4 | NO4— | NP4— | 80–89 |
| Time delay | 9 | 2—time<br>—signal to be delayed | Q—an inverse sequence of input values from the last Q steps | input function delayed by Q time steps | 1—delay in time steps (Q) | 90 |
| Gain | 10 | 2—time<br>—input signal | 1—not used | response function | 1—gain coefficient | — |
| Ideal integrator | 11 | 2—time<br>—signal to be integrated | 2—signal in the previous time step<br>—integral in the previous time step | integral of the input signal computed by the trapezoidal rule | 1—time step size | — |

Table A5—1 (continued)

| Unit name | Type index | Inputs | States (1) | Output | Parameters | Error code |
|---|---|---|---|---|---|---|
| Second-order process with unit gain (3) | 12 | 2—time<br>—input signal | 2—output in the previous time step<br>—derivative in the previous time step | response function | 2—time constant<br>—damping factor | — |
| Ideal differentiator | 13 | 2—time<br>—signal to be differentiated | 2—input in the previous time step<br>—arbitrary | input signal differentiated by the two-point rule | 1—time step size | — |
| Weighted linear (4) smoothing | 14 | 2—time<br>—signal to be filtered | K—previous input values | filtered signal | 1+K—order (K)<br>—weighting coefficients | 140 |
| First-order process with unit gain (5) | 15 | 2—time<br>—input signal | 1—output in the previous time step | response function | 1—time constant | — |
| Addition of two signals | 16 | 3—time<br>—signal1<br>—signal2 | 1—not used | sum of two input signals | 1—not used | — |
| Sign inversion | 17 | 2—time<br>—input signal | 1—not used | input*(−1) | 1—not used | — |
| Multiplication of two signals | 18 | 3—time<br>—signal1<br>—signal2 | 1—not used | product of two input signals | 1—not used | — |
| On-off controller | 19 | 2—time<br>—signal defining on-off | 1—not used | response (+A or −A) | 1—amplitude A | — |
| Gain with constraints (6) | 20 | 2—time<br>—input signal | 1—not used | response function | 2—lower bound on the output<br>—upper bound on the output | — |

| | | | | | | |
|---|---|---|---|---|---|---|
| Ideal integral (1) controller (7) | 21 | 2—time<br>—input signal | 2—input in the previous step<br>—integral in the previous step | response function computed by the trapezoidal rule | 2—gain factor<br>—integration time constant | — |
| Ideal proportional-integral (PI) controller (3) | 22 | 2—time<br>—input signal | 2—input in the previous step<br>—integral in the previous step | response function | 2—gain coefficient<br>—integration time constant | — |
| Proportional-integral-differential controller (9, 10) | 23 | 2—time<br>—input signal | 2—input in the previous step<br>—integral in the previous step | response function | 3—gain coefficient<br>—integration time constant<br>—derivation time constant | — |

## *Notes to Table A5—1*

1. To define a dynamic unit, you should select its type index, specify the name of its output variable and values for the parameters. For the sake of simplicity zero initial states are automatically assumed.

Non-zero state behaviour should be simulated by creating appropriate inputs with unit types 1 to 4. State variables are preserved after runs and hence you can continue the simulation. To repeat it from the original zero initial states you should select the modified option and redefine the units.

2. Type indices 5–8 are reserved for user-written subroutines. Four new models can be built into the program by inserting the appropriate FORTRAN lines into the dummy subroutines USER1–USER4 supplied with the TACS coding. Preparation of such models is described in Section A5—1.4. In *Table A5—1* the variables NI, NR × NC, NO and NP represent the actual number of inputs, states, outputs and parameters in user-written models.

3. The second-order process is given by the transfer function

$$G(s)=\frac{1}{T^2s^2+2\xi Ts+1}$$

where $T$ and $\xi$ represent the time constant and the damping factor, respectively. The corresponding differential equation

$$T^2\ddot{y}+2\xi T\dot{y}+y=u$$

is converted to the state-space representation

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix}=\begin{bmatrix}0 & 1\\ -\frac{1}{T^2} & -\frac{2}{T}\xi\end{bmatrix}\cdot\begin{bmatrix}x_1\\ x_2\end{bmatrix}+\begin{bmatrix}0\\ \frac{1}{T^2}\end{bmatrix}u$$

$$y=x_1$$

where $x_2=\dot{y}$. From this model a discrete-time description is constructed assuming $u=$const. within the time step $\Delta t$.

4.

$$y(t)=\sum_{i=0}^{K-1}\lambda_i u(t-i\cdot\Delta t)$$

5. The first-order process is given by the transfer function

$$G(s)=\frac{1}{Ts+1}$$

converted to the discrete-time state-space representation

$$x(t+\Delta t)=e^{-\Delta t/T}x(t)+(1-e^{-\Delta t/T})u(t)$$

$$y(t)=x(t)$$

6. To simulate the behaviour of measuring devices

$$y(t)=\begin{cases}LB & \text{if } u(t)\leq LB\\ u(t) & \text{if } LB\leq u(t)\leq UB\\ UB & \text{if } u(t)\geq UB.\end{cases}$$

7. To retain similarity with other controllers, the *I*-controller is given by the transfer function

$$G(s)=\frac{K}{T_I \cdot s}$$

8. The transfer function of the ideal *PI*-controller is

$$G(s)=K\left(1+\frac{1}{T_I \cdot s}\right)$$

9. The *PID* controller is described by

$$G(s)=K\frac{T_I \cdot s+1}{T_I \cdot s}\frac{T_D \cdot s+1}{\frac{T_D \cdot s}{\alpha}+1}$$

where $\alpha=10$, for details see [3].

10. The set point is zero for each controller. This situation can always be attained by using a deviation variable for controller input.

*Remark*: Some units return the error code $10*I$, where $I$ is the type index, if no space is available to place all the variables.

## A5—1.4 Additional units in ETACS

Table A5—2. Additional units in ETACS

| Function | Type index | Title | Error codes |
|---|---|---|---|
| Process models | 24 | steam heater | — |
| | 25 | steam boiler | 250 |
| | 26 | stirred tank reactor with first-order process | 260 |
| | 27 | shell and tube heat exchanger | 270 |
| | 28 | valve model for incompressible fluid | — |
| | 29 | valve model for compressible fluid | 290 |
| Discrete-time identification | 30 | off-line identification by least squares | 300 |
| | 31 | off-line identification by instrumental variables | 310—311 |
| | 32 | on-line identification by least squares | — |
| | 33 | on-line identification by instrumental variables | — |
| Frequency plot | 34 | Nyquist plot | — |
| Auxiliary | 35 | pseudo-random binary signal generator | — |
| | 36 | discrete-time linear SISO unit | — |

In addition to the basic units in *Table A5—1*, ETACS includes the units listed in *Table A5—2*. Since these units are more complex, their functions are described separately.

*Steam heater (type index 24)*

A steam-heated stirred tank with constant fluid holdup is modelled by the equation

$$Mc_p \frac{dT}{dt} = Wc_p(T_{in} - T) + W_s \lambda$$

where the variables and their typical values are as follows

Input variables: 4

$t$ — time, s

$W$ — fluid inlet mass flow rate, 125 kg s$^{-1}$,

$T_{in}$ — inlet temperature, 293.2 K,

$W_S$ — steam mass flow rate, 14 kg s$^{-1}$,

State variable: 1

$T$ — fluid temperature, 353.577 K,

Output variable: 1

$T$ — fluid temperature, 353.577 K,

Parameters: 3

$M$ — fluid mass holdup, 6395.7 kg,

$c_p$ — specific heat of the fluid, 4187 J kg$^{-1}$ · K$^{-1}$,

$\lambda$ — latent heat of the fluid, 2 257 200 J kg$^{-1}$.

*Steam boiler (type index 25)*

The unit is shown in *Fig. A5—1* and is described by the following equations.

$$\frac{dM}{dt} = F - V$$

$$\frac{dm}{dt} = V - v_e$$

$$\frac{d}{dt}(Mc_pT) = Fc_pT_{in} + q - V(c_pT + \lambda)$$

$$pV = \frac{m}{\text{Mol}} RT$$

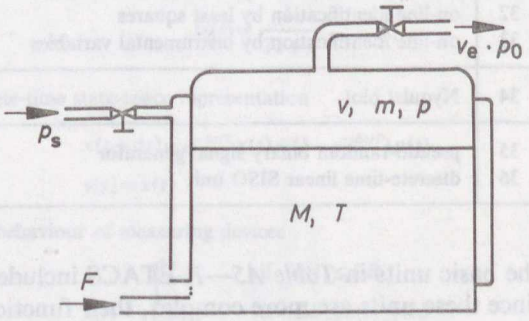



Fig. A5—1. Steam boiler

$$V = V_0 - \frac{M}{\varrho}$$

$$T = \frac{c_2}{\ln p - c_1} - c_3$$

$$v_e = k_v \sqrt{p(p - p_0)}$$

$$q = \alpha(T - T)\left(\frac{2M}{\varrho}\sqrt{\frac{3.1415}{A}} + A\right)$$

$$T = \frac{c_2^s}{\ln p_s - c_1^s} - c_3^s$$

We use a system of differential equations derived from these equations. The variables and their typical values are as follows:

Input variables:

$t$ — time, s
$F$ — mass flow rate of the fluid, 0.1 kg $s^{-1}$
$T_{in}$ — fluid inlet temperature, 320 K
$p_s$ — steam pressure, 2.0e+7 kg $m^{-1}$ $s^{-2}$
$p_0$ — vapour load pressure, 1.5e+5 kg $m^{-1}$ $s^{-2}$

State variables:

$M$ — fluid mass holdup, kg
$m$ — vapour mass holdup, kg
$T$ — vapour temperature, K
$p$ — vapour pressure, kg $m^{-1}$ $s^{-2}$
$V$ — volume occupied by vapour, $m^3$

Output variables:

$T$ — vapour temperature, 395.03074 K
$p$ — vapour pressure, 165 138.76 kg $m^{-1}$ $s^{-2}$
$v_e$ — outlet vapour flow rate, 0.1 kg $s^{-1}$
$h$ — $M/(A\varrho)$, fluid level, 0.801282 m

Parameters:

$c$ — specific heat of the fluid, 3800 J $kg^{-1}$ $K^{-1}$
$\alpha$ — heat transfer coefficient, 500 J $s^{-1}$ $m^{-2}$ $K^{-1}$
$\lambda$ — latent heat of the fluid, 1.2e+6 J $kg^{-1}$
Mol — molar mass, 0.046 kg $mol^{-1}$
$V$ — tank volume, 1.57 $m^3$
$\varrho$ — fluid mass density, 800 kg $m^{-3}$
$A$ — tank cross-sectional area, 0.78 $m^2$
$k_v$ — valve constant, 2.0e−6 m s
$c_1$, $c_2$, $c_3$, $c_1^s$, $c_2^s$, $c_3^s$ — Antoine coefficients, 23.804657, −3803.98, −41.679993, 23.19635, −3816.4399, −46.12999;
relative error tolerance for the differential equation solver, 0.02;
maximum number of step in the ODE solver, 200;

Details of the model can be found in [4].

*Stirred tank reactor with a first-order process (type index 26)*

The process is described by the equations

$$\frac{dx}{dt} = -\frac{q}{V}x + \frac{r(x,T)}{a_0}$$

$$\frac{dT}{dt} = -\frac{q}{V}(T - T_{in}) + \frac{-H}{c}r(x,T) - \frac{U}{V\varrho c}$$

$$r(x,T) = A\exp\left(-\frac{E}{RT}\right)a_0(1-x)$$

$$U = Qc_w(T - T_w)$$

where the variables and their typical values are as follows

Input variables:

- $t$ — time, s
- $a_0$ — reactant concentration in feed, 800 mol $m^{-3}$
- $q$ — flow rate of reacting mixture, 0.0001 $m^3$ $s^{-1}$
- $T_{in}$ — feed temperature, 303.2 K,
- $Q$ — cooling water mass flow rate, 0.02 kg $s^{-1}$
- $T_w$ — cooling water feed temperature, 293.2 K

State variables:

- $x$ — conversion of the reactant 0.425
- $T$ — temperature, 345 K

Output variables:

- $a_0(1-x)$ — outlet reactant concentration, 460 mol $m^{-3}$
- $T$ — temperature, 345 K

Parameters:

- $c$ — heat capacity per unit mass, 4187 J $kg^{-1}$ $K^{-1}$
- $\Delta H$— heat of reaction, $-5.75e+5$ J $mol^{-1}$
- $V$ — reactor volume, 0.0005 $m^3$,
- $A$ — pre-exponential factor, $6.853e+10$ $s^{-1}$
- $E$ — energy of activitation, 76491 J $mol^{-1}$
- $\varrho$ — mass density of the mixture, 1000 kJ $m^{-3}$

relative error tolerance for the differential equation solver, 0.02,
maximum number of step in the ODE solver, 200.

*Shell and tube exchanger (type index 26)*

The shell, the parallel-flow tube and the counter-flow tube are described by the partial differential equations

$$C_3\frac{\partial T_s}{\partial t} + W_s C_s\frac{\partial T_s}{\partial x} + \frac{Sh}{2}(T_s - T_{t1}) + \frac{Sh}{2}(T_s - T_{t2}) = 0$$

$$\frac{C_t}{2}\frac{\partial T_{t1}}{\partial t} + W_t C_t\frac{\partial T_{t1}}{\partial x} + \frac{Sh}{2}(T_{t1} - T_s) = 0$$

$$\frac{C_t}{2}\frac{\partial T_{t2}}{\partial t} - W_t C_t\frac{\partial T_{t2}}{\partial x} + \frac{Sh}{2}(T_{t2} - T_s) = 0$$

respectively, and the boundary and initial conditions are

$$T_s(0, t) = T_{si}(t); \qquad T_{t2}(L, t) = T_{ti}(t); \qquad T_{t1}(0, t) = T_{t2}(0, t);$$

$$T_{t1}(x, 0) = T_{t2}(x, 0) = T_s(x, 0) = T_{si}(x)$$

see *Fig. A5—2* and [3] for details.

The derivatives with respect to the spatial coordinate $x$ are approximated by finite differences on a grid and the resulting set of ordinary differential equations is solved. The variables specifying the unit are as follows.

Input variables:

$t$ — time, s,
$T_{si}(t)$ — inlet shell side temperature, 292.7 K
$T_{ti}(t)$ — inlet tube side temperature, 338.3 K
$W_s$ — shell side mass flow rate, 2.54 kg $s^{-1}$
$W_t$ — tube side mass flow rate, 5.24 kg $s^{-1}$

State variables: $3*N$

temperatures at the grid points in the shell, in the co-current and the counter-current passes, resp.

Output variables:

$T_s(L, t)$ — outlet shell side temperature, 292.7 K
$T_{t2}(0, t)$ — outlet tube side temperature, 292.7 K

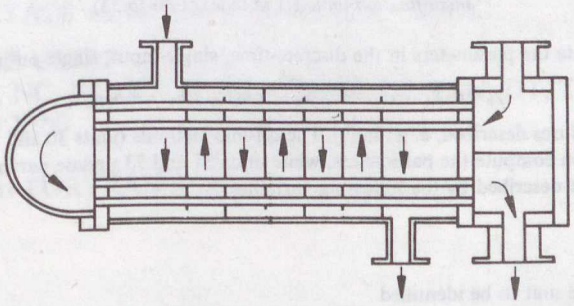


Fig. A5—2. Shell and tube exchanger

Parameters:

$L$ — length 4.1 m;
$D_s$ — inside shell diameter, 0.254 m
$N$ — number of tubes, 52
$D_{t,i}$ — inside tube diameter, 0.017 m
$D_{t,0}$ — outside tube diameter, 0.0191 m
$S$ — heat transfer surface area per unit length 3.1202 m
$h$ — heat transfer coefficient, 900 J $m^{-1}$ $s^{-2}$ $K^{-1}$
$c_s$ — specific heat of shell side fluid, 3162 J $kg^{-1}$ $K^{-1}$
$c_t$ — specific heat of tube side fluid, 4187 J $kg^{-1}$ $K^{-1}$

$\varrho_s$ — mass density of shell side fluid, 1301.62 kg m$^{-3}$
$\varrho_t$ — mass density of tube side fluid, 1000 kg m$^{-3}$

relative error tolerance for the differential equation solver, 0.1,
maximum number of steps in the ODE solver, 200,

From these parameters the unit computes the specific heats $c_s$ and $c_t$ per unit length where

$$C_s=\pi(D_s^2-D_{t,0}^2N)\frac{\varrho_s c_s}{4} \qquad C_t=\pi D_{t,i}^2 N\frac{\varrho_t c_s}{4}$$

*Remark—Error codes*:
As for the units in TACS the error code is $10*I$ if insufficient space is available, where $I$ is the type index. In addition, units 24 to 27 return the error code $10*I+1$ if the maximum allowed number of steps was insufficient to solve the differential equations over the actual step of the simulation.

*Valve models* (*type index* 28 *and* 29)

Depending on the control signal $u$ the flow rate $Q$ is calculated by the nonlinear characteristics

$$Q=kQ_{max}\,[1+(0.1|u|+0.3u^2+0.6|u|^3)\,\mathrm{sign}\,(u)]$$

where $k=1$ for incompressible fluids (type index: 28) and $k=\sqrt{t(p_1-p_2)}$ for compressible fluids (type index: 29), where $p_1$ and $p_2$ denote the pressure before and after the valve, respectively. The only parameter of the unit is the maximum flow rate $Q$. For unit 29 the error code is 291 if $p_1<p_2$.

*Identification units* (*type indexes* 30 *to* 33)

The units estimate the parameters in the discrete-time, single-input, single-output linear models

$$y_n+a_1y_{n-1}+\ldots+a_ky_{n-k}=b_1u_{n-1}+\ldots+b_ku_{n-k}$$

by standard procedures described, e. g., in [1]. The off-line methods (units 30 and 32) sample the data in $N$ cycles and then compute the parameters, while units 31 and 33 update parameter values in each cycle. The units are described by the following variables.

Input variables: 3

$t$ — time, s
$u_i$ — input of the unit to be identified
$y_i$ — output of the unit to be identified

State variables: auxiliary quantities, dependly on the method
Output variables: $2K$ vector of parameter estimates

$$(a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_k)^T$$

*Parameters*: Their number depends on the method. In each case specify the order $k$ of the model. For off-line algorithms the $k$ is followed by $N$, the number of sample points to be processed, whereas the on-line methods require a real number to form a diagonal matrix that is considered as the initial estimate of the inverse covariance matrix of the parameter vector. These latter methods also demand initial estimates for the parameters.

Units 30 and 31 have the following error codes. The code is $10*I+1$ if the matrix to be inverted is singular. Unit 31 returns the value 311 if the allowed number of iterations is insufficient to achieve the required accuracy.

*Frequency response plots (type index* 34)

The unit samples $2**K$ input-output pairs from the network and constructs a Nyquist plot by fast Fourier transformation. The three parameters are

K — defines the number of input-output pairs in $2**K$ form;
IWS — number of time steps to take before sampling input;
OWS — delay (in time steps) of output sampling with respect to input sampling.

*Pseudo-random binary signal generator (type index* 35)

The output is $+1$ or $-1$, and the generated sequence used as an input signal when identifying a network or constructing its frequency diagrams.

Discrete time single-input, single-output linear unit (type index: 36)

The unit computes the output $y_n$ of the linear difference equation also considered in identification units and hence can be used to generate input-output pairs to study these latter procedures.

## A5—1.5 Preparation of new model subroutines

The function of each unit subroutine is to calculate new state and output variables from the actual state and input for one time step DT. You can create up to 4 new dynamic units by inserting FORTRAN lines into the dummy subroutines USER1—USER4. To show how your internal variables are transferred to the other parts of TACS we list here the subroutines USER1.

```
SUBROUTINE USER1
1 (N, NI, NR, NC, NO, NP, ITIME, JTIME, IH, TAU, DT, T, X)
DIMENSION X(N)
:
Here you insert FORTRAN lines.
:
RETURN
END
```

Subroutines USER1—USER4 are completely analogous. The variables in the argument list are as follows:

N = NI + NR*NC + NO + NP
NI — number of input variables,
NR — number of rows in the state matrix,
NC — number of columns in the state matrix; Remark: With distributed models in mind, state variables are stored in an NR * NC array. If you need a state vector, set NC = 1 in input data.
NO — number of output variables,
NP — number of parameters,
ITIME — total number of time steps, ITIME = T/DT,

JTIME — actual simulation time in time steps,
IH — error code

$$IH \begin{cases} 10*(\text{type index})+(\text{type of error}) \\ 0 \quad \text{otherwise,} \end{cases}$$

T — end time of simulation,
DT — time step size,
TAU — actual simulation time,
$X$ — $N$-vector containing input, state and output variables and parameters as shown in *Fig. A5—3*. Notice the first entry $X(1)$ should be reserved for actual simulation time, placed there by TACS.

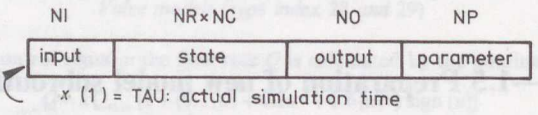



Fig. A5—3. Model variables

Models defined by subroutines USER1—USER4 are referred to by type indices 5—8, respectively. The dummy subroutines are given in a separate file USERS on the diskette supplied. If you insert your own FORTRAN lines this part of the program must be recompiled, linked, and loaded with the existing system.

ETACS includes the subroutine ROW4AB for solving systems of ordinary differential equations. ROW4AB is an extension of the program ROW4A written by Gottwald and Wanner [2]. It efficiently solves both stiff and non–stiff differential equations and hence can be used in user-supplied subroutines. The subroutine head is given by

SUBROUTINE ROW4AB
1 (N, NL, NU, T, P, X, TEND, TOL, HMAX, H, IPAS, DIFFU, DIAGS)

N — number of differential equations,
NL — number of lower co-diagonals of the Jacobian matrix
NU — number of upper co-diagonals of the Jacobian matrix
T — initial time,
P(.) — parameter vector to user supplied subroutines (not used in ROW4AB),
X(N) — IN: values of independent variables at $T$,
OUT: the same at final time (TEND),
TEND — end time,
TOL — relative error tolerance for the ODE solution,
HMAX — maximum size of step,

H — beginning size of step,
IPAS — IN: maximum number of steps,
OUT: succesfull steps,
DIFFU — auxiliary subroutine for computing time derivatives (DX(I) = FI),
DIAGS — auxiliary subroutine for computing Jacobian matrix with respect to X (PD(I, J) = dFI/dX(J)),

You can choose the name of auxiliary subroutines DIFFU and DIAGS. The head of DIFFU is given by

SUBROUTINE DIFFU (N, T, P, X, DX)

N — number of differential equations,
T — initial time,
P(.) — parameter vector,
X(N) — vector of independent variables,
DX(N) — vector of time derivatives,

The head of DIAGS is given by

SUBROUTINE DIAGS (N, NL, NU, T, P, X, DX, PD)

N — number of differential equations,
NL — number of lower co-diagonals of the Jacobian matrix
NU — number of upper co-diagonals of the Jacobian matrix
T — initial time,
P(.) — parameter vector,
X(N) — vector of independent variables,
DX(N) — vector of time derivatives,
PD(N, N) or PD(N*(NL + NU + 1)) — Jacobian with respect to X.

When NU or/and NL have non-zero values the Jacobian is a band matrix. The elements of the band matrix are stored row-wise in a one-dimensional array, assuming NW = NL + NU + 1 elements for each row. Only the non-zero elements are actually stored in the one-dimensional array, beginning at the first position for each row. For example the non-symmetric band matrix

$$\mathbf{PD} = \begin{bmatrix} 1 & 2 & 3 & 0 & 0 \\ -1 & 1 & 2 & 3 & 0 \\ 0 & -1 & 1 & 2 & 3 \\ 0 & 0 & -1 & 1 & 2 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

with N = 5, NU = 2, NL = 1

is stored in a one-dimensional array, also named PD, as follows

| | | | |
|---|---|---|---|
| PD(1)=1 | PD(2)=2 | PD(3)=3 | PD(4)=is not set |
| PD(5)=−1 | PD(6)=1 | PD(7)=2 | PD(8)=3 |
| PD(9)=−1 | PD(10)=1 | PD(11)=2 | PD(12)=3 |
| PD(13)=−1 | PD(14)=1 | PD(15)=2 | PD(16)=is not set |
| PD(17)=−1 | PD(18)=1 | PD(19)=and | PD(20)=are not set. |

## A5—2 HOW TO USE THE PROGRAM

Though the program is self-explanatory and very easy to use, some remarks and a summary of input data may further facilitate its application.

### A5—2.1 Input-output options and general remarks

Input options are as follows:

— control network is specified by answering questions appearing on the display; or

— control network is specified by reading data from an input file. Following data input a summary of the network is printed. If required, you may modify this network and simulation control parameters by answering some further questions.

All results appear on the display, including echoes of entered values. In addition, you may request

— saving results in a file; or

— creating a data file for further use.

The latter option assumes that you specify the network on the display, since otherwise you already have an input file.

Input and output options can be selected from a menu of execution modes at the beginning of program run. The following points should be noted:

— Data should be separated by semicolon (for terminate string), comma or space (for numerical data).

— Decimal points are only necessary when the data has a fractional part. Exponent marks E or D can be used.

— Variable names must not contain semicolon separator and must occupy at most the first 4 characters of the line.

— 'Break' is $$$$ in the first 4 characters of the line. This string enables you to return to an earlier point in the program and to re-enter erroneous data. Notice

that data can also be corrected at later stages of the program before the simulation run.

— To give a positive answer for a Yes/No question use y or Y. Any other characters (except $$$$ are recognized as N).

## A5—2.2 Summary of input data

*Table A5—3* gives a summary of input data. Notice that for the units supplied, output No. in Item is 1 (single—output units). Input No. is 2 for all units except units 16 and 18, where you may select inputs 2 and 3. (Input No. 1 is reserved for actual simulation time.)

## A5—2.3 Additional options in ETACS

While in TACS all variables are plotted as functions of time, ETACS provides two further options. First, when plotting a column of a state variable matrix we can plot distributions of variables along the spatial coordinate in distributed parameter units. Second, plotting an output as a function of another output variable, phase trajectories can be constructed. Profiles and phase trajectories are specified following item 16 in *Table A5—3* as follows.

Profile specifications:

Each plot may contain up to 3 distribution functions. Each function is specified by the number of the unit, the index of the column in the state variable matrix (i.e., index of the state variable to be plotted), the time point considered, and the character used for plotting.

Phase trajectory specifications:

Each plot may contain up to 3 phase trajectories. Each trajectory is specified by four integers $i1$, $i2$, and $j1$, $j2$, such that output $i2$ of unit $i1$ is plotted versus output $j2$ of unit $j1$. The character to be used is also selected. After specifying all the desired phase trajectories the last data item is an integer $q$ such that values for phase trajectories will be sampled in each $q$th computational cycle. Each specification step is terminated by entering an empty line.

## A5—2.4 Example

Consider the stirred tank heater with temperature control system shown in *Fig. A5—2* and let $T_{set}$ denote the desired temperature. Introducing deviations from the steady state as variables of the model, the system can be represented by the block diagram shown in *Fig. A5—3*, where SETP = 0. Variables OPEN and CONT denote responses of the uncontrolled process and that of the closed-loop system, respectively.

Table A5—3. Summary of input data

| Section | Item | Question | Response | Remark |
|---|---|---|---|---|
| Mode select | 1 | select an execution mode | 0, 1, 2, 3, 4 | |
| ( ) 0, 4: | 2 | unit number for batch file | from 1 to 99 except 10, 11 | |
| | 3 | name of batch file | file name | |
| ( ) 3: | 4 | unit number for printer file | from 1 to 99 except 10, 11 | |
| | 5 | name of printer file | file name | go to Item 8. |
| ( ) 2: | 6 | unit number for batch file | from 1 to 99 except 10, 11 | |
| | 7 | name of batch file | file name | |
| 1: | 8 | | | opening of data files |
| Menu | 9 | Do you want the menu of dynamic units | *Y* or *N* | *Y*: list units |
| | | SPECIFICATION OF CONTROL NETWORK | | |
| Time specification | 10 | enter time step size and terminal time | 2 floats | |
| Units | | SPECIFICATION OF DYNAMIC UNITS<br>to terminate an empty line | | |
| | 11 | specification of *x*th unit Enter type index | *i* integer | items 11–13 are repeated for all units |
| | 12 | output: Enter name and value | 4 characters +1 float | items 12 and 13 are repeated for all units |
| | 13 | parameter: Enter name and value | 4 characters +1 float | outputs and parameters |
| Connection | | SPECIFY CONNECTIONS<br>if unit 11, output 12 is connected to unit J1, input J2, enter I1, I2, J1, J2.<br>To terminate enter an empty line | | |
| | 14 | enter *x*th connection | 4 integers | item 14 is repeated for all connections |
| Plot | | PLOT SPECIFICATIONS<br>To terminate enter an empty line | | |
| | 15 | enter *x*th function to be plotted (unit, output, min., max.) | 2 integers + 2 floats | item 15 and 16 are repeated for all functions to be plotted |
| | 16 | enter character to be used | 1 character | |
| Print | | PRINT SPECIFICATIONS<br>to terminate enter an empty line | | |

Table A5—3 (continued)

| Section | Item | Question | Response | Remark |
|---|---|---|---|---|
| | 17 | enter $x$th function to be printed (unit, output) | 2 integers | item 17 is repeated for all functions to be printed |
| | 18 | enter step, between two outputs | 1 integer | |
| Profile ( ) | | PROFILE SPECIFICATIONS to terminate enter an empty line | | |
| | 19 | enter $x$th profile of $y$th draw (unit, number of column of state matrix, time) | 2 integers + 2 floats | item 19 and 20 are repeated for all functions and draws to be printed |
| | 20 | enter character to be used | 1 character | |
| Phase trajectory ( ) | | PHASE TRAJECTORY SPECIFICATIONS | | |
| | 21 | enter $x$th phase trajectory of $y$th draw (unit, output versus unit output) | 4 integers | item 21 and 22 are repeated for all functions and draws to be printed |
| | 22 | enter character to be used | 1 character | |
| | 23 | enter steps between two outputs | 1 integer | |
| | 24 | do you want to list the data of network | Y or N | Y: list the data |
| Break point 1 | 25 | do you want a simulation run | Y or N | Y: Perform simulation<br>N: go to Item 27 |
| | 26 | do you want to list the collected data of functions | Y or N | Y: list the data |
| Break point 2 | 27 | do you want to modify the data of network | Y or N | Y: go to Item 29 |
| | 28 | do you want to restart | Y or N | Y: go to Item 1<br>N: end of program |
| Modify | | MODIFYING OF CONTROL NETWORK | | |
| | 29 | enter time step size and terminal time | 2 floats or Enter | enter: Old values are actual numbers |
| | 30 | do you want to modify units | Y or N | N: Go to Item 27 |
| | 31 | enter unit to respecify | 1 integer | actual number of the unit in the network |
| | 32 | do you want to add new unit | Y or N | Y: define a new unit |
| | 33 | do you want to modify connections | Y or N | Y: repeat Item 14 |
| | 34 | do you want to modify output | Y or N | Y: repeat Items 15–17 |
| | 35 | enter steps between two outputs | 1 integer | return to Item 19 |
| ( ) | 36 | do you want to modify data for profiles | Y or N | Y: repeat Items 19–20 |
| ( ) | 37 | do you want to modify data for phase trajectories | Y or N | Y: repeat Items 21–23 |

( )—unit number 11 is reserved only in ETACS
( )—only in ETACS

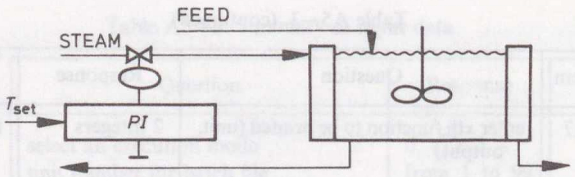



Fig. A5—4. Construction of a network

Compute:

(i) response of the uncontrolled process to unit step change in load variable;

(ii) closed-loop response to unit step change in load variable with *PI*-controller settings

$$K_c = 3 \quad \text{and} \quad T_I = 9.$$

For simplicity we assume $K=1$ and $T=0$ and solve the problem by TACS.

Following program initiation you may request the menu of dynamic units, presently available in the system. According to these indices, the analysis requires construction of the network shown in *Fig. A5—4*. Since the set point (STEP on *Fig. A5—5*) is null, the corresponding input is dropped. After constructing the control network we proceed to its specification. The first input data are simulation time, step size, and terminal time. To select a dynamic unit enter its type index. The response is a request for variable names and values. You should enter a 4-character name and value for the output variable, as well as for the parameters of the unit. The process displays the meaning of parameters to be specified. For example, in case of a step function generator (type index: 1) you should enter names and values for step time (before this time point the output remains 0) and the amplitude of the step. All other units of *Fig. A5—4* are similarly defined. For some units (e.g., for addition of two signals or for sign inversion) only output name and initial value should be specified. In terminate the specification of value should be specified. To terminate the specification of units, enter an empty line.

Now proceed to the network shown in *Fig. A5—6* by specifying connections. If output $j1$ of unit $i1$ is the input $j2$ of unit $i2$, enter the values of the 4 integers $i1$, $j1$, $i2$, $j2$. It should be emphasized that for each unit the first input variable is time,

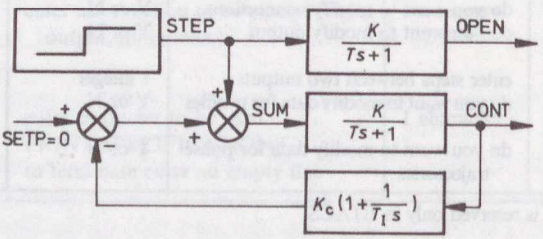



Fig. A5—5. Construction of a network

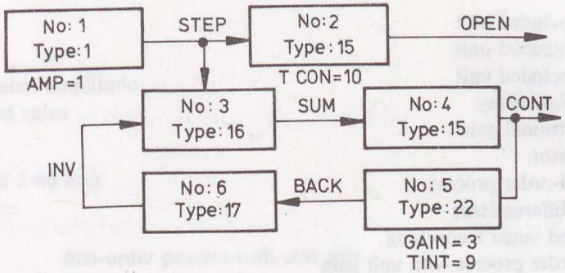



Fig. A5—6. Construction of a network

and hence 'free' inputs are indexed by 2, 3, . . . . For example, to connect the step function (output 1 of unit 1) to the input of the first-order unit (input 2 of unit 2), enter 1, 1, 2, 2. To terminate the specification of connections enter an empty line.

It remains to specify the requested output, thus select output variables to be plotted and printed during the simulation. To select a function to be plotted, enter No. of the unit, No. of the output, the expected minimum and maximum values for the variable, and a character to be used for plotting the function considered. To terminate the specification of functions to be plotted enter an empty line.

```
EEEEEEEE  TTTTTTTT   AAAAAA    CCCCCC   SSSSSS
EEEEEEEE  TTTTTTTT  AAAAAAAA  CCCCCCCC  SSSSSSSS
EE            TT    AA    AA  CC    CC  SS
EEEEE         TT    AA    AA  CC        SSSSSSS
EEEEE         TT    AAAAAAAA  CC         SSSSSSS
EE            TT    AAAAAAAA  CC    CC       SS
EEEEEEEE      TT    AA    AA  CCCCCCCC  SSSSSSSS
EEEEEEEE      TT    AA    AA   CCCCCC   SSSSSS
```

Extended Teaching Aid for Control Studies on IBM PC version 1.00 by L. Farkas, S. Vajda

Do you want menu of dynamic units (Y/N)?

Y

MENU OF AVAILABLE DYNAMIC UNITS

| type index | unit |
|---|---|
| 1 | step function generator |
| 2 | sinus function generator |
| 3 | impulse function generator |
| 4 | ramp function generator |
| 5 | user-included unit |

| | |
|---|---|
| 6 | user-included unit |
| 7 | user-included unit |
| 8 | user-included unit |
| 9 | pure time delay |
| 10 | proportional gain |
| 11 | integrator |
| 12 | second-order process |
| 13 | ideal differentiator |
| 14 | weighed linear smoothing |
| 15 | first-order process with unit gain |
| 16 | addition of two signals |
| 17 | sign inversion |
| 18 | multiplication of two signals |
| 19 | relay (on–off controller) |
| 20 | proportional gain with constraints |
| 21 | ideal integral controller |
| 22 | ideal proportional-integral controller |
| 23 | proportional-integral-differential controller |
| 24 | steam heater |
| 25 | steam boiler |
| 26 | stirred tank reactor with first-order process |
| 27 | shell and tube heat exchanger |
| 28 | valve modell for incompressible fluid |
| 29 | valve modell for compressible fluid |
| 30 | off-line identification by least squares |
| 31 | off-line identification by instrumental variables |
| 32 | on-line identification by least squares |
| 33 | on-line identification by instrumental variables |
| 34 | Nyquist and Bode plots |
| 35 | pseudorandom binary signal generator |
| 36 | discrete-time linear SISO unit |

SPECIFICATIONS OF CONTROL NETWORK

enter time step size and terminal time

0.1 30

SPECIFICATION OF DYNAMIC UNITS

to terminate: enter an empty line

specification of 1-st unit

enter type index

1

1 step function generator

number of inputs: 1

1-st time (s)

1-st output: step function

enter name and value

STEP 0.

1-st parameter: siso time

```
STIM 0.
      2-nd parameter: amplitude
    enter name and value
AMP 1.
    specification of 2-nd unit
    enter type index
15
            15          first-order process with unit gain
    number of inputs: 2
      1-st    time (s)
      2-nd    input signal
     1-st output: response function
    enter name and value
OPEN 0.
      1-st parameter: time constant
    enter name and value
TCON 10.
    specification of 3-rd unit
    enter type index
16
            16          addition of two signals
    number of inputs: 3
      1-st    time (s)
      2-nd    signal 1
      3-rd    signal 2
     1-st output: sum of two signals
    enter name and value
SUM 0.
    specification of 4-th unit
    enter type index
15
            15          first-order process with unit gain
    number of inputs: 2
      1-st    time (s)
      2-nd    input signal
     1-st output: response function
    enter name and value
CONT 0.
      1-st parameter: time constant
    enter name and value
TCON 10.
    specification of 5-th unit
    enter type index
22
            22          ideal proportional-integral controller
```

```
    number of inputs: 2
        1-st    time (s)
        2-nd    input signal

      1-st output: control signal
    enter name and value
DACK 0.

      1-st parameter: gain factor
    enter name and value
GAIN 3.

      2-nd parameter: integration time constant
    enter name and value
TINT 9.

    specification of 6-th unit
    enter type index
17
              17          sign inversion

    number of inputs: 2
        1-st    time (s)
        2-nd    input signal

      1-st output: inverted input signal
    enter name and value
INV 0.

    specification of 7-th unit
    enter type index

    SPECIFY CONNECTIONS
    if unit I1, output I2 is connected to unit J1, input J2, enter I1, I2, J1, J2 to terminate: enter an
    empty line

    enter    1-st connection
1 1 2 2
    enter    2-nd connection
1 1 3 2
    enter    3-rd connection
3 1 4 2
    enter    4-th connection
4 1 5 2
    enter    5-th connection
5 1 6 2
    enter    6-th connection
6 1 3 3
    enter    7-th connection

    PLOT SPECIFICATIONS

    to terminate: enter an empty line
    enter 1-st function to be plotted (unit, output, min., max.)
```

```
1 1 0 1
    enter character to be used
I
    enter 2-nd function to be plotted (unit, output, min., max.)
2 1 0 1
    enter character to be used
+
    enter 3-rd function to be plotted (unit, output, min., max.)
4 1 0 1
    enter character to be used

    enter 4-th function to be plotted (unit, output, min., max.)

    PRINT SPECIFICATIONS, (enter unit, output)

    to terminate: enter an empty line
    enter 1-st function to be printed

    enter steps between two outputs
10

    PROFILE SPECIFICATIONS
    to terminate: enter an empty line
    enter 1-st profile of 1-st draw
        (unit, number of column of state matrix, time)

    PHASE TRAJECTORY SPECIFICATION
    to terminate: enter an empty line
    enter 1-st phase trajectory of 1-st draw
        (unit output versus unit output)

    ***** READING END OF FILE *****

    *** MODE CHANGING, THE NEW MODE IS 3 ***

    PHASE TRAJECTORY SPECIFICATION
    to terminate: enter an empty line
    enter 1-st phase trajectory of 1-st draw
        (unit output versus unit output)

    DYNAMIC UNITS IN THE NETWORK:

        1    step function generator
        15   first-order process with unit gain
        16   addition of two signals
        22   ideal proportional-integral controller
        17   sign inversion

  Do you want to list the data of network (Y/N)?
Y
```

```
UNIT:  1
          1    step function generator
inputs:   1,   state matrix:   1*    1,   outputs:   1,   parameters:   2
                               TIME   —   .000000
                               ARB1   —   .000000
                               STEP   —   .000000
                               STIM   —   .000000
                               AMP    —   1.00000
UNIT:  2
         15    first-order process with unit gain
inputs:   2,   state matrix:   1*    1,   outputs:   1,   parameters:   1
                               TIME   —   .000000
                               INP2   —   .000000
                               ARB1   —   .000000
                               OPEN   —   .000000
                               TCON   —   10.0000
UNIT:  3
         16    addition of two signals
inputs:   3,   state matrix:   1*    1,   outputs:   1,   parameters:   1
                               TIME   —   .000000
                               INP2   —   .000000
                               INP3   —   .000000
                               ARB1   —   .000000
                               SUM    —   .000000
                               ARB3   —   .000000
UNIT:  4
         15    first-order process with unit gain
inputs:   2,   state matrix:   1*    1,   outputs:   1,   parameters:   1
                               TIME   —   .000000
                               INP2   —   .000000
                               ARB1   —   .000000
                               CONT   —   .000000
                               TCON   —   10.0000
UNIT:  5
         22    ideal proportional-integral controller
inputs:   2,   state matrix:   1*    2,   outputs:   1,   parameters:   2
                               TIME   —   .000000
                               INP2   —   .000000
                               ARB1   —   .000000
                               ARB2   —   .000000
                               BACK   —   .000000
                               GAIN   —   3.00000
                               TINT   —   9.00000
         17    sign inversion
inputs:   2,   state matrix:   1*    1,   outputs:   1,   parameters:   1
```

| | | |
|---|---|---|
| TIME | — | .000000 |
| INP2 | — | .000000 |
| ARB1 | — | .000000 |
| INV | — | .000000 |
| ARB3 | — | .000000 |

THE 6 CONNECTIONS IN THE NETWORK:

| | in unit | output | will be | in unit | input |
|---|---|---|---|---|---|
| 1-st | 1 | ! | — | 2 | 2 |
| 2-nd | 1 | ! | — | 3 | 2 |
| 3-rd | 3 | ! | — | 4 | 2 |
| 4-th | 4 | ! | — | 5 | 2 |
| 5-th | 5 | ! | — | 6 | 2 |
| 6-th | 6 | ! | — | 3 | 3 |

number of functions to be plotted: 3
number of functions to be printed: 0
number of profiles: 0
number of phase trajectories: 0

PLOTS:

| unit | output | min. value | max. value | character: |
|---|---|---|---|---|
| 1 | 1: STEP | .00000 | 1.0000 | I |
| 2 | 1: OPEN | .00000 | 1.0000 | + |
| 4 | 1: CONT | .00000 | 1.0000 | * |

Do you want a simulation run (Y/N)?

Y

PLOTTED 3 FUNCTION(S)

| | | | | | | |
|---|---|---|---|---|---|---|
| 1-st | unit | STEP | min.: | 1.00000 | max.: | 1.00000 |
| 2-nd | unit | OPEN | min.: | .000000 | max.: | .919712 |
| 4-th | unit | CONT | min.: | .000000 | max.: | .196969 |

| TIME | STEP | OPEN | CONT |
|---|---|---|---|
| 1.00000 | 1.00000 | .860688E–01 | .739233E–01 |
| 2.00000 | 1.00000 | .173041 | .136446 |
| 3.00000 | 1.00000 | .251736 | .172737 |
| 4.00000 | 1.00000 | .322943 | .190968 |
| 5.00000 | 1.00000 | .387373 | .196969 |
| 6.00000 | 1.00000 | .445673 | .194890 |
| 7.00000 | 1.00000 | .498424 | .187625 |
| 7.99999 | 1.00000 | .546155 | .177250 |
| 9.00000 | 1.00000 | .589344 | .165171 |
| 10.0000 | 1.00000 | .628423 | .152356 |
| 11.0000 | 1.00000 | .663783 | .139454 |
| 12.0000 | 1.00000 | .695778 | .126887 |
| 13.0000 | 1.00000 | .724729 | .114917 |
| 14.0000 | 1.00000 | .750925 | .103696 |

| | | | |
|---|---|---|---|
| 15.0000 | 1.00000 | .774627 | .932991E–01 |
| 16.0000 | 1.00000 | .796074 | .837493E–01 |
| 17.0000 | 1.00000 | .815480 | .750380E–01 |
| 18.0000 | 1.00000 | .833040 | .671318E–01 |
| 19.0000 | 1.00000 | .848928 | .599858E–01 |
| 20.0000 | 1.00000 | .863304 | .535480E–01 |
| 21.0000 | 1.00000 | .876313 | .477631E–01 |
| 22.0000 | 1.00000 | .888083 | .425756E–01 |
| 23.0001 | 1.00000 | .898733 | .379316E–01 |
| 24.0001 | 1.00000 | .908370 | .337796E–01 |
| 25.0001 | 1.00000 | .917090 | .300716E–01 |
| 26.0001 | 1.00000 | .924980 | .267631E–01 |
| 27.0001 | 1.00000 | .932119 | .238129E–01 |
| 28.0001 | 1.00000 | .938579 | .211840E–01 |
| 29.0001 | 1.00000 | .944424 | .188424E–01 |
| 30.0001 | 1.00000 | .949712 | .167574E–01 |

Do you want to modify the data of network (Y/N)?

Do you want to restart (Y/N)?

END OF PROGRAM ETACS

## A5—3 PROGRAM IMPLEMENTATION

The present versions are dimensioned in order to handle
50 dynamic units,
100 connections,
25 functions to be printed or plotted,
in ETACS only
25 profile curves to be plotted,
25 phase trajectory curves to be plotted, TACS and ETACS require 95 KBy and 180 KBy of memory, respectively. If you add your own models with a large number of variables, the bottleneck of the programs may be either array W of length 2000 storing all input, output and state variables and parameters, or array S of length 1200 storing all work variables of plotter subroutines and the extended subroutines of ETACS. If you exceed these limits you receive error messages. In such rare cases you should increase the dimensions of W, NEV and S, and the value of KLIM (3) and KLIM (5), respectively.

The present versions assume that BO characters are available on the printer. To obtain larger plots change the value of KLIM (6) to the actual record length of your device (minimum is 80).

RESULTS

**********

PLOTS 3 FUNCTION(S)
TIME STEP SIZE: .1000 10 STEP(S) BETWEEN TWO OUTPUTS

1-st UNIT STEP — CHARACTER: I
2-nd UNIT OPEN — CHARACTER: +
4-th UNIT CONT — CHARACTER: *

SCALES:

```
 I: .00000        .25000        .50000        .75000       1.0000
      I----------I----------I----------I----------I----------I
 +: .00000        .25000        .50000        .75000       1.0000
      I----------I----------I----------I----------I----------I
 *: .00000        .25000        .50000        .75000       1.0000
TIME  I----------I----------I----------I----------I----------I
1.000      *+                                                 I
2.000       * +                                               I
3.000        *  +                                             I
4.000         *   +                                          I
5.000          *    +                                        I
6.000           *     +                                      I
7.000            *      +                                    I
8.000             *       +                                  I
9.000              *        +                                I
10.00               *         +                              I
11.00                *          +                            I
12.00                 *           +                          I
13.00                  *            +                        I
14.00                   *             +                      I
15.00                    *              +                    I
16.00                    *               +                   I
17.00                     *               +                  I
18.00                     *                +                 I
19.00                      *                +                I
20.00                      *                 +               I
21.00                       *                 +              I
22.00                       *                  +             I
23.00                        *                 +             I
24.00                        *                  +            I
25.00                         *                 +            I
26.00                         *                  +           I
27.00                         *                  +           I
28.00                         *                   +          I
29.00                         *                   +          I
30.00                         *                   +          I
```

Do you want to list collected data of functions (Y/N) ?

## A5—4 FILES ON THE DISKETTE

TACS

| | |
|---|---|
| TACS.FOR | — main program and initialization, |
| TACS1.FOR | — subroutines for network simulation and for data input to units, |
| TACS2.FOR | — subroutines to input data of network and output specifications, plot or print data, collect data for plot and print, |
| TACS3.FOR | — subroutines for units and to list data of network and output specifications, |
| TACS4.FOR | — auxiliary subroutines and functions, |
| USERS.FOR | — subroutines for user-written units, |
| TACS.EXE | — executable program, |
| MANUAL.DAT | — example input data file. |

ETACS

| | |
|---|---|
| ETACS.FOR | — main program and initialization, |
| ETACS1.FOR | — subroutines for network simulation and for data input to units, |
| ETACS2.FOR | — subroutines to input data of network and output specifications, to plot or print data, collect data for plot, print and phase trajectory; |
| ETACS3.FOR | — subroutines for units, |
| ETACS4.FOR | — subroutines for units and auxiliary tasks, |
| ETACS5.FOR | — subroutines to list data of network and output specifications, and create profile and phase trajectory plots, |
| ETACS6.FOR | — subroutines to create Nyquist diagrams, |
| ETACS7.FOR | — auxiliary subroutines and functions, |
| USERS.FOR | — subroutines for user-written units, |
| ETACS.EXE | — executable program, |
| MANUAL.DAT | — example input data file. |

## A5—5 REFERENCES

[1] Goodwin, S. S. and Payne, R. C. (1977): *Dynamic System Identification Experiment Design, and Data Analysis.* Academic Press, New York.

[2] Gottwald, B. A. and Wanner S. (1982): *Simulation*, **37**, 169.

[3] Gould, L. A. (1969): *Chemical Process Control: Theory and Application.* Addision-Wesley, Reading, MA.

[4] Franks, R. G. E. (1972): *Modelling and Simulation in Chemical Engineering.* Wiley-Interscience, New York.

# APPENDIX A6

## Appendix to Chapter 6

### EURECHA

European Committee For Computers in Chemical Engineering Education

### STATCHAR

Statistical Characterization of Measurement Results

MANUAL

Program Version 1
May, 1984

Manual authors: G. E. VERESS and M. KENDERESY-SZŰCS

Department of General and Analytical Chemistry
Technical University of Budapest

Budapest
Szt. Gellért tér 4.
Hungary
H-1111

# CONTENTS


A6—1 INTRODUCTION 601

A6—2 SECTION CHAR 602
A6—2.1 Introduction 602
A6—2.2 Point estimation 602
A6—2.2.1 Point estimation for true value 602
A6—2.2.2 Point estimation of systematic error 602
A6—2.2.3 Point estimation of precision 603
A6—2.3 Test for normality 603
A6—2.4 Hypothesis testing 603
A6—2.4.1 Hypothesis testing for true value 604
A6—2.4.2 Hypothesis testing for systematic error 604
A6—2.4.3 Hypothesis testing for precision 604
A6—2.5 Interval estimation 604
A6—2.5.1 Interval estimation for true value 605
A6—2.5.2 Interval estimation for systematic error 605
A6—2.5.3 Interval estimation for precision 605

A6—3 SECTION COMP 606
A6—3.1 Introduction 606
A6—3.2 Comparison of precisions 606
A6—3.3 Comparison of true values 607
A6—3.4 Comparison of systematic errors 607

A6—4 SECTION CALIN 607
A6—4.1 Introduction 607
A6—4.2 Estimation of parameters of calibration line 608
A6—4.3 Interval estimation of the unknown component 609

A6—5 PROGRAM USE 610

A6—6 EXAMPLES 616
A6—6.1 Example for section CHAR 616
A6—6.2 Example for section COMP 618

A6—7 NOTES ON PROGRAM IMPLEMENTATION 620

## A6—1 INTRODUCTION

In order to obtain the maximum information from experimental data it is necessary to characterize and know the magnitude of the errors involved, so that the accuracy and precision of the measurement must be known, as also must be known the bias and deviation of the mean of replicate measurements.

The program in the STATCHAR package has been developed in an analytical chemistry department in order to obtain maximum information from experimental data. The program goes further than the simple mean and standard deviation calculation packages commonly available, since various statistical tests are performed to identify the magnitude of both the random and systematic error. The ultimate reliability of any measurement method depends on the systematic and random errors.

Three sections are presented in this program:

CHAR — for the statistical CHARACTERIZATION of replicate measurements
COMP — for the COMPARISON of sets of replicate measurements
CALIN — for the statistical characterization of CALIBRATION line and of the measuring result using this line.

Section CHAR can be used to characterize the precision and systematic error of measuring methods and is also used to give the statistical characterization of the measured results of unknown material.

Section COMP performs the statistical comparison of two measurement series and is suitable for comparing the precision and accuracy of measuring methods or of giving the mathematical statistical comparison of the composition of materials.

Section CALIN performs the construction and statistical characterization of the calibration line and the point and interval estimation of unknown material quantities using the calibration line.

The program has been developed for extracting the maximum information from measured data. It is also useful in the teaching environment to demonstrate the statistical characterization of measured data, and to link courses in statistical theory with the handling of laboratory data.

STATCHAR can be operated in interactive and batch modes. The user first selects among sections, and can then choose procedures from within the section. The procedures can be used repeatedly.

# A6—2 SECTION CHAR

## A6—2.1 Introduction

Section CHAR performs the mathematical statistical characterization of replicate analytical chemical measurement results.

The user is offered four optional services:

— point estimation for true value, systematic error and precision,
— test for normality of measurement results,
— hypothesis testing for true value, systematic error and precision,
— interval estimation for true value, systematic error and precision.

It is assumed that the measured result is a sum of the true value, systematic error and random error, that is

$$\text{measured result} = \text{true value} + \text{systematic error} + \text{random error}$$

The section operates only if the number of replicate measurements exceeds five.

## A6—2.2 Point estimation

The subprogram performs the point estimation of true value, systematic error and precision for replicate analytical chemical measurement results. The user can choose to estimate the true value or the systematic error or precision.

### A6—2.2.1 Point estimation for true value

If the user wishes to give a point estimation for true value then besides the measured results the value of systematic error should be given as an input to the program. The output, i.e. the point estimation of the true value, will be the difference between the mean of measured results and systematic error:

$$\text{point estimation of true value} = \text{mean of measured results} - \text{systematic error}$$

The program will also give the estimated value of precision, see Section A6—2.2.3.

### A6—2.2.2 Point estimation of systematic error

If the user wishes to determine the point estimation of systematic error, i.e. the accuracy of the method, the measured results and the known true value must be given as an input. The output is then the point estimation of systematic error given as a difference of the mean of measured results and the true value:

Point estimation of systematic error = mean of measured results − true value

In addition the program estimates the value of precision, too, see Section A6—2.2.3.

### A6—2.2.3 Point estimation of precision

If the point estimation of precision is required by the user then the program input consists only of the measured data. The point estimation of precision is given as the variance of measured results, that is:

Point estimation of precision = variance of measured results

This service is also provided when the point estimation of true value or systematic error is carried out (see Sections A6—2.2.1 and A6—2.2.2).

## A6—2.3 Test for normality

Since most mathematical statistical procedures require normality of the data it is of considerable interest to learn whether the measured data are normally distributed or not.

If the user wishes to test measured data for normality the TEST for NORMALITY will require an input of the measured data and the confidence level of the test. On basis of the $\chi^2$ test the program performs the test for normality on measured results. As a result the program tells the user whether the data can be regarded as being normally distributed or not.

## A6—2.4 Hypothesis testing

The section CHAR performs the hypothesis testing for true value, systematic error and precision from replicate analytical chemical measurement results presuming the model of measured results introduced in Section A6—2.1.

The applied statistical tests are based on the assumption that the data are normally distributed. The section therefore checks whether the test for normality was carried out. If it was not, the test for normality is performed for which the confidence level must be given as an input (see Section A6—2.3). If the data are normally distributed then hypothesis testing is activated. If the measured data cannot be regarded as normally distributed hypothesis testing is performed only at the explicit request of the user.

The user may choose to implement hypothesis testing for either the true value, systematic error or precision.

### A6—2.4.1 Hypothesis testing for true value

If the user desires to perform hypothesis testing for true value the program asks for an input of replicate measured results, the value of systematic error, the presumed value of true value, and the confidence level.

Using the point estimation for true value (see Section A6—2.2.1) the program executes hypothesis testing on basis of the one-sample *t*-test. The output indicates whether the presumed value of the true value is acceptable or not at the given confidence level and the point estimation for the true value.

### A6—2.4.2 Hypothesis testing for systematic error

If hypothesis testing for systematic error (i.e. accuracy) is required then user gives input data containing the replicate measured results, true value, presumed value of systematic error and the confidence level of the hypothesis test.

Using the point estimation for systematic error (Section A6—2.2.2) the program accomplishes hypothesis testing on the basis of the one-sample *t*-test. The output indicates whether the presumed value from systematic error is acceptable or not at the given confidence level and the point estimation for the systematic error.

### A6—2.4.3 Hypothesis testing for precision

If the user wants to perform hypothesis testing for precision the program input will require the replicate measured data, the presumed value of precision and the confidence level of the hypothesis testing.

Using point estimation for precision (Section A6—2.2.3) the hypothesis testing for precision applies the $\chi^2$ test. The output indicates the acceptance or rejection of the presumed value for precision at the given confidence level and the point estimation for precision.

## A6—2.5 Interval estimation

The section CHAR also gives the interval estimation for true value, systematic error and precision from replicate analytical chemical measured results assuming the model introduced in Section 2.1.

Whereas the statistics applied for interval estimation assume the normality of measured data. The program checks in every case whether the test for normality had been completed the program asks for a confidence level and runs the test (see Section A6—2.3). If the normality of measured data can be assumed by the test, then the program continues to execute interval estimation. When the data cannot

be regarded as normally distributed, the interval estimation is only performed at the explicit request of the user.

The user may ask for the interval estimation for either true value or systematic error or precision.

### A6—2.5.1 Interval estimation for true value

If the user wants an interval estimation for true value then the program calls for the input of the value of systematic error and the confidence level of the interval estimation besides the measured data.

The program carries out interval estimation for true value using point estimation for true value (Section A6—2.2.1) and the *t*-distribution. The output gives the point estimation for true value and the lower and upper value of the interval at the given confidence level.

### A6—2.5.2 Interval estimation for systematic error

If the user wants an interval estimation for systematic error (i.e. accuracy) then the program calls for the input of the true value and confidence level for estimation besides replicate measured results.

The program carries out interval estimation for systematic error using point estimation for systematic error (Section A6—2.2.2) and *t*-distribution. The output gives the lower and upper value of the interval at the given confidence level besides the point estimation of the systematic error.

### A6—2.5.3 Interval estimation for precision

If the user wants an interval estimation for precision (i.e. variance of the measured data) then the program calls for the input of the confidence level for interval estimation besides the replicate measured results.

The program performs interval estimation for precision (see Section A6—2.2.3) on basis of the $\chi^2$ distribution. The output gives the lower and upper value of the interval at the given confidence level and the point estimation for precision.

# A6—3 SECTION COMP

## A6—3.1 Introduction

The section COMP performs the statistical comparison of two measurement series consisting of replicate analytical chemical measurements.

The user is offered the following options:

— comparison of true values
— comparison of systematic errors
— comparison of precisions.

The section operates only if the number of replicate measurements is at least five. It is assumed that the measured result is a sum of the true value, systematic error and random error:

$$\text{Measured result} = \text{true value} + \text{systematic error} + \text{random error}.$$

In the section COMP all statistical tests used assume normality of the data sets. For this reason the program always executes normality testing at the given confidence level (see Section A6—2.3).

If normality of the data is accepted then the program continues operation. If the data are not normally distributed then the program continues to run only at the users' request.

After testing the normality of both data sets the program proceeds with their comparison.

## A6—3.2 Comparison of precisions

If the user wishes to give a comparison of precisions for two sets of measured values then besides the measured results the confidence level for the hypothesis testing should be given as an input to the program.

Using the point estimations for precisions (Section A6—2.2.3) of the two measured result sets section COMP executes hypothesis testing for the equality of the precisions on the basis of the *F*-test.

The output gives the estimated precisions for the two data sets, the calculated *F*-statistics and the acceptance or rejection of the hypothesis for the equality of the precisions at the given confidence level.

### A6—3.3 Comparison of true values

Section COMP achieves the statistical comparison of true values for two measurement series consisting of replicate analytical chemical results.

The section first gives the characterization of each measurement series. It calls for the value of the systematic error and gives point estimation of true value and precision (Section A6—2.2) for both measurement series.

The comparison of true values for two measurement series is carried out using the two-sample $t$-test if the deviations are equivalent, and the Welch test if the deviations differ. The required confidence level must be given as an input for hypothesis testing.

The output of the section contains the point estimation of true values of the two measurement series, the calculated value of the $t$-statistics and the acceptance or rejection of the hypothesis regarding the equivalence of the true values at given confidence level.

### A6—3.4 Comparison of systematic errors

The section COMP, at the request of the user, performs the statistical comparison of systematic errors i.e. the accurracies for two measurement series consisting of replicate measured results.

The program first gives a characterization of each data set. The true values are input and the point estimation of systematic error and precision are output (Section A6—2.2).

The two-sample t-test is used for the comparison of systematic errors if the deviations of the data sets are equivalent, and the Welch test is applied if the deviations differ. Hypothesis testing calls for the value of confidence level.

The program gives an output containing the point estimation of systematic errors for the two measurement series, the calculated value of $t$-statistics and the acceptance or the rejection of the hypothesis regarding the equivalence of systematic errors.

## A6—4 SECTION CALIN

### A6—4.1 Introduction

The section CALIN performs the construction, statistical characterization and application of the calibration line which is of significant importance in analytical chemical practice. The operation is twofold:

1. the parameters of the calibration line are determined during calibration (see Section A6—4.2);
2. point and interval estimation is given for an unknown material quantity on the basis of the calibration line and its statistical characteristics (see Section A6—4.3).

Since the construction of the calibration line can be regarded as a univariate linear regression, the section is also suitable for solving arbitrary univariate linear regression problems.

## A6—4.2 Estimation of parameters of calibration line

Section CALIN estimates the coefficients of linear calibration functions, i.e. calibration lines, assuming that the following relation holds between material quantity $x$ and measured analytical property $y$:

$$y = a + bx$$

where $a$ is the axial intercept section of the calibration line, $b$ is the slope of the calibration line or in other words, the sensitivity of the analytical method.

Subprogram CALIN gives an estimation of parameters $a$ and $b$ using as input calibration data pairs $\{x_i, y_i\}$ $i=1, \ldots, N$ where $x_i$ is the concentration (amount) and $y_i$ is the measured value of the $i$th calibrating material. The section reads data in the following order

$$x_1, \ldots, x_N, y_1, \ldots, y_N$$

Parameter estimation is carried out using the least squares method, which is the principle of classical calibration:

$$\min_{a,b} \sum_{i=1}^{N} (y_i - a - bx_i)^2$$

This principle can be applied if it can be assumed that $x_i$ is accurate and only $y_i$ is affected by error. Consequently the estimated values of the parameters are:

$$\hat{b} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

and

$$\hat{a} = \hat{y} - \hat{b}\bar{x}$$

where

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$$

The section additionally calculates the correlation coefficient of the calibration line according to the following relationship:

$$r = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{N} (x_i - \bar{x})^2\right]^{1/2} \left[\sum_{i=1}^{N} (y_i - \bar{y})^2\right]^{1/2}}$$

The output of section CALIN contains the estimated value of parameters $a$ and $b$ and the correlation coefficient.

## A6—4.3 Interval estimation of the unknown component

Section CALIN is suitable for estimating the unknown quantity of material from measurements on its analytical property using the calibration line.

The subprogram first determines the mean of measured values of the analytical property, $y_1, \ldots, y_L$

$$\hat{\bar{y}} = \frac{\sum_{l=1}^{L} y_l}{L}$$

and uses this value to calculate an estimate for the quantity of material according to

$$\hat{x} = \frac{\hat{\bar{y}} - \hat{a}}{\hat{b}}$$

Section CALIN assumes that the error effecting the analytical property $y_i$ is normally distributed. In this case the program can determine an interval estimation for unknown quantity $x$ at a confidence level given by input. The interval estimation is

$$P(L \leq x \leq U) = 1 - \alpha$$

where $L$ is the lower and $U$ the upper value of the interval defined as:

$$L = \hat{x} - \Delta$$

$$U = \hat{x} + \Delta$$

where

$$\Delta = t_{1-\alpha, N-2} \cdot \frac{S}{\hat{b}} \sqrt{\frac{1}{N} + \frac{1}{L} + \frac{(\hat{x} - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}}$$

$$s = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{a} - \hat{b}x_i)^2}{N-2}}$$

and $t$ is the value of the $t$ (Student)-distribution at level $1-\alpha$ and for $N-2$ degrees of freedom.

The output of section CALIN for analysis contains the estimated quantity $\hat{x}$ of unknown material and values $L$ and $U$ of the confidence interval at the given confidence level.

## A6—5 PROGRAM USE

The input is menu driven and many options are available to the user. In the interactive operation the prompts clearly define the way in which the input data are to be presented. In batch operation the prompts must be anticipated. This input sequence can be extracted from *Figures A6—1 to A6—5*. The output examples (Section 6) further clarify the input requirement.

*Input specification sheets*

First item: Select Section 1 to 3

*Inputs for section CHAR* (See *Figs A6—1* to *A6—3*)

Number of replicate measurements (N)
Set of measurements (A($i$))
Type of statistical method (MST 1)
Type of parameter to be estimated (IMOD)
Known systematic error
Known true value
Confidence level for normality test ($\alpha$)
Confidence level for interval estimation ($\alpha$)
Confidence level for hypothesis testing ($\alpha$)
Presumed deviation
Presumed true-value
Presumed systematic error

*Inputs for section COMP* (See *Fig. A6—4*)
(For first data set)

Number of replicate measurements (N)
Set of measurements (A($i$))
Confidence level for normality test ($\alpha$)
(For second data set)

Number of replicate measurements (M)
Set of measurements (B($i$))
Type of parameter to be compared (MST2)
Confidence level for hypothesis testing ($\alpha$)

Presumed systematic error of first data set
Presumed systematic error of second data set
Presumed true value of first data set
Presumed true value of second data set

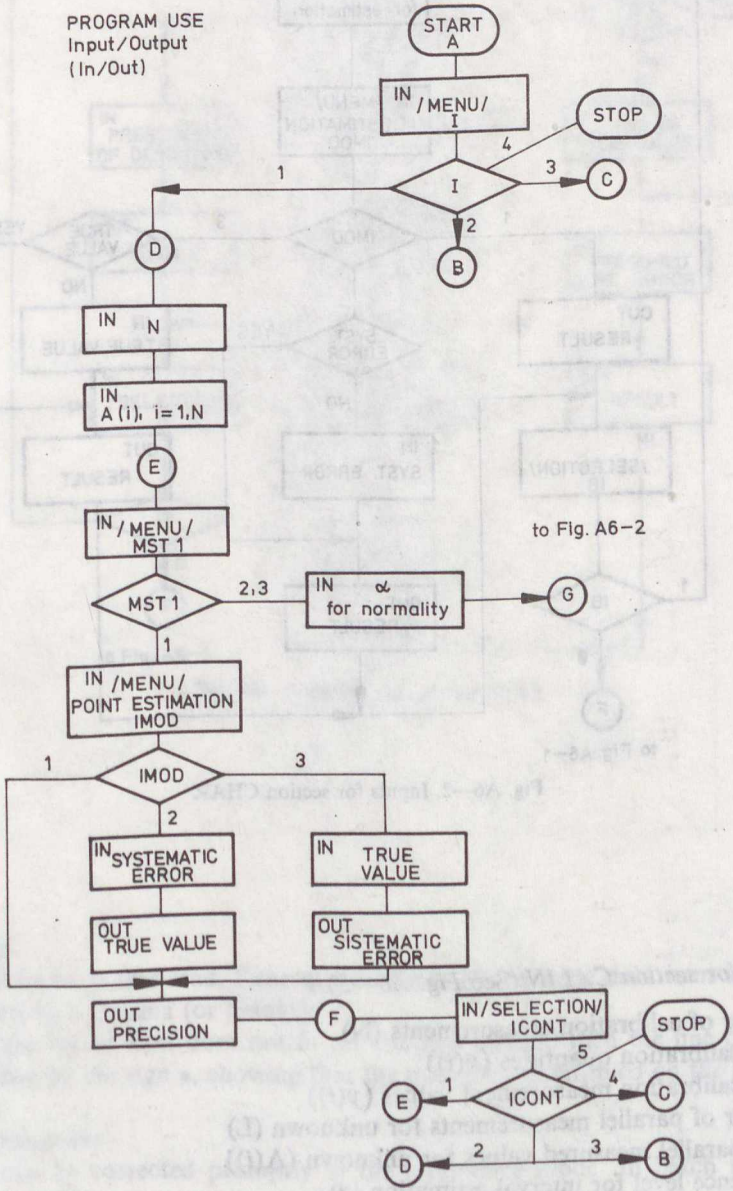



Fig. A6—1. Inputs for section CHAR

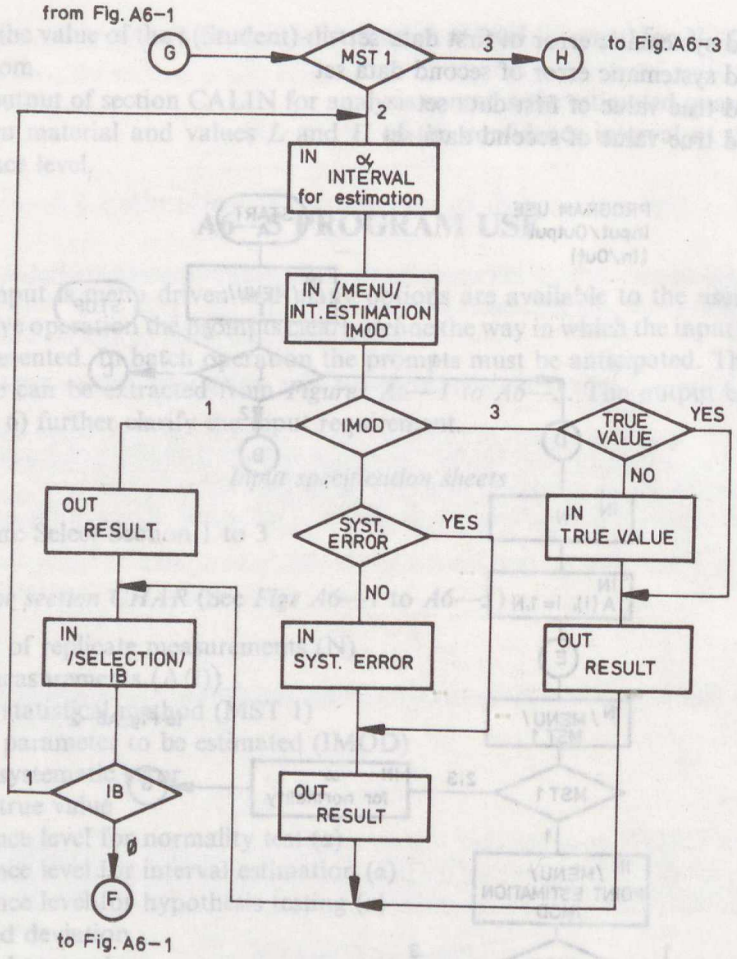



Fig. A6—2. Inputs for section CHAR

*Inputs for section CALIN* (See *Fig. A6—5*)

Number of calibration measurements (N)
Set of calibration quantities ($x(i)$)
Set of calibration measurement values ($y(i)$)
Number of parallel measurements for unknown (L)
Set of parallel measured values for unknown (A($i$))
Confidence level for interval estimation ($\alpha$)

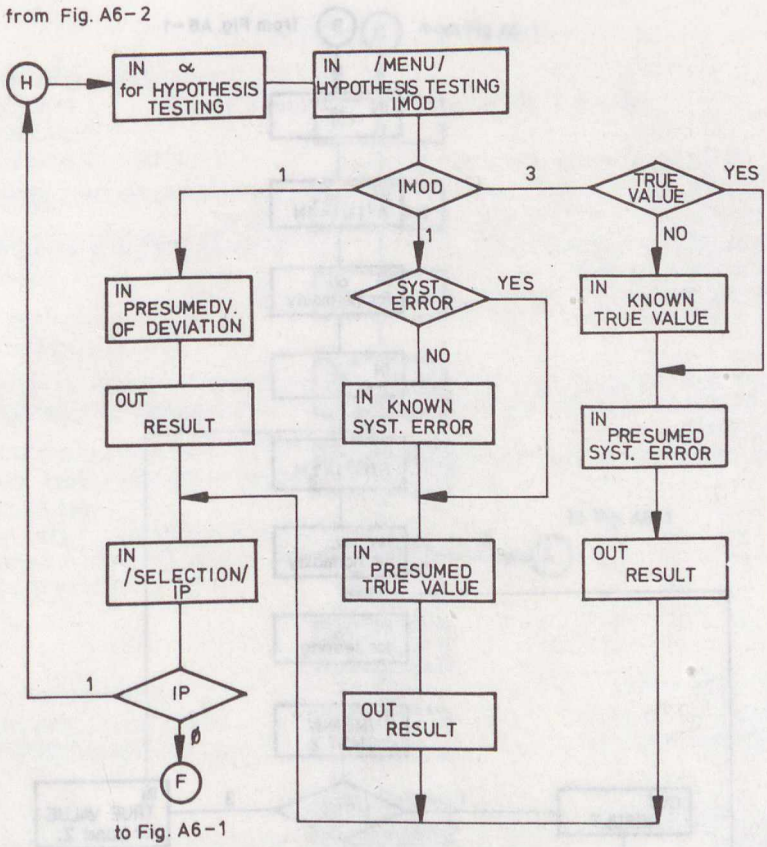



Fig. A6—3. Inputs for section CHAR

*Notes*:

1. All input is free-read. Subsequent values of the same data item should be separated by a comma (or blanks).

2. If the list of data does not fit on one line or card then the line or card is terminated by the sign *, showing that the data list is continued on the next card or line.

3. *Error codes*

Errors can be corrected promptly in the interactive mode. In batch mode the execution of the program stops with a display of the error message.

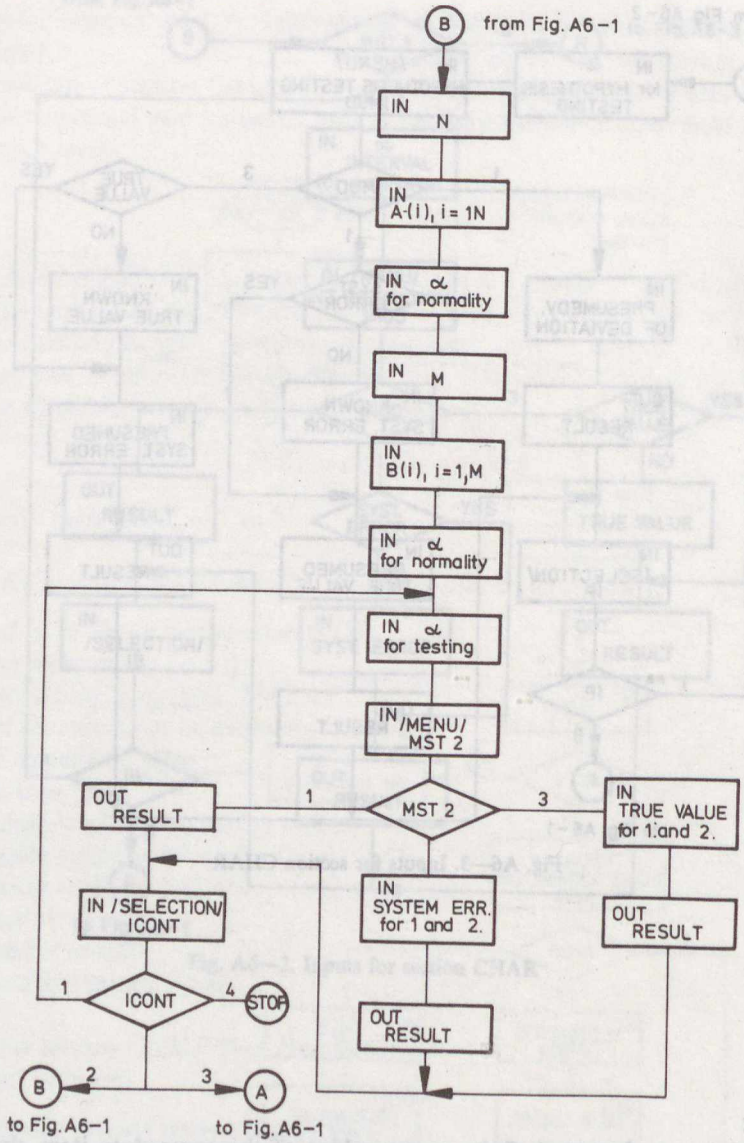



Fig. A6—4. Inputs for section COMP

The program indicates error in the following cases:

1. Error in menu code
2. if (a) data number exceeds 500,

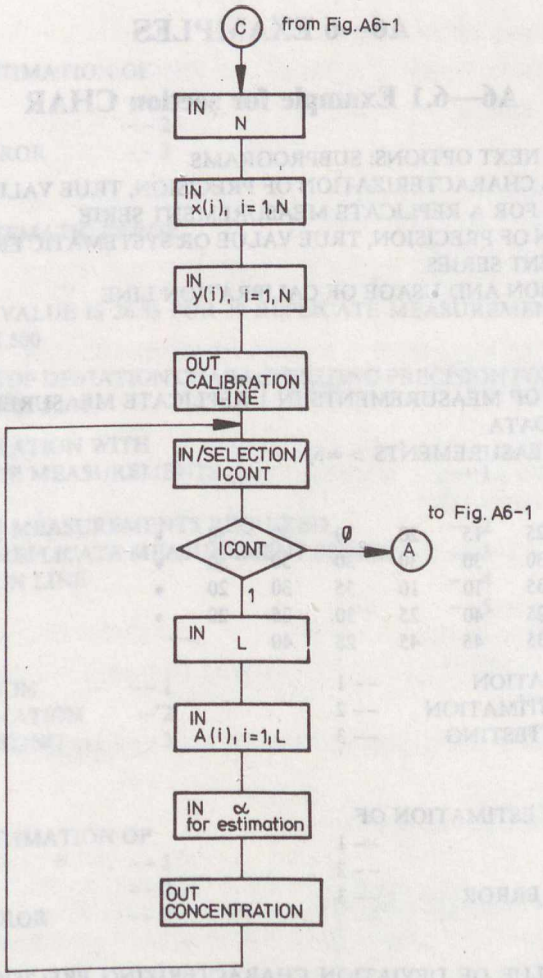



Fig. A6—5. Inputs for section CALIN

(b) data number is smaller than 5 for subprogram CHAR and COMP,

(c) data number is smaller than 3 for subprogram CALIN.

3. Significance level is out of the range 0.–0.999.

# A6—6 EXAMPLES

## A6—6.1 Example for section CHAR

```
CHOOSE FROM NEXT OPTIONS: SUBPROGRAMS
— STATISTICAL CHARACTERIZATION OF PRECISION, TRUE VALUE OR SYSTEM-
  ATIC ERROR FOR A REPLICATE MEASUREMENT SERIE                    -- 1
— COMPARISON OF PRECISION, TRUE VALUE OR SYSTEMATIC ERROR FOR TWO
  MEASUREMENT SERIES                                              -- 2
— CONSTRUCTION AND USAGE OF CALIBRATION LINE                      -- 3
— END                                                             -- 4
1

TYPE NUMBER OF MEASUREMENTS IN 1 REPLICATE MEASUREMENT SERIES AND
  MEASURED DATA
(NUMBER OF MEASUREMENTS > =5)
39
   30   25   25   15   20   30   40   30   *
   40   25   30   30   30   30   30   30   *
   30   25   35   10   10   35   30   20   *
   30   25   25   40   25   30   25   20   *
   25   15   35   45   45   25   40
— POINT ESTIMATION                -- 1
— INTERVAL ESTIMATION             -- 2
— HYPOTHESIS TESTING              -- 3
1

WANTED POINT ESTIMATION OF
— PRECISION                       -- 1
— TRUE VALUE                      -- 2
— SYSTEMATIC ERROR                -- 3
1

ESTIMATED VALUE OF DEVIATION CHARACTERIZING PRECISION FOR 39 REPLI-
  CATE MEASUREMENTS IS 8.298

CONTINUE CALCULATION WITH
— THESE REPLICATE MEASUREMENTS                                    -- 1
SUBPROGRAMS:
— NEW REPLICATE MEASUREMENTS REQUIRED                             -- 2
— COMPARE TWO REPLICATE MEASUREMENT SERIES                        -- 3
— USE CALIBRATION LINE                                            -- 4
— END                                                             -- 5
1
— POINT ESTIMATION                -- 1
— INTERVAL ESTIMATION             -- 2
— HYPOTHESIS TESTING              -- 3
```

```
1

WANTED POINT ESTIMATION OF
— PRECISION          -- 1
— TRUE VALUE         -- 2
— SYSTEMATIC ERROR   -- 3

2

TYPE KNOWN SYSTEMATIC ERROR

1.5

ESTIMATED TRUE VALUE IS 26.83 FOR 39 REPLICATE MEASUREMENTS IF SYSTEM-
  ATIC ERROR IS 1.500

ESTIMATED VALUE OF DEVIATION CHARACTERIZING PRECISION FOR 39 REPLICATE
  MEASUREMENTS IS 8.298

CONTINUE CALCULATION WITH
— THESE REPLICATE MEASUREMENTS                    -- 1
SUBPROGRAMS:
— NEW REPLICATE MEASUREMENTS REQUIRED             -- 2
— COMPARE TWO REPLICATE MEASUREMENT SERIES        -- 3
— USE CALIBRATION LINE                            -- 4
— END                                             -- 5

1

— POINT ESTIMATION        -- 1
— INTERVAL ESTIMATION     -- 2
— HYPOTHESIS TESTING      -- 3

1

WANTED POINT ESTIMATION OF
— PRECISION          -- 1
— TRUE VALUE         -- 2
— SYSTEMATIC ERROR   -- 3

3

TYPE KNOWN VALUE

30.

ESTIMATED SYSTEMATIC ERROR IS -1.667 FOR 39 REPLICATE MEASUREMENTS IF
  TRUE VALUE IS 30.00

ESTIMATED VALUE OF DEVIATION CHARACTERIZING PRECISION FOR 39 REPLICATE
  MEASUREMENTS IS 8.298

OTHER ESTIMATION REQUIRED ON MEASURED DATA? YES — 1, NO — 0

1

TYPE CONFIDENCE LEVEL (E.G. 0.95)

0.95
```

```
WANTED INTERVAL ESTIMATION OF
— PRECISION                    -- 1
— TRUE VALUE                   -- 2
— SYSTEMATIC ERROR             -- 3

3

INTERVAL ESTIMATION OF SYSTEMATIC ERROR IS -2.357<==-1.667<==1.025 FOR 39
   MEASUREMENTS AT LEVEL 0.950
   IF TRUE VALUE IS 30.00

OTHER ESTIMATION REQUIRED ON MEASURED DATA? YES — 1, NO — 0
1

CONTINUE CALCULATION WITH
— THESE REPLICATE MEASUREMENTS                    -- 1
SUBPROGRAMS:
— NEW REPLICATE MEASUREMENTS REQUIRED             -- 2
— COMPARE TWO REPLICATE MEASUREMENT SERIES         -- 3
— USE CALIBRATION LINE                            -- 4
— END                                             -- 5

1

— POINT ESTIMATION             -- 1
— INTERVAL ESTIMATION          -- 2
— HYPOTHESIS TESTING           -- 3

3

TYPE CONFIDENCE LEVEL FOR NORMALITY TEST OF MEASURED DATA (E.G. 0.95)

0.95

TYPE CONFIDENCE LEVEL (E.G. 0.95)

0.95

WANTED HYPOTHESIS TESTING OF
— PRECISION                    -- 1
— TRUE VALUE                   -- 2
— SYSTEMATIC ERROR             -- 3

1

TYPE PRESUMED VALUE OF DEVIATION

6.55
```

## A6—6.2 Example for section COMP

```
CHOOSE FROM NEXT OPTIONS: SUBPROGRAMS
— STATISTICAL CHARACTERIZATION OF PRECISION, TRUE VALUE OR SYSTEM-
   ATIC ERROR FOR A REPLICATE MEASUREMENT SERIE                -- 1
— COMPARISON OF PRECISION, TRUE VALUE OR SYSTEMATIC ERROR FOR TWO
   MEASUREMENT SERIES                                          -- 2
— CONSTRUCTION AND USAGE OF CALIBRATION LINE                   -- 3
— END                                                          -- 4
```

```
2
TYPE NUMBER OF MEASUREMENTS IN 1 REPLICATE MEASUREMENT SERIES AND
  MEASURED DATA
(NUMBER OF MEASUREMENTS > = 5)
16
   38   43   44   44   46   51   50   49   *
   44   53   51   55   55   57   54   62
TYPE CONFIDENCE LEVEL FOR NORMALITY TEST OF MEASURED DATA (E.G. 0.95)
0.95
TYPE NUMBER OF MEASUREMENTS IN 2 REPLICATE MEASUREMENT SERIES AND
  MEASURED DATA
(NUMBER OF MEASUREMENTS > = 5)
16
   38   45   46   46   48   55   54   53   *
   53   57   55   61   61   63   60   70
TYPE CONFIDENCE LEVEL FOR NORMALITY TEST OF MEASURED DATA (E.G. 0.95)
0.95
TYPE CONFIDENCE LEVEL (E.G. 0.95)
WANTED COMPARISON OF TWO DATA SETS
— PRECISION                    -- 1
— TRUE VALUE                   -- 2
— BIAS (SYSTEMATIC ERROR)      -- 3
1
VALUE OF F TEST 1.781
TABULATED VALUES  0.3487    2.867
ESTIMATED PRECISION OF DATA SET 1. FROM 16 DATA 6.071
ESTIMATED PRECISION OF DATA SET 2. FROM 16 DATA 8.103
EQUALITY OF PRECISION AT LEVEL 0.950 IS ACCEPTABLE
PLEASE TYPE KNOWN TRUE VALUE OF DATA SET 2
53
VALUE OF 2-SAMPLE T-TEST 0.3950
TABULATED VALUE 2.042
ESTIMATED SYSTEMATIC ERROR OF DATA SET 1. FROM 16 DATA 0.6250 E-01
ESTIMATED SYSTEMATIC ERROR OF DATA SET 2. FROM 16 DATA 1.063
EQUALITY OF SYSTEMATIC ERRORS AT LEVEL 0.950 IS ACCEPTABLE
CALIBRATION LINE NEEDED? YES — 1, NO — 0
1
TYPE NUMBER OF PARALLEL MEASUREMENTS FOR THE UNKNOWN (L)
TYPE MEASURED VALUES FOR THE UNKNOWN (L)
TYPE CONFIDENCE LEVEL OF INTERVAL ESTIMATION FOR
  UNKNOWN CONCENTRATION/QUANTITY
```

```
3
  16.1   16.2   15.8
0.95
POINT ESTIMATION FOR CONCENTRATION/QUANTITY OF UNKNOWN 7.881
INTERVAL ESTIMATION FOR CONCENTRATION/QUANTITY OF UNKNOWN
   7.682    <=    7.881    <=    8.085
FROM 3 PARALLEL MEASUREMENTS AT LEVEL 0.950
CALIBRATION LINE NEEDED? YES — 1, NO — 0
0
CHOOSE FROM NEXT OPTIONS: SUBPROGRAMS
— STATISTICAL CHARACTERIZATION OF PRECISION, TRUE VALUE OR SYSTEM-
  ATIC ERROR FOR A REPLICATE MEASUREMENT SERIE          -- 1
— COMPARISON OF PRECISION, TRUE VALUE OR SYSTEMATIC ERROR FOR TWO
  MEASUREMENT SERIES                                    -- 2
— CONSTRUCTION AND USAGE OF CALIBRATION LINE            -- 3
— END                                                   -- 4
4
```

## A6—7 NOTES ON PROGRAM IMPLEMENTATION

The program has been designed to work in three modes. These are the following:

— BATCH mode (0)
— INTERACTIVE mode (1)
— INTERACTIVE + PRINTER (2)

The most usual way of running the program is in mode 0. The student has to prepare all the input data and control information necessary to run the program before running the program. The program operates on this prepared data without the intervention of the student and provides the result or the error messages at the end of the run. In case of a fatal error occurring during program execution, the program stops and displays an error message. To correct or modify the input data the student has to revise the data set and run the program again to obtain new results.

The usual input device for the prepared data is a card reader, the output device is usually a printer. However, either could equally be a disk file.

In mode 1, the student has a terminal, with which he keys in the input data, and on which he gets back the results. The program works interactively, i. e. keying in the data and return of the results are carried out by a man-machine dialogue. Correcting the faults, or modifying data can be done immediately. Prompt texts and immediate error checks help the student present the input data correctly.

Functionally mode 2 is the same as mode 1 except that a printer is also available as well as the terminal. In this mode results appear on the printer as well as on the terminal.

The most comprehensive mode is 2. The other two modes have been included to enable the programs to be used when hardware capacity is limited.

The mode is selected by activating one of the three lines in the BLOCK DATA segment.

```
DATA IR(5), IW(6), IPMT(6), MODE(0)
DATA IR(5), IW(5), IPMT(5), MODE(1)
DATA IR(5), IW(7), IPMT(5), MODE(2)
```

IR, IW and IPMT are the unit identifiers for input, output and prompt peripherals respectively. The first line belongs to the batch mode, the second to the interactive mode, and the third to the interactive + printer mode. From the three lines let the chosen one remain active, and transform the other two into comments by writing a letter C in the first position on the line.

If you want to use the program in another mode, change the position of the letters C and recompile BLOCK DATA.

The total coding of programs STATCHAR extends to some 3500 FORTRAN statements.

The program and its subroutines are written in standard FORTRAN IV.

# SUBJECT INDEX


abstract object 3
actuator 285
adaptive control 378
aggregate 7
analytical chemical
— correlation 391
— sign process 391
— systems 388, 469
annual cost 191, 192
assessing profitability 209

badly posed problem 147—151
basic chemical technological connection 17
— serial 17
biased estimators 39
block diagram 286, 328
Bode plot 289
break-even point 214
bundle 6

calibration line 458, 608
capacity process 303
CAPCOS program 264
capital
— cost calculation 195
— estimation 253
— investment 231
cascade connection 19
characteristic
— polynominal 329
— roots 329
classification of elements according to
— effect of element 16
— number of input-output 13
— state relation 15
— time dependence 15
— type of energy transformation 16
— type of equation 15
closed loop
— frequency functions 367
— transfer functions 367
component 104—106, 108, 110
conditional stability 366
confidence
— ellipsoid 37
— region 37
connectivity 100
— equation 96
— strong 99, 102, 104, 119
control
— signal 284
— valve 285
controller 284
— tuning 372
convolution 410
cost estimation 250, 255
credibility 479
critical frequency 290
critically damped system 351
curve fitting 413
cycle 99, 111, 119, 122
cycling method 313

dead time 292, 357
degree of freedom 97, 125—147
— condenser 142
— distillation column 139
— flash drum 132
— flowsheet 144
— gas compressor 133
— heat exchanger 136
— isothermal-isobaric flash separator 131
— mixer 128
— multicomponent split unit 138
— pressure relief valve 132
— process furnace 135

degree of freedom (cont.)
— pump 132
— reboiler 141
— simple reactor 136
— stream 125
— stream splitter 130
— unit 128
density function-like signals 427
depreciation 197
depth-first search 110, 111
derivative element 310
deterministic indentifiability 52
direct integral
— least-squares 63
— methods 62
— distributed parameter system 315, 359

eigenvalue-eigenvector analysis 77
elements in
— parallel 287
— series 287
elements of the system 11
equation-oriented approach 97, 102, 104, 105, 161—167
equivalent max. investment period (*EMIP*) 215
error 152
— absolute 152
— relative 152
— root mean square 152
expansion in partial fractions 325
extent of nonlinearity 46

factor analysis 418
feedback 280
— connection 20
feedforward 280
final value theorem 321, 370
first-order
— element 341
— lag 320
fixed-capital investment 229
flowsheeting 124
flowsheeting systems 159—181
— ASCEND-II 165
— ASPEN 167
— CAPES 176—178
— CHESS 161, 167—174
— CONCEPT 161, 185
— FLOWPACK 161
— FLOWTRAN 161
— FORSIM 161
— GENDER 164, 165
— MASTEP 167
— PACER 161
— PROCESS 161
— QUASILIN 164, 165, 179—185
— SIMUL 161
— SPEEDUP 164, 165
— UNICORN 161
Fourier transformation 401
frequency function 287, 335

gain
— crossover frequency 365
— margin 365
Gauss–Marquardt estimation algorithm 42
goodness of the fit 36

Haar transformation 408
hierarchy levels 23
— complex operational unit 23
— composite operational unit 23
— simple operational unit 23
— steps of technological process 23

identification 316
impulse 316
indegree 99
indirect least squares 51
inflation 230
information theory 462
initial value theorem 321
input-output pairs 5
input-output relation of 7
— a branch 12
— chemical transformer 13
— mixer 11
— separation 12
— technological operator in general 9
input-output equation state 7
instrumental variable method 65
integral controller 299
interacting multicapacity system 353
inversion of Laplace transforms 325
iteration
— Broyden's method 159, 163
— direct (substitution) 120, 121, 153, 156, 163
— Newton's method 157, 163
— quasi-Newton methods 164
— secant method 154—157, 163
— Wegstein's method, see secant method

labelling 110, 118
Laplace transformation 319
least squares method 35
line 98
linearization 44, 317
— transformations 46
lumped parameter system 315

matching 103, 106, 116
mathematical system theory 2
matrix
— adjacency 100, 106, 109
— full 110
— identity 101
— index 101
— Jacobian 147, 153—155, 158, 159, 163
— reachability 107, 109
— symmetric 100
measurement error 449
MIMO system 283
model reduction 77
modular technique 241
module 97, 119, 120
moments 433
multicapacity systems 353
multiple correction coefficient 36
multi-response
— estimation 49
— model 48

near-singularity 64
nearly-singular estimation problems 82
negative feedback control loop 287
network 98, 102
— acyclic 99, 104, 105, 120
— bipartite 106, 116
— directed 99
— of components 104
— simple 98
— undirected 99, 103, 116
networking 96
noninteracting multicapacity systems 353
Nyquist
— plot 289
— stability criterion 290, 366

object $A$ 2, 5
— linear 5
— physical 3
optimal control 378
optimum controller 313
ordered segment pairs 3
outdegree 99
output assignment 103, 105, 106, 116
output-input relations 3
overdamped system 350

parallel
— connection 20
— lines 98
parameter sensitivity 51
partitioning 98, 102, 116, 163
path 99, 103, 107
— augmenting 116—119
— directed 99, 118
pattern recognition 435
PD-controller 311
performance
— characteristics of measuring systems 445
— criteria 372
phase margin 365
PI-controller 306
PID-controller 311
point 98
Poisson weighting 58
precedence structure 102, 105
precision 453, 456, 603—606
principal component
— analysis 77
— estimator 40
process
— line 25
— reaction curve 377
profitability 190, 201
— calculations 217
proportional controller 295
pulse 316

quality assurance systems 484
quarter-amplitude damping 292
quasi-steady-state assumption 85

ramp 316
random error 450, 455
ratio methods 234
recirculation connection 21
reparametrization 46
replacement rule 121—123
residual variance 36
ridge regression estimator 40
Routh–Hurwitz stability criterion 332

sales revenue 226
scaling 39
second-order system 350
segment 3
selfloop 98
sensitivity analysis 225
sequential–modular approach 98, 102, 104, 105, 119, 161—167
serial–bypass connection 20
set point 284
settings 313
signal transformations 397
simultaneous–modular approach 161, 166, 167
singular value decomposition 37
SISO system 283
smoothing 411
sparse system 97, 103
stability 281, 329
— limit 290
state
— equation 380
— of a system 7, 379
— space method 379
— variable 379
statistical decision theory 471
steady state 124, 159
— offset 298, 370
step 316
stream 125
— inlet 125
— outlet 125
symmetric bootstrap (*SB*) estimator 65
subnetwork 99, 104
system, general definition 2
system of factories and combines 25
systematic errors 450, 453, 455, 456, 602, 604, 605, 607

tear set 119, 163
— canonical 123
— family (of tear sets) 122
— — nonredundant 122, 124
— nonredundant 122, 123
— redundant 122
— relatives 122
tearing 98, 119
technological operators 9
time-shift theorem 321
transducer 283
transfer function 319
transformation matrix of the technological operator 10
transmitters 284
true value 451, 454, 458, 602, 604, 605
tuning 313

unbiased estimator 36
underdamped system 352
unit cell 96
utilities costs 258

variable
— decision 97, 124
— design 97, 124
— iterated 119

walk 99
— directed 99
Walsh transformation 407

Bibliotheca Academiae Scientiarum Hungaricae
MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

The development, design and control of the technology of processes in chemical industry demands knowledge of the methods necessary for modelling, measurement, model-identification and control. The aim of the book is to facilitate orientation in this complex field by surveying and summarizing the most important chemical, technological mathematical, control and measurement methods.

By the application of the classical computer methods and by its unique, novel approach, the book provides significant help in orientation among the numerous methods necessary in the design and control of chemical engineering processes.

The book is a useful tool not only for experts interested in development and design but it can also be used efficiently in teaching.

5079545
5B

PALLAI
VERESS
(Editors)

# STUDIES IN COMPUTER-AIDED MODELLING, DESIGN AND OPERATION