tanulmányok

195/1987

MTA Számítástechnikai és Automatizálási Kutató Intézet Budapest

MAGYAR TUDOMÁNYOS AKADÉMIA
SZÁMITÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZETE

BINÁRIS VÁLTOZÓK STRUKTÚRÁJÁNAK VIZSGÁLATA

TELEGDI LÁSZLÓ

Tanulmányok 195/1987

A kiadásért felelős:

DR. KEVICZKY LÁSZLÓ

Főosztályvezető:

Dr. MAROS ISTVÁN

ISBN 963 311 224 9
ISSN 0324-2951

Készült: az Alfaprint Nyomdaip.Kisszöv.-ben
Munkaszám: 243/1987.
Felelős vezető: Barabás Gábor

ELŐSZÓ

A tanulmány olyan bináris (valószinüségi) változók strukturájának statisztikai vizsgálatával foglalkozik, amelyek dichotom jellegek indikátor változói. A változókat nagyszámu objektumon figyeljük meg. Vizsgálatuk azon az adatmezőn alapul, amelyik megadja, hogy az egyes objektumok mely jellegekkel rendelkeznek. Tetszőleges objektumot az $\underline{A}_1, \underline{A}_2, \ldots, \underline{A}_n$ jellegek egy részhalmaza jellemez. Az objektumok kisérletek kimenetelének is tekinthetőek, a jellegek (pontosabban a meglétük) pedig eseményeknek, amelyek a kisérletek során vagy bekövetkeztek, vagy nem. Ilymódon a változók strukturájának vizsgálata ezen események egymás közti kapcsolatainak vizsgálatát jelenti. Ehhez szükséges a 2^n számu $\underline{A}_{i_1}\underline{A}_{i_2}\ldots\underline{A}_{i_k}$ esemény valószinüségének meghatározása. Ezt a feladatot 2^n már tizes nagyságrendü $\underline{n}$ mellett is igen nagy volta teszi bonyolulttá.

Legegyszerübb dolgunk akkor lenne, ha az $\underline{A}_i$ események - tehát a jellegek, ill. a változók - (teljesen) függetlenek lennének: ekkor tetszőleges $\underline{A}_{i_1}\underline{A}_{i_2}\ldots\underline{A}_{i_k}$ esemény valószinüsége az $\underline{A}_{i_1}, \underline{A}_{i_2}, \ldots, \underline{A}_{i_k}$ események valószinüségeinek szorzata. A tanulmány 2. fejezete (az 1. fejezet a Bevezetés) a változók függetlenségvizsgálatával foglalkozik, amelyet nagyobb $\underline{n}$ és/vagy kis $\underline{P}(\underline{A}_i)$-k

esetén nem könnyü elvégezni.

Ha a változók nem függetlenek, jellemeznünk kell függőségüket. A többdimenziós normális eloszlás egyszerüsége és gyakori előfordulása adta az ötletet, hogy vizsgáljan a normális küszöb modellt. Ez feltételezi, hogy a jellegeknek mindegyik objektumra nézve van egy-egy valós számmal kifejezhető mértéke, vagyis az i-edik (bináris) változóhoz hozzá van rendelve egy L_i háttér változó. A modell szerint ezek együttes normális eloszlásuak, egy-egy küszöb tartozik hozzájuk, továbbá egy objektum akkor és csak akkor rendelkezik az A_i jellegel, ha L_i rajta felvett értéke eléri a megfelelő küszöböt. A 3. fejezet ezzel a modellel foglalkozik.

A normális küszöb modellnek a vázolt feladatban van egy hátránya: feltételezése esetén igen nehézkes sajátos jellegkombinációk és ilyenekkel jellemzett objektumcsoportok meghatározása, az objektumok klaszteranalizise. A 4. fejezet a (dichotom jellegekkel, ill. bináris változókkal jellemzett) objektumok klaszteranalizisére kidolgozott uj eljárással, az un. többszörös többdimenziós skálázással (többszörös MDS, MMDS) foglalkozik, ami a változók (közönséges) MDS-e révén sorolja klaszterekbe az objektumokat. A változók (közönséges) MDS-e azt a problémát vizsgálja, hogy a változók között értelmezett távolságok alapján hogyan lehet a változókat megjeleniteni valamely R^k alacsony dimen-

ziós euklideszi térben, vagyis hogyan lehet R^k-ban olyan pont-n-est konstruálni hozzájuk, hogy a pontok euklideszi távolsága minél jobban tükrözze a változók távolságát. Ahhoz, hogy a változók jól skálázhatóak legyenek, konzisztenseknek kell lenniük a következő értelemben: ha két változó "közel" van egy harmadikhoz, egymáshoz is "közel" kell lenniük. Az MDS, amelyet a tanulmány legjelentősebb és legszélesebb körben használható eredményének tartok, azt a problémát vizsgálja, amelyik akkor merül fel, ha a fenti konzisztencia nem teljesül: hogyan lehet az objektumokat minél homogénebb klaszterekbe sorolni, amikoris egy-egy klaszter homogenitását a változók ezen klaszter mellett történő MDS-ének jóságával mérjük?

A tanulmány több szempontból is alkalmazott matematikai. i) A (matematikai) statisztika és a klaszteranalizis ilyen diszciplinák. - ii) Nem egy diszciplina egy fejezetével foglalkozom, hanem egy - általánosan megfogalmazott - feladat gyakorlatban történő megoldása során fellépő problémákkal. Irodalmi áttekintést ezért csak a többdimenziós skálázáshoz adok, ahhoz sem a Bevezetésben. - iii) A tanulmány 5. fejezete a számítógépek egyik fontos orvosbiológiai alkalmazásával, a veleszületett rendellenességek statisztikai vizsgálatával foglalkozik. Hangsulyozni szeretném azonban, hogy itt többről van szó, mint a 2-4. fejezetben is-

mertetésre kerülő matematikai eredmények alkalmazásáról egy konkrét feladatban: az eredmények jelentős része (igy például az MMDS is) ezen konkrét feladatból nőtt ki. A rendelkezésre álló adatok mennyisége indokolja a számitógépes feldolgozást. Az 5. fejezet eredményei is tanusitják, hogy a felhasznált uj módszerek segitségével uj következtetéseket lehet levonni.

A tanulmányban kétszeres aláhuzással jelzem a vektorokat és mátrixokat. Az egyszeres aláhuzás semmit sem jelent, célja csupán a szövegből való kiemelés.

Az a kutató munka, melynek során ismertetésre kerülő eredményeim megszülettek, része volt annak a tevékenységnek, amelyet Czeizel Endrével, az orvostudomány doktorával és Tusnády Gáborral, a matematikai tudomány kandidátusával több mint egy évtizede, Simonovits Miklóssal, a matematikai tudomány doktorával és Dávidné dr. Bolla Marianna matematikus kollégámmal körülbelül fél évtizede végzünk. Az együttes munka során mindegyiküktől sok segitséget kaptam, de különösen Tusnády Gábortól. A tanulmány alapjául szolgáló kandidátusi értekezés munkahelyi birálatáért ugyancsak Tusnády Gábornak, valamint Hollósné dr. Marosi Judit matematikusnak, opponálásáért néhai Sarkadi Károlynak, a matematikai tudomány doktorának, valamint Csirik Jánosnak és Michaletzky Györgynek, a matematikai tudomány kandidátusainak tartozom hálával.

Az értekezés és a tanulmány az MTA SZTAKI Alkalmazott Matematikai Főosztályának Statisztika Osztályán készült; vezetőimtől, különösen Prékopa Andrástól, az MTA rendes tagjától minden támogatást megkaptam a munka elvégzéséhez. A tanulmány eltérése a (sikeresen megvédett) értekezéstől három forrásból származik. i) Elhagytam az értekezés részletes futási eredményeket tartalmazó Függelékét. - ii) Figyelembe vettem az opponensi véleményeket és ezekre adott válaszaimat. - iii) Beépitettem az értekezés beadása óta elért eredményeimet.

TARTALOMJEGYZÉK

1. BEVEZETÉS

Attól függően, hogy milyen tulajdonságra vonatkozik, a valószínüségi változónak több tipusa van. Kiemelkedő fontosságu a következő kettő:

a) Kvantitativ (mennyiségi) változó. Mérhető és ennek következtében mértékszámokkal is kifejezhető tulajdonságra, mennyiségre vonatkozik. Lehetséges megfigyelési értékei valós számok. Ha ezek folytonos sokaságot képeznek, a változó folytonos, ha megszámlálhatóan sokan vannak, a változó diszkrét.

b) Kvalitativ (minőségi, kategórikus) változó. Nem mérhető és igy mértékszámokkal ki sem fejezhető tulajdonságra vonatkozik. Lehetséges megfigyelési értékei egymástól minőségileg különböző osztályok (kategóriák).

Az olyan diszkrét változót, amelyik csak az 1 és 0 értékeket veheti fel, bináris változónak, az olyan kvalitativ változót, amelyiknek csak két különböző megfigyelési értéke fordulhat elő (igen-nem, férfi-nő, élő-nem élő, beteg-egészséges stb.), dichotom (alternativ) változónak nevezik. A dichotom változó két lehetséges megfigyelési értéke közül az egyik sok esetben ki van tüntetve, és valamely jelleg (pl. betegség) meglétét jelenti. Az ilyen dichotom változóhoz természetes módon hozzárendelhető egy bináris változó: a jelleg (meg-

létének mint eseménynek az) indikátor változója. Ilyen változók képezik értekezésem tárgyát. Olyan bináris változók strukturájának vizsgálatával foglalkozom tehát, amelyek dichotom jellegekre vonatkoznak. Jelölje ezek számát n, a változókat $W_1, W_2, \ldots, W_n$, a jellegeket $A_1, A_2, \ldots, A_n$, a megfigyelések számát M, azokat az objektumokat pedig, amelyeken a változókat megfigyeljük, $Y_1, Y_2, \ldots, Y_M$.

Az 1.1. paragrafus az adatok számitógépes tárolásával foglalkozik. Ezt illetően kézenfekvő a következő két követelmény: i) az adatok a memória ne tul nagy részét foglalják el; ii) tetszőleges

$$G_g = \{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\} \qquad (1.1)$$

jellegkombinációhoz azon objektumok

$$\sigma(G_g) = \sigma(i_1, i_2, \ldots, i_k)$$

és

$$\sigma_T(G_g) = \sigma_T(i_1, i_2, \ldots, i_k)$$

száma, amelyek rendelkeznek a G_g-hez tartozó jellegekkel, de más jelleggel nem, ill. amelyek rendelkeznek a G_g-hez tartozó jellegekkel és esetleg továbbiakkal is, gyorsan kikereshető legyen. E célból a fenti gyakoriságokat és címüket, ill. az értékükre vonatkozó infor-

mációt egy egyindexes NS egész tömbben helyezem el oly módon, hogy NS-ből tetszőleges k elemü, más szóval k méretü G_g jellegkombináció esetén az $\sigma(G_g)$, $\sigma_T(G_g)$ értékek legfeljebb (k+1) lépésben meghatározhatóak.

A 2. fejezet a W_i változók függetlenségvizsgálatával foglalkozik. Tetszőleges (1.1) alaku G_g jellegkombináció esetén jelölje

$$P_T(G_g) = P_T(i_1,i_2,\ldots,i_k)$$

annak valószinüségét, hogy egy objektum rendelkezik a G_g-hez tartozó jellegekkel és esetleg továbbiakkal is. Az a kérdés, hogy a W_i változók (teljesen) függetlennek tekinthetőek-e, elméletileg a valamennyi legalább 2 méretü, (1.1) alaku jellegkombinációra

$$P_T(i_1,i_2,\ldots,i_k) = \prod_{j=1}^{k} P_T(i_j)$$

nullhipotézishez tartozó χ^2-próbával dönthető el, a megfelelő 2^n-es kontingenciatáblázat alapján. A számomra érdekes esetekben azonban - még a ritka jellegek lehetőség szerinti összevonása után is - 2^n nagysága és/vagy a $P_T(i)$ valószinüségek kicsisége miatt a χ^2-próba elvégzése illuzórikus, ezért máshogyan kell a függetlenséget

ellenőrizni.

Jelölje $\underline{\sigma}^{(k)}$ azon objektumok számát, amelyek pontosan $\underline{k}$ számu jelleggel rendelkeznek, és legyen $\underline{\sigma} = \underline{\sigma}^{(0)}$. Megmutatom, hogy a $\underline{W}_i$-k függetlensége esetén teljesülnie kell, hogy

$$\sigma = M \prod_{i=1}^{n} \frac{\sigma}{\sigma+\sigma(i)}.$$

Ezen egyenlőség ellenőrzése lehet a függetlenség vizsgálatának első lépése.

Jelölje $\underline{P}^{(k)}$ annak valószinüségét, hogy egy objektum pontosan $\underline{k}$ számu jelleggel rendelkezik, és legyen $\underline{P} = \underline{P}^{(0)}$. Vezessük be a következő jelöléseket:

$$q_i = \frac{P_T(i)}{1-P_T(i)}, \quad i = 1, 2, \ldots, n,$$

$$S_0 = 1,$$

$$S_k = \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} \prod_{j=1}^{k} q_{i_j}, \quad k = 1, 2, \ldots, n.$$

Megmutatom, hogy a $\underline{P}^{(k)}$ valószinüségek előállitásához

$\underline{P}$-n kivül elég az $\underline{S}_k$-kat kiszámitani, amelyek - a $\underline{q}_i$-kon keresztül - csak a $\underline{P}_T(\underline{i})$ valószinüségektől függenek. [Az $\underline{S}_k$-k rekurzivan, Newton formulájával számithatók: $\underline{k}$ = = 1, 2, ..., $\underline{n}$ esetén

$$S_k = \frac{1}{k} \sum_{j=1}^{k} (-1)^{j-1} T_j S_{k-j},$$

ahol

$$T_j = \sum_{i=1}^{n} q_i^j, \quad j = 1, 2, \ldots, n.]$$

Jelölje $\underline{N}_T^{(k)}$ az előforduló különböző $\underline{k}$ méretü jellegkombinációk számát, $\underline{N}^{(k)}$ pedig ezek közül azokét, amelyek a pontosan $\underline{k}$ számu jelleggel rendelkező objektumok körében fordulnak elő. A jellegkombinációk mérete tapasztalati eloszlásának jellemzésében ezek a mennyiségek is fontos szerepet játszanak. $\tilde{\underline{N}}_T^{(k)}$ és $\tilde{\underline{N}}^{(k)}$ várható értéküket közelitőleg határozom meg oly módon, hogy az $\underline{\sigma}_T(\underline{i}_1,\underline{i}_2,\ldots,\underline{i}_k)$, ill. $\underline{\sigma}(\underline{i}_1,\underline{i}_2,\ldots,\underline{i}_k)$ változók binomiális eloszlását Poisson-eloszlással közelitem. Megmutatom, hogy $\tilde{\underline{N}}_T^{(k)}$ és $\tilde{\underline{N}}^{(k)}$ ilyen közelitése is csak a $\underline{P}$ és $\underline{P}_T(\underline{i})$ valószinüségektől függ. A $\underline{W}_i$ változók függetlenségvizsgálatának következő lépése a $\underline{P}^{(k)}$, $\tilde{\underline{N}}_T^{(k)}$, $\tilde{\underline{N}}^{(k)}$ mennyiségek $\hat{\underline{P}}^{(k)}$, $\hat{\tilde{\underline{N}}}_T^{(k)}$, $\hat{\tilde{\underline{N}}}^{(k)}$ becslésének meghatáro-

zása és az $\{\underline{\sigma}^{(k)}\}$, $\{\underline{N}_T^{(k)}\}$, $\{\underline{N}^{(k)}\}$, ill. $\{\underline{\hat{MP}}^{(k)}\}$, $\{\underline{\hat{N}}_T^{(k)}\}$, $\{\underline{\hat{N}}^{(k)}\}$ sorozatok összehasonlitása. Döntő az $\{\underline{\sigma}^{(k)}\}$ és $\{\underline{\hat{MP}}^{(k)}\}$ sorozatok összehasonlitása; megmutatom, hogyan célszerü ezt elvégezni.

Ha a változók függetlensége elfogadhatatlannak bizonyul, felmerülhet az a gondolat, hogy ezt csak néhány jelleg okozza. Ezért megmutatom, hogyan lehet megvizsgálni, hogy miként változik az egyes jellegek adott méretü jellegkombinációk közötti gyakorisága a méret változásával. Ha ezek a gyakoriságok lényegében egymáshoz hasonlóan változnak, akkor a változók függetlenségének elfogadhatatlanságát nem csupán néhány jelleg okozza.

A 3. fejezet a normális küszöb modellel [lásd pl. Tusnády (1969)] foglalkozik. Ez feltételezi, hogy a vizsgált dichotom jellegeknek mindegyik objektumra nézve van egy-egy valós számmal kifejezhető mértéke, vagyis a $\underline{X}_i$ bináris változókhoz hozzá van rendelve egy-egy $\underline{L}_i$ folytonos háttér változó. A modell szerint ezek az $\underline{L}_i$-k együttes normális eloszlásu valószinüségi változók, amelyekhez egy-egy $\underline{T}_i$ küszöb tartozik, továbbá egy objektum akkor és csak akkor rendelkezik az $\underline{A}_i$ jelleggel, ha $\underline{L}_i$ rajta felvett értéke eléri $\underline{T}_i$-t. Az $\underline{L}_i$-kről feltesszük, hogy standardak (lévén általában nem megfigyelhetőek, ez a technikai jellegü feltétel nem jelent megszoritást), ezért együttes eloszlásukat meghatározzák az $(\underline{L}_i, \underline{L}_j)$ változópárok közötti $\underline{r}_{ij}$ korrelációs együtthatók. A mo-

dell a függetlenségnél lényegesen gyengébb feltevésen alapul, viszont nagy változószámnál jelentős csökkenést eredményez a paraméterek számában. A modell paraméterei a $\underline{T}_i$-k és az $\underline{r}_{ij}$-k, számuk összesen $\binom{n+1}{2}$. Maximum likelihood (ML) becslésük $\underline{n} > 2$ esetén nem ismert. $\underline{T}_i$-t a

$$\Phi^{-1}[1 - \frac{\sigma_T(i)}{M}]$$

kifejezés értékével becslem [$\Phi(\underline{z})$ az (1-dimenziós) standard normális eloszlásfüggvény; $\underline{n} = 1$ esetén ez a $\underline{T}_i$ küszöb ML becslése], $\underline{r} = \underline{r}_{ij}$-t pedig az

$$F(\hat{T}_i, \hat{T}_j, r) = \frac{\sigma_T(i,j)}{M}$$

egyenlet megoldásával [itt

$$F(\tilde{T}, T, r) = P(\xi \geq \tilde{T}, \eta \geq T) =$$

$$= \frac{1}{2\pi\sqrt{1-r^2}} \int_{\tilde{T}}^{\infty} \int_{T}^{\infty} \exp[-\frac{u^2-2ruv+v^2}{2(1-r^2)}] \, du \, dv,$$

ξ és η standard, $\underline{r}$ korrelációs együtthatóju, együttes normális eloszlásu valószinüségi változók (lásd pl. Anderson, 1958); $\underline{n} = 2$ esetén a fentiek szerint megha-

tározott $\hat{T}_i$ becslésekkel ez az ML becslést adja]. F sorfejtés alapján történő gyors számitógépes meghatározására a 70-es évek közepén eljárást dolgoztam ki, amely az akkor elérhető eljárásoknál hatékonyabbnak bizonyult. Bebizonyitom, hogy a sorfejtés $|r| < 1/\sqrt{2}$ esetén konvergens, és megmutatom, hogyan lehet $F(\tilde{T},T,r)$-et tetszőleges r esetén meghatározni. Mivel $F(\tilde{T},T,r)$ r-ben monoton növő, az r_{ij}-becslések numerikus meghatározása ezután már nem okoz nehézséget.

A modell kézenfekvő ellenőrzése annak vizsgálata, mennyire felel meg a 2-nél nagyobb méretü jellegkombinációk előfordulása a modellnek. Mivel itt nem sikerült meghatároznom a $p^{(k)}$ valószinüségek becslését, ezért a

$$\{\sum_{g:|G_g|=k} \sigma_T(G_g)\}$$

sorozat elemeit hasonlitom össze az

$$\{M \times \sum_{g:|G_g|=k} P_T(G_g)\}$$

sorozat elemeivel ($|G_g|$-vel a G_g jellegkombináció méretét jelölöm). A 2-nél nem nagyobb méretü jellegkombinációk modelltől való eltérését egyenként is jellemzem.

Mivel az $\underline{\underline{R}} = (\underline{r}_{ij})$ korrelációs mátrix ML becslése nem ismeretes, $\underline{\underline{R}}$-et elemenként becslem. Ezért előfordulhat, hogy $\underline{\underline{R}}$ igy nyert $\hat{\underline{\underline{R}}}$ becslése nem pozitiv szemidefinit. Kétféle módon is előállitok olyan, $\hat{\underline{\underline{R}}}$-hoz "közeli" $\tilde{\underline{\underline{R}}}$ mátrixot, amelyik pozitiv szemidefinit és $\underline{n}$ nyomu. [Tudomásom szerint nem ismert, hogy az ilyen tulajdonságu $\tilde{\underline{\underline{R}}}$-ok $\hat{\underline{\underline{R}}}$-tól mátrix normában való távolságát milyen mátrix minimalizálja.] Mindkét esetben olyan $\tilde{\underline{\underline{R}}}$ mátrixot kapunk, amely ugyan pozitiv szemidefinit és $\underline{n}$ nyomu, de rendelkezhet $\tilde{\underline{r}}_{ii} \neq 1$ vagy $|\tilde{\underline{r}}_{ij}| > 1$ elemekkel. Ismertetem, hogyan juthatunk olyan $\tilde{\underline{\underline{R}}}$-hoz, amely már "igazi" korrelációs mátrix.

A 4. fejezet <u>többdimenziós skálázással</u> [angolul "<u>m</u>ulti<u>d</u>imensional <u>s</u>caling", MDS; lásd például Kruskal (1977a,b)] foglalkozik. A változók MDS-e azt a problémát vizsgálja, hogy a változók között értelmezett távolságok alapján hogyan lehet a változókat térképszerüen ábrázolni, kirajzolni valamely $\underline{R}^k$ alacsony dimenziós euklideszi térben, vagyis hogyan lehet $\underline{R}^k$-ban olyan pont-$\underline{n}$-est konstruálni hozzájuk, hogy a változók távolságai minél kevésbé térjenek el a megfelelő pontok euklideszi távolságaitól. Ha sikerül jól skálázni a változókat, "térképük" ránézésre sok információt adhat strukturájukról. Ahhoz, hogy a változók jól skálázhatóak legyenek, konzisztenseknek kell lenniük a következő értelemben: ha két változó "közel" van egy harmadikhoz, egymáshoz is "kö-

zel" kell lenniük. Az MDS témakörébe tartozó feladatok osztályozása, az MDS irodalmában fellelhető összefoglaló munkák áttekintése, valamint az MDS un. klasszikus megoldását ismertető 4.1. paragrafus után a 4.2. paragrafus az általam bevezetett többszörös többdimenziós skálázással (multiple MDS, MMDS) foglalkozik. Az MMDS, amelyet az értekezés legfontosabb elméleti eredményének tartok, azt a problémát vizsgálja, amelyik akkor merül fel, ha a fenti konzisztencia nem teljesül: hogyan lehet az objektumokat minél homogénebb klaszterekbe sorolni, amikoris egy-egy klaszter homogenitását a változók ezen klaszter mellett történő MDS-ének jóságával mérjük? Keresendő tehát a p természetes szám, az $Y_1, Y_2, \ldots, Y_p$ (diszjunkt) objektumklaszterek és azon $\underline{x}_i^{(m)} \in R^k$ pontok, amelyek a változókat reprezentálják abban az értelemben, hogy $\underline{x}_i^{(m)}$ és $\underline{x}_j^{(m)}$ közelsége a W_i, W_j változók Y_m melletti közelségének felel meg. (Az MMDS hazai előzményei közé tartozik a következő, Tusnády Gábor által vizsgált probléma: hogyan lehet az objektumokat klaszterezni és a változókhoz az egyes klaszterek mellett pontokat rendelni ugy, hogy egy-egy objektum 1 értékü változóihoz az objektumot tartalmazó klaszter mellett tartozó pontok minél közelebb legyenek egymáshoz?)

Legyen e_{gi} a W_i változónak az y_g objektumon megfigyelt értéke (1 vagy 0 aszerint, hogy y_g rendelkezik-e

$\underline{A}_i$-vel vagy sem),

$$\underline{\underline{e}}_g = (e_{g1}, e_{g2}, \ldots, e_{gn}),$$

$$U_g^{(m)} = \sum_{\substack{i=1 \\ e_{gi}=1}}^{n-1} \sum_{\substack{j=i+1 \\ e_{gj}=1}}^{n} \|\underline{\underline{x}}_i^{(m)} - \underline{\underline{x}}_j^{(m)}\|^2,$$

$\underline{\underline{DC}}$ olyan eljárás, amellyel tetszőleges $\underline{Z}$ objektumhalmazhoz - a $\underline{Z}$ objektumaihoz tartozó $\underline{\underline{e}}_g$-k és $\underline{\underline{G}}_g$-k alapján - $\underline{d}_{ij}(\underline{Z})$ távolságokat és $\underline{c}_{ij}(\underline{Z})$ sulyozó tényezőket rendelek hozzá,

$$d_{ij}^{(m)} = d_{ij}(Y_m), \quad c_{ij}^{(m)} = c_{ij}(Y_m)$$

és

$$V^{(m)} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c_{ij}^{(m)} [d_{ij}^{(m)} - \|\underline{\underline{x}}_i^{(m)} - \underline{\underline{x}}_j^{(m)}\|]^2.$$

Az MMDS-problémát azzal a $\underline{T}$ algoritmussal oldom meg, amelyik a következő két lépést alkalmazza váltakozva: i) az egyes $\underline{y}_g$ objektumokat abba a klaszterbe sorolja, amelyikre $\underline{U}_g^{(m)}$ minimális (amelyik klaszterben a $\underline{G}_g$ jellegkombináció jellegeinek megfelelő változókhoz tartozó pontok a legközelebb vannak egymáshoz); ii) az egyes klaszterek mellett meghatározza

$$V^{(m)} = V^{(m)}[\underline{\underline{x}}_1^{(m)}, \underline{\underline{x}}_2^{(m)}, \ldots, \underline{\underline{x}}_n^{(m)}]$$

negativ gradiens vektorát, majd ennek irányában irány-menti minimalizálást végezve kiszámitja az $\underline{\underline{x}}_i^{(m)}$ pontok uj koordinátáit.

Az MMDS-t az

$$E = \sum_{m=1}^{p} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \{n_{ij}^{(m)} \| \underline{\underline{x}}_i^{(m)} - \underline{\underline{x}}_j^{(m)} \|^q + [K - \| \underline{\underline{x}}_i^{(m)} - \underline{\underline{x}}_j^{(m)} \|]^q\}$$

mennyiséggel jellemzem, ahol

$$n_{ij}^{(m)} = n_{ij}(Y_m),$$

$\underline{n}_{ij}(\underline{Z})$ azon $\underline{Z}$-beli objektumok száma, amelyek rendelkeznek az $\underline{A}_i$ és $\underline{A}_j$ jelleggel, $\underline{q}$ és $\underline{K}$ alkalmas állandók. [Simonovits Miklós hivta fel a figyelmemet arra, hogy $\underline{q} = 2$ esetén $\underline{E}$ a következő, 2 direkciós erejü rugókból álló rendszer potenciálja: $\underline{n}_{ij}^{(m)}$ számu 0 hosszuságu és egy $\underline{K}$ hosszuságu rugó $\underline{\underline{x}}_i^{(m)}$ és $\underline{\underline{x}}_j^{(m)}$ között ($1 \leq \underline{i} < \underline{j} \leq \underline{n}$, $\underline{m} = 1, 2, \ldots, \underline{p}$).] Az MMDS $\underline{E}$-vel történő jellemzése a következő $\underline{\underline{DC}}$-nek felel meg:

$$c_{ij}(Z) = n_{ij}(Z) + 1, \quad d_{ij}(Z) = K \,/\, c_{ij}(Z).$$

Bebizonyitom, hogy ezen DC és $q = 2$ esetén a T algoritmus folyamán E monoton nem nő. Ismertetem, hogyan célszerü p kezdő és további értékeit, q-t és K-t választani, valamint kezdő klasztereket és pontkonfigurációt előállitani.

Valamely adatmező MMDS-ét két szempontból is jó lenne értékelni. Annak matematikai értékelését, hogy az MMDS mennyire jó más klaszterező módszerekhez képest, elvileg kivihetetlennek tartom. Az egyes klaszterező módszerek ugyanis döntően éppen abban különböznek egymástól, hogy milyen kritériumot adnak a klaszterezés jóságára. A maga kritériuma szerint mindegyik módszer a legjobb, a kritériumok viszont nem hasonlithatóak objektiven össze. (Különböző klaszterező módszereket heurisztikusan persze összehasonlithatunk: megnézzük, hogy adott klasztereket hogyan adnak vissza.) Annak értékelése, hogy a konkrét adatmező objektumai mennyire jól klaszterezhetőek az MMDS szempontjából, elvileg a következőképpen végezhető el. Tegyük fel, hogy az adatmező valamilyen q és K mellett végrehajtott MMDS-ének befejezésekor a klaszterszám p, és E értéke

$$E^* = E^*(n, M, \mu, p, q, K),$$

ahol

$$\mu = \frac{\sum_{m=1}^{p} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} n_{ij}^{(m)}}{\binom{n}{2}} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sigma_T(i,j)}{\binom{n}{2}}.$$

Legyen $\underline{E}_{opt}$ és $\underline{E}_{pessz}$ az $\underline{E}$ mennyiség értéke az $(\underline{n},\underline{M},\underline{\mu})$ értékhármashoz tartozó, az MMDS szempontjából optimális, ill. pesszimális adatmező $\underline{q}$ és $\underline{K}$ fenti értéke mellett végrehajtott MMDS-ének $\underline{p}$ fenti értéke mellett történő befejezésekor. Nyilvánvaló módon $\underline{E}^*$ az $[\underline{E}_{opt},\underline{E}_{pessz}]$ intervallumban helyezkedik el. Annak, hogy a konkrét adatmező mennyire jó az MMDS szempontjából, kézenfekvő mérőszáma az

$$\frac{E^*-E_{opt}}{E_{pessz}-E_{opt}}$$

mennyiség, amely nem lehet 0-nál kisebb és 1-nél nagyobb. Meghatározásához azonban tudni kellene $\underline{E}_{opt}$ és $\underline{E}_{pessz}$ értékét. Ezeket azonban nem tudom, mivel nem ismerem az optimális és pesszimális adatmezőt.

A 4.3. paragrafus a változók függetlensége esetén vizsgálja az MMDS-t. Ennek müködésében meghatározó szerepe van az őt jellemző $\underline{E}$ mennyiségben szereplő $\underline{f}$ függvénynek. Ennek megválasztásához a 4.2. paragrafusban ugy konkretizálom a távolságokat és sulyozó tényezőket meghatározó $\underline{\underline{DC}}$ eljárást, hogy azok csak az egyes jellegpárok együttes előfordulásától függenek. Emiatt az egyes jellegek előfordulási valószinüségeinek különbözősége esetén az eljárás nem veszi észre a függetlenséget: akkor is közel hoz egymáshoz két jelleget, ha azok csak azért fordulnak elő sürün együtt, mert mindketten gyakoriak. (Ez azonban

célom is volt: az 5.6. paragrafusban ismertetésre kerülő konkrét feladat során, amelyből az MMDS kinőtt, nem akartam észrevétlenül hagyni jellegzetes jelleg-kombinációkat.) Azt, hogy az eljárás észrevegye a függetlenséget, $\underline{n}_{ij}(\underline{Z})$ definiciójának módositásával lehet megvalósitani. Ha azonban az egyes jellegek előfordulási valószinüségei megegyeznek, erre nincs szükség. Ezért a függetlenséget ebben az esetben vizsgálom. Nevezetesen olyan adatmezőt generálok, amelyre

$$\sigma_T(i,j) = \sigma(i,j) \equiv \mu, \quad 1 \leq i < j \leq n.$$

Ismertetem, hogy ebben az esetben milyen objektumklasztereket ad az MMDS, hogyan helyezkednek el a jellegeknek megfelelő pontok, továbbá milyen összefüggést kapunk az $\underline{E}$ mennyiség $\tilde{\underline{E}}$ végértékére általában és a $\mu = 0$, $\underline{p} = \underline{K} = 1$ feltételek mellett, amikoris

$$E = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (1-\|\underline{x}_i - \underline{x}_j\|)^2.$$

A 4.4. paragrafus <u>hipergráfok euklideszi térbe ágyazásával és particionálásával</u> foglalkozik. Az $\underline{y}_g$ objektumok összességéhez ugyanis természetes módon hozzárendelhető egy $\underline{H}$ hipergráf ugy, hogy $\underline{H}$ szögpontjai a $\underline{W}_i$ válto-

zóknak felelnek meg, a hiperélek az objektumoknak [a g-edik hiperél pontosan azokat a szögpontokat köti össze, amely szögpontokhoz tartozó változóknak megfelelő jellegekkel y_g rendelkezik. Ekkor az i-edik és j-edik szögpontot $\sigma_T(i,j)$ közönséges él köti össze]. Megmutatom, hogy a H hipergráf k-dimenziós euklideszi térbe ágyazását, vagyis $\underline{x}_i \in R^k$ pontok H szögpontjaihoz való rendelését a

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} [\sigma_T(i,j)\|\underline{x}_i - \underline{x}_j\|^2 - (K - \|\underline{x}_i - \underline{x}_j\|)^2]$$

mennyiség minimalizálásával lehet kézenfekvő módon elvégezni, ami viszont ekvivalens a változók

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c_{ij}(d_{ij} - \|\underline{x}_i - \underline{x}_j\|)^2$$

minimalizálásával történő MDS-évei, ahol

$$c_{ij} = \sigma_T(i,j) + 1, \quad d_{ij} = K / c_{ij},$$

és K alkalmas állandó. Uj kritériumot javaslok H particionálására, nevezetesen azt, hogy a H_m rész-hipergráfok mi-

nél jobban beágyazhatóak legyenek. Megmutatom, hogy ez természetes módon az $\underline{E}$ mennyiség $\underline{q} = 2$ mellett történő minimalizálásához vezet [az $\underline{i}$-edik és $\underline{j}$-edik szögpontot $\underline{H}_m$-ben $\underline{n}_{ij}^{(m)}$ közönséges él köti össze], igy nyilvánvalóan ekvivalens az MMDS-sel.

A 4.5. paragrafus klaszterek MDS-ével foglalkozik, pontosabban a következő probléma klaszter-MDS segitségével történő megoldásával: hogyan lehet az objektumok különböző (módszerek alkalmazásával kapott) klaszterezéseiből egy "eredő" klaszterezést előállitani? Képezem az $\underline{r}$-edik klaszterezés melletti $\underline{m}_1$-edik és $\underline{m}_2$-edik klaszter $\underline{d}_{m_1 m_2}^{(r)}$ távolságát, és az $\underline{Y}_1^{(r)}, \underline{Y}_2^{(r)}, \ldots, \underline{Y}_{p_r}^{(r)}$ klasztereknek ($\underline{p}_r$ a klaszterek száma az $\underline{r}$-edik klaszterezésben) a

$$\underline{\underline{D}}^{(r)} = [d_{m_1 m_2}^{(r)}]_{m_1, m_2 = 1}^{p_r}$$

távolságmátrix alapján történő MDS-ével a klaszterekhez $\underline{\underline{z}}_1^{(r)}, \underline{\underline{z}}_2^{(r)}, \ldots, \underline{\underline{z}}_{p_r}^{(r)}$ pontokat konstruálok a $\underline{k}_r$-dimenziós euklideszi térben. Ezek segitségével $\underline{K}_1$-dimenziós

$$(K_1 = \sum_{r=1}^{\nu} k_r,$$

ν a klaszterezések száma) $\underline{\underline{v}}_g$ pontokat feleltetek meg az $\underline{Y}_g$ objektumoknak. Megmutatom, hogyan lehet a $\underline{\underline{v}}_g$-k klasztere-

zésével az $\underline{y}_c$-k kivánt "eredő" klaszterezését előállitani. Ismertetem a klaszterek MDS-ét, amelyet a klaszterek inkonzisztenciája esetén célszerü alkalmazni.

A 4.6. paragrafus a <u>sávszélesség-redukcióval</u> rokon következő problémával foglalkozik: hogyan lehet adott (nxn)-es szimmetrikus, sok zérus-elemet tartalmazó $\underline{\underline{A}}$ mátrixhoz azt a $\underline{\underline{P}}$ permutációs mátrixot meghatározni, amely mellett a

$$(b_{ij})_{i,j=1}^{n} = \underline{\underline{P}}\,\underline{\underline{A}}\,\underline{\underline{P}}^{T}$$

mátrixhoz tartozó

$$\sum_{i,j:\,b_{ij}\neq 0} |i-j| \;/\; \sum_{i,j:\,b_{ij}\neq 0} 1$$

mennyiség minimális? (Szemben a sávszélesség-redukcióval itt tehát nem maximumot, hanem átlagot kell minimalizálni. A motiváció azonban ugyanaz: egy szimmetrikus ritka mátrixot a számitógép memóriájának minél kisebb részében elhelyezni.) Ismertetem, hogyan lehet ezt a problémát MDS-sel heurisztikusan - és valószinüleg közelitően - megoldani.

Az 5. fejezet a <u>veleszületett rendellenességek</u> statisztikai vizsgálatával foglalkozik. Ismertetem a veleszületett rendellenesség (congenital anomaly, congenital abnormality, CA; congenital malformation, CM; a CM-ek a CA-k részhalmazát képezik) definicióját, majd indokolom, miért fontos a CA-k statisztikai vizsgálata, a 2-4. fejezet eredményeinek ismertetésre kerülő alkalmazása. Is-

mertetem az izolált CA és a többszörös CA (multiple CA, MCA) definicióját (egy CA önmagában való előfordulása; több CA együttes előfordulása ugyanannál a személynél), majd a CA-k gyakoriságuk alapján történő értékelését.

Az 5.1. paragrafus kilenc izolált gyakori CM (isolated common CM, ICCM) öröklődésének vizsgálatával foglalkozik. Az egyes ICCM-ek öröklődésének leírására egy speciális normális küszöb modellt választok. Eszerint az adott ICCM-hez hozzárendelt L háttér változóra, amelyet itt szokás hajlamnak is nevezni,

$$L = G + E,$$

ahol G a genetikai, E a környezeti hatást jelenti, G és E független, 0 várható értékü, h^2, ill. $(1-h^2)$ szórásnégyzetü, normális eloszlásu valószinüségi változók (h^2 az un. örökölhetőségi együttható), és mindegyikük multifaktoriális hatásnak, tehát sok kis tényező hatásának az eredménye, továbbá a rokonok hajlamai együttes normális eloszlásuak, és d-edfoku rokonok hajlamainak ($h^2/2^d$) a korrelációs együtthatója. A modellben, amelyet ilyen formában Czeizel Endre és Tusnády Gábor vezetett be, a hajlam mellett legdöntőbb a normális (Gaussian), additiv (additive), multifaktoriális (multifactorial) hatás és a küszöb (threshold), ezért GAMT-modellnek nevezik. Különböző rokoncsoportokra, ill. ilyenek bizonyos összességeire eljárást adok a h^2 becslésére és a becslés megfelelő szintü konfidenciaintervallum-

ba foglalására, továbbá annak ellenőrzésére, hogy a különböző (rokoncsoportokhoz tartozó) h^2-becslések eltérése szignifikáns-e vagy sem (amivel tulajdonképpen a modellt ellenőrzöm).

Az 5.2. paragrafus az MCA-k értékelésével foglalkozik. Ismertetem az $n = 40$ CA-t és az adatokat, amelyeket a Veleszületett Rendellenességek Országos Nyilvántartása alapján Czeizel Endre bocsátott rendelkezésemre. M = = 1 186 776 gyermek közül 30 850 volt CA-val rendelkező (közülök 921 különböző) és ezen belül 2762 MCA-val rendelkező (közülük $N = 881$ különböző). Az adatok tárolása a fentebb említett, az 1. fejezet következő paragrafusában ismertetendő módon történt. NS méretét az ott leírt módokon csökkentve az előfordult 2568-féle G_g rendellenességkombinációhoz egy 9268 elemü NS volt szükséges. Ismertetem az $\sigma^{(k)}$, $N^{(k)}$, $N_T^{(k)}$ mennyiségek értékét.

Az 5.3. paragrafus a CA-k un. függetlenségi koncepciójával foglalkozik. Ennek lényege, hogy ugyanaz a CA különböző esetekben különböző, multifaktoriális vagy oligofaktoriális kóreredetü lehet (a két tipus az MCA-k statisztikai vizsgálatában megkülönböztethetetlen), és valamennyi multifaktoriális CA független a többi CA-tól. Tetszőleges MCA tehát p_1, ill. p_2 valószinüséggel multi- vagy oligofaktoriális, vagyis az MCA-k valószinüségeloszlása két eloszlás keveréke, de sem a p_1, p_2 valószinüségeket nem ismerjük, sem azt nem tudjuk, hogy az egyes MCA-k a keverék

melyik eloszlásához tartoznak. A függetlenségi koncepció nem teszi fel a CA-k bármelyik kóreredet melletti függetlenségét. Először ez utóbbi hipotézist vizsgálom meg, a 2. fejezet alapján. Ennek eredménye, hogy indokolt a hipotézis elutasitása. Ugyanez azonban a függetlenségi koncepcióra is érvényes: ennek fennállása esetén ugyanis az MCA-knak csak 11 százaléka lenne multifaktoriális, ami orvosilag irreális. Megvizsgálom, hogy a két hipotézis elfogadhatatlanságát nem csupán néhány CA okozza-e. Az eredmények azt mutatják, hogy nem.

Az 5.4. paragrafus a függetlenségi koncepció módositásával foglalkozik. A koncepció "duálisa" az a feltételezés, amely szerint az oligofaktoriális MCA-kban a CA-k függetlenek, de a multifaktoriálisak nem feltétlen azok. Megmutatom, hogyan célszerü ezen hipotézis vizsgálatához a $P_T(i)$ valószinüségeket becsülni, és eljárást adok a becslések meghatározásához. A hipotézis elfogadhatónak bizonyul, ahhoz azonban, hogy használható is legyen, ki kell egésziteni a multifaktoriális CA-kra vonatkozó valamilyen feltevéssel. Ezt teszi a feltételes függetlenségi koncepció, amely a függetlenségi koncepciónak és duálisának mintegy "metszete". Eszerint tetszőleges MCA p_1, ill. p_2 valószinüséggel multi- vagy oligofaktoriális kóreredetü, és a CA-k a kóreredet mint feltétel mellett függetlenek. Ez azt jelenti, hogy az MCA-k valószinüségeloszlása $m = 2$ számu olyan eloszlás keveréke, amelyek mindegyikében a CA-k függetlenek.

Mivel a függetlenségi koncepció elfogadhatatlan, a feltételes függetlenségi koncepció is az. Viszont természetes módon általánosítható: nem tesszük fel, hogy akár a multi-, akár az oligofaktoriális kóreredetü MCA-khoz a keverék egyetlen eloszlása tartozik, tehát azt, hogy $m = 2$. Az igy konstruált keverékeloszlásos modellt Tusnády Gábor vizsgálta [lásd Tusnády (1978-1982)].

Az 5.5. paragrafus a GAMT-modell kiterjesztésével foglalkozik. Felteszem, hogy a különböző CA-kat előidéző genetikai és környezeti tényezők (természetesen ugyanannál a személynél) korreláltak, és a multifaktoriális MCA-k ennek a korrelációnak a következményei. Mivel az oligofaktoriális és multifaktoriális MCA-k megkülönböztethetetlenek, ezért a paragrafus további része azon a hipotézisen alapul, amely szerint az egyes CA-k öröklődésmenete a GAMT-modellel leirható, az őket előidéző genetikai és környezeti tényezők korreláltak, és valamennyi MCA ennek a korrelációnak a következménye. Felteszem, hogy az egyes $\underline{A}_i$ rendellenességekhez - ugyanannál a gyereknél - hozzárendelt $\underline{L}_i$ hajlamok együttes normális eloszlásuak. Ez a GAMT-modell kiterjesztését jelenti, hiszen az eredeti GAMT-modell egy rendellenesség több családtagra vonatkozó hajlamát irja le, itt viszont egy gyermeknek több hajlama van (az n számu CA mindegyikéhez egy), és ezeknek a hajlamoknak a rendszere alakitja ki a gyerek CA-inak rendszerét. A modellt és a megfelelő hipotézist a 3. fejezet alapján vizsgálom meg. Ennek ered-

ménye, hogy a modell elfogadható. A modell segotségével jellemzem a CA-k multifaktoriális és oligofaktoriális közelségét.

Az 5.6. paragrafus a CA-k MMDS-ével foglalkozik. Az MCA-k statisztikai vizsgálatának ugyanis fő céljai közé tartozott a rendellenes gyerekek csoportositása a CA-k együttes előfordulása alapján és ezáltal jellegzetes CA-kombinációk felderitése, és éppem e feladat megoldására dolgoztam ki az MMDS-t. Ismertetem, hogyan lehet ennek adekvát voltáról meggyőződni: véletlen módon generált vagy szisztematikusan kijelölt CA-klaszterekhez véletlen adatmezőt generálunk, és megnézzük, ugy csoportositja-e az MMDS a generált gyerekeket, hogy a különböző gyerekklaszterek mellett kirajzolódnak az egyes CA-klaszterek. (A generálást a keverékeloszlásos modell alapján végzem.) Ismertetem a véletlen adatmezőt és többszörös többdimenziós skálázását, majd az "igazi" adatmező ez alapján adekvátnak mutatkozó MMDS-ének eredményét.

A rendellenes gyerekek csoportositását és jellegzetes CA-kombinációk ezáltal történő felderitését keverékfelbontással Tusnády Gábor, orvosi megfontolások alapján Czeizel Endre is elvégezte. A háromféle klaszterezésből klaszter-MDS segitségével egy "eredő" klaszterezést állitottam elő. Az eredmények orvosi értékelése, ellenőrzése (genetikai családvizsgálatok) és felhasználása (genetikai tanácsadás) Czeizel et al. (1987)-ben kerül kifejtésre. - A rendellenessé-

gek Boole-faktoranalizisét Rejtő Lidia, loglineáris elemzését Rudas Tamás végezte el, korrespondencia-analizisükkel pedig Dávidné Bolla Marianna foglalkozik.

1.1. Bináris adatok tárolása

Legyen

$$G_g = \{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}, \quad 1 \le k \le n, \quad 1 \le i_1 < i_2 < \ldots < i_k \le n,$$

tetszőleges jellegkombináció,

$$\sigma(G_g) = \sigma(i_1, i_2, \ldots, i_k)$$

és

$$\sigma_T(G_g) = \sigma_T(i_1, i_2, \ldots, i_k)$$

azon objektumok száma, amelyek rendelkeznek a $\underline{G}_g$-hez tartozó jellegekkel, de más jelleggel nem, ill. amelyek rendelkeznek a $\underline{G}_g$-hez tartozó jellegekkel és esetleg továbbiakkal is. Mivel (2^n-1) számu különböző $\underline{G}_g$ van, azért feltehető, hogy $1 \le g \le 2^n-1$. [Az $\underline{\sigma}(\underline{G}_g)$ gyakoriságok éppen a megfelelő $\underline{n}$-dimenziós kontingenciatáblázat elemei azzal a különbséggel, hogy az "üres halmaz" (az az esemény, hogy valamennyi változó értéke 0) nem szerepel a jellegkombinációk között.

A számomra érdekes esetekben a jellegek ritkán fordulnak elő, ezért a kontingenciatáblázat elemeinek tulnyomó többsége zérus.] Nyilvánvaló a következő összefüggés:

$$\sigma_T(G_g) = \sum_{v:G_v \geq G_g} \sigma(G_v).$$

Az adatok számitógépes kezelését illetően kézenfekvő a következő két követelmény: i) az adatok a memória ne tul nagy részét foglalják el; ii) tetszőleges $\underline{G}_g$ jellegkombinációhoz az $\underline{\sigma}(\underline{G}_g)$, $\underline{\sigma}_T(\underline{G}_g)$ gyakoriságok gyorsan kikereshetőek legyenek. E célból a fenti gyakoriságokat és címüket, ill. az értékükre vonatkozó információt egy egyindexes $\underline{NS}$ egész tömbben helyezem el oly módon, hogy tetszőleges

$$G_g = \{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}$$

esetén az $[\underline{\sigma}_T(\underline{G}_g), \underline{\sigma}(\underline{G}_g)]$ számpárt $\underline{NS}$-ben

$$\sigma_T(i_1,i_2,\ldots,i_k,i_k+1),\ \sigma_T(i_1,i_2,\ldots,i_k,i_k+2),$$
$$\ldots,\ \sigma_T(i_1,i_2,\ldots,i_k,n)$$

címei követik. [Tulajdonképpen tehát egy keresőfával dolgozom. A keresőfák irodalmából megemlitem Knuth (1973)-at és Aho et al. (1982)-t.] $\underline{NS}$ feltöltéséhez az alapadatok az olyan különböző $\underline{G}_g$ jellegkombinációk, amelyek önmagukban,

más jelleg nélkül is előfordultak [amelyekre tehát $\sigma(G_g) > 0$], és a megfelelő $\sigma(G_g)$ gyakoriságok. A feltöltés során mindegyik fenti g-re egyrészt $\sigma(G_g)$ - végleges - értékét helyezem el NS megfelelő elemében, másrészt G_g összes nem üres G_v részhalmaza NS megfelelő elemében elhelyezett vagy ekkor elhelyezendő $\sigma_T(G_v)$ gyakoriságának értékét megnövelem $\sigma(G_g)$-vel. Az σ_T gyakoriságok tehát - szemben az σ gyakoriságokkal - általában csak a feltöltési algoritmus végére kapják meg végleges értéküket. - NS-t egy EU egész változó segitségével töltöm fel, amelynek értéke mindig NS első feltöltetlen elemének sorszáma, részletesen az alábbiak szerint:

1.) EU:=n+1.

2.) $i = 1, 2, \ldots, n$-re

NS[i]:=EU,
NS[EU]:=NS[EU+1[:=σ(i)

{a feltöltés befejezése után NS[EU] = NS[NS[i]]-ben az $\sigma_T(i)$ gyakoriság értékét kapjuk meg},

EU:=EU+2+n-i;

3.) $G_g = \{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}$, $k \geq 2$, $\sigma(G_g) > 0$,

esetén [egy olyan $\underline{G}_g$ jellegkombinációt, amelyre ugyan $\underline{\sigma}_T(\underline{G}_g) > 0$, de $\underline{\sigma}(\underline{G}_g) = 0$, csak a $\underline{G}_h \supset \underline{G}_g$ és $\underline{\sigma}(\underline{G}_h) > 0$ feltételeknek eleget tevő $\underline{G}_h$ jellegkombinációk részhalmazaként veszünk figyelembe] a

$$G_v = \{A_{i_1}\}, \{A_{i_2}\}, \ldots, \{A_{i_k}\},$$
$$\{A_{i_1}, A_{i_2}\}, \{A_{i_1}, A_{i_3}\}, \ldots, \{A_{i_{k-1}}, A_{i_k}\},$$
$$\ldots, \{A_{j_1}, A_{j_2}, \ldots, A_{j_r}\}, \ldots, G_g$$

részhalmazokra $(\underline{j}_1 < \underline{j}_2 \ldots < \underline{j}_r \in \{\underline{i}_1, \underline{i}_2, \ldots, \underline{i}_k\})$

$$S_1 := j_1,$$
$$S_2 := NS[S_1] + 1 + j_2 - j_1,$$
$$\vdots$$
$$S_r := NS[S_{r-1}] + 1 + j_r - j_{r-1}$$

{a $\underline{G}_v$ részhalmazok fenti sorrendje miatt $\underline{NS}[\underline{S}_1]$, $\underline{NS}[\underline{S}_2], \ldots, \underline{NS}[\underline{S}_{r-1}] > 0$}, $\underline{NS}[\underline{S}_r] = 0$ esetén

$$NS[S_r] := EU,$$
$$EU := EU + 2 + n - j_r,$$

továbbá

$$NS[NS[S_r]] := NS[NS[S_r]] + \sigma(G_g)$$

{a feltöltés befejezése után $\underline{NS}[\underline{NS}[\underline{S}_r]]$-ben az $\underline{\sigma}_T(\underline{G}_v)$ gyakoriság értékét kapjuk meg}, és $\underline{G}_v = \underline{G}_g$ esetén

$$NS[NS[S_r]+1]:=\sigma(G_g).$$

Megmutatom, hogy igaz a következő állitás: a fentiek szerint feltöltött $\underline{NS}$ tömbből tetszőleges

$$G_g = \{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}, \quad 1 \leq k \leq n, \quad 1 \leq g \leq 2^n-1,$$

esetén az $\underline{\sigma}_T(\underline{G}_g)$, $\underline{\sigma}(\underline{G}_g)$ gyakoriságok legfeljebb $(\underline{k}+1)$ lépésben kikereshetőek. Definiáljuk ugyanis a következő $\underline{T}_j$ sorozatot:

$$\begin{aligned}
&T_1:=NS[i_1]+1+i_2-i_1,\\
&T_j:=NS[T_{j-1}]+1+i_{j+1}-i_j, \quad j = 2, 3, \ldots, k-1,\\
&T_k:=NS[T_{k-1}].
\end{aligned}$$

Ha bármelyik $\underline{T}_j$ értéke 0, akkor $\underline{\sigma}_T(\underline{G}_g) = \underline{\sigma}(\underline{G}_g) = 0$, egyébként pedig $\underline{\sigma}_T(\underline{G}_g) = \underline{NS}[\underline{T}_k]$ és $\underline{\sigma}(\underline{G}_g) = \underline{NS}[\underline{T}_k+1]$. Ezzel az állitás már adódik.

Mint láttuk, az $\underline{NS}$ tömbben tetszőleges

$$G_g = \{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}$$

jellegkombináció esetén az $[\sigma_T(G_g), \sigma(G_g)]$ számpárt az $(n-i_k)$ számu

$$\sigma_T(i_1,i_2,\ldots,i_k,i_k+1),\ \sigma_T(i_1,i_2,\ldots,i_k+2),$$
$$\ldots,\ \sigma_T(i_1,i_2,\ldots,i_k,n)$$

gyakoriság cimei követik (minél nagyobb i_k értéke, annál kevesebb cim). Ebből következik, hogy NS méretét minimalizálandó annál nagyobbnak célszerü választani valamely A_i jelleg sorszámát, minél több olyan különböző G_g jellegkombinációban szerepel, amelyre $\sigma_T(G_g) > 0$, vagyis minél nagyobb az

$$N_T(i) = \sum_{\substack{g: A_i \in G_g, \\ \sigma_T(G_g)>0}} 1$$

kifejezés értéke. Viszont *a priori* nem ismerjük az $N_T(i)$ mennyiségeket, csak az $\sigma(i)$ gyakoriságokat. Mivel azonban a két mennyiség általában pozitivan, méghozzá erősen pozitivan korrelált, azért az NS tömb mérete az esetek többségében, tehát átlagosan csökkenthető azzal, hogy a memóriában az A_i jellegeket az $\sigma(i)$ gyakoriságok növekvő nagysága szerint rendezzük.

NS mérete, ami függ az A_i jellegek sorrendjétől, de a G_g jellegkombinációkétól nem, további két módon is csök-

kenthető:

i) ha az $\sigma_T(G_g)$ gyakoriságok várható nagyságrendje ezt megengedi, NS minden j számú elemébe $k(>j)$ számú értéket helyezünk el;

ii) a nagyméretü G_g-khez, ha ilyen kevés van, manuálisan határozzuk meg az $\sigma_T(G_g)$, $\sigma(G_g)$ gyakoriságokat.

$k = 1, 2, \ldots, n$ esetén jelölje $\sigma^{(k)}$ azon objektumok számát, amelyek pontosan k számu jelleggel rendelkeznek, akkor $r = 1, 2, \ldots, k$ mellett a pontosan k számu jelleggel rendelkező objektumok körében előforduló r méretü jellegkombinációk száma $\binom{k}{r}\sigma^{(k)}$, az összes (M számu) objektum körében előfordulóké pedig

$$\sigma_T^{(r)} = \sum_{k=r}^{n} \binom{k}{r}\sigma^{(k)}$$

($r = 1, 2, \ldots, n$). Legyen $\sigma = \sigma^{(0)}$, akkor $\sigma_T^{(1)}$ értéke azt mutatja, hogy az $(M-\sigma)$ számu olyan objektumnak, amelyik rendelkezik legalább egy jelleggel, összesen hány jellege van. Ez azt jelenti, hogy amig egy tetszőleges objektumnak átlagosan $[\sigma_T^{(1)}/M]$, addig egy olyannak, amelyik rendelkezik legalább egy jelleggel, átlagosan $[\sigma_T^{(1)}/(M-\sigma)]$ jellege van.

$k = 0, 1, \ldots, n$ esetén az elméletileg lehetséges különböző k méretü jellegkombinációk száma $\binom{n}{k}$. Jelölje $N_T^{(k)}$

az előfordulók számát, $\underline{N}^{(k)}$ pedig ezek közül azokét, amelyek a pontosan $\underline{k}$ számu jelleggel rendelkező objektumok körében fordulnak elő. Nyilvánvaló módon

$$N^{(k)} \leq N_T^{(k)} \leq \binom{n}{k}.$$

Ha a jellegek között vannak egymást definició szerint kizáróak, akkor

$$N_T^{(k)} < \binom{n}{k}.$$

Jelölje $\underline{N}$ a jellegek alapján különböző, legalább 2 jelleggel rendelkező objektumok számát, akkor

$$N = \sum_{k=2}^{n} N^{(k)}.$$

Az $\underline{\sigma}(\underline{G}_g)$, $\underline{\sigma}_T(\underline{G}_g)$, $\underline{\sigma}^{(k)}$, $\underline{N}^{(k)}$ és $\underline{N}_T^{(k)}$ mennyiségeket a következő, függetlenségvizsgálattal foglalkozó fejezetben fogom felhasználni.

2. A VÁLTOZÓK FÜGGETLENSÉGÉNEK VIZSGÁLATA

Tetszőleges

$$G_g = \{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}, \quad 1 \leq k \leq n, \quad 1 \leq g \leq 2^n - 1,$$

jellegkombináció esetén jelölje

$$P_T(G_g) = P_T(i_1, i_2, \ldots, i_k)$$

annak valószínüségét, hogy egy objektum rendelkezik a G_g-hez tartozó jellegekkel és esetleg továbbiakkal is. Jelölje $P^{(k)}$ annak valószínüségét, hogy egy objektum pontosan k számu jelleggel rendelkezik ($k = 0, 1, \ldots, n$), és legyen $P = P^{(0)}$. Vezessük még be a következő jelöléseket:

$$q_i = \frac{P_T(i)}{1 - P_T(i)}, \quad i = 1, 2, \ldots, n,$$

$$S_0 = 1,$$

$$S_k = \sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq n} \prod_{j=1}^{k} q_{i_j}, \quad k = 1, 2, \ldots, n.$$

Tegyük most fel, hogy a $\underline{W}_i$ változók függetlenek. Ekkor tetszőleges $\underline{k} = 1, 2, \ldots, \underline{n}$ és $1 \leq \underline{i}_1 < \underline{i}_2 < \ldots < \underline{i}_k \leq \underline{n}$ esetén

$$P_T(i_1, i_2, \ldots, i_k) = \prod_{j=1}^{k} P_T(i_j).$$

[Az a kérdés, hogy a $\underline{W}_i$-k függetlenek-e, elméletileg a fenti egyenlőségeket feltételező nullhipotézishez tartozó χ^2-próbával dönthető el, a megfelelő 2^n-es kontingenciatáblázat alapján. A számomra érdekes esetekben azonban - még a ritka jellegek lehetőség szerinti összevonása után is - 2^n nagysága és/vagy a $\underline{P}_T(\underline{i})$ valószinüségek kicsisége miatt a χ^2-próba elvégzése illuzórikus. A függetlenséget tehát máshogyan kell ellenőrizni.] $\underline{P}^{(k)}$ és $\underline{S}_k$ definiciójából egyszerü számolással következik, hogy

$$P^{(k)} = PS_k, \quad k = 0, 1, \ldots, n,$$

tehát az $\underline{\sigma}^{(k)}/\underline{M}$ relativ gyakoriságoknak megfelelő $\underline{P}^{(k)}$ valószinüségek előállitásához $\underline{P}$-n kivül elég az $\underline{S}_k$-kat kiszámitani, amelyek - a $\underline{q}_i$-ken keresztül - csak a $\underline{P}_T(\underline{i})$ valószinüségektől függenek. [Az $\underline{S}_k$-k rekurzivan, Newton formulájával számithatók:

$$S_k = \frac{1}{k} \sum_{j=1}^{k} (-1)^{j-1} T_j S_{k-j}, \quad k = 1, 2, \ldots, n,$$

ahol

$$T_j = \sum_{i=1}^{n} q_i^j, \quad j = 1, 2, \ldots, n;$$

lásd pl. Kuros (1956)].

A jellegkombinációk mérete tapasztalati eloszlásának jellemzésében az $\underline{N}_T^{(k)}$, $\underline{N}^{(k)}$ mennyiségek is fontos szerepet játszanak. Jelölje $\widetilde{\underline{N}}_T^{(k)}$, ill. $\widetilde{\underline{N}}^{(k)}$ a várható értéküket, akkor $\widetilde{\underline{N}}^{(k)}$ és $\widetilde{\underline{N}}_T^{(k)}$ az

$$\{\sigma(i_1, i_2, \ldots, i_k) > 0\},$$

ill. az

$$\{\sigma_T(i_1, i_2, \ldots, i_k) > 0\}$$

események valószínüségeinek $1 \leq \underline{i}_1 < \underline{i}_2 < \ldots < \underline{i}_k \leq \underline{n}$ mellett vett összege. Jelölje $\underline{E}(\underline{i}_1, \underline{i}_2, \ldots, \underline{i}_k)$, ill. $\underline{E}_T(\underline{i}_1, \underline{i}_2, \ldots, \underline{i}_k)$ az $\underline{\sigma}(\underline{i}_1, \underline{i}_2, \ldots, \underline{i}_k)$, $\underline{\sigma}_T(\underline{i}_1, \underline{i}_2, \ldots, \underline{i}_k)$ változók várható értékét. $\underline{\sigma}(\underline{i}_1, \underline{i}_2, \ldots, \underline{i}_k)$ és $\underline{\sigma}_T(\underline{i}_1, \underline{i}_2, \ldots, \underline{i}_k)$ binomiális eloszlását $\underline{E}(\underline{i}_1, \underline{i}_2, \ldots, \underline{i}_k)$, ill. $\underline{E}_T(\underline{i}_1, \underline{i}_2, \ldots, \underline{i}_k)$ paraméterü Poisson-eloszlással közelitve a fenti két összeg közelitőleg

$$\sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq n} \{1 - \exp[-E(i_1, i_2, \ldots, i_k)]\},$$

ill.

$$\sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} \{1 - \exp[-E_T(i_1, i_2, \ldots, i_k)]\}$$

lesz. [Ez abból következik, hogy tetszőleges, λ paraméterü Poisson-eloszlást követő ξ valószinüségi változó esetén

$$P(\xi > 0) = 1 - \exp(-\lambda).]$$

A változók függetlensége miatt tetszőleges $1 \le \underline{i}_1 < \underline{i}_2 < \ldots < \underline{i}_k \le \underline{n}$, $1 \le \underline{k} \le \underline{n}$ mellett egyrészt

$$E(i_1, i_2, \ldots, i_k) = MP \prod_{j=1}^{k} q_{i_j},$$

másrészt

$$E_T(i_1, i_2, \ldots, i_k) = M \prod_{j=1}^{k} P_T(i_j),$$

tehát ekkor $\tilde{\underline{N}}^{(k)}$ és $\tilde{\underline{N}}_T^{(k)}$ fenti közelitése is csupán a $\underline{P}$ és $\underline{P}_T(\underline{i})$ $(\underline{i} = 1, 2, \ldots, \underline{n})$ valószinüségektől függ.

Tetszőleges $\underline{A}_i$ jelleg esetén jelölje $\underline{P}(\underline{i})$ annak valószinüségét, hogy egy objektum rendelkezik $\underline{A}_i$-vel, de más jelleggel nem, és legyen $\underline{e}_{gi}$ a $\underline{W}_i$ változónak az $\underline{y}_g$ objektumon megfigyelt - 1 vagy 0 - értéke. A változók függetlensége miatt

$$P(e_{gj}=0,\ j\neq i \mid e_{gi}=1) = P(e_{gj}=0,\ j\neq i \mid e_{gi}=0),$$

vagyis

$$\frac{P(i)}{P_T(i)} = \frac{P}{1-P_T(i)},$$

ahonnan

$$P_T(i) = \frac{P(i)}{P+P(i)}.$$

Legyen $\underline{\sigma} = \underline{\sigma}^{(0)}$, és becsüljük a $\underline{P}_T(\underline{i})$ valószinüségeket ugy, hogy $\hat{\underline{P}} = \underline{\sigma} / \underline{M}$ és

$$\hat{P}(i) = \frac{\sigma(i)}{M}, \quad i = 1, 2, \ldots, n,$$

teljesüljön. Ekkor

$$\hat{P}_T(i) = \frac{\sigma(i)}{\sigma+\sigma(i)}, \quad i = 1, 2, \ldots, n.$$

Ezekből a $\hat{\underline{P}}_T(\underline{i})$-kből a Newton-formula segitségével előállitom az $\hat{\underline{S}}_k$ becsléseket, és a $\underline{P}^{(k)}$ valószinüségek becslését a

$$\hat{P}^{(k)} = \hat{P}\hat{S}_k$$

egyenlőséggel határozom meg. Mivel a változók függetlensége miatt

$$P = \prod_{i=1}^{n} [1-P_T(i)],$$

azért teljesülnie kell, hogy

$$\frac{\sigma'}{M} = \hat{P} = \prod_{i=1}^{n} [1-\hat{P}_T(i)] = \prod_{i=1}^{n} [1 - \frac{\sigma(i)}{\sigma+\sigma(i)}] = \prod_{i=1}^{n} \frac{\sigma}{\sigma+\sigma(i)},$$

ahonnan

$$\sigma' = M \prod_{i=1}^{n} \frac{\sigma}{\sigma+\sigma(i)}. \qquad (2.1)$$

A $\underline{P}_T(\underline{i})$ valószinüségek efficiens becslései az $[\underline{\sigma}_T(\underline{i}) / \underline{M}]$ relativ gyakoriságok. Mellettük a változók függetlensége esetén az

$$\sigma' = M \prod_{i=1}^{n} \frac{M-\sigma_T(i)}{M} \qquad (2.2)$$

egyenlőségnek kell teljesülnie. Az ebben szereplő $\underline{\sigma}_T(\underline{i})$ gyakoriságok azonban - ellentétben a (2.1) egyenlőségben szereplő $\underline{\sigma}(\underline{i})$ gyakoriságokkal - általában nem alapadatok, hanem csak a számitógépes feldolgozás során kerülnek meghatározásra. Ebből következik, hogy mig a (2.1) egyenlőség

jobboldalát általában közvetlenül ki tudjuk számitani, a (2.2) egyenlőségét általában nem. Ezért a (2.1) egyenlőség ellenőrzése lehet a függetlenség vizsgálatának első lépése, amelyet az $\{\underline{\sigma}^{(k)}\}$, $\{\underline{N}^{(k)}\}$, $\{\underline{N}_T^{(k)}\}$, ill. $\{\underline{\hat{MP}}^{(k)}\}$, $\{\underline{\hat{N}}^{(k)}\}$, $\{\underline{\hat{N}}_T^{(k)}\}$ sorozatok összehasonlitása követhet. Döntő az $\{\underline{\sigma}^{(k)}\}$ és $\{\underline{\hat{MP}}^{(k)}\}$ sorozatok összehasonlitása, ami a megfelelő kontingenciatáblázat kisgyakoriságu celláinak radikális összevonását jelenti. Valószinüleg lehetséges az ilyen cellák kevésbé radikális összevonása is; ez további kutatás tárgya lehet. - Az összehasonlitást a

$$\sum_{k=0}^{n} \frac{[\sigma^{(k)} - \hat{MP}^{(k)}]^2}{\hat{MP}^{(k)}} \qquad (2.3)$$

kifejezés alapján végzem el. Ennek aszimptotikus viselkedésével kapcsolatban a következők mondhatók. Ha a $\underline{P}^{(k)}$ valószinüségek ismertek lennének, akkor a

$$\sum_{k=0}^{n} \frac{[\sigma^{(k)} - MP^{(k)}]^2}{MP^{(k)}}$$

kifejezés aszimptotikusan $\underline{n}$ szabadságfoku χ^2-eloszlást követne [lásd pl. Vincze (1968), 143-144. oldal]. Esetünkben azonban nem ismertek, becsüljük őket. Ilyenkor (lásd u.o.,

148. oldal) a szabadságfokot a becsült paraméterek számával csökkenteni kell. Az a sejtésem, hogy a valószinüségek általam ismertetett becslése két paraméter becslésére vezethető vissza, és igy a (2.3) kifejezés aszimptotikusan (n-2) szabadságfoku X^2-eloszlást követ, de ezt bizonyitani nem tudom. Független változók mellett generált véletlen adatmezőkön végzett számitásaim alátámasztják ezt a sejtést.

Ha a változók függetlensége elfogadhatatlannak bizonyul, felmerülhet az a gondolat, hogy ezt csak néhány jelleg okozza. Ezért célszerü megvizsgálni, hogyan változik az egyes jellegek adott méretü jellegkombinációk közötti gyakorisága a méret változásával. k, i = 1, 2, ..., n esetén jelölje $\sigma^{(k)}(i)$ az olyan objektumok számát, amelyek pontosan k számu jelleggel rendelkeznek, köztük A_i-vel. Tekintsük az $\sigma^{(1)}(i)$, $\sigma^{(2)}(i)$, ..., $\sigma^{(n)}(i)$ értékeket (i = 1, 2, ..., n). i = 1, 2, ..., n mellett legyen $p_i(k)$ a megfelelő

$$\sigma^{(k)}(i) \Big/ \sum_{i=1}^{n} \sigma^{(k)}(i)$$

relativ gyakoriság [annak relativ gyakorisága, hogy egy k számu jelleggel rendelkező objektum rendelkezik A_i-vel, épen $kp_i(k)$, k = 1, 2, ..., n]. Legyen m_i a $p_i(k)$-k

$$\sum_{k=1}^{n} \sigma^{(k)}(i)p_i(k) \;/\; \sum_{k=1}^{n} \sigma^{(k)}(i)$$

sulyozott átlaga, a_i az $\sigma^{(k)}(i)$ multiplicitásu $p_i(k)$ pontokhoz ($k = 1, 2, \ldots, n$) a legkisebb négyzetek módszerével illesztett $r_i(k)$ egyenes meredeksége, és határozzuk meg az

$$S_i = \sum_{k=1}^{n} \sigma^{(k)}(i) \left\{ \frac{[p_i(k)-m_i]^2}{m_i} - \frac{[p_i(k)-r_i(k)]^2}{r_i(k)} \right\} \quad (2.4)$$

kifejezés értékét ($i = 1, 2, \ldots, n$). a_i azt mutatja, hogyan változik a $p_i(k)$ relativ gyakoriság az A_i-t tartalmazó jellegkombináció méretének növekedésével, S_i pedig azt, hogy mennyivel becsüli $r_i(k)$ jobban $p_i(k)$-t, mint m_i. Ha az adott méretü jellegkombinációk közötti egyes jelleggyakoriságok a méret változásával lényegében egymáshoz hasonlóan változnak, akkor a változók függetlenségének elfogadhatatlanságát nem csupán néhány jelleg okozza.

A változók függetlensége a (2.3) kifejezés alapján akkor bizonyul elfogadhatatlannak, ha a nagyobb méretü jellegkombinációk tul sokan vagy tul kevesen vannak. Az, hogy ezt csak néhány jelleg okozza, azt jelenti, hogy az ezen jellegeket tartalmazó kombinációkból van tul sok vagy tul kevés. Ekkor azonban ezeknek a jellegeknek az adott méretü

jellegkombinációk közötti gyakoriságai a méret változásával a többi jellegéitől eltérően változnak. Ezt a változást méri - az egyes jellegekre - az $\underline{S}_i$ kifejezés. Ennek aszimptotikus viselkedéséről az alábbiakat lehet mondani. Legyen

$$S_i^{(1)} = \sum_{k=1}^{n} \sigma^{(k)}{}_{(i)} \frac{[p_i(k)-m_i]^2}{m_i}$$

és

$$S_i^{(2)} = \sum_{k=1}^{n} \sigma^{(k)}{}_{(i)} \frac{[p_i(k)-r_i(k)]^2}{r_i(k)},$$

akkor $\underline{S}_i = \underline{S}_i^{(1)} - \underline{S}_i^{(2)}$. Ha az $\underline{m}_i$ és $\underline{r}_i(\underline{k})$ maximum likelihood becslések helyett a megfelelő χ^2 minimum becslések állanának [$\underline{m}_i$ esetén a $\underline{p}_i(\underline{k})$-k

$$\sum_{k=1}^{n} \sigma^{(k)}{}_{(i)} p_i^2(k) \; / \; \sum_{k=1}^{n} \sigma^{(k)}{}_{(i)}$$

sulyozott négyzetátlagának négyzetgyöke, $\underline{r}_i(\underline{k})$ esetén explicit módon nem felirható], akkor $\underline{S}_i^{(1)}$ és $\underline{S}_i^{(2)}$ aszimptotikusan χ^2-eloszlást követne [lásd pl. Cramér (1946)]. A megfelelő hipotéziseket $\underline{H}_1$, $\underline{H}_2$-vel jelölve nyilvánvaló módon $\underline{H}_1 \subset \underline{H}_2$. Ismeretes, hogy bizonyos feltételek mellett i-

lyenkor a maximum likelihood-ok különbségének kétszerese aszimptotikusan r szabadságfoku χ^2-eloszlást követ [lásd pl. Serfling (1980); r a paraméterek számának különbsége, esetünkben 2 - 1 = 1]. A χ^2 minimumokra azonban nem találtam hasonló állitást.

3. A NORMÁLIS KÜSZÖB MODELL

A küszöb modell feltételezi, hogy a vizsgált dichotom jellegeknek mindegyik objektumra nézve van egy-egy valós számmal kifejezhető mértéke, vagyis a W_i bináris változóhoz hozzá van rendelve egy L_i folytonos háttér változó ($i = 1, 2, \ldots, n$). A küszöb modell szerint L_i folytonos valószínüségi változó, amelyhez egy T_i küszöb tartozik. Valamely objektum akkor és csak akkor rendelkezik az A_i jelleggel, ha L_i rajta megfigyelt értéke nem kisebb T_i-nél. Feltesszük, hogy az L_i változók átlaga 0, szórása 1 (mivel a háttér változók általában nem megfigyelhetőek, ez a technikai jellegü feltétel nem jelent megszoritást).

Tetszőleges

$$G_g = \{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}, \quad 1 \le k \le n, \quad 1 \le i_1 < i_2 < \ldots < i_k \le n,$$

jellegkombináció a küszöb modellben ekvivalens a

$$\tilde{G}_g = \begin{Bmatrix} L_{i_j} \ge T_{i_j}, & j = 1, 2, \ldots, k; \\ L_r < T_r, & r \ne i_1, i_2, \ldots, i_k \end{Bmatrix}$$

eseménnyel, ezért

$$P(G_g) = P(\tilde{G}_g)$$

és

$$P_T(G_g) = P(L_{i_j} \geq T_{i_j}, \quad j = 1, 2, \ldots, k),$$

ahol $P(G_g)$ annak valószinüségét jelöli, hogy egy objektum rendelkezik a G_g-hez tartozó jellegekkel, de más jelleggel nem.

A normális küszöb modell a fentieken kivül feltételezi, hogy az

$$(L_1, L_2, \ldots, L_n)$$

valószinüségi (vektor)változó (n-dimenziós) normális eloszlásu (ezt a feltevést alátámaszthatja például a centrális határeloszlás-tétel). Mivel az L_i-k feltevés szerint standardak, azért együttes eloszlásukat meghatározzák az

$$r_{ij} = E(L_i L_j), \quad 1 \leq i < j \leq n,$$

korrelációs együtthatók. A normális küszöb modell paraméterei tehát a T_i küszöbök és az r_{ij}-k, számuk összesen

$$n + \binom{n}{2} = \binom{n+1}{2}.$$

Maximum likelihood (ML) becslésük $n > 2$ esetén nem ismert.

A fenti, többdimenziós normális küszöb modell veleszületett rendellenességek (CA-k) öröklődését leiró speciális változatának [amely egy CA több családtagra vonatkozó hajlamát irja le, és amelyben d-edfoku rokonok hajlamainak

($\underline{h}^2 / 2^d$) a korrelációs együtthatója] eredetéről az 5.1. paragrafusban irok. Bármilyen meglepő viszont, de a fenti, általánosan megfogalmazott modellel nem találkoztam az irodalomban. További vizsgálatát, mindenekelőtt a korrelációs mátrix maximum likelihood becslésének - számomra igen nehéznek tünő - meghatározását fontos és érdekes feladatnak tartom.

A $\underline{T}_i$ küszöböket ugy becslem, hogy $\hat{\underline{T}}_i$ az

$$MP_T(i) = \sigma_T(i)$$

egyenlet megoldása, ahonnan

$$\hat{T}_i = \Phi^{-1}\left[1 - \frac{\sigma_T(i)}{M}\right], \quad i = 1, 2, \ldots, n$$

($\underline{n} = 1$ esetén ez a küszöb ML becslése). A korrelációs együtthatókat az

$$MP_T(i,j) = \sigma_T(i,j), \quad 1 \leq i < j \leq n,$$

egyenletekből becslem ($\underline{n} = 2$ esetén a fenti $\hat{\underline{T}}_i$-kkal ez az ML becslést adja). Ekkor $\underline{r} = \hat{\underline{r}}_{ij}$ az

$$F(\hat{T}_i, \hat{T}_j, r) = \frac{\sigma_T(i,j)}{M} \qquad (3.1)$$

egyenlet megoldása $(1 \le i < j \le n)$, ahol tetszőleges véges valós $\tilde{T}$, T és $-1 < r < 1$ esetén

$$F(\tilde{T},T,r) = \frac{1}{2\pi\sqrt{1-r^2}} \int_{\tilde{T}}^{\infty}\int_{T}^{\infty} \exp\left[-\frac{u^2-2ruv+v^2}{2(1-r^2)}\right] du\, dv,$$

$r = \pm 1$ esetén

$$F(\tilde{T},T,r) = Q[\max(\tilde{T},T)],$$

továbbá

$$Q(z) = 1 - \Phi(z), \quad -\infty < z < \infty.$$

Jelölje $Q^{(k)}$ a Q függvény k-adik deriváltját $(k = 0, 1, \ldots)$, akkor $|r| < 1$ esetén a Pearson-sorfejtés szerint

$$F(\tilde{T},T,r) = \sum_{k=0}^{\infty} Q^{(k)}(\tilde{T})Q^{(k)}(T)\,\frac{r^k}{k!} \qquad (3.2)$$

[lásd pl. Anderson (1958)]. $F(\tilde{T},T,r)$ gyors számítógépes meghatározását a következőképpen végzem el. Legyen

$$s = \sqrt{1-r^2}, \qquad P_r(\tilde{T},T) = \frac{F(\tilde{T},T,r)}{Q(T)}$$

és $\varphi(z)$ a standard normális sürüségfüggvény.

3.1. TÉTEL. Ha

$$t = \frac{r}{s}, \qquad Z = \frac{\tilde{T}-rT}{s},$$

akkor $|r| < 1/\sqrt{2}$ esetén

$$P_r(\tilde{T},T) = Q(Z) + \varphi(Z) \sum_{k=1}^{\infty} a_k b_k \frac{t^k}{k!},$$

ahol a_k és b_k az

$$a_1 = 1, \quad a_2 = Z, \ \ldots, \ a_k = Z a_{k-1} - (k-2) a_{k-2},$$

$$b_1 = \frac{\varphi(T)}{Q(T)} - T, \quad b_2 = 1 - T b_1, \ \ldots, \ b_k = (k-1) b_{k-2} - T b_{k-1}$$

rekurziókkal számolható.

Bizonyitás.

$$P_r(\tilde{T},T) = \frac{1}{Q(T)} \int_T^{\infty} \varphi(v) \int_{(\tilde{T}-rv)/s}^{\infty} \varphi(z)\, dz\, dv =$$

$$= Q(Z) + \frac{1}{Q(T)} \int_T^{\infty} \varphi(v) \int_0^{t(v-T)} \varphi(Z-z)\, dz\, dv.$$

Jelöljük $\underline{h}(\underline{t})$-vel a második tagot, akkor

$$h(t) = \frac{1}{Q(T)} \int_0^{\infty} \varphi(T+u) \int_0^{tu} \varphi(Z-v)\, dv\, du.$$

Rögzitett $\underline{T}$ és $\underline{Z}$ mellett fejtsük Taylor-sorba a $\underline{h}(\underline{t})$ függvényt $\underline{t} = 0$ körül. Ekkor $\underline{h}(0) = 0$, továbbá

$$h'(t) = \frac{1}{Q(T)} \int_0^{\infty} u\varphi(T+u)\varphi(Z-tu)\, du,$$

$$\frac{h'(0)}{\varphi(Z)} = \frac{\varphi(T)}{Q(T)} - T = b_1 = 1b_1 = a_1 b_1.$$

A második deriváltra

$$h''(t) = \frac{1}{Q(T)} \int_0^{\infty} u^2 \varphi(T+u)(Z-tu)\varphi(Z-tu)\, du$$

adódik, általában pedig

$$h^{(k)}(t) = \frac{1}{Q(T)} \int_0^{\infty} u^k \varphi(T+u) B_k(Z-tu)\, du,$$

ahol egyrészt a $\underline{B}_k$ függvény olyan, hogy $\underline{B}_k(\underline{Z}-\underline{tu})_{t=0}$ nem függ $\underline{u}$-tól és $\varphi(\underline{Z})\underline{a}_k$-val egyenlő, másrészt

$$\int_0^\infty u^k\varphi(T+u)\,du = \int_0^\infty u^{k-1}(T+u)\varphi(T+u)\,du -$$

$$- T\int_0^\infty u^{k-1}\varphi(T+u)\,du = \ldots = b_k.$$

Ezek után bebizonyitom, hogy $|\underline{r}|<1/\sqrt{2}$ esetén a Taylor-sor konvergens. Nyilvánvaló módon

$$|r|<\frac{1}{\sqrt{2}} \longleftrightarrow |t|<1.$$

Szükségem lesz a következő két lemmára.

3.1. LEMMA. *Ha* $\underline{c}>0$ *és* $\underline{d}>1$ *valós számok, akkor létezik olyan* $\underline{k}^*$ *természetes szám, hogy* $\underline{k}>\underline{k}^*$ *esetén*

$$c^2d^2 + 2cd\sqrt{k-1} - (k-1) \le d^4k,$$

$$cd(d^2+1) + \sqrt{k-1} \le d^4\sqrt{k+1}.$$

Bizonyitás. Legyen

$$f(k) = (d^4+1)k - 2cd\sqrt{k-1} - (c^2d^2+1),$$

$$g(k) = d^4\sqrt{k+1} - \sqrt{k-1} - cd(d^2+1).$$

Nyilvánvaló módon

$$\lim_{k\to\infty} f(k) = \lim_{k\to\infty} g(k) = \infty,$$

amiből már adódik is a lemma.

3.2. LEMMA. *Legyen* $\{a_k\}$ *és* $\{b_k\}$ *a 3.1. Tételben szereplő számsorozat,* $c>0$ *és* $d>1$ *valós számok,* k^* *a 3.1. Lemmában szereplő természetes szám. Ha* $C>0$ *olyan valós szám, hogy* $k \le \max(2,k^*)$ *esetén*

$$|a_k b_k| \le Cd^{2k}k!,$$

$$|a_k b_{k+1}|,\ |a_{k+1} b_k| \le Cd^{2k+1}k!\sqrt{k+1},$$

akkor a két egyenlőtlenség tetszőleges k*-ra teljesül.*

Bizonyitás. A lemma teljes indukcióval adódik a 3.1. Lemmából.

Legyen $c = \max(|Z|,T)$ és $1 < d < \sqrt{1/|t|}$. A 3.2. Lemma miatt tetszőleges k-ra

$$|a_k b_k \frac{t^k}{k!}| \le C(d^2|t|)^k,$$

márpedig $\underline{d}$ definiciójából következik, hogy $\underline{d}^2|\underline{t}| < 1$. Ezzel a tétel már adódik.

Az exponenciális tulcsordulás eékerülése végett célszerü a következő rekurzióval dolgozni:

$$e_1 = tb_1;$$

$$c_2 = Ze_1; \qquad D_2 = \frac{t-Te_1}{2}; \qquad e_2 = tZD_2;$$

$$\cdot \qquad \cdot \qquad \cdot$$

$$\cdot \qquad \cdot \qquad \cdot$$

$$\cdot \qquad \cdot \qquad \cdot$$

$$c_k = Ze_{k-1} - (k-2)tD_{k-1}; \qquad D_k = \frac{tc_{k-1}-Te_{k-1}}{k}; \qquad e_k =$$

$$= t[ZD_k + \frac{(k-2)t(TD_{k-1}-e_{k-2})}{k}].$$

3.3. LEMMA. Legyen $\{\underline{a}_k\}$ és $\{\underline{b}_k\}$ a 3.1. Tételben szereplő, $\{\underline{e}_k\}$ a fent definiált számsorozat, akkor

$$e_k = a_k b_k \frac{t^k}{k!}, \quad k = 1, 2, \ldots$$

Bizonyitás. A lemma teljes indukcióval adódik.

$\underline{d}$ és $\underline{C}$ egy lehetséges választása a következő:

$$d = \sqrt[4]{1/|t|}, \quad C = \max(C_1, C_2, \ldots, C_{k^*}),$$

ahol

$$C_k = \max[\,|a_k b_k|/(d^{2k}k!),\ |a_k b_{k+1}|/(d^{2k+1}k!\sqrt{k+1}),$$
$$|a_{k+1} b_k|/(d^{2k+1}k!\sqrt{k+1})], \quad k = 1, 2, \ldots, k^*.$$

Legyen ε_1 megfelelően kis pozitiv szám és

$$K = \min[k:\ k > k^*, \quad C(\sqrt{|t|})^k < \varepsilon_1],$$

akkor

$$F(\tilde{T},T,r) \approx Q(T)[Q(Z)+\varphi(Z)\sum_{k=1}^{K} e_k], \quad |r| < \frac{1}{\sqrt{2}}.$$

Az $\underline{F}(\tilde{\underline{T}},\underline{T},\underline{r})$ mennyiség $|\underline{r}| \geq 1/\sqrt{2}$ mellett történő meghatározásához szükségem lesz a következő, egyszerü számolással adódó állitásokra:

i) $$F(\tilde{T},T,r) + F(\tilde{T},-T,-r) = Q(\tilde{T});$$

ii) $$F(\tilde{T},T,r) = F(\tilde{T},\frac{T-\tilde{T}}{\sqrt{2(1-r)}},-\sqrt{\frac{1-r}{2}}) + F(\frac{\tilde{T}-T}{\sqrt{2(1-r)}},T,-\sqrt{\frac{1-r}{2}});$$

iii) $|\underline{r}| > 1/2$, és igy $|\underline{r}| \geq 1/\sqrt{2}$ esetén

$$\sqrt{\frac{1-|r|}{2}} < \frac{1}{2}.$$

A fenti állitások alapján megfelelően kis $\varepsilon_2 > 0$ mellett, $1/2 < \underline{r} < 1-\varepsilon_2$ és igy $1/\sqrt{2} \leq \underline{r} < 1-\varepsilon_2$ esetén

$$F(\tilde{T},T,r) = F(\tilde{T},\frac{T-\tilde{T}}{\sqrt{2(1-r)}},-\sqrt{\frac{1-r}{2}}) + F(\frac{\tilde{T}-T}{\sqrt{2(1-r)}},T,-\sqrt{\frac{1-r}{2}}),$$

$1/2 < -\underline{r} < 1-\varepsilon_2$ és igy $1/\sqrt{2} \leq -\underline{r} < 1-\varepsilon_2$ esetén

$$F(\tilde{T},T,r) = Q(\tilde{T}) -$$

$$- F(\tilde{T},-\frac{\tilde{T}+T}{\sqrt{2(1+r)}},-\sqrt{\frac{1+r}{2}}) - F(\frac{\tilde{T}+T}{\sqrt{2(1+r)}},-T,-\sqrt{\frac{1+r}{2}}),$$

$|\underline{r}| \geq 1-\varepsilon_2$ esetén pedig

$$F(\tilde{T},T,r) \approx Q[\max(\tilde{T},T)].$$

A (3.1) egyenleteket kielégitő $\underline{r} = \hat{\underline{r}}_{ij}$ korrelációs együttható becsléseket explicit módon nem tudom meghatározni. Mivel azonban $\underline{F}(\tilde{\underline{T}},\underline{T},\underline{r})$ $\underline{r}$-ben monoton növő, az egyenletek numerikus megoldása a fenti eljárás birtokában nem okoz nehézséget.

A normális küszöb modell kézenfekvő ellenőrzése annak vizsgálata, mennyire felel meg a 2-nél nagyobb méretü jellegkombinációk előfordulása a modellnek. Tetszőleges

$$G_g = \{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}, \quad 3 \le k \le n, \quad 1 \le i_1 < i_2 < \ldots < i_k \le n,$$

jellegkombináció mellett

$$|\hat{r}_{j_1 j_2}| < \frac{1}{k-1}, \quad j_1, j_2 = i_1, i_2, \ldots, i_k; \quad j_1 \neq j_2, \tag{3.3}$$

esetén

$$\hat{P}_T(G_g) =$$

$$= \sum_{v_{12}=0}^{\infty} \sum_{v_{13}=0}^{\infty} \cdots \sum_{v_{k-1,k}=0}^{\infty} \{ \Big(\prod_{j_1=1}^{k-1} \prod_{j_2=j_1+1}^{k} \frac{\hat{r}_{i_{j_1} i_{j_2}}^{v_{j_1 j_2}}}{v_{j_1 j_2}!} \Big) \times \tag{3.4}$$

$$\times \Big[\prod_{j=1}^{k} Q^{(u_j)}(T_{i_j}) \Big] \},$$

ahol

$$u_j = \sum_{j_1=1}^{j-1} v_{j_1 j} + \sum_{j_2=j+1}^{k} v_{j j_2}, \quad j = 1, 2, \ldots, k$$

[lásd Taqqu (1977); ez (3.2) általánositása]. $\hat{\underline{P}}_T(\underline{G}_g)$ (3.4) szerinti gyors számitógépes meghatározására Tusnády Gábor dolgozott ki eljárást. Ha (3.3) nem teljesül, $\underline{P}_T(\underline{G}_g)$ Monte-Carlo módszerrel becsülhető [lásd pl. Deák (1980). Annak, hogy csak ebben az esetben dolgozom Monte-Carlo módszerrel, az az oka, hogy a számomra érdekes esetekben a $\underline{P}_T$ valószinüségek általában igen kicsik]. A $\hat{\underline{P}}_T(\underline{G}_g)$ becslések birtokában meghatározhatóak - legalábbis néhány kisebb $\underline{k}$-ra - az

$$\{M \times \sum_{g:|G_g|=k} \hat{P}_T(G_g); \quad k = 3, 4, \ldots, n\}$$

sorozat elemei, és ezek összehasonlithatóak a megfelelő

$$\{\sum_{g:|G_g|=k} \sigma_T(G_g); \quad k = 3, 4, \ldots, n\}$$

sorozat elemeivel [$|\underline{G}_g|$ $\underline{G}_g$ mérete; eltérően attól az esettől, amikor a változók függetlenek, a $\underline{P}^{(k)}$ valószinüségek becslését a normális küszöb modellben nem sikerült meghatároznom].

Tetszőleges $\underline{G}_g$ jellegkombináció esetén legyen $\underline{E}_T(\underline{G}_g)$ azon objektumok számának várható értéke, amelyek rendelkeznek a $\underline{G}_g$-hez tartozó jellegekkel és esetleg továbbiakkal is, $\underline{G}_g = \{\underline{A}_i\}$ vagy $\underline{G}_g = \{\underline{A}_i, \underline{A}_j\}$ esetén pedig

$$\sigma_C^{(3)}(G_g) = \sum_{\substack{v:\,|G_v|=3 \\ G_v \supseteq G_g}} \sigma_T(G_v), \qquad E_C^{(3)}(G_g) = \sum_{\substack{v:\,|G_v|=3 \\ G_v \supseteq G_g}} E_T(G_v),$$

$$d(G_g) = \frac{[\sigma_C^{(3)}(G_g) - E_C^{(3)}(G_g)]^2}{E_C^{(3)}(G_g)},$$

akkor $\underline{d}(\underline{G}_g)$ [amelyik egyébként nem χ^2-eloszlásu, hiszen az $\underline{\sigma}_C^{(3)}(\underline{G}_g)$ összeg tagjai nem függetlenek] azt mutatja, hogy $\underline{G}_g$ nagyobb méretü jellegkombinációkban való előfordulása mennyire tér el a normális küszöb modell szerint várttól.

Végül még egy problémáról szólok. Mivel az

$$\underline{\underline{R}} = (r_{ij})_{i,j=1}^{n}$$

korrelációs mátrix maximum likelihood becslése $\underline{n} > 2$ esetén nem ismeretes, $\underline{\underline{R}}$-et elemenként becslem. Ezért előfordulhat, hogy az $\underline{\underline{R}}$ mátrix igy nyert $\hat{\underline{\underline{R}}}$ becslése nem lesz pozitiv szemidefinit, ami bizonyos további vizsgálatoknál (pl. faktoranalizisnél) akadályt jelent, ezért szükséges a becslés módositása, pozitiv szemidefinitté tétele. Ez többféleképpen is elvégezhető. Az $\hat{\underline{\underline{R}}}$-hoz mátrix norma szerint legközelebbi $\underline{\underline{R}}^*$ pozitiv szemidefinit mátrix ugy határozható meg, hogy $\hat{\underline{\underline{R}}}$ spektrálfelbontásában a negativ sajátértékeket 0-val helyettesitjük. Az igy nyert

$$\underline{\underline{R}}^* = (r_{ij}^*)_{i,j=1}^n$$

mátrix sajátértékeinek

$$S = \operatorname{tr} \underline{\underline{R}}^* = \sum_{i=1}^{n} r_{ii}^*$$

összege viszont nagyobb $\underline{n}$-nél. Legyen

$$\tilde{r}_{ij} = \frac{n r_{ij}^*}{S}, \quad i, j = 1, 2, \ldots, n,$$

akkor az

$$\underline{\underline{\tilde{R}}} = (\tilde{r}_{ij})_{i,j=1}^n$$

mátrix pozitiv szemidefinit, és

$$\operatorname{tr} \underline{\underline{\tilde{R}}} = \sum_{i=1}^{n} \tilde{r}_{ii} = n,$$

az azonban tudomásom szerint nem ismert, hogy az ilyen tulajdonságu mátrixok $\underline{\underline{\hat{R}}}$-tól mátrix normában való távolságát milyen mátrix minimalizálja.

Jelölje $\underline{\underline{\hat{R}}}$ sajátértékeit $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. $\underline{\underline{\hat{R}}}$-hoz "közeli", $\underline{n}$ nyomu, pozitiv szemidefinit $\underline{\underline{\tilde{R}}}$ mátrix ugy is meghatározható, hogy $\underline{\underline{\hat{R}}}$ spektrálfelbontásában a λ_i-ket olyan μ_i-

kkel helyettesitjük, melyekre

$$\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n;$$

$$\mu_i \begin{cases} > 0, & \text{ha } \lambda_i \geq \varepsilon_3, \\ = 0 & \text{egyébként;} \end{cases}$$

továbbá

$$\sum_{i=1}^{n} f(\lambda_i - \lambda_n)(\mu_i - \lambda_i)^2 = \text{minimum},$$

ahol $\varepsilon_3 > 0$ olyan, hogy

$$\sum_{i=1}^{n} \mu_i = \left(\sum_{i=1}^{n} \lambda_i =\right) n$$

teljesüljön, $\underline{f}(\underline{x})$ pedig alkalmas függvény (pl. $\sqrt{\underline{x}}$). Ez a kvadratikus programozási feladat a Beale-módszerrel oldható meg [lásd pl. Bernau (1977)].

Az $\hat{\underline{\underline{R}}}$ módositására megadott mindkét módszer alkalmazása során olyan $\tilde{\underline{\underline{R}}}$ mátrixot kapunk, amely ugyan pozitiv szemidefinit és $\underline{n}$ nyomu, de rendelkezhet $\tilde{\underline{r}}_{ii} \neq 1$ vagy $|\tilde{\underline{r}}_{ij}| > 1$ elemekkel. Olyan $\tilde{\underline{\underline{R}}}$-hoz, amely már "igazi" korrelációs mátrix, iterativ uton juthatunk, a következő két lépés ismételt alkalmazásával: a) a mátrixot pozitiv szemi-

definitté tesszük; b) a főátlóba 1-eket irunk, a főátlón kivüli, 1-nél nagyobb vagy (-1)-nél kisebb elemeket 1-gyel, ill. (-1)-gyel helyettesitjük. Tapasztalataim szerint az iteráció néhány lépés után "igazi" korrelációs mátrixhoz vezet.

4. TÖBBDIMENZIÓS SKÁLÁZÁS

Tegyük fel, hogy M számu objektumunk és n számu - most tetszőleges, nem feltétlenül bináris - változónk van. A többváltozós statisztikai módszerek legnagyobb része a változóknak az objektumokon tett megfigyelési adataival dolgozik. Ezekkel a módszerekkel ellentétben többdimenziós skálázás (angolul "multidimensional scaling", röviditve MDS) esetén az adatpontokat nem tudjuk - vagy nem akarjuk - közvetlenül mint az n-dimenziós tér M számu pontját megfigyelni, csak közvetett információval rendelkezünk róluk. Ez az információ az objektumok vagy/és változók különbözőségére (távolságára) vagy - ellenkezőleg - hasonlóságára (közelségére) vonatkozik. Az MDS azzal a problémával foglalkozik, hogy egy (MxM)-es, (nxn)-es vagy (Mxn)-es távolság- vagy hasonlóságmátrix (az MDS adatmátrixa) alapján hogyan lehet az objektumokat vagy/és változókat a térben megjeleniteni, más szóval hogyan lehet az alacsony dimenziós euklideszi térben olyan pont-M-est vagy/és pont-n-est konstruálni hozzájuk, hogy a pontok euklideszi távolsága minél jobban tükrözze az objektumok vagy/és változók távolságát (különbözőségét).

A többdimenziós skálázás témakörébe tartozó feladatokat több kritérium szerint is szokták osztályozni. Talán a legfontosabb kritérium, hogy az adatok (amelyek az MDS ese-

tében tehát távolságok vagy hasonlóságok) hányféle dologra vonatkoznak. Ha egyfélére (vagy objektumokra, vagy változókra), egyféle (angolul "one-mode") adatu, ha kétfélére (objektumokra és változókra), kétféle (angolul "two-mode") adatu MDS-ről beszélünk.

Egyféle adatu MDS esetén az egyes sorok ugyanazoknak a dolgoknak felelnek meg, mint a megfelelő oszlopok. Ilyenkor az MDS adatmátrixa négyzetes és szimmetrikus. Attól függően, hogy a sorok és oszlopok az objektumoknak vagy a változóknak felelnek meg, az objektumok vagy a változók többdimenziós skálázásáról beszélünk.

Kétféle adatu MDS esetén az egyes sorok nem ugyanazoknak a dolgoknak felelnek meg, mint a megfelelő oszlopok. Ilyenkor téglalapmátrixszal van dolgunk. A kétféle adatu többdimenziós skálázást többdimenziós kiteritésnek (angolul "multidimensional unfolding", röviditve MDU) is nevezik. - A továbbiakban változók egyféle adatu többdimenziós skálázásával fogok foglalkozni.

A többdimenziós skálázás irodalmában a legkorábbi összefoglaló munka Torgerson (1958). A későbbiek közül a következőket emelem ki: Shepard et al. (1972), Romney et al. (1972), Kruskal (1977a,b), Kruskal és Wish (1978), Mardia et al. (1979) 14. fejezete, Schiffman et al. (1981), Gordon (1981) 5. fejezete, De Leeuw és Heiser (1982), Wish és Carroll (1982), Gower (1984). Magyar nyelven Füstös (1981) és Telegdi (1984, 1986) említhető.

4.1. A többdimenziós skálázás klasszikus megoldása

DEFINICIÓ. Egy $\underline{\underline{D}} = (\underline{\underline{d}}_{rs})_{r,s=1}^{n}$ mátrixot távolságmátrixnak nevezünk, ha szimmetrikus és

$$d_{rr} = 0, \quad d_{rs} \geq 0, \quad r \neq s.$$

DEFINICIÓ. Egy $\underline{\underline{D}}$ távolságmátrixot euklideszinek mondunk, ha valamely $\underline{R}^p$ euklideszi térben van olyan pontkonfiguráció, melynek pontok közötti távolságait $\underline{\underline{D}}$ adja meg; más szóval ha valamely p egész szám mellett vannak olyan $\underline{\underline{x}}_1, \underline{\underline{x}}_2, \ldots, \underline{\underline{x}}_n \in \underline{R}^p$ pontok, amelyekre

$$d_{rs}^2 = (\underline{\underline{x}}_r - \underline{\underline{x}}_s)^T(\underline{\underline{x}}_r - \underline{\underline{x}}_s).$$

A következő tétel [lásd pl. Mardia et al. (1979)] segítségével tetszőleges $\underline{\underline{D}} = (\underline{\underline{d}}_{rs})$ távolságmátrixról eldönthető, hogy euklideszi-e, és ha igen, hogyan lehet hozzá megfelelő pontkonfigurációt találni. Ehhez bevezetek néhány jelölést. Legyen $\underline{\underline{I}}_n$ az n-dimenziós egységmátrix,

$$\underline{\underline{1}}_n = (1, 1, \ldots, \overset{n}{1})^T, \quad \underline{\underline{H}}_n = \underline{\underline{I}}_n - \frac{1}{n}\underline{\underline{1}}_n\underline{\underline{1}}_n^T,$$

$$a_{rs} = -\frac{1}{2}d_{rs}^2, \quad \underline{\underline{A}} = (a_{rs}).$$

TÉTEL. Legyen $\underline{\underline{D}} = (\underline{\underline{d}}_{rs})$ tetszőleges távolságmátrix és

$$\underline{\underline{B}} = (b_{rs})_{r,s=1}^{n} = \underline{\underline{H}}_n \underline{\underline{A}} \underline{\underline{H}}_n .$$

$\underline{\underline{D}}$ akkor és csak akkor euklideszi, ha $\underline{\underline{B}}$ pozitiv szemidefinit. Nevezetesen a következők igazak:

/i/ Ha $\underline{\underline{D}}$ egy $\underline{\underline{Z}} = (\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_n)^T$ konfiguráció pontok közötti euklideszi távolságainak mátrixa, akkor

$$b_{rs} = (\underline{z}_r - \bar{\underline{z}})^T (\underline{z}_s - \bar{\underline{z}}), \quad r, s = 1, 2, \ldots, n, \qquad (4.1)$$

ahol

$$\bar{\underline{z}} = \frac{1}{n} \sum_{r=1}^{n} \underline{z}_r = \frac{1}{n} \underline{\underline{Z}}^T \underline{1}_n .$$

(4.1) mátrix alakban $\underline{\underline{B}} = (\underline{\underline{H}}_n \underline{\underline{Z}})(\underline{\underline{H}}_n \underline{\underline{Z}})^T$, tehát $\underline{\underline{B}}$, amely a $\underline{\underline{Z}}$ konfiguráció centrált skalárszorzat-mátrixa, pozitiv szemidefinit.

/ii/ Megforditva, ha $\underline{\underline{B}}$ p rangu pozitiv szemidefinit, akkor egy megfelelő konfiguráció konstruálható az alábbi módon. Legyenek $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ $\underline{\underline{B}}$ pozitiv sajátértékei

$$\underline{\underline{x}}_{(i)}^T \underline{\underline{x}}_{(i)} = \lambda_i, \quad i = 1, 2, \ldots, p,$$

szerint normált $\underline{\underline{x}}_{(1)}, \underline{\underline{x}}_{(2)}, \ldots, \underline{\underline{x}}_{(p)}$ (egyszeres sajátérték mellett egyértelmüen meghatározott) or-

<u>togonális sajátvektorokkal. Legyen</u>

$$\underline{\underline{X}} = (x_{rs}) = [\underline{\underline{x}}(1), \underline{\underline{x}}(2), \ldots, \underline{\underline{x}}(p)],$$

<u>akkor az</u>

$$\underline{\underline{x}}_r = (x_{r1}, x_{r2}, \ldots, x_{rp})^T$$

<u>koordinátáju</u> $\underline{P}_r \in \underline{R}^p$ <u>pontok távolságainak mátrixa</u> $\underline{\underline{D}}$, <u>továbbá a pontkonfiguráció sulypontja az origo, skalárszorzat-mátrixa</u> $\underline{\underline{B}}$.

Legyen $\underline{\underline{D}} = (\underline{d}_{rs})$ tetszőleges távolságmátrix. Ehhez keresendőek olyan $\underline{P}_1, \underline{P}_2, \ldots, \underline{P}_n$ $\underline{k}$-dimenziós pontok, amelyekre teljesül, hogy $\underline{P}_r$ és $\underline{P}_s$ euklideszi távolságát $\hat{\underline{d}}_{rs}$-val jelölve $(\hat{\underline{d}}_{rs})$ valamilyen értelemben "hasonló" $\underline{\underline{D}}$-hez. Általában nemcsak a $\underline{P}_r$ pontok, hanem a $\underline{k}$ dimenzió is ismeretlen. Ez utóbbit a gyakorlatban rendszerint 1-nek, 2-nek vagy 3-nak választják, mert igy az objektumok a pontok révén ténylegesen kirajzolódnak. Ha $\underline{\underline{D}}$ euklideszi, valamely $\underline{p}$-dimenziós euklideszi térben definició szerint van olyan pontkonfiguráció, amelynek pontok közötti távolságait éppen $\underline{\underline{D}}$ adja meg. Ahhoz, hogy ez az MDS-probléma megoldása legyen, az kell, hogy $\underline{p}$-t $\underline{k}$-nak lehessen választani. A gyakorlatban azonban $\underline{p}$ általában tul nagy ehhez.

Egy lehetséges konfigurációválasztást sugall az előző tétel. Válasszuk az $\underline{R}^k$ azon pontjaiból álló konfigurációt,

melyek koordinátáit a tételben szereplő $\underline{\underline{B}}$ mátrix első $\underline{k}$ számu (egyszeres sajátérték mellett egyértelmüen meghatározott) sajátvektora adja meg. Minél nagyobb pozitiv $\underline{\underline{B}}$ első $\underline{k}$ számu sajátértéke és minél kisebb abszolut értékü a többi, annál jobban fogja közeliteni a konfiguráció pontok közötti távolságmátrixa $\underline{\underline{D}}$-t. Ezt a konfigurációt az MDS-probléma $\underline{k}$-dimenziós klasszikus megoldásának nevezik [optimalitási tulajdonságait lásd pl. Mardia et al. (1979)-ben]. Számitásmenete röviden a következő:

/i/ $\underline{\underline{D}}$-ből - amelyről nem kell feltenni, hogy euklideszi - meghatározzuk az $\underline{\underline{A}}$ mátrixot.

/ii/ Képezzük a $\underline{b}_{rs} = \underline{a}_{rs} - \bar{\underline{a}}_{r.} - \bar{\underline{a}}_{.s} + \bar{\underline{a}}_{..}$ elemekből álló $\underline{\underline{B}}$ mátrixot, ahol

$$\bar{a}_{r.} = \frac{1}{n}\sum_{s=1}^{n} a_{rs}, \qquad \bar{a}_{.s} = \frac{1}{n}\sum_{r=1}^{n} a_{rs},$$

$$\bar{a}_{..} = \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} a_{rs}.$$

/iii/ Előállitjuk $\underline{\underline{B}}$ $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ $\underline{k}$ számu legnagyobb sajátértékét (feltesszük, hogy ezek pozitivak; ha ez nem teljesül, az MDS-problémának nem létezik $\underline{k}$-dimenziós klasszikus megoldása) és az

$$\underline{\underline{x}}_{(i)}^{T}\underline{\underline{x}}_{(i)} = \lambda_i, \quad i = 1, 2, \ldots, k,$$

szerint normált $\underline{\underline{x}}_{(1)}, \underline{\underline{x}}_{(2)}, \ldots, \underline{\underline{x}}_{(k)}$ sajátvektorokat.

/iv/ A $\underline{P}_r$ pontok kivánt koordinátái az

$$\underline{\underline{X}} = [\underline{\underline{x}}_{(1)}, \underline{\underline{x}}_{(2)}, \ldots, \underline{\underline{x}}_{(k)}]$$

mátrix $\underline{\underline{x}}_r = (\underline{x}_{r1}, \underline{x}_{r2}, \ldots, \underline{x}_{rk})^T$ sorainak elemei ($\underline{r} = 1, 2, \ldots, \underline{k}$).

Tetszőleges ($\underline{d}_{rs}$) távolságmátrix és valamely $\underline{\underline{X}}$ pontkonfiguráció ($\hat{\underline{d}}_{rs}$) euklideszi távolságmátrixa közötti eltérés többdimenziós skálázás során leginkább használt mérőszámai a

$$\sum_{r=1}^{n-1}\sum_{s=r+1}^{n}(d_{rs}-\hat{d}_{rs})^2, \tag{4.2}$$

$$\sum_{r=1}^{n-1}\sum_{s=r+1}^{n}c_{rs}(d_{rs}-\hat{d}_{rs})^2 \tag{4.3}$$

és

$$\sum_{r=1}^{n-1}\sum_{s=r+1}^{n}(d_{rs}-\hat{d}_{rs})^2 \Big/ \sum_{r=1}^{n-1}\sum_{s=r+1}^{n}\hat{d}_{rs}^2$$

mennyiségek, ahol a $\underline{c}_{rs}$-ek valamilyen adott sulyozó tényezők [(4.2) a (4.3) mennyiség $\underline{c}_{rs} \equiv 1$-hez tartozó speciális esete]. Mindhárom mennyiség az $\underline{\underline{X}}$ pontkonfiguráció, tehát az $\underline{x}_{11}$, $\underline{x}_{12}$, ..., $\underline{x}_{1k}$, $\underline{x}_{21}$, $\underline{x}_{22}$, ..., $\underline{x}_{2k}$, ..., $\underline{x}_{n1}$, $\underline{x}_{n2}$, ..., $\underline{x}_{nk}$ ismeretlen változók függvénye. Az MDS-probléma megoldása ekkor egy ($\underline{n}$x$\underline{k}$)-változós függvény feltétel nélküli minimalizálásával történik. A minimalizáló eljárások általában gradiens- vagy Fletcher-tipusu algoritmusok [lásd pl. Abaffy (1976)], amelyek nem tudnak különbséget tenni lokális és globális minimum között, ezért jó kezdeti konfigurációt igényelnek. Ilyennek választható az MDS-probléma klasszikus megoldása.

4.2. Többszörös többdimenziós skálázás

Tegyük fel, hogy az $\underline{Y}_1$, $\underline{Y}_2$, ..., $\underline{Y}_M$ objektumokat jellemző, az első három fejezetben szereplő $\underline{W}_1$, $\underline{W}_2$, ..., $\underline{W}_n$ bináris változó többdimenziós skálázása a feladat. A korábbiakkal összhangban ez a következőket jelenti: valamilyen módon távolságokat kell konstruálni a változók között, és a változókat ugy kell kirajzolni az alacsony dimenziós euklideszi térben, hogy a változóknak megfelelő pontok euklideszi távolságai minél kevésbé különbözzenek a változók távolságaitól. Ahhoz, hogy a változók jól skálázhatóak legyenek, konzisztenseknek kell lenniük a következő értelem-

ben: ha két változó "közel" van egy harmadikhoz, egymáshoz is "közel" kell lenniük. Tegyük fel például (1. példa), hogy $\underline{n} = 3$, $\underline{M} = 44$, az első 19 objektum az $\underline{A}_1$ és $\underline{A}_2$, a következő 14 az $\underline{A}_1$ és $\underline{A}_3$, az utolsó 11 pedig az $\underline{A}_2$ és $\underline{A}_3$ jellegekkel rendelkezik. Ezekhez az objektumokhoz hozzárendelhető a változók

$$\begin{pmatrix} 0 & 3 & 4 \\ 3 & 0 & 5 \\ 4 & 5 & 0 \end{pmatrix}$$

távolságmátrixa, ami alapján a változók jól skálázhatóak:

A_3

A_1 A_2

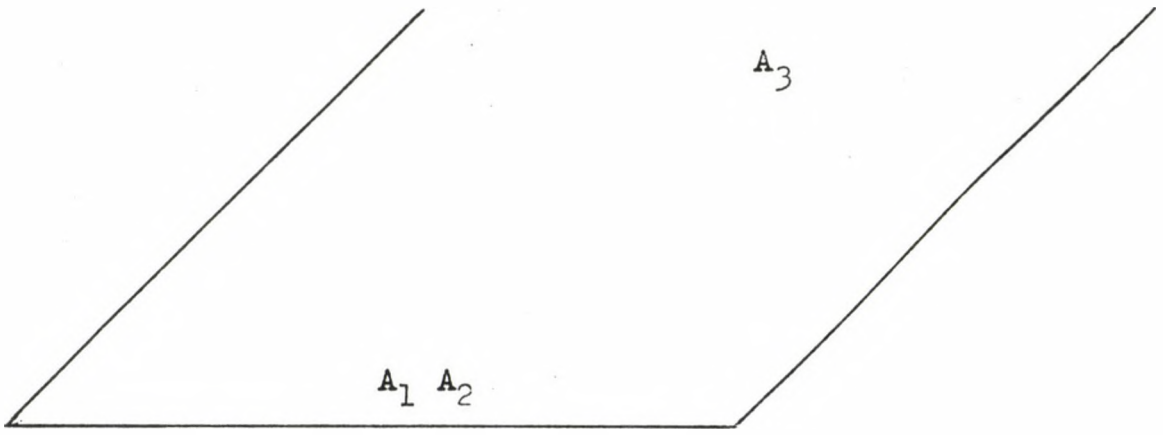
Legyen most (2. példa) $\underline{M} = 33$, és tegyük fel, hogy az első 19 objektum az $\underline{A}_1$ és $\underline{A}_2$, a további 14 pedig az $\underline{A}_1$ és $\underline{A}_3$ jellegekkel rendelkezik. Ezekhez az objektumokhoz a változók

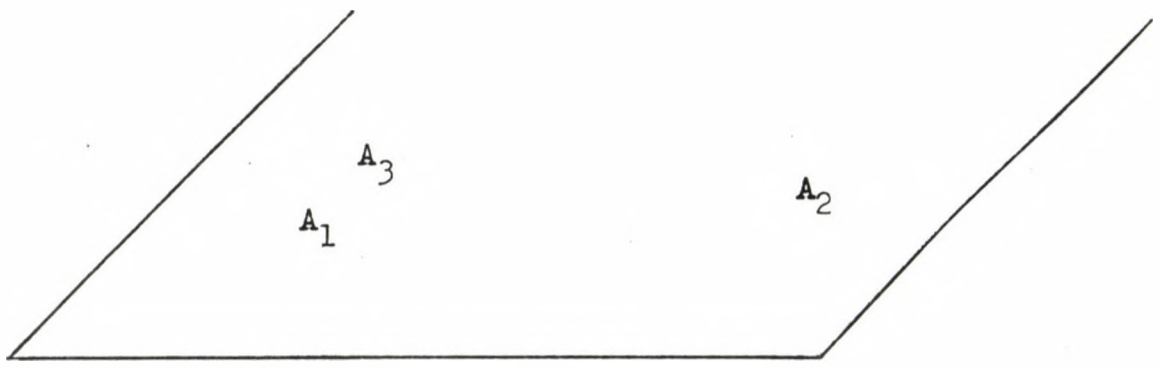
$$\begin{pmatrix} 0 & 3 & 4 \\ 3 & 0 & 60 \\ 4 & 60 & 0 \end{pmatrix}$$

távolságmátrixa rendelhető hozzá, ami alapján a változók csak nagyon rosszul skálázhatóak. Próbáljuk ezért őket külön az első 19 és külön a további 14 objektum alapján skálázni! Ehhez a két objektumhalmazhoz (amelyek klaszteroknek tekinthetőek) a változók

$$\begin{pmatrix} 0 & 3 & 60 \\ 3 & 0 & 60 \\ 60 & 60 & 0 \end{pmatrix} \qquad \text{és} \qquad \begin{pmatrix} 0 & 60 & 4 \\ 60 & 0 & 60 \\ 4 & 60 & 0 \end{pmatrix}$$

távolságmátrixai rendelhetőek hozzá, amik alapján viszont a változók - két sikon! - már jól skálázhatóak:





A fenti példákban szereplő gyakoriságokhoz természetesen többféle távolság rendelhető hozzá. Jelöljük n_{ij}-vel azon objektumok számát, amelyek rendelkeznek az A_i és A_j jellegekkel. Egy lehetséges távolságdefiníció a következő:

$$d_{ij} = \sqrt{n_{ii}+n_{jj}-2n_{ij}}.$$

Ez az n_{ij} hasonlóságból távolságba való standard transzformáció [lásd pl. Hardia et al. (1979)] eredménye. Annak ellenére, hogy ennek a transzformációnak vannak bizonyos előnyös tulajdonságai (pl. ha a hasonlóságmátrix pozitiv szemidefinit, akkor a megfelelő távolságmátrix euklideszi lesz), használatát sem általában, sem a konkrét esetben nem tartom jónak: tulságosan "kisimitja" a távolságstruktúrát, szemléletesen nyilvánvalóan nagyon különböző hasonlóságokhoz nem eléggé különböző távolságokat rendel hozzá [a 2. példában a (19,0) hasonlóságpárhoz a $(\sqrt{14},\sqrt{33})$ távolságpárt]. A változók inkonzisztenciájából fakadó problémát ezzel inkább megkerüli, de nem oldja meg. Ezért a későbbiekkel összhangban a

$$d_{ij} = \frac{K}{n_{ij}+1}$$

távolságdefiníciót választom ($K = 60$ mellett). Itt azonban ez nem lényeges, hiszen magukból a gyakoriságokból is látszik, hogy az említett konzisztencia az első példában

teljesül, a másodikban nem.

A <u>többszörös többdimenziós skálázás</u> [multiple MDS, MMDS; Telegdi (1982), Simonovits et al. (1982), Telegdi (1984, 1986)] a 2. példához hasonló esetekkel foglalkozik, azzal a problémával, amelyik akkor merül fel, ha a konzisztencia nem teljesül: hogyan lehet az objektumokat minél homogénebb klaszterekbe sorolni, amikoris egy-egy klaszter homogenitását a változók ezen klaszter mellett történő (közönséges) többdimenziós skálázásának jóságával mérjük? Keresendő tehát a p természetes szám, az

$$Y_1, Y_2, \ldots, Y_p \subset \{y_1, y_2, \ldots, y_M\} = Y$$

diszjunkt klaszterek (amelyekre

$$\bigcup_{m=1}^{p} Y_m = Y)$$

és a k-dimenziós euklideszi tér azon $\underline{x}_i^{(m)}$ pontjai ($i = 1, 2, \ldots, n$; $m = 1, 2, \ldots, p$), amelyek a változókat reprezentálják abban az értelemben, hogy $\underline{x}_i^{(m)}$ és $\underline{x}_j^{(m)}$ közelsége W_i és W_j Y_m melletti közelségének felel meg.

Legyen

$$\underline{e}_g = (e_{g1}, e_{g2}, \ldots, e_{gn})$$

[e_{gi} a W_i változónak az y_g objektumon megfigyelt - 1 vagy

0 - értéke ($g = 1, 2, \ldots, M$; $i = 1, 2, \ldots, n$)]. Mivel megegyező $\underline{e}_g$-jü objektumok megkülönböztethetetlenek, tegyük fel, hogy

$$\underline{e}_{g_1} = \underline{e}_{g_2} \Rightarrow g_1 = g_2,$$

viszont az objektumoknak multiplicitása van. Mivel az egyes objektumok klaszterbe sorolásának értelemszerüen a rajtuk 1 értékü változóktól kell függniük, az MDS az olyan objektumokkal, amelyeken legfeljebb egy változónak 1 az értéke, nem tud mit kezdeni. Ezért tegyük fel, hogy minden objektumon legalább 2 változó értéke 1. Ekkor a (változók alapján különböző) objektumok száma az 1.1. paragrafusban definiált N, minden objektum egy legalább 2 méretü jellegkombinációnak felel meg, és az Y_g objektum multiplicitása $\sigma(G_g)$ ($g = 1, 2, \ldots, N$), ahol G_g az a jellegkombináció, amelyiknek Y_g megfelel.

Legyen $\underline{DC}$ olyan eljárás, amellyel az objektumok tetszőleges $Z \subset Y$ halmazához - a Z objektumaihoz tartozó $\underline{e}_g$-k és $\sigma(G_g)$-k alapján - $d_{ij}(Z)$ távolságokat és $c_{ij}(Z)$ súlyozó tényezőket rendelek hozzá ($1 \leq i < j \leq n$). Legyen

$$d_{ij}^{(m)} = d_{ij}(Y_m), \quad c_{ij}^{(m)} = c_{ij}(Y_m),$$

$$V^{(m)} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c_{ij}^{(m)} [d_{ij}^{(m)} - \| \underline{x}_i^{(m)} - \underline{x}_j^{(m)} \|]^2$$

és

$$V = \sum_{m=1}^{p} V^{(m)},$$

akkor - a közönséges MDS-sel összhangban - természetesnek látszik az MMDS-t a V mennyiséggel jellemezni.

Legyen

$$V_g^{(m)} = \sum_{\substack{i=1 \\ e_{gi}=1}}^{n-1} \sum_{\substack{j=i+1 \\ e_{gj}=1}}^{n} c_{ij}^{(m)} [d_{ij}^{(m)} - \|\underline{x}_i^{(m)} - \underline{x}_j^{(m)}\|]^2$$

($g = 1, 2, \ldots, N$; $m = 1, 2, \ldots, p$). Az MMDS-probléma kézenfekvő módon oldható meg a következő két lépés váltakozva történő alkalmazásával:

/i/ $g = 1, 2, \ldots, N$-re az y_g objektumot abba a klaszterbe soroljuk, amelyikre $V_g^{(m)}$ minimális (amelyik klaszterben azoknak a változóknak a távolságai, amelyeknek y_g-n 1 az értéke, a legkevésbé térnek el a nekik ott megfelelő pontok euklideszi távolságaitól);

/ii/ $m = 1, 2, \ldots, p$-re meghatározzuk (explicit módon kiszámitjuk) az (nxk)-változós

$$V^{(m)} = V^{(m)}[\underline{x}_1^{(m)}, \underline{x}_2^{(m)}, \ldots, \underline{x}_n^{(m)}]$$

függvény gradiens vektorát, majd ezen vektor (-1)-

szeresének irányában iránymenti minimalizálást végezve kiszámitjuk az $\underline{\underline{x}}_i^{(m)}$ pontok uj koordinátáit ($\underline{i}$ = = 1, 2, ..., $\underline{n}$; egy lépésen belül esetleg kétszer-háromszor is).

Mivel a /ii/-es lépések során nem minimalizáljuk $\underline{V}^{(m)}$-et, csak csökkentjük az értékét, számitógépes realizációnál érdemes $\underline{V}^{(m)}$ helyett a

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} c_{ij}^{(m)} \left[d_{ij}^{(m)} - \| \underline{\underline{x}}_i^{(m)} - \underline{\underline{x}}_j^{(m)} \| \right]^2$$

függvények ($\underline{i}$ = 1, 2, ..., $\underline{n}$) gradiensét venni egymás után ($\underline{n}$ számu pont helyett egyszerre csak egyet mozgatni. Ezzel a memóriaigény és a futási idő is csökken).

A fenti algoritmusnak - és ezáltal a többszörös többdimenziós skálázás $\underline{V}$-vel való jellemzésének - gyengéje $\underline{V}_g^{(m)}$ $\underline{m}$-ben való minimalizálása az /i/-es lépések során. Az objektumok optimális osztályozásához igy ugyanis az $\underline{\underline{x}}_i^{(m)}$ pontokon kivül a $\underline{d}_{ij}^{(m)}$ távolságokra és a $\underline{c}_{ij}^{(m)}$ sulyozó tényezőkre is folyamatosan szükség van, ezért amint egy objektumot az egyik klaszterből a másikba átteszünk, azonnal módositanunk kell a megfelelő $\underline{d}_{ij}^{(m)}$ és $\underline{c}_{ij}^{(m)}$ értékeket. Ezért célszerü $\underline{V}_g^{(m)}$ helyett egy olyan $\underline{U}_g^{(m)}$ mennyiséget minimalizálni $\underline{m}$-ben az /i/-es lépések során, amelyik csak az $\underline{\underline{x}}_i^{(m)}$ pontoktól

függ, a $\underline{d}_{ij}^{(m)}$ távolságoktól és a $\underline{c}_{ij}^{(m)}$ sulyozó tényezőktől nem. Legyen

$$U_g^{(m)} = \sum_{\substack{i=1 \\ e_{gi}=1}}^{n-1} \sum_{\substack{j=i+1 \\ e_{gj}=1}}^{n} \|\underline{x}_i^{(m)} - \underline{x}_j^{(m)}\|^2,$$

akkor ennek a mennyiségnek $\underline{m}$-ben való minimalizálása a következőt jelenti: az $\underline{y}_g$ objektumot abba a klaszterbe soroljuk, amely mellett az $\underline{y}_g$-n 1 értékü változók párjaihoz tartozó pontok távolságainak négyzetösszege a legkisebb. Az MMDS-t jellemzendő

$$E = \sum_{m=1}^{p} E^{(m)} = \sum_{m=1}^{p} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f[d_{ij}^{(m)}, c_{ij}^{(m)}, \|\underline{x}_i^{(m)} - \underline{x}_j^{(m)}\|]$$

mennyiségben szereplő $\underline{f}$ függvénynek olyannak kell lennie, hogy

/a/ az

$$E_g^{(m)} = \sum_{\substack{i=1 \\ e_{gi}=1}}^{n-1} \sum_{\substack{j=i+1 \\ e_{gj}=1}}^{n} f[d_{ij}^{(m)}, c_{ij}^{(m)}, \|\underline{x}_i^{(m)} - \underline{x}_j^{(m)}\|]$$

jelölés mellett $[\underline{E}_g^{(m)} - \underline{U}_g^{(m)}]$ értéke ne függjön $\underline{m}$-től ($\underline{g}$ = 1, 2, ..., $\underline{N}$),

/b/ $[\underline{E}^{(m)}-\underline{V}^{(m)}]$ és igy $(\underline{E}-\underline{V})$ értéke ne függjön az $\underline{\underline{x}}_i^{(m)}$ pontoktól $(\underline{m} = 1, 2, \ldots, \underline{p};\ \underline{i} = 1, 2, \ldots, \underline{n})$.

Az az algoritmus, amelyik az /i/-es lépések során $\underline{U}_g^{(m)}$-et minimalizálja $\underline{m}$-ben, a /ii/-es lépések során pedig $\underline{V}^{(m)}$ értékét csökkenti, $\underline{E}$ feltétel nélküli (sajnos nem mindig globális) minimumát állitja elő.

Az $\underline{E}$ definiciójában szereplő $\underline{f}$ függvény célszerü megválasztásához már szükség van a $\underline{\underline{DC}}$ eljárás konkretizálására. Legyen $\underline{\underline{DC}}$ olyan, hogy tetszőleges $\underline{Z} \subset \underline{Y}$-ra

$$c_{ij}(Z) = n_{ij}(Z) + 1, \quad d_{ij}(Z) = K \,/\, c_{ij}(Z),$$

ahol

$$n_{ij}(Z) = \sum_{\substack{g:\ y_g \in Z \\ e_{gi}, e_{gj}=1}} \sigma(G_g)$$

és $\underline{K}$ alkalmas állandó. Legyen $\underline{n}_{ij}^{(m)} = \underline{n}_{ij}(\underline{Y}_m)$ és

$$f[d_{ij}^{(m)}, c_{ij}^{(m)}, \|\underline{\underline{x}}_i^{(m)} - \underline{\underline{x}}_j^{(m)}\|] =$$

$$= f\{d_{ij}^{(m)}[n_{ij}^{(m)}], c_{ij}^{(m)}[n_{ij}^{(m)}], \|\underline{\underline{x}}_i^{(m)} - \underline{\underline{x}}_j^{(m)}\|\} =$$

$$= f[n_{ij}^{(m)}, \|\underline{\underline{x}}_i^{(m)} - \underline{\underline{x}}_j^{(m)}\|] =$$

$$= n_{ij}^{(m)} \|\underline{\underline{x}}_i^{(m)} - \underline{\underline{x}}_j^{(m)}\|^2 + [K - \|\underline{\underline{x}}_i^{(m)} - \underline{\underline{x}}_j^{(m)}\|]^2.$$

4.1. TÉTEL. Tekintsük azt a T algoritmust, amelyik az /i/-es lépések során $U_g^{(m)}$-et minimalizálja m-ben, a /ii/-es lépések során pedig $V^{(m)}$ értékét csökkenti. A DC eljárás és az f függvény fenti megválasztása esetén az algoritmus folyamán E monoton nem növő.

Bizonyitás. Legyen

$$E_1 = \sum_{m=1}^{p} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} n_{ij}^{(m)} \|\underline{x}_i^{(m)} - \underline{x}_j^{(m)}\|^2$$

és

$$E_2 = \sum_{m=1}^{p} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} [K - \|\underline{x}_i^{(m)} - \underline{x}_j^{(m)}\|]^2,$$

akkor $E = E_1 + E_2$. [Simonovits Miklós hivta fel a figyelmemet arra, hogy E a következő, 2 direkciós erejü rugókból álló rendszer potenciálja: $n_{ij}^{(m)}$ számu 0 hosszuságu és egy K hosszuságu rugó $\underline{x}_i^{(m)}$ és $\underline{x}_j^{(m)}$ között ($1 \le i < j \le n$, $m = 1, 2, \ldots, p$).] Az /i/-es lépések során az $\underline{x}_i^{(m)}$ pontok nem változnak, igy E_2 sem. Viszont

$$E_1 = \sum_{m=1}^{p} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{\substack{g: y_g \in Y_m \\ e_{gi}, e_{gj} = 1}} \sigma(G_g) \|\underline{x}_i^{(m)} - \underline{x}_j^{(m)}\|^2 =$$

$$= \sum_{g=1}^{N} \{\sigma(G_g)[\sum_{m=1}^{p} \sum_{\substack{i=1 \\ e_{gi}=1}}^{n-1} \sum_{\substack{j=i+1 \\ e_{gj}=1}}^{n} \|\underline{x}_i^{(m)} - \underline{x}_j^{(m)}\|^2]\} =$$

$$= \sum_{g=1}^{N} \{\sigma(G_g)[\sum_{m=1}^{p} U_g^{(m)}]\},$$

amiből következik, hogy az /i/-es lépések során $\underline{E}_1$ és igy $\underline{E}$ értéke nem nő. Másrészt egyszerü számolással kapjuk, hogy

$$E - V = K^2 \sum_{m=1}^{p} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{n_{ij}^{(m)}}{n_{ij}^{(m)}+1},$$

ami a /ii/-es lépések során nem változik. Ebből következik, hogy $\underline{E}$ értéke a /ii/-es lépések során sem nő. Ezzel a tétel már adódik.

A $\underline{T}$ algoritmus során egyes klaszterek majdnem vagy akár teljesen is "kiürülhetnek". Az ilyen klasztereket célszerü megszüntetni (ami által természetesen $\underline{p}$ csökken) oly módon, hogy valamely hozzájuk tartozott $\underline{y}_g$ objektumot a fennmaradó klaszterek közül abba sorolunk, amelyikre $\underline{U}_g^{(m)}$ minimális. Ebből is következik, hogy $\underline{p}$ kezdő értékét nagy-

nak érdemes választani [lásd Telegdi és Simonovits (1983)]. Kezdő klaszterekt az $\underline{v}_g$ objektumoknak az $\underline{e}_g$-k alapján [pl. $\underline{k}$-közép (lásd Gulyás, 1983), pontosabban itt most $\underline{p}$-közép eljárás által] történő klaszterezésével állithatunk elő. Kezdő pont-konfigurációnak az egyes kezdő klaszterek melletti MDS-problémák klasszikus megoldásai vehetőek.

A fenti kezdés mellett is előfordulhat, hogy a $\underline{T}$ algoritmus $\underline{E}$-nek csak lokális minimumát állitja elő. Ennek veszélye csökkenthető, ha - a /ii/-es lépések során $\underline{V}^{(m)}$ helyett $\underline{E}^{(m)}$ gradiensét véve - az $\underline{f}$ és $\underline{U}_g^{(m)}$ függvényekben négyzetösszeg helyett $\underline{q}$-adik hatványösszeget irunk és $\underline{q}$ értékét az algoritmus folyamán megváltoztatjuk. Számitógépes tapasztalataim szerint az algoritmus akkor a leghatékonyabb, ha a fent leirt kezdés után $\underline{q}$-t 3-nak választjuk, majd amikor $\underline{E}$ már alig csökken [az /i/-es lépések során már kevés objektum kerül át másik klaszterbe, a /ii/-es lépések során pedig az $\underline{x}_i^{(m)}$ pontok alig változnak], értékét előbb 2-nek, azután 1-nek vesszük. - Ami a $\underline{K}$ állandót illeti, tapasztalataim szerint ugy célszerü választani, hogy $\frac{K}{\mu/p}$ [itt

$$\mu = \frac{\sum_{m=1}^{p} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} n_{ij}^{(m)}}{\binom{n}{2}} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sigma_T(i,j)}{\binom{n}{2}},$$

$\underline{\sigma}_T(\underline{i},\underline{j})$ azon objektumok száma, amelyek rendelkeznek az $\underline{A}_i$,

$\underline{A}_j$ jellegekkel és esetleg továbbiakkal is] kicsi, 10^{-1} nagyságrendü legyen.

Az MMDS eredményének áttekintését megkönnyiti, ha a kapott klasztereket összevonjuk. Ez az alábbiak szerint hajtható végre. Legyen

$$z_{mi} = \sum_{\substack{g:y_g \in Y_m \\ e_{gi}=1}} \sigma(G_g) \; / \sum_{g:y_g \in Y_m} \sigma(G_g)$$

(annak relativ gyakorisága $\underline{Y}_m$-ben, hogy $\underline{W}_i$ értéke 1), és

$$\underline{\underline{z}}_m = (z_{m1}, z_{m2}, \ldots, z_{mn}),$$

$\underline{m} = 1, 2, \ldots, \underline{p}$. Definiáljuk az $\underline{Y}_m$ klaszterek hasonlóságát mint a megfelelő $\underline{\underline{z}}_m$-ek korrelációs együtthatóját. Tekintsük most azt a gráfot, amelyiknek szögpontjai az egyes klaszthereknek felelnek meg, és amelyikben a "nem nagyon hasonló" (valamely $\underline{R}_1$-nél, pl. 0,6-nál nem nagyobb korrelációs együtthatóju $\underline{\underline{z}}_m$-ekhez tartozó) klaszterpároknak megfelelő szögpontok között vannak élek. Szinezzük ezt a gráfot oly módon, hogy az azonos szinü szögpontoknak megfelelő klaszterekbe tartozó objektumok multiplicitással vett számának szórása minimális legyen. Vonjuk össze az azonos szinü szögpontoknak megfelelő klasztereket, és az igy kapott klaszterek között a fentiekhez hasonlóan definiáljunk ha-

sonlóságot. Tekintsük azt a gráfot, amelyiknek szögpontjai az uj klaszereknek felelnek meg, és amelyikben a "nem nagyon különböző" (valamely $\underline{R}_2$-nél, pl. -0,45-nél nem kisebb korrelációs együtthatóju $\underline{z}_m$-ekhez tartozó) klaszterpároknak megfelelő szögpontok között vannak élek. Szinezzük ezt a gráfot a fentiek szerint, és vonjuk össze az azonos szinü szögpontoknak megfelelő klasztereket. [A "hasonló" klaszterek összevonásának értelme nyilvánvaló. A "különböző" klaszterek összevonásáé a következő: valamely $(\underline{Y}_{m_1}, \underline{Y}_{m_2})$ klaszterpárhoz tartozó $\underline{z}_{m_1}$ és $\underline{z}_{m_2}$ korrelációs együtthatója akkor erősen negativ, ha az $\underline{Y}_{m_1}$-ben "közeli" változók csoportjai majdnem vagy teljesen különböznek az $\underline{Y}_{m_2}$-ben "közeli" változók csoportjaitól; ekkor azonban az $\underline{n}$ számu változó reprezentálható a két klaszter együttese mellett egy pont-$\underline{n}$-essel.] Az igy nyert klaszterek melletti MDS-problémák klasszikus megoldásait kezdő pont-konfigurációnak véve ismételjük meg a $\underline{T}$ algoritmust. Ha $\underline{E}$ már $\underline{q} = 1$ mellett sem nagyon csökken, az algoritmust és az MMDS-t a /ii/-es lépés $\underline{q} = 2$ mellett történő egyszeri alkalmazásával (ez a változók még jobb kirajzolását segiti elő) fejezzük be.

Valamely adatmező MMDS-ét két szempontból is jó lenne értékelni. Annak matematikai értékelését, hogy az MMDS mennyire jó más klaszterező módszerekhez képest, elvileg kivihetetlennek tartom. Az egyes klaszterező módszerek ugyanis döntően éppen abban különböznek egymástól, hogy mi-

lyen kritériumot adnak a klaszterezés jóságára. A maga kritériuma szerint mindegyik módszer a legjobb, a kritériumok viszont nem hasonlithatóak objektiven össze. (Különböző klaszterező módszereket heurisztikusan persze összehasonlithatunk: megnézzük, hogy adott klasztereket hogyan adnak vissza.) Annak értékelése, hogy a konkrét adatmező objektumai mennyire klaszterezhetőek jól az MMDS szempontjából, elvileg a következőképpen végezhető el. Tegyük fel, hogy az adatmező valamilyen K mellett végrehajtott MMDS-ének $q = 2$ mellett történő befejezésekor a klaszterszám p, és E értéke

$$E^* = E^*(n,M,\mu,p,K).$$

Legyen E_{opt} és E_{pessz} az E mennyiség értéke az (n,M,μ) értékhármashoz tartozó, az MMDS szempontjából optimális, ill. pesszimális adatmező K fenti értéke mellett végrehajtott MMDS-ének $q = 2$ és p fenti értéke mellett történő befejezésekor. Nyilvánvaló módon E^* az $[E_{opt},E_{pessz}]$ intervallumban helyezkedik el. Annak, hogy a konkrét adatmező mennyire jó az MMDS szempontjából, kézenfekvő mérőszáma az

$$\frac{E^*-E_{opt}}{E_{pessz}-E_{opt}}$$

mennyiség, amely nem lehet 0-nál kisebb és 1-nél nagyobb. Meghatározásához azonban tudni kellene E_{opt} és E_{pessz} ér-

tékét. Ezeket azonban nem tudom, mivel nem ismerem az optimális és pesszimális adatmezőt.

4.3. Többszörös többdimenziós skálázás a változók függetlensége esetén

Az MMDS működésében meghatározó szerepe van az őt jellemző E mennyiségben szereplő f függvénynek. Ennek megválasztásához az előző paragrafusban úgy konkretizáltam a távolságokat és sulyozó tényezőket meghatározó DC eljárást, hogy azok csak az egyes jellegpárok együttes előfordulásától függenek. Emiatt az egyes jellegek előfordulási valószinüségeinek különbözősége esetén az eljárás nem veszi észre a függetlenséget: akkor is közel hoz egymáshoz két jelleget, ha azok csak azért fordulnak elő sürün együtt, mert mindketten gyakoriak. (Ez azonban célom is volt: az 5.6. paragrafusban ismertetésre kerülő konkrét feladat során, amelyből az MMDS kinőtt, nem akartam észrevétlenül hagyni jellegzetes jellegkombinációkat.) Azt, hogy az eljárás észrevegye a függetlenséget, könnyen meg lehet valósitani: a 76. oldalon $n_{ij}(Z)$ definiciójában az összeget osztani kell a

$$\sum_{\substack{g:\ y_g \in Z \\ e_{gi}=1}} \sigma(G_g) \times \sum_{\substack{g:\ y_g \in Z \\ e_{gj}=1}} \sigma(G_g) \Big/ \sum_{g:\ y_g \in Z} \sigma(G_g)$$

kifejezéssel. Ha azonban az egyes jellegek előfordulási valószinüségei megegyeznek, erre nincs szükség. Ezért a függetlenséget ebben az esetben vizsgáltam. Nevezetesen olyan adatmezőt generáltam, amelyre

$$\sigma_T(i,j) = \sigma(i,j) \equiv \mu, \quad 1 \leq i < j \leq n.$$

Ekkor a 79. oldalon definiált g hatványkitevőt végig 2-nek és az objektumklaszterek p számát állandónak véve az MDS olyan objektumklasztereket adott, amelyek mellett nem rajzolódtak ki jellegklaszterek, és a jellegeknek megfelelő pontok - mindegyik objektumklaszter mellett - koncentrikus szabályos sokszögek csucsain helyezkedtek el. A következő táblázat néhány n-re megadja az egyes sokszögek csucsainak n_ν számát (kivülről befelé haladva).

n	m_1	m_2	m_3	m_4
3	3	-	-	-
4	4	-	-	-
5	5	-	-	-
6	6	-	-	-
7	6	1	-	-
10	9	1	-	-
15	12	3	-	-
20	15	5	-	-
25	17	7	1	-
30	19	9	2	-
35	21	10	4	-
40	24	11	5	-
45	25	12	7	1
50	28	13	8	1
55	29	15	8	3
60	30	16	10	4
65	32	17	12	4

Az $\underline{E}$ mennyiség $\tilde{\underline{E}}$ végértékére az

$$\tilde{\underline{E}} = \frac{K^2 p}{p+\mu}\left[C_n p\binom{n-2}{2}+\mu\binom{n}{2}\right]$$

összefüggést kaptam, amely $\mu = 0$ és $\underline{p} = \underline{K} = 1$ mellett, amikoris

$$E = V = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(1-\|\underline{\underline{x}}_i-\underline{\underline{x}}_j\|)^2,$$

az

$$\tilde{\underline{E}} = C_n\binom{n-2}{2}$$

alakra egyszerüsödik. [$\underline{K}$ a 76. oldalon szereplő állandó; $\underline{C}_n$-re $4 \leq \underline{n} \leq 65$ esetén a

$$0{,}1716 \leq C_n \leq 0{,}1807$$

egyenlőtlenséget kaptam ($0{,}1716 = 3 - 2\sqrt{2} = \underline{C}_4$; $0{,}1807 = \frac{75-36\sqrt{3}}{70} = \underline{C}_7$; ahogy $\underline{n}$ értéke 65-höz közeledik, $\underline{C}_n$ értéke 0,176-hoz).]

4.4. Hipergráfok euklideszi térbe ágyazása és particionálása

Feleltessünk meg az n számu bináris változóval jellemzett $Y_1, Y_2, \ldots, Y_N$ objektum Y összességének egy H *értékelt hipergráfot* oly módon, hogy H szögpontjai az egyes változóknak felelnek meg, a hiperélek az objektumoknak (a g-edik hiperél pontosan azokat a szögpontokat köti össze, amely szögpontokhoz tartozó változóknak Y_g-n 1 az értéke; $g = 1, 2, \ldots, N$), a hiperélek értéke pedig a megfelelő objektumok multiplicitásának. Ekkor az i-edik és j-edik szögpontot $n_{ij} = \sigma_T(i,j)$ összértékü (közönséges) él köti össze $(1 \leq i < j \leq n)$. *Ágyazzuk be* a H hipergráfot a k-dimenziós euklideszi térbe, vagyis rendeljünk $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n \in R^k$ pontokat H egyes szögpontjaihoz (k természetes szám). Jó beágyazással szemben kézenfekvő azt a követelményt támasztani, hogy két pont annál közelebb legyen egymáshoz, minél nagyobb összértékü él köti össze a megfelelő szögpontokat, vagyis hogy az $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ pontokra

$$S_1 = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} n_{ij} \|\underline{x}_i - \underline{x}_j\|^2$$

minimális legyen. Nyilvánvaló módon azonban S_1 $\underline{x}_i \equiv \underline{x}$ esetén 0, vagyis minimális. A beágyazástól tehát azt is meg

kell követelni, hogy ne engedje az n számu pontot egy pontba összehuzódni. Keressük ezért azokat az $\underline{\underline{x}}_1$, $\underline{\underline{x}}_2$, ..., $\underline{\underline{x}}_n$ pontokat, amelyekre $S = S_1 + S_2$ minimális, ahol

$$S_2 = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (K - \|\underline{\underline{x}}_i - \underline{\underline{x}}_j\|)^2,$$

K alkalmas állandó. Legyen

$$c_{ij} = n_{ij} + 1, \quad d_{ij} = K / c_{ij}$$

($1 \leq i < j \leq n$), akkor a 4.1. Tételből következik, hogy a H hipergráf S minimalizálásával történő euklideszi térbe ágyazása ekvivalens a változók

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c_{ij}(d_{ij} - \|\underline{\underline{x}}_i - \underline{\underline{x}}_j\|)^2 \qquad (4.4)$$

minimalizálásával történő (közönséges) többdimenziós skálázásával.

Tegyük most fel, hogy H particionálása a feladat. Végezzük ezt azon, tudomásom szerint mások által nem publikált kritérium alapján, hogy a H_m rész-hipergráfok minél jobban beágyazhatóak legyenek ($m = 1, 2, \ldots, p$; p ter-

mészetes szám), nevezetesen ugy, hogy $\underline{E} = \underline{E}_1 + \underline{E}_2$ minimális legyen, ahol $\underline{E}_1$ és $\underline{E}_2$ a 4.1. Tétel bizonyitásában szereplő összegek $\lfloor \underline{n}_{ij}^{(m)}$ az $\underline{i}$-edik és $\underline{j}$-edik szögpontot $\underline{H}_m$-ben összekötő élek összértéke, $\underline{x}_i^{(m)} \in \underline{R}^k$ az $\underline{i}$-edik szögpontnak $\underline{H}_m$-ben megfelelő pont]. A $\underline{H}$ hipergráf $\underline{E}$ minimalizálásával történő particionálása és euklideszi térbe ágyazása nyilvánvaló módon ekvivalens a többszörös többdimenziós skálázással. $\lfloor \underline{H}$ más módon történő particionálását lásd Bolla és Tusnády (1982)-ben, hipergráf modellen alapuló klaszteranalizist Futó (1978)-ban. A klaszteranalizis irodalmának összefoglaló munkái közül a következőket emelem ki: Jardine és Sibson (1971), Anderberg (1973), Hartigan (1975), Van Ryzin (1977), Everitt (1980), Späth (1980), Gordon (1981), Jambu és Lebeaux (1983).]

4.5. Klaszterek többdimenziós skálázása

Tegyük fel, hogy a változók többszörös többdimenziós skálázásával és/vagy egyéb módszerekkel ν-féleképpen (diszjunkt) klaszterekbe soroltuk az objektumokat. Kézenfekvő módon merül fel az a probléma, hogy miképpen lehet ezekből a klaszterezésekből egy "eredő" klaszterezést előállitani. A probléma megoldása céljából először jellemzem az ugyanazon klaszterezés melletti klaszterek távolságát. Jelölje $\underline{p}_r$ a klaszterek számát az $\underline{r}$-edik klaszterezésben, $\underline{Y}_m^{(r)}$ pe-

dig az ugyanezen klaszterezés melletti $\underline{m}$-edik klasztert ($\underline{m} = 1, 2, \ldots, \underline{p}_r$; $\underline{r} = 1, 2, \ldots, \nu$). $\underline{Y}_{m_1}^{(r)}$ és $\underline{Y}_{m_2}^{(r)}$ távolságát a következő módon jellemzem. Legyen

$$K_{m_1 m_2}^{(r)}(i) = \sum_{m} \frac{[n_i^{(rm)} \sum_{m} N^{(rm)} - N^{(rm)} \sum_{m} n_i^{(rm)}]^2}{N^{(rm)}[\sum_{m} n_i^{(rm)}] \sum_{m} [N^{(rm)} - n_i^{(rm)}]}, \qquad (4.5)$$

ahol $\underline{N}^{(rm)}$ és $\underline{n}_i^{(rm)}$ az $\underline{Y}_m^{(r)}$ klaszter összes, ill. azon objektumainak száma, amelyek rendelkeznek az $\underline{A}_i$ jelleggel, és az $\underline{m}$-re történő összegezés (4.5)-ben $\underline{m}_1$, $\underline{m}_2$-re terjed ki ($1 \leq \underline{m}_1 < \underline{m}_2 \leq \underline{p}_r$; $\underline{r} = 1, 2, \ldots, \nu$). Rögzített $\underline{i}$ mellett $\underline{j} = 1, 2$-re legyen $\nu_{j1} = \underline{n}_i^{(rm_j)}$, $\nu_{j2} = \underline{N}^{(rm_j)} - \underline{n}_i^{(rm_j)}$. Egyszerű számolással adódik, hogy ekkor

$$K_{m_1 m_2}^{(r)}(i) = \frac{(\nu_{11}+\nu_{12}+\nu_{21}+\nu_{22})(\nu_{11}\nu_{22}-\nu_{12}\nu_{21})^2}{(\nu_{11}+\nu_{12})(\nu_{21}+\nu_{22})(\nu_{11}+\nu_{21})(\nu_{12}+\nu_{22})},$$

ami azt méri, hogy egy objektumnak az $\underline{Y}_{m_1}^{(r)}$ vagy $\underline{Y}_{m_2}^{(r)}$ klaszterbe esése mennyire függ össze az $\underline{A}_i$ jelleg előfordulásával, más szóval mennyire különböző a $\underline{W}_i$ változó 1 értékének relativ gyakorisága a két klaszterben. $K_{m_1 m_2}^{(r)}(\underline{i})$ [lásd pl. Vincze (1968), 152. oldal] aszimptotikusan 1 szabadságfoku χ^2-eloszlást követ. Ebből következik, hogy ha a változók

függetlenek, akkor a

$$d^{(r)}_{m_1 m_2} = \sum_{i=1}^{n} K^{(r)}_{m_1 m_2}(i)$$

statisztika, ami azt jellemzi, hogy mennyire különböző az egyes változók 1 értékének gyakorisága az $\underline{Y}^{(r)}_{m_1}$ és $\underline{Y}^{(r)}_{m_2}$ klaszterekben, aszimptotikusan $\underline{n}$ szabadságfokú χ^2-eloszlást követ. A változók függősége esetén ez nem teljesül, de az eloszlással nem dolgozunk, ezért $\underline{d}^{(r)}_{m_1 m_2}$-et választom $\underline{Y}^{(r)}_{m_1}$ és $\underline{Y}^{(r)}_{m_2}$ távolságának. Az $\underline{Y}^{(r)}_1$, $\underline{Y}^{(r)}_2$, ..., $\underline{Y}^{(r)}_{p_r}$ klaszterek-nek a

$$\underline{\underline{D}}^{(r)} = [d^{(r)}_{m_1 m_2}]^{p_r}_{m_1, m_2 = 1}$$

távolságmátrix alapján történő (közönséges) többdimenziós skálázásával konstruáljunk a klaszterekhez olyan $\underline{\underline{z}}^{(r)}_1$, $\underline{\underline{z}}^{(r)}_2$, ..., $\underline{\underline{z}}^{(r)}_{p_r}$ pontokat a $\underline{k}_r$-dimenziós euklideszi térben, hogy a pontok euklideszi távolsága minél jobban tükrözze a klaszterek távolságát ($\underline{r} = 1, 2, \ldots, \nu$). Az $\underline{r}$-edik klaszterezés mellett rendeljük az [$\underline{\sigma}(\underline{G}_g)$ multiplicitású] $\underline{y}_g$ objektumhoz azt a $\underline{\underline{z}}^{(r)}_m$ pontot, amelyik az őt tartalmazó klaszternek megfelel, majd ezeknek a pontoknak képezzük azokat a $\underline{\underline{w}}^{(r)}_g$ lineáris transzformáltjait ($\underline{g} = 1, 2, \ldots, \underline{N}$), amelyekre

$$\sum_{g=1}^{N} \sigma(G_g)\underline{w}_g^{(r)} = \underline{0}, \qquad \sum_{g=1}^{N} \sigma(G_g)\|\underline{w}_g^{(r)}\|^2 = 1$$

($r = 1, 2, \ldots, \nu$; az összegezés g-re, tehát az objektumokra történik, és nem m-re, vagyis a klaszterekre). Legyen

$$\underline{v}_g = [\underline{w}_g^{(1)T}, \underline{w}_g^{(2)T}, \ldots, \underline{w}_g^{(\nu)T}], \quad g = 1, 2, \ldots, N,$$

és jelölje p^* a klaszterek kivánt számát az "eredő" klaszterezésben, C_m^* pedig az ugyanezen klaszterezés melletti m-edik klasztert ($m = 1, 2, \ldots, p^*$). Legyen

$$K_1 = \sum_{r=1}^{\nu} k_r.$$

Az C_m^* klaszterek az $\underline{v}_g$ objektumoknak a (K_1-dimenziós) $\underline{v}_g$ pontok alapján (pl. k-közép, pontosabban itt most p^*-közép eljárás által) történő klaszterezésével állithatóak elő. Ennek során az egyes klaszterezések jósága is figyelembe vehető oly módon, hogy a K_1-dimenziós távolságok, ill. normák számitásánál az r-edik koordináta-k_r-es az r-edik klaszterezés jóságának megfelelő sullyal szerepel. Ezt a sulyozást azonban nem javaslom. Az egyes klaszterező módszerek ugyanis - a 4.2. paragrafus végén mondottakkal össz-

hangban - nem hasonlithatóak objektiven össze. Nem is szükséges azonban a sulyozás, mert az "eredő" klaszterek kialakitásához a jobb klaszterezéseknek megfelelő koordináta-k_r-esek enélkül is nagyobb mértékben járulnak hozzá.

Az "eredő" klaszterezés előállitásának döntő mozzanata az objektumklaszterek MDS-e. Ez a korábbiakkal összhangban a következőket jelenti: távolságokat konstruálunk a klaszterek között, és a klaszterekét ugy próbáljuk meg beágyazni az alacsony dimenziós euklideszi térbe, hogy a klaszterekneknek megfelelő pontok euklideszi távolságai minél kevésbé különbözzenek a klaszterek távolságaitól. Ahhoz, hogy a klaszterek jól skálázhatóak legyenek, a változókhoz hasonlóan (ld. 4.?. paragrafus) konzisztenseknek kell lenniük. Ha ez nem teljesül, megkisérelhetjük a klaszterek MMDS-ét. Ez a következő problémával foglalkozik: hogyan lehet a jellegeket minél homogénebb csoportokba sorolni, amikoris egy-egy csoport homogenitását a klaszterek ezen csoport mellett történő (közönséges) MDS-ének jóságával mérjük? Rögzitett klaszterezés (r) mellett tehát keresendő az s természetes szám, a

$$B_1, B_2, \ldots, B_s \subset \{A_1, A_2, \ldots, A_n\} = A$$

diszjunkt jellegcsoportok (amelyekre

$$\bigcup_{\mu=1}^{s} B_\mu = A)$$

és a k_r-dimenziós euklideszi tér azon $\underline{\underline{x}}_m^{(\mu)}$ pontjai ($m = 1, 2, \ldots, p_r$; $\mu = 1, 2, \ldots, s$), melyek a klasztereket reprezentálják abban az értelemben, hogy $\underline{\underline{x}}_m^{(\mu)}$ és $\underline{\underline{x}}_j^{(\mu)}$ közelsége az f_m és f_j klaszterek B_μ melletti közelségének felel meg.

Dichotomizáljuk a klaszter-jelleg kapcsolatot a következőképpen: az A_i jelleget az Y_m klaszterre tipikusnak mondjuk, ha Y_m A_i-vel rendelkező objektumainak relativ gyakorisága nagyobb, mint egy alkalmas K_2 állandó. Legyen e_{im} értéke 1 vagy 0 aszerint, hogy A_i tipikus Y_m-re vagy sem ($m = 1, 2, \ldots, p_r$), és

$$\underline{\underline{e}}_i = (e_{i1}, e_{i2}, \ldots, e_{ip_r})$$

($i = 1, 2, \ldots, n$). Az MDS ezek után a 4.2. paragrafusban leirt módon végezhető el. - A K_2 állandót tapasztalataim szerint ugy célszerü választani, hogy azon jellegek átlagos száma, amelyek tipikusak egy klaszterre, közelitőleg $\sqrt{n}$ legyen.

4.6. Egy sávszélesség-redukcióval rokon probléma

Legyen

$$\underline{\underline{A}} = (a_{ij})_{i,j=1}^{n}$$

szimmetrikus, sok zérus-elemet tartalmazó mátrix. A sávszélesség-redukció problémája [lásd pl. George és Liu (1981), Chinn et al. (1982), Arany (1984)] a következő: melyik az a $\underline{\underline{P}}$ permutációs mátrix, amely mellett a

$$\underline{\underline{B}} = (b_{ij})_{i,j=1}^{n} = \underline{\underline{P}}\,\underline{\underline{A}}\,\underline{\underline{P}}^{T} \qquad (4.6)$$

mátrix

$$\max_{i,j:\,b_{ij}\neq 0} |i-j|$$

sávszélessége minimális? Ezzel rokon a következő probléma: melyik az a $\underline{\underline{P}}$ permutációs mátrix, amely mellett a (4.6) szerint meghatározott $\underline{\underline{B}}$ mátrixhoz tartozó

$$\sum_{i,j:\,b_{ij}\neq 0} |i-j| \;/\; \sum_{i,j:\,b_{ij}\neq 0} 1 \qquad (4.7)$$

mennyiség minimális? (Szemben a sávszélesség-redukcióval itt tehát nem maximumot, hanem átlagot minimalizálunk. A motiváció azonban ugyanaz: egy szimmetrikus ritka mátrixot a számítógép memóriájának minél kisebb részében elhelyez-

ni.) Ez a probléma heurisztikusan - és valószinüleg közelitően - (közönséges) többdimenziós skálázással oldható meg, az alábbi módon. Legyen

$$c_{ij} = \begin{cases} 1, & \text{ha } a_{ij} = 0, \\ 2, & \text{egyébként;} \end{cases}$$

továbbá $\underline{d}_{ij} = \underline{K} / \underline{c}_{ij}$ $(1 \leq \underline{i} < \underline{j} \leq \underline{n}$; $\underline{K}$ alkalmas állandó). Minimalizáljuk a (4.4) kifejezést, és jelölje $\underline{\underline{x}}_1, \underline{\underline{x}}_2, \ldots, \underline{\underline{x}}_n$ az igy kapott pontokat. Legyen $(\hat{\underline{d}}_{ij})$ az

$$[\underline{\underline{x}}_1, \underline{\underline{x}}_2, \ldots, \underline{\underline{x}}_n]$$

pontkonfiguráció euklideszi távolságmátrixa. "Füzzük fel" az $\underline{\underline{x}}_i$ pontokat ugy, hogy a $\hat{\underline{d}}_{ij}$ távolságok növekvő sorrendje szerint összekötjük az egyes pontpárokat, de csak akkor, ha i) a két pont egyike sincs még 1-nél több ponttal összekötve, ii) legfeljebb $(\underline{n}-2)$ számu pontpár van még csak összekötve. Ily módon sorba rendezem a pontokat, és ezzel kijelölök egy $\underline{\underline{P}}$ permutációs mátrixot. Tapasztalataim szerint az igy nyert $\underline{\underline{P}}$ mellett (4.6) szerint meghatározott (4.7) menynyiség általában kisebb, mint ha $\underline{\underline{P}}$-t ugy állitom elő, hogy a $\hat{\underline{d}}_{ij}$ távolságok helyett magukkal a $\underline{d}_{ij}$-kkel dolgozom (és jóval kisebb, mint ha $\underline{\underline{P}}$-nek az egységmátrixot választom).

Ha a (4.4) kifejezést 2 helyett más, kellően nagy hat-

ványkitevő mellett minimalizálom, magára a sávszélesség-redukció problémájára kapok - természetesen heurisztikus és általában közelitő - megoldást. A viszonylag jelentős gépidőigény miatt ez azonban csak elméleti szempontból érdekes.

5. VELESZÜLETETT RENDELLENESSÉGEK STATISZTIKAI VIZSGÁLATA

A fogamzást követően a születésig alaki (szerkezeti), müködési és/vagy biokémiai fogyatékosságok alakulhatnak ki az embrióban vagy a magzatban. Egy ilyen, a születés pillanatában létező - esetleg csak később észlelt - fogyatékosságot veleszületett anomáliának (angolul "congenital anomaly") nevezünk. Leggyakoribb fajtája a veleszületett abnormitás (angolul "congenital abnormality"), amely a hibás egyedfejlődés következménye, a születés pillanatában létező szerkezeti fogyatékosság. A veleszületett abnormitás leggyakoribb fajtája a veleszületett malformáció (angolul "congenital malformation", röviditve CM), amely abnormális benső fejlődési folyamat eredményeként létrejövő alaki fogyatékosság. A magyar nyelvü szakirodalomban a veleszületett anomália, abnormitás és malformáció kifejezések nem terjedtek el, hanem többnyire mindhármójuk helyett a veleszületett rendellenesség kifejezést használják. A továbbiakban a veleszületett rendellenesség kifejezés azt a veleszületett abnormitásnál valamivel szélesebb, a veleszületett anomáliánál viszont lényegesen szükebb kategóriát fogja jelenteni, amelyik hazánkban bejelentésköteles. Ezt a fajta veleszületett rendellenességet a "congenital anomaly" és "congenital abnormality" kifejezések közös kezdőbetüivel CA-nak fogom röviditeni.

Jelenleg Magyarországon a regisztrált CA-k gyakorisága mintegy 4,5 százalék. A CA-k tényleges gyakorisága valószínűen nagyobb: a születések 6-8 százalékában kell velük számolni. A csecsemőhalandóság (1973 és 1982 között 2,706%) egyötödét CA-k okozzák. (Ez hatszor annyi csecsemőhalált jelent, mint amennyit az összes fertőző betegség együtt előidéz.) Az életben maradottaknak gyakran számottevő orvosi és társadalmi problémáik vannak, hiszen a CA-k többnyire nem gyógyithatóak teljesen. Ezért is a legjobb megoldás a megelőzés lenne. Ennek hatékonysága azonban erősen függ a CA-k okának és általában a CA-knak az ismeretétől. Ez tette szükségessé a CA-k statisztikai vizsgálatát, aminek Magyarországon orvosi oldalról Czeizel Endre, matematikai oldalról Tusnády Gábor volt a megindítója és irányitója, és aminek része ez az értekezés is, az előző három fejezetnek a továbbiakban ismertetésre kerülő alkalmazása.

Magyarországon 1962-ben inditották be (másodiknak a világon) a CA-k folyamatos és kötelező bejelentését, valamint a Veleszületett Rendellenességek Országos Nyilvántartását (VRONY). A CA-k osztályozása a VRONY-ban annak 1970-es megujitása óta a Betegségek Nemzetközi Osztályozásán (jelenleg annak 9. Módositásán) alapul, bizonyos változtatásokkal.

Egy CA önmagában való előfordulása esetén <u>izolált</u>, több CA ugyanannál a személynél történő együttes előfordulása esetén <u>többszörös veleszületett rendellenességről</u> (angolul "multiple" CA, MCA) beszélünk. Utóbbiak különös fon-

tossággal birnak [lásd pl. Cohen (1977), Czeizel (1981)].

A CA-k gyakoriságuk alapján is értékelhetőek. Eszerint három csoportjukról szokás beszélni. A gyakori CA-k születéskori gyakorisága meghaladja az 1‰-et. A ritka CA-k előfordulása nem éri el a 0,1‰ születéskori gyakoriságot. A kettő között foglalnak helyet a közepes gyakoriságu CA-k. - Az izolált (angolul "isolated") gyakori (angolul "common") CA-eket ICCM-nek fogom rövidíteni.

5.1. ICCM-ek öröklődésének vizsgálata

Ismert tény, hogy egy sor veleszületett rendellenesség több családtagnál, halmozottan fordulhat elő [lásd pl. Czeizel (1984)]. Ezért is fontos ezen betegségek öröklődésének vizsgálata. A következő kilenc ICCM-mel foglalkoztam ebből a szempontból: koponyahiány és /vagy nyitott gerinc, nyulajak, a gyomorkimenet szűkülete, szivkamra-sövény rendellenesség, csipőficam, dongaláb, lágyéksérv, hugycsőhasadék, le nem szállt here. Hazánkban ezek a nyilvántartott összes CA születéskori gyakoriság 51,5‰-át képezik. Valódi gyakoriságukat figyelembe véve részesedésük még magasabb is lehet, szemléletesen érzékeltetve az ICCM-ek közegészségügyi jelentőségét.

Az egyes rendellenességek öröklődésmenetének leirására egy speciális normális küszöb modellt választok. Ez feltételezi, hogy a vizsgált rendellenességnek mindenkire nézve

van egy valós számmal kifejezhető mértéke, vagyis a betegséghez hozzá van rendelve egy L háttér változó. (Szokás ezt az L-et hajlamnak is nevezni.) A modell szerint L standard normális eloszlásu valószinüségi változó, amelyre

$$L = G + E,$$

ahol G a genetikai, E a környezeti hatást jelenti, G és E független, 0 várható értékü, h^2, ill. $(1-h^2)$ szórásnégyzetü, normális eloszlásu valószinüségi változók, továbbá az, hogy valaki beteg, azt jelenti, hogy rendellenességének mértéke elér egy - a vizsgált egyedet tartalmazó populációra jellemző - küszöböt. [h^2 az örökölhetőség vagy örökletesség, közismertebb néven örökölhetőségi együttható (lásd pl. Tusnády, 1969; Sváb, 1971; Fischer et al., 1974). G és E mindegyikéről felteszem, hogy multifaktoriális hatásnak, tehát sok kis tényező hatásának az eredménye. Felteszem, hogy az un. domináló variancia (más néven dominanciavariancia, lásd pl. Sváb, 1971; Tusnády et al., 1978b) 0, és mivel igy G varianciája teljes mértékben az un. additiv varianciából (lásd pl. Sváb, 1971; Tusnády et al., 1978b) származik, G-t tisztán additivnek tekintem.] A modellben a rokonok rendellenességének mértékei együttes normális eloszlásuak és a rokonsági foktól függően korreláltak: d-edfoku rokonok rendellenessége mértékeinek ($h^2/2^d$) a korrelációs együtthatója.

A fenti, ilyen formában Czeizel Endre és Tusnády Gábor által bevezetett modellt [v.ö. Carter (1961), Falconet (1965), Schuler et al. (1974)], amelyben a rendellenesség mértéke mellett legdöntőbb a normális (Gauss-eloszlású, angolul "Gaussian"), additiv (angolul "additive"; két értelemben is: egyrészt G tisztán additiv, másrészt G-t és E-t additiven vesszük figyelembe), multifaktoriális (angolul "multifactorial") hatás és a küszöb (angolul "threshold"), GAMT-modellnek nevezik [lásd Tusnády et al. (1978b), Czeizel és Tusnády (1984)]. Az ICCM-ek vizsgálatánál célom ezt figyelembe véve különböző rokoncsoportokra, ill. ilyenek bizonyos összességeire a h^2 becslése és a becslés megfelelő szintü konfidenciaintervallumba foglalása volt, továbbá annak ellenőrzése, hogy a különböző h^2-becslések eltérése szignifikáns-e vagy sem. Ez utóbbi a modell ellenőrzését jelentette. Azon rendellenességekre, amelyekre a modell illeszkedett a genetikai családvizsgálat adataihoz, a kapott h^2-értékeket jól lehetett használni a genetikai tanácsadásban.

Rögzitett ICCM esetén jelölje p_1 és p_2 a betegség relativ gyakoriságát a vizsgált betegeknek megfelelő férfi, ill. női kontrollcsoportban, $\hat{T}_1$ és $\hat{T}_2$ a küszöbök ezen adatok által a GAMT-modell szerint meghatározott becslését, akkor $v = 1, 2$ esetén

$$\hat{T}_v = \Phi^{-1}(1-p_v).$$

Adataink lehetnek a betegek nulladfoku (egypetéjü ikrek), elsőfoku (szülők, testvérek), másodfoku (szülők testvérei, testvérek gyermekei) és harmadfoku (unokatestvérek) rokonairól. A zárójelben emlitett rokoncsoportokat rétegeknek nevezem. Jelölje $\tilde{p}_{2R-1}$, $\tilde{T}_{2R-1}$, valamint $\tilde{p}_{2R}$, $\tilde{T}_{2R}$ a betegség relativ gyakoriságát és a küszöb becslését az R-edik rétegnek megfelelő férfi, ill. női kontrollcsoportban, akkor $v = 0, 1$ esetén

$$\tilde{T}_{2R-v} = \Phi^{-1}(1-\tilde{p}_{2R-v}), \quad R = 1, 2, \ldots, 6.$$

Legyen $u = 1, 2$, ha fiu-, $u = 3, 4$, ha leánygyermekek férfi, ill. női rokonait vizsgáljuk, és $j = 4(R-1) + u$. Nyilván $1 \leq j \leq 24$. Jelölje m_j és M_j a vizsgált betegek összes, ill. rendellenes (a szóbanforgó ICCH-mel rendelkező) rokonainak számát a j-edik esetben. Az egyes veleszületett rendellenességekre vonatkozóan a következő adatok álltak rendelkezésemre: p_v, $\tilde{p}_{2R-v}$, m_j, M_j. Az adatokból a h^2-et a maximum likelihood módszerrel becsültem. M_j binomiális eloszlását az alább definiálandó λ_j paraméterü Poisson-eloszlással közelitettem, és feltettem, hogy a különböző betegek rokonai között végzett megfigyelések függetlenek (tehát hogy a különböző gyerekek rokonai között esetleg fennálló rokoni kapcsolatok elhanyagolhatóak).

Definiáljuk $P_r(\tilde{T},T)$-t mint a 46. oldalon, és legyen

$$p_j(r) = P_r(\tilde{T}_{2R-j(\mathrm{mod}\ 2)}, \hat{T}_{1+\mathrm{entier}(\frac{j-1}{2})(\mathrm{mod}\ 2)}),$$

$$\lambda_j = \lambda_j(x) = m_j p_j(x/2^d), \qquad w_j(x) = \log \frac{\lambda_j(x)^{M_j} e^{-\lambda_j(x)}}{M_j!}$$

($\underline{d}$ a $\underline{j}$-edik esethez tartozó, $\underline{R}$-edik rokoncsoport rokonsági foka). Jelölje $\underline{J}$ a $\underline{j}$-k egy olyan összességét, amely mellett a $\underline{h}^2$ becsülendő, akkor a likelihood függvény $\underline{J}$-re vonatkozólag

$$G_J(x) = \sum_{j \in J} w_j(x).$$

Jelöljük $\underline{ML}_J$-vel a likelihood függvény maximumát, és legyen $\underline{h}_J^2$ az a szám, ahol ez a maximum felvétetik (ez nyilván $\underline{h}^2$ becslése $\underline{J}$ mellett). Az ehhez tartozó $\underline{h}_{JA}^2$, $\underline{h}_{JF}^2$ alsó és felső konfidenciahatárt a

$$G_J(x) = ML_J - \tfrac{1}{2}\chi_1^2(1-\varepsilon) \qquad (5.1)$$

egyenletből határoztam meg, ahol ε megfelelően kis pozitiv szám. Ha teljesülnek Serfling (1980) 4.2.2. paragrafusának (144-149. oldal) regularitási feltételei, akkor (lásd u.o., 155-156. oldal) $2[\underline{G}_J(\underline{x})-\underline{ML}_J]$ aszimptotikusan 1 szabadságfoku χ^2-eloszlást követ. Az az érzésem, hogy ezek a feltételek teljesülnek, de ezt bizonyitani nem tudom. Az (5.1) egyenlet által megadott konfidenciaintervallumot mindazonáltal használhatónak vélem.

Ha $\underline{J} = \{j\}$ egyelemü, akkor

$$G_J(x) = w_j(x),$$

ez pedig

$$\lambda_j(x) = M_j \tag{5.2}$$

esetén maximális. Ebből következik, hogy

$$ML_J = ML_j = \log \frac{M_j^{M_j} e^{-M_j}}{M_j!},$$

$\underline{h}_J^2 = \underline{h}_j^2$ pedig az (5.2) egyenletből számolható. Mivel $\lambda_j(\underline{x})$ monoton növő, ez a számolás, valamint a többelemü $\underline{J}$ esetén történő maximumkeresés és az (5.1) egyenlet numerikus megoldása (vagyis a $\underline{h}^2$ pont- és intervallumbecslése) az

$$F(\tilde{T},T,r) = P_r(\tilde{T},T)[1-\Phi(T)]$$

mennyiséget meghatározó, a 3. fejezetben ismertetett eljárás birtokában nem okoz nehézséget.

A különböző $\underline{h}^2$-becslések azonosságát a $\underline{h}_J^2 \equiv \underline{h}^2$ hipotézis vizsgálatával ellenőriztem.

$$J = \bigcup_{k=1}^{k^*} J_k$$

esetén

$$2\left(\sum_{k=1}^{k^*} ML_{J_k} - ML_J\right)$$

a próbafüggvény.

Nevezzük az egyes rétegek rögzitett u-hoz tartozó eseteinek összességét az u-adik oszlopnak ($u = 1, 2, 3, 4$). Ha a veleszületett rendellenesség mindkét nemre jellemző, és a nemek nem ömlesztve szerepelnek, a fenti, bizonyos értelemben horizontális hierarchián kivül vertikálisan is összevontam: a rétegekhez teljesen hasonlóan az oszlopokra is számoltam a megfelelő h^2, h_A^2, h_T^2 és ML értékeket. Az oszlopok öröklhetőségi együtthatóinak összehasonlitása a

$$2\left(\sum_{u=1}^{4} ML_{u\text{-adik oszlop}} - ML_{\text{összes eset}}\right)$$

próbafüggvény értékének meghatározásával történt. Ha teljesülnek Serfling (1980) emlitett regularitási feltételei, akkor (lásd u.o., 156-160. oldal) a fenti két próbafüggvény aszimptotikusan (k^*-1), ill. 3 szabadságfoku X^2-eloszlást követ. Az az érzésem, hogy ezek a feltételek teljesülnek, de ezt bizonyitani nem tudom.

A kilenc ICCM-et megvizsgálva általános volt, hogy

$h^2_{testvér} > h^2_{1.foku}$, $h^2_{2.foku} > h^2_{1.foku}$, $h^2_{3.foku} > h^2_{1.foku}$.

Az első egyenlőtlenség a domináló variancia pozitiv voltával volt magyarázható, a másik kettőt az együttes normalitás gyengitésével (csak a peremeloszlásokat feltételezve normálisnak) sikerült az egyenlőség felé közeliteni [lásd Tusnády et al. (1981), Czeizel és Tusnády (1984)].

5.2. Az MCA-k értékelése

A VRONY-ban pontosan és részletesen szereplő MCA-k értékeléséhez a ritka rendellenességek lehetőség szerinti összevonásával és bizonyos további meggondolások alapján a CA-k $n = 40$ csoportba lettek sorolva [lásd Czeizel et al. (1981), Telegdi (1983). Az egyes CA-csoportokat két-két nagybetüvel jelölöm]. Ezek a CA-csoportok a következők: koponyahiány (AN), agysérv (EN), nyitott gerinc (SB), nyulajak (CL), farkastorok (CP), végtagredukció (LR), többujjuság (PI), összenőtt ujjak (SY), nyitott has (EX), nyelőcsőelzáródás vagy -szükület (OA), végbélelzáródás vagy -szükület (AA), kisszemüség vagy szemnélküliség (AM), szürkehályog (CT), hugycsőhasadék (HS), a külső nemi szervek egyéb rendellenességei (EG), szivrendellenességek (HD), rekeszizomsérv (DI), vesehiány (RA), veseciszta (CK), vékonybélelzáródás (AI), kisfejüség (MC), vizfejüség (HY), egyéb

agyi és idegrendszeri rendellenességek (ON), az arc, orr, nyelv, nyak és koponya egyéb rendellenességei (FN), egyéb szemrendellenességek (EY), fülrendellenességek (EA), ferde nyak (TC), nyaki ciszták, sipolyok és fül előtti függelék (BR), csipőficam (CD), dongaláb (CF), egyéb végtagrendellenességek (OL), a légzőszervek rendellenességei (RS), a gyomorkimenet szűkülete (PS), egyéb zsigeri rendellenességek (OD), le nem szállt here (UT), egyéb hugy-ivarrendszeri rendellenességek (OU), lágyéksérv (IH), egyéb mozgásszervi rendellenességek (MS), az endokrin szervek rendellenességei (EO), bőr-, haj- és körömrendellenességek, daganatok, valamint egyéb rendellenességek (ST). A CA-k mindössze 40 csoportba történő besorolása (az egyszerüség kedvéért a továbbiakban ezeket a CA-csoportokat fogom CA-knak nevezni) természetesen bizonyos információveszteséget jelent, és csökkenti az azonositás pontosságát, ugyanakkor azonban megkönnyiti az MCA-k statisztikai leirását.

Magyarországon 1970-76 között 1 188 529 gyermek született. Közülük 32 603 szerepel a VRONY-ban CA-val rendelkező gyermekként. Ezek közül 4515-nek (13,8 százalék) több CA-ja volt. Az ilyen esetek 34,8 százalékában (1573 gyermek) valamilyen szindróma (feltehetően közös kóreredetü CA-k felismert együttese) megállapitása miatt, 4,0 százalékában (180 gyermek) egyéb okból a konkrét CA-k nincsenek részletezve; ezt az 1753 gyermeket ki kellett zárni a statisztikai értékelésből. Igy 30 850 CA-val rendelkező

(közülük 921 különböző) és ezen belül 2762 MCA-val rendelkező (közülük $\underline{N}$ = 881 különböző) gyermek maradt $\underline{M}$ = = 1 186 776 gyermek közül a többszörös veleszületett rendellenességek statisztikai értékeléséhez. [Az időkorlátozásnak, annak, hogy az utolsó vizsgált év 1976, az a magyarázata, hogy a munka 1977-ben kezdődött (lásd Tusnády et al., 1978a; Bolla et al., 1979) az akkor meglévő, 1970-76-os adatokkal, és az adatmezőt később nem lett volna célszerű megváltoztatni. Később megtörtént az MCA-k 1977-82-es adatok alapján történő értékelése, de természetesen felhasználva a korábbi időszak vizsgálatának eredményeit.] Az 1753 gyermek kizárásából fakadó hiba, melynek korrekciójára nem volt lehetőség, egy irányban torzit: a nagyobb méretű MCA-k számát csökkenti jobban. Ugyanebben az irányban és valószínüleg még erősebben torzit a természet: nagyobb méretű MCA-k esetén gyakoribb a spontán abortusz. Ezen torzitás mértéke tudomásom szerint nem ismert.

Az adatok tárolása az 1.1. paragrafusban leirt módon történt. Az ott definiált $\underline{NS}$ tömb méretét az ott leirt módokon csökkentve az előfordult 2568-féle G_g rendellenességkombinációhoz egy 9268 elemü $\underline{NS}$ volt szükséges.

Jelölje $\underline{K}$ a G_g-k méretének maximumát, vagyis azt a számot, ahány CA-val egy gyermek maximálisan rendelkezik. Esetünkben $\underline{K}$ értéke 7 volt, a CA-val rendelkező gyermekek száma

$$\sum_{k=1}^{K} \sigma^{(k)} = 30\ 850,$$

az MCA-val rendelkezőké

$$\sum_{k=2}^{K} \sigma^{(k)} = 2762,$$

a pontosan 1 CA-val rendelkezőké

$$\sigma^{(1)} = 30\ 850 - 2762 = 28\ 088,$$

az egészséges (CA-val nem rendelkező) gyermekeké pedig

$$\sigma = N - 30\ 850 = 1\ 155\ 936.$$

A pontosan $\underline{k}$ számú CA-val rendelkező gyermekek $\underline{\sigma^{(k)}}$ száma $\underline{k}$ = 2, 3, ..., 7 esetén rendre a következő volt: 2128, 448, 172, 50, 11 és 3. A 30 850 CA-val rendelkező gyermek összesen $\underline{\sigma_T^{(1)}}$ = 34 513 CA-val rendelkezett, így egy tetszőleges gyermek átlagosan $\underline{\sigma_T^{(1)}}$ / $\underline{N}$ = 0,0291, egy CA-val rendelkező gyermek pedig átlagosan $\underline{\sigma_T^{(1)}}$ / 30 850 = 1,119 CA-val. A gyermekek között előfordult pontosan $\underline{r}$ elemű CA-kombinációk $\underline{\sigma_T^{(r)}}$ száma $\underline{r}$ = 2, 3, ..., 7 esetén rendre 4932,

1761, 642, 179, 32 és 3 volt, a pontosan $\underline{k}$ számú CA-val rendelkező gyermekek körében és az összes gyermek közt előfordult különböző $\underline{k}$ elemü CA-kombinációk $N^{(k)}$ és $N_T^{(k)}$ száma pedig a következő:

$$
\begin{aligned}
N^{(0)} &= N_T^{(0)} = 1, \\
N^{(1)} &= N_T^{(1)} = n = 40, \\
N^{(2)} &= 358, \quad N_T^{(2)} = 562, \quad [\binom{n}{2} = 780,] \\
N^{(3)} &= 340, \quad N_T^{(3)} = 1141, \quad [\binom{n}{3} = 9880,] \\
N^{(4)} &= 119, \quad N_T^{(4)} = 612, \\
N^{(5)} &= 50, \quad N_T^{(5)} = 178, \\
N^{(6)} &= 11, \quad N_T^{(6)} = 32, \\
N^{(7)} &= N_T^{(7)} = 3.
\end{aligned}
$$

Összegükre

$$\sum_{k=0}^{K} N^{(k)} = 1 + n + N = 922$$

és

$$\sum_{k=0}^{K} N_T^{(k)} = 1 + 2568 = 2569.$$

5.3. A függetlenségi koncepció

Ugyanaz a CA különböző esetekben különböző kóreredetű lehet (a kóreredet általában nincs bejelentve, ezért az MCA-k statisztikai vizsgálatánál ismeretlen volt). Az ICCM-eknek az 5.1. paragrafusban ismertetett vizsgálata is hozzájárult annak tisztázásához, hogy a CA-val rendelkező gyerekek jelentős részénél a CA(k) sok tényező hatásának az eredménye(i), tehát multifaktoriális kóreredetű(ek). Ellenkező esetben, amikoris a CA(k) kevés tényező hatásának az eredménye(i), oligofaktoriális kóreredetről fogok beszélni (jobb hiján). Az MCA-k egy kézenfekvő és sokáig elfogadott magyarázatát a CA-k un. függetlenségi koncepciója nyujtja [lásd pl. Czeizel és Tusnády (1984)]. Ennek lényege, hogy a multifaktoriális kóreredet szoros kapcsolatban van a CA-k véletlen előfordulásával. Egy multifaktoriális kóreredetü CA független a többi CA-tól, igy az ilyen CA-t tartalmazó MCA-k véletlen kombinációk. A többi MCA oligofaktoriális kóreredetü. A két tipus azonban - a kóreredet bejelentetlensége miatt - az MCA-k statisztikai vizsgálatában megkülönböztethetetlen. Tulajdonképpen tehát arról van szó, hogy tetszőleges MCA p_1, ill. p_2 valószinüséggel multi- vagy oligofaktoriális, vagyis az MCA-k valószinüségeloszlása két eloszlás keveréke, de sem a p_1, p_2 valószinüségeket nem ismerjük, sem azt nem tudjuk, hogy az egyes MCA-k a keverék melyik eloszlásához tartoznak.

A függetlenségi koncepció nem teszi fel a CA-k bármelyik kóreredet melletti függetlenségét, tehát azt, hogy minden MCA véletlen kombináció. Ez utóbbi hipotézist egy viszonylag kis, 232 MCA-s gyermeket tartalmazó vizsgálati anyag alapján Roberts és Powell (1975) elutasitotta, mivel két vagy több CA ugyanannál a gyermeknél gyakrabban fordult elő, mint az függetlenség esetén várható lett volna. Lényegesnek látszott azonban ennek az elutasitásnak jogosságát a fentebb ismertetett hétéves, országos anyagon ellenőrizni. Ezen tulmenően fontos jellemezni az MCA-k és a véletlen kombinációk gyakorisága eltérésének mértékét és ennek fényében dönteni a függetlenségi koncepció elfogadásáról.

Mivel a CA-k száma $n = 40$, azért a megfelelő kontingenciatáblázat celláinak száma 2^{40}. Ezen cellaszám mellett a 33. oldal χ^2-próbával kapcsolatos megállapitása messzemenően érvényes, a próba elvégzése illuzórikus. Ezért a 2. fejezettel összhangban a CA-k bármelyik kóreredet melletti függetlensége vizsgálatának első lépéseként a (2.1) egyenlőség teljesülését ellenőriztem. (2.1) jobb oldalának 1 153 534 volt az értéke (σ-nak 1 155 966). Válasszuk ezután M^*-ot ugy, hogy (2.1) jobb oldalán M helyett M^*-ot irva már teljesüljön az egyenlőség. Ekkor $M^* = 1\,184\,320$. Az M mintaelemszámnak M^* (a következő paragrafusban majd $\tilde{M}$; mindkettő kisebb mint M) értékre történő cserélése egy arra irányuló kisérlet első lépése, hogy az MCA-k két eloszlás keverékének feltételezett valószinüség-

eloszlásából leválasszuk azt az eloszlást, amelyben a CA-k nem függetlenek. (Ez most - a függetlenségi koncepciónak megfelelően - az oligofaktoriális, a következő paragrafusban a multifaktoriális kóreredetü MCA-k elhagyását jelentené.) A kisérlet második lépése az elhagyandó objektumok specifikálása és a $\underline{p}^{(k)}$ valószinüségek becslésének módositása lenne, amit a két lépés váltakozva történő ismételt alkalmazása követne addig, amig a (2.1) képlet két oldalának különbsége kellően kicsivé nem válik. Az iteráció alaposabb vizsgálatát a következő paragrafusban említésre kerülő, Tusnády Gábortól származó általánosabb keverékfelbontó eljárás tette szükségtelenné.

Legyen $\underline{p}^{(k)}$ annak valószinüsége, hogy egy gyermek pontosan $\underline{k}$ számu CA-val rendelkezik. $\underline{L}^{*}\hat{\underline{p}}^{(k)}$ értéke $\underline{k}$ = 2, 3, ..., 7 esetén rendre a következő volt: 303,78; 1,97; $3{,}7\cdot10^{-3}$; $2{,}8\cdot10^{-5}$; $6{,}8\cdot10^{-8}$; valamint $1{,}3\cdot10^{-10}$. Ezek az értékek igen nagy mértékben különböznek a megfelelő $\underline{o}^{(k)}$ gyakoriságoktól: a sejtésem szerint aszimptotikusan $\underline{n}$ - 2 = = 38 szabadságfoku X^2-eloszlást követő (2.3) kifejezés értéke - $\underline{L}$ helyett $\underline{L}^{*}$-gal számolva - 10^{10} nagyságrendü [$X^2_{38}(0{,}05) = 53{,}384$]. Ez azt mutatja, hogy a CA-k bármelyik kóreredet melletti függetlensége igen nagy mértékben nem teljesül. {Ez a mérték azonban erősen függ az MCA-mérettől: $\underline{L}^{*}\hat{\underline{p}}^{(k)}$ és $[\underline{o}^{(k)}-\underline{L}^{*}\hat{\underline{p}}^{(k)}]$ aránya $\underline{k}$ = 2, 3, 4 mellett rendre 1 : 6, 1 : 226 és 1 : $1{,}4\ 10^4$, továbbá

$$\sum_{k=5}^{K} M^{*}\hat{P}^{(k)} : \sum_{k=5}^{K} [\sigma^{(k)} - M^{*}\hat{P}^{(k)}] = 1 : 2{,}3\cdot 10^{6}.\}$$

$\underline{k}$ = 2, 3 mellett $\underline{N}^{(k)}$ és $\underline{N}_T^{(k)}$ várható értékének becslése a következő volt:

$$\hat{N}^{(2)} = 130{,}80; \quad \hat{N}_T^{(2)} = 132{,}57;$$
$$\hat{N}^{(3)} = 1{,}96; \quad \hat{N}_T^{(3)} = 1{,}99.$$

Mindenképpen indokolt tehát a CA-k bármelyik kóreredet melletti függetlenségének elutasitása. Ugyanez azonban a függetlenségi koncepcióra is érvényes: ennek fennállása esetén ugyanis az MCA-k 89 százaléka oligofaktoriális MCA lenne, ami orvosilag irreális.

A 2. fejezetben leirt módon megvizsgáltam, hogy a CA-k bármelyik kóreredet melletti függetlenségének, ill. a függetlenségi koncepciónak az alfogadhatatlanságát nem csak néhány CA okozza-e. Azt kaptam, hogy nem, mivel a (2.4) kifejezés csupán három CA mellett volt - 1 szabadságfoku χ^2-eloszlást feltételezve - szignifikáns: AA és EA esetén, amikor az $\underline{a}_i$ meredekség pozitiv, valamint CD esetén, amikor $\underline{a}_i$ negativ.

5.4. A függetlenségi koncepció módosítása

Az 5.2. paragrafusban említettem azokat a torzításokat, amelyek a nagyobb méretü MCA-k száma nagyobb mértékü csökkentésének irányában hatnak. A függetlenségi koncepción tulmenő vizsgálatokat az indokolta, hogy a valóság a modelltől az ellenkező irányban tért el.

A függetlenségi koncepció "duálisa" az a - lényegében Källen és Winberg (1969)-ből származó - feltételezés, amely szerint az oligofaktoriális MCA-kban a CA-k függetlenek, de a multifaktoriális kóreredetüek nem feltétlen azok. Tetszőleges $\{\underline{A}_{i_1}, \underline{A}_{i_2}, \ldots, \underline{A}_{i_k}\}$ CA-kombináció esetén jelölje $\underline{P}_T(\underline{i}_1,\underline{i}_2,\ldots,\underline{i}_k)$ annak valószinüségét, hogy egy gyermek rendelkezik $\underline{A}_{i_1}$-gyel, $\underline{A}_{i_2}$-vel, ..., $\underline{A}_{i_k}$-val és esetleg további CA-kkal is. $\underline{k} = 1, 2, \ldots, \underline{n}$, $\underline{j} = 1, 2, \ldots, \underline{k}$ és $1 \leq \underline{i}_1 < \underline{i}_2 < \ldots < \underline{i}_j \leq \underline{n}$ esetén jelölje $\underline{P}^{(k)}(\underline{i}_1,\underline{i}_2,\ldots,\underline{i}_j)$ annak valószinüségét, hogy egy gyerek pontosan $\underline{k}$ számu CA-val rendelkezik, köztük $\underline{A}_{i_1}$-gyel, $\underline{A}_{i_2}$-vel, ..., $\underline{A}_{i_j}$-vel, $\underline{\sigma}^{(k)}(\underline{i}_1,\underline{i}_2,\ldots,\underline{i}_j)$ pedig az ilyen gyerekek számát. Szem előtt tartva, hogy egyrészt a nagyméretü MCA-k még a multifaktoriális kóreredetü CA-k függősége esetén is - orvosi megfontolások alapján - azért zömmel oligofaktoriálisak, másrészt a legalább 5 CA-val rendelkező gyerekek száma csak 64, a duális koncepció ellenőrzéséhez a $\underline{P}_T(\underline{i})$ valószinüségeket ugy becsültem, hogy a CA-k bármelyik kóreredet melletti függetlenségét feltételezve $\hat{\underline{P}}^{(3)} = \underline{\sigma}^{(3)} / \underline{M}$ és

$$\hat{P}^{(4)}(i) = \sigma^{(4)}(i) / M, \quad i = 1, 2, \ldots, n, \qquad (5.3)$$

teljesüljön. Legyen $\underline{P} = \underline{P}^{(0)}$, és vezessük még be a következő jelöléseket:

$$q_i = P_T(i) / [1-P_T(i)], \quad i = 1, 2, \ldots, n,$$

$$S_0 = 1,$$

$$S_k = \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} \prod_{j=1}^{k} q_{i_j}, \quad k = 1, 2, \ldots, n,$$

$$\bar{S}_k(i) = \sum_{\substack{1 \le i_1 < i_2 < \ldots < i_k \le n \\ i_1, i_2, \ldots, i_k \neq i}} \prod_{j=1}^{k} q_{i_j}, \quad \begin{matrix} k = 1, 2, \ldots, n-1, \\ i = 1, 2, \ldots, n. \end{matrix}$$

Amint azt a 2. fejezetben megemlitettem, a CA-k függetlensége miatt $\underline{P}^{(3)} = \underline{P}\underline{S}_3$. Ugyancsak a függetlenségből, valamint $\underline{P}^{(k)}(\underline{i})$ és $\underline{\bar{S}}_k(\underline{i})$ definiciójából egyszerü számolással következik, hogy

$$P^{(4)}(i) = Pq_i\bar{S}_3(i), \quad i = 1, 2, \ldots, n,$$

ezért

$$\frac{P^{(4)}(i)}{P^{(3)}} = q_i \frac{\overline{S}_3(i)}{S_3} = \frac{P_T(i)}{1-P_T(i)} \cdot \frac{\overline{S}_3(i)}{S_3}, \quad i = 1, 2, \ldots, n,$$

ahonnan egyszerű számolással adódik, hogy

$$P_T(i) = \frac{P^{(4)}(i)}{P^{(4)}(i) + P\overline{S}_3(i)}, \quad i = 1, 2, \ldots, n.$$

tehát

$$\hat{P}_T(i) = \frac{\alpha^{(4)}(i)}{\alpha^{(4)}(i) + \hat{P}\hat{S}_3(i)}, \quad i = 1, 2, \ldots, n.$$

Mivel a CA-k függetlensége miatt

$$P = \prod_{i=1}^{n} [1-P_T(i)] = \prod_{i=1}^{n} \left[1 - \frac{P^{(4)}(i)}{P^{(4)}(i) + P\overline{S}_3(i)}\right] =$$

$$= \prod_{i=1}^{n} \frac{P\overline{S}_3(i)}{P^{(4)}(i) + P\overline{S}_3(i)},$$

azért teljesülnie kell, hogy

$$\alpha^{(3)} = [M\hat{P}^{(3)} = M\hat{P}\hat{S}_3 =] M\hat{S}_3 \prod_{i=1}^{n} \frac{\hat{P}\hat{S}_3(i)}{\hat{P}^{(4)}(i) + \hat{P}\hat{S}_3(i)}. \qquad (5.4)$$

Ez nem pontosan teljesül. Válasszuk $\tilde{M}$-ot úgy, hogy (5.3)-ban és (5.4)-ben M helyett $\tilde{M}$-ot írva $\hat{P}^{(3)} = \sigma^{(3)} / \tilde{M}$ mellett már teljesüljön (5.4). Ezt a következő iteráció segítségével érem el:

$$\tilde{M}_0 = M - M^*, \qquad \hat{P}_{T,0}(i) = \frac{\sigma_T(i)-\sigma(i)}{\tilde{M}_0}, \quad i = 1, 2, \ldots, n,$$

$$\hat{P}_0 = \prod_{i=1}^{n} [1-\hat{P}_{T,0}(i)],$$

$$\tilde{M}_{m+1} = \frac{\sigma^{(3)}}{\hat{P}_m \hat{S}_{3,m}},$$

$$\hat{P}_{T,m+1}(i) = \frac{\sigma^{(4)}(i)}{\sigma^{(4)}(i)+\tilde{M}_{m+1}\hat{P}_m \hat{S}_{3,m}(i)}, \quad i = 1, 2, \ldots, n,$$

$$\hat{P}_{m+1} = \prod_{i=1}^{n} [1-\hat{P}_{T,m+1}(i)], \quad m = 1, 2, \ldots$$

[$\sigma_T(i)$ és $\sigma(i)$ az olyan gyermekek száma, akik rendelkeznek A_i-vel és esetleg további CA-kkal is, ill. akik rendelkeznek A_i-vel, de más CA-val nem]. Ekkor $\tilde{M} = 5078$. $\hat{MP}^{(k)}$ értéke $k = 0, 1, 2, 5, 6, 7$ esetén rendre a következő volt:

1426, 1889, 1163, 25, 4, valamint 1. Mivel a $\underline{k} = 0, 1, 2$ melletti értékek itt nem különösebben érdekesek - hiszen az oligofaktoriális MCA-k általában nagyobb méretüek -, a duális koncepció ránézésre nem elfogadhatatlan. Ahhoz azonban, hogy használható is legyen, feltétlenül ki kell egészíteni a multifaktoriális kóreredetü CA-kra vonatkozó valamilyen feltevéssel. Ezt teszi a feltételes függetlenségi koncepció, amely a függetlenségi koncepciónak és duálisának mintegy "metszete". Eszerint tetszőleges MCA p_1, ill. p_2 valószinüséggel multi- vagy oligofaktoriális kóreredetü, és a CA-k a kóreredet mint feltétel mellett függetlenek. Ez azt jelenti, hogy az MCA-k valószinüségeloszlása $\underline{m} = 2$ számu olyan eloszlás keveréke, amelyek mindegyikében a CA-k függetlenek. Mivel a függetlenségi koncepció (amelyik csak a multifaktoriális CA-k függetlenségét teszi fel) elfogadhatatlan volt, nyilván a feltételes függetlenségi koncepció is az. Viszont természetes módon általánositható: nem teszszük fel, hogy akár a multi-, akár az oligofaktoriális kóreredetü MCA-khoz a keverék egyetlen eloszlása tartozik, tehát azt, hogy $\underline{m} = 2$. Az igy konstruált, Tusnády Gábortól származó keverékeloszlásos modell Telegdi et al. (1981)-ben, Tusnády (1978-1982)-ben és Czeizel et al. (1987)-ben kerül kifejtésre.

5.5. A GAMT-modell kiterjesztése

Mint az 5.3. paragrafusban utaltam rá, az esetek nagy részében a veleszületett rendellenesség multifaktoriális kóreredetü. Az ilyen rendellenességek közül néhány ICOM öröklődésmenetét a GAMT-modellel sikerült leirni (v.ö. 5.1. paragrafus).

Az 5.4. paragrafusban a függetlenségi koncepció néhány módositását vizsgáltam meg. Egy további módositást jelent, ha feltesszük, hogy a különböző CA-kat előidéző genetikai és környezeti tényezők (természetesen ugyanannál a személynél) korreláltak, és a multifaktoriális kóreredetü MCA-k ennek a korrelációnak a következményei. Ez a hipotézis csak a multifaktoriális MCA-król mond valamit. Előfordulhat azonban, hogy néhány ismeretlen, oligofaktoriális MCA ugyanilyen korrelációnak a következménye. Mivel a két tipus megkülönböztethetetlen, ezért a paragrafus további része azon a hipotézisen alapul, amely szerint az egyes CA-k öröklődésmenete a GAMT-modellel leirható, az őket előidéző genetikai és környezeti tényezők korreláltak, és valamennyi MCA ennek a korrelációnak a következménye [lásd Czeizel et al. (1987)]. Mivel az MCA-k jelentős részben multifaktoriális kóreredetüek, ez a hipotézis ésszerü és ránézésre nem elfogadhatatlan.

Jelölje L_i és T_i az A_i rendellenességhez tartozó hajlamot, ill. küszöböt. A küszöb modellnek megfelelően egy

gyermek akkor és csak akkor rendelkezik $\underline{A}_i$-vel, ha $\underline{L}_i$ rajta felvett értéke nem kisebb $\underline{T}_i$-nél. Legyen

$$\underline{\underline{L}} = (L_1, L_2, \ldots, L_n),$$

és tegyük fel, hogy - a normális küszöb modellnek megfelelően - $\underline{\underline{L}}$ $\underline{n}$-dimenziós normális eloszlásu (és az $\underline{L}_i$-k standardak). A küszöböket és az $\underline{\underline{L}}$ eloszlását meghatározó $\underline{r}_{ij}$ korrelációs együtthatókat a 3. fejezetben leirt módon becsültem.

A fenti modell a GAMT-modell kiterjesztése. Az eredeti, az ICCM-ek öröklődésének leirására kidolgozott GAMT-modell egy rendellenesség több családtagra vonatkozó hajlamát irja le. A fenti modellben viszont egy gyermeknek több hajlama van - az $\underline{n}$ számu CA mindegyikéhez egy -, és ezeknek a hajlamoknak a rendszere alakitja ki a gyerek CA-inak rendszerét. A modellt és igy a megfelelő hipotézist a 3. fejezettel összhangban ugy ellenőriztem, hogy megnéztem, mennyire felel meg a 2-nél nagyobb méretü CA-kombinációk előfordulása az $(\underline{r}_{ij})$ korrelációs mátrix alapján vártnak.

Legyen $\underline{\sigma}_T(\underline{i},\underline{j},\underline{k})$ és $\underline{E}_T(\underline{i},\underline{j},\underline{k})$ az olyan gyermekek száma, ill. számának várható értéke, akik rendelkeznek $\underline{A}_i$-vel, $\underline{A}_j$-vel, $\underline{A}_k$-val és esetleg további CA-kkal is. A legnagyobb $\underline{\sigma}_T(\underline{i},\underline{j},\underline{k})$-érték 26 volt, a TC, CD és CF rendellenességek esetén. A megfelelő korrelációs együtthatók becsült értéke 0,71, 0,50 és 0,44 volt, az ezek alapján meghatározott

$\hat{E}_T(i,j,k)$-érték pedig 32,19, ami azt mutatja, hogy ezt a CA-hármast a modell jól irja le. A kisebb $\sigma_T(i,j,k)$-kra rátérve azt találtam, hogy a megfelelő $\hat{E}_T(i,j,k)$-k eloszlása egyre szélesebb. Az eltérések általában nem voltak szignifikánsak, a lehetséges kombinációk igen nagy száma miatt azonban nem az egyes $(\sigma_T, \hat{E}_T)$ értékek, hanem a megfelelő összegek egymáshoz való viszonya érdekes. Ezért meghatároztam a

$$\sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} \sigma_T(i,j,k)$$

és - a 3. fejezetben emlitett eljárás segitségével - a

$$\sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} \hat{E}_T(i,j,k) = M \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} \hat{P}_T(i,j,k)$$

kifejezések értékét; ez 2006, ill. 1876 volt.

$G_g = \{A_i\}$, $i = 1, 2, \ldots, n$, valamint $G_g = \{A_i, A_j\}$, $i, j = 1, 2, \ldots, n$, $i < j$ mellett meghatároztam, ill. becsültem a 3. fejezetben definiált $\sigma_C^{(3)}(G_g)$, $E_C^{(3)}(G_g)$, valamint a modelltől való eltérést jellemző $d(G_g)$ mennyiségek értékét. Ez utóbbi a következő CA-k és CA-párok esetén volt nagy: SB, AM, RA, MC, FN és MS, ill. CP-PS, PY-PS, AA-CF,

AI-OL és BR-IH [mindegyik esetben $\sigma_C^{(3)}(G_g)$ értéke nagyobb volt, mint $\hat{E}_C^{(3)}(G_g)$-é]. Meghatároztam a

$$\sum_{g:|G_g|=2} \sigma_C^{(3)}(G_g), \qquad \sum_{g:|G_g|=2} \hat{E}_C^{(3)}(G_g)$$

kifejezések értékét ($|G_g|$ a G_g rendellenességkombináció mérete); ez 5283, ill. 5627 volt. Ez a kiterjesztett GAMT-modell és a megfelelő hipotézis elfogadhatóságát mutatja.

Tetszőleges $G_g = \{A_i, A_j\}$ esetén legyen $a_{ij} = d(G_g)$ $(1 \leq i < j \leq n)$. Amig az r_{ij} korrelációs együtthatók a CA-k multifaktoriális közelségeinek tekinthetőek, az a_{ij}-k a CA-k oligofaktoriális közelségeinek.

5.6. Egy véletlen és az "igazi" adatmező többszörös többdimenziós skálázása

Az MCA-k statisztikai vizsgálatának fő céljai közé tartozott a rendellenes gyerekek csoportositása a CA-k együttes előfordulása alapján és ezáltal jellegzetes CA-kombinációk felderitése. E feladat megoldására dolgoztam ki a 4.2. paragrafusban részletesen tárgyalt többszörös többdimenziós skálázást. Ennek adekvát voltáról ugy lehet meggyőződni, hogy véletlen módon generált vagy szisztema-

tikusan kijelölt CA-klaszterekhez véletlen adatmezőt generálunk, és megnézzük, ugy csoportositja-e az MMDS a generált gyerekeket, hogy a különböző gyerekklaszterek mellett kirajzolódnak az egyes CA-klaszterek. A véletlen gyerekek generálásához először természetesen specifikálni kell azon (a többszörös veleszületett rendellenességeket leiró) hipotézist és modellt, amely alapján a generálás azután történik. Mivel a függetlenségi és a feltételes függetlenségi koncepció elfogadhatatlannak bizonyult, a kiterjesztett GAMT-modell mellett pedig nem merül fel a W_i változók inkonzisztenciájának MMDS-t igénylő problémája, célszerü a generálást a keverékeloszlásos modell alapján végezni. Legyen a CA-k száma 16, és jelölje i az i-edik CA-t. Tekintsük a következő hat CA-klasztert:

1, 2, 3, 5, 8, 9, 11, 12, 15;
4, 7, 10, 14, 15;
2, 4, 5, 6, 8, 10, 13, 14, 16;
1, 2, 4, 7, 8, 11, 12, 16;
3, 6, 7, 9, 10, 11, 13, 16;
1, 3, 5, 6, 9, 12, 13, 14, 15.

Ezekhez a klaszterekhez egymástól függetlenül 150-150 gyereket generáltam oly módon, hogy az m-edik klaszterhez generált g-edik gyerek egymástól függetlenül 0,3-0,3 valószinüséggel rendelkezik a klaszterhez tartozó, 0,01-0,01 való-

szinüséggel a klaszterhez nem tartozó CA-kkal ($\underline{m}$ = 1, 2, ..., 6; $\underline{g}$ = 1, 2, ..., 150). A kettőnél kevesebb CA-val rendelkező gyerekeket elhagyva 666 számu, 428-féle gyerek maradt. Ezek többszörös többdimenziós skálázását a gyerekklaszterek számának kezdő értékét 15-nek választva végeztem el. Az MMDS hat gyerekklasztert adott, amelyek mellett világosan kirajzolódtak a CA-klaszterek. A kapott gyerekklaszterekben az egyes CA-k gyakorisága a következő volt:

55	45	47	1	37	0	0	43	57	1	58	39	0	0	43	5
0	0	0	29	0	0	29	0	2	19	1	3	0	20	26	0
0	50	1	79	38	47	0	50	1	50	6	7	36	43	0	42
37	37	1	28	1	3	43	32	0	1	38	30	1	0	1	38
3	1	26	0	0	21	47	0	27	41	28	0	43	0	0	29
43	2	45	0	60	60	1	1	55	2	0	48	52	65	58	0

A táblázatból is látható, hogy a gyerekklaszterek "visszaadják" a CA-klasztereket. Hasonlóan kedvező eredményre vezetett a fenti módon generált más véletlen adatmezők többszörös többdimenziós skálázása is.

Az "igazi" adatmező fentiek alapján adekvátnak látszó többszörös többdimenziós skálázását a gyerekklaszterek számának kezdő értékét 11-nek és 21-nek választva is elvégeztem. A kapott, $\underline{C}_m$-mel jelölt gyerekklaszterek melletti CA-képek és -gyakoriságok első esetben a

C_1: {SB, CL, EX, AA, HS, HD, DI, UT, OU, IH};

C_2: {CL, LR, AA, HY, CF, EG, HD, EA, UT};

C_3: {SB, AA, HS, RA, OU}, {HD, CD, UT};

C_4: {CL, SY, MC, HY, AM, CF, HS, HD, CK, FN, EA, IH, MS};

C_5: {HS, UT};

C_6: {AN, EN, CL, LR, PY, SY, EX, CF, HS};

C_7: {CF, TC, CD, IH};

C_8: {CP, CF, HD, CD};

C_9: {AN, EN, PY, CF, HD, CK, RS, OU};

C_{10}: {EN, SB, CP, PY, SY, HY, AM, HD, DI, FN, EA, CD, MS};

C_{11}: {AN, CL, PY, HD, CK, CD, PS, OD};

második esetben a

C_1: {SY, HD, UT, IH};

C_2: {AN, CL, EX};

C_3: {AN, EN, CL, LR, PY, AA, HY, CF, HS, EG, HD, CK};

C_4: {TC, CD, OL};

C_5: {AN, CL, LR, EX, OA, AA, CF, HS, HD, DI, RA, EA, CD, OL, OD, IH, MS};

C_7: {CP, LR, PY, AA, MC, AM, EY, CF, HS, EG, HD, FN, EA, UT, MS, ST};

C_8: {CP, PY, AA, HD, FN, CD, PS, UT, OU};

C_9: {SB, PY, EX, OA, AA, HY, HD, RA, RS, UT, OU}, {SY, CF, HS, PS, OD, IH, MS};

C_{10}: {HD, OU};

C_{11}: {SB, CL, EX, MC, HD, DI, RA, CK, OU, MS, EO};
C_{12}: {EM, SB, CP, PY, SY, AM, CF, HS, HD, DI, FN, EA, UT};
C_{13}: {CL, SY, CF, TC, CD, IH};
C_{14}: {CL, SY, MC, HY, AM, CF, HD, EA};
C_{15}: {AM, SY, HD, RA, FN, EO};
C_{16}: {EM, CL, LR, EX, AA, HY, HD, UT};
C_{17}: {AM, HD, AI, FN}, {MC, CD, IH};
C_{18}: {HD, RS};
C_{19}: {CP, HY, CT, UD, MS};
C_{20}: {CL, PY, HY, EA}, {CF, CD, UT};
C_{21}: {CL, SY, EX, HY, AM, FN, MS},
{PY, MC, CF, EG, HD, RA, EA, TC, CD, OD, UT}

CA-klaszterekre (jellegzetes CA-kombinációkra) vezettek (a klaszterek leggyakoribb CA-it aláhuzással jelölöm; nem minden gyerekklaszter mellett rajzolódott ki CA-klaszter, voltak viszont olyanok, amelyek mellett több is).

A rendellenes gyerekek csoportositását és jellegzetes CA-kombinációk ezáltal történő felderitését keverékfelbontással Tusnády Gábor [lásd Tusnády (1978-1982)], orvosi megfontolások alapján Czeizel Endre [lásd Czeizel et al. (1987)] is elvégezte. A háromféle klaszterezésből a 4.5. paragrafusban leirt módon egy "eredő" klaszterezést állitottam elő. Az eredmények orvosi értékelése, ellenőrzése (genetikai családvizsgálatok) és felhasználása (genetikai tanácsadás) Czeizel et al. (1987)-ben kerül kifejtésre. - A

rendellenességek Boole-faktoranalizisét Rejtő Lidia, log-lineáris elemzését Rudas Tamás végezte el, korrespondancia-analizisükkel pedig Dávidné Bolla Marianna foglalkozik.

IRODALOM

1. Abaffy J. (1976) A duális mátrixok módszerének egy osztályáról. Alk. Mat. Lapok, 2, 351-358.
2. Aho, A. V., Hopcroft, J. E., Ullman, J. D. (1982) Számitógép-algoritmusok tervezése és analizise. Müszaki Könyvkiadó, Budapest.
3. Anderberg, M. R. (1973) Cluster Analysis for Applications. Academic Press, New York.
4. Anderson, T. W. (1958) An Introduction to Multivariate Statistical Analysis. Wiley, New York.
5. Arany I. (1984) Ritka mátrixu lineáris egyenletrendszerek hatékony számitógépes kezelése. Kandidátusi értekezés. Budapest.
6. Bernau H. (1977) Felső korlát technikák a kvadratikus programozáshoz. Alk. Mat. Lapok, 3, 161-170.
7. Bolla M., Czeizel E., Telegdi L., Tusnády G. (1979) Többszörös veleszületett rendellenességek statisztikai vizsgálata. Muszka D., Madarász I., Székely S., szerk.: Számitástechnikai és kibernetikai módszerek alkalmazása az orvostudományban és a biológiában, 9. Kollokvium Közleményei, Neumann János Számitógéptudományi Társaság, Szeged, 154-165.
8. Bolla M., Tusnády G. (1982) Hipergráfok euklideszi térbe való beágyazása veleszületett rendellenességek

clusterezéséhez. Győri I., Csirik J., Eller J., Madarász I., szerk.: Számitástechnikai és kibernetikai módszerek alkalmazása az orvostudományban és a biológiában, 11. Kollokvium Közleményei, Neumann János Számitógéptudományi Társaság, Szeged, 169-173.

9. Carter, C. O. (1961) The inheritance of congenital pyloric stenosis. Br. Med. Bull., 17, 251-267.

10. Chinn, P. Z., Chvátalová, J., Dewdney, A. K., Gibbs, N. E. (1982) The bandwidth problem for graphs and matrices - A survey. J. of Graph Theory, 6, 223-254.

11. Cohen, M. M., Jr. (1977) On the nature of syndrome delineation. Acta Genet. Med. Gemellol., 103, 103-119.

12. Cramér, H. (1946) Mathematical Methods of Statistics. Princeton University Press.

13. Czeizel, A. (1981) The definition of multiple congenital abnormalities. Acta Morph. Acad. Sci. Hung., 29, 251-258.

14. Czeizel E. (1984) Az emberi öröklődés. Gondolat, Budapest.

15. Czeizel, A., Pázsy, A., Telegdi, L., Tusnády, G. (1981) Classification and registration of multiple congenital abnormalities. Acta Morph. Acad. Sci. Hung., 29, 377-390.

16. Czeizel, A., Telegdi, L., Tusnády, G. (1987) Multiple Congenital Abnormalities. Akadémiai Kiadó, Budapest (elfogadva).

17. Czeizel, A., Tusnády, G. (1984) Aetiological Studies of Isolated Common Congenital Abnormalities in Hungary. Akadémiai Kiadó, Budapest.

18. Deák I. (1980) Monte-Carlo módszerek a többdimenziós térben elhelyezkedő halmazok valószínüségeinek meghatározására normális eloszlás esetén. Kandidátusi értekezés. Budapest.

19. De Leeuw, J., Heiser, W. (1982) Theory of multidimensional scaling. [34], 285-316.

20. Everitt, B. (1980) Cluster Analysis, 2nd Ed. Heinemann Educational Books, London.

21. Falconer, D. S. (1965) The inheritance of liability to certain diseases, estimated from the incidence among relatives. Ann. Hum. Genet., 29, 51-67.

22. Fischer J., Telegdi L., Csukás B. (1974) A növekedési görbéken alapuló sztochasztikus predikció egy uj módszere. Mérés és Automatika, 22, 212-215.

23. Futó I. (1978) Hipergráf modellen alapuló klaszterelemzés. Tanulmány. Építéstudományi Intézet, Budapest.

24. Füstös L. (1981) A sokdimenziós skálázás módszerei. Módszertani Füzetek, MTA Szociológiai Kutatóintézet, Budapest.

25. George, J. A., Liu, J. W-H. (1981) Computer Solution of

Large Sparse Positive Definite Systems. Prentice Hall, Englewood Cliffs, New Jersey.

26. Gordon, A. D. (1981) Classification. Chapman and Hall, London.

27. Gower, J. (1984) Multivariate Analysis: Ordination, Multidimensional Scaling and Allied Topics. Lloyd, E., ed.: Handbook of Applicable Mathematics, Volume VI: Statistics, Part B, Wiley-Interscience, Chicester, 727-781.

28. Gulyás O. (1983) Osztályozási eljárások. [40], I., 7/1-24.

29. Hartigan, J., A. (1975) Clustering Algorithms. Wiley, New York.

30. Jambu, M., Lebeaux, M-O. (1983) Cluster Analysis and Data Analysis. North-Holland, Amsterdam.

31. Jardine, N., Sibson, R. (1971) Mathematical Taxonomy. Wiley, New York.

32. Källen, B., Winberg, J. (1969) Multiple malformations studied with a national register of malformations. Pediatrics, 44, 410-417.

33. Knuth, D. E. (1973) Sorting and Searching. Addison-Wesley, Reading, Massachusets.

34. Krishnaiah, P. R., Kanal, L. N., eds. (1982) Handbook of Statistics, Volume 2: Classification, Pattern Recognition and Reduction of Dimensionality. North-Holland, Amsterdam.

35. Kruskal, J. B. (1977a) Multidimensional scaling and other methods for discovering structure. Enslein, K., Ralston, A., Wilf, H. S., eds.: Mathematical Methods for Digital Computers, Volume III: Statistical Methods for Digital Computers, Wiley-Interscience, New York, 296-339.

36. Kruskal, J. B. (1977b) The relationship between multidimensional scaling and clustering. [63], 17-44.

37. Kruskal, J. B., Wish, M. (1978) Multidimensional Scaling. Sage Publications, Beverly Hills.

38. Курош, А. Г. (1956) Курс Высшей Алгебры. Гостехиздат, Москва.

39. Mardia, K. V., Kent, J. T., Bibby, J. M. (1979) Multivariate Analysis. Academic Press, New York.

40. Rejtő L., szerk. (1983-1984) Többváltozós statisztikai módszerek. Jegyzet. Bolyai János Matematikai Társulat, Budapest.

41. Roberts, C. J., Powell, R. G. (1975) Interrelation of the common congenital malformations. Lancet, 1975/ii, 848-854.

42. Romney, A. K., Shepard, R. N., Nerlove, S. B., eds. (1972) Multidimensional Scaling, Volume II. Seminar Press, New York.

43. Schiffman, S. S., Reynolds, M. L., Young, F. W. (1981) Introduction to Multidimensional Scaling. Academic Press, New York.

44. Schuler D., Fekete Gy., Krause I., Fischer J., Bene B., Telegdi L. (1974) A leukémia és a rosszindulatu daganatos betegségek gyakorisága különböző betegcsoportok rokonságában. Anthrop. Közl., 18, 167-173.

45. Serfling, R. J. (1980) Approximation Theorems of Mathematical Statistics. Wiley, New York.

46. Shepard, R. N., Romney, A. K., Nerlove, S. B., eds. (1972) Multidimensional Scaling, Volume I. Seminar Press, New York.

47. Simonovits, M., Telegdi, L., Tusnády, G. (1982) Classification of objects via multidimensional scaling of variables. Caussinus, H., Ettinger, P., Mathieu, J. R., eds.: COMPSTAT 1982, Proceedings in Computational Statistics, Part II, Physica-Verlag, Wien, 243-244.

48. Späth, H. (1980) Cluster Analysis Algorithms for Data Reduction and Classification of Objects. Ellis Horwood, Chicester.

49. Sváb J. (1971) A populációgenctika alapjai. Mezőgazdasági Kiadó, Budapest.

50. Taqqu, M. S. (1977) Law of the iterated logarithm for sums of nonlinear functions of Gaussian variables that exhibit a long range dependence. Z. Wahrscheinlichkeitsth. und verw. Gebiete, 40, 203-238.

51. Telegdi, L. (1982) Multiple multidimensional scaling:

a new approach to the analysis of multidimensional contingency tables with application to congenital abnormalities. Metron, 40, 277-288.

52. Telegdi L. (1983) Veleszületett rendellenességek függetlenségének vizsgálata. Alk. Mat. Lapok, 9, 421-436.

53. Telegdi L. (1984) Többdimenziós skálázás. [40], II., 8/1-58.

54. Telegdi L. (1986) Többdimenziós skálázás. Móri F. T., Székely J. G., szerk.: Többváltozós statisztikai analizis, Műszaki Könyvkiadó, Budapest, 233-250.

55. Telegdi, L., Czeizel, A., Tusnády, G., Bolla, M., Pázsy, A. (1981) Statistical Study on the Multiple Congenital Abnormalities in Hungary, 1970-1976. Working Paper, MS/6. MTA SZTAKI, Budapest.

56. Telegdi, L., Simonovits, M. (1983) Initialization of multiple multidimensional scaling. Horvitz, D. G., Sánchez-Crespo, J. L., eds.: Contributed Papers of the 44th Session of the International Statistical Institute, Volume 2, International Statistical Institute, Madrid, 575-578.

57. Torgerson, W. S. (1958) Theory and Methods of Scaling. Wiley, New York.

58. Tusnády G. (1969) A multifaktoriális öröklődés. Mat. Lapok, 20, 389-396.

59. Tusnády G. (1978-1982) Keverékek felbontása. Mat. La-

pok, 30, 59-67.

60. Tusnády, G., Czeizel, A., Telegdi, L. (1981) ML-fitting of multifactorial threshold models. Periodica Math. Hung., 12, 205-216.

61. Tusnády, G., Csiszár, A., Telegdi, L., Czeizel, E., Bolla, M. (1978a) Statistical investigation of multiple congenital malformations. Kozesnik, J., ed.: Transactions of the Eight Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Volume B, Academia, Prague, 301-308.

62. Tusnády G., Telegdi L., Czeizel E. (1978b) Gyakori veleszületett rendellenességek öröklődésmenetének vizsgálata. Alk. Mat. Lapok, 4, 1-35.

63. Van Ryzin, J., ed. (1977) Classification and Clustering. Academic Press, New York.

64. Vincze I. (1968) Matematikai statisztika ipari alkalmazásokkal. Müszaki Könyvkiadó, Budapest.

65. Wish, M., Carroll, J. D. (1982) Multidimensional scaling and its applications. [34], 317-345.

1986-BAN MEGJELENTEK:

179/1986 Terlaky Tamás: Egy véges criss-cross módszer és alkalmazásai

180/1986 K.N. Čimev: Separable sets of arguments of functions

181/1986 Renner Gábor: Kör approximációja a számitógépes geometriai tervezésben

182/1986 Proceedings of the Joint Bulgarian-Hungarian Workshop on "Mathematical Cybernetics and Data Processing" Scientific Station of Sofia University, Giulecica /Bulgaria/, May 6-10, 1985 /Editors: J. Denev, B. Uhrin/ Vol I

183/1986 Proceedings of the Joint Bulgarian-Hungarian Workshop on "Mathematical Cybernetics and Data Processing" Scientific Station of Sofia University, Giulečica /Bulgaria/, May 6-10, 1985 /Editors: J. Denev, B. Uhrin/ Vol II

184/1986 HO THUAN: Contribution to the theory of relational databases

185/1986 Proceedings of the 4th International Meeting of Young Computer Scientists IMICS'86 /Smolenice, 1986/ /Editors: J. Demetrovics, J. Kelemen/

186/1986 PUBLIKÁCIÓK - PUBLICATIONS 1985
Szerkesztette: Petróczy Judit

187/1986 Proceedings of the Winter School on Conceptual modelling /Visegrád, 27-30 January, 1986/ /Editors: E. Knuth, A. Márkus/

188/1986 Lengyel Tamás: A Cluster analizis néhány kombinatorikai és valószinüségszámitási problémája

189/1986 Bernus Péter: Gyártórendszerek funkcionális analizise és szintézise

190/1986 Hernádi Ágnes: A tipus fogalma, és szerepe a modellezésben

191/1986 VU DUC THI: Funkcionális függőséggel kapcsolatos néhány kombinatorikai jellegü vizsgálat a relációs adatmodellben

192/1986 Márkusz Zsuzsanna: P a p e r s on Many-stored logic as a tool for modelling

193/1986 KNVVT Conference on Automation of Information Processing on Personal Computers
Budapest, May 5-9, 1986 Vol I.
Szerkesztette: Ratkó István

194/1986 KNVVT Conference on Automation of Information Processing on Personal Computers
Budapest, May 5-9, 1986 Vol II.
Szerkesztette: Ratkó István