Alkalmazott matematikai lapok

1987–88/1–2

A MAGYAR TUDOMÁNYOS AKADÉMIA
MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK
OSZTÁLYÁNAK KÖZLEMÉNYEI

13.

KÖTET

AKADÉMIAI KIADÓ, BUDAPEST

ALKALMAZOTT MATEMATIKAI LAPOK

A MAGYAR TUDOMÁNYOS AKADÉMIA MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK OSZTÁLYÁNAK KÖZLEMÉNYEI

XIII. kötet 1—2. szám

Szerkesztőség: 1502 Budapest XII., Kende u. 13—17.
Kiadóhivatal: 1055 Budapest V., Alkotmány u. 21.

Az Alkalmazott Matematikai Lapok változó terjedelmű füzetekben jelenik meg, és olyan eredeti tudományos cikkeket publikál, amelyek a gyakorlatban, vagy más tudományokban közvetlenül felhasználható új matematikai eredményt tartalmaznak, illetve már ismert, de színvonalas matematikai apparátus újszerű és jelentős alkalmazását mutatják be. A folyóirat közül cikk formájában megírt, új tudományos eredménynek számító programokat, és olyan, külföldi folyóiratban már publikált dolgozatokat, amelyek magyar nyelven történő megjelentetése elősegítheti az elért eredmények minél előbbi, széles körű hazai felhasználását.

A folyóirat feladata a Magyar Tudományos Akadémia III. (Matematikai és Fizikai) Osztályának munkájára vonatkozó közlemények, könyvismertetések stb. publikálása is.

A kéziratok a főszerkesztőhöz, vagy a szerkesztő bizottság bármely tagjához beküldhetők. A főszerkesztő címe:

Prékopa András, főszerkesztő
1502 Budapest, Kende u. 13—17.

Közlésre el nem fogadott kéziratokat a szerkesztőség lehetőleg visszajuttat a szerzőhöz, de a beküldött kéziratok megőrzéséért vagy továbbításáért felelősséget nem vállal.

Az Alkalmazott Matematikai Lapok előfizetési ára kötetenként 128 forint. Belföldi megrendelések az Akadémiai Kiadó, 1055 Budapest V., Alkotmány u. 21. címen (pénzforgalmi jelzőszám 215—11 488), külföldi megrendelések a Kultúra Külkereskedelmi Vállalat, H-1389 Budapest, Pf. 149. címen (pénzforgalmi jelzőszám 218—10 990) lehetségesek.

A Magyar Tudományos Akadémia III. (Matematikai és Fizikai) Osztálya a következő idegen nyelvű folyóiratokat adja ki:

1. Acta Mathematica Hungaricae,
2. Acta Physica Hungaricae,
3. Studia Scientiarum Mathematicarum Hungarica.

Kedves Olvasóink!

Folyóiratunk 13. kötete 1987--88-as évszámmal összevontan jelenik meg. A Kiadó így tudja az MTA elnökségének határozatát végrehajtani, s az elmaradást 1989-re megszüntetni. A 14. kötettől változatlanul jelenik meg a folyóirat. Az 1987-re befizetett előfizetési díj 1988-ra is érvényes.

A gazdasági és technikai nehézségekre való tekintettel kérjük megértésüket és türelmüket.

A szerkesztőség

Alkalmazott Matematikai Lapok **13** (1987—88) 1—7

A BLAND SZABÁLY A PRIMÁL ÉS A DUÁL SZIMPLEX MÓDSZER ESETÉN

KLAFSZKY EMIL—TERLAKY TAMÁS

Budapest

Cikkünkben megmutatjuk, hogy a BLAND által [1]-ben megfogalmazott pivotálási szabály nem csak a primál, hanem a duál szimplex módszer esetén is biztosítja az algoritmus végességét. A teljességre törekvés, valamint az analógia megmutatása céljából megismételjük a primál szimplex módszerre adott BLAND bizonyítást is.

1. Bevezetés

R. G. BLAND [1] cikkében új pivotálási szabályt fogalmazott meg a szimplex módszerhez, amely alkalmazásával bizonyítani lehet a szimplex algoritmus végességét. Dolgozatunkban a továbbiakban az alábbi lineáris programozási feladattal foglalkozunk:

$$\begin{aligned} &\min cx \\ &Ax = b \\ &x \geqq 0 \end{aligned} \tag{1.1}$$

Ahol $x, c \in R^n$, $b \in R^m$ és A: $m \times n$ teljes sorrangú mátrix.

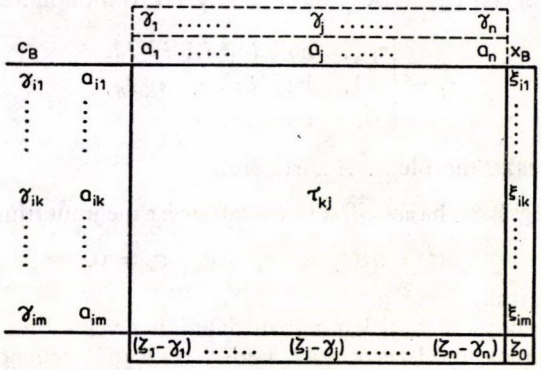
A továbbiakban a mátrixokat nagy, a vektorokat kis latin betűkkel jelöljük, a vektorok komponenseit pedig görög betűkkel. Így pl. $x=(\xi_1, \ldots, \xi_n)$ és τ_{rs} a szimplex tábla r-edik sorának s-edik eleme.

Feltételezzük, hogy az olvasó ismeri a primál és a duál szimplex módszert, de az egységes és teljes tárgyalásmód érdekében a primál és duál szimplex módszer legfontosabb tulajdonságait a következő fejezetben röviden összefoglaljuk.

2. A primál és a duál szimplex módszer alaptulajdonságai

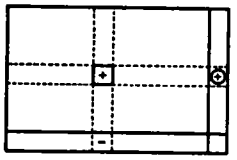
a) A primál szimplex módszer

A szimplex tábla adatait az alábbi ábra szemlélteti:



1. ábra

Azt a szituációt, amikor éppen pivotálunk, az alábbi sémán szemléltetjük:



2. ábra

ahol a „$-$" szimbólum negatív számot; a „$+$" pozitívat és a „$\oplus$" nem-negatívat szimbolizál.

Legyen B az aktuális bázis vektorok mátrixa és I_B a bázisindexek halmaza. A primál szimplex módszer jellemzői a következők:

2.1. *Megjegyzés.* Ha a_s bevonandó a bázisba, akkor $\zeta_s - \gamma_s > 0$.

2.2. *Megjegyzés.* Ha τ_{rs} pivot elem, akkor $\tau_{rs} > 0$.

2.3. *Megjegyzés.* Tetszőleges B bázis esetén $\zeta_i - \gamma_i = 0$, $i \in I_B$.

2.4. *Megjegyzés.* Tetszőleges B bázis esetén $Bt_s = a_s$, ahol t_s a szimplex tábla s-edik oszlopa, mivel $t_s = B^{-1}a_s$.

2.5. *Megjegyzés.* Tetszőleges pivottranszformáció esetén

$$\zeta_0^{\text{új}} = \zeta_0^{\text{régi}} - \frac{\xi_r}{\tau_{rs}}(\zeta_s - \gamma_s).$$

Amennyiben $\zeta_0^{uj} = \zeta_0^{\text{régi}}$, ami ciklizálás esetén mindig fennáll, akkor $\xi_r = 0$ kell hogy legyen, mivel $\zeta_s - \gamma_s > 0$ a 2.2. megjegyzés miatt.

2.6. LEMMA. Tetszőleges B bázis esetén a szimplex tábla bármely sora, valamint az alsó sorban szereplő z vektor lin A-ban van, ahol lin A az A sorvektorai által generált altér.

Bizonyítás. Ismert, hogy $z = c_B B^{-1}A$, így $z \in \text{lin } A$, valamint a 2.4. megjegyzés miatt a táblázat r-edik sora $e_r \cdot B^{-1}A$ alakban írható, (ahol e_r az m dimenziós r-edik egységvektor) ami bizonyítja állításunkat.

2.7. LEMMA. Legyen egy $t \in R^n$ vektor a következő módon adott:

$$\tau_i = \begin{cases} \tau_{i_j s}, & \text{ha} \quad i = i_j,\ i_j \in I_B, \\ -1, & \text{ha} \quad i = s,\ s \notin I_B, \\ 0, & \text{máskor.} \end{cases}$$

Ekkor $t \perp \text{lin } A$, azaz t merőleges A sorterére.

Bizonyítás. Elégséges, ha az $At = 0$ összefüggést megmutatjuk:

$$At = BB^{-1}a_s - a_s = a_s - a_s = 0.$$

Így lemmánkat beláttuk.

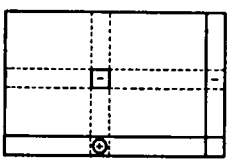
Megjegyezzük, hogy a fenti lemmában definiált t vektor megegyezik a lexikografikus duál szimplex módszerben az s-edik vektornál fellépő oszlopvektorral. Így a fenti két lemmából azonnal adódik a következő 2.8. tétel:

2.8. TÉTEL. A primál szimplex tábla tetszőleges sora (beleértve a z vektort is) merőleges a lexikografikus duál szimplex módszer táblázatának tetszőleges oszlopára.

Ezután rátérünk a duál szimplex módszer tulajdonságainak ismertetésére.

b) A duál szimplex módszer

Azt a szituációt, amikor éppen pivotálunk, az alábbi sémán szemléltetjük:



3. ábra

2.9. *Megjegyzés.* Ha a_r a bázisból távozó vektor, akkor $\xi_r<0$.

2.10. *Megjegyzés.* Ha τ_{rs} a pivot elem, akkor $\tau_{rs}<0$.

2.11. *Megjegyzés.* Tetszőleges B bázis esetén $\tau_{rj}=0$ $j\in I_B$ esetén, kivéve $\tau_{rr}=1$, ahol $1\leqq r\leqq m$.

2.12. *Megjegyzés.* Tetszőleges x bázismegoldás esetén $\xi_i=0$ $i\notin I_B$ esetén.

2.13. *Megjegyzés.* Tetszőleges pivot transzformáció esetén

$$\zeta_0^{\text{új}} = \zeta_0^{\text{régi}} - \frac{\zeta_s-\gamma_s}{\tau_{rs}}\,\xi_r$$

Amennyiben $\zeta_0^{uj}=\zeta_0^{\text{régi}}$, ami ciklizálás esetén minden lépésben fennáll, akkor $\zeta_s-\gamma_s=0$ kell hogy legyen, mivel $\xi_r<0$ a 2.9. megjegyzés miatt.

2.14. LEMMA. Tetszőleges két x^1 és x^2 megoldás különbsége merőleges lin A-ra.

Bizonyítás. Mivel x^1 és x^2 is megoldás, így

$$A(x^1-x^2) = Ax^1-Ax^2 = b-b = 0.$$

Így lemmánkat beláttuk.

2.15. *Megjegyzés.* Fenti lemmánk azt is biztosítja a 2.6. lemmával együtt, hogy két megoldás különbsége merőleges a szimplex tábla tetszőleges sorára.

KÖVETKEZMÉNY. Ha x^1 és x^2 bázismegoldások, akkor

$$\xi_r^1 = \xi_r^2 + \sum_{k\notin I_{B^1}} \tau_{rk}\,\xi_k^2,$$

ahol τ_{rk} az x^1-nek megfelelő szimplex tábla r-edik sorának k-adik eleme.

Bizonyítás. Legyen t^r a B^1 bázisnak megfelelő szimplex tábla r-edik sora. Ekkor előző megjegyzésünk szerint $t^r(x^2-x^1)=0$. A bázis tulajdonságot felhasz-

nálva

$$t^r(x^2-x^1)=\sum_{k\notin I_{B^1}}\tau_{rk}(\xi_k^2-\xi_k^1)+\tau_{rr}(\xi_r^2-\xi_r^1)=\sum_{k\notin I_{B^1}}\tau_{rk}\xi_k^2+\xi_r^2-\xi_r^1=0,$$

azaz

$$\xi_r^1=\xi_r^2+\sum_{k\notin I_{B^1}}\tau_{rk}\xi_k^2.$$

3. A Bland szabály a primál és a duál szimplex módszerhez

Ebben a fejezetben megfogalmazzuk a *Bland szabályt* a primál és a duál szimplex módszer esetében is, valamint a szabály alkalmazásának néhány közvetlen következményét. A pivotálási szabály a primál szimplex módszer esetében a következő:

P-szabály. a) a lehetséges bevonandó vektorok közül a legalacsonyabb indexűt vonjuk be a bázisba; b) a lehetséges távozó vektorok közül a legalacsonyabb indexű vektor távozik a bázisból.

Az alábbiak nyilvánvaló következményei a *P-szabálynak.*

3.1. *Megjegyzés.* Ha a_q-t vonjuk be a bázisba, akkor $\zeta_k-\gamma_k\leqq 0$, $k<q$.

3.2. *Megjegyzés.* Ha $\xi_q=0$ és a_q távozik a bázisból, valamint a_s jön be a bázisba, akkor $\tau_{ks}\leqq 0$ minden $k<q$-ra amikor $\xi_k=0$.

A 4. fejezetben bizonyítjuk az alábbi fontos tételt.

3.3. TÉTEL. A primál szimplex módszer véges lépésben befejeződik, ha a *P-szabályt* alkalmazzuk a pivot elem kiválasztásakor.

Most a duál szimplex módszerre vonatkozó pivotálási szabályt, majd annak következményeit fogalmazzuk meg.

D-szabály. a) a lehetséges távozó vektorok közül a legalacsonyabb indexű vektor távozik a bázisból, *b)* a lehetséges bevonandó vektorok közül a legalacsonyabb indexűt vonjuk be a bázisba.

3.4. *Megjegyzés.* Ha a_q vektor távozik a bázisból, akkor $\xi_k\geqq 0$, $k<q$.

3.5. *Megjegyzés.* Ha $\zeta_q-\gamma_q=0$ és a_q a bázisba bevonandó vektor és a_r távozik a bázisból, akkor $\tau_{rj}\geqq 0$, $j<q$ esetén, amikor $\zeta_j-\gamma_j=0$.

A 3. tétellel analóg állítás igaz a duál szimplex módszer esetében is. Bizonyítását szintén a 4. fejezetben végezzük el.

3.6. TÉTEL. A duál szimplex módszer véges lépésben végetér, ha a *D-szabályt* alkalmazzuk a pivot elem kiválasztásakor.

4. A primál és a duál szimplex módszer végességének bizonyítása

Először 3.3. tételt bizonyítjuk be a szimplex módszer és a pivotálási szabály tulajdonságaira támaszkodva.

A 3.3. *tétel bizonyítása:*

Tegyük fel indirekt, hogy egy adott megengedett bázisból indulva a szimplex eljárás nem ér véget, azaz ciklizál, mivel a lehetséges bázisok száma véges. Mivel a P-szabály egyértelműen meghatározza a bázisba bejövő és onnan távozó vektort, így a kör egyértelműen meghatározott. Legyen I^* azon vektorok indexeinek halmaza, melyek a ciklizálás során kikerülnek a bázisból. (Természetesen ezek ugyanazok, mint amelyek a kör során valamikor bekerülnek a bázisba. Megjegyezzük, hogy ha $s \notin I^*$, akkor a_s vagy mindig a bázisban volt, vagy soha sem volt a bázisban a kör során. Jelöljük q-val a maximális indexet I^*-ban. A kör során a_q valamikor bejön, valamikor meg távozik a bázisból. Legyen B' az a bázis, amikor a_q bejön a bázisba, B'' pedig az a bázis, amikor távozik a_q a bázisból. Ekkor B'-re és B''-re is fennállnak a 2.1.—2.8.-ban leírt tulajdonságok, valamint B'-re a 3.1. és B''-re a 3.2-ben leírtak. Tekintsük a B'-nek megfelelő z' és a B''-nek megfelelő t'' vektort, ahol z'-t a 2.6. lemma, t''-t a 2.7. lemmának megfelelően készítettük el. (Legyen s_s a bázisba bejövő vektor az utóbbi esetben.) Ekkor $z' \perp t''$, azaz $z' \cdot t'' = 0$.

$$z' t'' = \sum_{k \in I_{B''}} \tau''_{ks} \zeta'_k - \zeta_s = 0. \tag{4.1}$$

A 2.1. megjegyzés miatt $\zeta''_s - \gamma_s > 0$, azaz $\sum_{k \in I_{B''}} \tau''_{ks} \gamma_k - \gamma_s > 0$, amit (4.1)-ből kivonva kapjuk, hogy

$$\sum_{k \in I_{B''}} \tau''_{ks} (\zeta'_k - \gamma_k) - (\zeta'_s - \gamma_s) < 0. \tag{4.2}$$

Nyilván $s \in I^*$, így $s < q$ így a 3.1. megjegyzés miatt $\zeta'_s - \gamma_s \leqq 0$, valamint $\tau''_{qs} > 0$ és $\zeta'_q - \gamma_q > 0$ a 2.2. és 2.1. megjegyzés miatt. Továbbá a 2.3. megjegyzés miatt $\zeta'_k - \gamma_k = 0$, $k \in I_B$, így (4.2)-ből nyerjük, hogy:

$$\sum_{k \in I_{B''} \setminus I_{B'} \setminus q} \tau''_{ks} (\zeta'_k - \gamma_k) < 0. \tag{4.3}$$

Mivel $I_{B''} \setminus I_{B'} \setminus q \subset \{k | k \in I^*,\ k < q\}$ így a 3.1. megjegyzés miatt $\zeta'_k - \gamma_k \leqq 0$, és a 3.2. megjegyzés miatt $\tau''_{ks} \leqq 0$, azaz (4.3) bal oldala nemnegatív, így ellentmondásra jutottunk, indirekt feltevésünk nem igaz. Tételünket tehát bebizonyítottuk.

Most rátérünk 3.6. tétel bizonyítására.

A 3.6. *tétel bizonyítása:*

Bizonyításunk hasonló módon indirekt, mint a 3.3. tétel bizonyítása. Tegyük fel, hogy a duál szimplex módszer a *D-szabály* alkalmazása mellett ciklizál. Legyen I^* ismét a kör során a bázisból kilépő indexek halmaza, és legyen q az I^*-ban levő indexek közül a maximális.

Legyen B' az a bázis, amikor a_q bejön a bázisba, és B'' az a bázis, amikor a_q távozik a bázisból. Ekkor B' és B'' bázisokra természetesen fennállnak a 2.9.—2.15-ben leírt tulajdonságok, továbbá B'-re a 3.5., B''-re a 3.4. megjegyzésben leírtak.

Tekintsük a B'-bázisnak megfelelő $t^{r'}$ vektort, ahol a_r a bázisból távozó vektor és az x', x'' megoldásvektorokat B' és B''-nek megfelelően. Ekkor a 2.15. megjegyzés miatt $t^{r'}(x''-x')=0$, azaz a 2.11. megjegyzés miatt:

$$\sum_{k \notin I_{B'}} \tau'_{rk}(\xi''_k - \xi'_k) + \tau'_{rr}(\xi''_r - \xi'_r) = 0. \tag{4.4}$$

A 2.11. megjegyzés miatt $\tau'_{rr}=1$ és a 3.4. megjegyzés miatt $\xi''_r \geqq 0$, a 2.9. megjegyzés miatt $\xi'_r<0$, így (4.4)-ből nyerjük, hogy

$$\sum_{k \notin I_{B'}} \tau'_{rk}(\xi''_k - \zeta'_k) < 0. \tag{4.5}$$

Mivel $\tau'_{rq}<0$, $\xi''_q<0$ és $\xi'_q=0$, valamint a 2.12. megjegyzés miatt $\xi'_k=0$, ha $k \notin I_{B'}$ és $\xi''_k=0$, ha $k \notin I_{B''}$, így (4.5)-ből

$$\sum_{k \in I_{B''} \setminus I_{B'} \setminus q} \tau'_{rk}\xi''_k < 0. \tag{4.6}$$

Mivel $I_{B''} \setminus I_{B'} \setminus q \subset \{k | k \in I^*, k<q\}$ így a 3.5. megjegyzés miatt $\tau'_{rk} \geqq 0$ és a 3.4. megjegyzés miatt $\xi''_k \geqq 0$, azaz a (4.6) egyenlőtlenség bal oldala nemnegatív, így ellentmondásra jutottunk, indirekt feltevésünk nem igaz, tehát tételünket bebizonyítottuk, a *D-szabály* alkalmazásával a duál szimplex módszer nem ciklizálhat.

5. Összefoglaló megjegyzések

Dolgozatunkban a *Bland szabállyal* módosított primál és duál szimplex módszer végességét bizonyítottuk. Bizonyításunk közvetlenül a szimplex táblában található adatokra támaszkodik, nem igényli a szimplex tábla bővítését, ahogy az BLAND [1] dolgozatában található. Az általunk adott bizonyítások oktatási célra is közvetlenül felhasználhatók, mivel bizonyításaink csak a primál és a duál szimplex módszer jól ismert tulajdonságaira, valamint a *P-* és *D-szabály* alkalmazásának elemi következményeire támaszkodnak.

A duál szimplex módszer végességét bizonyíthattuk volna úgy is, hogy a duál szimplex módszer nem más, mint a duál feladatra alkalmazott primál szimplex módszer, megszabadítva a nem lényeges alkotóelemektől. Ebből közvetlenül következne a duál szimplex módszer végessége is. Ez azonban kevéssé ismert, így indokoltnak tartottuk a duál szimplex módszer végességének rövid, direkt bizonyítását.

IRODALOM

[1] BLAND, R. G., "A new pivoting rule for the simplex method", *Mathematics of Operations Research* **2** (1977) 102—108.
[2] ROCKEFELLAR, R. T., "The elementary vectors of a subspace of R^n", in: *Combinatorial Mathematics and its Applications* ed. R. G. Bore and T. A. Dowling, University of North-Carolina, Chapel Hill, 104—127.

(Beérkezett: 1983. október 14.)

KLAFSZKY EMIL
NEHÉZIPARI MŰSZAKI EGYETEM MATEMATIKAI INTÉZET
3515 MISKOLC, EGYETEMVÁROS

TERLAKY TAMÁS
ELTE TTK OPERÁCIÓKUTATÁSI TANSZÉK
1088 BUDAPEST, MÚZEUM KRT. 6—8.

THE "BLAND" PIVOTING RULE FOR THE PRIMAL AND DUAL SIMPLEX METHOD

E. KLAFSZKY and T. TERLAKY

In our paper we show that the pivoting rule — formulated by R. G. BLAND in [1] — garantees the finiteness both of the primal and dual simplex method. To show the analogy we repeat the proof of Bland for the primal simplex method.

Alkalmazott Matematikai Lapok 13 (1987—88) 9—15

A FELÜLETI DISZJUNKTÍV PROGRAMOZÁSI FELADATRÓL

FÜLÖP JÁNOS

Budapest

A dolgozatban megmutatjuk, hogy tetszőleges lineáris diszjunktív programozási feladat egy ekvivalens felületi diszjunktív programozási feladat alakjára transzformálható. Véges metszősík módszert adunk a felületi feladat megoldására. Ennél a módszernél, az eddig ismertektől eltérően, a megengedett pontok halmazának korlátossága nincs megkövetelve.

1. Bevezetés

Tekintsük a lineáris diszjunktív programozási feladatot *konjunktív normál alakjában* megadva [1, 3, 9]:

$$\min \mathbf{c}'\mathbf{x}$$

$$\mathbf{x}\in X = \{\mathbf{x}|\mathbf{D}\mathbf{x} \geqq \mathbf{d}, \mathbf{x} \geqq \mathbf{0}\}, \tag{1.1}$$

$$\mathbf{x}\in Y = \bigcap_{h\in H} \bigcup_{i\in Q_h} \{\mathbf{x}|\mathbf{a}_i^{h\prime}\mathbf{x} \geqq b_i^h\},$$

ahol $\mathbf{D}$ $m\times n$-es mátrix, $\mathbf{c}$ és $\mathbf{a}_i^h$ n-dimenziós vektorok, $\mathbf{d}$ m-dimenziós vektor, b_i^h skalár és $\mathbf{x}$ a változók n-dimenziós vektora. A H és Q_h, $h\in H$ indexhalmazok végesek. Feltesszük, hogy $X\neq\emptyset$, de sem X, sem Y korlátossága nincs kikötve. Megmutatható, hogy (1.1) alakjára transzformálható tetszőleges olyan matematikai programozási feladat, ahol a célfüggvény lineáris és a megengedett pontok halmaza véges sok halmazelméleti egyesítés és metszet képzési művelet véges sok zárt féltérre való alkalmazásával jön létre [3, 9]. Számos jól ismert matematikai programozási feladat írható fel (1.1) alakjában, pl. a lineáris vegyes 0—1-es feladat, a lineáris komplementaritási feladat, a szeparábilis programozási feladat, az extremális pont optimalizálási feladat stb. [2, 3, 9].

Az (1.1) alakú diszjunktív programozási feladatot *felületi feladatnak* nevezzük, ha $X\cap\{\mathbf{x}|\mathbf{a}_i^{h\prime}\mathbf{x}\geqq b_i^h\}$ vagy üres, vagy az X poliéder egy határoló felülete (*face*, [8]) minden $i\in Q_h$, $h\in H$ esetén. Az X poliéder egy F részhalmazát akkor nevezzük X határoló felületének, ha $F=X$ vagy létezik az X poliédernek olyan E érintő hipersíkja, hogy $F=X\cap E$.

Legyen $P=\text{cl conv}\,(X\cap Y)$ az $X\cap Y$ megengedett tartomány konvex burkának lezártja. BALAS megmutatta [1, 3, 9], hogy P konvex poliéder és hogy az (1.1) feladat helyett elég a $\min\{\mathbf{c}'\mathbf{x}|\mathbf{x}\in P\}$ feladattal foglalkozni. Ha a P poliédert elő tudjuk állítani explicit alakban, akkor (1.1) helyettesíthető egy lineáris programozási feladat megoldásával. Tegyük fel, hogy a P poliéder q-dimenziós. Ekkor P előáll a P poliédert tartalmazó q-dimenziós sokaság és bizonyos, a P $(q-1)$-dimenziós

határoló felületein átfektetett, R^n-beli hipersíkok által meghatározott félterek metszeteként. BALAS módszert adott ezen $(q-1)$-dimenziós határoló felületek előállítására [1, 3, 9], azonban ha $\prod_{h \in H} |Q_h|$ nagy, akkor ez a módszer olyan nagyméretű lineáris egyenlőtlenség rendszer által meghatározott poliéder összes csúcsának meghatározását jelenti, amely feladatot a mai módszerekkel és számítógépekkel elfogadható időn belül nem tudunk végrehajtani. A felületi diszjunktív programozási feladatok esetén azonban a BALAS által adott módszer jelentősen egyszerüsödik és a P poliéder is bizonyos speciális tulajdonságokkal rendelkezik. Ezt felhasználva JEROSLOW [6], valamint SHERALI és SHETTY [11] véges metszősík módszert adtak a felületi feladat megoldására.

Ebben a dolgozatban megmutatjuk, hogy tetszőleges (1.1) alakú, de nem felületi diszjunktív programozási feladat új változók és feltételek bevezetése árán egy ekvivalens felületi feladattá konvertálható. Ezután egy új, véges metszősík módszert mutatunk be a felületi feladat megoldására. Ennél a módszernél, az eddig ismertektől eltérően, a megengedett pontok halmazának korlátossága nincs megkövetelve.

2. Lineáris diszjunktív programozási feladat felületi feladat alakjára történő transzformálása

Legyen egy lineáris diszjunktív programozási feladat az (1.1) konjunktív alakban megadva. Legyen T azon (i, h) indexpárok halmaza, ahol $X \cap \{\mathbf{x} | \mathbf{a}_i^{h\prime} \mathbf{x} \geqq b_i^h\}$ nem határoló felülete az X poliédernek, a továbbiakban az üres halmazt is határoló felületnek tekintve. Minden $(i, h) \in T$ indexpárhoz bevezetünk egy nemnegatív u_i^h változót. Tetszőleges $\mathbf{x}$ vektor esetén teljesíthető az $\mathbf{a}_i^{h\prime} \mathbf{x} + u_i^h \geqq b_i^h$ egyenlőtlenség, ha az u_i^h változót elég nagynak választjuk. Azonban $\mathbf{a}_i^{h\prime} \mathbf{x} \geqq b_i^h$ csak akkor teljesül, ha az előző egyenlőtlenségben u_i^h nempozitívnak is választható. Jelölje $\mathbf{u}$ az u_i^h, $(i, h) \in T$ komponensekből álló vektort. Könnyen látható, hogy

$$\begin{gathered} \min \mathbf{c}' \mathbf{x} \\ \mathbf{D}\mathbf{x} \geqq \mathbf{d}, \ \mathbf{x} \geqq \mathbf{0} \qquad (2.1) \\ \mathbf{a}_i^{h\prime} \mathbf{x} + u_i^h \geqq b_i^h, \ u_i^h \geqq 0, \quad (i, h) \in T \\ \begin{pmatrix} \mathbf{x} \\ \mathbf{u} \end{pmatrix} \in \bigcap_{h \in H} \left[\bigcup_{\substack{(i,h) \in T \\ i \in Q_h}} \left\{ \begin{pmatrix} \mathbf{x} \\ \mathbf{u} \end{pmatrix} \middle| -u_i^h \geqq 0 \right\} \cup \bigcup_{\substack{(i,h) \notin T \\ i \in Q_h}} \left\{ \begin{pmatrix} \mathbf{x} \\ \mathbf{u} \end{pmatrix} \middle| \mathbf{a}_i^{h\prime} \mathbf{x} \geqq b_i^h \right\} \right] \end{gathered}$$

felületi diszjunktív programozási feladat és ekvivalens az (1.1) feladattal. A (2.1) feladatban az (1.1) alakhoz képest $|T|$ számú új változó és feltétel szerepel.

3. Az X poliéder megengedett pontot már nem tartalmazó határoló felületének felderítése

A következőkben (1.1)-ről feltesszük, hogy felületi feladat. Ebben az esetben a megengedett pontok $X \cap Y$ halmaza az X poliéder bizonyos határoló felületeinek egyesítése. Ha az (1.1) felületi feladat esetén valamely (i, h) indexpárra $\{\mathbf{x} | \mathbf{a}_i^{h\prime} \mathbf{x} \geqq b_i^h\} \supset X$ teljesül, akkor ezen h index elhagyható a H indexhalmazból. Ha maga a H

indexhalmaz üressé válik, akkor (1.1) lineáris programozási feladattá redukálódik, ellenkező esetben könnyen látható, hogy a megmaradt (1.1) feladat esetén $X \subset \{\mathbf{x} | \mathbf{a}_i^{h\prime}\, \mathbf{x} \leqq b_i^h\},\ \forall i \in Q_h,\ h \in H$.

Az X poliéder egy F határoló felületét *nem-megengedettnek* nevezzük, ha $F \cap Y = \emptyset$. Ha az (1.1) feladat megengedett pontját kereső eljárás során egy nem-megengedett F határoló felületre bukkanunk, akkor érdemes F pontjait kizárni a további keresésből. Metszősík feltételt fogunk előállítani, amely levágja F pontjait, de meghagyja (1.1) megengedett pontjait. A későbbiekben egy ún. *célfüggvény metszést* is fogunk alkalmazni, amely kizár a keresésből minden olyan pontot, amely az addig talált legjobb megengedett célfüggvényértéknél nagyobb célfüggvényértékkel rendelkezik. Az eljárás egy adott lépésében jelölje

$$\mathbf{G}\mathbf{x} \geqq \mathbf{g} \tag{3.1}$$

az eddig előállított k számú metszősík feltételt, ahol $\mathbf{G}$ $(k \times n)$-es mátrix és $\mathbf{g}$ k-dimenziós vektor. Az eljárás indulásakor nyilván $k = 0$. Egy adott lépésben az X poliéder további vizsgálat alá eső része $X \cap Z$, ahol $Z = \{\mathbf{x} | \mathbf{G}\mathbf{x} \geqq \mathbf{g}\}$. Legyen $\mathbf{x}^0 \in X \cap Z$ és $S = \{h \in H | \mathbf{a}_i^{h\prime}\, \mathbf{x}^0 < b_i^h,\ \forall i \in Q_h\}$. Minden $h \in H \setminus S$ index esetén válasszunk ki egy olyan $i(h)$ indexet, amelyre $\mathbf{a}_{i(h)}^{h\prime}\, \mathbf{x}^0 = b_{i(h)}^h$ teljesül. Legyen $F = X \cap \{\mathbf{x} | \mathbf{a}_{i(h)}^{h\prime}\, \mathbf{x} = b_{i(h)}^h,\ \forall h \in H \setminus S\}$. Ha $S = \emptyset$, akkor F *megengedett* határoló felület, azaz $F \subset Y$. Az $S \neq \emptyset$ esetben megpróbáljuk az X poliédernek egy olyan megengedett határoló felületét megtalálni, amely az F határoló felületben fekszik.

Rögzítsünk le egy $p \in S$ indexet, válasszunk ki egy $j \in Q_p$ indexet és oldjuk meg a

$$\max \{\mathbf{a}_j^{p\prime}\, \mathbf{x} | \mathbf{x} \in F \cap Z\} \tag{3.2}$$

lineáris programozási feladatot. Mivel (1.1) felületi feladat, így (3.2) véges z_j^p optimumértékkel rendelkezik és nyilván $z_j^p \leqq b_j^p$. Ha $z_j^p = b_j^p$, akkor távolítsuk el a p indexet az S indexhalmazból, legyen $i(p) = j$ és tekintsük most az új $H \setminus S$ indexhalmaz által meghatározott F határoló felületet. A $z_j^p < b_j^p$ esetben válasszunk egy új $j \in Q_p$ indexet és oldjuk meg az új (3.2) feladatot. Ha az derül ki, hogy $z_j^p < b_j^p$ minden $j \in Q_p$ esetén, akkor $F \cap Z \cap Y = \emptyset$ és egy metszősík feltételt fogunk előállítani, amely kizárja $F \cap Z$ pontjait a további keresésből.

4. A metszősík feltétel előállítása

A (3.2) feladat, megfelelő kiegészítő változók bevezetésével, a következő alakban is felírható:

$$\max z = \mathbf{a}_j^{p\prime}\, \mathbf{x}$$

$$\mathbf{D}\mathbf{x} - \mathbf{y} = \mathbf{d},\quad \mathbf{G}\mathbf{x} - \mathbf{w} = \mathbf{g},\quad \mathbf{x} \geqq \mathbf{0},\quad \mathbf{y} \geqq \mathbf{0},\quad \mathbf{w} \geqq \mathbf{0}, \tag{4.1}$$

$$\mathbf{a}_{i(h)}^{h\prime}\, \mathbf{x} + v_{i(h)}^h = b_{i(h)}^h,\quad v_{i(h)}^h = 0,\quad h \in H \setminus S.$$

Gyűjtsük össze a (4.1)-ben szereplő változókat az $(m+n+k+|H \setminus S|)$-dimenziós $\bar{\mathbf{x}}$ vektorba, a jobb oldalon szereplő komponenseket az $(m+k+|H \setminus S|)$-dimenziós $\bar{\mathbf{b}}$ vektorba, a célfüggvény komponenseket az $(m+n+k+|H \setminus S|)$-dimenziós $\mathbf{c}_j^p$ vektorba és a feltételi együtthatókat az $(m+k+|H \setminus S|) \times (m+n+k+|H \setminus S|)$ méretű

$\mathbf{A}$ mátrixba. Jelölje J a $v^h_{i(h)}$, $h \in H \setminus S$ változók $\bar{\mathbf{x}}$ vektorbeli indexhalmazát. Írjuk a (4.1) feladatot

$$\max \{z = \bar{\mathbf{c}}_j^{p\prime} \bar{\mathbf{x}} | \mathbf{A}\bar{\mathbf{x}} = \bar{\mathbf{b}}, \bar{\mathbf{x}} \geqq \mathbf{0}, \bar{x}_i = 0, \ i \in J\} \tag{4.2}$$

alakba. Tegyük fel, hogy a (4.1) feladatot a (4.2) feladat alakjában szimplex módszerrel oldottuk meg. Legyen $\mathbf{B}$ a (4.2) optimális bázisa és jelölje I_B, ill. I_R a bázisváltozók, ill. a bázison kívüli változók indexhalmazát. Legyen $\mathbf{r}^{(j,p)}$ az optimális bázishoz tartozó redukált költségek vektora. Nyilván $r_i^{(j,p)} \geqq 0$, $\forall i \in I_R \setminus J$ és $r_i^{(j,p)} = 0$, $\forall i \in I_B$. Az optimális szimplex tábla célfüggvény sorából kapjuk, hogy

$$z = z_j^p - \sum_{i \in I_R} r_i^{(j,p)} \bar{x}_i = z_j^p - \mathbf{r}^{(j,p)\prime} \bar{\mathbf{x}}.$$

Nyilvánvaló, hogy $\bar{\mathbf{c}}_j^{p\prime} \bar{\mathbf{x}} \geqq b_j^p$, $\bar{\mathbf{x}} \geqq 0$ esetén teljesül az

$$(1/(z_j^p - b_j^p))\mathbf{r}^{(j,p)\prime} \bar{\mathbf{x}} \geqq 1, \quad \bar{\mathbf{x}} \geqq \mathbf{0} \tag{4.3}$$

feltételrendszer is. Mivel $X \cap Z$ pontjait tekintve $\bar{\mathbf{x}} \geqq 0$, ezért $\bar{\mathbf{r}}^{(j,p)} \geqq 0$ esetén (4.3) inkonzisztens. Ebben az esetben $\mathbf{a}_j^{p\prime} \mathbf{x} < b_j^p$, $\forall \mathbf{x} \in X \cap Z$, így a j indexet elhagyhatjuk a Q_p indexhalmazból. Nyilvánvaló, hogy tetszőleges $\mathbf{x} \in X \cap Y \cap Z$ teljesíti a (4.3) feltételrendszert legalább egy $j \in Q_p$ esetén. Állítsuk elő minden $j \in Q_p$ esetén a megfelelő (4.3) feltételrendszert. Ha mindegyik inkonzisztens, akkor $X \cap Y \cap Z = \emptyset$ és leállhatunk. Különben, BALAS diszjunktív metszés ötletét felhasználva [1, 2, 3, 9] állítsuk elő az

$$\mathbf{r}' \bar{\mathbf{x}} \geqq 1 \tag{4.4}$$

egyenlőtlenséget, ahol

$$r_i = \max_{j \in Q_p} [r_i^{(j,p)}/(z_j^p - b_j^p)].$$

Az $X \cap Y \cap Z$ halmaz pontjaiból képzett $\bar{\mathbf{x}}$ vektorok nyilván teljesítik a (4.4) egyenlőtlenséget. Mivel $z_j^p < b_j^p$, $\forall j \in Q_p$ és $r_i^{(j,p)} \geqq 0$, $\forall i \notin J$, $j \in Q_p$, ezért $r_i \leqq 0$ minden $i \notin J$ index esetén. Az $F \cap Z$ halmazbeli pontoknak megfelelő $\bar{\mathbf{x}}$ vektorokban $\bar{\mathbf{x}}_i = 0$, $\forall i \in J$, így ezen pontokra nem teljesül a (4.4) egyenlőtlenség. Megjegyezzük, hogy a (4.4) metszés mélyíthető GLOVER negatív él kiterjesztési ötletének felhasználásával [5, 9]. Mielőtt a (4.4) egyenlőtlenséget (3.1)-hez csatolnánk az $F \cap Z$ pontjainak kizárása céljából, a (4.4) egyenlőtlenséget a (3.1)-ben megkívánt alakra kell transzformálni. Ez könnyen megtehető a $\mathbf{w}$, $\mathbf{y}$ és $v^i_{i(h)}$ változók eliminálásával, felhasználva definíciójukat.

5. Optimalizálás megengedett felületeken

Tegyük fel, hogy közvetlenül vagy (3.2) feladatok megoldása után az $S = \emptyset$ esethez jutunk. Ekkor az $F = X \cap \{\mathbf{x} | \mathbf{a}^h_{i(h)\prime} \mathbf{x} = b^h_{i(h)}, \ \forall h \in H\}$ határoló felület megengedett, azaz $F \subset Y$, továbbá $F \cap Z \neq \emptyset$. Azt is tudjuk, hogy a Z halmazt előállító $\mathbf{Gx} \geqq \mathbf{g}$ feltételek közül legfeljebb csak a célfüggvény metszés vág bele az F határoló felületbe, így az $F \setminus Z$ halmazbeli pontok (1.1) optimális megoldása szempontjából nem érdekesek. Oldjuk meg a

$$\min \{\mathbf{c}' \mathbf{x} | \mathbf{x} \in F\} \tag{5.1}$$

lineáris programozási feladatot. Ha (5.1) célfüggvénye alulról nem korlátos az F határoló felületen, akkor nyilván (1.1) azonos célfüggvénye sem korlátos alulról az $X \cap Y$ megengedett halmazon. Ha (5.1)-nek véges optimuma van, akkor az F egy csúcspontján, így X egy csúcspontján is felvétetik. Jelölje $\mathbf{x}^0$ ezt a csúcspontot. Legyen

$$\bar{Q}_h = \{i \in Q_h | \mathbf{a}_i^{h\prime} \mathbf{x}^0 = b_i^h\}, \quad h \in H.$$

$\bar{Q}_h \neq \emptyset$, mivel $i(h) \in \bar{Q}_h$, $h \in H$. Nyilvánvaló, hogy tetszőleges $j(h) \in \bar{Q}_h$, $h \in H$ indexek választása esetén $\bar{F} = X \cap \{\mathbf{x} | \mathbf{a}_{j(h)}^{h\prime} \mathbf{x} = b_{j(h)}^h, h \in H\}$ az X poliéder megengedett határoló felülete. A már tekintett F kivételével vizsgáljuk meg egymás után az így előálló $\bar{F}$ határoló felületeket, vajon csökkenthető-e tovább (5.1) célfüggvénye, ha az F megengedett halmazt ideiglenesen az $\bar{F}$ határoló felületre cseréljük fel. Könnyen látható, hogy ez az $\mathbf{x}^0$ ponthoz tartozó optimális szimplex táblában végzett vizsgálattal eldönthető. A csökkentés lehetősége esetén cseréljük fel véglegesen a két határoló felületet, oldjuk meg az új (5.1) feladatot és végezzük el újból a vizsgálatokat. Mivel az X poliédernek véges sok határoló felülete van, így véges sok célfüggvény értéket javító határoló felület csere után vagy az derül ki, hogy (1.1) célfüggvénye az $X \cap Y$ halmazon alulról nem korlátos, vagy X olyan $\mathbf{x}^0$ csúcspontjához jutunk, ahonnan a fenti határoló felület cserével nem tudunk továbblépni.

Tegyük fel, hogy az utóbbi esetnél vagyunk. Az (1.1) optimális megoldásának meghatározása szempontjából a továbbiakban csak azon $\mathbf{x}$ pontok érdekesek, melyekre $\mathbf{c}'\mathbf{x} < \mathbf{c}'\mathbf{x}^0$. Ha pusztán a $-\mathbf{c}'\mathbf{x} \geqq -\mathbf{c}'\mathbf{x}^0$ egyenlőtlenséget adnánk hozzá (3.1)-hez, akkor a legutolsó (5.1) feladat optimális megoldásai, amelyek X egy határoló felületét alkotják, nem lennének kirekesztve a további keresésből. Válaszszunk egy pozitív ε számot és adjuk hozzá a

$$-\mathbf{c}'\mathbf{x} \geqq \mathbf{c}'\mathbf{x}^0 + \varepsilon \tag{5.2}$$

célfüggvény metszést (3.1)-hez, vagy ha ilyen már volt, akkor cseréljük fel a régit (5.2)-vel. Az (5.2) feltétellel a legutolsó $\mathbf{x}^0$ csúcspont által generált összes $\bar{F}$ megengedett határoló felületet levágjuk.

Ezután a maradék $X \cap Z$ poliéderen újból megengedett pontot tartalmazó határoló felületet keresünk. Ha közben kiderül, hogy $X \cap Y \cap Z = \emptyset$, akkor még nem állíthatjuk, hogy a legutoljára talált $\mathbf{x}^0$ csúcspont az (1.1) optimális megoldása. Meg kell vizsgálnunk, hogy a $V = X \cap Y \cap \{\mathbf{x} | \mathbf{c}'\mathbf{x}^0 - \varepsilon < \mathbf{c}'\mathbf{x} < \mathbf{c}'\mathbf{x}^0\}$ halmaz üres vagy sem. Ha $V = \emptyset$, akkor a legutoljára talált $\mathbf{x}^0$ az (1.1) feladat optimális megoldása. A $V \neq \emptyset$ esetben a

$$\min \{\mathbf{c}'\mathbf{x} | \mathbf{x} \in V\} \tag{5.3}$$

feladat optimális megoldása lesz azonos (1.1) optimális megoldásával. Mivel $X \cap Y \cap Z \cap \{\mathbf{x} | \mathbf{c}'\mathbf{x} \leqq \mathbf{c}'\mathbf{x}^0 - \varepsilon\} = \emptyset$ és $X \cap Y$ az X poliéder bizonyos határoló felületeinek egyesítése, $V \neq \emptyset$ esetben (5.3) optimuma X egy csúcspontján is felvétetik. Hagyjuk el az (5.2) feltételt (3.1)-ből és legyen Z az így kapott (3.1) rendszernek eleget tevő pontok halmaza. Nyilván $X \cap Y \subset Z$. Jelölje vert X az X poliéder csúcspontjainak halmazát és legyen $W = Z \cap \{\mathbf{x} | \mathbf{c}'\mathbf{x}^0 - \varepsilon \leqq \mathbf{c}'\mathbf{x} \leqq \mathbf{c}'\mathbf{x}^0\}$. Az (5.3) feladat optimális megoldása $W \cap \text{vert } X$ egy eleme. A $W \cap \text{vert } X$ halmaz meghatározása egy extremális pont optimalizálási feladat megengedett pontjainak meghatározásával azonos [4, 7, 9, 12].

Tekintsük most a

$$\min \{\mathbf{c}'\mathbf{x} | \mathbf{x} \in W \cap Y \cap \text{vert } X\} \tag{5.4}$$

feladatot. Könnyen belátható, hogy az utoljára talált $\mathbf{x}^0$ csúcspont (5.4) megengedett pontja, továbbá (5.4) optimum értéke pontosan akkor kisebb a $\mathbf{c}'\mathbf{x}^0$ értéknél, ha $V \neq \emptyset$. Tegyük fel, hogy ismerjük a $W \cap \text{vert } X$ véges halmazt. Ekkor (5.4) megoldása elvégezhető úgy, hogy a $W \cap \text{vert } X$ halmazból elhagyjuk az Y halmazba nem tartozó elemeket és meghatározzuk a minimális $\mathbf{c}'\mathbf{x}$ értéket a maradék véges halmazon. A $W \cap \text{vert } X$ halmaz elemeinek előállítására véges metszősík módszerek léteznek [4, 7, 9]. Ezáltal véges lépésben megoldhatjuk az (5.4) feladatot, amelynek optimális megoldása egyben (1.1) optimális megoldása is.

6. A végesség

Az (1.1) feladat megoldására javasolt eljárás minden főiterációjában véges számú (3.2) feladat megoldása után az X olyan F határoló felületéhez jutunk, melyre $F \cap Z \neq \emptyset$, és vagy $F \subset Y$ vagy $F \cap Z \cap Y = \emptyset$. Az első esetben az 5. fejezetben leírtak szerint megengedett felületeken való optimalizálást hajtunk végre. A második esetben a 4. fejezetben leírtak szerint metszősík feltételt állítunk elő $F \cap Z$ pontjainak elhagyása céljából. Mindkét esetben legalább eggyel csökken X azon határoló felületeinek a száma, amelyek tartalmaznak pontot a maradék $X \cap Z$ poliéderből. Mivel az X poliédernek véges sok határoló felülete van, véges sok főiteráció után az alábbi esetek egyikéhez jutunk:

(i) az (5.1) feladat célfüggvényértéke alulról nem korlátos a megengedett pontok halmazán és ugyanez teljesül az (1.1) feladatra is;
(ii) $X \cap Z = \emptyset$;
(iii) az inkonzisztens (4.3) rendszerekhez tartozó $j \in Q_p$ indexek elhagyása után $Q_p = \emptyset$.

Az (i) esetben leállhatunk. Ha az (ii)—(iii) eseteknél vagyunk és előzőleg az (1.1) feladat egyetlenegy megengedett pontját sem állítottuk elő, akkor (1.1)-nek nincs megengedett megoldása, leállhatunk. Az ellenkező esetben oldjuk meg az (5.4) feladatot az 5. fejezet végén leírtak szerint. Az (5.4) optimális megoldása egyben (1.1) optimális megoldása is lesz.

IRODALOM

[1] BALAS, E., "Disjunctive programming: Properties of the convex hull of feasible points", MSRR 384, GSIA, Carnegie-Mellon University, 1974.
[2] BALAS, E., "Disjunctive programming: Cutting Planes from Logical Conditions", in: *Nonlinear Programming*, MANGASARIAN, O. L., MEYER, R. R. and ROBINSON, S. M. eds. (Academic Press, New York, 1975).
[3] BALAS, E., "Disjunctive programming", *Annals of Discrete Mathematics* 5 (1979) 3—51.
[4] FÜLÖP, J., „Eredeti csúcspont keresése és alkalmazása konkáv függvény lineáris feltételek melletti minimalizálásakor", *Alkalmazott Matematikai Lapok* 9 (1983) 51—72.
[5] GLOVER, F., "Polyhedral convexity cuts and negative edge extension", *Zeitschrift für Operations Research* 18 (1974) 181—186.
[6] JEROSLOW, R. G., "A cutting plane game and its algorithms", Discussion Paper No. 7724, Center for Operations Research and Econometrics, University of Catholique de Louvain, June, 1977.
[7] MAJTHAY, A. and WHINSTON, A., "Quasi-concave minimization subject to linear constraints", *Discrete Mathematics* 9 (1974) 35—59.
[8] ROCKAFELLAR, R. T., *Convex Analysis* (Princeton University Press, Princeton, N. Y., 1970).

[9] SHERALI, H. D. and SHETTY, C. M., *Optimization with disjunctive constraints* (Springer-Verlag, New York, 1980).
[10] SHERALI, H. D. and SHETTY, C. M., "A finitely convergent algorithm for bilinear programming problems using polar cuts and disjunctive face cuts", *Mathematical Programming* **19** (1980) 14—31.
[11] SHERALI, H. D. and SHETTY, C. M., "A finitely convergent procedure for facial disjunctive programs", *Discrete Applied Mathematics* **4** (1982) 135—148.
[12] SHERALI, H. D. and SEN, S., "A disjunctive cutting plane algorithm for the extreme point mathematical programming problem", *Operations Research* **22** (1985) 83—94.

(Beérkezett: 1986. december 2.)

FÜLÖP JÁNOS
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1111 BUDAPEST, KENDE U. 8—10.

ON THE FACIAL DISJUNCTIVE PROGRAMMING PROBLEM

J. FÜLÖP

In the paper it is shown that any linear disjunctive programming problem can be transformed into the form of an equivalent facial disjunctive programming problem. A finite cutting plane method is proposed for the facial problem. At this method, contrary to other ones, the boundedness of the feasible region is not assumed.

Alkalmazott Matematikai Lapok **13** (1987—88) 17—42

BINÁRIS VÁLTOZÓK STRUKTÚRÁJÁNAK VIZSGÁLATA

TELEGDI LÁSZLÓ

Budapest

A dolgozat olyan, nagyszámú objektumon megfigyelt bináris valószínűségi változók struktúrájának statisztikai vizsgálatával foglalkozik, amelyek dichotom jellegek indikátor változói. Vizsgálja a változók függetlenségét abban az esetben, ha a változószám nagy és/vagy az egyes jellegek valószínűsége kicsi. Ha a változók nem függetlenek, függőségüket a normális küszöb modellel jellemezzük. Ez feltételezi, hogy az egyes változókhoz hozzá van rendelve egy-egy háttér változó. Ezek együttes normális eloszlásúak, egy-egy küszöb tartozik hozzájuk, továbbá egy objektum akkor és csak akkor rendelkezik valamely jelleggel, ha a megfelelő háttér változó rajta felvett értéke eléri a megfelelő küszöböt. A dolgozat ismerteti a (dichotom jellegekkel, ill. bináris változókkal jellemzett) objektumok klaszteranalízisére kidolgozott új eljárást, a többszörös többdimenziós skálázást, amely a változók közönséges többdimenziós skálázása révén sorolja klaszterekbe az objektumokat. Végül a számítógépek egyik fontos orvosbiológiai alkalmazásával, a veleszületett rendellenességek statisztikai vizsgálatával foglalkozik.

1. Bevezetés

Attól függően, hogy milyen tulajdonságra vonatkozik, a valószínűségi változónak több típusa van. Kiemelkedő fontosságú a következő kettő:

a) *Kvantitatív (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) *Kvalitatív (minőségi, kategorikus) változó.* Nem mérhető és így 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 kvalitatív változót, amelyiknek csak két különböző megfigyelési értéke fordulhat elő (igen-nem, férfi-nő, élő-nem élő, beteg-egészséges stb.), *dichotom (alternatív) 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 (meglétének mint eseménynek az) *indikátor változója.* Ilyen változók képezik jelen dolgozat tárgyát. Olyan bináris változók struktúráját vizsgáljuk 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, $y_1, y_2, \ldots, y_M$. A dolgozatban azzal az esettel foglalkozunk, amikor a változókat nagyszámú objektumon figyeljük meg. Vizsgálatuk azon az adatmezőn alapul, amelyik megadja, hogy

az egyes objektumok mely jellegekkel rendelkeznek. Tetszőleges objektumot a jellegek egy részhalmaza jellemez. Az objektumok olyan kísérletek kimenetelének is tekinthetőek, amelyek során a jellegek (pontosabban a meglétük) mint események vagy bekövetkeztek, vagy nem. Ily módon a változók struktúrájának vizsgálata ezen események egymás közti kapcsolatainak vizsgálatát jelenti. Ehhez szükséges a 2^n számú $A_{i_1}A_{i_2}...A_{i_k}$ esemény valószínűségének meghatározása. Ezt a feladatot 2^n már tízes nagyságrendű n mellett is igen nagy volta teszi bonyolulttá.

Legegyszerűbb dolgunk akkor lenne, ha az A_i események — tehát a jellegek, ill. a változók — (teljesen) függetlenek lennének: ekkor tetszőleges $A_{i_1}A_{i_2}...A_{i_k}$ esemény valószínűsége az A_{i_1}, A_{i_2}, ..., A_{i_k} események valószínű égeinek szorzata. A 2. szakasz a változók függetlenségvizsgálatával foglalkozik, amelyet nagyobb n és/vagy kis $P(A_i)$-k esetén nem könnyű elvégezni.

Ha a változók nem függetlenek, jellemeznünk kell függőségüket. A többdimenziós normális eloszlás egyszerűsége és gyakori előfordulása adta az ötletet, hogy vizsgáljam 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 eloszlásúak, egy-egy küszöb tartozik hozzájuk, továbbá egy objektum akkor és csak akkor rendelkezik az A_i jelleggel, ha L_i rajta felvett értéke eléri a megfelelő küszöböt. A 3. szakasz 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 klaszteranalízise. A 4. szakasz a (dichotom jellegekkel, ill. bináris változókkal jellemzett) objektumok klaszteranalízisére kidolgozott új eljárással, az ún. 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 megjeleníteni valamely R^k alacsony dimenziós 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 MMDS azt a problémát vizsgálja, amelyik akkor merül fel, ha a fenti konzisztencia nem teljesül: hogyan lehet az objektumokat minél homogénebb klaszterekbe sorolni, amikor is egy-egy klaszter homogenitását a változók ezen klaszter mellett történő MDS-ének jóságával mérjük?

A dolgozat 5. szakasza a számítógépek egyik fontos orvosbiológiai alkalmazásával, a veleszületett rendellenességek statisztikai vizsgálatával foglalkozik. Hangsúlyozni szeretném azonban, hogy itt többről van szó, mint a 2—4. szakaszban ismertetésre kerülő matematikai eredmények alkalmazásáról egy konkrét feladatban: az eredmények jelentős része ezen konkrét feladatból nőtt ki. A rendelkezésre álló adatok mennyisége indokolja a számítógépes feldolgozást. Az 5. szakasz eredményei is tanúsítják, hogy a felhasznált új módszerek segítségével új következtetéseket lehet levonni.

2. A változók függetlenségének vizsgálata

Legyen

$$G_g = \{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}, \quad 1 \leqq k \leqq n \quad (k \; G_g \text{ mérete}),$$

$$1 \leqq i_1 < i_2 < \ldots < i_k \leqq n,$$

tetszőleges jellegkombináció, és $O(G_g)$ azon objektumok száma, amelyek rendelkeznek a G_g-hez tartozó jellegekkel, de más jelleggel nem. Mivel (2^n-1) számú különböző G_g van, azért feltehető, hogy $1 \leqq g \leqq 2^n-1$. [Az $O(G_g)$ gyakoriságok éppen a megfelelő 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 számunkra érdekes esetekben a jellegek ritkán fordulnak elő, ezért a kontingenciatáblázat elemeinek túlnyomó többsége zérus.] Tetszőleges fenti alakú G_g jellegkombináció esetén $(1 \leqq g \leqq 2^n-1)$ jelölje $P_T(i_1, i_2, \ldots, i_k)$ annak valószínűségét, hogy egy objektum rendelkezik a G_g-hez tartozó jellegekkel és esetleg továbbiakkal is. Jelölje $P^{(k)}$ annak valószínűségét, hogy egy objektum pontosan k számú jelleggel rendelkezik, $O^{(k)}$ az ilyen objektumok számát $(k=0, 1, \ldots, n)$, és legyen $P=P^{(0)}$. Vezessük még be a következő jelöléseket:

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

$$S_0 = 1,$$

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

Tegyük most fel, hogy a W_i változók függetlenek. Ekkor tetszőleges $k=1, 2, \ldots, n$ és $1 \leqq i_1 < i_2 < \ldots < i_k \leqq n$ esetén

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

[Az a kérdés, hogy a 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 számunkra érdekes esetekben azonban — még a ritka jellegek lehetőség szerinti összevonása után is — 2^n nagysága és/vagy a $P_T(i)$ valószínűségek kicsisége miatt a χ^2-próba elvégzése illuzórikus. A függetlenséget tehát máshogyan kell ellenőrizni.] [13]-ban megmutattam, hogyan lehet az S_k-kat rekurzívan számítani, továbbá hogy

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

tehát az $O^{(k)}/M$ relatív gyakoriságoknak megfelelő $P^{(k)}$ valószínűségek előállításához P-n kívül elég az S_k-kat kiszámítani, amelyek — a q_i-ken keresztül — csak a $P_T(i)$ valószínűségektől függenek.

Tetszőleges A_i jelleg esetén jelölje $P(i)$ annak valószínűségét, hogy egy objektum rendelkezik A_i-vel, de más jelleggel nem, $O(i)$ az ilyen objektumok számát, és legyen e_{gi} a W_i változónak az y_g objektumon megfigyelt — 1 vagy 0 — értéke. A változók

függetlensége miatt

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

vagyis

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

ahonnan

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

Legyen $O=O^{(0)}$. Becsüljük a $P_T(i)$ valószínűségeket úgy, hogy $\hat{P}=O/M$ és

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

teljesüljön. Ekkor

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

Ezekből a $\hat{P}_T(i)$-kből az említett, rekurzív módon előállítjuk az $\hat{S}_k$ becsléseket, és a $P(k)$ valószínűségek becslését a

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

egyenlőséggel határozzuk meg. Mivel a változók függetlensége miatt

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

azért teljesülnie kell, hogy

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

ahonnan

$$O = M \prod_{i=1}^{n} \frac{O}{O+O(i)}. \tag{2.1}$$

Legyen $O_T(i)$ azon objektumok száma, amelyek rendelkeznek A_i-vel és esetleg további jellegekkel is. A $P_T(i)$ valószínűségek efficiens becslései az $[O_T(i)/M]$ relatív gyakoriságok. Mellettük a változók függetlensége esetén az

$$O = M \prod_{i=1}^{n} \frac{M-O_T(i)}{M} \tag{2.2}$$

egyenlőségnek kell teljesülnie. Az ebben szereplő $O_T(i)$ gyakoriságok azonban — ellentétben a (2.1) egyenlőségben szereplő $O(i)$ gyakoriságokkal — általában nem alapadatok, hanem csak a számítógépes feldolgozás során kerülnek meghatározásra. Ebből következik, hogy míg a (2.1) egyenlőség jobb oldalát általában közvetlenül ki tudjuk számítani, a (2.2) egyenlőségét általában nem. Ezért a (2.1) egyenlőség ellenőrzése lehet a függetlenség vizsgálatának első lépése, amelyet az $\{O^{(k)}\}$ és $\{M\hat{P}^{(k)}\}$

sorozatok összehasonlítása követhet. Ez a megfelelő kontingenciatáblázat kis gyakoriságú celláinak radikális összevonását jelenti. Valószínűleg lehetséges az ilyen cellák kevésbé radikális összevonása is; ez további kutatás tárgya lehet. — Az összehasonlítást a

$$\sum_{k=0}^{n} \frac{[O^{(k)} - M\hat{P}^{(k)}]^2}{M\hat{P}^{(k)}} \tag{2.3}$$

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

$$\sum_{k=1}^{n} \frac{[O^{(k)} - MP^{(k)}]^2}{MP^{(k)}}$$

kifejezés aszimptotikusan n szabadságfokú χ^2-eloszlást követne (lásd pl. [18], 143—144. oldal). Esetünkben azonban nem ismertek, becsüljük őket. Ilyenkor (lásd uo., 148. oldal) a szabadságfokot a becsült paraméterek számával csökkenteni kell. Az a sejtésem, hogy a valószínűségek ismertetett becslése két paraméter becslésére vezethető vissza, és így a (2.3) kifejezés aszimptotikusan $(n-2)$ szabadságfokú χ^2-eloszlást követ, de ezt bizonyítani nem sikerült. Független változók mellett generált véletlen adatmezőkön végzett számításaim alátámasztják ezt a sejtést.

Ha a változók függetlensége elfogadhatatlannak bizonyul, felmerülhet az a gondolat, hogy ezt csak néhány jelleg okozza. Ezért megvizsgáltam (lásd [13]), hogyan változik az egyes jellegek adott méretű jellegkombinációk közötti gyakorisága a méret változásával. $k, i=1, 2, \ldots, n$ esetén jelölje $O^{(k)}(i)$ az olyan objektumok számát, amelyek pontosan k számú jelleggel rendelkeznek, köztük A_i-vel. Tekintsük az $O^{(1)}(i), O^{(2)}(i), \ldots, O^{(n)}(i)$ értékeket $(i=1, 2, \ldots, n)$. $i=1, 2, \ldots, n$ mellett legyen $p_i(k)$ a megfelelő

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

relatív gyakoriság [annak relatív gyakorisága, hogy egy k számú jelleggel rendelkező objektum rendelkezik A_i-vel, éppen $kp_i(k)$, $k=1, 2, \ldots, n$]. Legyen m_i a $p_i(k)$-k

$$\sum_{k=1}^{n} O^{(k)}(i)\, p_i(k) \Big/ \sum_{k=1}^{n} O^{(k)}(i)$$

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

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

kifejezés értékét $(i=1, 2, \ldots, n)$. a_i azt mutatja, hogyan változik a $p_i(k)$ relatív 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 méretű 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 méretű jellegkombinációk túl sokan vagy túl kevesen vannak. Az, hogy ezt csak néhány jelleg okozza, azt jelenti, hogy az ezen jellegeket tartalmazó kombinációkból van túl sok vagy túl kevés. Ekkor azonban ezeknek a jellegeknek az adott méretű jellegkombinációk közötti gyakoriságai a méret változásával a többi jellegéitől eltérően változnak. Ezt a változást méri — az egyes jellegekre — az S_i kifejezés. Ennek aszimptotikus viselkedéséről az alábbiakat lehet mondani. Legyen

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

és

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

akkor $S_i = S_i^{(1)} - S_i^{(2)}$. Ha az m_i és $r_i(k)$ maximum likelihood becslések helyett a megfelelő χ^2 minimum becslések állanának [m_i esetén a $p_i(k)$-k

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

súlyozott négyzetátlagának négyzetgyöke, $r_i(k)$ esetén explicit módon nem felírható], akkor $S_i^{(1)}$ és $S_i^{(2)}$ aszimptotikusan χ^2-eloszlást követne (lásd pl. [4]). A megfelelő hipotéziseket H_1, H_2-vel jelölve nyilvánvaló módon $H_1 \subset H_2$. Ismeretes, hogy bizonyos feltételek mellett ilyenkor a maximum likelihood-ok különbségének kétszerese aszimptotikusan r szabadságfokú χ^2-eloszlást követ (lásd pl. [9]; r a paraméterek számának különbsége, esetünkben $2-1=1$). A χ^2 minimumokra azonban nem találtam hasonló állítást.

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 valószínűségi 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 jellegű feltétel nem jelent megszorítást).

Tetszőleges

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

$$1 \leqq i_1 < i_2 < \ldots < i_k \leqq n,$$

jellegkombináció a küszöb modellben ekvivalens a

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

eseménnyel, ezért annak valószínűsége, hogy egy objektum rendelkezik a G_g-hez tartozó jellegekkel, de más jelleggel nem, megegyezik $\tilde{G}_g$ valószínűségével, továbbá

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

ahol $P_T(G_g)$ $[= P_T(i_1, i_2, \ldots, i_k)]$ annak valószínűségét jelöli, hogy egy objektum rendelkezik a G_g-hez tartozó jellegekkel és esetleg továbbiakkal is.

A *normális küszöb modell* a fentieken kívül feltételezi, hogy az

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

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

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

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 öröklődését leíró speciális változatának (lásd pl. [17]) eredetéről [6]-ban esik szó. Bármilyen meglepő viszont, de az általánosan megfogalmazott modellel nem találkoztam az irodalomban. További vizsgálata, mindenekelőtt a korrelációs mátrix maximum likelihood becslésének meghatározása nehéz, de fontos és érdekes feladatnak látszik.

A T_i küszöböket úgy becsüljük, hogy $\hat{T}_i$ az

$$MP_T(i) = O_T(i)$$

egyenlet megoldása, ahonnan

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

($n = 1$ esetén ez a küszöb ML becslése). Legyen $O_T(i, j)$ azon objektumok száma, amelyek rendelkeznek A_i-vel, A_j-vel és esetleg további jellegekkel is. A korrelációs együtthatókat az

$$MP_T(i, j) = O_T(i, j), \quad 1 \leqq i < j \leqq n,$$

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

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

egyenlet megoldása ($1 \leqq i < j \leqq n$), 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

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

(lásd pl. [1]). $F(\tilde{T}, T, r)$ gyors számítógépes meghatározását például a [17]-ben ismertetett eljárás segítségével végezhetjük el.

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

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

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

$$1 \leqq i_1 < i_2 < \ldots < i_k \leqq n,$$

jellegkombináció mellett

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

$$j_1 \neq j_2,$$

esetén [a (3.3) feltétel alapvető fontosságú]

$$\hat{P}_T(G_g) = \sum_{v_{12}=0}^{\infty} \sum_{v_{13}=0}^{\infty} \cdots \sum_{v_{k-1,k}=0}^{\infty} \left\{ \left(\prod_{j_1=1}^{k-1} \prod_{j_2=j_1+1}^{k} \frac{\hat{r}_{j_1 j_2}^{v_{j_1 j_2}}}{v_{j_1 j_2}!} \right) \times \left[\prod_{j=1}^{k} Q^{(u_j)}(T_{i_j}) \right] \right\}, \tag{3.4}$$

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 pl. [11]); ez (3.2) általánosítása. $\hat{P}_T(G_g)$ (3.4) szerinti gyors számítógépes meghatározására TUSNÁDY GÁBOR dolgozott ki eljárást. Ha (3.3) nem teljesül, $P_T(G_g)$ *Monte-Carlo módszerrel* becsülhető (lásd pl. [7]). (Annak, hogy csak ebben az esetben dolgozunk *Monte-Carlo módszerrel*, az az oka, hogy a számunkra érdekes esetekben a P_T valószínűségek általában igen kicsik.) A $\hat{P}_T(G_g)$ becslések birtokában meghatározhatóak — legalábbis néhány kisebb k-ra — az

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

sorozat elemei, és ezek összehasonlíthatóak a megfelelő

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

sorozat elemeivel [$|G_g|$ a G_g jellegkombináció mérete, $O_T(G_g)$ pedig azon objektumok száma, amelyek rendelkeznek a G_g-hez tartozó jellegekkel és esetleg továbbiakkal is; eltérően attól az esettől, amikor a változók függetlenek, a $P^{(k)}$ valószínűségek becslését a normális küszöb modellben nem sikerült meghatározni].

Tetszőleges G_g jellegkombináció esetén legyen $E_T(G_g)$ a fenti $O_T(G_g)$ várható értéke, $G_g=\{A_i\}$ vagy $G_g=\{A_i, A_j\}$ esetén pedig

$$O_C^{(3)}(G_g) = \sum_{\substack{v:|G_v|=3 \\ G_v \supset G_g}} O_T(G_v)$$

és

$$E_C^{(3)}(G_g) = \sum_{\substack{v:|G_v|=3 \\ G_v \supset G_g}} E_T(G_v),$$

akkor

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

[amelyik egyébként nem χ^2-eloszlású, hiszen az $O_C^{(3)}(G_g)$ összeg tagjai nem függetlenek] azt mutatja, hogy G_g nagyobb méretű jellegkombinációkban való előfordulása mennyire tér el a normális küszöb modell szerint várttól.

Végül még egy problémáról szólok. Mivel az

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

korrelációs mátrix maximum likelihood becslése $n>2$ esetén nem ismeretes, $\mathbf{R}$-et elemenként becsüljük. Ezért előfordulhat, hogy az $\mathbf{R}$ mátrix így nyert $\hat{\mathbf{R}}$ becslése nem lesz pozitív szemidefinit, ami bizonyos további vizsgálatoknál (pl. faktoranalízisnél) akadályt jelent, ezért szükséges a becslés módosítása, pozitív szemidefinitté tétele. Ez többféleképpen is elvégezhető. Az $\hat{\mathbf{R}}$-hoz mátrix norma szerint legközelebbi $\mathbf{R}^*$ pozitív szemidefinit mátrix úgy határozható meg, hogy $\hat{\mathbf{R}}$ spektrálfelbontásában a negatív sajátértékeket 0-val helyettesítjük. Az így nyert

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

mátrix sajátértékeinek

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

összege viszont nagyobb n-nél. Legyen

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

akkor az

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

mátrix pozitív szemidefinit, és

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

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

Jelölje $\hat{\mathbf{R}}$ sajátértékeit $\lambda_1 \geqq \lambda_2 \geqq \ldots \geqq \lambda_n$. $\hat{\mathbf{R}}$-hoz „közeli", n nyomú, pozitív szemidefinit $\tilde{\mathbf{R}}$ mátrix úgy is meghatározható, hogy $\hat{\mathbf{R}}$ spektrálfelbontásában a λ_i-ket olyan μ_i-kkel helyettesítjük, amelyekre

$$\mu_1 \geqq \mu_2 \geqq \ldots \geqq \mu_n;$$

$$\mu_i \begin{cases} > 0, & \text{ha} \quad \lambda_i \geqq \varepsilon, \\ = 0 & \text{egyébként;} \end{cases}$$

továbbá

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

ahol $\varepsilon > 0$ olyan, hogy

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

teljesüljön, $f(x)$ pedig alkalmas függvény (pl. $\sqrt{x}$). Ez a kvadratikus programozási feladat a *Beale-módszerrel* oldható meg (lásd pl. [3]).

Az $\hat{\mathbf{R}}$ módosítására megadott mindkét módszer alkalmazása során olyan $\tilde{\mathbf{R}}$ mátrixot kapunk, amely ugyan pozitív szemidefinit és n nyomú, de rendelkezhet $\tilde{r}_{ii} \neq 1$ vagy $|\tilde{r}_{ij}| > 1$ elemekkel. Olyan $\tilde{\mathbf{R}}$-hoz, amely már „igazi" korrelációs mátrix, iteratív úton juthatunk, a következő két lépés ismételt alkalmazásával: a) a mátrixot pozitív szemidefinitté tesszük; b) a főátlóba 1-eket írunk, a főátlón kívüli, 1-nél nagyobb vagy (-1)-nél kisebb elemeket 1-gyel, ill. (-1)-gyel helyettesítjük. Tapasztalataim szerint az iteráció néhány lépés után „igazi" korrelációs mátrixhoz vezet.

4. Többszörös többdimenziós skálázás

Tegyük fel, hogy a változók többdimenziós skálázása (*m*ulti*d*imensional *s*caling, MDS) a feladat. Ez a következőket jelenti: valamilyen módon távolságokat kell konstruálni a változók között, és a változókat úgy 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 értelemben: ha két változó „közel" van egy harmadikhoz, egymáshoz is „közel" kell lenniük. Tegyük fel például (1. példa), hogy $n=3$, $M=44$, az első 19 objektum az A_1 és A_2, a következő 14 az A_1 és A_3, az utolsó 11 pedig az A_2 és 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:

$\underline{A_3}$

$\underline{A_1}$ $\underline{A_2}$

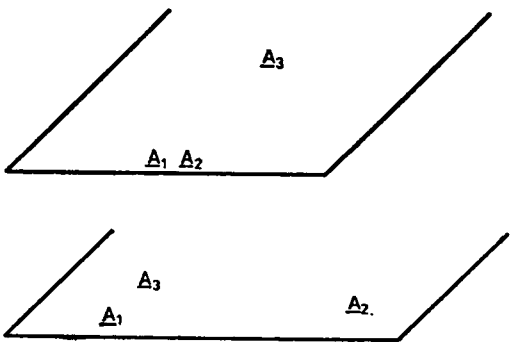
Legyen most (2. példa) $M=33$, és tegyük fel, hogy az első 19 objektum az A_1 és A_2, a további 14 pedig az A_1 és A_3 jellegekkel rendelkezik. Ezekhez az objektumokhoz a változók

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

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! Ezen két objektumhalmazhoz (amelyek klaszthereknek tekinthetőek) a változók

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

távolságmátrixai rendelhetőek hozzá, amik alapján viszont a változók — két síkon! — már jól skálázhatóak (lásd 1. ábra).



1. ábra

A fenti példákban szereplő gyakoriságokhoz természetesen többféle távolság rendelhető hozzá. Legyen

$$n_{ij} = \begin{cases} O_T(i), & \text{ha } i = j; \\ O_T(i,j) & \text{ha } i \neq j. \end{cases}$$

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. [8]) eredménye. Annak ellenére, hogy ennek a transzformációnak vannak bizonyos előnyös tulajdonságai (pl. ha a hasonlóságmátrix pozitív szemidefinit, akkor a megfelelő távolságmátrix euklideszi lesz), használatát sem általában, sem a konkrét esetben nem tartom jónak: túlságosan „kisimítja" 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 választjuk ($K=60$ mellett). Itt azonban ez nem lényeges, hiszen magukból a gyakoriságokból is látszik, hogy az említett konzisztencia az első példában teljesül, a másodikban nem.

A többszörös többdimenziós skálázás (*m*ultiple MDS, MMDS; lásd [10, 12, 14]) 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, amikor is egy-egy klaszter homogenitását a változók ezen klaszter mellett történő MDS-ének jóságával mérjük? Keresendő tehát a p természetes szám, az

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

diszjunkt klaszterek (amelyekre

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

és azon $\mathbf{x}_i^{(m)} \in R^k$ pontok, amelyek a változókat reprezentálják abban az értelemben, hogy $\mathbf{x}_i^{(m)}$ és $\mathbf{x}_j^{(m)}$ közelsége a W_i és W_j változók Y_m melletti közelségének felel meg.

Legyen

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

Mivel megegyező $\mathbf{e}_g$-jű objektumok megkülönböztethetetlenek, tegyük fel, hogy

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

viszont az objektumoknak multiplicitása van. Mivel az egyes objektumok klaszterbe sorolásának értelemszerűen a rajtuk 1 értékű változóktól kell függniük, az MMDS az olyan objektumokkal, amelyeken legfeljebb egy változónak 1 az értéke, nem tud mit kezdeni. Ezért tegyük fel, hogy minden objektumon legalább két változó értéke 1. Legyen a (változók alapján különböző ilyen) objektumok száma N. Ekkor minden objektum egy legalább 2 méretű jellegkombinációnak felel meg, és az y_g objektum multiplicitása $O(G_g)$ ($g=1, 2, \ldots, N$), ahol G_g az a jellegkombináció, amelyiknek y_g megfelel.

Legyen **DC** olyan eljárás, amellyel az objektumok tetszőleges $Z \subset Y$ halmazához — a Z objektumaihoz tartozó $\mathbf{e}_g$-k és $O(G_g)$-k alapján — $d_{ij}(Z)$ távolságokat

és $c_{ij}(Z)$ súlyozó tényezőket rendelünk hozzá,

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

és

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

Az MMDS-problémát azzal a T algoritmussal oldjuk meg (lásd [14]), amelyik a következő két lépést alkalmazza váltakozva: i) az egyes y_g objektumokat abba a klaszterbe sorolja, amelyikre

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

minimális; ii) az egyes klaszterek mellett meghatározza

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

negatív gradiens vektorát, majd ennek irányában irány menti minimalizálást végezve kiszámítja az $\mathbf{x}_i^{(m)}$ pontok új koordinátáit.

Az MMDS-t az

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

mennyiséggel jellemezzük, ahol

$$n_{ij}^{(m)} = n_{ij}(Y_m), \quad n_{ij}(Z) = \sum_{\substack{g: y_g \in Z \\ e_{gi}, e_{gj}=1}} O(G_g),$$

q és K alkalmas állandó. [Simonovits Miklós hívta fel a figyelmemet arra, hogy $q=2$ esetén E a következő, 2 direkciós erejű rugókból álló rendszer potenciálja: $n_{ij}^{(m)}$ számú 0 hosszúságú és egy K hosszúságú rugó $\mathbf{x}_i^{(m)}$ és $\mathbf{x}_j^{(m)}$ között.] Az MMDS E-vel történő jellemzése a következő **DC**-nek felel meg:

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

[14]-ben bebizonyítottam, hogy ezen **DC** és $q=2$ esetén a T algoritmus folyamán E monoton nem nő. [14, 15]-ben ismertettük, hogyan célszerű p kezdő és további értékeit, q-t és K-t választani, valamint kezdő klasztereket és pontkonfigurációt előállítani.

Valamely adatmező MMDS-ét két szempontból is jó lenne értékelni. Annak matematikai értékelése, hogy az MMDS mennyire jó más klaszterező módszerekhez képest, korrekten kivihetetlen. 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 hasonlíthatóak objektíven össze. (Különböző klaszterező módszereket heurisztikusan persze összehasonlíthatunk: 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 K mellett végrehajtott MMDS-ének $q=2$ mellett történő befejezésekor a klaszterszám p, és E értéke

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

ahol

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

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 $[E_{\text{opt}}, E_{\text{pessz}}]$ intervallumban helyezkedik el. Annak, hogy a konkrét adatmező mennyire jó az MMDS szempontjából, kézenfekvő mérőszáma az

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

mennyiség, amely nem lehet 0-nál kisebb és 1-nél nagyobb. Meghatározásához azonban tudni kellene E_{opt} és E_{pessz} értékét. Ezeket azonban nem tudjuk, mivel nem ismerjük az optimális és pesszimális adatmezőt.

Többszörös többdimenziós skálázás a változók függetlensége esetén

Az MMDS működésében meghatározó szerepe van az őt jellemző E mennyiségnek. Ennek megválasztásához úgy konkretizáltuk a távolságokat és súlyozó tényezőket meghatározó **DC** eljárást, hogy azok csak az egyes jellegpárok együttes előfordulásától függenek. Emiatt az egyes jellegek előfordulási valószínűségeinek különbözősége esetén az eljárás nem veszi észre a függetlenséget: akkor is közel hoz egymáshoz két jelleget, ha azok csak azért fordulnak elő sűrűn együtt, mert mindketten gyakoriak. (Ez azonban szándékos: a bevezetésben már említett konkrét feladat során, amelyből az MMDS kinőtt, nem akartunk észrevétlenül hagyni jellegzetes jellegkombinációkat.) Azt, hogy az eljárás észrevegye a függetlenséget, könnyen meg lehet valósítani: $n_{ij}(Z)$ definíciójában az összeget osztani kell a

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

kifejezéssel. Ha azonban az egyes jellegek előfordulási valószínűségei megegyeznek, erre nincs szükség. Ezért a függetlenséget ebben az esetben vizsgáltam, mégpedig determinisztikusan. Nevezetesen azt az adatmezőt vizsgáltam, amelyre

$$O_T(i,j) = E_T(i,j) \equiv \mu, \quad 1 \leqq i < j \leqq n$$

[$E_T(i,j)$ azon objektumok $O_T(i,j)$ számának várható értéke, amelyek rendelkeznek A_i-vel és A_j-vel, de más jelleggel nem]. Ekkor a q hatványkitevőt végig 2-nek és az objektumklaszterek p számát állandónak véve az MMDS olyan objektumklasztere-

ket adott, amelyek mellett nem rajzolódtak ki jellegklaszterek, és a jellegeknek megfelelő pontok — mindegyik objektumklaszter mellett — koncentrikus szabályos sokszögek csúcsain helyezkedtek el. Az 1. táblázat néhány n-re megadja az egyes sokszögek csúcsainak m_v számát (kívülről befelé haladva). Az E mennyiség $\tilde{E}$ végértékére az

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

összefüggést kaptam, amely $\mu=0$ és $p=K=1$ mellett, amikor is

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

az

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

alakra egyszerűsödik. [C_n-re $4\leqq n\leqq 65$ esetén a

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

egyenlőtlenséget kaptam $(0{,}1716=3-2\sqrt{2}=C_4$; $0{,}1807=\dfrac{75-36\sqrt{3}}{70}=C_7$; ahogy n értéke 65-höz közeledik, C_n értéke 0,176-hoz).] A fenti, speciális E minimalizálása a következő feladat megoldását jelenti: határozzunk meg a síkon n számú pontot úgy, hogy távolságaik átlaga 1, szórása minimális legyen!

Hipergráfok euklideszi térbe ágyazása és partícionálása

Feleltessünk meg az n számú 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 $n_{ij}=O_T(i,j)$ összértékű (közönséges) él köti össze $(1\leqq i<j\leqq n)$. *Ágyazzuk be* a H hipergráfot a k-dimenziós euklideszi térbe, vagyis rendeljünk $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\in R^k$ pontokat H egyes szögpontjaihoz (k természetes szám). Jó beágyazással szemben kézenfekvő azt a követelményt támasztani, hogy két pont annál közelebb legyen egymáshoz, minél nagyobb összértékű él köti össze a megfelelő szögpontokat, vagyis hogy az $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ pontokra

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

minimális legyen. Nyilvánvaló módon azonban az S_1 összeg $\mathbf{x}_i\equiv\mathbf{x}$ esetén 0, vagyis minimális. A beágyazástól tehát azt is meg kell követelni, hogy ne engedje az n számú pontot egy pontba összehúzódni. Keressük ezért azokat az $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ pontokat,

amelyekre

$$S = S_1 + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (K - \|\mathbf{x}_i - \mathbf{x}_j\|)^2$$

minimális, ahol K alkalmas állandó. Legyen

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

$(1 \leqq i < j \leqq n)$, akkor [14] 7.1. tételéből következik, hogy a H hipergráf S minimalizálásával történő euklideszi térbe ágyazása ekvivalens a változók

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c_{ij}(d_{ij} - \|\mathbf{x}_i - \mathbf{x}_j\|)^2 \tag{4.1}$$

minimalizálásával történő (közönséges) többdimenziós skálázásával.

Tegyük most fel, hogy H *partícionálása* a feladat. Végezzük ezt azon kritérium alapján, hogy a H_m rész-hipergráfok minél jobban beágyazhatóak legyenek $(m = 1, 2, \ldots, p$; p természetes szám), nevezetesen úgy, hogy E minimális legyen [$n_{ij}^{(m)}$ az i-edik és j-edik szögpontot H_m-ben összekötő élek összértéke, $\mathbf{x}_i^{(m)} \in R^k$ az i-edik szögpontnak H_m-ben megfelelő pont]. A H hipergráf E minimalizálásával történő partícionálása és euklideszi térbe ágyazása nyilvánvaló módon ekvivalens az MMDS-sel.

Klaszterek többdimenziós skálázása

Tegyük fel, hogy a változók MMDS-évl és/vagy egyéb módszerekkel v-félképpen (diszjunkt) klaszterekbe soroltuk az objektumokat. Kézenfekvő módon merül fel az a probléma, hogy miképpen lehet ezekből a klaszterezésekből egy un. konszenzus klaszterezést előállítani. A probléma megoldása céljából először jellemezzük az ugyanazon klaszterezés melletti klaszterek távolságát. Jelölje p_r a klaszterek számát az r-edik klaszterezésben, $Y_m^{(r)}$ pedig az ugyanezen klaszterezés melletti m-edik klasztert $(m = 1, 2, \ldots, p_r$; $r = 1, 2, \ldots, v)$. $Y_{m_1}^{(r)}$ és $Y_{m_2}^{(r)}$ távolságát a következő módon jellemezzük. Legyen

$$K_{m_1 m_2}^{(r)}(i) = \sum_m \frac{\left[n_i^{(rm)} \sum_l N^{(rl)} - N^{(rm)} \sum_l n_i^{(rl)}\right]^2}{N^{(rm)}\left[\sum_l n_i^{(rl)}\right] \sum_l \left[N^{(rl)} - n_i^{(rl)}\right]}, \tag{4.2}$$

ahol $N^{(rm)}$ és $n_i^{(rm)}$ az $Y_m^{(r)}$ klaszter összes, ill. azon objektumainak száma, amelyek rendelkeznek az A_i jelleggel, és az m, l-re történő összegezés (4.2)-ben m_1, m_2-re terjed ki $(1 \leqq m_1 < m_2 \leqq p_r$; $r = 1, 2, \ldots, v)$. Rögzített i mellett $j = 1$, 2-re legyen $v_{j1} = n_i^{(rm_j)}$, $v_{j2} = N^{(rm_j)} - n_i^{(rm_j)}$. Egyszerű számolással adódik, hogy ekkor

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

ami azt méri, hogy egy objektumnak az $Y_{m_1}^{(r)}$ vagy $Y_{m_2}^{(r)}$ klaszterbe esése mennyire függ össze az A_i jelleg előfordulásával, más szóval mennyire különböző a W_i változó

1 értékének relatív gyakorisága a két klaszterben. $K^{(r)}_{m_1 m_2}(i)$ (lásd pl. [18], 152. oldal) aszimptotikusan 1 szabadságfokú 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 $Y^{(r)}_{m_1}$ és $Y^{(r)}_{m_2}$ klaszterekben, aszimptotikusan n szabadságfokú χ^2-eloszlást követ. A változók függősége esetén ez nem teljesül, de az eloszlással nem dolgozunk, ezért $d^{(r)}_{m_1 m_2}$-et választjuk $Y^{(r)}_{m_1}$ és $Y^{(r)}_{m_2}$ távolságának. Az $Y^{(r)}_1, Y^{(r)}_2, \ldots, Y^{(r)}_{p_r}$ klaszteréknek a

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

távolságmátrix alapján történő MDS-ével konstruáljunk a klaszterekhez olyan $\mathbf{z}^{(r)}_1, \mathbf{z}^{(r)}_2, \ldots, \mathbf{z}^{(r)}_{p_r}$ pontokat a k_r-dimenziós euklideszi térben, hogy a pontok euklideszi távolsága minél jobban tükrözze a klaszterek távolságát ($r = 1, 2, \ldots, v$). Az r-edik klaszterezés mellett rendeljük az $[O(G_g)$ multiplicitású$]$ y_g objektumhoz azt a $\mathbf{z}^{(r)}_m$ pontot, amelyik az őt tartalmazó klaszternek megfelel, majd ezeknek a pontoknak képezzük azokat a $\mathbf{w}^{(r)}_g$ lineáris transzformáltjait ($g = 1, 2, \ldots, N$), amelyekre

$$\sum_{g=1}^{N} O(G_g)\mathbf{w}^{(r)}_g = \mathbf{0}, \quad \sum_{g=1}^{N} O(G_g)\|\mathbf{w}^{(r)}_g\|^2 = 1$$

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

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

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

és p^* a klaszterek kívánt száma a konszenzus klaszterezésben. A konszenzus klaszterek az y_g objektumoknak a (K_1-dimenziós) $\mathbf{v}_g$ pontok alapján (pl. k-közép, pontosabban itt most p^*-közép eljárás által) történő klaszterezésével állíthatóak elő. Ennek során az egyes klaszterezések jósága is figyelembe vehető oly módon, hogy a K_1-dimenziós távolságok, ill. normák számításánál az r-edik koordináta-k_r-es az r-edik klaszterezés jóságának megfelelő súllyal szerepel. Ezt a súlyozást azonban nem javaslom. Az egyes klaszterező módszerek ugyanis — a fentebb mondottakkal összhangban — nem hasonlíthatóak objektíven össze. Nem is szükséges azonban a súlyozás, mert a konszenzus klaszterek kialakításához a jobb klaszterezéseknek megfelelő koordináta-k_r-esek enélkül is nagyobb mértékben járulnak hozzá.

A konszenzus klaszterezés előállításának döntő mozzanata az objektumklaszterek MDS-e. Ez a korábbiakkal összhangban a következőket jelenti: távolságokat konstruálunk a klaszterek között, és a klaszterekét úgy próbáljuk meg beágyazni az alacsony dimenziós euklideszi térbe, hogy a klaszteréknek megfelelő pontok euklideszi távolságai minél kevésbé különbözzenek a klaszterek távolságaitól. Ahhoz, hogy a klaszterek jól skálázhatóak legyenek, a változókhoz hasonlóan (lásd fent) konzisztenseknek kell lenniük. Ha ez nem teljesül, megkísérelhetjük a klaszterek MMDS-ét. Ez a következő problémával foglalkozik: hogyan lehet a jellegeket minél homo-

génebb csoportokba sorolni, amikor is egy-egy csoport homogenitását a klaszterek ezen csoport mellett történő MDS-ének jóságával mérjük? Rögzített 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 $\mathbf{x}_m^{(\mu)}$ pontjai ($m=1, 2, \ldots, p_r$; $\mu=1, 2, \ldots, s$), amelyek a klasztereket reprezentálják abban az értelemben, hogy $\mathbf{x}_m^{(\mu)}$ és $\mathbf{x}_j^{(\mu)}$ közelsége az Y_m és Y_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 relatív 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

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

($i=1, 2, \ldots, n$). Az MMDS ezek után a fentebb vázolt módon végezhető el. — A K_2 állandót úgy célszerű választani, hogy azon jellegek átlagos száma, amelyek tipikusak egy klaszterre, közelítőleg $\sqrt{n}$ legyen.

Egy sávszélesség-redukcióval rokon probléma

Legyen

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

szimmetrikus, sok zérus-elemet tartalmazó mátrix. *A sávszélesség-redukció problémája* (lásd pl. [2]) a következő: melyik az $\mathbf{P}$ a permutációs mátrix, amely mellett a

$$\mathbf{B} = (b_{ij})_{i,j=1}^{n} = \mathbf{PAP}^T \tag{4.3}$$

mátrix

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

sávszélessége minimális? Ezzel rokon a következő probléma: melyik az a $\mathbf{P}$ permutációs mátrix, amely mellett a (4.3) szerint meghatározott $\mathbf{B}$ mátrixhoz tartozó

$$\frac{\sum_{i,j:\, b_{ij}\neq 0} |i-j|}{\sum_{i,j:\, b_{ij}\neq 0} 1} \tag{4.4}$$

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

Ez a probléma heurisztikusan — és valószínűleg közelítően — (közönséges) MDS-sel oldható meg, az alábbi módon. Legyen

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

továbbá $d_{ij}=K/c_{ij}$ $(1 \leqq i < j \leqq n$; K alkalmas állandó). Minimalizáljuk a (4.1) kifejezést, és jelölje $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ az így kapott pontokat. Legyen $(\hat{d}_{ij})$ az

$$(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$$

pontkonfiguráció euklideszi távolságmátrixa. „Fűzzük fel" az $\mathbf{x}_i$ pontokat úgy, hogy a $\hat{d}_{ij}$ 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 $(n-2)$ számú pontpár van még csak összekötve. Ily módon sorba rendezzük a pontokat, és ezzel kijelölünk egy **P** permutációs mátrixot. Az így nyert **P** mellett (4.3) szerint meghatározott (4.4) mennyiség általában kisebb, mint ha **P**-t úgy állítjuk elő, hogy a $\hat{d}_{ij}$ távolságok helyett magukkal a d_{ij}-kkel dolgozunk (és jóval kisebb, mint ha **P**-nek az egységmátrixot választjuk).

Ha a (4.1) kifejezést 2 helyett más, kellően nagy hatványkitevő mellett minimalizáljuk, magára a sávszélesség-redukció problémájára kapunk — természetesen heurisztikus és általában közelítő — megoldást. A viszonylag jelentős gépidőigény miatt ez azonban csak elméleti szempontból érdekes.

5. Veleszületett rendellenességek statisztikai vizsgálata

[13]-ban ismertettem a *veleszületett rendellenesség* (*c*ongenital *a*nomaly, *c*ongenital *a*bnormality, CA) definícióját: a születés pillanatában létező alaki, működési és/vagy biokémiai fogyatékosság. Jelenleg Magyarországon a regisztrált CA-k gyakorisága mintegy 4,5 százalék. A CA-k tényleges gyakorisága vélhetően 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 gyógyíthatóak 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 irányítója, és aminek része ez a dolgozat is, az előző két szakasznak a továbbiakban ismertetésre kerülő alkalmazása.

A következő CA-kat vizsgáltuk:
koponyahiány (AN), agysérv (EN), nyitott gerinc (SB), nyúlajak (CL), farkastorok (CP), végtagredukció (LR), többujjúság (PY), összenőtt ujjak (SY), nyitott has (EX), nyelőcsőelzáródás vagy -szűkület (OA), végbélelzáródás vagy -szűkület (AA), kisszeműség vagy szemnélküliség (AM), szürkehályog (CT), húgycsőhasadék (HS), a külső nemi szervek egyéb rendellenességei (EG), szívrendellenességek (HD), rekeszizomsérv (DI), vesehiány (RA), veseciszta (CK), vékonybélelzáródás (AI), kisfejűség (MC), vízfejűség (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), csípőficam (CD), dongaláb (CF), egyéb végtagrendellenességek (OL), a légzőszervek rendellenességei (RS), a gyomorkimenet szűkülete (PS), egyéb zsigeri rendellenességek (OD), le nem szállt here (UT), egyéb húgy-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-kon belül különös fontossággal bírnak a *többszörös veleszületett rendellenességek* (*m*ultiple CA, MCA). Definíciójuk megtalálható például [13]-ban: több CA ugyanannál a személynél történő együttes előfordulása.

Ugyanaz a CA különböző esetekben különböző kóreredetű lehet (a kóreredet általában nincs bejelentve, ezért az MCA-k statisztikai vizsgálatánál ismeretlen volt). A [17]-ben ismertetett vizsgálat is hozzájárult annak tisztázásához, hogy a CA-val rendelkező gyerekek jelentős részénél a CA(k) sok tényező hatásának az eredményei(i), tehát *multifaktoriális kóreredetű(ek)*. Ellenkező esetben, amikor is a CA(k) kevés tényező hatásának az eredménye(i), *oligofaktoriális kóreredetről* fogunk beszélni.

A GAMT-modell kiterjesztése

Mint arra fentebb utaltam, az esetek nagy részében a CA multifaktoriális kóreredetű. Az ilyen CA-k közül néhánynak az öröklődésmenetét az ún. GAMT-modellel sikerült leírni (vö. [17]).

[13]-ban megvizsgáltam az ún. függetlenségi koncepciót és néhány módosítását. Egy további módosítást jelent, ha feltesszük, hogy a különböző CA-kat előidéző genetikai és környezeti tényezők (természetesen ugyanannál a személynél) korreláltak, és a multifaktoriális kóreredetű 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 típus megkülönböztethetetlen, ezért a paragrafus további része azon a hipotézisen alapul, amely szerint az egyes CA-k öröklődésmenete a GAMT-modellel leírható, az őket előidéző genetikai és környezeti tényezők korreláltak, és valamennyi MCA ennek a korrelációnak a következménye. Mivel az MCA-k jelentős részben multifaktoriális kóreredetűek, ez a hipotézis ésszerű 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 A_i-vel, ha L_i rajta felvett értéke nem kisebb T_i-nél. Legyen

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

és tegyük fel, hogy — a normális küszöb modellnek megfelelően — $\mathbf{L}$ n-dimenziós normális eloszlású (és az L_i-k standardak). A küszöböket és az $\mathbf{L}$ eloszlását meghatározó r_{ij} korrelációs együtthatókat a 3. szakaszban leírt módon becsültem.

A fenti modell a GAMT-modell kiterjesztése. Az eredeti GAMT-modell egy CA több családtagra vonatkozó hajlamát írja le. A fenti modellben viszont egy gyermeknek több hajlama van — az n számú CA mindegyikéhez egy —, és ezeknek

a hajlamoknak a rendszere alakítja ki a gyerek CA-inak rendszerét. A modellt és így a megfelelő hipotézist a 3. szakasszal összhangban úgy ellenőriztem, hogy megnéztem, mennyire felel meg a 2-nél nagyobb méretű CA-kombinációk előfordulása az (r_{ij}) korrelációs mátrix alapján vártnak.

Legyen $O_T(i, j, k)$ és $E_T(i, j, k)$ az olyan gyermekek száma, ill. számának várható értéke, akik rendelkeznek A_i-vel, A_j-vel, A_k-val és esetleg további CA-kal is. A legnagyobb $O_T(i, j, k)$-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 írja le. A kisebb $O_T(i, j, k)$-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 $(O_T, \hat{E}_T)$ értékek, hanem a megfelelő összegek egymáshoz való viszonya érdekes. Ezért meghatároztam a

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

és — a 3. szakaszban említett eljárás segítségével — 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. szakaszban definiált $O_C^{(3)}(G_g)$, $E_C^{(3)}(G_g)$, 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 $O_C^{(3)}(G_g)$ értéke nagyobb volt, mint $\hat{E}_C^{(3)}(G_g)$-é]. Meghatároztam a

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

kifejezések értékét; 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 $a_{ij} = d(G_g)$ $(1 \leqq i < j \leqq n)$. Amíg az r_{ij} korrelációs együtthatók a CA-k multifaktoriális közelségeinek tekinthetőek, az a_{ij}-k a CA-k oligofaktoriális közelségeinek.

Egy véletlen és az „igazi" adatmező többszörös többdimenziós skálázása

Az MCA-k statisztikai vizsgálatának fő céljai közé tartozott a rendellenes gyerekek csoportosítása a CA-k együttes előfordulása alapján és ezáltal jellegzetes CA-kombinációk felderítése. E feladat megoldására dolgoztam ki az MMDS-t. Ennek adekvát voltáról úgy lehet meggyőződni, hogy véletlen módon generált vagy szisztematikusan kijelölt CA-klaszterekhez véletlen adatmezőt generálunk, és meg-

nézzük, úgy csoportosítja-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 (az MCA-kat leíró) 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, a generálást az ún. keverékeloszlásos modell (lásd [16]) alapján végeztem. 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 valószínűséggel rendelkezik a klaszterhez tartozó, 0,01—0,01 valószínűséggel a klaszterhez nem tartozó CA-kkal ($m=1, 2, \ldots, 6$; $g=1, 2, \ldots, 150$). A kettőnél kevesebb CA-val rendelkező gyerekeket elhagyva 666 számú, 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, C_m-mel jelölt gyerekklaszterek melleti CA-képek és -gyako-

riságok 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}: {EN, 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}: {*AN*, *SY*, *HD*, RA, FN, EO};

C_{16}: {EN, *CL*, *LR*, EX, AA, HY, *HD*, *UT*};

C_{17}: {AM, *HD*, *AI*, FN}, {TC, CD, IH};

C_{18}: {HD, RS};

C_{19}: {CP, HY, CT, HD, 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 dőlt betűvel jelöljük; nem minden gyerekklaszter mellett rajzolódott ki CA-klaszter, voltak viszont olyanok, amelyek mellett több is).

A rendellenes gyerekek csoportosítását és jellegzetes CA-kombinációk ezáltal történő felderítését keverékfelbontással TUSNÁDY GÁBOR (lásd [16]), orvosi megfontolások alapján CZEIZEL ENDRE (lásd [5]) is elvégezte. A háromféle klaszterezésből a 4. szakaszban leírt módon egy konszenzus klaszterezést állítottam elő. Az eredmények orvosi értékelése, ellenőrzése (genetikai családvizsgálatok) és felhasználása (genetikai tanácsadás) [5]-ben kerül kifejtésre.

Köszönetnyilvánítás

Az a kutató munka, amelynek során a dolgozatban ismertetésre kerülő eredmények 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, valamint DR. BOLLA MARIANNA matematikussal körülbelül fél évtizede végzünk. Az együttes munka során mindegyiküktől sok segítséget kaptam, de különösen TUSNÁDY GÁBORtól. A dolgozat alapjául a hasonló című kandidátusi értekezés magyarul még nem publikált részei szolgáltak. Az értekezés munkahelyi bírálatáért 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 az MTA SZTAKI Alkalmazott Matematikai Főosztályának Statisztika Osztályán készült. A munka elvégzéséhez minden támogatást megkaptam vezetőimtől, különösen PRÉKOPA ANDRÁStól, az MTA rendes tagjától.

IRODALOM

[1] ANDERSON, T. W., *An Introduction to Multivariate Statistical Analysis* (Wiley, New York, 1958).
[2] ARANY, I., „Nagyméretű, ritka, szimmetrikus mátrixok hatékony számítógépes kezelése", *Alk. Mat. Lapok* **11** (1986) 1—90.
[3] BERNAU, H., „Felső korlát technikák a kvadratikus programozáshoz", *Alk. Mat. Lapok* **3** (1977) 161—170.
[4] CRAMÉR, H., *Mathematical Methods of Statistics* (Princeton University Press, 1946).
[5] CZEIZEL, A., TELEGDI, L., TUSNÁDY, G., *Multiple Congenital Abnormalities* (Akadémiai Kiadó, Budapest, 1988, nyomdában).
[6] CZEIZEL, A., TUSNÁDY, G., *Aetiological Studies of Isolated Common Congenital Abnormalities in Hungary* (Akadémiai Kiadó, Budapest, 1984).
[7] DEÁK, I., *Véletlenszám-generátorok és alkalmazásuk* (Akadémiai Kiadó, Budapest, 1986).
[8] MARDIA, K. V., KENT, J. T., BIBBY, J. M., *Multivariate Analysis* (Academic Press, New York, 1979).
[9] SERFLING, R. J., *Approximation Theorems of Mathematical Statistics* (Wiley, New York, 1980).
[10] SIMONOVITS, M., TELEGDI, L., TUSNÁDY, G., "Classification of objects via multidimensional scaling of variables", in: *COMPSTAT 1982, Proceedings in Computational Statistics, Part II* Ed. H. Caussinus, P. Ettinger, J. R. Mathieu (Physica-Verlag, Wien, 1982) 243—244.
[11] TAQQU, M. S., „Law of the iterated logarithm for sums of nonlinear functions of Gaussian variables that exhibit a long range dependence", *Z. Warscheinlichkeitstsh. und verw. Gebiete* **40** (1977) 203—238.
[12] TELEGDI, L., „Multiple multidimensional scaling: a new approach to the analysis of multidimensional contingency tables with application to congenital abnormalities", *Metron* **40** (1982) 277—288.
[13] TELEGDI, L., „Veleszületett rendellenességek függetlenségének vizsgálata", *Alk. Mat. Lapok* **9** (1983) 421—436.
[14] TELEGDI, L., „Többdimenziós skálázás", in: *Többváltozós statisztikai analízis* Szerk. Móri F. T., Székely J. G. (Műszaki Könyvkiadó, Budapest, 1986) 233—250.
[15] TELEGDI, L., SIMONOVITS, M., "Initialization of multiple multidimensional scaling", in: *Contributed Papers of the 44th Session of the International Statistical Institute, Volume* 2 Ed. D. G. Horvitz, J. L. Sánchez-Crespo (International Statistical Institute, Madrid, 1983) 575—578.
[16] TUSNÁDY, G., „Keverékek felbontása", *Mat. Lapok* **30** (1978—1982) 59—67.
[17] TUSNÁDY, G., TELEGDI, L., CZEIZEL, E., „Gyakori veleszületett rendellenességek öröklődésmenetének vizsgálata", *Alk. Mat. Lapok* **4** (1978) 1—25.
[18] VINCZE, I., *Matematikai statisztika ipari alkalmazásokkal* (Műszaki Könyvkiadó, Budapest, 1968).

(Beérkezett: 1987. január 9.)

TELEGDI LÁSZLÓ
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZETE
1111 BUDAPEST, KENDE U. 13—17.

1. TÁBLÁZAT

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

INVESTIGATION OF THE STRUCTURE OF BINARY VARIABLES

L. TELEGDI

The paper deals with the statistical investigation of binary random variables observed on a large number of objects and indicating the presence or absence of dichotomous characters. The *independence* of the variables is investigated in the case when the number of variables is large and/or the probabilities of the various characters are small. If the variables are not independent, their dependence is characterized by the *Gaussian threshold model*. This assumes that background variables with a joint normal distribution are assigned to the various variables. There are thresholds belonging to them, furthermore one of the objects has one of the characters if and only if the value of the corresponding background variable on the object exceeds the corresponding threshold. A new method for clustering the objects is reported. It is called *multiple multidimensional scaling*, and puts the objects into clusters by ordinary multidimensional scaling of the variables. Finally the paper deals with an important biomedical application of computer, the statistical investigation of *congenital abnormalities*.

Alkalmazott Matematikai Lapok **13** (1987—88) 43—56

PRIMA: ÚJ ELLENŐRZÖTT OSZTÁLYOZÓ MÓDSZER

JURICSKAY ISTVÁN, VERESS GÁBOR E.

Pécs Budapest

Az általunk javasolt új ellenőrzött osztályozó módszer, a PRIMA módszer az osztálytávolság fogalmán alapszik. Lényege az, hogy az osztálysúlypontok és osztályinhomogenitások alapján minden osztályhoz külön távolságfogalmat, az osztálytávolság fogalmát rendeljük és az osztályozás ezen távolságok alapján történik.

Az osztálytávolság fogalmán alapuló PRIMA osztályozó módszer igen sokféle, bonyolult esetben is alkalmazható. Az alkalmazás feltételei más módszerekével összevetve kevésbé szigorúak. Algoritmusa igen egyszerű, hatékonysága és stabilitása jónak bizonyult. Egyszerűsége, ugyanakkor a bonyolult esetekre, mint pl. nagyon sok változós feladatokra, vagy sok osztályba sorolásra, zajos és hiányos adatok feldolgozására stb. való alkalmazhatósága miatt a meglevő hatékony alakfelismerési módszerek mellett is figyelemre méltó.

1. Bevezetés

Jelen közleményben új alakfelismerő ellenőrzött osztályozó módszerünk, a *Pattern Recognition by Independent Multicategory Analysis*, PRIMA módszer matematikai alapjait ismertetjük. Előző közleményünkben [5] elsősorban a módszer analitikai kémiai alkalmazásait tárgyaltuk.

A többváltozós statisztikai módszerek fontos csoportjának, az alakfelismerő *(pattern recognition)* módszereknek (pl. [7]) ma még nem alakult ki az egységes magyar nevezéktana. Hívják alakfelismerésnek [10], osztályozási módszereknek [9], de klasszifikációs módszereknek is [3].

Az alakfelismerő módszerek fontos csoportját alkotják az ellenőrzött osztályozó *(supervised classification)*, vagy más néven tanítóval történő tanuló *(supervised training)* módszerek. Ismertetésünknél feltételezzük az alapelvek ismeretét, hivatkozva a fontosabb irodalmi összefoglalókra (pl. [7], [10]).

Tudomásunk szerint módszerünk elve új, csak a SIMCA módszer [11] alapszik hasonló elveken, ezért dolgozatunkban külön kitérünk e két módszer összehasonlítására.

2. A PRIMA módszer elve

2.1. *Alapfogalmak*

Az *objektumok* indexe legyen i, $i=1, \ldots, I$. Az objektumok ismert, vagy ismeretlen *osztályokhoz* tartoznak, az osztályok indexe legyen k, (bizonyos esetekben l), k, ill. $l=1, \ldots, K$.

Az objektumokat *tulajdonságaikkal* jellemezzük, értelmezzük a J-dimenziós

X tulajdonságteret. A k-adik osztály i-edik objektuma j-edik tulajdonságának értékét jelölje x_{ij}^k, ha az osztály nem ismert, vagy tetszés szerinti, akkor x_{ij}, ha jelöljük, hogy az osztály nem ismert, akkor x_{ij}^*. A tulajdonságértékekből alkotott $\mathbf{x}_i$ vektort tulajdonságvektornak nevezzük, azaz

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iJ} \end{pmatrix}, \quad i = 1, \ldots, I,$$

ahol $\mathbf{x}_i \in X$. Minden objektum tehát úgy értelmezhető, mint a tulajdonságtérnek a tulajdonságvektorral megadott pontja.

Az objektumok két csoportra oszthatók: egyrészt ismert osztályú objektumok, amelyek a tananyag, vagy az ellenőrző anyag objektumai lehetnek, másrészt ismeretlen osztályú, vagy osztályozandó objektumok. Ennek megfelelően a tananyagból értelmezzük az $\mathbf{X}^L$ tananyag-mátrixot:

$$\mathbf{X}^L = \begin{Bmatrix} \begin{pmatrix} x_{11}^1, & \ldots, x_{1J}^1 \\ \vdots & \\ x_{I^1 1}^1, & \ldots, x_{I^1 J}^1 \end{pmatrix} \\ \vdots \\ \begin{pmatrix} x_{11}^k, & \ldots, x_{1J}^k \\ \vdots & \\ x_{I^k 1}^k, & \ldots, x_{I^k J}^k \end{pmatrix} \\ \vdots \\ \begin{pmatrix} x_{11}^K, & \ldots, x_{1J}^K \\ \vdots & \\ x_{I^K 1}^K, & \ldots, x_{I^K J}^K \end{pmatrix} \end{Bmatrix},$$

ahol I^k, $k=1, \ldots, K$ a k-adik osztály objektumainak a száma. Teljesen hasonló módon értelmezzük az $\mathbf{X}^T$ ellenőrzőanyag-mátrixot. Értelmezzük továbbá az $\mathbf{X}^*$ osztályozandó-anyagmátrixot az alábbi módon:

$$\mathbf{X}^* = \begin{pmatrix} x_{11}^*, & \ldots, x_{1J}^* \\ \vdots & \\ x_{I^* 1}^*, & \ldots, x_{I^* J}^* \end{pmatrix},$$

ahol I^*az osztályozandó objektumok száma.

2.2. *Az osztályok súlypontja és inhomogenitása*

Az osztályokat a tanulás során két paraméterrel, a súlypontjukkal és az inhomogenitásukkal jellemezzük.

Az osztályok tulajdonságtérbeli elhelyezkedését az osztályok súlypontjával jellemezhetjük. Adott *osztály* adott *tulajdonságának átlaga*, $\bar{x}_j^k$ a tulajdonságok

egyszerű átlagával számolható, azaz

$$\bar{x}_j^k = \frac{\sum_{i=1}^{I^k} x_{ij}^k}{I^k} \qquad j = 1, \ldots, J, \quad k = 1, \ldots, K.$$

Adott osztály adott tulajdonságának az átlagát a tananyagmátrixból az osztályt jellemző részmátrix adott oszlopának az összeadásával nyerjük. Adott *osztály súlypontja*, $\bar{\mathbf{x}}^k$, vagy centruma a tulajdonságtérben a tulajdonságátlagokból képezett vektor, azaz

$$\bar{\mathbf{x}}^k = \begin{pmatrix} \bar{x}_1^k \\ \vdots \\ \bar{x}_J^k \end{pmatrix} \qquad k = 1, \ldots, K.$$

Az osztályok tulajdonságtérbeni kiterjedtségét a tulajdonságok inhomogenitásával vagy szórásával jellemezhetjük. Adott *osztály* adott *tulajdonságának inhomogenitása* s_j^k az adott tulajdonság szórásaként számolható, azaz

$$s_j^k = \left(\frac{\sum_{i=1}^{I^k} (x_{ij}^k - \bar{x}_j^k)^2}{I^k - 1} \right)^{1/2} \qquad j = 1, \ldots, J, \quad k = 1, \ldots, K.$$

Az inhomogenitással egyenértékű fogalomként értelmezhetjük az adott *osztály* adott *tulajdonságának súlyát* w_j^k-t, mint az osztály tulajdonság-inhomogenitás négyzetének reciprokát, azaz

$$w_j^k = \frac{1}{(s_j^k)^2};$$

Adott *osztály inhomogenitás-vektora* $\mathbf{s}^k$, vagy kiterjedtsége a tulajdonságtérben a tulajdonság-inhomogenitásokból képezett vektor, azaz

$$\mathbf{s}^k = \begin{pmatrix} s_1^k \\ \vdots \\ s_J^k \end{pmatrix}, \qquad k = 1, \ldots, K.$$

Analóg módon bevezetjük az *osztály súlyvektorát* $\mathbf{w}^k$, amely a tulajdonságsúlyokból képezett vektor, azaz

$$\mathbf{w}^k = \begin{pmatrix} w_1^k \\ \vdots \\ w_J^k \end{pmatrix}, \qquad k = 1, \ldots, K.$$

Megjegyezzük, hogy a fenti fogalmak hiányos adatok esetén is a rendelkezésre álló adatok alapján egymástól függetlenül egyszerűen számolhatók, természetesen a hiányzó adatokkal csökkentve az adatszám I^k értékét.

2.3. *Osztálytávolság*

A tulajdonságtér pontjai között sokféle távolság értelmezhető. Módszerünk az ún. „osztálytávolság" fogalmán alapszik. Minden osztályra vonatkozóan külön-külön értelmezzük a vizsgált objektum és az adott osztály centrumának a távolságát. Az osztálytávolság definiálására bármilyen távolságfogalom alkalmas, az alábbiakban az *euklideszi távolságot* használjuk.

A módszerünk alapfogalmát jelentő osztálytávolság fogalma előtt bevezetjük az osztálytávolság ún. tulajdonságkomponensének fogalmát. A k-adik osztályra vonatkozóan az *osztálytávolság j-edik tulajdonságkomponense* a

$$d_j^k(x_{ij}, \bar{x}_j^k) = \frac{1}{s_j^k}(x_{ij} - \bar{x}_j^k), \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, K$$

összefüggéssel értelmezhető.

Az osztálytávolság tulajdonságkomponensei alapján értelmezhetjük módszerünk alapvető fogalmát, az ún. *osztálytávolság* fogalmát

$$d^k(\mathbf{x}_i, \bar{\mathbf{x}}^k) = \left(\frac{1}{J}\sum_{j=1}^{J}[d_j^k(x_{ij}, \bar{x}_j^k)]^2\right)^{1/2}, \qquad i = 1, \ldots, I, \quad k = 1, \ldots, K,$$

vagyis az osztálytávolság négyzete a tulajdonságkomponensek négyzetének átlaga. Figyelembe véve a tulajdonságkomponensek definícióját, az osztálytávolság a

$$d^k(\mathbf{x}_i, \bar{\mathbf{x}}^k) = \left(\frac{1}{J}\sum_{j=1}^{J}\frac{1}{(s_j^k)^2}(x_{ij} - \bar{x}_j^k)^2\right)^{1/2}, \qquad i = 1, \ldots, I, \quad k = 1, \ldots, K,$$

vagy ezzel analóg módon a

$$d^k(\mathbf{x}_i, \bar{\mathbf{x}}^k) = \left(\frac{1}{J}\sum_{j=1}^{J} w_j^k(x_{ij} - \bar{x}_j^k)^2\right)^{1/2}, \qquad i = 1, \ldots, I, \quad k = 1, \ldots, K,$$

alakban is írható.

Egy pont egy adott osztálytávolsága tehát a pontnak az adott osztálysúlyponttól mért, az osztály súlyvektorával súlyozott, egy tulajdonságra vonatkoztatott átlagos távolsága.

Az osztálytávolság úgy is értelmezhető, hogy a tulajdonságtérben minden osztály számára külön-külön olyan koordinátarendszert értelmezünk, amelynek origója az osztálysúlypont és egysége az osztály tulajdonságinhomogenitása. Ilyen értelmezésben az osztálytávolság az osztályonkénti z transzformáció, vagy *„autoscaling"* alapján is értelmezhető, azaz

$$d^k(\mathbf{x}_i, \bar{\mathbf{x}}^k) = \left(\frac{1}{J}\sum_{j=1}^{J}(z_{ij}^k)^2\right)^{1/2}, \qquad i = 1, \ldots, I, \quad k = 1, \ldots, K,$$

ahol

$$z_{ij}^k = \frac{x_{ij} - \bar{x}_j^k}{s_j^k}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, K,$$

az x_{ij} tulajdonságérték k-adik osztálybeli z-transzformáltja, vagy standardizált értéke.

A fenti távolságfogalom elvben csak az $s_j^k \neq 0$ esetben értelmezhető, a gyakorlati alkalmazhatóság érdekében olyan speciális esetekben, amikor $s_j^k = 0$, a feladat természetétől (mérési hiba stb.) függő, zérustól eltérő önkényes értéket célszerű használni.

A fenti távolságfogalom, mint átlagos távolság hiányos adatok esetén is alkalmazható, ekkor J értékét értelemszerűen csökkenteni kell. Az $\frac{1}{J}$ normálási tényező éppen azért szerepel a definícióban, hogy a távolságok a dimenziószámtól függetlenül is összehasonlíthatók maradjanak.

2.4. *A PRIMA módszer matematikai statisztikai értelmezése*

Ahhoz, hogy a döntéselméletileg igazoltan legjobb osztályozást, a Bayes-döntést alkalmazhassuk, nemcsak az egyes osztályok előfordulási valószínűségeit, hanem minden osztály minden tulajdonságra megadott feltételes valószínűségeit is ismernünk kellene. A pontmódszerek (KNN, potenciálfüggvényes módszer) ezt azzal közelítik, hogy minden egyes tananyagobjektumból nyerhető információt azonos súllyal vesznek figyelembe. Bebizonyítható, hogy ennek a közelítésnek a tárkapacitás és a gépidőigény szab határt. A súlypont-módszerek ezeket a valószínűségeloszlásokat egyetlen adattal, az osztály súlypontjával jellemzik.

Az ismertetett PRIMA módszer a valódi eloszlások közelítésére az inhomogenitást is figyelembe veszi, annak megfelelelően, mintha a valódi eloszlások szimmetrikus eloszlások lennének. Az átlagok az illető tulajdonságok *várható értékeinek becsléseként* foghatók fel, és az inhomogenitás mértéke (a korrigált tapasztalati szórás) a *második centrális momentum becslése*. E becslések jogosságára és jóságára az egyváltozós matematikai statisztikai módszerek szabják meg az ismert kritériumokat. A transzformált tulajdonságtér változóiról elmondhatjuk, hogy ezek várható értéke zérus, ha az illető objektum az osztályba tartozik. Az így definiált osztálytávolság fogalmunk alapján tehát jó becslést kapunk.

Az eloszlások jellemzésére további momentumok figyelembevételével a *Bayes döntés* egyre jobb közelítéseit kaphatnánk meg. Ez olyan torzító transzformációnak felelne meg, ahol az osztálytávolság fogalom anizotrop (irányfüggő) lenne, ami éppen a lényegkiemelés elve. Ennek részleteivel itt nem foglalkozunk.

3. Osztályozás a PRIMA módszerrel

3.1. *Tanulás*

A PRIMA módszerrel történő tanulás folyamata az $\mathbf{X}^L$ tananyag mátrixból az $\bar{\mathbf{x}}^1, \ldots, \bar{\mathbf{x}}^K$ osztálysúlypontok és az $\mathbf{s}^1, \ldots, \mathbf{s}^K$ osztályinhomogenitás-vektorok, illetve az $\mathbf{w}^1, \ldots, \mathbf{w}^K$ osztálysúlyvektorok előállítása.

3.2. *Felismerés*

A PRIMA módszerrel történő felismerés lépése kettős. A felismerés első lépése a felismerendő objektumot reprezentáló $\mathbf{x}^*$ pont $d^1(\mathbf{x}^*, \bar{\mathbf{x}}^1), \ldots, d^K(\mathbf{x}^*, \bar{\mathbf{x}}^K)$ osztálytávolságainak a meghatározása.

A felismerés második lépése az osztálytávolságok alapján annak az eldöntése, hogy az ismeretlen osztályú objektum, amely a tulajdonságtérben az $\mathbf{x}^*$ pont, mely osztályba vagy osztályokba tartozik. Az osztály meghatározására több lehetőség kínálkozik.

Az objektum osztályba sorolásának egyik elvi lehetősége az, hogy az objektumot abba az osztályba soroljuk, amelynek a legkisebb az osztálytávolsága, azaz a

$$\min_k \{d^1, \ldots, d^K\}$$

feltételt kielégítő k-adik osztályt tekintjük az objektum osztályának.

Az objektum osztályba sorolásának egy másik, sokkal inkább gyakorlati lehetősége az, hogy az objektumot a k-adik osztályba tartozónak tekintjük, ha teljesül a

$$d^k(\mathbf{x}^*, \bar{\mathbf{x}}^k) \leqq d$$

feltétel, ahol d alkalmasan választott ún. osztálytávolság-küszöbszám. E feltétel használata esetén egy objektum természetesen több osztályba is tartozhat, illetve kívülálló is lehet. Az osztálybatartozást eldöntő osztálytávolság-küszöbszám meghatározása a feladat természetétől függően célszerűen a felismerési képesség alapján iteratív úton történhet.

4. Az osztályozás jellemzése

4.1. *Az osztályok közötti távolság*

Az osztálytávolság fenti definíciójából nyilvánvaló az, hogy az egyik osztály súlypontjának a távolsága egy másik osztály súlypontjától az egyik osztály súlyvektorától függ, így az egyik osztály távolsága a másik súlypontjától általában nem egyezik meg a másik osztály távolságával az egyik súlypontjától, azaz a

$$d^k(\bar{\mathbf{x}}^l, \bar{\mathbf{x}}^k) = \left(\frac{1}{J}\sum_{j=1}^{J}\frac{1}{(s_j^k)^2}(\bar{x}_j^l - \bar{x}_j^k)^2\right)^{1/2}$$

és a

$$d^l(\bar{\mathbf{x}}^k, \bar{\mathbf{x}}^l) = \left(\frac{1}{J}\sum_{j=1}^{J}\frac{1}{(s_j^l)^2}(\bar{x}_j^k - \bar{x}_j^l)^2\right)^{1/2}$$

távolságok általában nem egyeznek meg. Értelmezzük ezért az osztályok közötti távolságot.

Az osztályok közötti távolság értelmezéséhez vezessük be az s_j^{kl} fogalmat, amelyet a j-edik tulajdonságnak a k-adik és az l-edik osztályokra vonatkoztatott *átlagos inhomogenitásának* nevezzük és az

$$\frac{1}{(s_j^{kl})^2} = \frac{1}{2}\left(\frac{1}{(s_j^k)^2} + \frac{1}{(s_j^l)^2}\right), \qquad j = 1, \ldots, J, \quad k, l = 1, \ldots, K,$$

azaz az

$$(s_j^{kl})^2 = \frac{2(s_j^k)^2 \cdot (s_j^l)^2}{(s_j^k)^2 + (s_j^l)^2} \qquad j = 1, \ldots, J, \quad k = 1, \ldots, K,$$

összefüggéssel definiálunk.

Az átlagos inhomogenitás alapján az osztálytávolsághoz hasonlóan itt is értelmezhetjük az *osztályok-közötti-távolság tulajdonságkomponenseit* a

$$d_j^{kl}(\bar{x}_j^l, \bar{x}_j^k) = \frac{1}{s_j^{kl}}(\bar{x}_j^l - \bar{x}_j^k) \qquad j = 1, \ldots, J, \quad k, l = 1, \ldots, K,$$

összefüggéssel, vagy az átlagos inhomogenitás definíciójának a figyelembevételével a

$$d_j^{kl}(\bar{x}_j^l, \bar{x}_j^k) = \left(\frac{1}{2}\left(\frac{1}{(s_j^k)^2} + \frac{1}{(s_j^l)^2}\right)\right)^{1/2}(\bar{x}_j^l - \bar{x}_j^k), \qquad j = 1, \ldots, J, \quad k, l = 1, \ldots, K,$$

összefüggéssel.

Megjegyezzük, hogy a *Fischer-hányados*

$$F_j^{kl} = \frac{1}{s_j^k + s_j^l}(\bar{x}_j^l - \bar{x}_j^k) \qquad j = 1, \ldots, J, \quad k, l = 1, \ldots, K,$$

szintén speciális osztályok-közötti-távolság tulajdonságkomponensének tekinthető.

Az osztályok súlypontjai között sokféle *osztályok-közötti-távolságot* értelmezhetünk. Az osztályok-közötti-távolság értelmezésének egyik lehetősége a

$$d^{kl} = (\bar{\mathbf{x}}^l, \bar{\mathbf{x}}^k) = \left(\frac{1}{J}\sum_{j=1}^{J}\frac{1}{2}\left(\frac{1}{(s_j^k)^2} + \frac{1}{(s_j^l)^2}\right)(\bar{x}_j^l - \bar{x}_j^k)^2\right)^{1/2}, \qquad k, l = 1, \ldots, K,$$

összefüggés, vagy átírva

$$d^{kl}(\bar{\mathbf{x}}^l, \bar{\mathbf{x}}^k) = \left(\frac{1}{J}\sum_{j=1}^{J}\frac{1}{(s_j^{kl})^2}(\bar{x}_j^l - \bar{x}_j^k)^2\right)^{1/2} \qquad k, l = 1, \ldots, K.$$

4.2. *A tulajdonságok jellemzése*

Az osztályok-közötti-távolság tulajdonságkomponense jellemző az adott tulajdonság jelentőségére. Az osztályok-közötti-távolság tulajdonságkomponenseiből felépíthetők az osztályok-közötti-távolság tulajdonságkomponens mátrixai, a

$$\mathbf{D}_j = \begin{pmatrix} d_j^{11}, \ldots, d_j^{1K} \\ \vdots \\ d_j^{K1}, \ldots, d_j^{KK} \end{pmatrix} \qquad j = 1, \ldots, J$$

mátrixok, melyek antiszimmetrikusak és diagonális elemei zérusok.

Ezen mátrixok segítségével jellemezhető az osztályozás szempontjából az adott tulajdonság jelentősége, ugyanis, ha a d_j^{kl} abszolút értéke nagy, akkor a j-edik tulajdonság jelentős a k és az l osztályok közötti megkülönböztetés szempontjából, míg, ha d_j^{kl} abszolút értéke kicsi minden k, l osztálypárra, akkor a j-edik tulajdonság nem jelentős, esetleg elhagyható.

4.3. *Az osztályok jellemzése*

Az osztályok-közötti-távolság fogalma alapján jellemezhető az osztályok elhelyezkedése a tananyagban. Az osztályok-közötti-távolság értékekből, azaz a d^{kl} értékekből felépíthető a **D** osztályok-közötti-távolság-mátrix, azaz

$$\mathbf{D} = \begin{pmatrix} d^{11}, \ldots, d^{1K} \\ \vdots \\ d^{K1}, \ldots, d^{KK} \end{pmatrix}.$$

Ez a mátrix szimmetrikus és a diagonális elemei zérusok.

Az osztályok-közötti-távolság-mátrix alkalmas az osztályok elhelyezkedésének a jellemzésére. Azok az osztályok, amelyekre a d^{kl} távolság nagy, a tulajdonságtérben jól szeparáltak, így a téves osztályozás valószínűsége kicsi. Ha a d^{kl} távolság kicsi, akkor az osztályok a tulajdonság-térben nem eléggé különülnek el.

4.4. *Az osztályozás jellemzése a tananyaggal*

Az osztályonkénti felismerőképesség megadja osztályonként azt, hogy a tananyag objektumainak hányadrésze kerül a helyes osztályba, azaz

$$r^k = \frac{f^k}{l^k}, \qquad k = 1, \ldots, K,$$

ahol f^k a tananyag k-adik osztályba sorolt és a k-adik osztályba tartozó objektumainak a száma. A fenti osztályonkénti felismerőképesség mellett értelmezhető a teljes tananyagra vonatkozó teljes felismerőképesség is, pl. az

$$r = \frac{\sum_{k=1}^{K} f^k}{\sum_{k=1}^{K} l^k}$$

összefüggés alapján.

A felismerőképesség az osztályok elkülönültsége mellett a tanulás jóságát is jellemzi, így például alkalmas mutató az osztálytávolság-küszöbérték iteratív módon történő meghatározására.

4.5. *Az osztályozás jellemzése az ellenőrző anyaggal*

Az alakfelismerő módszerek osztályozó hatékonysága a felismerőképesség mellett az előrebecslőképességgel is jellemezhető. Az osztályonkénti előrebecslőképesség azt adja meg, hogy az ellenőrző anyag objektumainak hányad része kerül a helyes osztályba, azaz

$$p^k = \frac{f^{kT}}{l^{kT}}, \qquad k = 1, \ldots, K,$$

ahol $\hat{l}^{kT}$ az ellenőrző anyag k-adik osztályba sorolt, k-adik osztályba tartozó objektumainak a száma, míg l^{kT} az ellenőrző anyag k-adik osztályba tartozó objektumainak a száma. A fenti osztályonkénti előrebecslőképesség mellett értelmezhető a teljes ellenőrző anyagra vonatkozó teljes előrebecslőképesség is, pl. a

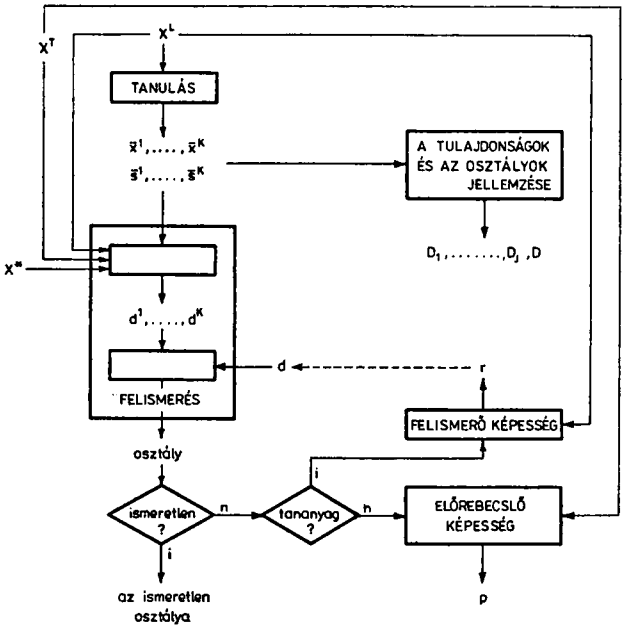
$$p = \frac{\sum_{k=1}^{K} \hat{l}^{kT}}{\sum_{k=1}^{K} l^{k}}$$

összefüggés alapján.

5. A PRIMA algoritmus

Az előzőekben ismertetett alapfogalmak felhasználásával új ellenőrzött osztályozó algoritmus készíthető. Az algoritmus folyamatábráját az 1. ábra mutatja. Az algoritmus az osztályozásból és az osztályozás jellemzéséből áll.

Az ellenőrzött osztályozó módszerek általános elvének megfelelően az osztályozás kétlépcsős. Először a tanulás során a tananyag segítségével meghatározzuk az osztályokat jellemző súlypontvektorokat és inhomogenitásvektorokat. A felismerés, vagy osztályozás a második lépés. Az osztályozandó objektum tulajdonságvektorát az egyes osztályok standardizált tulajdonságterébe transzformáljuk és ennek



1. ábra. A PRIMA algoritmus folyamatábrája

felhasználásával minden osztályra meghatározzuk az objektum osztálytávolságait, melyek felhasználásával az osztálytávolság-küszöb-érték alapján elvégezzük az osztályozást.

Az algoritmus a tulajdonságok és az osztályok jellemzése mellett elvégzi a felismerőképesség és az előrebecslőképesség meghatározását is. A felismerőképesség alapján mód nyílik az osztálytávolság-küszöbérték iteratív meghatározására.

A PRIMA módszer algoritmusa alapján készített program jelenleg VIDEOTON 1010 mini-számítógépre FORTRAN és PROPER—16 személyi számítógépre, illetve ROSY 80 és SINCLAIR ZX—81 mikroszámítógépekre BASIC nyelven áll rendelkezésre.

6. A PRIMA módszer értékelése

6.1. *A módszer alkalmazásai*

A PRIMA módszerrel számos osztályozási feladatot oldottunk meg. Előző közleményünkben [5] részletesen ismertettük az analitikai kémia irodalmában ismert IRIS feladat [2], továbbá egy konformációs feladat [8] megoldását. Bemutattuk a PRIMA módszer alkalmazhatóságát aszfaltinanyagok érettségének vizsgálatára is, ahol 17 tulajdonság alapján 5 osztályba soroláskor jónak nevezhető, 90%-os felismerőképességet kaptunk [1]. A PRIMA módszert sikerrel alkalmaztuk almalevek minősítésére is [6], továbbá régészeti feladatok megoldására.

A PRIMA módszert több orvosdiagnosztikai feladat megoldására is alkalmaztuk. A vizelet gázkromatográfiával meghatározott 12 szteroidkomponense alapján beteg gyerekeket 8 osztályba soroltunk. A felismerőképesség ennél a feladatnál 94%-nak adódott [4].

Megjegyezzük, hogy a PRIMA módszer alkalmas bizonyos képfelismerési feladatok megoldására is. Ezzel kapcsolatban egy 35 komponensű bináris tulajdonságot adó mozaik betűfelismerési feladat megoldása során nyert eredményeinek publikációját tervezzük.

Eddigi tapasztalataink szerint a PRIMA módszer alkalmazási kritériumai az ismert ellenőrzött osztályozó módszereknél kevésbé szigorúak, így az alkalmazási feladatok igen nagy részében kielégíthetők.

6.2. *Kívülállók, több osztályba tartozó objektumok*

Az osztálytávolság definícióján alapuló módszerek (SIMCA, PRIMA) egyik legnagyobb előnye a többivel szemben az, hogy az osztálytávolságok terébe képezik le az objektumokat, és alkalmasan választott távolságkritériumokkal a tulajdonságok terét olyan térrészekre osztják, melyekbe eső objektumokról egyértelműen eldönthető, hogy az valamely tanult osztály eleme, esetleg több osztályba is beletartozhat, vagy az osztályokon kívüleső. Ha az osztályok nem kölcsönösen kizáróak és együttesen nem teljesek, akkor csak ezek a módszerek alkalmazása lehet indokolt.

6.3. *Folytonos és diszkrét változók*

Az osztályozási feladatok megoldása során sok problémát jelent az olyan tulajdonságvektorok kezelése, melyek komponenseinek egy része folytonos, más része diszkrét (speciális esetben bináris, „igen" — „nem") jellegű változókat tartalmaz.

Az alapvetően folytonos változókra kidolgozott PRIMA módszer tisztán bináris változókra való alkalmazása során azt tapasztaltuk, hogy nagymértékben hiányos és zajos adatkészlet esetén is jó hatásfokú osztályozást biztosít, ha gondoskodunk arról, hogy az osztályjellemző paraméterek között az inhomogenitások zérustól különbözzenek.

A folytonos és diszkrét (bináris) változók szimultán kezelhetősége véleményünk szerint a PRIMA egyik fontos érdeme, de konkrét feladatok esetén külön mérlegelendő a folytonos és a diszkrét komponensek számszerű aránya, mivel ilyen különleges alkalmazásokról eddig kevés tapasztalat áll rendelkezésre.

6.4. *Hiányos adatrendszer*

Az alkalmazási feladatok jelentős hányadában nem biztosítható az adatmátix, ill. a felismerendő objektum tulajdonságvektorának hiánytalan ismerete. A probléma olyan áthidalása, hogy a tanulásból, illetve az értékelésből kihagyjuk az objektumot, vagy a tulajdonságot, nyilvánvaló információveszteséggel jár. Az adathiányra legérzékenyebbek a mátrix-számítást alkalmazó módszerek (diszkriminancia-analízis, főkomponens-analízis), mivel a mátrixműveletek (inverzió, transzponálás stb.) teljes mátrixot igényelnek. A hiányzó adatok betöltése az átlaggal, vagy interpolált adatokkal csak kevéssé hiányos adatkészlet esetén engedhető meg. A betöltés ilyen módja nagy felelősséget igényel, mivel a módszerek könnyen bizonytalanná válnak, így pl. a korrelációk, varianciák értelmüket vesztik. Ha a hiányos adat éppen a felismerendő objektumnál fordult elő, akkor a koordináták közös transzformációján alapuló módszerek (diszkriminancia analízis, főkomponens analízis) egyike sem alkalmazható az adat önkényes betöltése nélkül.

A PRIMA módszer osztálytávolság fogalma lehetővé teszi, hogy az osztályjellemzők újratanítása nélkül egy alacsonyabb dimenziószámú altérben képezzük az osztálytávolságokat és ezek, mivel normáltak, összehasonlíthatók a teljes adatkészlettel nyert hasonló adatokkal. A hiányos adatokkal történő osztályozásnak különösen fontos szerepe van olyan esetekben, amikor az esetleg hosszadalmas, vagy igen költséges tulajdonságmeghatározó mérések elvégzése előtt már elvégezzük az osztályozást és az eredménytől függően döntünk a tényleges mérés végrehajtásáról. E speciális képességnek, az ún. hierarchikus osztályozásnak diagnosztikai szűrő jellegű, esetleg képfelismerő feladatoknál lehet jelentősége.

6.5. *Kiegészítő adatok*

Egészen speciális adathiányra vezet, ha a kutatás során a tulajdonságok mérésében alkalmazott módszereket megváltoztatjuk, vagy új tulajdonságot vezetünk be. A legtöbb osztályozási módszer esetén ilyenkor az eddig összegyűlt összes információt kénytelenek vagyunk törölni és az új, esetleg másikat helyettesítő tulajdonságot

csak új tananyaggal vezethetjük be az osztályozási feladatba. A módszerek nagy része tehát az adatrendszer szempontjából merev, az adatmátrix sem oszlopaiban (tulajdonságok), sem soraiban (tananyag-objektumok) nem bővíthető.

A PRIMA módszer alkalmazásakor az adatok közé új tulajdonságok bevezetése csak azt kívánja meg, hogy néhány új tanítóobjektum esetén meghatározzuk az új tulajdonság osztályátlagait és inhomogenitásait. A további objektumok osztályozása már az új osztályjellemzők betoldásával végezhető el, ugyanakkor a korábbi ismeretek megőrizhetők. A tanítás algoritmusa lehetővé teszi, hogy egy-egy objektumot az eddigi tananyaghoz bármikor hozzáfűzzünk. Ezáltal folyamatosan tanuló adaptív algoritmus építhető fel, az adatmátrix sorai is folyamatosan bővíthetők, s nem kell az egész tanulási folyamatot újra kezdeni. Ebből az is következik, hogy a többször előforduló „kívülálló" objektumokat esetleg új osztály képzésére is felhasználhatjuk, vagyis az osztályok száma is bővíthető. A PRIMA ezen különleges sajátságát főként a viszonylag kis objektumszámú, vagy folyamatszabályozási jellegű alkalmazásoknál használhatjuk ki.

6.6. *Számítástechnikai jellemzők*

Az osztályozási feladatok megoldását a számítógép memóriakapacitása, vagy a viszonylag nagy gépidőigény gyakran nehezíti.

A pontmódszerek esetén a memóriaigény közelítőleg az $I \cdot J$ szorzattal becsülhető, ahol I az objektumok száma, J a figyelembe vett lényeges tulajdonságok száma. A felismerési fázisban, mivel I darab távolságszámítást és ebből minimumkeresést kell végrehajtani, a gépidő igény is nagy, $I \cdot J$-vel arányos.

A változók közös transzformációján alapuló módszerek memóriaigénye még kibővül a $J \cdot J$ méretű mátrixok kezeléséhez szükséges tárkapacitással. Ez alól kivétel a SIMCA módszer, ahol az osztályonként végzett főkomponensanalízis nem igényli valamennyi tananyag-objektum elérhetőségét. A gépidőigény zömét a tanítási algoritmus konvergálási sebessége és a mátrixműveletek adják, a felismerési algoritmusok gépideje ehhez képest elhanyagolható.

A PRIMA módszer memóriaigénye a $J \cdot K$ érték néhányszorosára tehető, ahol K az osztályok száma, amely általában töredéke a többi módszer memóriaigényének. A tanulási és a felismerési folyamat gépidőigénye egyenletesen oszlik meg és szintén $J \cdot K$ értékével arányos. Mindezeket figyelembe véve a PRIMA módszerrel az osztályozás igen könnyen elvégezhető még kis személyi számítógépekkel is. Ennek illusztrálására megemlítjük, hogy az alkalmazásoknál említett diagnosztikai feladatot 16 K byte-os Sinclair ZX—18 mikroszámítógéppel oldottuk meg.

6.7. *A SIMCA és a PRIMA módszerek összehasonlítása*

Mint ismeretes, a SIMCA módszer [11] az osztályozó feladatok jelentős részének megoldására kiválóan alkalmas. Tapasztalatunk szerint a PRIMA módszer is igen sok feladat megoldására előnyösen használható.

Mind a SIMCA, mind a PRIMA módszer alapvető jellemzője az osztályonként értelmezett koordináta-rendszer, az osztálytávolság fogalma. A két módszer közötti alapvető különbség abban áll, hogy míg a SIMCA a főkomponens-térben, addig a PRIMA a tulajdonságtérben értelmezi az osztálytávolságot. Ennek megfelelően

a SIMCA figyelembe veszi a tulajdonságok közötti korrelációt is, a PRIMA csak ezek szórását. Ebből eredően a SIMCA csökkenti a tulajdonságtér dimenzióját, a PRIMA nem. Ugyanakkor a PRIMA hiányos adatok esetén is könnyen használható. A SIMCA memória és számítási igénye jelentősebb, a PRIMA igen egyszerűen kivitelezhető.

A két módszer jellemző tulajdonságai alapján összefoglalva megállapítható, hogy ha kicsi a tananyag, vagy kicsi a rendelkezésre álló számítógép, vagy hiányosak az adatok, akkor csak a PRIMA módszer alkalmazható, a SIMCA nem. Elegendően nagy tananyag és elegendően nagy számítástechnikai kapacitás esetén, ha a tulajdonságok között jelentős korreláció tételezhető fel, akkor a SIMCA módszert célszerű használni, míg ha a tulajdonságok között nem jelentős a korreláció, akkor a PRIMA módszer használata ajánlható.

A szövegben előforduló jelölések értelmezése

d	osztálytávolság-küszöbszám
$d^k(\mathbf{x}_i, \bar{\mathbf{x}}^k)$	az i-edik objektum k-adik osztálytávolsága
$d^k_j(x_{ij}, \bar{x}^k_j)$	az i-edik objektum k-adik osztálytávolságának j-edik tulajdonságkomponense
$d^{kl}(\bar{\mathbf{x}}^l, \bar{\mathbf{x}}^k)$	a k-adik és az l-edik osztályok közötti távolság
$d^{kl}_j(\bar{x}^l_j, \bar{x}^k_j)$	a k-adik és az l-edik osztályok közötti távolság j-edik tulajdonságkomponense
$\mathbf{D}_j$	az osztályok közötti távolságok j-edik tulajdonságkomponensének mátrixa
$\mathbf{D}$	az osztályok közötti távolságok mátrixa
F^{kl}_j	a j-edik tulajdonság k-adik és l-edik osztályokra vonatkozó *Fischer-hányadosa*
i	az objektumok indexe
I	az objektumok száma
I^k	a tananyag k-adik osztályba tartozó objektumainak a száma
$\hat{I}^k$	a tananyag k-adik osztályba sorolt objektumainak a száma
I^{kT}	az ellenőrző anyag k-adik osztályba tartozó objektumainak a száma
$\hat{I}^{kT}$	az ellenőrző anyag k-adik osztályba sorolt objektumainak a száma
I^*	az osztályozandó objektumok száma
j	a tulajdonságok indexe
J	a tulajdonságok száma
k	az osztályok indexe
K	az osztályok száma
l	az osztályok indexe
r	teljes felismerőképesség
r^k	a k-adik osztályra vonatkozó felismerőképesség
p	teljes előrebecslőképesség
p^k	a k-adik osztályra vonatkozó előrebecslőképesség
s^k_j	a j-edik tulajdonság k-adik osztályra vonatkozó inhomogenitása
$\mathbf{s}^k$	a k-adik osztály inhomogenitásvektora
s^{kl}_j	a j-edik tulajdonság k-adik és l-edik osztályokra vonatkozó átlagos inhomogenitása
w^k_j	a j-edik tulajdonság k-adik osztályra vonatkozó súlya
$\mathbf{w}^k$	a k-adik osztály súlyvektora
x_{ij}	az i-edik objektum j-edik tulajdonsága
x^k_{ij}	a k-adik osztály i-edik objektumának j-edik tulajdonsága
x^*_{ij}	az ismeretlen osztályba tartozó i-edik objektum j-edik tulajdonsága
$\mathbf{x}$	az objektum tulajdonságvektora
$\mathbf{x}_i$	az i-edik objektum tulajdonságvektora
$\bar{x}^k_j$	a j-edik tulajdonság átlaga a k-adik osztályban
$\bar{\mathbf{x}}^k$	a k-adik osztály súlypontja
$\mathbf{X}^L$	tananyag-mátrix
$\mathbf{X}^T$	ellenőrző-anyag-mátrix
$\mathbf{X}^*$	osztályozandó-anyag-mátrix
X	J-dimenziós tulajdonságtér
z^k_{ij}	a k-adik osztályba tartozó i-edik objektum j-edik tulajdonságának standardizált értéke

IRODALOM

[1] BRUCKNER-WEIN, A., JURICSKAY, I. és VERESS, G. E., „Aszfaltinanyagok érettségének vizsgálata IR spektroszkópiával", Kemometria '83 Konferencia, Csopak, 1983 (Előadás).
[2] FISCHER, R. A., "The use of multiple measurements in taxonomic problems", *Annales of Eugenics* 7 (1936) 179.
[3] FÜSTÖS, L., MESZÉNA, GY. és SIMONNÉ MOSOLYGÓ, N., *A sokváltozós adatelemzés statisztikai módszerei* (Akadémiai Kiadó, Budapest, 1986).
[4] JURICSKAY, I., MRS. JURICSKAY, I. and VERESS, G. E., "A new pattern recognition method for interpreting chromatographic results", Budapest Chromatography Conference, Budapest, 1983 (Lecture).
[5] JURICSKAY, I. and VERESS, G. E., "PRIMA a new pattern recognition method", *Anal. Chim. Acta* **171** (1985) 61—76.
[6] KERESZTURI-KOVÁCS, M., „Almalevek minősítése alakfelismeréssel", Dimplomamunka, Budapesti Műszaki Egyetem, Vegyészmérnöki Kar, 1985.
[7] KRISHNAIAH, P. R. and KANAL, L. N., *Classification, Pattern Recognition and Reduction of Dimensionality* (North-Holland, Amsterdam, New York, 1982).
[8] MECKE, R. and NOACK, K., "Strukturbestimmungen von ungesattigen Ketonen mit Hilfe von Infrarot- und Ultraviolett Spektren", *Chem. Ber.* 93 (1960) 210.
[9] MÓRI, F. T. és SZÉKELY, J. G. (szerkesztők), *Többváltozós statisztikai analízis* (Műszaki Könyvkiadó, Budapest, 1986).
[10] RÉVÉSZ, P. és FRITZ, J., *Az alakfelismerés statisztikus módszerei,* Jegyzet, MTA Matematikai Kutató Intézet, Budapest, 1974.
[11] WOLD, S., "Pattern recognition by means of disjoint principal components models", *Pattern Recognition* **8** (1976) 127—139.

(Beérkezett: 1985. július 1.)

JURICSKAY ISTVÁN
PÉCSI ORVOSTUDOMÁNYI EGYETEM
7643 PÉCS, IFJÚSÁG U. 13.

DR. VERESS GÁBOR E.
BUDAPESTI MŰSZAKI EGYETEM
1111 BUDAPEST, GELLÉRT TÉR 4.

PRIMA: A NEW SUPERVISED CLASSIFICATION METHOD

I. JURICSKAY, G. E. VERESS

This paper deals with the new supervised classification method called *Pattern Recognition by Independent Multicategory Analysis* (PRIMA). This method uses the concept of class distance defined on the basis of centre of gravity and inhomogeneity for each class. The PRIMA algorithm is very simple, its efficiency and stability are good. It can be applied in different, complex cases, the conditions of its applicability are less strict than those of other methods.

Alkalmazott Matematikai Lapok 13 (1987—88) 57—68

TÍPUSSPECIFIKÁCIÓK HELYESSÉGÉNEK VIZSGÁLATA

VARGA LÁSZLÓ

Budapest

Az adattípusnak azokat a tulajdonságait, amelyek a felhasználó célját szolgálják, absztrakt specifikációval adhatjuk meg. Amikor viszont megvalósítjuk az adattípust, egy megfelelő ábrázolást választunk és megadjuk az adattípus konkrét specifikációját. Így két specifikációt adunk meg ugyanarra az adattípusra. Ebben a dolgozatban a konkrét specifikáció absztrakt specifikáció szerinti helyességét fogalmazzuk meg és a kettős specifikáció három esetében vizsgáljuk a helyesség elégséges feltételét.

1. Bevezetés

Adattípusok specifikálására különböző módszereket használnak. Ezek között a legismertebbek a procedurális specifikáció [1], az elő- és utófeltételekkel megadott specifikáció [2] és az algebrai specifikáció [3]. Az utóbbi különösen alkalmas adattípusok reprezentációtól és implementációtól független definiálására, azaz absztrakt adattípusok leírására. Ennek a megállapításnak az képezi az alapját, hogy az algebrai specifikáció felírásához nincs szükségünk segédrendszerre, a segédrendszerrel kapcsolatos ismeretekre. Az elő- és utófeltételekkel megadott specifikáció a segédrendszertől függő módon használható fel absztrakt, illetve konkrét adattípusok leírására. Konkrét adattípusok specifikálása esetén ez a leírási forma alkalmas a reprezentáció és az implementáció kellő mélységű támogatására. A procedurális specifikációt rendszerint egy programozási nyelv keretében adjuk meg. Erre nyújt például lehetőséget a *Simula nyelv Class konstrukciója* [4] és a legújabb nyelvek között az *Ada Package,* illetve *Generic* [5] lehetősége.

Közismert, hogy az absztrakt specifikáció elsősorban a felhasználó, a konkrét specifikáció pedig az implementációt végző szakember igényeit hivatott kielégíteni. Ennek megfelelően a kétféle követelményt gyakran két specifikációval, azaz specifikáció párral elégítjük ki. A specifikációk párosításai rendszerint a következők:

absztrakt:	konkrét:
1. elő- és utófeltétellel	procedurális
2. elő- és utófeltétellel	elő- és utófeltétellel
3. algebrai	elő- és utófeltétellel

A kettős specifikáció esetén felmerül a konkrét specifikáció absztrakt specifikáció szerinti helyességének a kérdése. Ezt a problémát először C. A. HOARE vetette fel a fenti 1. esetben és megadta a helyesség egy elégséges feltételét [1]. A módszert

később a 2. esetre is kiterjesztették. A 3. eset visszavezethető a 2. esetben felmerülő probléma megoldására [6].

Ebben a dolgozatban az adattípusok helyességének egy egységes tárgyalását adjuk korlátozva magunkat a fenti három eset vizsgálatára.

2. Definíciók és tételek

1. *Definíció.* A $d=(A, F)$ kettőst adattípusnak nevezzük, ha $A=\{A_1, A_2, \ldots, A_m\}$ és $F=\{f_0, f_1, \ldots, f_n\}$, ahol A_i megszámlálható halmaz ($i=1, 2, \ldots, n$), továbbá

$$f_0: \to A_1,$$

$$f_i: A_{i_1} \times A_{i_2} \times \ldots \times A_{i_k} \to A_{i_0}; \quad i = 1, 2, \ldots, n.$$

parciális leképezések. Az A_1 halmazról feltesszük, hogy az *konstruktív*, azaz az A_1 halmaz minden eleme előállítható az $f_0, f_1, \ldots, f_n$ műveletek megfelelő sorrendben történő egymás utáni véges számú alkalmazásával. Az A_1 halmazt a d adattípusba tartozó objektumok halmazának nevezzük.

2. *Definíció.* Az F halmaznak azt a legkisebb elemszámú részhalmazát, amelynek műveletei elégségesek az A_1 halmaz minden elemének az előállítására, konstrukciós részhalmaznak, a benne szereplő műveleteket pedig *konstruktoroknak* nevezzük.

Megjegyzés. Az F halmaznak általában több konstrukciós részhalmaza létezhet. Ilyenkor a konstrukciós részhalmazok közül kiválasztunk egyet és csak azt tekintjük konstrukciós részhalmaznak.

3. *Definíció.* Legyen F_c az F halmaz egy konstrukciós részhalmaza, akkor $F_s=F \setminus F_c$ különbséghalmazt nem konstrukciós vagy szelekciós részhalmaznak nevezzük. F_s elemei nem konstrukciós műveletek, illetve *szelekciós* műveletek.

4. *Definíció.* Legyen $d=(A, F)$ egy adattípus, ahol $A=\{A_1, A_2, \ldots, A_m\}$. Ha az $A_2, \ldots, A_m$ halmazok közül egy vagy több egy másik adattípus objektumainak halmazát jelöli, akkor a d típust paraméteres adattípusnak nevezzük.

5. *Definíció.* Legyen $d=(A, F)$ egy adattípus. A C halmazt az A egy reprezentációjának nevezzük, ha létezik egy olyan $\varphi: C \to A$ leképezés, hogy

$$(\forall a \in A)(\exists c \in C) \quad (a = \varphi(c)).$$

Itt $\varphi(c)=(\varphi_1(c), \varphi_2(c), \ldots, \varphi_m(c))$; $\varphi_i(c) \in A_i$, $i=1, 2, \ldots, m$. Mi olyan esetekkel foglalkozunk, amikor

$$\varphi_i(c) = c_i, \quad i = 2, 3, \ldots, m;$$

ezért a tárgyalás során $\varphi(c)$-t és $\varphi_1(c)$-t jelölésben nem különböztetjük meg egymástól.

1. TÉTEL. Adva van a $d_a=(A, F)$ és a $d_c=(C, G)$ adattípus, ahol $F=\{f_0, f_1, \ldots, f_h, \ldots, f_n\}$ és a $G=\{g_0, g_1, \ldots, g_h, \ldots, g_n\}$. Tegyük fel, hogy itt $f_0, f_1, \ldots, f_h$ és $g_0, g_1, \ldots, g_h$ konstrukciós műveletek. Tegyük fel továbbá, hogy

$A=\{A_1\ A_2, \ldots, A_m\}$, $C=\{C_1, A_2, \ldots, A_m\}$. Ha a φ leképezésre teljesülnek a

$$f_0 = \varphi(g_0) \wedge g_0 \in C$$

$$(\forall c \in C)(\forall_i, 1 \leqq i \leqq k)(f_i(\varphi(c)) = \varphi(g_i(c)) \wedge g_i(c) \in C$$

összefüggések, akkor C az A egy reprezentációja.

Bizonyítás. Azt kell kimutatni, hogy minden $a \in A$ esetén létezik olyan $c \in C$, amelyre $a=\varphi(c)$ teljesül. A bizonyítást az a-ban szereplő konstrukciós műveletek száma szerinti teljesindukcióval végezhetjük el.

Ha $a=f_0$ akkor $c=g_0$ esetén az $a=\varphi(c)$ egyenlet fennáll. Tegyük fel tehát, hogy

$$a = f_c(a'),$$

ahol f_c egy konstrukciós művelet és a'-re már fennáll a tétel:

$$a' = \varphi(c').$$

Ámde

$$a = f_c(\varphi(c')) = \varphi(g_c(c'))$$

ahol g_c az f_c-nek megfelelő konstrukciós művelet. Így tehát $c=g_c(c')$ választással a tétel a esetében is teljesül, ami igazolja a tétel állításának helyességét.

Az adattípus meghatározásánál az A objektum halmazt és az F művelethalmazt kell meghatározni.

Az A objektum halmaz elemeit a szokásos módon, egy olyan állítással adjuk meg, amely csakis az A halmaz elemeire igaz:

$$f_i\colon\ A_{i_1} \times A_{i_2} \times \ldots \times A_{i_k} \to A_{i_0}$$

ahol tehát $I_a(a)$ egy állítás. Ez az állítás a gyakorlatban különböző formákban fogalmazódik meg. Rendszerint egy adott programozási nyelv keretében a deklarációs rész és egy elsőrendű logikában megfogalmazott predikátum együtt szolgáltatja ezt az állítást. A deklarációs résznek, megfelelő predikátummal való helyettesítésével azonban a fenti két rész egy predikátumban egyesíthető. Például az integer i; deklaráció helyett az integer (i) logikai függvényt használhatjuk.

Az A halmaz megadása tehát számunkra azt jelenti, hogy adva van az $I_a(a)$ predikátum.

Az F művelethalmaz meghatározása a benne szereplő műveletek szintaxisának és szemantikájának a megadását jelenti. A szintaxis meghatározása a

$$f_i\colon A_{i_1} \times A_{i_2} \times \ldots \times A_{i_k} \to A_{i_0}$$

forma meghatározását kívánja meg. Ez rendszerint a gyakorlatban is a fenti forma leírásával történik. A szemantika meghatározására azonban a különböző specifikációs módszerek használatosak. Ezeket három csoportba sorolhatjuk:

— procedurális specifikáció;
— elő- és utófeltételekkel adott specifikáció;
— algebrai spacifikáció.

Procedurális specifikáció. Adva van egy műveleti rendszer, például egy programozási nyelv. Procedurális specifikáció esetén az

$$f_i(a_1, a_2, \ldots, a_k)$$

függvényeket egy kiszámítási szabály határozza meg az adott műveleti rendszerben. Ilyen kiszámítási szabály lehet például egy eljárás törzse.

Elő- és utófeltételekkel adott specifikáció esetén az

$$a' = f_i(a)$$

leképezést a bemenő adatokat leíró $\mathrm{pre}_i(a)$ predikátummal és a bemenő, valamint az eredményadatok kapcsolatát leíró $\mathrm{post}_i(a, a')$ predikátummal adjuk meg. Ezt a következő formában írjuk fel:

$$\{\mathrm{pre}_i(a)\}\, a' = f_i(a)\, \{\mathrm{post}_i(a, a')\}.$$

Ez tehát azt jelenti, hogy az f_i függvény értelmezési tartománya a

$$\{a|\ \mathrm{pre}_i(a)\}$$

halmaz. A reláció pedig a

$$\{(a, a')|\ \mathrm{post}_i(a, a')\}$$

halmaz.

Algebrai specifikáció esetén a F halmaz műveleteinek szemantikáját a közöttük fennálló kapcsolatokkal adjuk meg. Ezeket a kapcsolatokat általában a következő formájú egyenletek írják le:

$$f_s(f_c(a)) = f_i(f_j(\ldots(f_k(a)))),$$

ahol f_c konstrukciós, f_s pedig nem konstrukciós művelet, továbbá $f_c, f_s, f_i, f_j, \ldots, \ldots, f_k \in F$.

6. *Definíció.* Adva van a $d_a=(A, F)$ és a $d_c=(C, G)$ adattípus, $F=\{f_0, f_1, \ldots, f_n\}$, $G=\{g_0, g_1, \ldots, g_n\}$, ahol C az A egy reprezentációja adott

$$\varphi: C \to A$$

leképezés mellett. A d_c adattípus specifikációját a d_a adattípus specifikációja szerint helyesnek nevezzük, ha valahányszor $f_i(a)$ értelmezve van, akkor $g_i(c)$, is értelmezve van minden olyan c-re, amelyre $a=\varphi(c)$, és

$$(\forall c \in C)(\forall i,\ 1 \leqq i \leqq n)(c' = g_i(c)) \wedge c' \in C \Rightarrow \varphi(c') = f_i(\varphi(c)) \wedge \varphi(c) \in A \wedge \varphi(c') \in A);$$

$$c_0 = g_0 \Rightarrow \varphi(c_0) = f_0.$$

2. Tétel. Adva van a $d_a=(A, F)$ adattípus, ahol $A=\{a|I_a(a)\}$, az F műveletei pedig elő- és utófeltételekkel adottak:

$$\{\text{igaz}\}\, a_0 = f_0\, \{\mathrm{post}_0(a_0)\},$$

$$\{\mathrm{pre}_i\ (a)\}\, a' = f_i(a)\, \{\mathrm{post}_i(a, a')\}, \quad i = 1, 2, \ldots, n.$$

Adva van továbbá a $d_c=(C, G)$ adattípus, ahol $C=\{c|I_c(c)\}$, a G műveletei pedig rendre a $q_0, q_1, \ldots, q_n$ programokkal adottak. Adva van a $\varphi: C \to A$ leképezés.

Ha bebizonyítjuk, hogy

$$(\forall c \in C)(I_c(c) \Rightarrow I_a(\varphi(c)))$$

és igazak a $q_0, q_1, \ldots, q_n$ programokra a következő teljeshelyességi tételek:

$$\{\text{igaz}\}\, q_0\, \{I_c(c) \wedge \text{post}_0(\varphi(c))\}$$

$$\{I_c(c) \wedge \text{pre}_i(\varphi(c))\}\, q_i\, \{I_c(c') \wedge \text{post}_i(\varphi(c), \varphi(c'))\}$$

akkor a d_c adattípus specifikációja a d_a specifikációja szerint helyes.

Bizonyítás. Először azt kell kimutatni, hogy a C az A egy reprezentációja az adott φ mellett. A 1. tétel értelmében ehhez elég kimutatni, hogy

$$f_0 = \varphi(g_0) \wedge g_0 \in C \quad \text{és} \quad f_i(\varphi(c)) = \varphi(g_i(c)) \wedge g_i(c) \in C$$

minden konstrukciós f_i, g_i műveletpárra. Ez viszont fennáll, ugyanis

$$I_c(c) \wedge \text{post}_0(\varphi(c)) \Rightarrow f_0 = \varphi(g_0) \wedge c \in C$$

$$I_c(g_i(c)) \wedge \text{post}_i(\varphi(c), \varphi(c')) \Rightarrow f_i(\varphi(c)) = \varphi(g_i(c)).$$

A teljeshelyességi tételekből pedig a következőt vezethetjük le:

a) Valahányszor $f_i(\varphi(c'))$ értelmezve van, értelmezve van a $g_i(c)$ is.
b) $I_c(c) \wedge \text{pre}_i(\varphi(c)) \wedge I_c(c') \wedge \text{post}_i(\varphi(c), \varphi(c')) \Rightarrow c \in C \wedge c' \in C \wedge \varphi(c') = f_i(\varphi(c))$

Másrészt feltételeztük, hogy

$$I_c(c) \Rightarrow I_a(\varphi(c)), \quad \text{azaz} \quad \varphi(c) \in A,$$

és

$$I_c(c') \Rightarrow I_0(\varphi(c')), \quad \text{azaz} \quad \varphi(c') \in A.$$

Ez a tétel tehát lehetővé teszi, hogy adattípusok helyességének a kimutatását a fenti esetben a programok teljeshelyességének a kimutatására vezessük vissza.

Vizsgáljuk most azt az esetet, amikor mindkét specifikáció elő- és utófeltételekkel adott.

3. TÉTEL. Legyen adva $d_a = (A, F)$, ahol

$$A = \{a \mid I_a(a)\}$$

$$\{\text{igaz}\}\, a_0 = f_0\, \{\text{post}_{f_0}(a)\},$$

$$\{\text{pre}_{f_i}(a)\}\, a' = f_i(a)\, \{\text{post}_{f_i}(a, a')\}, \quad i = 1, 2, \ldots, n,$$

és

$$d_c = (C, G),$$

ahol

$$C = \{c \mid I_c(c)\}$$

$$\{\text{igaz}\}\, c_0 = g_0\, \{\text{post}_{g_0}(c)\},$$

$$\{\text{pre}_{g_i}(c)\}\, c' = g_i(c)\, \{\text{post}_{g_i}(c, c')\} \quad i = 1, 2, \ldots, n.$$

Adva van továbbá a $\varphi\colon C \to A$ leképezés. Ha bebizonyítjuk, hogy

a) $\quad \text{post}_{g_0}(c) \Rightarrow I_c(c),$

b) $\quad I_c(c) \Rightarrow I_a(\varphi(c)),$

c) $\quad \text{post}_{g_0}(c) \Rightarrow \text{post}_{f_0}(\varphi(c)),$

d) $\quad I_c(c) \wedge \text{pre}_{f_i}(\varphi(c)) \Rightarrow \text{pre}_{g_i}(c),$

e) $\quad I_c(c) \wedge \text{pre}_{f_i}(\varphi(c)) \wedge \text{post}_{g_i}(c, c') \Rightarrow \text{post}_{f_i}(\varphi(c), \varphi(c')),$

továbbá minden konstrukciós g_i műveletre kimutatjuk, hogy

$$I_c(c) \wedge \text{pre}_{g_i}(c) \wedge \text{post}_{g_i}(c, c') \Rightarrow I_c(c'),$$

akkor a d_c specifikációja a d_a szerint helyes.

Bizonyítás. Először belátjuk, hogy C az A egy reprezentációja. Mivel

$$\text{post}_{g_0}(c) \wedge \text{post}_{f_0}(\varphi(c)) \wedge I_c(c) \Rightarrow f_0 = \varphi(g_0) \wedge g_0 \in C$$

és minden konstrukciós műveletre kapjuk, hogy

$$I_c(c) \wedge \text{pre}_{g_i}(c) \wedge \text{post}_{g_i}(c, c') \wedge \text{post}_{f_i}(\varphi(c), \varphi(c')) \wedge I_c(c')$$
$$\Rightarrow f_i(\varphi(c)) = \varphi(g_i(c)) \wedge g_i(c) \in C$$

ezért az 1. tétel értelmében C az A egy reprezentációja. A d) tétel alapján nyilvánvaló, hogy valahányszor $\varphi(c)$-re az f_i értelmezve van, akkor a megfelelő g_i és értelmezve van c-re, ugyanakkor e) szerint

$$I_c(c) \wedge \text{pre}_{f_i}(\varphi(c)) \wedge \text{post}_{g_i}(c, c') \wedge \text{post}_{f_i}(\varphi(c), \varphi(c')) \Rightarrow$$
$$\Rightarrow c' = g_i(c) \wedge c \in C \wedge c' \in C \wedge \varphi(c') = f_i(\varphi(c))$$

és b) alapján pedig következik, hogy

$$\varphi(c) \in A \wedge \varphi(c') \in A.$$

Így a 6. definíció alapján d_c a d_a adattípus szerint helyes.

Vizsgáljuk meg végül a 3. esetet.

Legyen adva a $d_a = (A, F)$ adattípus specifikációja algebrai egyenletekkel:

$$f_s(f_c(a)) = f_i(f_j(\ldots(f_k(a)))),$$

és legyen adva

$$\text{atr}\colon a \to N$$

attributum függvény, a következő szemantikával

$$\text{atr}\,(f_0) = n_0$$
$$\text{atr}\,(f_c(a)) = h_{f_c}(\text{atr}\,(a)),$$

ahol $h_{f_c}\colon a \to N$ adott leképezés, és N a nem negatív egész számok halmaza.

Legyen továbbá adott az absztrakt invariáns

$$I_a(a)\colon\ n_0 \leqq \mathrm{atr}\,(a) \leqq n_1$$

formában.

A d_a adattípus specifikációját a következő egyszerű átalakítással elő- és utófeltételekkel megadott formára hozhatjuk:

Minden f_s szelekciós műveletre:

$$\{b = f_c(a)\}\quad b' = f_s(b)\quad \{b' = f_i(f_j(\ldots(f_k(a))))\},$$

az f_0 konstrukciós függvényre:

$$\{\text{igaz}\}\quad a = f_0\{\mathrm{atr}\,(a) = n_0\}$$

és f_c konstrukciós függvényre pedig

$$\{n_0 \leqq h_{f_c}(\mathrm{atr}\,(a)) \leqq n_1\}\quad a' = f_c(a)\quad \{a' = f_c(a)\}.$$

Ezzel a módszerrel tehát a 3. esetünkben a helyességbizonyítás problémáját visszavezettük a 2. eset problémájára.

3. Egy példa

A táblázat egy lehetséges absztrakciója a következő:

$d_a = (\{\text{table, index, elem}\}, \{\text{null, assign, delete, read}\})$, ahol az egyszerűség kedvéért legyen

index = $\{1, 2, \ldots, n)$, $n > 0$

elem = integer

Szintaxis:

null: → table

assign: table x index x elem → table

delete: table x index → table

read: table x index → elem

Szemantika:

delete (null, p) = null

delete (assign (t, p, e), q) =

ha $p = q$, akkor delete (t, q)

különben assign (delete (t, q), p, e)

read (null, p) = 0

read (assign (t, p, e), q) =

ha $p = q$ akkor e különben read (t, q)

Attributum függvény:

length: table → N

length (null) = 0

length (assign (t, p, e)) = length (delete (t, p)) + 1

Absztrakt invariáns:

$I_a(t)$: $0 \leqq \mathrm{length}(t) \leqq n$

Egyenlőség axiómája:

$t_1 = t_2 \equiv (t_1 = \text{null} \wedge t_2 = \text{null}) \vee$
$(\forall i,\ 1 \leqq i \leqq n)\ (\text{read}(t_1, i) = \text{read}(t_2, i))$

Specifikáció analízis:

TÉTEL. $\text{read}(\text{delete}(t, p), q) =$
ha $p = q$, akkor 0 különben $\text{read}(t, q)$.

Bizonyítás. Alkalmazzuk a konstrukciós műveletek számára vonatkozó teljesindukciót.

a) $t = \text{null}$ esetén $\text{read}(\text{delete}(t, p), q) = \text{read}(\text{null}, q) = 0$

b) Legyen $t = \text{assign}(t_0, j, e)$, ahol feltesszük, hogy a tétel t_0-ra már igaz.

$\text{read}(\text{delete}(\text{assign}(t_0, j, e), p), q) =$
ha $j = p$, akkor $\text{read}(\text{delete}(t_0, p), q)$
különben $\text{read}(\text{assign}(\text{delete}(t_0, p), j, e), q)$

Mivel

$$\text{read}(\text{delete}(t_0, p), q) = \begin{cases} 0, & \text{ha } p = q \\ \text{read}(t_0, q), & \text{ha } p \neq q \end{cases}$$

és

$\text{read}(\text{assign}(\text{delete}(t_0, p), j, e), q) =$
$= e$, ha $j = q$
$= \text{read}(\text{delete}(t_0, p), q)$, ha $j \neq q$

azaz

$$\text{read}(\text{delete}(t, p), q) = \begin{cases} 0, & \text{ha } p = q \\ e, & \text{ha } q = j \\ \text{read}(t_0, q) & \text{ha } q \neq p \wedge q \neq j \text{ m} \end{cases}$$

ámde

$\text{read}(t, q) =$ ha $q = j$ akkor e különben $\text{read}(t_0, q)$,

ami mutatja a tétel helyességét.

Hasonló módon bizonyíthatók be a specifikáció analizálására vonatkozó további tételek is:

$p \neq q \Rightarrow \text{assign}(\text{assign}(t, p, e), q, c) = \text{assign}(\text{assign}(t, q, c), p, e)$,
$\text{delete}(\text{delete}(t, p), q) = \text{delete}(\text{delete}(t, q), p)$,

stb.

Reprezentáció:

konkrét táblázat = integer vector $v[1 : n]$

Legyen $h(v, 0) = 0$,
$h(v, i) =$ ha $v[i] = 0$ akkor $h(v, i)$
különben $h(v, i) + 1$

és a konkrét attributumfüggvény

$\text{length}_c(v) = h(v, n)$

A konkrét attributumfüggvény tehát v-ben a nem zérus elemek számával azonos.

Legyen

$$I_c(v)\colon\ 0 \leqq h(v,n) \leqq n$$

és

$$\varphi(v) = \text{null} \quad \underline{\text{ha}} \quad h(v,n) = 0$$
$$= \text{assign}\,(\varphi(\text{delete}_c\,(v,p)),\ p,\ A[p]),$$
$$\underline{\text{ha}} \quad h(v,n) \neq 0 \wedge A[p] \neq 0.$$

Nyilvánvaló, hogy

$$h(v,n) \neq 0 \Rightarrow (\exists p,\ 1 \leqq p \leqq m)\quad (A[p] \neq 0).$$

Implementáció:

$$\{\text{igaz}\}\quad v = \text{null}_c\,\{(\forall p,\ 1 \leqq p \leqq n)\quad (A[p] = 0)\}$$
$$\{w_1(v,p,e)\}\quad v' = \text{assign}_c\,(v,p,e)$$
$$\{v'[p] = e \wedge (\forall i,\ i \neq p)\quad (v'[i] = v[i])\}$$
$$\{w_2(v,p)\}\quad v' = \text{delete}_c\,(v,p)$$
$$\{v'[p] = 0 \wedge (\forall i,\ i \neq p)\quad (v'[i] = v[i])\}$$
$$\{w_2(v,p)\}\quad x = \text{read}_c\,(v,p)\quad \{x = v[p]\}$$

ahol

$$w_1(v,p,e)\colon\ I_c(v) \wedge 1 \leqq p \leqq n \wedge \text{integer}\,(e),$$
$$w_2(v,p)\colon\ I_c(v) \wedge 1 \leqq p \leqq n.$$

Reprezentáció analízis:

1. TÉTEL. $(\forall p,\ 1 \leqq p \leqq n)\ (v[p]=0) \equiv h(v,n)=0$
Az állítás $h(v,n)$ definíciója alapján nyilvánvaló.

2. TÉTEL. $\varphi(v)$ függvény, azaz
$\varphi_1(v) =$ ha $h(v,n) = 0$, akkor null
különben $\text{assign}\,(\varphi_1(\text{delete}_c\,(v,p)),\ p,\ v[p])$,
$\varphi_2(v) =$ ha $h(v,n) = 0$, akkor null
különben $\text{assign}\,(\varphi_2(\text{delete}_c\,(v,q)),\ q,\ v[q])$

mellett

$$A[p] \neq \wedge 0 A[q] \neq 0 \Rightarrow \varphi_1(v) = \varphi_2(v).$$

Bizonyítás. Végezzük el a bizonyítást a $h(v,n)$ szerinti teljesindukcióval.

a) Ha $h(v,n)=0$ akkor $\varphi_1(v)=\varphi_2(v)$ nyilvánvalóan igaz.
Ha $h(v,n)=1$, akkor csak $p=q$ lehet és az egyenlőség ismét triviálisan teljesül.
b) Tegyük tehát fel, hogy $h(v,n)=k$, $k \geqq 2$ és minden olyan v-re, amelyre $h(v,n) \leqq k-1$ a tétel fennáll.

Bizonyítsuk be, hogy a tétel $h(v,n)=k$-ra is igaz.

$$A[j] \neq 0 \Rightarrow h(\text{delete}_c\,(v,j),\ n) = h(v,n) - 1,$$

ezért

$$\varphi_1(\text{delete}_c\,(v,j)) = \varphi_2(\text{delete}_c\,(v,j)).$$

Tehát a bizonyítandó állítás

$p \neq q =$ assign $(\varphi_1(\text{delete}_c(A, p)), p, A[p]) =$
assign $(\varphi_1(\text{delete}_c(A, q)), q, A[q])$.

Ezt viszont az egyenlőség axiómája alapján láthatjuk be:

read (assign $(\varphi_1(\text{delete}_c(A, p)), p, A[p], j)) =$
ha $j=p$, akkor $A[p]$
különben read $(\varphi_1(\text{delete}_c(A, p)), j)$.

Legyen $u = \text{delete}_c(\text{delete}_c(v, p), q)$, akkor

read (assign $(\varphi_1(\text{delete}_c(A, p), p, A[p]), j)) =$
$= A[p]$, ha $j=p$
$= A[q]$, ha $j \neq p \wedge j = q$
$= \text{read}(\varphi_1(u), j)$, ha $j \neq p \wedge j \neq q$

p és q felcserélésével azonnal kapjuk, hogy

read $(\varphi_1(\text{delete}_c(v, q), q, A[q]), j) =$
$= A[q]$, ha $j=q$
$= A[p]$, ha $j \neq q \wedge j = p$
$= \text{read}(\varphi_1(u), j)$, ha $j \neq q \wedge j \neq p$

ami igazolja a tételt.

3. TÉTEL. $h(v, n) = \text{length}(\varphi(v))$.

Bizonyítás. Ismét $h(v, n)$ szerinti teljesindukcióval dolgozhatunk.

a) $h(v, n) = 0 \Rightarrow v \Rightarrow \text{null} \Rightarrow \varphi(v) = \text{null} = \text{length}(\varphi(v)) = 0$

b) Legyen tehát $h(v, n) = k$ és tegyük fel, hogy a tétel fennáll minden olyan u-ra, amelyre $h(u, n) < k$.
Legyen $v = \text{assign}_c(u, p, v[p])$, ahol $v[p] \neq 0$.

Akkor feltevésünk szerint

$$h(u, n) = \text{length}(\varphi(u)),$$

és

$$h(v, n) = h(u, n) + 1.$$

Másrészt

$$\begin{aligned} \text{length}(\varphi(v)) &= \text{length}(\varphi(\text{assign}_c(u, p, v[p]))) = \\ &= \text{length}(\text{assign}(\varphi(\text{delete}_c(u)), p, v[p])) = \\ &= \text{length}(\varphi(\text{delete}_c(u)) + 1 = \\ &= \text{length}(\varphi(u)) + 1, \end{aligned}$$

ami igazolja a tételt.

Ebből azonnal kapjuk, hogy

4. TÉTEL. $I_c(v) = I_a(\varphi(v))$.

5. TÉTEL. Az assign_c konstrukciós művelet nem visz ki a konkrét objektumok teréből, azaz

$$I_c(v) \wedge w_1(w, p, e) \wedge v'[p] = e \wedge (\forall i,\ i \neq p) \quad (v'[i] = v[i]) \Rightarrow I_c(v').$$

A tétel állítása $h(v', n)$ definíciója alapján nyilvánvaló.

Implementáció analizis:

Hozzuk először az absztrakt specifikációt elő- és utófeltételekkel megadott formára:

$\{\text{igaz}\}\ t=\text{null}\{\text{length}(t)=0\}$

$\{\text{length}(\text{delete}(t,p))+1\leqq n\}$
$t'=\text{assign}(t,p,e)\ \{t'=\text{assign}(t,p,e)\}$

$\{t=\text{null}\}\ t'=\text{delete}(t,p)\ \{t'=\text{null}\}$

$\{t=\text{assign}(t_0,p,e)\}\ t'=\text{delete}(t,q)$
$\{\underline{\text{ha}}\ p=q,\ \underline{\text{akkor}}\ t'=\text{delete}(t,q)$
$\underline{\text{különben}}\ t'=\text{assign}(\text{delete}(t,q),p,e)\}$

$\{t=\text{null}\}\ x=\text{read}(t,p)\ \{x=0\}$

$\{t=\text{assign}(t_0,p,e)\}\ x=\text{read}(t,q)$
$\{\underline{\text{ha}}\ p=q\ \underline{\text{akkor}}\ x=e\ \underline{\text{különben}}\ x=\text{read}(t,q)\}.$

Az implementációra vonatkozó tételek közül két tétel bizonyítását mutatjuk be, a többi hasonlóképpen bizonyítható.

1. TÉTEL

$$0\leqq h(v,n)\leqq n\wedge\text{length}\,(\text{delete}\,(\varphi(v),p))\leqq$$

$$\leqq n-1\wedge v'[p]=e\wedge(\forall i,\ i\neq p)\ \ (v'[i]=v[i])\ \ +0\leqq h(v',n)\leqq n\Rightarrow$$

$$\Rightarrow\varphi(v')=\text{assign}\,(\varphi(v),p,e).$$

Bizonyítás.

$\varphi(v')=\text{assign}\,(\varphi(\text{delete}_c(v',p)),p,v'[p])$.

Ezért

$\text{read}\,(\varphi(v'),j)=\underline{\text{ha}}\ j=p\ \underline{\text{akkor}}\ v'[p]$
$\underline{\text{különben}}\ \text{read}\,(\varphi(\text{delete}_c(v',p)),j)$.

Hasonlóképpen

$\text{read}\,(\text{assign}\,(\varphi(v),p,e),j)$
$\underline{\text{ha}}\ j=p,\ \underline{\text{akkor}}\ e$
$\underline{\text{különben}}\ \text{read}\,(\varphi(v,),j)$

Figyelembe véve, hogy $e=v'[p]$ és

$$j\neq p\Rightarrow\text{read}\,(\varphi(\text{delete}_c(v',p),j))=\text{read}\,(\varphi(v),j)$$

kapjuk a tétel állítását.

2. TÉTEL

$$0\leqq h(v,n)\leqq n\wedge\varphi(v)=\text{assign}\,(t_0,p,e)\wedge p=q\wedge$$

$$\wedge v'[p]=0\wedge(\forall i,\ 1\leqq i\leqq n)\ \ (v'[i]=v[i])\Rightarrow\varphi(v')=\text{delete}\,(\varphi(v),q).$$

Bizonyítás. Ismét az egyenlőség axióma és $\varphi(v)$ definíciója alapján keressük a megoldást.

$$\text{read}\,(\varphi(v'),j)=\underline{\text{ha}}\quad\varphi(v')=\text{null},\ \ \underline{\text{akkor}}\ \ 0$$

$$\underline{\text{különben}}\ \ \text{read}\,(\text{assign}\,(\varphi(\text{delete}_c(v',k)),k,v'[k]),j).$$

A tétel baloldala alapján

$$\varphi(v') = \text{null} \Rightarrow \varphi(v) = \text{assign}(\text{null}, p, e) \Rightarrow \text{delete}(\varphi(v), p) = \text{null} \Rightarrow$$

$$\Rightarrow (\forall j, 1 \leqq j \leqq n) \quad (\text{read}(\text{delete}(\varphi(v), p), j) = 0).$$

Nézzük a $\varphi(v') \neq \text{null}$ esetet. Ekkor

$$\text{read}(\varphi(v'), j) =$$
$$\text{read}(\text{assign}(\varphi(\text{delete}_c(v', k)), k, v'[k]), j) =$$
$$\underline{\text{ha}}\ j=k,\ \underline{\text{akkor}}\ v'[k]$$
$$\underline{\text{különben}}\ \text{read}(\varphi(\text{delete}_c(v', k)), j),$$

azaz

$$\text{read}(\varphi(v'), j) = \underline{\text{ha}}\ j=p,\ \underline{\text{akkor}}\ 0$$
$$\underline{\text{különben}}\ v[j],$$

ami ugyanaz, mint $\text{read}(\text{delete}(\varphi)v, q), j))$.

IRODALOM

[1] HOARE, C. A. R., "Proof of correctness of data representations", *Acta Informatics* (1972) 271—281.
[2] SHAW, M., WULF, W. A. and LONDON, R. L., "Abstraction and verification in Alphard", *Communications of the ACM* (1977) 553—564.
[3] EHRIG, H. and MAHR, B., *Fundamentals of Algebraic Specification* 1 (Springer-Verlag, Berlin, 1985).
[4] DAHL, OLE-JOHAN, MYHRHAUG, BJORN and NYGAARD, KRISTEN, "SIMULA 67 common base language", Publication N. S—22, Oslo, Norwegian Computing Center, October 1970.
[5] *Reference Manuale for the ADA Programming Language* (ANSI/MIL—STD—1815A, January 1983).
[6] VARGA, L., "Specification of reliable software", *MTA SZTAKI Tanulmányok* **113** (1980) 302—325.

(Beérkezett: 1986. december 10.)

VARGA LÁSZLÓ
ELTE TTK ÁLTALÁNOS SZÁMÍTÁSTUDOMÁNYI TANSZÉK
1088 BUDAPEST, MÚZEUM KRT. 6—8.

STUDY OF THE CORRECTNESS OF DATA TYPE SPECIFICATIONS

L. VARGA

The properties of a data type relevant to programs using it could be described by an abstract specification. For implementing a data type we may choose an appropriate representation and give a concrete data type specification. Hence double specification is given for a data type. In this paper the correctness of a concrete specification according to an abstract specification is formulated and sufficient condition for the correctness of three double specifications are discussed.

Alkalmazott Matematikai Lapok **13** (1987—88) 69—76

A BOOLE-FÜGGVÉNYEK BOOLE ALGEBRÁJÁNAK STRUKTURÁLIS TULAJDONSÁGAIT FELHASZNÁLÓ BOOLE FÜGGVÉNY OPTIMALIZÁCIÓS MÓDSZER

PÁSZTORNÉ VARGA KATALIN

Budapest

A cikkben ismertetjük az *n-változós Boole-függvények* halmaza (C_n^n) *Boole algebrájának* azon néhány tulajdonságát, ami szükséges az eredmény megértéséhez. Részletezzük azokat a dualitásból következő, nem közismert kapcsolatokat, amelyek rendkívüli egyszerűsítést jelentenek a szintézisben. Bemutatunk egy olyan szintézis algoritmust, ami döntésmechnizmusát tekintve sokkal egyszerűbb a korábbiaknál.

1. Fogalmak és jelölések

Jelölje

$B = \{0, 1\}$ — az egydimenziós,

$B^n = \{0, 1\}^n$ — az n dimenziós *Boole-teret* (más elnevezések szerint n-dimenziós (0, 1)-teret vagy n-dimenziós *Boole-kockát*),

$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ — a B^n tér egy vektorát (pontját),

$\bar{\alpha} = (\bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_n)$ — az α vektor negáltját,

$\mathbf{x} = (x_1, x_2, \ldots, x_n)$ — a B^n teret befutó változót *(Boole-vektorváltozót)*,

$\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$ — az $\mathbf{x}$ vektorváltozó negáltját,

$h(\alpha, \beta)$ — α és β Hamming távolságát, (α, β különböző koordinátáinak számát),

$\vee$ — a diszjunkciót

konkatenáció vagy $\wedge$ — a konjunkciót,

$f(x_1, x_2, \ldots, x_n) = f(\mathbf{x})$ — egy n-változós *Boole-függvényt*,

$\bar{f}$ vagy $\neg f$ — az f negáltját,

$f^*(\mathbf{x})$ vagy f^* — az f duálisát,

$f_{xi=\alpha}$ — az $f(x_1, x_2, \ldots, x_{i-1}, \alpha, x_{i+1}, \ldots, x_n)$ függvényt,

F — az f egy diszjunktív normálformáját (DNF),

f_φ — egy nem teljesen meghatározott (parciális) *Boole-függvényt*,

α-t az f 1 (0 vagy határozatlan)-pontjának nevezzük aszerint, hogy $f(\alpha) = 1$ (0 vagy határozatlan).

f_1, f_0, g — az f_φ-t meghatározó függvények (f_φ komponensei)

$$f_\varphi = \begin{cases} 1, & \text{ha } f_1 = 1 \\ 0, & \text{ha } f_0 = 1 \\ \text{határozatlan}, & \text{ha } g = 1 \end{cases}$$

Vagyis f 1-pontjai f_1 1-pontjaival, f 0-pontjai f_0 1-pontjaival, f határozatlan pontjai g 1-pontjaival esnek egybe.
$\underset{\wedge}{f}=f_1$ az f_φ alsó határa, $\hat{f}=f_1 \vee g$ az f_φ felső határa.
Az f_φ-vel kompatibilis az f teljesen meghatározott *Boole-függvény*, ha $\underset{\wedge}{f} \to f$, $f \to \hat{f}$ teljesül. f megadható $f=\underset{\wedge}{f} \vee \gamma g$ vagy $f=\underset{\wedge}{f} \vee \gamma \bar{f}_0$ alakban is, ahol γ a megfelelő *Boole-függvény*.

$$x^\alpha = \begin{cases} x, & \text{ha } \alpha = 1 \\ \bar{x}, & \text{ha } \alpha = 0 \end{cases}$$

A term kifejezés elemi konjunkciót, a minterm kifejezés teljes elemi konjunkciót jelöl. Ha $p=x_1^{a_1} x_2^{a_2} \ldots x_k^{a_k}$, akkor $p^- = x_1^{\bar{a}_1} x_2^{\bar{a}_2} \ldots x_k^{\bar{a}_k}$. A p term fedi a q termet, ha $pq=q$.
További jelölések:

$\underset{\wedge}{F}=t_1 \vee t_2 \vee \ldots \vee t_n$	— $\underset{\wedge}{f}$ egy DNF-je,
$\hat{F}^*=d_1 \vee d_2 \vee \ldots \vee d_s$	— $\hat{f}^*$ egy DNF-je.
(F/G)	— egy az f_φ-vel kompatibilis DNF.
$\hat{F} \diagdown p$ (vagy $f \diagdown p$)	— F-ből a p term által lefedett termek elhagyásával kapott DNF,
T	— az az $s \times n$-es táblázat, amelynek oszlopaihoz $\underset{\wedge}{F}$, soraihoz pedig $\hat{F}^*$ termjeit rendeljük, és T-nek az e_{ij} eleme a d_i és t_j termek közös tényezőinek halmaza (1. ábra).
t_i-hez tartozó $x_{k,\min}^{a_k}$	— egy minimális elemszámú e_{qi} egy eleme.
E_j	— olyan minimális számú változót tartalmazó halmaz, hogy

$$E_j \cap e_{ij} \neq \emptyset, \quad 1 \leqq i \leqq s.$$

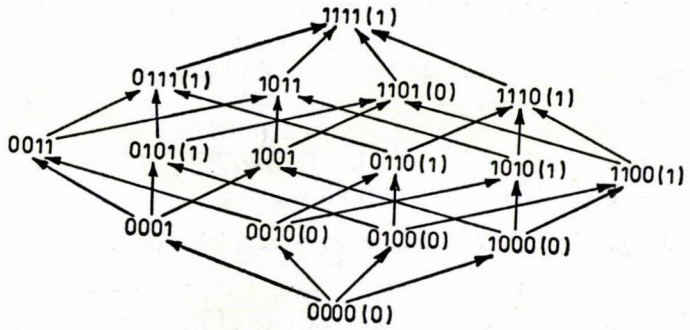
Példa: $f_1=abc \vee ab\bar{d} \vee \bar{a}bd \vee a\bar{b}c\bar{d} \vee \bar{a}bc$, $f_0=\bar{b}c\bar{d} \vee \bar{a}\bar{b}\bar{d} \vee \bar{a}c\bar{d} \vee abc\bar{d}$, $g=\bar{b}d$ az f_φ komponensei.

	abc	*ab$\bar{d}$*	*$\bar{a}$bd*	*a$\bar{b}$c$\bar{d}$*	*$\bar{a}$bc*
$\bar{a}$bcd	*c*	*d*	*$\bar{a}$*	*b, c, d*	*$\bar{a}$, c*
abd	*a, b*	*a, b*	*b, d*	*a*	*b*
bcd	*b, c*	*b*	*b, d*	*c*	*b, c*
acd	*a, c*	*a*	*d*	*a, c*	*c*

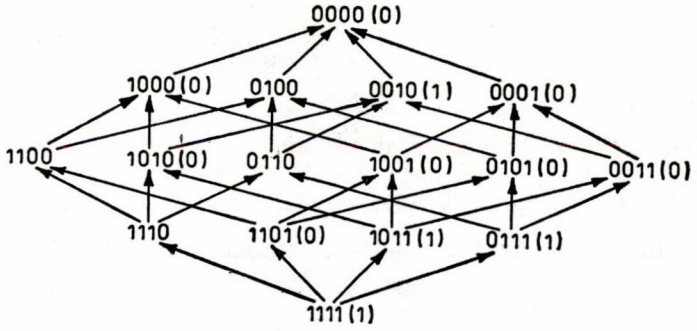
1. ábra

2. A dualitás és a T táblázat

Néhány, az [1, 2]-ben már leírt tulajdonságot röviden ismertetünk a megértés könnyítése céljából.
Tekintsük B-n a $0<1$ relációból adódó rendezést. Tetszőleges α, $\beta \in B^n$-re $\alpha \leqq \beta$ azt jelenti, hogy $\alpha_i \leqq \beta_i$, $1 \leqq i \leqq n$. E reláció szerint B^n hálót alkot. E háló hálódiagramjában a $h(\alpha, \beta)=1$ esetben az α-val és a β-val jelölt csúcsot a nagyobb



2/a ábra. Az f_φ függvény ábrázolása



2/b ábra. Az f_φ függvény duálisának ábrázolása a duális hálóban

elem felé irányuló él köti össze. A háló minimális illetve, maximális eleme a **0**, illetve az **1** vektor. Ábrázoljunk egy adott $f(x_1, \ldots, x_n)$ függvényt a B^n-ben úgy, hogy a tér pontjaihoz hozzárendeljük a függvény ott felvett értékét (2a ábra).

A B^n-beli $\min(\alpha, \beta)$ és $\max(\alpha, \beta)$ műveletek, mint hálóműveletek, valamint a **0** és **1** egymás duálisai. A műveletek szerepét felcserélve az $1 < 0$ duális reláció alapján keletkező duális hálót kapjuk. A duális háló hálódiagramja az eredeti hálóéból az elemeknek duálisukra való cserélésével kapható (2b ábra). Egy függvény és a duálisa közötti kapcsolat a következő:

$$f^*(\alpha) = \bar{f}(\bar{\alpha}). \tag{2.1}$$

A fenti ábrázolási mód szerint egy f függvény f^* duálisának ábrázolásánál a duális háló egy pontjához az eredeti háló ugyanolyan helyzetű pontjához rendelt f függvény értékének duálisát kell írni. Ez szemléletesen is mutatja azt, a (2.1)-ből is látható tényt, hogy f^* 1-pontjainak száma megegyezik $\bar{f}$ 1-pontjainak számával. Nem teljesen meghatározott függvények esetén ugyanez vonatkozik $\hat{f}^*$ és f_0 1-pontjaira.

A háló pontjait elemi konjunkciókként tekintve az F és F^*-ra is érvényesek az eddigiek alapján könnyen belátható alábbi állítások és az azokat követő tételek:
a) f^* ($\hat{f}^*$) 1-pontjainak száma megyegyezik f (f_0) 1-pontjainak számával.
b) Tetszőleges α n-dimenziós *Boole vektorra* $h(\alpha, \bar{\alpha})=n$.
c) $f^*(\alpha)=\beta$ akkor és csak akkor, ha $f(\bar{\alpha})=\bar{\beta}$ $(\beta \in B)$.
d) $f^*(\alpha)=f^*(\bar{\alpha})$ akkor és csak akkor, ha $f(\alpha)=f(\bar{\alpha})$ $(=\bar{f}^*(\alpha))$.
e) Tetszőleges α, β-ra, $h(\bar{\alpha}, \beta)=n-h(\alpha, \beta)$.

2.1. TÉTEL. ([2], 1. tétel) T-ben egyetlen e_{ij} sem lehet üres.

Bizonyítás. Tegyük fel, hogy legalább egy e_{ij} üres. Tartozzon a j-edik oszlophoz α, ekkor a feltevés miatt az i-edik sorhoz csak $\bar{\alpha}$ tartozhat. Ez azt jelenti, hogy $f(\alpha)=1$ és $\hat{f}^*(\bar{\alpha})=1$, ami a c) miatt lehetetlen. Ezzel a tételt bebizonyítottuk.

2.2. TÉTEL. Ha $f^*(\alpha)=1$, $f(\beta)=1$ és $h(\alpha, \beta)=n-1$, ahol $\alpha_i=\beta_i$, akkor f-nek van olyan, a β-t fedő prímimplikánsa, amelyben $x_i^{\alpha i}$ szerepel.

Bizonyítás. e) miatt $h(\bar{\alpha}, \beta)=1$ és $f(\bar{\alpha})=0$. $\bar{\alpha}$ és β α_i szerint szomszédos. Mivel az f-nek $\bar{\alpha}$ 0-pontja, β 1-pontja, az $x_i^{\beta_i}$ változó nem egyszerűsíthető ki a β-t adó $x_1^{\beta_1}x_2^{\beta_2}\ldots x_n^{\beta_n}$ konjunkcióból.

2.1. KÖVETKEZMÉNY. Minden olyan $x_i^{\beta_i}$ változó szerepel a β-t lefedő prímimplikánsban, amelyhez van olyan α, hogy $f^*(\alpha)=$ és $\bar{\alpha}$ szomszédos β-val β_i szerint. (Más szóval ahol α és β közös változóinak száma egy, azaz $\|e_{\alpha\beta}\|=1$ T-ben.)

2.2. KÖVETKEZMÉNY. Azok a változók, amelyek a T táblázatban az $\hat{f}$ egy t_i termjének oszlopában legalább egyszer egyedül szerepelnek, biztosan előfordulnak a t_i-t lefedő prímimplikánsokban.

2.3. TÉTEL. Ha $f(\beta)=1$, $f^*(\alpha)=1$ és $h(\alpha, \beta)=n-k$, $k>1$, továbbá $n-k$ maximális f^* összes 1-pontját tekintve, valamint $\alpha_{i1}=\beta_{i1}, \ldots, \alpha_{ik}=\beta_{ik}$, akkor
1. az $x_{i1}^{\alpha i1}, x_{i2}^{\alpha i2}, \ldots, x_{ik}^{\alpha ik}$ változók bármelyike szerepelhet β-t lefedő prímimplikánsban;
2. nincs olyan β-t lefedő prímimplikáns, amely e k változó mindegyikét tartalmazná.

Bizonyítás. 1. e) miatt $h(\bar{\alpha}, \beta)=k$ és $f(\bar{\alpha})=0$, tehát van legalább egy 0-pont (pl. $\bar{\alpha}$) a β-nak $\beta_{i1}, \beta_{i2}, \ldots, \beta_{ik}$-ra nézve k „sugarú" környezetében. Vagyis nem egyszerűsíthető ki egyszerre minden közös változó [1].
2. Ahhoz, hogy a k közös változó egyike sem legyen kiegyszerűsíthető, az kell, hogy az $\bar{\alpha}$-tól e változókra nézve $k-1$ *Hamming távolságú* $\gamma_1, \ldots, \gamma_k$ pontok 0-pontok legyenek. Ebben az esetben viszont $\bar{\gamma}_1, \bar{\gamma}_2, \ldots, \bar{\gamma}_k$ f^*-nak 1-pontjai, és ezekre $h(\bar{\gamma}_i, \beta)=n-1$, ami ellentmond annak, hogy k minimális volt β-ra nézve.

2.3. KÖVETKEZMÉNY. 1. A T táblázat i-edik oszlopában levő minimális, de 1-nél nagyobb, elemszámú e_{qi}-ban szereplő változók bármelyikéhez van olyan, a t_i-t lefedő prímimplikáns, amelyben az illető változó szerepel.
2. Nem szerepelhet e_{qi} minden eleme egyazon prímimplikánsban.

2.4. TÉTEL. Ha a 2.2. tétel alapján egy $x_i^{\alpha i}$ változót mint egy β-t lefedő p prímimplikáns változóját rögzítettük, akkor e p elérését célzó további döntésekhez nincs szükség az f^* azon 1-pontjaira, amelynek i-edik komponense α_i.

Bizonyítás. Egy $x_i^{\alpha_i}$ rögzítése azt jelenti, hogy $p = x_i^{\alpha_i} q$ (q x_i-t nem tartalmazó elemi konjunkció), vagyis p az $x_i = \alpha_i$ féltérben helyezkedik el. Tehát $\bar{f}$-nak azon 1-pontjai, amelyekben α az i-edik komponens, a p további pontosítása során nem jönnek szóba. Ezzel a tételt bebizonyítottuk.

Megjegyzés. A bizonyításban leírt folyamat szemléletesen azt jelenti, hogy q meghatározása egy olyan $(n-1)$-dimenziós térben történik, amelynek pontjai az n-dimenziós térbeli, α_i i-edik komponensű pontok utódai.

2.4. KÖVETKEZMÉNY. Ha a T táblázat t_j oszlopában megjelöltük $x_i^{\alpha_i}$-t, mint egyedüli elemét egy e_{qj}-nek, akkor minden olyan e_{rj} $(1 \leqq r \leqq s)$ elhagyható, amely tartalmazza $x_i^{\alpha_i}$-t.

2.5. KÖVETKEZMÉNY. A 2.4. tétel igaz akkor is, ha $x_i^{\alpha_i}$-t a 2.3. tétel alapján (nem egyelemű e_{qj}-ból) választottuk ki.

2.5. TÉTEL. A β-t lefedő p prímimplikánst (mint az eljárás során megjelölt változók konjunkcióját) akkor kapjuk meg, ha a változók megjelölése során f^* minden 1-pontját elhagytunk már.

Bizonyítás. Egy $x_i^{\alpha_i}$ változó megjelölése annak ki nem egyszerűsíthetőségét, az f^* függvény α_i i-edik komponensű 1-pontjainak elhagyása pedig a kiegyszerűsíthetetlenség feltételeinek megszüntetését jelenti. Ha már nincs f^*-nak pontja, akkor nincsenek korlátozó feltételek, tehát β nem érintett változói mind kiegyszerűsíthetők.

2.6. TÉTEL. A 2.2—2.5. tételek alkalmazhatók olyan T táblázatokra is, ahol a t_i és d_j termek tetszőleges elemi konjunkciók.

Bizonyítás. Egy k változós elemi konjunkciót 2^{n-k} teljes elemi konjunkció összevonásával (az $n-k$ változó kiegyszerűsítésével) kaphatunk meg. Ha ilyenek szerepelnek a t_i-k, illetve d_j-k között, akkor a megfelelő teljes elemi konjunkciók szerepelnek f, illetve f^* mintermes felírásban. Mivel a T táblázatból az egyes változók kiegyszerűsíthetetlenségére vonatkozó információt használjuk fel, a kiegyszerűsíthető változókra (a mintermes T táblázatban) nyújtott információ érdektelen.

2.7. TÉTEL. Legyen T az f_φ-hez tartozó táblázat. Ekkor

1. $$f_T = \bigvee_{j=1}^{n} \bigwedge_{i=1}^{s} e_{ij} \quad \text{kompatibilis } f_\varphi\text{-vel},$$

2. $$f_{T\min} = \bigvee_{j=1}^{n} \bigwedge_{x_k^{\alpha_k} \in E_j} x_k^{\alpha_k} \quad \text{kompatibilis } f_\varphi\text{-vel},$$

3. $$\bigwedge_{x_k^{\alpha_k} \in E_j} x_k^{\alpha_k} \quad \text{primimplikánsa } f_\varphi\text{-nek}.$$

Bizonyítás. 1. A $K = \bigwedge_{i=1}^{s} e_{ij}$ konjunkcióra igaz, hogy $t_j \to K$. A $K \to f_\varphi$-hez még azt kell belátni, hogy K nem fed le 0-pontot f_φ-ben. Ez a dualitás miatt fennáll,

mivel e_{ij} minden eleme d_i-ben is szerepel, a dualitás miatt viszont az f_0 termjei $d_1^-, \ldots, d_s^-$.

2. A $K_{\min} = \bigwedge_{x_a^{\alpha_k} \in E_j} x_k^{\alpha_k}$ konjunkcióra igaz, hogy $t_j \to K_{\min}$. A $K_{\min} \to f_\varphi$ hasonlóan látható be, mint 1-ben.

3. E_j definíciójából következik, hogy a konjunkcióból változó nem hagyható el.

3. T-rekurzív definíció és algoritmus

Ezek után a T táblázat tulajdonságait felhasználva megadjuk f_φ egy nem-redundáns lefedéséhez szükséges prímimplikánsok *rekurzív* definícióját. Másfajta rekurzív definíciókra lásd [3, 4, 5]-öt.

Jelölje $\mathrm{IRRP}(f/\hat{f}^*)$ az f_φ irredundáns lefedéséhez szükséges prímimplikánsok halmazát. 1, illetve 0 jelölje azt a függvényt, ami mindenütt 1, illetve sehol sem 1. $\mathrm{PT}(t/\hat{f}^*)$ jelöljön egy, a t termet lefedő prímimplikánst, amelyet az $\hat{f}^*$ korlátoz. A T-rekurzív definíció:

1. $\mathrm{IRRP}(0/h) = 0$
2. $\mathrm{IRRP}(\underset{\sim}{f}/\hat{f}^*) = \mathrm{PT}(t_i/\hat{f}^*) \vee \mathrm{IRRP}(\underset{\sim}{f} \setminus \mathrm{PT}(t_i/\hat{f}^*)/\hat{f}^*)$
3. $\mathrm{PT}(t_i/0) = 1$
4. $\mathrm{PT}(1/0) = 1$
5. $\mathrm{PT}(t_i/\hat{f}^*) = x_{i,\min}^{\alpha_i} \in t_i \wedge \mathrm{PT}(t_{i,x_i=\alpha_i}/\hat{f}^* \setminus \{d_k | x_i^{\alpha_i} \in d_k\})$

Szavakban az egyes sorok jelentése. 1. Az 1-pont nélküli függvénynek nincs prímimplikánsa. 2. Az irredundáns lefedés prímimplikánsai a függvény egy valódi termjét lefedő p prímimplikánsból és az f-ből a p-által lefedett 1 pontok elhagyása után kapott függvény irredundáns lefedését adó prímimplikánsokból áll. 3. és 4. azt fejezi ki, hogy egy 0-pont nélküli függvény bármely termjét az 1 függvény lefedi. 5. A t_i termet lefedő prímimplikáns tényezői az $x_{i,\min}^{\alpha_i}$-ből és $t_{i,x_i=\alpha_i}$ termet a módosított korlátozó tartomány mellett lefedő prímimplikáns tényezőiből állnak.

A következő algoritmus a rekurzív definíció alapján, adott T táblázat esetén előállít egy irredundáns lefedést.

1. $k=0$.
2. Ha T-nek nincs oszlopa, akkor a prímimplikánsok száma k és $P(1)-P(k)$-ban találhatók. Vége.
3. $k=k+1$, jelölje O_1 a T első oszlopát. $P(k)=1$.
4. Válasszunk ki O_1-ből egy $x_{i,\min}^{\alpha_i}$ változót. $P(k)=P(k) \wedge x_i^{\alpha_i}$. Töröljük O_1-ből az $x_i^{\alpha_i}$-t tartalmazó elemeket.
5. Ha O_1 nem üres, akkor a 4 lépés következik.
6. Hagyjuk el T azon oszlopait, amelyeket $P(k)$ lefed. A 2. lépés következik.

4. A T-rekurzív közelítés értékelése

A T-rekurzív definíció az eddig elért legegyszerűbb alakú, és szinte változtatás nélkül adja a logikai programozással való megoldást. Ennek leírásához vezessük be a következő jelöléseket:

I	— a táblázat oszlopainak sorszáma,
Ti	— az I-edik oszlophoz rendelt konjunkció (kiválasztásának jele $k(I, Ti)$),
Im	— prímimplikáns jele,
oszl $(I, \ldots)$	— a táblázat I-edik oszlopa,
minvalt (I, X)	— X az I-edik oszlopban az $x_i^{\alpha i}{}_{\min}$,
elokeszit (I, Ti)	— a táblázat I-edik oszlopának előkészítése,
prim (Im, I)	— az I-edik oszlop alapján kapott prímimplikáns (a T-rekurzív definícióban PT) összeszerkesztése.

Ezek után egy MPROLOG leírás a következő:

1. *irrp*: — *pt*(*Im*), *tabl_mod*(*Im*), *irrp*.
2. *irrp*: — *not k*(*I*, *L*), *result*.
   (A T-rekurzív definíció 2. és 1. sorának megfelelően.)
3. *pt*(*Im*): — *elokeszit*(*I*, *Ti*), *prim*(*Im*, *Ti*).
4. *prim*(*nil*, *I*): — *not oszl*(*I*, *J*, *L*, *S*), !.
5. *prim*($X \cdot L$, *I*): — *minvalt*(*I*, *X*), *prim*(*L*, *I*).
   (A T-rekurzív definíció 3., 4., 5. sorát fejezik ki.)

A program részletezése túlmutat a hely és tematika keretein. A program léte viszont mutatja, hogy a logikai programozás eszközeit valódi feladatok megoldására itt is lehet használni, ha megtaláljuk a megfelelő megfogalmazást.

A T-rekurzív definíció alapján kidolgozott hagyományos algoritmus minden korábbinál egyszerűbb, bonyolultsága kisebb az előzőknél, és számítógépes realizációjában mindössze az adatkezelést kell megoldani.

A korábbi algoritmusok két csoportba oszthatók.

1. Az $ax \vee a\bar{x} = a$ azonosságot felhasználó, [6]-ra épülő változatok.
2. A [4]-beli alaperedményt finomító változatok [5, 7].

Jelöljük $b(n)$-nel az algoritmus bonyolultságát n-változós függvény esetén, ahol N az f_1, K az f_0 1-pontjainak száma. Az algoritmusok bonyolultságára a következő becslés adható:

— Az 1. esetben $\binom{N}{2} \leqq b(n) < N^{2^n}$.

— A 2. esetben $n \cdot (N \cdot K + N + K) \leqq b(n) < n \cdot N(N \cdot K + N + K)$.

— Az itt megadott algoritmusban $N + K < b(n) < n \cdot N \cdot K + \sum_{i=1}^{n} (N - i)$.

IRODALOM

[1] PÁSZTORNÉ Varga, K., „Módszerek Boole-függvények minimális vagy nemredundáns {∧, ∨, ¬}, vagy {NOR}, vagy {NAND} bázisbeli zárójeles, vagy zárójel nélküli formuláinak előállítására", *MTA SZTAKI Tanulmányok*, 1/1973.

[2] PÁSZTORNÉ VARGA, K., „Nem teljesen meghatározott Boole-függvények szintézise {∧, ∨, ¬}, {NAND}, illetve {NOR} bázisban", *MTA Számítástechnikai Központ Közlemények*, 1972/3, 17—31.

[3] MORREALE, E., "Partitioned list algorithms for prime implicant determination from canonical form", *IEEE Trans. E. C.* EC—16 (1967) 611—620.

[4] MORREALE, E., "Recursive operators for prime implicants and irredundant normal form determination", *IEEE Transaction on Computers* C—19 (1970) 504—509.

[5] TORGUE, I. C., PÁSZTOR, K. and AZEMA, P., «Couverture irredondante des fonctions booléennes définies par leur monômes vrais et faux-fonctions simultanéèes», *Acta Cybernetica* 2 (1975).

[6] QUINE, W. V., "The problem of simplifying truth functions", *American Mathematical Monthly* 59 (1952) 521—531.

[7] BRAYTON, R. K., HACHTEL, G. D., MCMULLEN, C. T. and SANGIOVANNI-VINCENTELLI, A. L., *Logic Minimisation Algorithms for VLSI Synthesis* (Kluwer Academic Publ., 1984).

(Beérkezett: 1986. szeptember 8.)

PÁSZTORNÉ VARGA KATALIN
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1111 BUDAPEST, KENDE U. 13—17.

AN OPTIMIZING METHOD FOR BOOLEAN FUNCTIONS BASED ON STRUCTURAL FEATURES OF THE BOOLEAN ALGEBRA OF BOOLEAN FUNCTIONS

K. PÁSZTOR-VARGA

As a preparation, some properties of the *Boolean-algebra* of C_1^n are described, and several new relations on duality are demonstrated. These results facilitate some important simplifications in the function synthesis. An effective synthesis algorithm is obtained and a realisation by logical program is shown. A comparison of the complexity of different kinds of synthesis algorithms is given.

Alkalmazott Matematikai Lapok 13 (1987—88) 77—81

ELÉGSÉGES KRITÉRIUM MÁSODRENDŰ DIFFERENCIÁLEGYENLET ZÉRUS EGYENSÚLYI HELYZETÉNEK GLOBÁLIS ASZIMPTOTIKUS STABILITÁSÁRA

KALOCSAI VIKTOR

Budapest

Két paramétert tartalmazó kvadratikus *Ljapunov-függvény* alkalmazásával igen általános elégséges kritériumot adunk az

$$y''+p(t,y,y')y'+q(t,y,y')y=0$$

egyenlet zérus egyensúlyi helyzetének globális aszimptotikus stabilitására.

Elméleti szempontból és az alkalmazásokat tekintve egyaránt igen fontos feladat, hogy az

$$y''=f(t,y,y');\quad f(t,0,0)\equiv 0;\quad t\geqq t_0$$

differenciálegyenlet zérus egyensúlyi helyzetének aszimptotikus stabilitására minél tágabb elégséges kritériumot adjunk. (Az irodalomban csak meglehetősen speciális elégséges kritériumok léteznek, lásd [1, 2, 3].) Jelen dolgozatban feltételezzük, hogy a fenti egyenlet

$$y''+p(t,y,y')y'+q(t,y,y')y=0;\quad t\geqq t_0 \tag{1}$$

alakban adott, ahol p és q folytonos függvények, és a megoldások $[t_0,\infty)$-ben értelmezettek. Két paramétert tartalmazó kvadratikus *Ljapunov-függvény* alkalmazásával eléggé általános és igen egyszerűen ellenőrizhető elégséges kritériumot sikerült adni.

Eredményünket az alábbi (fő) tételben foglaljuk össze.

1. TÉTEL. Legyenek ϱ, δ és ε tetszőleges pozitív számok. Tekintsük az alábbi H halmazt:

$$H=\left\{(u,v)\in\mathbf{R}^2\middle| u=\frac{((\varrho+1)\tilde{u}+\delta\tilde{v})}{\sqrt{(\varrho+1)^2+\delta^2}};\quad v=\frac{(-\delta\tilde{u}+(\varrho+1)\tilde{v})}{\sqrt{(\varrho+1)^2+\delta^2}};\quad \tilde{u}\in\mathbf{R};\right.$$

$$\left.\tilde{v}>\frac{\sqrt{(\varrho+1)^2+\delta^2}}{4\delta\varrho}\left(\tilde{u}-\frac{\varrho(\varrho+1)-\delta^2}{\sqrt{(\varrho+1)^2+\delta^2}}\right)^2+\frac{\delta(\varrho+1)}{\sqrt{(\varrho+1)^2+\delta^2}}+\frac{\varepsilon}{4\delta\varrho\sqrt{(\varrho+1)^2+\delta^2}}\right\}. \tag{2}$$

Ha bármely $t_0\leqq t$-re, és bármely $r,s\in\mathbf{R}$ számokra igaz, hogy

$$(q(t,r,s),\ p(t,r,s))\in H,$$

akkor (1) zérus egyensúlyi helyzete globálisan aszimptotikusan stabilis.

Bizonyítás. (1)-re alkalmazzuk a

$$v(r, s) = (rs)\mathbf{V}\begin{pmatrix} r \\ s \end{pmatrix}$$

Ljapunov-függvényt, ahol

$$\mathbf{V} = \begin{bmatrix} \frac{1+\varrho}{\delta} + \frac{\delta}{\varrho} & \frac{1}{\varrho} \\ \frac{1}{\varrho} & \frac{1+\varrho}{\delta\varrho} \end{bmatrix}. \tag{3}$$

Nyilvánvalóan

$$\dot{v}_{(1)}(r, s) = (rs)\mathbf{W}\begin{pmatrix} r \\ s \end{pmatrix},$$

ahol

$$\mathbf{W} = \begin{bmatrix} \alpha & \gamma \\ \gamma & \beta \end{bmatrix},$$

$$\alpha = -\frac{q}{\varrho},$$

$$\beta = \frac{1}{\varrho} - \frac{1+\varrho}{\delta\varrho} p,$$

$$\gamma = \frac{1}{2}\left(\frac{1+\varrho}{\delta} + \frac{\delta}{\varrho} - \frac{p}{\varrho} + \frac{1+\varrho}{\delta\varrho} q\right).$$

Itt

$$p = p(t, r, s)$$

és

$$q = q(t, r, s),$$

t az a szám, amelyre $r=y(t)$, $s=y'(t)$; $y(t)$ (1)-nek az a megoldása, amely mentén v-t deriváljuk.

A zérus egyensúlyi helyzet globális aszimptotikus stabilitásának egy elégséges feltétele, hogy létezzék $\varepsilon_1>0$, amelyre

$$\dot{v}_{(1)}(r, s) \leqq -\varepsilon_1(r^2+s^2).$$

Az utóbbi viszont akkor teljesül, ha

$$\operatorname{tr}\mathbf{W} + \sqrt{(\operatorname{tr}\mathbf{W})^2 - 4\det\mathbf{W}} \leqq -2\varepsilon_1,$$

ahol $\operatorname{tr}\mathbf{W}=\alpha+\beta$ és $\det\mathbf{W}=\alpha\beta-\gamma^2$. Ilyen ε_1 létezik, ha létezik $\varepsilon_2>0$ és $\varepsilon>0$, amelyre

$$\operatorname{tr}\mathbf{W} < -\varepsilon_2,$$

$$\det\mathbf{W} > \varepsilon.$$

A két alábbi egyenlőtlenség a következő egyenlőtlenségekkel ekvivalens:

$$f(q, p) = -\frac{q}{\varrho} + \frac{1}{\varrho} - \frac{1+\varrho}{\delta\varrho} p < -\varepsilon_2, \tag{4}$$

$$g(q, p) = -\frac{q}{\varrho}\left(\frac{1}{\varrho} - \frac{1+\varrho}{\delta\varrho} p\right) - \left[\frac{1}{2}\left(\frac{1+\varrho}{\delta} + \frac{\delta}{\varrho} - \frac{p}{\varrho} - \frac{1+\varrho}{\delta\varrho} q\right)\right]^2 > \varepsilon. \tag{5}$$

Hosszadalmas és fáradságos számításokkal (e számítások ugyanakkor elemiek, így a részletektől eltekintünk) meggyőződhetünk arról, hogy (5) teljesülése esetén létezik $\varepsilon_2>0$, amellyel (4) igaz. Az (u, v) síkon az $f(u, v)=0$ és $g(u, v)=0$ egyenletű vonalak az 1. ábrán láthatók. A $g(u, v)>\varepsilon>0$ tartomány az 1. ábrán látható parabolával határolt $\det \mathbf{W}>0$ tartomány valódi részhalmaza. Ennek meghatározásához nem kell mást tennünk, mint a $\det \mathbf{W}-\varepsilon$ kifejezésen (amely u-ra és v-re kvadratikus alak) főtengelytranszformációt kell végrehajtani. A keresett tartomány határát a $\det \mathbf{W}-\varepsilon=0$ egyenletnek megfelelően azok az (u, v) koordinátájú pontok alkotják. amelyekre:

$$(6) \qquad -(\varrho+1)^2u^2-\delta^2v^2+(2\delta\varrho+2\delta)uv+(2\delta^2\varrho+2\varrho^3+4\varrho^2+2\varrho-2\delta^2)u+$$
$$+(2\varrho\delta+2\varrho^2\delta+2\delta^3)v-\varrho^4-\varrho^2-2\varrho^3-\delta^4-2\varrho\delta^2-2\varrho^2\delta^2-\varepsilon=0.$$

Ha a főtengely-transzformációnak megfelelően új, $\tilde{u}$, $\tilde{v}$ változókra térünk át

$$\begin{pmatrix} u \\ v \end{pmatrix} = \frac{1}{\sqrt{(\varrho+1)^2+\delta^2}} \begin{pmatrix} \varrho+1 & \delta \\ -\delta & \varrho+1 \end{pmatrix} \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix},$$

akkor ugyancsak hosszadalmas, de elemi átalakítások után a (6) egyenlet eképpen módosul:

$$(8) \qquad \tilde{v}=\frac{\sqrt{(\varrho+1)^2+\delta^2}}{4\delta\varrho}\left(\tilde{u}-\frac{\varrho(\varrho+1)-\delta^2}{\sqrt{(\varrho+1)^2+\delta^2}}\right)^2+$$
$$+\frac{\delta(\varrho+1)}{\sqrt{(\varrho+1)^2+\delta^2}}+\frac{\varepsilon}{4\delta\varrho\sqrt{(\varrho+1)^2+\delta^2}},$$

amelyből a tétel állítása már egyszerűen következik.

Megjegyzések.

1. Ha $p\equiv\delta$ és $q\equiv\varrho$, akkor $\mathbf{W}=-\mathbf{I}$. Éppen az az összefüggés vezetett $\mathbf{V}$ (3) szerinti meghatározásához.

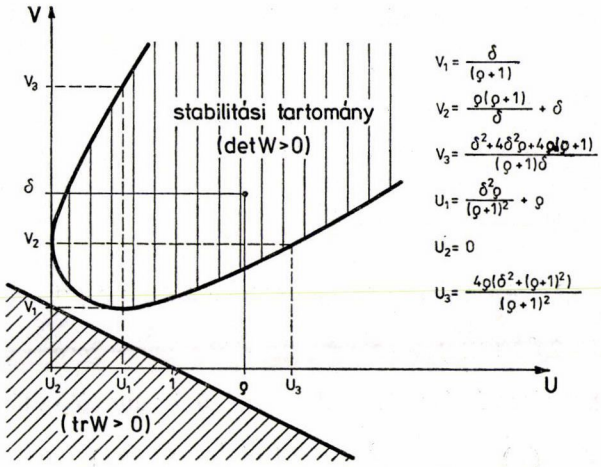
2. Az 1. megjegyzés tükrében tételünk a következőképpen is interpretálható. Az $y''+\delta y'+\varrho y=0$ állandó együtthatós homogén lineáris differenciálegyenlet $y=y'=0$ egyensúlyi helyzete nyilván aszimptotikusan stabilis (ne felejtsük el, hogy $\delta>0$, $\varrho>0$!). Ezt a (3) együtthatójú kvadratikus *Ljapunov-függvénnyel* ki lehet mutatni. Ha most ugyanezt a *Ljapunov-függvényt* alkalmazzuk az

$$y''+\delta y'+\varrho y+(p(t, y, y')-\delta)y'+(q(t, y, y')-\varrho)y=0$$

perturbált egyenletre, akkor meglepő módon az derül ki, hogy igen nagy, akár nem korlátos perturbációk esetén is megmarad az aszimptotikus stabilitás. A stabilitási tartomány az 1. ábra $\det \mathbf{W}>0$ tartománya.

3. Adott p, q függvények esetén a tétel alkalmazhatósága ϱ és δ alkalmas választásától is függ. Többféle (ϱ, δ) párok választásával *előre számított stabilitási tartományok* vehetők fel az (u, v) síkon. Ennek komoly gyakorlati jelentősége lehet.

4. Ha a tételbe $\varepsilon=0$-t helyettesítünk és az aszimptotikus stabilitást stabilitásra cseréljük, akkor az állítás igaz marad.



1. ábra

A δ és ϱ paraméterek optimalizálásának egy lehetséges módját példaképpen arra az esetre mutatjuk meg, amikor

$$p(t, y, y') = p = \text{állandó} > 0.$$

Hajtsuk végre a $\tau = at$; $a > 0$ időtranszformációt. Ekkor egyenletünk a következő alakot nyeri:

$$\ddot{y} + \frac{p}{a}\dot{y} + \frac{q(t, y, y')}{a^2} y = 0,$$

ahol

$$\dot{y} = \frac{dy}{dt}.$$

Legyen $\frac{p}{a} = \frac{\varrho(\varrho+1)}{\delta} + \delta$, amelyből

$$a = \frac{\delta p}{\delta^2 + \varrho(\varrho+1)}.$$

A zérus egyensúlyi helyzet globális aszimptotikus stabilitásának nyilvánvaló feltétele (lásd az 1. ábrát), hogy létezzék $\varepsilon > 0$, amelyre

$$\frac{\varepsilon}{a^2} < \frac{q(t, y, y')}{a^2} < \frac{4\varrho(\delta^2 + (1+\varrho)^2)}{(1+\varrho)^2} - \frac{\varepsilon}{a^2},$$

vagyis

$$\varepsilon < q(t, y, y') < \frac{4\varrho(\delta^2 + (1+\varrho)^2)\delta^2}{(1+\varrho)^2(\delta^2 + \varrho(\varrho+1))^2} p^2 - \varepsilon. \tag{10}$$

Elemi számítások sorozata után kimutatható, hogy

$$\delta^2 = \frac{\varrho(1+\varrho)^2}{1-\varrho}$$

választás mellett a (10)-ben szereplő felső határ maximális lesz. Ekkor

$$\varepsilon < q(t, y, y') < p^2 - \varepsilon.$$

Eredményünket az alábbi tételben foglalhatjuk össze:

2. TÉTEL. (1)-ben legyen $p(t, y, y') \equiv p = \text{áll.} > 0$. Ha létezik $\varepsilon > 0$, amelyre teljesül, hogy

$$\varepsilon < q(t, r, s) < p^2 - \varepsilon, \quad t \geqq t_0, \quad r \in \mathbf{R}, \quad s \in \mathbf{R},$$

akkor (1) zérus egyensúlyi helyzete globálisan aszimptotikusan stabilis.

IRODALOM

[1] HATVANI, L., „Nem-autonom differenciálegyenlet-rendszerek megoldásainak stabilitása és parciális stabilitása", *Alkalmazott Matematikai Lapok* **5** (1979) 1—48.
[2] KARSAI, J., "On the global asymptotic stability of the zero solution of the equation $\ddot{x}+g(t, x, \dot{x})\dot{x}+f(x)=0$", *Studia Scientiarum Mathematicae Hungarica* **19** (1984) 385—393.
[3] KERTÉSZ, V., „Pozitív definit kvadratikus Ljapunov-függvények alkalmazása stabilitási vizsgálatokhoz", *Alkalmazott Matematikai Lapok* **9** (1983) 375—386.

(Beérkezett: 1987. február 3.)

KALOCSAI VIKTOR
1098 BUDAPEST, BÖRZSÖNY U. 2/C II. 15.

SUFFICIENT CONDITION ON THE GLOBAL ASYMPTOTICAL STABILITY OF THE ZERO EQUILIBRIUM OF A SECOND ORDER DIFFERENTIAL EQUATION

V. KALOCSAI

Applying a quadratic *Lyapunof function* of two parameters, there is given a rather general sufficient condition for the global asymptotic stability of the zero equilibrium of the second order differential equation $y''+p(t, y, y')y'+q(t, y, y')y=0$.

Alkalmazott Matematikai Lapok 13 (1987—88) 83—96

ÁLTALÁNOS HESTENES—STIEFEL TÍPUSÚ KONJUGÁLT IRÁNY REKURZIÓK I. A DIREKT REKURZIÓ

HEGEDŰS CSABA J.

Budapest

Tetszőleges **A** mátrixra nézve konjugált vektorpárokat generáló általános rekurziót ismertetünk, mely elméletileg akkor fejeződik be, ha az utolsónak generált pár elemei egyidejűleg az $\mathbf{A}^H$, ill. **A** mátrix nullterébe esnek.

1. Bevezetés

A konjugált irány módszerek vizsgálata HESTENES és STIEFEL [8], valamint LÁNCZOS [11] ma már klasszikusnak tekinthető munkáival vette kezdetét. A módszerek vizsgálata az utóbbi időben nagyon intenzívvé vált. Ennek oka, hogy a nagy ritkamátrixos feladatoknál sokrétűen alkalmazhatók, így lineáris egyenletrendszerek, szinguláris érték feladatok és sajátérték-feladatok megoldására; de e módszerekkel sikerre számíthatunk nemlineáris feladatoknál is, mint például a nemlineáris egyenletrendszerek és szélsőérték feladatok megoldásánál.

Nagy ritkamátrixú lineáris egyenletrendszereknél az egyik közkedvelt módszer szerint az $\mathbf{Ax}=\mathbf{b}$ lineáris egyenletrendszert beszorozzuk egy **B** ún. prekondícionáló mátrixszal, amelyre **BA** közel egységmátrix, legalábbis a kondíciószáma jóval kisebb, mint az **A** mátrixé. A konjugált gradiens módszert ezután a $\mathbf{BAx}=\mathbf{Bb}$ egyenletre alkalmazzuk. Ha a **B** mátrix elég jó, akkor elegendő néhány lépés a megoldás eléréséhez. A gyakorlati kivitelezésnél nem szükséges a **BA** mátrixszorzatot elkészíteni. A **B** mátrix meghatározásának egyik sikeres módja szerint az **A** mátrix LU-faktorizációját $\tilde{\mathbf{L}}\tilde{\mathbf{U}}$ alakban közelítjük, $(\mathbf{B}=\tilde{\mathbf{U}}^{-1}\tilde{\mathbf{L}}^{-1})$, ahol $\tilde{\mathbf{L}}$ és $\tilde{\mathbf{U}}$ úgy készül, hogy a mátrixok *ij*-edik eleme zérus, ha az **A** mátrix a_{ij} eleme zérus. Innen a nemteljes faktorizáció *(incomplete factorization)* elnevezés. A módszert illetően l. [1], [4], [13], [15], [16].

A konjugált irány módszerek különféle változatairól és alkalmazásairól jó képet kaphatunk STOER [15] áttekintő cikke nyomán, valamint HESTENES [9] könyvéből.

Az alkalmazásoknál leggyakrabban a konjugált gradiens módszerrel találkozunk, amely szimmetrikus, pozitív definit együtthatómátrix mellett alkalmazható lineáris egyenletrendszer megoldására. A konvergenciát ekkor az

$$\|\mathbf{x}_k-\mathbf{h}\|_2 \leqq \alpha\{(\sqrt{\varkappa}-1)/(\sqrt{\varkappa}+1)\}^k$$

egyenlőtlenség jellemzi, ahol k a lépésszám, α valamely állandó, $\mathbf{x}_k$ a megoldás k-adik közelítése, **h** az $\mathbf{Ax}=\mathbf{b}$ egyenlet megoldása és $\varkappa$ az **A** mátrix kondíciószáma [2], [10].

A konjugált gradiens módszer közvetlenül indefinit szimmetrikus mátrixokra is alkalmazható, ekkor azonban nincs garancia arra vonatkozóan, hogy a formulákban a nevező nem válhat zérussá. Pedig a módszer csábító volna, mert az áttérés az $\mathbf{A}^H\mathbf{A}\mathbf{x}=\mathbf{A}^H\mathbf{b}$ normálegyenletre numerikus többletmunkát igényel, és a kondíciószám négyzetelésével óhatatlanul csökkenti a konvergencia sebességét.

LUENBERGER [12] 1969-ben javasolta a hiperbolikus párok módszerét arra az esetre, amikor a konjugált gradiens módszert indefinit valós szimmetrikus mátrixra alkalmazzuk, és a nevező valamely lépésben pontosan zérus. A módszerrel később FLETCHER [3] foglalkozott, és nem találta megfelelőnek, minthogy abban a fontos gyakorlati esetben, amikor a nevező közel zérus, nem alkalmazható.

Ebben a cikksorozatban olyan *általánosított Hestenes—Stiefel-típusú konjugált irány rekurziókat* fogunk ismertetni, amelyek téglalap alakú mátrixokra a kondíciószám négyzetelődése nélkül alkalmazhatók, s közel zérus nevező esetére is adunk eljárást az előnyösebb folytatásra. A sorozat első néhány cikke a rekurziók fajtáival és azok tulajdonságaival foglalkozik.

A továbbiakban ismertetünk néhány, a konjugált irányokra vonatkozó fogalmat és állítást, amelyekre később szükségünk lesz. A tárgyalást az $\mathbf{A}\in C^{m,n}$ komplex elemű mátrixokra végezzük. A második fejezetben adott rekurzió-családot a „direkt" jelzővel láttuk el, mert ehhez képest a következő cikkben ismertetendő (fordított rekurzió-) családban bizonyos elemek fordított sorrendben fognak szerepelni. A direkt rekurzió tulajdonképpen a [7]-ben adott *Hestenes—Stiefel-típusú rekurzió* javítása, amiből speciális esetben kiadódik a klasszikus konjugált gradiens módszer. A fordított rekurzió új, eddig még nem alkalmazott módszereket tartalmaz.

Az $\mathbf{A}\in C^{m,n}$ mátrixra nézve a $\mathbf{v}_j\in C^m$ és $\mathbf{u}_k\in C^n$ vektorokat konjugált irányoknak vagy pároknak nevezzük, ha minden j, k indexre teljesül

$$\mathbf{v}_j^H\mathbf{A}\mathbf{u}_k = \alpha_j\delta_{jk}, \quad \alpha_j \neq 0, \quad j,k = 1,2,\ldots, \tag{1.1}$$

ahol H a transzponált konjugáltat jelöli és δ_{jk} a *Kronecker delta*. Annak hangsúlyozására, hogy egy vektorpár-rendszer elemei az $\mathbf{A}$ mátrixra konjugáltak, az *„A-konjugált"* kifejezést fogjuk használni. Ha $\mathbf{A}$ hermitikus, pozitív definit és $\mathbf{v}_j=\mathbf{u}_j$ minden j-re, akkor a $\mathbf{v}_j$ (vagy $\mathbf{u}_j$) vektorokat *A-ortogonális vektoroknak* nevezzük. Ha még $\alpha_j=1$ minden j-re, akkor valós esetben *A-ortonormált* rendszerről, komplex vektoroknál pedig *A-unitér* vektorrendszerről beszélhetünk, [14, 200. old.].

Maximálisan annyi *A-konjugált pár* állítható elő, mint amennyi az $\mathbf{A}$ mátrix rangja [5], [6], [7].

1.1. TÉTEL ([5], [6], [7]). Ha bevezetjük a

$$\mathbf{P}_i^l = \mathbf{I}_m - \sum_{j=1}^{i} \mathbf{A}\mathbf{u}_j\mathbf{v}_j^H/\mathbf{v}_j^H\mathbf{A}\mathbf{u}_j \tag{1.2}$$

$$\mathbf{P}_i^r = \mathbf{I}_n - \sum_{j=1}^{i} \mathbf{u}_j\mathbf{v}_j^H\mathbf{A}/\mathbf{v}_j^H\mathbf{A}\mathbf{u}_j \tag{1.3}$$

projektorokat, ahol $\mathbf{I}_m$ m-edrendű egységmátrix, akkor az i-elemű konjugált pár

rendszer tetszőleges $\mathbf{r}_{i+1}$ és $\mathbf{q}_{i+1}$ segédvektorokkal a

$$\mathbf{v}_{i+1}^H = \mathbf{r}_{i+1}^H \mathbf{P}_i^l \tag{1.4}$$

$$\mathbf{u}_{i+1} = \mathbf{P}_i^r \mathbf{q}_{i+1} \tag{1.5}$$

összefüggések szerint eggyel bővíthető, amennyiben

$$\mathbf{r}_{i+1}^H \mathbf{P}_i^l \mathbf{A} \mathbf{P}_i^r \mathbf{q}_{i+1} \neq 0. \tag{1.6}$$

1.2. TÉTEL ([6], [7]). Ha az $\mathbf{r}_{i+1}$ és $\mathbf{q}_{i+1}$ vektorokat is vetítéssel állítjuk elő, akkor valamely nemzérus $\mathbf{r}_0 \in C^m$ és $\mathbf{q}_0 \in C^n$ kezdővektorokból kiindulva az

$$\begin{gathered} \mathbf{r}_{i+1} = \mathbf{P}_i^l \mathbf{r}_i, \quad \mathbf{q}_{i+1}^H = \mathbf{q}_i^H \mathbf{P}_i^r, \\ \mathbf{v}_{i+1}^H = \mu_{i+1} \mathbf{r}_{i+1}^H \mathbf{C} \mathbf{P}_i^l, \quad \mathbf{u}_{i+1} = \lambda_{i+1} \mathbf{P}_i^r \mathbf{K} \mathbf{q}_{i+1} \end{gathered} \tag{1.7}$$

összefüggésekkel rekurzíve egy konjugált pár rendszert készíthetünk, ahol $\mathbf{C} \in C^{m,m}$ és $\mathbf{K} \in C^{n,n}$ hermitikus, pozitív definit, de egyébként tetszőleges mátrixok és μ_i, λ_i skálázást végző nemzérus skalárok.

Az idézett dolgozatokban megmutattuk, hogy (1.7) annyira egyszerűsíthető, hogy az $\mathbf{r}_{i+1}$, $\mathbf{q}_{i+1}$, $\mathbf{v}_{i+1}$ és $\mathbf{u}_{i+1}$ vektorok előállításához elegendő $\mathbf{r}_i$, $\mathbf{q}_i$, $\mathbf{v}_i$, $\mathbf{u}_i$ ismerete, továbbá ha a vektorokat az

$$\begin{gathered} \mathbf{R} = [\mathbf{r}_1 \mathbf{r}_2 \ldots \mathbf{r}_i], \quad \mathbf{Q} = [\mathbf{q}_1 \mathbf{q}_2 \ldots \mathbf{q}_i], \\ \mathbf{V} = [\mathbf{v}_1 \mathbf{v}_2 \ldots \mathbf{v}_i], \quad \mathbf{U} = [\mathbf{u}_1 \mathbf{u}_2 \ldots \mathbf{u}_i] \end{gathered} \tag{1.8}$$

mátrixokba rendezzük, akkor az (1.7) rekurzió a következő mátrixegyenletekkel is reprezentálható [7]:

$$\mathbf{A} \mathbf{U} \mathbf{D}_a^{-1} \mathbf{M} \mathbf{D}_r = \mathbf{R} \mathbf{S}^{-1} - \mathbf{r}_{i+1} \mathbf{e}_i^T, \tag{1.9}$$

$$\mathbf{D}_q \Lambda \mathbf{D}_a^{-1} \mathbf{V}^H \mathbf{A} = \mathbf{S}^{-T} \mathbf{Q}^H - \mathbf{e}_i \mathbf{q}_{i+1}^H, \quad \mathbf{S}^{-T} = (\mathbf{S}^{-1})^T, \tag{1.10}$$

$$\mathbf{D}_r \mathbf{S}^{-1} \mathbf{D}_r^{-1} \mathbf{M}^{-1} \mathbf{V}^H = \mathbf{R}^H \mathbf{C}, \tag{1.11}$$

$$\mathbf{U} \Lambda^{-1} \mathbf{D}_q^{-1} \mathbf{S}^{-T} \mathbf{D}_q = \mathbf{K} \mathbf{Q}, \tag{1.12}$$

$$\mathbf{D}_a = \mathbf{V}^H \mathbf{A} \mathbf{U}, \tag{1.13}$$

$$\mathbf{D}_r = \mathbf{R}^H \mathbf{C} \mathbf{R} \tag{1.14}$$

$$\mathbf{D}_q = \mathbf{Q}^H \mathbf{K} \mathbf{Q}, \tag{1.15}$$

ahol $\mathbf{e}_i$ az i-edik egységvektor és $\mathbf{S}$ i-edrendű speciális alsó háromszögmátrix, amelynek alsó háromszögét csupa 1 elem tölti ki, tehát az inverze olyan bidiagonális mátrix, ahol a főátlóban 1-ek, a főátló alatt pedig -1-ek állnak. A $\mathbf{D}_a$, $\mathbf{D}_r$, $\mathbf{D}_q$ mátrixok diagonálmátrixok és a λ_i, μ_i skalárokat a $\mathbf{\Lambda} = \langle \lambda_j \rangle$, $\mathbf{M} = \langle \mu_i \rangle$ diagonálmátrixokba foglaltuk össze.

Az (1.7) rekurziót *Hestenes—Stiefel-típusú rekurziónak* nevezzük, egyrészt mivel speciális esetben innen kiadódnak a HESTENES és STIEFEL által adott konjugált gradiens módszer rekurziói [8], másrészt megkülönböztetésül a *Lánczos-típusú rekurziótól*, amelynél a $\mathbf{v}_{i+1}$ és $\mathbf{u}_{i+1}$ vektorok előállítása a

$$\mathbf{v}_{i+1}^H = \mu_{i+1} \mathbf{q}_{i+1}^H \mathbf{B} \mathbf{P}_i^l, \quad \mathbf{u}_{i+1} = \lambda_{i+1} \mathbf{P}_i^r \mathbf{B} \mathbf{r}_{i+1} \tag{1.16}$$

összefüggésekkel történik, ahol $\mathbf{B} \in C^{n,m}$ tetszőleges nemzérus mátrix. A *Hestenes—Stiefel-típusú rekurzió* az **A** mátrix szinguláris-érték feladatával rokon, míg a *Lánczos-típusú rekurzió* az **A** mátrix sajátértékproblémájával áll kapcsolatban. Szimmetrikus mátrix esetén a két rekurzió-típus egybeesik [7].

Jelölések. A mátrixokat nagybetűkkel, a vektorokat kisbetűkkel, a skalárokat — a dimenziók és indexek kivételével — görög kisbetűkkel jelöljük. A $\mathbf{v}_j$, $j=1, 2, \ldots, i$ vektorokat, vagy a $\mathbf{v}_j$, $\mathbf{u}_j$, $j=1, 2, \ldots, i$ vektorpárok rendszerét rövidített írásmóddal $\{\mathbf{v}_j\}_{j=1}^{i}$, ill. $\{\mathbf{v}_j, \mathbf{u}_j\}_{j=1}^{i}$ formában is jelölni fogjuk.

2. A direkt rekurzió

2.1. A rekurziós egyenletek általános alakja

Tekintsük az alábbi (1.9)—(1.15)-tel rokon rekurziós egyenletrendszert:

$$\mathbf{L}_i \mathbf{V}^H = \mathbf{D}_r^{-1} \mathbf{R}^H \mathbf{C}, \quad \mathbf{A}\mathbf{U}\mathbf{D}_a^{-1} = \mathbf{R}\mathbf{L}_i + \mathbf{r}_{i+1}\mathbf{l}_{i+1}^T, \tag{2.1}$$

$$\mathbf{D}_a^{-1} \mathbf{V}^H \mathbf{A} = \mathbf{F}_i \mathbf{Q}^H + \mathbf{f}_{i+1}\mathbf{q}_{i+1}^H, \quad \mathbf{U}\mathbf{F}_i = \mathbf{K}\mathbf{Q}\mathbf{D}_q^{-1}, \tag{2.2}$$

$$\mathbf{D}_a = \mathbf{V}^H \mathbf{A}\mathbf{U}, \quad \mathbf{D}_r = \mathbf{R}^H \mathbf{C}\mathbf{R}, \quad \mathbf{D}_q = \mathbf{Q}^H \mathbf{K}\mathbf{Q}, \tag{2.3}$$

ahol új mennyiségek az $\mathbf{L}_i$, $\mathbf{F}_i$ mátrixok és az $\mathbf{l}_{i+1}$, $\mathbf{f}_{i+1}$ vektorok. Általános esetben $\mathbf{L}_i$ i-edrendű alsó háromszögmátrix, $\mathbf{F}_i$ pedig i-edrendű felső háromszögmátrix és az eggyel magasabb rendű $\mathbf{L}_{i+1}$, $\mathbf{F}_{i+1}$ mátrixok az alábbi séma alapján készülnek:

$$\mathbf{L}_{i+1} = \begin{bmatrix} \mathbf{L}_i & 0 \\ \mathbf{l}_{i+1}^T & 1 \end{bmatrix}, \quad \mathbf{F}_{i+1} = \begin{bmatrix} \mathbf{F}_i & \mathbf{f}_{i+1} \\ 0 & 1 \end{bmatrix}, \tag{2.4}$$

ahol az $\mathbf{l}_{i+1}$ és $\mathbf{f}_{i+1}$ építővektorok i-elemű vektorok.

Összehasonlítva (2.1)—(2.3)-at az (1.9)—(1.15) egyenletrendszerrel, azt látjuk, hogy ott $\mathbf{L}_i = \mathbf{S}_i^{-1}\mathbf{D}_r^{-1}\mathbf{M}^{-1}$ alsó, $\mathbf{F}_i = \mathbf{\Lambda}^{-1}\mathbf{D}_q^{-1}\mathbf{S}^{-1}$ pedig felső bidiagonális mátrix, s a diagonálelemek sem szükségképpen egyenlőek 1-gyel.

Ez a fejezet az $\mathbf{L}_i$, $\mathbf{F}_i$ mátrixok olyan (2.4) szerinti bővítéseivel foglalkozik, amelyek mellett biztosított, hogy a $\{\mathbf{v}_j, \mathbf{u}_j\}$ rendszer *A-konjugált*, az $\{\mathbf{r}_j\}$ rendszer *C-ortogonális*, a $\{\mathbf{q}_j\}$ rendszer pedig *K-ortogonális*, továbbá a nullával osztás elkerülhető. A szemléletmód olyan, hogy mindig a (2.1)—(2.4) mátrixos alakokat vesszük alapul, s a kiadódó rekurziós összefüggéseket csak a mátrixegyenletek folyományának tekintjük. A rekurziós összefüggéseket az $\mathbf{L}_i$, $\mathbf{F}_i$ mátrixok viszik, emiatt ezeket a mátrixokat *rekurziós mátrixoknak* fogjuk nevezni.

Kényelmes bevezetni a következő konvenciót: Az, hogy az egyenletekben i-edrendű $\mathbf{L}_i$ és $\mathbf{F}_i$ mátrix szerepel, egyben jelentse azt is, hogy az **R**, **Q**, **V** és **U** mátrixok i számú oszlopvektort tartalmaznak, tehát, ha $\mathbf{A} \in C^{m,n}$, akkor **R**, $\mathbf{V} \in C^{m,i}$ és **Q**, $\mathbf{U} \in C^{n,i}$.

Célunk, hogy a rekurziós mátrixok minél egyszerűbbek legyenek. A nemzérus elemek a vektorok hosszát skálázzák, így szabadon választhatók. Emiatt a diagonálelemeket 1-nek vettük. Ez azt jelenti, hogy a $\{\mathbf{v}_j, \mathbf{u}_j\}$ *A-konjugált vektorok* hosszát

nem skálázzuk, csak az $\{\mathbf{r}_j, \mathbf{q}_j\}$ ortogonális vektorokét. A numerikus tapasztalatok szerint ez elegendő.

A továbbiakban közölt elmélet nem igényel többet az 1.1 tétel ismeretéénél, amelyet az Olvasó egyszerű ellenőrzéssel maga is bizonyíthat.

A rekurziós mátrixok speciális szerkezetéből következik, hogy valamely $\mathbf{r}_1, \mathbf{q}_1$ vektorokkal indulva a többi vektor rekurzíve számítható. Induláskor $i=1$-re (2.1) első, valamint (2.2) második egyenletéből kapjuk:

$$\mathbf{v}_1^H = \mathbf{r}_1^H \mathbf{C}/\|\mathbf{r}_1\|_C^2, \quad \mathbf{u}_1 = \mathbf{K}\mathbf{q}_1/\|\mathbf{q}_1\|_K^2, \tag{2.5}$$

ahol $\|\mathbf{r}_1\|_C^2 = \mathbf{r}_1^H \mathbf{C}\mathbf{r}_1$ és $\|\mathbf{q}_1\|_K^2 = \mathbf{q}_1^H \mathbf{K}\mathbf{q}_1$ energetikai vektornormákat jelölnek. Amennyiben $\mathbf{v}_1^H \mathbf{A}\mathbf{u}_1 \neq 0$, $\{\mathbf{v}_1, \mathbf{u}_1\}$ képezheti az első *A-konjugált párt*. Ha azonban $\mathbf{v}_1^H \mathbf{A}\mathbf{u}_1 = 0$, akkor át kell térni új vektorokra. Később látni fogjuk, hogy mindig tehető valami a rekurzió folytatására, ha az $\mathbf{A}\mathbf{u}_i$ vagy $\mathbf{v}_i^H \mathbf{A}$, $i=1, 2, \ldots$ vektoroknak legalább az egyike nemzérus vektor.

Zérus $\mathbf{v}_i^H \mathbf{A}\mathbf{u}_i$ érték két szempontból sem elfogadható. Egyrészt az *A-konjugált párok* definíciója ezt nem engedi meg, másrészt a (2.1)—(2.2) egyenletekben a $\mathbf{D}_a$ diagonálmátrix inverze szerepel, így a benne szereplő $\mathbf{v}_i^H \mathbf{A}\mathbf{u}_i$ mennyiségeknek nem szabad zérust adniok.

Az, hogy milyen lehetőségeink vannak a rekurzió folytatására, általában az i indexű vektorok mellett elmondható, emiatt $i=1$-et csak speciális esetnek tekintjük. Feltételezésünk szerint az $1, 2, \ldots, i$ indexű vektorok ismertek, rendelkeznek a kívánt ortogonalitási tulajdonságokkal, nevezetesen: a $\{\mathbf{v}_j, \mathbf{u}_j\}_{j=1}^i$ rendszer *A-konjugált,* az $\{\mathbf{r}_j\}_{j=1}^i$ rendszer *C-ortogonális* és a $\{\mathbf{q}_j\}_{j=1}^i$ rendszer *K-ortogonális,* továbbá még feltesszük, hogy az $\mathbf{L}_i$, $\mathbf{F}_i$ i-edrendű rekurziós mátrixok el vannak készítve. A következőkben az i-elemű vektorrendszerek eggyel bővítésével fogunk foglalkozni. Az alábbiakban három módszert ismertetünk.

2.2. Bidiagonális építés módszere

Ekkor mind az $\mathbf{L}$, mind az $\mathbf{F}$ mátrixot bidiagonálisan folytatjuk az $i+1$-edik lépésnél, így az $\mathbf{l}_{i+1}$ és $\mathbf{f}_{i+1}$ vektorok választása:

$$\mathbf{l}_{i+1}^T = -\lambda_{i+1}\mathbf{e}_i^T, \quad \mathbf{f}_{i+1} = -\varphi_{i+1}\mathbf{e}_i. \tag{2.6}$$

Szorozzuk (2.1) második egyenletét jobbról az $\mathbf{e}_i$ egységvektorral, (2.2) első egyenletét pedig balról az $\mathbf{e}_i^T$ vektorral. Felhasználva, hogy $\mathbf{D}_a$, $\mathbf{D}_r$ és $\mathbf{D}_q$ diagonálmátrixok az ortogonalitási feltevések folytán, az eredmény:

$$\lambda_{i+1}\mathbf{r}_{i+1} = \mathbf{r}_i - \mathbf{A}\mathbf{u}_i/\mathbf{v}_i^H \mathbf{A}\mathbf{u}_i, \quad \varphi_{i+1}\mathbf{q}_{i+1}^H = \mathbf{q}_i^H - \mathbf{v}_i^H \mathbf{A}/\mathbf{v}_i^H \mathbf{A}\mathbf{u}_i. \tag{2.7}$$

Feltéve, hogy $\mathbf{r}_{i+1}$ és $\mathbf{q}_{i+1}$ nemzérus vektorok, a λ_{i+1} és φ_{i+1} számok értékét szabadon rögzíthetjük: például 1-nek, vagy úgy, hogy $\|\mathbf{r}_{i+1}\|_C = 1$, $\|\mathbf{q}_{i+1}\|_K = 1$ legyen.

Ezután a $\mathbf{v}_{i+1}$ és $\mathbf{u}_{i+1}$ vektorokat úgy készítjük, hogy (2.4) szerint képezzük az $\mathbf{L}_{i+1}$ és $\mathbf{F}_{i+1}$ mátrixokat, s az $i+1$ indexre kiírjuk (2.1) első egyenletének utolsó sorát, valamint (2.2) második egyenletének utolsó oszlopát:

$$\mathbf{v}_{i+1}^H = \lambda_{i+1}\mathbf{v}_i^H + \mathbf{r}_{i+1}^H \mathbf{C}/\|\mathbf{r}_{i+1}\|_C^2, \quad \mathbf{u}_{i+1} = \varphi_{i+1}\mathbf{u}_i + \mathbf{K}\mathbf{q}_{i+1}/\|\mathbf{q}_{i+1}\|_K^2. \tag{2.8}$$

Ha ilymódon $\mathbf{v}_{i+1}^H \mathbf{A}\mathbf{u}_{i+1} \neq 0$, akkor a vektorrendszereket sikerrel bővítettük.

2.3. *Diagonál-folytatás módszere*

Ekkor az **L** és **F** mátrix közül az egyiket diagonálisan, a másikat bidiagonálisan folytatjuk az $i+1$-edik lépésben. A módszer alkalmazásának feltétele, hogy az $\mathbf{r}_{i+1}$ és $\mathbf{q}_{i+1}$ vektorok egyike bidiagonális építéssel zérusnak adódjék. Látni fogjuk, hogy egyszerre csak az egyik mátrix folytatható diagonálisan. Tegyük fel, hogy (2.7)-ben $\mathbf{q}_{i+1}\neq\mathbf{0}$, de az első egyenletre $\lambda_{i+1}\mathbf{r}_{i+1}=\mathbf{r}_i-\mathbf{A}\mathbf{u}_i/\mathbf{v}_i^H\mathbf{A}\mathbf{u}_i=\mathbf{0}$ adódik. Válasszuk ekkor a λ_{i+1} paramétert zérusnak, mert ekkor nem szükséges, hogy az $\mathbf{r}_{i+1}$ vektor zérus értéket vegyen fel. Az építővektorok tehát

$$\mathbf{l}_{i+1}^T=\mathbf{0},\quad \mathbf{f}_{i+1}=-\varphi_{i+1}\mathbf{e}_i, \tag{2.9}$$

továbbá tegyük fel, hogy

$$\mathbf{l}_{i+2}^T=\mathbf{0}. \tag{2.10}$$

Ez azt jelenti, hogy az **L** mátrixot diagonálmátrixként építjük — ez a *bal oldali diagonál-folytatás*. Helyettesítsük (2.1) egyenleteibe az $\mathbf{L}_{i+1}$ mátrixot és a zérus $\mathbf{l}_{i+2}$ vektort, ezzel a (2.1) egyenletek $i+1$-edik oszlopa, ill. sorából a következő két egyenletet nyerjük:

$$\mathbf{r}_{i+1}=\mathbf{A}\mathbf{u}_{i+1}/\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_{i+1},\quad \mathbf{v}_{i+1}^H=\mathbf{r}_{i+1}^H\mathbf{C}/\|\mathbf{r}_{i+1}\|_C^2. \tag{2.11}$$

Ennek kétféle megoldását is felírhatjuk a szabad skálázás folytán:

$$\begin{aligned}&\text{a)}\ \mathbf{r}_{i+1}=\mathbf{A}\mathbf{u}_{i+1}/\|\mathbf{A}\mathbf{u}_{i+1}\|_C,\quad \mathbf{v}_{i+1}^H=\mathbf{r}_{i+1}^H\mathbf{C},\\ &\text{b)}\ \mathbf{r}_{i+1}=\mathbf{A}\mathbf{u}_{i+1},\quad \mathbf{v}_{i+1}^H=\mathbf{r}_{i+1}^H\mathbf{C}/\|\mathbf{r}_{i+1}\|_C^2,\end{aligned} \tag{2.12}$$

amelynek helyettesítéssel ellenőrizhetők. Az elsőnél $\|\mathbf{r}_{i+1}\|_C=1$, a másodiknál $\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_{i+1}=1$. Ha $\mathbf{A}\mathbf{u}_{i+1}=\mathbf{0}$, akkor $\mathbf{v}_{i+1}=\mathbf{0}$, amennyiben úgy tekintjük, hogy $\mathbf{v}_{i+1}$ a $\mathbf{C}\mathbf{A}\mathbf{u}_{i+1}$ vektorral arányos, s ekkor a rekurzió befejeződik.

Hasonló módon a *jobb oldali diagonál-folytatás* formulái:

$$\mathbf{l}_{i+1}^T=-\lambda_{i+1}\mathbf{e}_i^T,\quad \mathbf{f}_{i+1}=\mathbf{0}, \tag{2.13}$$

$$\begin{aligned}&\text{a)}\ \mathbf{q}_{i+1}^H=\mathbf{v}_{i+1}^H\mathbf{A}/\|\mathbf{v}_{i+1}^H\mathbf{A}\|_K,\quad \mathbf{u}_{i+1}=\mathbf{K}\mathbf{q}_{i+1},\\ &\text{b)}\ \mathbf{q}_{i+1}^H=\mathbf{v}_{i+1}^H\mathbf{A},\quad \mathbf{u}_{i+1}=\mathbf{K}\mathbf{q}_{i+1}/\|\mathbf{q}_{i+1}\|_K^2,\end{aligned} \tag{2.14}$$

ahol $\mathbf{v}_{i+1}^H\mathbf{A}=\mathbf{0}$ lesz a befejeződés feltétele. Az $\mathbf{f}_{i+2}=\mathbf{0}$ feltevést nem is írtuk ki, mert a (2.14) vektorokkal (2.7) második egyenlete $\varphi_{i+2}=0$-t eredményez, tehát $\mathbf{f}_{i+2}=\mathbf{0}$ következmény lesz.

2.4. *Oldallépés módszere: bidiagonális építés és csere az ortogonális vektoroknál az egyik oldalon*

E módszer bevezetését az tette indokolttá, hogy a bidiagonális építéssel készített $i+1$-edik vektorokkal a $\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_{i+1}$ belső szorzat közel zérusnak, vagy éppen zérusnak adódhat, ami a numerikus hiba erőteljes növekedését eredményezi, vagy meghiúsítja a rekurzió folytatását. Itt olyan változtatással próbálkozunk, amely a meglevő vektorokból újakat készít, ezzel lehetővé téve egy esetleg előnyösebb folytatást.

Most is létezik bal és jobb oldali verzió. A módszer lényege, hogy az egyik oldalon a bidiagonális építés után az i és $i+1$-edik ortogonális vektorokat felcseréljük, és a másik oldalon a bidiagonális építést csak ezután tesszük meg. Az ortogonális vektorok cseréje maga után vonja a velük egy oldalon levő i és $i+1$-edik konjugált vektorok megváltozását is. Az egyértelműség érdekében a megváltozó mennyiségeket felül hullámmal fogjuk jelölni.

Kidolgozzuk a bal oldali oldallépést. Írjuk (2.1) egyenleteit a következő particionált mátrixos alakba:

$$\begin{bmatrix} \mathbf{L}_{i-1} & 0 & 0 \\ \tilde{\mathbf{l}}_i^T & 1 & 0 \\ 0 & -\tilde{\lambda}_{i+1} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{V}_{i-1}^H \\ \tilde{\mathbf{v}}_i^H \\ \tilde{\mathbf{v}}_{i+1}^H \end{bmatrix} = \begin{bmatrix} \mathbf{D}_{r,i-1}^{-1} \mathbf{R}_{i-1}^H \mathbf{C} \\ \tilde{\mathbf{r}}_i^H \mathbf{C} / \|\tilde{\mathbf{r}}_i\|_C^2 \\ \tilde{\mathbf{r}}_{i+1}^H \mathbf{C} / \|\tilde{\mathbf{r}}_{i+1}\|_C^2 \end{bmatrix}, \tag{2.15}$$

$$\mathbf{A}\mathbf{U}_i \tilde{\mathbf{D}}_a^{-1} = [\mathbf{R}_{i-1} \tilde{\mathbf{r}}_i \tilde{\mathbf{r}}_{i+1}] \begin{bmatrix} \mathbf{L}_{i-1} & 0 \\ \tilde{\mathbf{l}}_i^T & 1 \\ 0 & -\tilde{\lambda}_{i+1} \end{bmatrix},$$

ahol $i+1$ elkészített vektor szerepel, és a mátrixoknál az alsó index arra utal, hogy hány vektorból készültek. Vezessük be a következő mátrixokat:

$$\mathbf{P}_{i+1} = \begin{bmatrix} \mathbf{I}_{i-1} & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{G}_{i+1} = \begin{bmatrix} \mathbf{I}_{i-1} & 0 & 0 \\ 0 & -\tilde{\lambda}_{i+1}^{-1} & 1 \\ 0 & 0 & \tilde{\lambda}_{i+1} \end{bmatrix}, \quad \mathbf{Z}_i = \begin{bmatrix} \mathbf{I}_{i-1} & 0 \\ 0 & -\tilde{\lambda}_{i+1}^{-1} \end{bmatrix}.$$

Ezután (2.15) első egyenletét szorozzuk balról a $\mathbf{P}_{i+1}$ mátrixszal, továbbá az $\mathbf{L}$- és $\mathbf{V}$-mátrix közé írjuk be a $\mathbf{G}_{i+1}\mathbf{G}_{i+1}^{-1}$ szorzatot és szorozzunk az egyikkel balra, a másikkal jobbra. Az eredmény:

$$\begin{bmatrix} \mathbf{L}_{i-1} & 0 & 0 \\ 0 & 1 & 0 \\ \tilde{\mathbf{l}}_i & -\tilde{\lambda}_{i+1}^{-1} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{V}_{i-1}^H \\ -\tilde{\lambda}_{i+1} \tilde{\mathbf{v}}_i^H + \tilde{\mathbf{v}}_{i+1}^H \\ \tilde{\lambda}_{i+1}^{-1} \tilde{\mathbf{v}}_{i+1}^H \end{bmatrix} = \begin{bmatrix} \mathbf{D}_{r,i-1}^{-1} \mathbf{R}_{i-1}^H \\ \tilde{\mathbf{r}}_{i+1}^H / \|\tilde{\mathbf{r}}_{i+1}\|_C^2 \\ \tilde{\mathbf{r}}_i / \|\tilde{\mathbf{r}}_i\|_C^2 \end{bmatrix} \mathbf{C}.$$

Hasonló módon (2.15) második egyenletében az $\mathbf{R}$- és $\mathbf{L}$-mátrixok közé az $\mathbf{I} = \mathbf{P}_{i+1}\mathbf{P}_{i+1}^{-1}$ szorzatot visszük be, és mindkét oldalt a $\mathbf{Z}_i$ mátrixszal szorozzuk jobbról:

$$\mathbf{A}\mathbf{u}_i \tilde{\mathbf{D}}_a^{-1} \mathbf{Z}_i = [\mathbf{R}_{i-1} \tilde{\mathbf{r}}_{i+1} \tilde{\mathbf{r}}_i] \begin{bmatrix} \mathbf{L}_{i-1} & 0 \\ 0 & 1 \\ \tilde{\mathbf{l}}_i & -\tilde{\lambda}_{i+1}^{-1} \end{bmatrix}.$$

A kapott alakokról leolvasható, hogy beleillenek a (2.1)—(2.4) általános sémába. Így az átrendezés eredményéül kapjuk:

$$\mathbf{l}_i^T = \mathbf{0}, \quad \lambda_{i+1} = \tilde{\lambda}_{i+1}^{-1}, \quad \mathbf{l}_{i+1}^T = [\tilde{\mathbf{l}}_i^T, -\lambda_{i+1}], \tag{2.16}$$

$$\mathbf{r}_i = \tilde{\mathbf{r}}_{i+1}, \quad \mathbf{r}_{i+1} = \tilde{\mathbf{r}}_i, \quad \mathbf{v}_{i+1}^H = \tilde{\lambda}_{i+1}^{-1} \tilde{\mathbf{v}}_{i+1}^H, \tag{2.17}$$

$$\mathbf{v}_i^H = \tilde{\mathbf{r}}_{i+1}^H / \|\tilde{\mathbf{r}}_{i+1}\|_C^2 = -\tilde{\lambda}_{i+1} \tilde{\mathbf{v}}_i^H + \tilde{\mathbf{v}}_{i+1}^H. \tag{2.18}$$

A $\mathbf{v}_i^H \mathbf{A}\mathbf{u}_i$ szorzatot nem kell újraszámolni, mert az *A-konjugáltság* miatt fennáll

$$\mathbf{v}_i^H \mathbf{A}\mathbf{u}_i = -\tilde{\lambda}_{i+1}\tilde{\mathbf{v}}_i^H \mathbf{A}\mathbf{u}_i + \tilde{\mathbf{v}}_{i+1}^H \mathbf{A}\mathbf{u}_i = -\tilde{\lambda}_{i+1}\tilde{\mathbf{v}}_i^H \mathbf{A}\mathbf{u}_i. \tag{2.19}$$

A jobb oldali $i+1$-edik hullám nélküli vektorokat ezután készítjük bidiagonális építéssel.

A jobb oldali oldallépés átrendezései a hullámozási konvenció betartásával:

$$\mathbf{f}_i = 0, \quad \varphi_{i+1} = \tilde{\varphi}_{i+1}^{-1}, \quad \mathbf{f}_{i+1}^T = [\tilde{\mathbf{f}}_i^T, -\varphi_{i+1}], \tag{2.20}$$

$$\mathbf{q}_i = \tilde{\mathbf{q}}_{i+1}, \quad \mathbf{q}_{i+1} = \tilde{\mathbf{q}}_i, \quad \mathbf{u}_{i+1} = \tilde{\varphi}_{i+1}^{-1}\tilde{\mathbf{u}}_{i+1}, \tag{2.21}$$

$$\mathbf{u}_i = \mathbf{K}\tilde{\mathbf{q}}_{i+1}/\|\tilde{\mathbf{q}}_{i+1}\|_K^2 = -\tilde{\varphi}_{i+1}\tilde{\mathbf{u}}_i + \tilde{\mathbf{u}}_{i+1}. \tag{2.22}$$

$$\mathbf{v}_i^H \mathbf{A}\mathbf{u}_i = -\tilde{\varphi}_{i+1}\mathbf{v}_i^H \mathbf{A}\tilde{\mathbf{u}}_i, \tag{2.23}$$

s ezután készülnek a bal oldali $i+1$-edik vektorok bidiagonális építéssel.

Mint látjuk, az i és $i+1$-edik ortogonális vektorok felcserélődnek, az i-edik konjugált vektor pedig olyan értéket kap, ami az i-edik lépésben tett féloldali újraindításnak felel meg, l. (2.5). Az oldallépés után az $\mathbf{l}_{i+1}$ vagy $\mathbf{f}_{i+1}$ építővektor már több nemzérus elemmel rendelkezhet, mert az oldallépés ugyanazon az oldalon többször egymás után eggyel növekvő indexek mellett megtehető. Ha például a k-adik szinttől kezdve minden lépésben oldallépéssel generáljuk a bal oldali vektorokat az i-edik szintig, akkor az $\mathbf{L}_i$ mátrix „fűrész"-alakú lesz:

$$\mathbf{L}_i = \begin{bmatrix} \cdot & & & & & & \\ & \cdot & & & & & \\ & & \cdot & & & & \\ -\lambda_{k-1} & & 1 & & & & \\ & & 0 & 1 & & & \\ & & \cdot & \cdot & 1 & & \\ & & \cdot & \cdot & & \cdot & \\ & & \cdot & \cdot & & & \cdot \\ & & -\lambda_k & -\lambda_{k+1} \ldots & -\lambda_i & & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ (k-1 \\ \\ \\ \\ \\ (i \end{matrix}$$

tehát a $-\lambda_k, -\lambda_{k+1}, \ldots$ elemek az utolsó sorba tolódtak.

Ezzel befejeztük az építési módszerek ismertetését.

2.5. A direkt rekurzió működésének alaptétele

Bevezetjük a következő konvenciót: A bidiagonális építéssel nyert vektorokat felül hullámmal jelöljük és őket próbavektoroknak tekintjük. Hullám nélkül az elfogadott vektorokat fogjuk jelölni. Elfogadható lehet bármely építéssel készített vektor, amely sikerrel bővít, tehát olyan $\mathbf{r}_{i+1}$ vektor, amelyre $\|\mathbf{r}_{i+1}\|_C \neq 0$, $\mathbf{q}_{i+1}$ vektor, amelyre $\|\mathbf{q}_{i+1}\|_K \neq 0$, és $\mathbf{v}_{i+1}$, $\mathbf{u}_{i+1}$ pár, amelyre $\mathbf{v}_{i+1}^H \mathbf{A}\mathbf{u}_{i+1} \neq 0$. Ha $\tilde{\mathbf{r}}_{i+1} = \mathbf{0}$, akkor úgy vesszük, hogy $\tilde{\mathbf{v}}_{i+1} = \mathbf{0}$. Hasonlóan $\tilde{\mathbf{q}}_{i+1} = \mathbf{0}$-ból $\tilde{\mathbf{u}}_{i+1} = \mathbf{0}$ következik. Most már megfogalmazhatjuk a direkt rekurzió működésére vonatkozó alaptételt.

2.1. TÉTEL. Amennyiben a $\tilde{\mathbf{v}}_{i+1}^H \mathbf{A}$ vagy a $\mathbf{A}\,\tilde{\mathbf{u}}_{i+1}$vektoroknak legalább egyike nemzérus, a (2.1)—(2.4) direkt rekurzió építési módszereinek valamelyikével az $\{\mathbf{r}_j\}_{j=1}^i$, $\{\mathbf{q}_j\}_{j=1}^i$ és $\{\mathbf{v}_j, \mathbf{u}_j\}_{j=1}^i$ vektor-rendszerek sikerrel bővíthetők. Mégpedig:

1. A bidiagonális építés sikeres, ha $\tilde{\mathbf{v}}_{i+1}^H \mathbf{A}\tilde{\mathbf{u}}_{i+1} \neq \mathbf{0}$;
2. A diagonál-folytatás sikeres a
   bal oldalon, ha $\tilde{\mathbf{r}}_{i+1} = \mathbf{0}$ és $\mathbf{A}\tilde{\mathbf{u}}_{i+1} \neq \mathbf{0}$,
   jobb oldalon, ha $\tilde{\mathbf{q}}_{i+1} = \mathbf{0}$ és $\tilde{\mathbf{v}}_{i+1}^H \mathbf{A} \neq \mathbf{0}$.
3. Ha $\tilde{\mathbf{v}}_{i+1}^H \mathbf{A} \neq \mathbf{0}$, akkor a bidiagonális építés vagy a bal oldallépés legalább egyike sikeres. Ha $\mathbf{A}\tilde{\mathbf{u}}_{i+1} \neq \mathbf{0}$, akkor a bidiagonális építés vagy a jobb oldallépés legalább egyike sikeres.

2.6. Az alaptétel igazolása

Több önmagában is érdekes tétel bizonyítása vezet majd el az alaptétel igazolásához.

2.2. TÉTEL. Az $\mathbf{r}_{i+1} \neq 0$ vektor *C-ortogonális* az $\mathbf{r}_j$, $j=1, 2, \ldots, i$ vektorokra, és a $\mathbf{q}_{i+1} \neq \mathbf{0}$ vektor *K-ortogonális* a $\mathbf{q}_j$, $j=1, 2, \ldots, i$ vektorokra.

Bizonyítás. Tegyük fel először, hogy $\mathbf{l}_{i+1} \neq \mathbf{0}$, $\mathbf{f}_{i+1} \neq \mathbf{0}$. Szorozzuk össze (2.1) két egyenletét, bal oldalt a bal oldallal, jobb oldalt a jobbal:

$$\mathbf{L}_i \mathbf{V}^H \mathbf{A}\mathbf{U}\mathbf{D}_a^{-1} = \mathbf{D}_r^{-1}\mathbf{R}^H \mathbf{C}\mathbf{R}\mathbf{L}_i + \mathbf{D}_r^{-1}\mathbf{R}^H \mathbf{C}\mathbf{r}_{i+1}\mathbf{l}_{i+1}^T,$$

s ez a $\mathbf{D}_a$, $\mathbf{D}_r$ mátrixok (2.3) előállítása miatt a

$$\mathbf{0} = \mathbf{R}^H \mathbf{C}\mathbf{r}_{i+1}\mathbf{l}_{i+1}^T$$

alakra egyszerűsödik. Mivel $\mathbf{l}_{i+1} \neq 0$, a kapott alakból kiolvasható, hogy az $\mathbf{r}_{i+1}$ vektor *C-ortogonális* az $\mathbf{r}_j$, $j=1, 2, \ldots, i$ vektorokra. A $\mathbf{q}_{i+1}$ vektorra vonatkozó állítás hasonlóan igazolható (2.2) egyenleteinek az összeszorzásával.

Vegyük észre, hogy az $\mathbf{l}_{i+1} \neq \mathbf{0}$, $\mathbf{f}_{i+1} \neq \mathbf{0}$ feltétel akkor teljesül, ha a direkt rekurzió $i+1$-edik lépését a bidiagonális építés, vagy az oldallépés módszerével végeztük. Így a tételt bizonyítani kell még arra az esetre, amikor a rekurziós lépés a diagonál-folytatás módszerével történik, azaz $\mathbf{l}_{i+1} = \mathbf{0}$ vagy $\mathbf{f}_{i+1} = \mathbf{0}$. Ennek az esetnek az igazolását majd a következő tétel bizonyítási eljárásának negyedik lépése adja meg.

2.3. TÉTEL. Az $i+1$-edik vektorok a következő projektor-összefüggésekkel állíthatók elő:

$$\mathbf{r}_{i+1} = \begin{cases} \lambda_{i+1}^{-1} \mathbf{P}_i^l \mathbf{r}_i, & \text{ha} \quad \mathbf{l}_{i+1} \neq \mathbf{0}, \\ (\mathbf{v}_{i+1}^H \mathbf{A}\mathbf{u}_{i+1})^{-1} \|\mathbf{q}_{i+1}\|_K^{-2} \mathbf{P}_i^l \mathbf{A}\mathbf{K}\mathbf{q}_{i+1}, & \text{ha} \quad \mathbf{l}_{i+1} = \mathbf{0}, \end{cases} \tag{2.24}$$

$$\mathbf{q}_{i+1}^H = \begin{cases} \varphi_{i+1}^{-1} \mathbf{q}_i^H \mathbf{P}_i^r, & \text{ha} \quad \mathbf{f}_{i+1} \neq \mathbf{0}, \\ (\mathbf{v}_{i+1}^H \mathbf{A}\mathbf{u}_{i+1})^{-1} \|\mathbf{r}_{i+1}\|_C^{-2} \mathbf{r}_{i+1}^H \mathbf{C}\mathbf{A}\mathbf{P}_i^r, & \text{ha} \quad \mathbf{f}_{i+1} = 0, \end{cases} \tag{2.25}$$

$$\mathbf{v}_{i+1}^H = \|\mathbf{r}_{i+1}\|_C^{-2} \mathbf{r}_{i+1}^H \mathbf{C}\mathbf{P}_i^l, \tag{2.26}$$

$$\mathbf{u}_{i+1} = \|\mathbf{q}_{i+1}\|_K^{-2} \mathbf{P}_i^r \mathbf{K}\mathbf{q}_{i+1}, \tag{2.27}$$

ahol $\mathbf{P}_i^l$ és $\mathbf{P}_i^r$ az (1.2)—(1.3)-ban bevezetett projektorok.

Bizonyítás. A bizonyítást hat lépésben tesszük meg.

1. Néhány szükséges összefüggés származtatásával kezdjük. A $\mathbf{P}_i^l$ projektor mátrixos alakban is felírható:

$$\mathbf{P}_i^l = \mathbf{I}_m - \mathbf{A}\mathbf{U}\mathbf{D}_a^{-1}\mathbf{V}^H, \tag{2.28}$$

ahol az i index jelzi, hogy a $\mathbf{V}$, $\mathbf{U}$ mátrixban i oszlopvektor található. Helyettesítsük ide (2.1)-ből $\mathbf{A}\mathbf{U}\mathbf{D}_a^{-1}$ kifejezését, majd az $\mathbf{L}_i\mathbf{V}^H$ mátrix előállítását (2.1) első egyenletéből:

$$\mathbf{P}_i^l = \mathbf{I}_m - \mathbf{R}\mathbf{L}_i\mathbf{V}^H - \mathbf{r}_{i+1}\mathbf{l}_{i+1}^T\mathbf{V}^H = \mathbf{I}_m - \mathbf{R}\mathbf{D}_r^{-1}\mathbf{R}^H\mathbf{C} - \mathbf{r}_{i+1}\mathbf{l}_{i+1}^T\mathbf{V}^H. \tag{2.29}$$

Hasonlóan adódik (1.3) és (2.2) összefüggéseinek felhasználásával

$$\mathbf{P}_i^r = \mathbf{I}_n - \mathbf{U}\mathbf{D}_a^{-1}\mathbf{V}^H\mathbf{A} = \mathbf{I}_n - \mathbf{K}\mathbf{Q}\mathbf{D}_q^{-1}\mathbf{Q}^H - \mathbf{U}\mathbf{f}_{i+1}\mathbf{q}_{i+1}^H. \tag{2.30}$$

Egy másik összefüggést kapunk, ha (2.1) első egyenletét jobbról az $\mathbf{R}$ mátrixszal szorozzuk. Figyelembe véve (2.3)-at kapjuk, hogy $\mathbf{L}_i$ és $\mathbf{V}^H\mathbf{R}$ egymás inverzei:

$$\mathbf{L}_i\mathbf{V}^H\mathbf{R} = \mathbf{I}_i = \mathbf{V}^H\mathbf{R}\mathbf{L}_i. \tag{2.31}$$

2. Feltesszük, hogy $\mathbf{l}_{i+1} \neq \mathbf{0}$. Ebből $\lambda_{i+1} \neq 0$ következik, tudjuk, ilyen a bidiagonális építés, ill. az oldallépés. Szorozzuk (2.1) második egyenletét jobbról az $\mathbf{e}_i$ vektorral, és használjuk fel (2.31)-et:

$$\lambda_{i+1}\mathbf{r}_{i+1} = (\mathbf{R}\mathbf{L}_i - \mathbf{A}\mathbf{U}\mathbf{D}_a^{-1})\mathbf{e}_i = (\mathbf{I}_m - \mathbf{A}\mathbf{U}\mathbf{D}_a^{-1}\mathbf{V}^H)\mathbf{R}\mathbf{L}_i\mathbf{e}_i = \mathbf{P}_i^l\mathbf{r}_i.$$

Az eredmény (2.24) első összefüggése. Hasonlóan igazolható az $\mathbf{f}_{i+1} \neq \mathbf{0}$ feltevés mellett a (2.25) első összefüggése.

3. Most is tegyük fel, hogy $\mathbf{l}_{i+1} \neq \mathbf{0}$. Szorozzuk (2.29)-et balról az $\mathbf{r}_{i+1}^H\mathbf{C}$ vektorral:

$$\mathbf{r}_{i+1}^H\mathbf{C}\mathbf{P}_i^l = \mathbf{r}_{i+1}^H\mathbf{C} - \|\mathbf{r}_{i+1}\|_C^2\mathbf{l}_{i+1}^T\mathbf{V}_i^H. \tag{2.32}$$

Itt felhasználtuk a 2.2 tételt, amelyet $\mathbf{l}_{i+1} \neq \mathbf{0}$ esetben bizonyítottunk. Ezután írjuk ki $\mathbf{v}_{i+1}^H$ általános kifejezését (2.1) első egyenletéből. E célból helyettesítsük ide $\mathbf{L}_{i+1}$ (2.4)-beli alakját, írjuk ki az utolsó sort és ismét használjuk fel, hogy a 2.2 tétel szerint az $i+1$-edrendű $\mathbf{D}_r$ mátrix diagonálmátrix:

$$\mathbf{l}_{i+1}^T\mathbf{V}_i^H + \mathbf{v}_{i+1}^H = \|\mathbf{r}_{i+1}\|_C^{-2}\mathbf{r}_{i+1}^H\mathbf{C}. \tag{2.33}$$

Összehasonlítva (2.32)-vel kapjuk (2.26)-ot. Hasonló módon igazolható (2.27) az $\mathbf{f}_{i+1} \neq \mathbf{0}$ esetére.

4. Befejezzük a 2.2 tétel bizonyítását azzal, hogy az $\mathbf{l}_{i+1} = \mathbf{0}$ esetre is igazoljuk az állítást. A bal oldali diagonál-folytatás mellett $\mathbf{f}_{i+1} \neq \mathbf{0}$, így (2.27) fennáll, ezt az előző lépésben láttuk be. A $\mathbf{P}_i^r$ projektor (2.30)-beli első kifejezéséből közvetlenül ellenőrizhető a

$$\mathbf{V}_i^H\mathbf{A}\mathbf{P}_i^r \equiv 0 \tag{2.24}$$

azonosság. Ebből (2.27) folytán a $\mathbf{V}_i^H\mathbf{A}\mathbf{u}_{i+1} = \mathbf{0}$ reláció következik. De (2.11) alapján $\mathbf{r}_{i+1}$ iránya $\mathbf{A}\mathbf{u}_{i+1}$ irányával egyezik, emiatt $\mathbf{V}_i^H\mathbf{r}_{i+1} = \mathbf{0}$ is érvényes. Végül helyettesítsük $\mathbf{V}_i^H$ kifejezését (2.1) első egyenletéből, hagyjuk el a nemszinguláris $\mathbf{L}_i$, $\mathbf{D}_r$ mátrixokat, s így az

$$\mathbf{R}^H\mathbf{C}\mathbf{r}_{i+1} = 0, \quad \mathbf{R}^H \in \mathbf{C}^{i,m} \tag{2.35}$$

összefüggésre jutunk, amellyel a kívánt célt elértük.

A jobb oldali diagonál-folytatás esetén a $\mathbf{q}_{i+1}$ vektor *K-ortogonalitását* hasonlóan bizonyíthatjuk, így a 2.2 tételt teljesen igazoltuk.

5. Minthogy a 2.2 tétel most már általánosan érvényes, megismételjük a 3. lépést, amivel (2.26) és (2.27) minden esetben igaz lesz.

6. (2.24) második formulája tulajdonképpen (2.11) első egyenletének bonyolultabb formában felírt alakja. Úgy keletkezett, hogy $\mathbf{u}_{i+1}$ (2.27)-ben található kifejezését (2.11) első egyenletébe helyettesítettük, és felhasználtuk az

$$\mathbf{A}\mathbf{P}_i^r \equiv \mathbf{P}_i^l\mathbf{A} \tag{2.36}$$

azonosságot. Az átalakítás nem öncélú, mert mutatja, hogy az $i+1$-edik vektorok mindenkor a $\mathbf{P}_i^l$ és $\mathbf{P}_i^r$ projektorokkal készülnek. Hasonló módon történik (2.25) második formulájának igazolása, s ezzel a 2.3 tétel bizonyítását befejeztük.

2.4. TÉTEL. Amennyiben $\mathbf{v}_{i+1}^H$ és $\mathbf{u}_{i+1}$ konvencióink szerint elfogadott vektorok, azaz $\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_{i+1}\neq 0$, akkor a $\mathbf{v}_{i+1}$, $\mathbf{u}_{i+1}$ vektorpár a $\{\mathbf{v}_j, \mathbf{u}_j\}_{j=1}^{i}$ *A-konjugált* rendszert eggyel bővíti, tehát $\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_j=0$, $\mathbf{v}_j^H\mathbf{A}\mathbf{u}_{i+1}=0$, $j\leqq i$.

Bizonyítás. Az 1.1 tétel alapján elegendő a (2.26) és (2.27) vetítéses előállításokat ellenőrizni.

Ezzel eljutottunk odáig, hogy igazoltuk a direkt rekurzióval előállított vektorok ortogonalitási tulajdonságait.

Az oldallépésre vonatkozóan teszünk néhány megjegyzést. Tegyük fel, az oldallépést az i és $i+1$-edik szint között a bal oldalon hajtjuk végre, amikor $\mathbf{f}_i=\mathbf{0}$, azaz jobb oldalon a diagonál-folytatással építünk. Ekkor az $\mathbf{u}_i$ és $\mathbf{q}_i$ vektorok ugyanúgy megtartják értéküket, mint egyébként, azonban a $\mathbf{v}_i$ vektor módosulása miatt a jobb oldali bidiagonális építés általában nem fog zérust adni φ_{i+1} és ezzel az $\mathbf{f}_{i+1}$ vektor értékére: Az $\mathbf{F}$ mátrix ekkor az $i+1$-edik oszloppal „kinyílik" bidiagonális szerkezetű mátrixba.

Egyszerre csak az egyik oldalon tehetünk oldallépést, mert a másik oldal i-edik vektorait fixen kell tartani. Azt azonban megtehetjük, hogy az i és $i+1$-edik szint között végrehajtott bal oldali oldallépés után az $i+1$-edik bal oldali vektorokat elhagyjuk, jobb oldali oldallépést teszünk az $i, i+1$-edik szinten és ezután bidiagonális építéssel képezünk új $i+1$-edik bal oldali vektorokat.

Az építővektorok általános alakja a következő lesz:

$$\mathbf{l}_{i+1} = \sum_{j=k}^{i} -\lambda_{j+1}\mathbf{e}_j, \quad \mathbf{f}_{i+1} = \sum_{j=h}^{i} -\varphi_{j+1}\mathbf{e}_j, \tag{2.37}$$

ahol a szummák alsó határára a következő két eset valamelyikének kell teljesülnie:

$$\begin{aligned} &\text{a)}\ k \leqq i, \quad h = i, \\ &\text{b)}\ k = i, \quad h \leqq i. \end{aligned} \tag{2.38}$$

2.5. TÉTEL. A bal oldali oldallépéskor fennáll

(2.39)

$$\varphi_{i+1}\|\mathbf{q}_{i+1}\|_K^2\,\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_{i+1} = \tilde{\lambda}_{i+1}^{-2}(\mathbf{u}_i^H\mathbf{A}^H\tilde{\mathbf{v}}_i)^{-1}\|\mathbf{A}^H\tilde{\mathbf{v}}_{i+1}\|_K^2 + \tilde{\lambda}_{i+1}^{-1}\|\tilde{\mathbf{q}}_{i+1}\|_K^2\,\tilde{\varphi}_{i+1}\tilde{\mathbf{v}}_{i+1}^H\mathbf{A}\tilde{\mathbf{u}}_{i+1},$$

a jobb oldali oldallépéskor pedig
(2.40)

$$\tilde{\lambda}_{i+1}\|\mathbf{r}_{i+1}\|_C^2\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_{i+1} = \tilde{\varphi}_{i+1}^{-2}(\tilde{\mathbf{u}}_i^H\mathbf{A}^H\mathbf{v}_i)^{-1}\|\mathbf{A}\tilde{\mathbf{u}}_{i+1}\|_C^2 + \tilde{\varphi}_{i+1}^{-1}\|\tilde{\mathbf{r}}_{i+1}\|_C^2\tilde{\lambda}_{i+1}\tilde{\mathbf{v}}_{i+1}^H\mathbf{A}\tilde{\mathbf{u}}_{i+1}.$$

Bizonyítás. Csak a bal oldali oldallépés formuláját igazoljuk. Az *A-konjugált* tulajdonság és (2.8) második egyenlete miatt

$$\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_{i+1} = \mathbf{v}_{i+1}^H\mathbf{A}\mathbf{K}\mathbf{q}_{i+1}/\|\mathbf{q}_{i+1}\|_K^2. \tag{2.41}$$

Ide (2.7)-ből $\mathbf{q}_{i+1}$ helyettesíthető, ami $\mathbf{q}_i$ és $\mathbf{A}^H\mathbf{v}_i$ lineáris kombinációja. A $\mathbf{q}_i$ vektor (2.41)-ből a $\mathbf{K}$ mátrixszal szorzódik, és (2.8)-ból $\mathbf{K}\mathbf{q}_i$ az $\mathbf{u}_i$ és $\mathbf{u}_{i-1}$ vektorok lineáris kombinációja, így az *A-konjugáltság* miatt a $\mathbf{q}_i$ vektor is kiesik:

$$\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_{i+1} = -\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{K}\mathbf{A}^H\mathbf{v}_i\frac{1}{\|\mathbf{q}_{i+1}\|_K^2\mathbf{u}_i^H\mathbf{A}^H\tilde{\mathbf{v}}_i\varphi_{i+1}}.$$

A kapott alakba helyettesítsük $\mathbf{v}_i$ kifejezését (2.18)-ból és $\mathbf{v}_{i+1}$-et (2.17)-ből:

$$\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_{i+1} = -\frac{\|\mathbf{A}^H\tilde{\mathbf{v}}_{i+1}\|_K^2}{\tilde{\lambda}_{i+1}\varphi_{i+1}\|\mathbf{q}_{i+1}\|_K^2\mathbf{u}_i^H\mathbf{A}^H\tilde{\mathbf{v}}_i} + \frac{\tilde{\mathbf{v}}_{i+1}^H\mathbf{A}\mathbf{K}\mathbf{A}^H\tilde{\mathbf{v}}_i}{\varphi_{i+1}\|\mathbf{q}_{i+1}\|_K^2\mathbf{u}_i^H\mathbf{A}^H\tilde{\mathbf{v}}_i}.$$

A második tagban szereplő $\mathbf{K}\mathbf{A}^H\tilde{\mathbf{v}}_i$ vektor az $\mathbf{u}_{i-1}$, $\mathbf{u}_i$ és $\tilde{\mathbf{u}}_{i+1}$ vektorok lineáris kombinációja, mivel $\mathbf{A}^H\tilde{\mathbf{v}}_i$ (2.7)-ből $\mathbf{q}_i$ és $\tilde{\mathbf{q}}_{i+1}$ segítségével kifejezhető, $\mathbf{K}\mathbf{q}_i$ és $\mathbf{K}\tilde{\mathbf{q}}_{i+1}$ pedig (2.8) alapján a fenti $\mathbf{u}$-vektorokkal állítható elő. A $\tilde{\mathbf{v}}_{i+1}^H\mathbf{A}$ vektorral képzett belső szorzatban csak $\tilde{\mathbf{u}}_{i+1}$ marad meg az *A-konjugáltság* miatt, így az együtthatóját a helyettesítésekből kihalászva kapjuk (2.39)-et (2.19) figyelembevételével.

A (2.39) összefüggésből kiolvasható, hogy $\|\mathbf{A}^H\tilde{\mathbf{v}}_{i+1}\|_K \neq 0$ mellett a $\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_{i+1}$ és $\tilde{\mathbf{v}}_{i+1}^H\mathbf{A}\tilde{\mathbf{u}}_{i+1}$ szorzatoknak csak az egyike lehet zérus. Ugyanezt mutatja (2.40) $\|\mathbf{A}\tilde{\mathbf{u}}_{i+1}\|_C \neq 0$ esetén, amiből a 2.1 tétel 3. állítása következik. A 2.1 tétel 1. és 2. állítása egyszerűen ellenőrizhető, így a 2.1 alaptétel bizonyítását is befejeztük.

Az $\mathbf{L}$ vagy $\mathbf{F}$ mátrixok inverzére szükségünk lehet a rekurzió elemzésekor. Az inverz könnyen elkészíthető, például az $\mathbf{L}$ mátrix esetén az

$$\begin{bmatrix}\mathbf{L}_j & 0\\ \mathbf{l}_{j+1}^T & 1\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{L}_j^{-1} & 0\\ -\mathbf{l}_{j+1}^T\mathbf{L}_j^{-1} & 1\end{bmatrix}$$

rekurzív építés segítségével. A direkt rekurzióhoz tartozó $\mathbf{L}$ mátrixnál nevezzük minden nemzérus $\mathbf{l}_j^T$ vektor első nemzérus elemét vezető elemnek, az utána következőket pedig belső elemnek. Ekkor bevezetve a

$$\nu_i = \begin{cases}\lambda_i, & \text{ha } \lambda_i \text{ vezető elem,}\\ 1, & \text{ha } \lambda_i \text{ belső elem,}\end{cases} \tag{2.43}$$

$$\gamma_i = \begin{cases}1, & \text{ha } \mathbf{l}_i \neq 0.\\ 0, & \text{ha } \mathbf{l}_i = 0,\end{cases} \tag{2.44}$$

$$\beta_i = \begin{cases}1. & \text{ha } \lambda_{i+1} \text{ vezető elem,}\\ \lambda_{i+1}, & \text{ha } \lambda_{i+1} \text{ belső elem}\end{cases} \tag{2.45}$$

jelöléseket, az inverz a következő formulával fejezhető ki:

$$(\mathbf{L}^{-1})_{ij} = \begin{cases} \gamma_i \beta_j \prod_{k=j+1}^{i} v_k, & i > j, \\ 1, & i = j, \\ 0, & i < j. \end{cases} \tag{2.46}$$

Az **F** mátrix inverze innen transzponálással nyerhető.

2.7. PÉLDA. Tekintsük az alábbi 3-rangú mátrixot:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & -1 \\ 2 & 0 & 1 \\ 3 & 1 & 1 \end{bmatrix},$$

és legyenek az induló vektorok $\mathbf{q}_1^T = \mathbf{u}_1^T = [1\ 0\ 0]$ és $\mathbf{r}_1^T = \mathbf{v}_1^T = [0\ 0\ 1]$. Ekkor a bidiagonális építés szerint készítsük el a további vektorokat, ahol a λ_j és φ_j skálázó paramétereket 1-nek választottuk:

$$\mathbf{r}_2 = \frac{1}{3}\begin{bmatrix} -1 \\ -2 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \frac{1}{5}\begin{bmatrix} -3 \\ -6 \\ 5 \end{bmatrix}, \quad \mathbf{r}_3 = \frac{1}{4}\begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_0 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix},$$

$$\mathbf{q}_2 = \frac{1}{3}\begin{bmatrix} 0 \\ -1 \\ -1 \end{bmatrix}, \quad \mathbf{u}_2 = \frac{1}{2}\begin{bmatrix} 2 \\ -3 \\ -3 \end{bmatrix}, \quad \mathbf{q}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{u}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

A harmadik **q**-vektor **0**-nak adódott. Ha ekkor a diagonál-folytatással járunk el (2.14b) szerint, akkor a $\mathbf{q}_3^T = \mathbf{v}_3^T \mathbf{A} = [0\ 2\ -2]$, $\mathbf{u}_3^T = [0\ 1/4\ -1/4]$ választással teljessé tudjuk tenni a rendszert. Az így kapott új $\mathbf{q}_3$ vektor ortogonális $\mathbf{q}_1$ és $\mathbf{q}_2$-re, valamint az így kapott $\mathbf{A}\mathbf{u}_3$ ortogonális $\mathbf{v}_1$ és $\mathbf{v}_2$-re. Ezek a vektorok mellett az **L** és **F** mátrixok alakja:

$$\mathbf{L} = \begin{bmatrix} 1 & & \\ -1 & 1 & \\ & -1 & 1 \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} 1 & -1 & \\ & 1 & \\ & & 1 \end{bmatrix}. \tag{2.47}$$

A zérus elkerülésére másik lehetőség a bal oldali oldallépés a második és a harmadik szint között. Ekkor $\mathbf{r}_2$ és $\mathbf{r}_3$ felcserélődik, $\mathbf{v}_3$ és $\mathbf{v}_2$ felveszi a (2.17)—(2.18) szerinti értéket:

$$\mathbf{r}_2 = \frac{1}{4}\begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \frac{4}{5}\begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}, \quad \mathbf{r}_3 = \frac{1}{3}\begin{bmatrix} -1 \\ -2 \\ 0 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}.$$

Igy (2.7) és (2.8) a $\mathbf{q}_3$ és $\mathbf{u}_3$ vektorra a következő értékeket adja:

$$\mathbf{q}_3 = \frac{1}{3}\begin{bmatrix} 0 \\ -1 \\ -1 \end{bmatrix} - \frac{2}{3}\begin{bmatrix} 0 \\ 2 \\ -3 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 0 \\ -5 \\ 5 \end{bmatrix}, \quad \mathbf{u}_3 = \frac{1}{2}\begin{bmatrix} 2 \\ -3 \\ -3 \end{bmatrix} + \frac{3}{10}\begin{bmatrix} 0 \\ -5 \\ 5 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} 5 \\ -9 \\ -6 \end{bmatrix}.$$

Most más vektorokat kaptunk, de ellenőrizhető, hogy az így kapott rendszer teljes. Ez esetben az **L** és **F** mátrixok:

$$\mathbf{L} = \begin{bmatrix} 1 & & \\ & 1 & \\ -1 & -1 & 1 \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} 1 & -1 & \\ & 1 & -1 \\ & & 1 \end{bmatrix}. \tag{2.48}$$

IRODALOM

[1] CONCUS, P., GOLUB, G. H., O'LEARY, D. P., "A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations", in *Sparse Matrix Computations Eds. J. R. BUNCH, D. J. ROSE* (Academic Press, New York, 1975) 309—332.

[2] DANIEL, J. W., "The conjugate gradient method for linear and nonlinear operator equations", *SIAM J. Numer. Anal.*, **4** (1967) 10—26

[3] FLETCHER, R., "Conjugate gradient methods for indefinite systems", in: *Proc. of the Dundee Biennial Conference on Numerical Analysis, Ed. G. A. Watson* (Lecture Notes in Mathematics 506, Springer, Berlin, 1976) 73—89.

[4] GUSTAFSSON, I., "On modified incomplete factorization methods", in: *Numerical Integration, of Differential Equations and Large Linear Systems Ed. J. Hinze* (Springer, Berlin, 1982) 334—351.

[5] HEGEDŰS, Cs. J., "Generalization of the method of conjugate gradients: The method of conjugate pairs", in: *Collection of Scientific Papers in Collaboration with the Joint Institute for Nuclear Research Institute for Physics,* Budapest, Hungary. Algorithms and Programs for Solution of Some Problems in Physics (Report KFKI 1979—82, MTA KFKI, Budapest, 1979) 199—209.

[6] HEGEDŰS, Cs. J., BODÓCS, L., "General recursions for A-conjugate vector pairs", Report KFKI 1982—56, MTA KFKI, Budapest, 1982.

[7] HEGEDŰS, Cs. J., BODÓCS, L., „Konjugált irányok előállítása — A konjugált párok módszere", *Alkalmaz. Mat. Lapok,* **11** (1985) 297—309.

[8] HESTENES, M. R., STIEFEL, E., "Methods of conjugate gradients for solving linear systems", *J. Res. Nat. Bur. Standards Sect.* B49 (1952) 409—436.

[9] HESTENES, M. R., *Conjugate Direction Methods in Optimization* (Springer, New York, 1980).

[10] KAMMERER, W. J., NASHED, M. Z., "On the convergence of the conjugate methods for singular operator equations", *SIAM J. Numer. Anal.*, 9 (1972) 165—181.

[11] LÁNCZOS, C., "An iteration method for the solution of the eigenvalue problem of linear differential and integral operators", *J. Res. Nat. Bur. Standards Sect.* B45 (1950) 255—283.

[12] LUENBERGER, D. G., "Hyperbolic pairs in the method of conjugate gradients", *SIAM J. Appl. Math.* **17** (1969) 1263—1267.

[13] MANTEUFFEL, T. A., "An incomplete factorization technique for positive definite linear systems", *Math. of Comput.*, **34** (1980) 437—497.

[14] RÓZSA, P., *Lineáris Algebra és Alkalmazásai* (Műszaki Könyvkiadó, Budapest, 1974).

[15] STOER, J., "Solution of large linear systems of equations by conjugate gradient type methods", in: *Mathematical Programming: The State of the Art, Eds. A. Bachem, M. Groetschel, B. Korte* (Springer, Berlin, 1983) 1—45.

[16] AXELSSON, O. AND LINDSKOG, G., "On the eigenvalue distribution of a class of preconditioning methods", *Numerische Mathematik* **48** (1986) 479—498.

(Beérkezett: 1985. december 2.)
(Átdolgozva beérkezett: 1987. április 16.)

HEGEDŰS CSABA J.
MTA KÖZPONTI FIZIKAI KUTATÓ INTÉZET
1525 BUDAPEST, PF. 49.

GENERAL RECURSIONS OF HESTENES—STIEFEL TYPE FOR CONJUGATE DIRECTIONS. I. THE DIRECT RECURSION

Cs. J. HEGEDŰS

A general recursion scheme is proposed for generating conjugate pairs with respect to an arbitrary matrix A. It comes to an end if the elements of the last pair are simultaneously in the zero space of matrices $\mathbf{A}^H$ and **A**.

Alkalmazott Matematikai Lapok 13 (1987—88) 97—125

SPECIÁLIS FÜGGVÉNYEK ÁLTALÁNOSÍTOTT PADÉ KÖZELÍTÉSE: SZINGULÁRIS FÜGGVÉNYEK

NÉMETH GÉZA

Budapest

Speciális függvényekre racionális típusú közelítéseket határozunk meg az *általánosított Padé módszer* alapján. Bár a közelítés intervallumában szingularitással rendelkező függvényeket vizsgálunk, a nyert (diagonális) törtek geometriai sebességgel konvergálnak a kifejtett függvényekhez a szingularitás kis környezetétől eltekintve.

1. Bevezetés

Speciális függvények polinomközelítései a numerikus analízis vagy számológépes felhasználás céljaira napjainkban már jól kidolgozottak és kézikönyvekben megtalálhatók [5—6]. Ugyanakkor sokkal kevésbé kidolgozott az elmélet racionális törtek esetén (polinomok hányadosa a közelítő kifejezés), bár az ilyen típusú kifejezések ugyanolyan könnyen és jól használhatók a gyakorlatban, mint a polinomok. Ma már sok tapasztalat alapján kialakult az a meggyőződés, hogy a racionális törtek néhány szempontból (elméletileg és gyakorlatilag) jobb közelítéseket tudnak nyújtani a polinomokhoz képest. Jelen munka tárgya racionális közelítések meghatározása néhány egyszerű, de fontos szinguláris függvény számára. Ezek az alábbiak:

(1.1) $f_1(x) = x^\mu$, $\mu > -1$, $0 < x \leqq 1$-ben és $0 < x < \infty$-ben,

(1.2) $f_2(x) = \ln x$, $0 < x \leqq 1$-ben és $0 < x < \infty$-ben,

(1.3) $f_3(x) = |x|^\varrho \operatorname{sgn}(x)$, $\varrho \geqq 0$, $-1 \leqq x \leqq 1$-ben és $-\infty < x < +\infty$-ben,

(1.4) $f_4(x) = \dfrac{x^{1-\alpha}}{\Gamma(1-\alpha)} \displaystyle\int_0^\infty \frac{t^{-\alpha}}{1+t} e^{-xt}\, dt$, $-\infty < \alpha < 1$, $0 < x < \infty$-ben.

A közelítések leveztésének módszere az *általánosított Padé approximáció*. A módszer ismertetéséhez célszerű a *Padé közelítésekről* néhány tényt említeni. Legyen adva valamely $f(x)$ függvény a *Taylor sorával*

$$f(x) = \sum_{k=0}^{\infty} c_k x^k, \tag{1.5}$$

(itt az egyenlőség nem fontos, elég az $x \to 0$-ra vonatkozó ún. aszimptotikus egyenlőség). A $P_m(x)/Q_n(x)$ racionális tört függvényt $f(x)$ *Padé közelítésének* nevezik, ha fennáll az

$$f(x) - \frac{P_m(x)}{Q_n(x)} = O(x^{n+m+1}), \quad x \to 0, \quad m = 0, 1, 2, \ldots, \quad n = 0, 1, 2, \ldots \tag{1.6}$$

reláció (ahol $P_m(x)=\sum_{j=0}^{m} p_j x^j$, $Q_n(x)=1+\sum_{j=1}^{n} q_j x^j$ polinomok). Az (1.6) relációt úgy is lehet fogalmazni, hogy a függvény és a racionális tört hatványsorát egyenlővé tesszük annyi tagig, ahányig ez egyáltalán lehetséges. Könnyen belátható, hogy a módszer olyan törteket ad, amelyek $x=0$-nál igen erős közelítést nyújtanak (de nagyobb x értékekre a hiba már természetesen nagyobbá válik). Az *általánosított Padé közelítés* célja, hogy egyenletesebb hibaeloszlást nyerjünk. Ez úgy érhető el, hogy a hatványfüggvények helyett valamilyen más polinomrendszerrel *(Jacobi, Laguerre)* dolgozunk. Ezért, kiindulva egy (adott)

$$f(x) = \sum_{k=0}^{\infty} d_k p_k(x) \tag{1.7}$$

$p_k(x)$ (polinomok) szerinti sorfejtésből határozzuk meg a $P_m(x)/Q_n(x)$ racionális tört függvényt. Két módszert említünk: *általánosított Padé módszert* [1] és a *Clenshaw —Lord módszert* [2]. Az előbbi megmagyarázásához átfogalmazzuk (1.6)-ot a következő alakba

$$Q_n(x)f(x) - P_m(x) = O(x^{n+m+1}), \quad x \to 0. \tag{1.8}$$

Tehát azt mondjuk, hogy $P_m(x)/Q_n(x)$ az $f(x)$ függvény *általánosított Padé approximációja*, ha fennáll az alábbi reláció

$$Q_n(x)f(x) - P_m(x) = O(p_{n+m+1}(x)). \tag{1.9}$$

Itt az $O(p_{n+m+1}(x))$ jelölés olyan függvényt jelent, amelynek a $p_k(x)$ polimonok szerinti sorfejtése a $k=n+m+1$-edik taggal kezdődik. A második módszer a direkt. Azt mondjuk, hogy $P_m(x)/Q_n(x)$ az $f(x)$ függvény *Clenshaw-Lord közelítése,* ha fennáll az alábbi reláció

$$f(x) - \frac{P_m(x)}{Q_n(x)} = O(p_{n+m+1}(x)). \tag{1.10}$$

A racionális közelítéseket az indexeik szerint kétdimenziós táblázatba, un. *Padé táblázatba* (ill. *általánosított Padé táblázatba*) lehet csoportosítani. Ennek első sora azonos a *Taylor sor* részletösszegeivel (ill. az (1.7) sor részletösszegeivel). Ilyen táblázatot megadni explicit képletekkel mindezideig csak igen kevés függvényre sikerült.

Végül két megjegyzést. Könnyen látható (1.9)-ből és (1.10)-ből — ha a $p_k(x)$ polinomok véges fokú rekurziós képletet elégítenek ki — hogy a *Padé approximáció* problémája lineáris egyenletrendszer megoldására vezet a az együtthatók meghatározásánál. Viszont (1.10) szerint a *Clenshaw-Lord módszer* szerinti közelítés meghatározása nem lineáris egyenletrendszerek megoldására vezet. Második megjegyzésként azt említjük meg még, hogy sok szinguláris függvény (így a cikkben következő némelyik) nem közelíthető *Padé módszerrel,* mert *Taylor sora* nincs, de kezelhető az *általánosított Padé módszerrel,* mert van polinom sorfejtése.

2. Az x^μ függvény

Az alábbi sorfejtést használjuk fel

$$x^\mu = \Gamma(1+\beta+\mu)\sum_{j=0}^{\infty}\frac{(-1)^j(2j+\lambda)\Gamma(j+\lambda)(-\mu)_j}{\Gamma(1+\beta+j)\Gamma(1+\lambda+\mu+j)}R_j^{(\alpha,\beta)}(x), \tag{2.1}$$

ahol

$$\lambda = \alpha+\beta+1,$$

$$\alpha > -1, \quad \beta > -1, \quad \mu > -1, \quad 0 \leqq x \leqq 1, \quad R_j^{(\alpha,\beta)}(x) \quad \text{az} \quad (1-x)^\alpha x^\beta$$

súlyfüggvény mellett a (0, 1) intervallumban *ortogonális Jacobi-féle polinomsorozat* ($j=0, 1, 2, \ldots, n$). Megmutatjuk, hogy az (1.9)-nek elegettevő *általánosított Padé közelítés*

$$\frac{P_m(x)}{Q_n(x)} = A_{n,m}\frac{{}_3F_2(-m, -n-\mu, n+m+\lambda+1;\ 1+\beta, 1-\mu;\ x)}{{}_3F_2(-n, -m+\mu, n+m+\lambda+\mu+1;\ 1+\mu, 1+\mu+\beta;\ x)}, \tag{2.2}$$

ahol

$$A_{n,m} = \frac{n!\Gamma(1+\beta+\mu)(1-\mu)_m\Gamma(n+m+\lambda+1)}{m!\Gamma(1+\beta)(1+\mu)_n\Gamma(n+m+\lambda+\mu+1)}. \tag{2.3}$$

A (2.1) képletben μ helyett $\mu+k$ értéket téve majd q_k számokkal szorozva és összegezve $k=0, 1, \ldots, n$-ig, kapjuk

$$x^\mu\sum_{k=0}^{n}q_kx^k = \tag{2.4}$$

$$= \sum_{j=0}^{\infty}\frac{(-1)^j(2_j+\lambda)\Gamma(j+\lambda)}{\Gamma(1+\beta+j)}R_j^{(\alpha,\beta)}(x)\sum_{k=0}^{n}q_k\frac{\Gamma(1+\beta+\mu+k)(-\mu-k)_j}{\Gamma(1+\lambda+\mu+k+j)}.$$

A kívánt közelítés meghatározásához a q_k számokat úgy kell meghatározni, hogy

$$\sum_{k=0}^{n}q_k\frac{\Gamma(1+\beta+\mu+k)(-\mu-k)_j}{\Gamma(1+\lambda+\mu+k+j)} = 0, \quad j = m+1, m+2, \ldots m+n, \tag{2.5}$$

(ahol természetesen n és m tetszőleges pozitív egész számok). A q_0 számot 1-nek véve (2.5) n egyenletet jelent n, $q_1, \ldots, q_n$ ismeretlenekre. Néhány egyszerűsítés után a (2.5) egyenlet

$$\sum_{k=0}^{n}q_k\frac{(1+\mu+\beta)_k(1+\mu)_k}{(1+\lambda+\mu+j)_k(1+\mu-j)_k} = 0 \tag{2.6}$$

lesz. Tegyük fel egy pillanatra, hogy

$$q_k = \frac{(-n)_k(-m+\mu)_k(n+m+\lambda+\mu+1)_k}{k!(1+\mu)_k(1+\beta+\mu)_k}, \quad k = 0, 1, \ldots n. \tag{2.7}$$

Ekkor (2.6) az alábbi alakra hozható

$$\sum_{k=0}^{n}\frac{(-n)_k(-m+\mu)_k(n+m+\lambda+\mu+1)_k}{k!(1+\lambda+\mu+j)_k(1+\mu-j)_k} = \tag{2.8}$$

$$= {}_3F_2(-n, -m+\mu, n+m+\lambda+\mu+1;\ 1+\lambda+\mu+j, 1+\mu-j;\ 1).$$

A nyert ${}_3F_2$ összeg ún. *Saalschütz-típusú hipergeometriai függvény,* melynek értéke faktoriálisokkal kifejezhető [4]:

$$S = \frac{(m+1-j)_n(-n-m-\lambda-j)_n}{(1+\mu-j)_n(-n-\mu-\lambda-j)_n}. \tag{2.9}$$

Elemien belátható, hogy S kifejezésében a szorzatok $j=m+1, \ldots, m+n$ értékekre zérustól különbözőek, kivéve az $(m+1-j)_n$ tagot, az viszont zérus, ezért q_k (2.7) kifejezése valóban a (2.5) egyenletrendszer megoldását adja. Ezzel természetesen meghatároztuk a (2.2) kifejezés nevezőjét. A számláló meghatározásához ki kell számítani az alábbi összeget

$$\sum_{j=0}^{m} \frac{(2j+\lambda)\Gamma(j+\lambda)}{\Gamma(1+\beta+j)} R_j^{(\alpha,\beta)}(x) \times \tag{2.10}$$

$$\times \frac{\Gamma(1+\beta+\mu)\Gamma(1+\mu)}{\Gamma(1+\lambda+\mu+j)\Gamma(1+\mu-j)} \frac{(m+1-j)_n(-n-m-\lambda-j)_n}{(1+\mu-j)_n(-n-\mu-\lambda-j)_n}.$$

Felhasználva a *Jacobi polinom* hatványokkal kifejezett alakját, kapjuk

$$\frac{\Gamma(1+\beta+\mu)\Gamma(1+\mu)\Gamma(1+\lambda)\Gamma(1+n+m)\Gamma(1+\lambda+n+m)}{\Gamma(1+\beta)\Gamma(1+\lambda+\mu+n)\Gamma(1+\mu+n)m!\Gamma(1+\lambda+m)}\cdot$$

$$\sum_{j=0}^{m} \frac{(\lambda/2)_j\left(\frac{\lambda}{2}+1\right)_j}{j!(\lambda+1)_j} \frac{(-n-\mu)_j(-m)_j(n+m+\lambda+1)_j}{(1+\lambda+\mu+n)_j(-n-m)_j(1+\lambda+m)_j} \sum_{i=0}^{j} \frac{(-j)_i(j+\lambda)_i}{i!(1+\beta)_i} x^i. \tag{2.11}$$

Az összeget x hatványai szerint rendezve

$$\sum_{i=0}^{m} \frac{(-x)^i}{i!} \frac{\Gamma(2i+\lambda+1)(-n-\mu)_i(-m)_i(n+m+\lambda+1)_i}{(\beta+1)_i(1+\lambda+\mu+n)_i(-n-m)_i(1+\lambda+m)_i}\cdot \tag{2.12}$$

$$\cdot {}_5F_4\left(2i+\lambda, i+\frac{\lambda}{2}+1, -m+i, -n-\mu+i, 1+\lambda+n+m+i; -n-m+i, i+\frac{\lambda}{2},\right.$$

$$\left. 1+\lambda+\mu+n+i, 1+\lambda+n+i; 1\right)$$

${}_5F_4$ hipergeometriai összeghez jutunk. Szerencsére ez *Dougall tétele* szerint kifejezhető faktoriálisokkal [3]. Egyszerűsítések után az

$$\frac{n!\Gamma(1+\beta+\mu)(1-\mu)_m\Gamma(n+m+\lambda+1)}{m!\Gamma(1+\beta)(1+\mu)_n\Gamma(n+m+\lambda+\mu+1)} \times \tag{2.13}$$

$$\times {}_3F_2(-m, -n-\mu, n+m+\lambda+1; 1+\beta, 1-\mu; x),$$

formulát kapjuk, ez pedig pontosan (2.2) képlet számlálója.

A $0<x<\infty$ intervallum esetén az

$$x^\mu = \Gamma(1+\beta+\mu) \sum_{j=0}^{\infty} \frac{(-\mu)_j}{\Gamma(1+\mu+j)} L_j^{(\beta)}(x), \quad \beta > -1, \tag{2.14}$$

sorfejtésből határozhatunk meg racionális törteket az *általánosított Padé-módszerrel.* Az előző módszer analógiájára végezhető el a számítás és az alábbi eredmény adódik

$$(2.15)\qquad x^{\mu}\sim\frac{n!}{m!}\,\frac{\Gamma(1+\beta+\mu)}{\Gamma(1+\beta)}\,\frac{(1-\mu)_m}{(1+\mu)_n}\,\frac{{}_2F_2(-m,\,-n-\mu;\;1+\beta,1-\mu;\;x)}{{}_2F_2(-n,\,-m+\mu;\;1+\mu,1+\beta+\mu;\;x)}.$$

Könnyen látható, hogy a (2.11) képlet határátmenettel nyerhető a (2.1) képletből. Ugyanis x helyett x/α-t téve mindkét oldalt (2.1)-ben szorozva α^{μ}-vel képezhetjük az $\alpha\to\infty$ határátmenetet. Elemien látható, hogy a súly függvény átmegy a *Lagrange súlyba,* a *Jacobi polinom* átmegy a *Laguerre polinomba* és a (0, 1) intervallum átmegy a (0, ∞)-be.

Visszatérve a (0, 1) intervallumhoz tartozó sorfejtésekhez, a (2.2) sorfejtésből α és β speciális értékeire megkaphatjuk a *Legendre* és *Csebisev sorokból* nyerhető közelítéseket stb. Ezek közül talán a *Csebisev sorból* nyert közelítés $\alpha=\beta=-\frac{1}{2}$ a legérdekesebb

$$(2.16)\qquad \frac{P_m(x)}{Q_n(x)}=\frac{\Gamma\left(\frac{1}{2}+\mu\right)}{\Gamma\left(\frac{1}{2}\right)}\,\frac{n!(1-\mu)_m\Gamma(n+m+1)}{m!(1+\mu)_n\Gamma(n+m+\mu+1)}\times$$

$$\times\frac{{}_3F_2(-m,\,-n-\mu,\,n+m+1;\;\frac{1}{2},1-\mu;\;x)}{{}_3F_2\left(-n,\,-m+\mu,\,n+m+\mu+1;\;1+\mu,\frac{1}{2}+\mu;\;x\right)}.$$

Ezzel az esettel kapcsolatos az (1.10) definíciónak megfelelő ún. *Clenshaw—Lord közelítés* (amely úgy tűnik, mindezidáig α és β-ban egyetlen lineáris egyenletekkel megoldható eset). Meg fogjuk mutatni, hogy ebben az esetben

$$(2.17)\qquad x^{\mu}\sim\frac{\Gamma\left(\frac{1}{2}+\mu\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma(1+\mu)}\,\frac{n!}{(1+2\mu)_n}\,\frac{(1-\mu)_m}{(1+\mu)_m}\times$$

$$\times\frac{{}_4F_3\left(-m,\,m+1,\,-n-\mu,\,n+\mu+1;\;1-\mu,1+\mu,\frac{1}{2};\;x\right)}{{}_4F_3\left(-n,\,n+2\mu+1,\,-m+\mu,\,m+\mu+1;\;\frac{1}{2}+\mu,1+\mu,1+2\mu;\;x\right)}.$$

A (2.17) képlet levezetéséhez [9] alapján az

$$(2.18)\qquad x^{\mu}=\frac{-\Gamma\left(\frac{1}{2}+\mu\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma(1+\mu)}\left\{1+2\sum_{k=1}^{\infty}(-1)^k\frac{(-\mu)_k}{(1+\mu)_k}T_k^*(x)\right\}$$

sort használjuk fel. Először a nevező meghatározását végezzük el. Meg kell oldani g_i $(i=0, 1, 2, \ldots, n)$-re

$$\sum_{i=0}^{n} a_{l+r+i+1} g_i = 0, \quad r = 0, 1, 2, \ldots, n-1, \tag{2.19}$$

egyenletrendszert, ahol $a_k=(-1)^k \frac{(-\mu)_k}{(1+\mu)_k}$. Ez az egyenlet az alábbi alakra hozható $(l=m-n)$

$$\sum_{i=0}^{n} (-1)^i g_i \frac{\Gamma(1+l+r-\mu+i)}{\Gamma(2+l+r+\mu+i)} = 0, \tag{2.20}$$

vagy

$$\int_0^1 (1-t)^{2\mu} t^{r+l-\mu} \left\{ \sum_{i=0}^{n} (-t)^i g_i \right\} dt = 0. \tag{2.21}$$

Ha a zárójelben álló polinom az $(1-t)^{2\mu} t^{l-\mu}$ súlyra vonatkozó *Jacobi polinom*, akkor az ortogonalitás miatt az összes feltétel teljesül. Így $g_0=1$ választással

$$\sum_{i=0}^{n} (-t)^i g_i = \frac{1}{(l-\mu+1)_n} (1-t)^{-2\mu} t^{-l+\mu} \frac{d^n}{dt^n} [(1-t)^{n+2\mu} t^{n+l-\mu}] = \tag{2.22}$$

$$= {}_2F_1(-n, n+l+\mu+1;\ l-\mu+1;\ t).$$

Tehát

$$g_i = \frac{(-n)_i (n+l+\mu+1)_i}{i!(l-\mu+1)_i} (-1)^i. \tag{2.23}$$

A nevezőpolinom konstansszorosa

$$N(x) = 2 \sum_{k=0}^{n} T_k^*(x) \sum_{k=0}^{n-k} g_i g_{i+k} = \tag{2.24}$$

$$= {}_2F_1\left(-n, n+l+\mu+1;\ l-\mu+1;\ 1-2x+2\sqrt{x^2-x}\right) \times$$

$$\times {}_2F_1\left(-n, n+l+\mu+1;\ l-\mu+1;\ 1-2x-2\sqrt{x^2-x}\right).$$

A hipergeometriai függvények szorzatát sorrá alakítjuk [4]

$$N(x) = \sum_{i=0}^{\infty} \frac{(-n)_i (n+l+\mu+1)_i (n+l-\mu+1)_i (-n-2\mu)_i}{i!(l-\mu+1)_i (l-\mu+1)_{2i}} \cdot \tag{2.25}$$

$$\cdot {}_2F_1(-n+i, n+i+l+\mu+1;\ l+2i-\mu+1;\ 1-4x).$$

A belső hipergeometriai függvényt átrendezzük x hatványai szerint

$${}_2F_1(-n+i, n+i+l+\mu+1;\ l+2i-\mu+1;\ 1-4x) = \tag{2.26}$$

$$= \frac{\Gamma(-2\mu)\Gamma(l+2i-\mu+1)}{\Gamma(i-n-2\mu)\Gamma(l+n+i-\mu+1)} {}_2F_1(-n+i, n+i+l+\mu+1;\ 1+2\mu;\ 4x).$$

Beírva ezt az előző sorba, az alaposan egyszerűsödik

$$N(x)\left[\frac{(-1)^n(1+2\mu)_n}{(l-\mu+1)_n}\right]^{-1} = \sum_{i=0}^{\infty}\frac{(-n)_i(n+l+\mu+1)_i}{i!(l-\mu+1)_i}\cdot \tag{2.27}$$

$$\cdot {}_2F_1(-n+i, n+i+l+\mu+1;\ 1+2\mu;\ 4x) =$$

$$= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\frac{(-n)_{i+j}(n+l+\mu+1)_{i+j}}{i!j!(l-\mu+1)_i(1+2\mu)_j}(4x)^j =$$

$$= \sum_{j=0}^{n}(4x)^j\frac{(-n)_j(n+l+\mu+1)_j}{j!(1+2\mu)_j}\sum_{i=0}^{n-j}\frac{(-n+j)_i(n+l+j+\mu+1)_i}{i!(l-\mu+1)_i}.$$

Az utóbbi második összeg az

$${}_2F_1(-s, b;\ c;\ 1) = \frac{(c-b)_s}{(c)_s},\ s = 0, 1, \ldots, \tag{2.28}$$

képlet segítségével felösszegezhető és értéke

$$\frac{1}{4^j}\frac{(-n-l+\mu)_j(n+2\mu+1)_j}{(\mu+1)_j\left(\mu+\frac{1}{2}\right)_j}\frac{(-1)^n(1+2\mu)_n}{(l-\mu+1)_n}. \tag{2.29}$$

Végül a nevező konstansszorosa

$$N(x) = \left[\frac{(1+2\mu)_n}{(l-\mu+1)_n}\right]^2\cdot \tag{2.30}$$

$$\cdot {}_4F_3\left(-n, n+2\mu+1, -n-l+\mu;\ n+l+\mu+1;\ \mu+\frac{1}{2}, \mu+1, 2\mu+1;\ x\right).$$

A számláló levezetéséhez megjegyezzük, hogy az (1.10) definíció szerint egyenlő az alábbi kifejezéssel

$$\sum_{k=0}^{n}\frac{(-n)_k(n+2\mu+1)_k(-m+\mu)_k(m+\mu+1)_k}{k!\left(\mu+\frac{1}{2}\right)_k(\mu+1)_k(2\mu+1)_k}\frac{\Gamma\left(k+\mu+\frac{1}{2}\right)}{\Gamma(k+\mu+1)\Gamma\left(\frac{1}{2}\right)}* \tag{2.31}$$

$$*\left\{1+2\sum_{j=1}^{m}(-1)^j\frac{(-\mu-k)_j}{(1+\mu+k)_j}T_j^*(x)\right\}.$$

Átalakítjuk kissé ezeket az összegeket

$$\frac{\Gamma\left(\mu+\frac{1}{2}\right)\Gamma(\mu+1)}{\Gamma(1/2)}\sum_{k=0}^{n}\frac{(-n)_k(n+2\mu+1)_k(-m+\mu)_k(m+\mu+1)_k}{k!(2\mu+1)_k}* \tag{2.32}$$

$$*\left\{\sum_{j=0}^{m}{}^*T_j^*(x)\frac{1}{\Gamma(1+\mu+k+j)\Gamma(1+\mu+k-j)}\right\}.$$

Először a kapcsos zárójelben álló összeget számítjuk ki. Rendezve x hatványai szerint kapjuk:

$$(2.33)\qquad \sum_{i=0}^{m} \frac{x^i}{i!\left(\frac{1}{2}\right)_i}\Gamma(2i+1)\sum_{j=0}^{m-i}\frac{(-1)^i(2i)_j(i+1)_j}{\Gamma(1+\mu+k+i+j)\Gamma(1+\mu+k-i-j)(i)_j j!}.$$

A belső összeg

$$(2.34)\qquad \frac{1}{\Gamma(1+\mu+k+i)\Gamma(1+\mu+k-i)}\sum_{j=0}^{m-i}\frac{(2i)_j(i+1)_j}{j!(i)_j}\frac{(-\mu-k+i)_j}{(1+\mu+k+i)_j},$$

egy alkalmas

$$\frac{(-m+i)_j}{(-m+i)_j}\frac{(1+i+m)_j}{(1+i+m)_j}\equiv 1$$

tényező beszúrásával ${}_5F_4$ hipergeometriai összeg, amely DOUGALL [3] tételével kiszámítható, és értéke

$$(2.35)\qquad \frac{\Gamma(1+i+m)(-1)^{m-i}}{(\mu+k-i)\Gamma(1+2i)\Gamma(1+\mu+k+m)\Gamma(\mu+k-m)(m-i)!}.$$

A számláló így az alábbi

$$(2.36)\qquad \sum_{i=0}^{m} x^i\frac{(-m)_i(m+1)_i}{i!\left(\frac{1}{2}\right)_i(\mu-i)}\frac{\Gamma\left(\mu+\frac{1}{2}\right)\Gamma(\mu+1)(-1)^m}{\Gamma\left(\frac{1}{2}\right)\Gamma(1+\mu+m)\Gamma(\mu-m)}*$$

$$*\sum_{k=0}^{n}\frac{(-n)_k(n+2\mu+1)_k(\mu-i)_k}{k!(2\mu+1)_k(1+\mu-i)_k}.$$

Az itt szereplő k szerinti összeg *Saalschütz típusú* ${}_3F_2$ *sor* és értéke

$$(2.37)\qquad \frac{(-1)^n n!(\mu+1+i)_n}{(2\mu+1)_n(-n-\mu+i)_n}.$$

Beírva ezt az előző összegbe, néhány egyszerűsítés után kapjuk

$$(2.38)\qquad \frac{\Gamma\left(\mu+\frac{1}{2}\right)}{\Gamma(\mu+1)\Gamma\left(\frac{1}{2}\right)}\frac{n!}{(1+2\mu)_n}\frac{(1-\mu)_m}{(1+\mu)_m}\times$$

$$\times{}_4F_3\left(-m, m+1, -n-\mu, n+\mu+1; \frac{1}{2}, 1+\mu, 1-\mu; x\right),$$

ez pedig azonos a (2.17) képlet számlálójával. (2.17) levezetésénél a g_i számok meghatározásánál feltettük hallgatólagosan, hogy $l\geqq 0$ azaz $m\geqq n$. Mégis az $m=0$, $\mu=1$ eset (vö. 1. Függelék) azt sugallja, hogy valószínűleg a teljes (1.10) *Padé táblázatot* ($m\geqq 0$, $n\geqq 0$) megadja a (2.17) képlet.

Végül egy egyszerű megjegyzés a (2.16) és (2.17) képletek összehasonlításáról. Tekintve az $m=n$ esetet (diagonális közelítés) elemi számítással adódik, hogy $n\to\infty$ esetén a konvergencia az $x=0$ helyen *Csebisev esetben* $O(n^{-3\mu})$, míg a *Clenshaw—Lord esetben* $O(n^{-4\mu})$. A pontosabb hibabecslésre ($x>0$) később még visszatérünk.

3. A logaritmus függvény

A (2.2) közelítés (valamint a (2.17) közelítés) rendelkezik egy igen érdekes tulajdonsággal. Nevezetesen $\mu\to 0$ esetén $m=n$ feltétel mellett a racionális törtek mind 1-hez tartanak. Ezért az

$$\ln x = \lim_{\mu\to 0}\frac{x^{\mu}-1}{\mu}$$

reláció felhasználásával a logaritmus függvényre racionális tört közelítéseket nyerhetünk. (Természetesen ugyanez az algoritmus alkalmazható $\{\ln x\}^k$-re is, ahol $k\geqq 1$ egész.)

A (2.2) képletből adódik

$$\ln x \sim \frac{\sum_{k=0}^{n}\dfrac{(-n)_k(-n)_k(2n+\lambda+1)_k}{k!\,k!\,(1+\beta)_k}\,h_k^{(n)}\,x^k}{{}_3F_2(-n,\,-n,\,2n+\lambda+1;\;1,\,1+\beta;\;x)}, \qquad 0<x\leqq 1, \tag{3.1}$$

ahol

$$h_k^{(n)} = 2\psi(k+1)-2\psi(n-k+1)-\psi(2n+\lambda+k+1)+\psi(k+\beta+1), \tag{3.2}$$

$$\ln x \sim \frac{\sum_{k=0}^{n}\dfrac{(-n)_k(-n)_k}{k!\,k!\,(1+\beta)_k}\,t_k^{(n)}\,x^k}{{}_2F_2(-n,\,-n;\;1,\,1+\beta;\;x)}, \qquad 0<x<\infty, \tag{3.3}$$

ahol

$$t_k = 2\psi(k+1)-2\psi(n-k+1)+\psi(k+\beta+1). \tag{3.4}$$

A (2.17) képletből adódik

$$\ln x \sim \frac{\sum_{k=0}^{n}\dfrac{(-n)_k^2(n+1)_k^2}{k!^3(1/2)_k}\,g_k^{(n)}\cdot x^k}{{}_4F_3\left(-n,\,-n,\,n+1,\,n+1;\;\dfrac{1}{2},\,1,\,1;\;x\right)}, \qquad 0<x\leqq 1, \tag{3.5}$$

ahol

$$g_k^{(n)} = \psi\left(k+\frac{1}{2}\right)+3\psi(k+1)-2\psi(n+k+1)-2\psi(n-k+1). \tag{3.6}$$

Az $n\neq m$ indexű közelítések határátmenettel nem nyerhetők, de az előző módszer kis módosítással mégis alkalmazható (amint azt rövidesen látni fogjuk).

Tekintsük $x^{\mu}\ln x$ *Jacobi sorát*

$$x^{\mu}\ln x = \frac{d}{ds}\Big|_{s=0} \cdot \sum_{j=0}^{\infty}(-1)^j \frac{(2j+\lambda)\Gamma(j+\lambda)}{\Gamma(1+\beta+j)} \times \tag{3.7}$$

$$\times R_j^{(\alpha,\beta)}(x)\frac{\Gamma(1+\beta+\mu+s)(-\mu-s)_j}{\Gamma(1+\lambda+j+\mu+s)}.$$

Végezzük el a $\mu \to l$ (l egész) helyettesítést, határátmenettel kapjuk

$$x^l \ln x = \sum_{j=0}^{l}(-1)^j \frac{(2j+\lambda)\Gamma(j+\lambda)}{\Gamma(1+\beta+j)} R_j^{(\alpha,\beta)}(x)\frac{\Gamma(1+\beta+l)}{\Gamma(1+\lambda+l+j)} \times \tag{3.8}$$

$$\times[\psi(1+\beta+l)-\psi(1+\lambda+l+j)+\psi(l+1)-\psi(l-j+1)](-l)_j -$$

$$-\sum_{j=l+1}^{\infty}(-1)^j \frac{(2j+\lambda)\Gamma(j+\lambda)}{\Gamma(1+\beta+j)} R_j^{(\alpha,\beta)}(x)\frac{\Gamma(1+\beta+l)}{\Gamma(1+\lambda+l+j)}(-1)^l l!(j-l-1)!.$$

A (2.4)—(2.9) alatti módon eljárva láthatjuk, hogy a q_k nevező együtthatóknak ki kell elégíteni az alábbi egyenletet

$$\sum_{k=0}^{n} q_k \frac{k!(1+\beta)_k}{(1-j)_k(1+j+\lambda)_k} = 0, \quad j = m+1, m+2, \ldots m+n. \tag{3.9}$$

Ez (2.6) $\mu=0$ esethez tartozó speciális egyenlete. Ezért

$$q_k = \frac{(-n)_k(-m)_k(n+m+\lambda+1)_k}{k!^2(1+\beta)_k}, \quad k = 0, 1, \ldots n. \tag{3.10}$$

A nevező ismeretében a számláló pedig könnyen nyerhető:

$$P_m(x) = \sum_{l=0}^{n}\frac{(-n)_l(-m)_l(n+m+\lambda+1)_l}{l!\,l!} \times \tag{3.11}$$

$$\times \sum_{j=0}^{l}(-1)^j \frac{(2j+\lambda)\Gamma(j+\lambda)\Gamma(1+\beta)(-l)_l}{\Gamma(1+\beta+j)\Gamma(1+\lambda+l+j)} \times$$

$$\times[\psi(1+\beta+l)-\psi(1+\lambda+l+j)+\psi(l+1)-\psi(l-j+1)]R_j^{(\alpha,\beta)}(x) -$$

$$-\sum_{l=0}^{n}\frac{(-n)_l(-m)_l(n+m+\lambda+1)_l}{l!}(-1)^l\Gamma(1+\beta)\times$$

$$\times \sum_{j=l+1}^{m}(-1)^j \frac{(2j+\lambda)\Gamma(j+\lambda)(j-l-1)!}{\Gamma(1+\beta+j)\Gamma(1+\lambda+l+j)} R_j^{(\alpha,\beta)}(x), \quad m \geqq n.$$

Hasonló gondolatmenettel nyerhető a *Clenshaw—Lord közelítés* is, tegyünk $\alpha=\beta=$

$= -\frac{1}{2}$-et a (3.8) képletbe, ekkor kapjuk

$$x^l \ln x = \frac{\Gamma\left(l+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma(l+1)} \sum_{j=0}^{l}{}^{*}(-1)^j T_j^*(x) \frac{(-l)_j}{(l+1)_j} \times \tag{3.12}$$

$$\times\left[\psi\left(l+\frac{1}{2}\right)+\psi(l+1)-\psi(l-j+1)-\psi(1+j+l)\right]-$$

$$-\sum_{j=l+1}^{\infty}(-1)^j T_j^*(x)(-1)^l \frac{(1/2)_l}{(l+1)_j}(j-l-1)!.$$

A (3.12) sor $l=0$ esete adja a (2.18)-cal analóg egyenletet a g_i számok meghatározásához

$$\sum_{i=0}^{n} b_{s+r+i+1} g_i = 0, \quad r = 0, 1, \ldots, n-1, \quad b_k = \frac{(-1)^k}{k}, \quad s = m-n. \tag{3.13}$$

Ez az egyenlet az alábbi alakra hozható

$$\int_0^1 t^{s+r}\left\{\sum_{i=0}^{n}(-t)^i g_i\right\} dt = 0. \tag{3.14}$$

Világos, hogy ez (2.21) $\mu=0$ esete, ezért a megoldás

$$g_i = \frac{(-n)_i (n+s+1)_i (-1)^i}{i!(s+1)_i}, \quad m \geqq n. \tag{3.15}$$

Innen a nevező már azonnal következik

$$ {}_4F_3\left(-n, n+1, -m, m+1; \frac{1}{2}, 1, 1; x\right). \tag{3.16}$$

A nevező ismeretében a számláló könnyen felírható (3.12) alapján

$$P_m(x) = \sum_{l=0}^{n} \frac{(-n)_l (n+1)_l (-m)_l (m+1)_l}{l!^4} \sum_{j=0}^{l}{}^{*}(-1)^j T_j^*(x) \frac{(-l)_j}{(l+1)_j} \times \tag{3.17}$$

$$\times\left[\psi\left(l+\frac{1}{2}\right)+\psi(l+1)-\psi(l-j+1)-\psi(l+j+1)\right]-$$

$$-\sum_{l=0}^{n} \frac{(-n)_l (n+1)_l (-m)_l (m+1)_l}{l!^3}(-1)^l \sum_{j=l+1}^{m}(-1)^j T_j^*(x) \frac{(j-l-1)!}{(l+1)_j}.$$

A $0<x<\infty$ intervallum esetén *Laguerre sorból* célszerű számolni

$$x^l \ln x = \sum_{j=0}^{l} \frac{\Gamma(1+\beta+l)}{\Gamma(1+\beta+j)} (-l)_j [\psi(1+\beta+l)+\psi(l+1)-\psi(l-j+1)] L_j^{(\beta)}(x) - \tag{3.18}$$

$$- \sum_{j=l+1}^{\infty} \frac{\Gamma(1+\beta+l)}{\Gamma(1+\beta+j)} (-1)^l l! (j-l-1)! L_j^{(\beta)}(x).$$

A nevező polinom q_l együtthatójára az alábbi egyenlet adódik

$$\sum_{l=0}^{n} q_l \frac{(1+\beta)_l l!}{(1-j)_l} = 0, \quad j = m+1, m+2, \ldots, m+n, \tag{3.19}$$

amelynek megoldása

$$q_l = \frac{(-n)_l (-m)_l}{l!^2 (1+\beta)_l}, \quad l = 0, 1, \ldots n, \quad m \geqq n. \tag{3.20}$$

Ezért a nevező

$$ {}_2F_2(-n, -m; 1, 1+\beta; x), \tag{3.21}$$

a számláló pedig (3.18) alapján

$$P_m(x) = \sum_{l=0}^{n} \frac{(-n)_l (-m)_l}{l!^2} \sum_{j=0}^{l} \frac{(-l)_j}{(1+\beta)_j} \times \tag{3.22}$$

$$\times [\psi(1+\beta+l)+\psi(1+l)-\psi(l-j+1)] L_j^{(\beta)}(x) -$$

$$- \sum_{l=0}^{n} \frac{(-n)_l (-m)_l}{l!} (-1)^l \sum_{j=l+1}^{m} \frac{(-l)_j}{(1+\beta)_j} (j-l-1)! L_j^{(\beta)}(x).$$

Az $\ln^2 x$, $\ln^3 x$, stb. függvények esetén az együtthatókban megjelennek ψ', ψ'' stb. loggamma deriváltak. Bár az algoritmus direkt, a képletek nagyon elbonyolódnak.

4. Signum függvény

A $\operatorname{sgn}(x)$ függvény

$$\operatorname{sgn}(x) = \begin{cases} -1, & -1 \leqq x < 0, \\ 0, & x = 0, \\ 1, & 0 < x \leqq 1, \end{cases} \tag{4.1}$$

racionális közelítéseit határozzuk meg a következőkben. Kissé általánosabban, a célszerűség miatt, az alábbi

$$\varphi(x) = |x|^{\varrho} \cdot \operatorname{sgn}(x), \quad \varrho \geqq 0 \tag{4.2}$$

függvény közelítéseit fogjuk tárgyalni. A

$$\varphi(x) = \frac{\Gamma(\sigma)}{\sqrt{\pi}} \Gamma\left(\frac{1}{2}+\frac{\varrho}{2}\right) \Gamma\left(1+\frac{\varrho}{2}\right) \sum_{j=0}^{\infty} \frac{(2j+\sigma+1)}{\Gamma\left(\frac{3}{2}+\sigma+\frac{\varrho}{2}+j\right) \Gamma\left(\frac{1}{2}+\frac{\varrho}{2}-j\right)} C_{2j+1}^{(\sigma)}(x), \tag{4.3}$$

sorfejtésből indulunk ki, ahol $C_i^{(\sigma)}(x)$ az i-edik *Gegenbauer polinom* $\sigma > 0$. Meg fogjuk mutatni, hogy $\varphi(x)$ *általánosított Padé közelítése*

$$(4.4)\qquad R_{n,m}(x) = \frac{2x}{\sqrt{\pi}}\Gamma\left(1+\frac{\varrho}{2}\right)\frac{n!}{m!}\frac{\Gamma(n+m+\sigma+2)\left(\frac{3}{2}-\frac{\varrho}{2}\right)_m}{\Gamma(n+m+\sigma+\varrho/2+3/2)\left(\frac{1}{2}+\frac{\varrho}{2}\right)_n} *$$

$$* \frac{{}_3F_2\left(-m,\ -n-\frac{\varrho}{2}+\frac{1}{2},\ n+m+\sigma+2;\ \frac{3}{2},\ \frac{3}{2}-\frac{\varrho}{2};\ x^2\right)}{{}_3F_2\left(-n,\ -m+\frac{\varrho}{2}-\frac{1}{2},\ n+m+\sigma+\frac{\varrho}{2}+\frac{3}{2};\ \frac{1}{2}+\frac{\varrho}{2},\ 1+\frac{\varrho}{2};\ x^2\right)}.$$

Ebből a közelítésből $-\infty < x < \infty$ intervallumra érvényes közelítések a $\sigma^{\varrho/2}\varphi\left(\frac{x}{\sigma^{1/2}}\right)$ kifejezésből $\sigma \to \infty$ esetén nyerhetők (természetesen ugyanez adódik a *Hermite sorból* is). Itt természetesen Sgn (x) fog szerepelni

$$(4.5)\qquad \operatorname{Sgn}(x) = \begin{cases} -1, & -\infty < x < 0, \\ 0, & x = 0, \\ 1, & 0 < x < \infty. \end{cases}$$

Az eredmény a következő:

$$(4.6)\qquad |x|^{\varrho}\operatorname{Sgn}(x) \sim \frac{2x}{\sqrt{\pi}}\Gamma\left(1+\frac{\varrho}{2}\right)\frac{n!}{m!}\frac{\left(\frac{3}{2}-\frac{\varrho}{2}\right)_m}{\left(\frac{1}{2}+\frac{\varrho}{2}\right)_n} *$$

$$* \frac{{}_2F_2\left(-m,\ -n-\frac{\varrho}{2}+\frac{1}{2};\ \frac{3}{2},\ \frac{3}{2}-\frac{\varrho}{2};\ x^2\right)}{{}_2F_2\left(-n,\ -m+\frac{\varrho}{2}-\frac{1}{2};\ \frac{1}{2}+\frac{\varrho}{2},\ 1+\frac{\varrho}{2};\ x^2\right)}.$$

A (4.4) képlet levezetése teljesen analóg az előzőkben használt módszerrel. Röviden összefoglalva: ϱ-t $\varrho+2k$-val helyettesítve, q_k-val szorozva és összegezve $0 \leqq k \leqq n$-re a (4.3) egyenletben, a q_k nevező együtthatókra egyenletrendszert nyerünk, ezt megoldva a *Saalschütz-tétel* segítségével nyerjük a (4.4)-ben szereplő nevező polinomot. A számláló polinom meghatározásához az ${}_5F_4$ függvényre vonatkozó *Dougall-tételt* lehet felhasználni.

A képleteket alkalmazásaként $\varrho = 0$, $\sigma \neq 0$ esetben az ún. *Gibbs konstans* meghatározására használtuk [8]. A következőkben a $\sigma = 0$ esetben, de a *Clenshaw—Lord közelítés* szerint határozunk meg racionális törteket a $\varphi(x)$ függvényhez. Ered-

ményünk a következő:

$$\varphi(x)\sim C_{n,m}\times \tag{4.7}$$

$$\times x\frac{{}_4F_3\left(-m,\, m+2,\, -n-\frac{\varrho}{2}+\frac{1}{2},\, n+\frac{\varrho}{2}+\frac{3}{2};\ \frac{3}{2},\, \frac{3}{2}-\frac{\varrho}{2},\, \frac{3}{2}+\frac{\varrho}{2};\ x^2\right)}{{}_4F_3\left(-n,\, n+\varrho+1,\, -m+\frac{\varrho}{2}-\frac{1}{2},\, m+\frac{\varrho}{2}+\frac{3}{2};\ \frac{1}{2}+\frac{\varrho}{2},\, 1+\frac{\varrho}{2},\, 1+\varrho;\ x^2\right)},$$

ahol

$$C_{n,m}=\frac{2}{\sqrt{\pi}}\frac{\Gamma\left(1+\frac{\varrho}{2}\right)}{\Gamma\left(\frac{3}{2}+\frac{\varrho}{2}\right)}\cdot\frac{\left(n+\frac{1}{2}+\frac{\varrho}{2}\right)n!}{(1+\varrho)_n}\frac{\left(\frac{3}{2}-\frac{\varrho}{2}\right)_m}{\left(\frac{1}{2}+\frac{\varrho}{2}\right)_{m+1}}(m+1). \tag{4.8}$$

A levezetés hasonló a (2.19)—(2.34)-nél részletezett számításhoz. Először a nevezőt határozzuk meg [9] alapján a (4.3)-ból vett együtthatók segítségével. Az itteni g_i együtthatók integrál előállításból adódnak az ortogonalitás figyelembevételével. Ezután [4] képlete használható a nevező együtthatóinak meghatározására. Végül a számláló kiszámításához az ${}_5F_4$ *Dougall-tételt*, valamint az ${}_3F_2$ *Saalschütz-tételt* kell használni.

A (4.7) képlet a $\varrho=0$ esetben *Gibbs konstansok* meghatározására volt alkalmazva [7]-ban.

5. Nem teljes gamma függvények

A nem teljes gamma függvényekre [6]

$$f(x)=x^{1-\alpha}e^x\Gamma(\alpha,x)=\frac{x^{1-\alpha}}{\Gamma(1-\alpha)}\int_0^\infty t^{-\alpha}\frac{e^{-t}}{1+t}\,dt, \tag{5.1}$$

határozunk meg racionális közelítéseket $0<x<\infty$-re. Az alábbi sorfejtést fogjuk felhasználni:

$$f(x)=x\sum_{k=0}^{\infty}\frac{1}{k+1}L_k^{(\alpha)}(x)=\frac{1+\alpha}{2}-(1-\alpha)\sum_{k=1}^{\infty}\frac{1}{(k+1)(k+2)}L_k^{(\alpha)}(x). \tag{5.2}$$

Az $f(x)$ függvény *általánosított Padé közelítése*

$$f(x)\sim\frac{P_m(x)}{Q_n(x)},\quad m\geqq n, \tag{5.3}$$

ahol

$$Q_n(x)={}_2F_2(-n,\, n+m+3;\ 2,\, 2-\alpha;\ -x), \tag{5.4}$$

$$P_m(x)=\sum_{j=0}^{n}\frac{(-n)_j(n+m+3)_j}{j!(2)_j(2-\alpha)_j}\sum_{l=0}^{j}(-x)^l(1-\alpha)_{j-l}- \tag{5.5}$$

$$-(-1)^n(1-\alpha)\frac{\Gamma(n+m+1)}{\Gamma(m+1)\Gamma(n+3)}\sum_{k=0}^{m}\frac{(-m)_k k!}{(n+3)_k(-n-m)_k}L_k^{(\alpha)}(x).$$

Az (5.4) és (5.5) levezetése $x^l f(x)$ *Laguerre sorának* képzésével végezhető el. Legyen

$$x^l f(x) = \sum_{k=0}^{\infty} c_k^{(l)} L_k^{(\alpha)}(x). \tag{5.6}$$

Látjuk (5.2)-ből, hogy

$$c_k^{(0)} = -(1-\alpha)\frac{1}{(k+1)(k+2)}, \quad k > 1. \tag{5.7}$$

Tegyük fel, hogy

$$c_k^{(l)} = \frac{A_l}{\prod_{j=1}^{l+2}(k+j)}, \quad k > l. \tag{5.8}$$

Szorozzuk az (5.6) egyenletet x-szel és alkalmazzuk az

$$xL_k^{(\alpha)}(x) = -(k+1)L_{k+1}^{(\alpha)}(x) + (2k+1+\alpha)L_k^{(\alpha)}(x) - (k+\alpha)L_{k-1}^{(\alpha)}(x) \tag{5.9}$$

rekurziós képletet tagonként az (5.6) jobb oldalán levő sor tagjaira, majd gyűjtsük össze $L_k^{(\alpha)}(x)$ együtthatóit, az lesz $c_k^{(l+1)}$ $(k>l+1)$

$$c_k^{(l+1)} = A_l \frac{1}{\prod_{j=1}^{l+3}(k+j)} \times \tag{5.10}$$

$$\times[-(k+l+2)(k+l+3) + (2k+1+\alpha)(k+l+3) - (k+1)(k+\alpha+1)].$$

A szögletes zárójel (5.10)-ben $-(l+2)(l+2-\alpha)$, ezért

$$A_{l+1} = -(l+2)(l+2-\alpha)A_l, \tag{5.11}$$

vagyis

$$A_l = (-1)^l (2)_l (2-\alpha)_l, \tag{5.12}$$

tehát

$$c_k^{(l)} = \frac{(-1)^l (2)_l (2-\alpha)_l}{\prod_{j=1}^{l+2}(k+j)}, \quad k > l. \tag{5.13}$$

A nevező polinom q_l együtthatójára fenn kell állni az alábbi egyenletrendszernek

$$\sum_{l=0}^{n} q_l \frac{(-1)^l (2)_l (2-\alpha)_l}{\Gamma(k+l+3)} k! = 0, \quad k = m+1, m+2, \ldots, m+n. \tag{5.14}$$

Rövid számolás után ezt az egyenletrendszert átírhatjuk az alábbiba

$$\int_0^1 t(1-t)^k \Big\{ \sum_{l=0}^{n} q_l (-t)^l (2-\alpha)_l \Big\} dt = 0. \tag{5.15}$$

Ha az integrálban az összeget úgy választjuk, hogy

$$\sum_{l=0}^{n} q_l (2-\alpha)_l (-t)^l = \frac{1}{(2)_l} t^{-1}(1-t)^{-m-1} \frac{d^n}{dt^n}[t^{n+1}(1-t)^{n+m+1}], \tag{5.16}$$

az (5.14) egyenletrendszer ki van elégítve, a megoldás

$$q_l = \frac{(-n)_l (n+m+3)_l (-1)^l}{l!\,(2)_l (2-\alpha)_l}, \quad l = 0, 1, \ldots, n,$$

és ezzel a nevező alakját bebizonyítottuk. Használjuk fel az (5.15) képletet $Q_n(x)f(x) - P_m(x)$ explicit meghatározásához. Évégből az (5.6) egyenletet szorozzuk q_l-lel és összegezzük $l = 0, 1, \ldots, n$-re

$$\sum_{k=n+m+1}^{\infty} L_k^{(\alpha)}(x) \sum_{l=0}^{n} q_l c_k^{(l)} = \sum_{k=n+m+1}^{\infty} F_k L_k^{(\alpha)}(x), \tag{5.17}$$

$$F_k = (\alpha-1) \int_0^1 t(1-t)^k \sum_{l=0}^{n} q_l (-t)^l (2-\alpha)_l \, dt =$$

$$= (\alpha-1) \frac{1}{(2)_n} \int_0^1 (1-t)^{k-m-1} \frac{d^n}{dt^n} [t^{n+1}(1-t)^{n+m+1}] \, dt = \tag{5.18}$$

$$= (\alpha-1) \frac{(-1)^n (1+m+k)_n}{(2)_n} \int_0^1 t^{n+1}(1-t)^k \, dt = (-1)^n \frac{(1+m-k)_n k!}{\Gamma(2n+k+3)} (\alpha-1).$$

Tehát

$$Q_n(x)f(x) - P_m(x) = (\alpha-1) \frac{n!\,\Gamma(n+m+2)}{\Gamma(2n+m+4)} \sum_{j=0}^{\infty} \frac{(n+1)_j (n+m+2)_j}{j!\,(2n+m+4)_j} L_{n+m+j+1}^{(\alpha)}(x). \tag{5.19}$$

Ezen képlet alapján $P_m(x)$ meghatározható. Sajnos nem triviális, hogy a két végtelen sor különbsége hogyan lesz polinom. Mégis, hogy ezt belássuk, *Laplace transzformáltakat* fogunk használni

$$\int_0^{\infty} P_m(x) x^{\alpha} e^{-px} \, dx = I_1 + I_2. \tag{5.20}$$

Ekkor

$$I_1 = \frac{1}{\Gamma(1-\alpha)} \int_0^{\infty} \frac{t^{-\alpha}}{1+t} \sum_{j=0}^{n} \frac{(-n)_j (n+m+3)_j}{j!\,(2-\alpha)_j} \frac{(-1)^j}{(p+t)^{j+2}} \, dt = \tag{5.21}$$

$$= \frac{1}{\Gamma(1-\alpha)} \int_0^{1/p} u^{1+\alpha}(1-pu)^{-\alpha} {}_2F_1(-n, n+m+3; 2-\alpha; -u) \frac{du}{1-u(p-1)} =$$

$$= \frac{1}{\Gamma(1-\alpha) p^{2+\alpha}} \int_0^1 u^{1+\alpha}(1-u)^{-\alpha} {}_2F_1\left(-n, n+m+3; 2-\alpha; -\frac{u}{p}\right) \frac{du}{1-u\left(1-\frac{1}{p}\right)}$$

Továbbá

$$I_2 = (1-\alpha)\frac{\Gamma(n+1)\Gamma(n+m+2+\alpha)}{\Gamma(2n+m+2)p^{1+\alpha}}\left(1-\frac{1}{p}\right)^{n+m+1}\times$$

$$\times {}_2F_1\left(n+1, n+m+2+\alpha;\ 2n+m+4;\ 1-\frac{1}{p}\right) = (1-\alpha)\frac{\Gamma(n+1)}{\Gamma(n+2-\alpha)p^{1+\alpha}}\times$$

$$\times\left(1-\frac{1}{p}\right)^{n+m+1}\int_0^1 t^{n+m+\alpha+1}(1-t)^{n+1-\alpha}\frac{dt}{\left[1-t\left(1-\frac{1}{p}\right)\right]^{n+1}} =$$

$$= (1-\alpha)\frac{(-1)^n}{\Gamma(n+2-\alpha)p^{1+\alpha}}\left(1-\frac{1}{p}\right)^{n+1}\int_0^1 \frac{d^n}{dt^n}[t^{n+m+\alpha+1}(1-t)^{n+1-\alpha}]\frac{dt}{1-t\left(1-\frac{1}{p}\right)}.$$

Használjuk fel a következő átalakításokat:

(3.52)
$$\frac{d^n}{dt^n}[t^{n+m+\alpha+1}(1-t)^{n+1-\alpha}] =$$

$$= (m+2+\alpha)_n t^{m+1+\alpha}(1-t)^{1-\alpha}\,{}_2F_1(-n, n+m+3;\ m+2+\alpha;\ t),$$

$${}_2F_1(-n, n+m+3;\ m+2+\alpha;\ t) = (-1)^n\frac{\Gamma(n+2-\alpha)}{\Gamma(2-\alpha)(m+2+\alpha)_n}\times$$

$$\times\,{}_2F_1(-n, n+m+3;\ 2-\alpha;\ 1-t),$$

kapjuk

(5.24)
$$I_2 =$$

$$\frac{1}{\Gamma(1-\alpha)p^{1+\alpha}}\left(1-\frac{1}{p}\right)^{m+1}\int_0^1 t^{m+1+\alpha}(1-t)^{1-\alpha}\,{}_2F_1(-n, n+m+3;\ 2-\alpha;\ 1-t)\times$$

$$\times\frac{dt}{1-t\left(1-\frac{1}{p}\right)}.$$

Tehát

(5.25)
$$I_1 + I_2 =$$

$$= \frac{1}{\Gamma(1-\alpha)p^{1+\alpha}}\sum_{j=0}^{n}\frac{(-n)_j(n+m+3)_j}{j!\,(2-\alpha)_j}\int_0^1 u^{\alpha}(1-u)^{-\alpha}\times$$

$$\times\frac{\left[\left(1-\frac{1}{p}\right)^{m+1}u^{m+1}(1-u)^{j+1} - \frac{(-u)^{j+1}}{p^{j+1}}\right]}{1-u\left(1-\frac{1}{p}\right)}du.$$

A szögletes zárójel osztása elvégezhető

(5.26)
$$\frac{1}{1-u\left(1-\frac{1}{p}\right)}[\ ] = \sum_{l=0}^{j} u^l(1-u)^{j-l}\left(-\frac{1}{p}\right)^l - \sum_{l=0}^{m} u^l(1-u)^{j+1}\left(1-\frac{1}{p}\right)^l,$$

ezért

(5.27)
$$I_1+I_2 =$$

$$= \frac{1}{\Gamma(1-\alpha)p^{1+\alpha}} \sum_{j=0}^{n} \frac{(-n)_j(n+m+3)_j}{j!(2-\alpha)_j}\left\{\sum_{l=0}^{j}\left(-\frac{1}{p}\right)^l \frac{\Gamma(l+1+\alpha)\Gamma(j-l+1-\alpha)}{\Gamma(j+2)} - \right.$$

$$\left. - \sum_{l=0}^{m}\left(1-\frac{1}{p}\right)^l \frac{\Gamma(l+1+\alpha)\Gamma(j+2-\alpha)}{\Gamma(l+j+3)}\right\}.$$

A második összegben a j szerinti összegzés könnyen elvégezhető az

(5.28)
$${}_2F_1(-n, n+m+3;\ l+3;\ 1) = (-1)^n \frac{(1+m-l)_n}{(l+3)_n},$$

képlet segítségével, és így azt kapjuk, hogy

(5.29)
$$\int_0^\infty x^\alpha P_m(x)e^{-px}dx =$$

$$= \frac{1}{\Gamma(1-\alpha)p^{1+\alpha}} \sum_{j=0}^{n} \frac{(-n)_j(n+m+3)_j}{j!(2)_j(2-\alpha)_j} \sum_{l=0}^{j}\left(-\frac{1}{p}\right)^l \Gamma(l+1+\alpha)\Gamma(j-l+1-\alpha) -$$

$$-(1-\alpha)(-1)^n \frac{\Gamma(n+m+1)}{\Gamma(m+1)\Gamma(n+3)} \sum_{l=0}^{m} \frac{(-m)_l\Gamma(1+\alpha+l)}{(n+3)_l(-n-m)_l} \frac{1}{p^{1+\alpha}}\left(1-\frac{1}{p}\right)^l,$$

amiből inverz transzformációval pontosan az (5.5) képlet adódik.

Kiszámítottuk $n=m$, $n=1, 2, 3, 4, 5$ esetekre $\alpha=0$ paraméterrel (exponenciális integrál) és az $\alpha=\frac{1}{2}$ paraméterrel (hibaintegrál, erfc) a közelítő racionális törteket és az együtthatókat a II. függelékben táblázatosan adjuk meg.

6. Diagonális törtek konvergenciája

Megmutatjuk, hogy a diagonális törtek $(m=n)$ $x>0$ esetén geometriai sebességgel konvergálnak. Az egyszerűség kedvéért a levezetést az $x^{1/2}$ függvényre mutatjuk be. A függvény *Clenshaw—Lord közelítése*

(6.1)

$$\frac{P_n(x)}{Q_n(x)} = \frac{2}{\pi}\frac{1}{(n+1)(2n+1)}\frac{{}_4F_3\left(-n,\ -n-\frac{1}{2},\ n+1,\ n+\frac{3}{2};\ \frac{1}{2},\ \frac{1}{2},\ \frac{3}{2};\ x\right)}{{}_4F_3\left(-n,\ -n+\frac{1}{2},\ n+2,\ n+\frac{3}{2};\ 1,\ \frac{3}{2}\ 2;\ x\right)},$$

hibája viszont

(6.2)

$$H_n(x) = \frac{S_n(x)}{Q_n(x)} = x^{1/2} - \frac{P_n(x)}{Q_n(x)} = \frac{\sum_{j=0}^{2n+1} \frac{(-2n-1)_j(2n+2)_j}{j!\,\Gamma\left(\frac{j}{2}+\frac{1}{2}\right)\Gamma\left(\frac{j}{2}+\frac{3}{2}\right)}\left(\frac{x}{4}\right)^{j/2}}{\frac{1}{2}\sum_{j=0}^{n}\frac{(-2n-1)_{2j+1}(2n+2)_{2j+2}}{(2j+1)!\,\Gamma(j+1)\Gamma(j+2)}\left(\frac{x}{4}\right)^{j}}.$$

A hibára integrálelőállítást fogunk adni. Alkalmazzuk ugyanis a

$$\sqrt{\pi}\,\Gamma(j+1) = 2^{-j}\Gamma\left(j/2+\frac{1}{2}\right)\Gamma(j/2+1)$$

relációt a számlálóra, kapjuk

(6.3)
$$\sum_{j=0}^{2n+1}\frac{(-2n-1)_j(2n+2)_j}{j!\,\Gamma\left(\frac{j}{2}+\frac{1}{2}\right)\Gamma\left(\frac{j}{2}+\frac{3}{2}\right)}(x/4)^{j/2} =$$

$$= \frac{1}{\sqrt{\pi}}\sum_{j=0}^{2n+1}\frac{(-2n-1)_j(2n+2)_j}{j!\,j!}\frac{\Gamma\left(\frac{j}{2}+1\right)}{\Gamma\left(\frac{j}{2}+\frac{3}{2}\right)}x^{j/2}.$$

Az összegben a gamma függvények hányadosára a beta integrált alkalmazva nyerjük a kívánt előállítást

(6.4)
$$\frac{1}{\pi}\int_0^1 (1-u)^{-1/2}\sum_{j=0}^{2n+1}\frac{(-2n-1)_j(2n+2)_j}{j!^2}(xu)^{j/2}\,du =$$

$$= \frac{2}{\pi\sqrt{x}}\int_0^{\sqrt{x}}\frac{s}{(x-s^2)^{1/2}}\sum_{j=0}^{2n+1}\frac{(-2n-1)_j(2n+2)_j}{j!^2}s^j\,ds.$$

A nevezőre teljesen hasonló gondolatmenettel kapjuk a következő előállítást

$$\frac{2}{\pi\sqrt{x}}\int_0^{\sqrt{x}}\frac{1}{(x-s^2)^{1/2}}\sum_{j=0}^{n}\frac{(-2n-1)_{2j+1}(2n+2)_{2j+1}}{(2j+1)!^2}s^{2j+1}\,ds. \tag{6.5}$$

Használjuk fel a (6.4) és (6.5)-ben szereplő összegekre az alábbi relációkat

(6.6)

$$\sum_{j=0}^{2n+1}\frac{(-2n-1)_j(2n+2)_j}{j!^2}s^j=\frac{1}{\Gamma(2n+2)}\frac{d^{2n+1}}{ds^{2n+1}}[s^{2n+1}(1-s)^{2n+1}]=P_{2n+1}^*(s),$$

$$\sum_{j=0}^{n}\frac{(-2n-1)_{2j+1}(2n+2)_{2j+1}}{(2j+1)!^2}s^{2j+1}= \tag{6.7}$$

$$=\frac{1}{2}\frac{1}{\Gamma(2n+2)}\frac{d^{2n+1}}{ds^{2n+1}}[s^{2n+1}\{(1-s)^{2n+1}-(1+s)^{2n+1}\}].$$

A (6.6) relációban szereplő $P_k^*(s)$ polinom a (0, 1) intervallumban *ortogonális Legendre polinom*, amelyre érvényes

$$|P_k^*(s)|\leqq 1,\quad 0\leqq s\leqq 1. \tag{6.8}$$

Tekintsük most az

$$I=\frac{2}{\pi\sqrt{x}}\int_0^{\sqrt{x}}\frac{1}{(x-s^2)^{1/2}}\frac{1}{\Gamma(2n+2)}\frac{d^{2n+1}}{ds^{2n+1}}[s^{2n+1}(1+s)^{2n+1}]\,ds \tag{6.9}$$

integrált. Célunk aszimptotikus kifejezésének meghatározása. Alakítsuk át I-t a *Cauchy integrál előállítás* segítségével

$$\frac{2}{\pi\sqrt{x}}\int_0^{\sqrt{x}}\frac{1}{(x-s^2)^{1/2}}\frac{1}{\Gamma(2n+2)}\frac{d^{2n+1}}{ds^{2n+1}}\frac{1}{2\pi i}\oint\frac{u^{2n+1}(1+u)^{2n+1}}{u-s}\,du\,ds= \tag{6.10}$$

$$=\frac{2}{\pi\sqrt{x}}\int_0^{\sqrt{x}}\frac{1}{(x-s^2)^{1/2}}\frac{1}{2\pi i}\oint\frac{u^{2n+1}(1+u)^{2n+1}}{(u-s)^{2n+2}}\,du\,ds=$$

$$=\frac{2}{\pi x}\int_0^{1}\frac{1}{(1-t^2)^{1/2}}\frac{1}{2\pi i}\oint\frac{u^{2n+1}(1+u)^{2n+1}}{(u-t\sqrt{x})^{2n+2}}\,du\,ds.$$

Ez az integrál a *Laplace-módszer* alapján kiértékelhető. A belső integrál lényeges járulékát az $u_0=\sqrt{t^2x+t\sqrt{x}}+t\sqrt{x}$ pont környezete adja minden $0\leqq t\leqq 1$ és $0<x\leqq 1$-re. Elemi számítással kapjuk

$$I=O(q^n),\quad n\to\infty,\quad q=\left(\sqrt{1+\sqrt{x}}+x^{1/4}\right)^4. \tag{6.11}$$

Vegyük figyelembe (6.8)-t és így végül is a kívánt hibabecsléshez jutunk

$$\lim_{n\to\infty}|H_n(x)|^{\frac{1}{n}}\leqq\frac{1}{\left(\sqrt{1+\sqrt{x}}+x^{1/4}\right)^4},\quad 0\leqq x\leqq 1. \tag{6.12}$$

Általánosított Padé közelítés esetén hasonló a helyzet. A közelítés

(6.13)

$$\frac{A_n(x)}{B_n(x)} = \frac{1}{\sqrt{\pi}\cdot(2n+1)} \frac{\Gamma(2n+1)}{\Gamma\left(2n+\frac{3}{2}\right)} \frac{{}_3F_2\left(-n, -n-\frac{1}{2}, 2n+1; \frac{1}{2}, \frac{1}{2}; x\right)}{{}_3F_2\left(-n, -n+\frac{1}{2}, 2n+\frac{3}{2}; 1, \frac{3}{2}, x\right)},$$

hibáját jelöljük $K_n(x)$-szel

(6.14) $$K_n(x) = x^{1/2} - \frac{A_n(x)}{B_n(x)}, \quad 0 \leqq x \leqq 1.$$

Az előzőkhöz hasonlóan nyerhetünk integrálelőállításokat

$$K_n(x) = \frac{\sum_{j=0}^{2n+1} \frac{(-2n-1)_j \Gamma(2n+1+j/2)}{j!\,\Gamma(j/2+1/2)} x^{j/2}}{\sum_{j=0}^{n} \frac{(-2n-1)_{2j+1} \Gamma\left(2n+\frac{3}{2}+j\right)}{(2j+1)!\,\Gamma(j+1)} x^j} =$$

(6 15)

$$= \frac{1 + \frac{x^{1/2}}{\Gamma(2n+1)} \int_0^x \frac{1}{(x-u)^{1/2}} \frac{d^{2n+1}}{du^{2n+1}} [u^{2n}(1-u^{1/2})^{2n+1}]\, du}{\frac{1}{2\Gamma(2n+1)} \int_0^x \frac{1}{(x-u)^{1/2}} \frac{d^{2n+1}}{du^{2n+1}} [u^{2n}\{(1-u^{1/2})^{2n+1} - (1+u^{1/2})^{2n+1}\}]\, du}.$$

A *Cauchy integrál* alkalmazása után az alábbi típusú integrálok aszimptotikus vizsgálatára van szükség

(6.16) $$I^{(\pm)} = (2n+1)x^{1/2} \int_0^1 \frac{1}{(1-u)^{1/2}} \frac{1}{2\pi i} \oint \frac{t^{2n}(1 \pm t^{1/2})^{2n+1}}{(t-xu)^{2n+2}}\, dt\, du.$$

A *Laplace-módszer* szerint eljárva könnyen láthatjuk, hogy a legnagyobb járulék azon pont környezetéből adódik az integrálban, amelyre

(6.17) $$t^{3/2} - 3xut^{1/2} \mp 2xu = 0.$$

A u szerinti integrálásnál az $u=1$ pont körül van a lényeges járulék. Elvégezve néhány elemi számítást, kapjuk

(6.18) $$I^{(+)} = O(Q^n), \quad n \to \infty,$$

(6.19) $$I^{(-)} = O(P^n), \quad n \to \infty,$$

ahol

$$Q = \left[1 + \frac{3}{2}(p+q)\right]^2 \qquad p = [x(1+\sqrt{1-x})]^{1/3},$$

(6.20)

$$P = \frac{1}{2} - \frac{9}{4}x + \left(\frac{1}{2} - \frac{3}{2}q\right)^2 + \left(\frac{1}{2} - \frac{3}{2}q\right)^2, \quad q = [x(1-\sqrt{1-x})]^{1/3}.$$

A (6.18)—(6.19) képletek alapján a keresett hibabecslés

$$\lim_{n\to\infty} |K_n(x)|^{1/n} \leqq \lim_{n\to\infty} \left(\frac{I^{(-)}}{I^{(+)}}\right)^{1/n} = \frac{P}{Q}, \quad 0 \leqq x \leqq 1. \tag{6.21}$$

Megjegyzések. A hibafüggvény x-ben ténylegesen hullámzó sajátságú. Előjelváltásainak száma pontosan $2n+1$. Ez azonos a legjobb közelítés előjelváltásainak a számával, de a maximális (abszolút értékben) eltérések nagysága távolról sem azonos. A bemutatott előállítások alapján az abszolút értékekre adott (6.12) és (6.21) becslések, a hiba aszimptotikus kifejezésévé fejleszthetők, lényegesen hosszadalmasabb számítással és bonyolultabb képletekkel. Természetesen a módszer alkalmazható az x^{μ} függvényre más kitevővel is, mint $\mu=\frac{1}{2}$, sőt a részletek közlése nélkül állítjuk, hogy lényegében ugyan ez az eredmény érvényes az $\ln x$ és $|x|^{\varrho}\,\mathrm{sgn}\,(x)$ függvények esetében is. Nem diagonális közelítésnél ($n\neq m$) bizonyos esetekben elvész a geometriai konvergencia. Így pl. $m>0$, $n=0$ esetén (polinom közelítés), valamint $m=0$, $n>0$ esetén (reciprok polinom közelítés) mint ismeretes, nem áll fenn a geometriai konvergencia. Az átmenet nyitott kérdés, ezzel egy másik cikkben kívánunk foglalkozni.

A következőkben az (5.1) alatti függvény (5.3) közelítésének hibabecslésével foglalkozunk. Először az (5.4) nevező aszimptotikus kifejtését határozzuk meg $n\to\infty$ esetére. A számításhoz segítségül *Laplace transzformációt* alkalmazunk

$$L_n(q) = \int_0^{\infty} x^{1-\alpha} Q_n(x) e^{-qx}\, dx. \tag{6.21}$$

Azonnal adódik (5.4)-ből, hogy

$$L_n(q) = \frac{\Gamma(2-\alpha)}{q^{2-\alpha}}\, {}_2F_1\left(-n, 2n+3;\ 2;\ -\frac{1}{q}\right). \tag{6.22}$$

Integrálalakot nyerhetünk $L_n(q)$-ra felhasználva az alábbi képletet

$${}_2F_1(-n, 2n+3;\ 2;\ -z) = \frac{1}{(2)_n} z^{-1}(1+z)^{-n-1} \frac{d^n}{dz^n}\{z^{n+1}(1+z)^{2n+1}\}, \tag{6.23}$$

és a *Cauchy integrálformulát*

$$L_n(q) = \frac{\Gamma(2-\alpha)}{n+1} \frac{q^{2n+\alpha+1}}{(q+1)^{n+2}} \frac{1}{2\pi i} \oint \frac{s^{n+1}(1+s)^{2n+1}}{(sq-1)^{n+1}}\, ds. \tag{6.24}$$

Alkalmazva a *Laplace-módszert* $L_n(q)$, $n\to\infty$ aszimptotikus alakját meghatározhatjuk a (6.24) képletből. Standard módon eljárva, kapjuk

$$L_n(q) = K^n O(n^{\alpha}), \quad n\to\infty, \quad K = K(q) = \frac{4q+9+3\sqrt{8q+9}}{4q+3-\sqrt{8q+9}}. \tag{6.25}$$

Továbbá felhasználva a *Mellin inverziós formulát* $Q_n(x)$ az alábbi lesz

$$Q_n(x) = \frac{1}{2\pi i x^{1-\alpha}} \int_{\sigma-i\infty}^{\sigma+i\infty} e^{qx} L_n(q)\, dq. \tag{6.26}$$

Az $x=\alpha n$ jelölést bevezetve a (6.26) integrálból $Q_n(x)$ aszimptotikus kifejezése könnyen megkapható, amiből az alábbi határérték adódik

$$\lim_{n\to\infty} \{Q_n(x)\}^{1/n} = e^{\alpha q} \frac{\left(1+\frac{1}{2}\alpha q\right)^2}{1-\alpha q}, \tag{6.27}$$

ahol q a

$$q^3+q^2+\frac{1}{\alpha}q-\frac{2}{\alpha^2}=0 \tag{6.28}$$

egyenlet pozitív megoldása.

A számláló aszimptotikus vizsgálata kissé bonyolultabb feladat. Az (5.19) képletet fogjuk felhasználni. Ismét egy *Laplace transzformáltat* alkalmazunk. Legyen

$$Z_n(p) = \int_0^\infty x^\alpha S_n(x) e^{-px}\, dx. \tag{6.29}$$

Az (5.19) sorban tagonként elvégezve a transzformációt, valamint használva az ${}_2F_1$ hipergeometriai függvényt integrálelőállítását kapjuk

$$Z_n(p) = (\alpha-1)\frac{\Gamma(n+1)\Gamma(2n+2+\alpha)}{\Gamma(3n+4)p^{2n+2+\alpha}}(p-1)^{2n+1}\,{}_2F_1\left(n+1, 2n+2+\alpha; 3n+4; 1-\frac{1}{p}\right) = (\alpha-1)\frac{\Gamma(n+1)}{\Gamma(n+2-\alpha)}\frac{(p-1)^{2n+1}}{p^{2n+2+\alpha}}\int_0^1 t^{2n+1+\alpha}(1-t)^{n+1-\alpha}\times$$

$$\times\left[1-t\left(1-\frac{1}{p}\right)\right]^{-n-1} dt. \tag{6.30}$$

A *Laplace-módszer* segítségével (6.30)-ból megkapható $Z_n(p)$, $n\to\infty$ aszimptotikus alakja:

$$Z_n(p) = L^n O(n^b), \quad n\to\infty$$

$$L = L(p) = \frac{4p+5-3\sqrt{8p+1}}{4p-1+\sqrt{8p+1}}. \tag{6.31}$$

A (6.25), illetve (6.31)-ben szereplő „a", ill. „b"n-től független állandók. Felhasználva az inverziós formulát $S_n(x)$-re érvényes

$$S_n(x) = \frac{1}{2\pi i x^\alpha} \int_{\sigma-i\infty}^{\tau+i\infty} e^{px} Z_n(p)\, dp. \tag{6.32}$$

A (6.31) képlet felhasználásával (6.32)-ből levezethető a hiba egy aszimptotikus becslése

$$\lim_{n\to\infty}\left|\frac{S_n(x)}{Q_n(x)}\right|^{1/n} \leqq H(\alpha), \quad \alpha=\lim_{n\to\infty}\frac{x}{n},$$

$$H(\alpha)=\begin{cases} e^{\alpha/2(1-3q)}\dfrac{(1-\alpha q)^{3/2}}{\left(1+\dfrac{1}{2}\alpha q\right)^3}, & 0<\alpha\leqq\alpha_0, \\ e^{\alpha(q_2-q_1+1)}\left[\dfrac{1+\dfrac{1}{2}\alpha q_2}{1+\dfrac{1}{2}\alpha q_1}\right]^2\dfrac{1-\alpha q_1}{1-\alpha q_2}, & \alpha_0\leqq\alpha<\infty \end{cases} \tag{6.33}$$

ahol

$$q, \quad 0<\alpha<\alpha_0\text{-ra}-\alpha_0=\frac{2}{9}(7\sqrt{7}+10)=6{,}337835373-a$$

$$q^3+q^2+\frac{1}{\alpha}q-\frac{2}{\alpha^2}=0, \tag{6.34}$$

egyenlet pozitív megoldása, továbbá q_1, ill. q_2 $\alpha_0\leqq\alpha<\infty$-re (6.34) azon megoldása, amelyre

$$q_1=\frac{1}{\alpha}+O\left(\frac{1}{\alpha^2}\right), \quad \alpha\to\infty, \tag{6.35}$$

$$q_2=-1+\frac{1}{\alpha}+O\left(\frac{1}{\alpha^2}\right), \quad \alpha\to\infty.$$

A $H(\alpha)$ függvény ((6.33)-mal implicit megadva) az alábbi tulajdonságokkal rendelkezik: $H(0)=1$, $0<H(\alpha)<1$, $\alpha>0$; $H(\alpha)$ $0<\alpha\leqq\alpha_1$-ben $-\alpha_1=\frac{9}{8}\left(\sqrt{\frac{35}{3}}-1\right)=$ $=2{,}717\,606\,538\ldots$ — monoton csökkenő, $\min H(\alpha)=H(\alpha_1)=1/16(117{,}15^{3/2}-$ $-367{,}7^{3/2})=0{,}009\,415\,904\,75\sim 1/106$, azon túl $\alpha\geqq\alpha_1$ esetén monoton növő és $H(\infty)=1/27$. Ebből következik, hogy az (5.4)—(5.5)-tel adott racionális tört geometriai sebességgel konvergál az (5.1) függvényhez az $x=0$ kis környezetének kivételével. Az $x=0$ helyen a hiba pontos értéke meghatározható. Az (5.19) képletből kapjuk

$$S_n(0)=(\alpha-1)\frac{n!\,\Gamma(n+m+2)}{\Gamma(2n+m+4)}\sum_{j=0}^{\infty}\frac{(n+1)_j(n+m+2)_j}{j!(2n+m+4)_j}L_{n+m+j+1}^{(\alpha)}(0)= \tag{6.36}$$

$$=\frac{\alpha-1}{\Gamma(\alpha+1)}\cdot\frac{n!\,\Gamma(n+m+\alpha+2)}{\Gamma(2n+m+4)}\sum_{j=0}^{\infty}\frac{(n+1)_j(n+m+2+\alpha)_j}{j!(2n+m+4)_j}.$$

Az utóbbi sor ${}_2F_1$ függvény egységnyi argumentummal és így gamma függvényekkel

felösszegezhető

(6.37)

$$S_n(0) = (\alpha-1)\frac{n!\,\Gamma(n+m+2+\alpha)}{\Gamma(\alpha+1)\Gamma(2n+m+4)}\,{}_2F_1(n+1, n+m+\alpha+2;\ 2n+m+4;\ 1) =$$

$$= -\frac{\Gamma(2-\alpha)}{\Gamma(1+\alpha)}\,\frac{\Gamma(n+1)}{\Gamma(n+2-\alpha)}\,\frac{\Gamma(n+m+2+\alpha)}{\Gamma(n+m+3)}.$$

Ezért $n=m$ esetén

(6.38) $$S_n(0) = O(n^{2\alpha-2}), \quad n \to \infty,$$

és így (nem geometriai) konvergencia $\alpha<1$ esetén az $x=0$ helyen is fennáll.

IRODALOM

[1] CHENEY, E. W., *Introduction to Approximation Theory* (McGraw-Hill, New York, 1966).
[2] CLENSHAW, C. W., LORD, K., "Rational Approximations from Chebyshev Series", In: *Studies in Numerical Analysis*, ed. *B. K. P. Scaife*, Academic Press, London (1974) 95—113.
[3] DOUGALL, J., "On Vandermonde's theorem and some more general expansions", *Proc. Edinburgh Math. Soc.* **25** (1907) 114—132.
[4] ERDÉLYI, A., MAGNUS, W., OBERHETTINGER, F. and TRICOMI, F. G., *Higher Transcendental Functions, Vols. I., II.* (McGraw-Hill, New York, 1953).
[5] HART, J. F., *Computer Approximations* (John Wiley, New York, 1969).
[6] LUKE, Y. L., *The Special Functions and Their Approximations Vols. I., II.* (Academic Press, New York, 1969).
[7] NÉMETH, G., PÁRIS, G., "The Gibbs phenomenon in generalized Padé Approximation", *Journal of Math. Phys.* **26** (1985) 1175—1178.
[8] NÉMETH, G., "The Gibbs phenomenon in generalized Padé Approximation", In: *Constructive Theory of Functions*, ed. *Bl. Sendov*, Publishing House of the Bulgarian Academy of Sciences, Sofia (1984) 634—639.
[9] PASZKOWSKI, S., *Zastosowania numeryczne wielomianow i szeregow Czebyszeva* (Panstwowe Wydawnictwo Naukowe, Warszawa, 1975).

(Beérkezett: 1986. június 13.)

NÉMETH GÉZA
MTA KÖZPONTI FIZIKAI KUTATÓ INTÉZET
1121 BUDAPEST XII., KONKOLY THEGE ÚT 29—33.

GENERALIZED PADÉ APPROXIMATION TO SPECIAL FUNCTIONS SINGULAR FUNCTIONS

G. NÉMETH

Rational approximants for some special functions are determined in sense of *generalized Padé approximation*. There are considered functions having singularity in the range of the approximation, however the main diagonal approximants converge to the function with geometric rate of convergence outside a small neighbourhood of this singularity.

I. Függelék

A (2.17) probléma megoldásánál kihasználtuk, hogy $m \geqq n$. De ugyanakkor könnyen belátható, hogy ezt a feltételt nem teljesítő $m=0$, $\mu=1$ esetben is helyesen adja (2.17) a *Clenshaw—Lord közelítést*. Számítsuk ki ugyanis az

$$1-x = \frac{1}{2}T_0^*(x) - \frac{1}{2}T_1^*(x) = P_0 / \sum_{r=0}^{n}{}^{*} q_r T_r^*(x) + O\left(T_{n+1}^*(x)\right)$$

probléma megoldását szolgáltató q_r, $r=0, 1, \ldots, n$ számokat. Ismeretes [2], hogy a

$$\frac{P_0}{\sum_{r=0}^{n}{}^{*} q_r T_r^*(x)} = \sum_{j=0}^{\infty}{}^{*} A_j T_j^*(x)$$

előállítás A_j együtthatójára fennáll a

$$\sum_{j=0}^{n}{}^{*} \gamma_j A_{r-j} = 0, \quad r = 1, 2, \ldots, n,$$

egyenletrendszer és

$$q_r = \mu \sum_{i=0}^{n-r}{}^{*} \gamma_i \gamma_{i+r}, \quad \mu = \frac{1}{2} \sum_{i=0}^{n}{}^{*} \gamma_i^2.$$

Az egyenletrendszert részletesen kiírva kapjuk

$$\begin{aligned}
-\frac{1}{2}\gamma_0 + \gamma_1 - \frac{1}{2}\gamma_2 &= 0, \\
-\frac{1}{2}\gamma_1 + \gamma_2 - \frac{1}{2}\gamma_3 &= 0, \\
&\cdots \\
-\frac{1}{2}\gamma_{n-2} + \gamma_{n-1} - \frac{1}{2}\gamma_n &= 0, \\
-\frac{1}{2}\gamma_{n-1} + \gamma_n &= 0.
\end{aligned}$$

Ennek az egyenletrendszernek megoldása — $\gamma_0=1$ feltevéssel —

$$\gamma_k = \frac{n+1-k}{n+1}.$$

Ebből

$$q_r = \frac{2(2n+3-3r)}{2n+3} + \frac{2r(r^2-1)}{(n+1)(n+2)(2n+3)}, \quad q_0 = 2, \quad p_0 = \frac{3}{4n+6}.$$

A közelítés nevezőjében a *Csebisev polinomok* szerinti összeget átírhatjuk x hatványai szerint haladó kifejezésbe. Ez az alábbi

$$\frac{1}{3}(4n+6)\cdot \sum_{r=0}^{n}{}^{*}(-1)^r q_r T_r^*(x) = (n+1)(n+2)_3F_2\left(-n, n+3, 1;\ \frac{3}{2}, 3;\ 1-x\right).$$

Ez utóbbi kifejezés azonos a (2.17)-ből $m=0$, $\mu=1$ helyettesítéssel nyerhető alakkal. Hasonlóan számolható végig néhány más eset is, $\mu=k$ egész, stb. A teljes bizonyítás még nem ismert.

II. Függelék

$$F(x) = x\int_0^{\infty} \frac{1}{1+t} e^{-xt}\,dt$$

általánosított Padé közelítései.

$$F(x) \sim \frac{P_n(x)}{Q_n(x)}.$$

$$P_n(x) = \sum_{j=0}^{n} \frac{(-n)_j (2n+3)_j}{j!(2)_j(2)_j} \sum_{l=0}^{j} (-x)^l (j-l)! -$$

$$-(-1)^n \frac{(2n)!}{n!(n+2)!} \sum_{j=0}^{n} \frac{(-n)_j\, j!}{(n+3)_j(-2n)_j} L_j(x),$$

$$Q_n(x) = {}_2F_2(-n, 2n+3;\ 2, 2;\ -x),$$

$$P_1(x) = \frac{1}{8} + \frac{29}{24}x, \quad Q_5(x) = 1 + \frac{65}{4}x + \frac{455}{9}x^2 + \frac{2275}{48}x^3 +$$

$$+\frac{91}{6}x^4 + \frac{1547}{1080}x^5,$$

$$P_2(x) = \frac{1}{18} + \frac{361}{180}x + \frac{559}{360}x^2, \quad Q_4(x) = 1 + 11x + 22x^2 +$$

$$+\frac{143}{12}x^3 + \frac{1001}{600}x^4,$$

$$P_3(x) = \frac{1}{32} + \frac{8677}{3360}x + \frac{38\,891}{6720}x^2 +$$

$$+\frac{11\,549}{6720}x^3, \quad Q_3(x) = 1 + \frac{27}{4}x + \frac{15}{2}x^2 + \frac{55}{32}x^3,$$

$$P_4(x) = \frac{1}{50} + \frac{38\,207}{12\,600}x + \frac{7039}{525}x^2 +$$

$$+ \frac{258\,269}{25\,200}x^3 + \frac{252\,251}{151\,200}x^4, \quad Q_2(x) = 1 + \frac{7}{2}x + \frac{14}{9}x^2,$$

$$P_5(x) = \frac{1}{72} + \frac{1571}{462}x + \frac{552\,733}{22\,176}x^2 + \frac{7\,003\,991}{199\,584}x^3 +$$

$$+ \frac{10\,964\,573}{798\,336}x^4 + \frac{5\,717\,711}{3\,991\,680}x^5, \quad Q_1(x) = 1 + \frac{5}{4}x,$$

$$G(x) = \frac{x^{1/2}}{\sqrt{\pi}} \int_0^\infty \frac{t^{-1/2}}{1+t} e^{-xt}\, dt$$

általánosított Padé közelítései.

$$G(x) \sim \frac{P_n^*(x)}{Q_n^*(x)},$$

$$P_n^*(x) = \sum_{j=0}^{n} \frac{(-n)_j (2n+3)_j}{j!(3/2)_j (2)_j} \sum_{l=0}^{j} (-x)^l \left(\frac{1}{2}\right)_{j-l} -$$

$$- \frac{(-1)^n}{2} \frac{(2n)!}{n!(n+2)!} \sum_{j=0}^{n} \frac{(-n)_j\, j!}{(n+3)_j (-2n)_j} L_j^{(1/2)}(x),$$

$$Q_n^*(x) = {}_2F_2\left(-n, 2n+3;\ 2, \frac{3}{2};\ -x\right).$$

$$P_1^*(x) = \frac{35}{96} + \frac{79}{48}x, \quad Q_5^*(x) = 1 + \frac{65}{3}x + \frac{728}{9}x^2 + \frac{260}{3}x^3 +$$

$$+ \frac{832}{27}x^4 + \frac{14\,144}{4455}x^5,$$

$$P_2^*(x) = \frac{77}{320} + \frac{829}{240}x + \frac{199}{80}x^2, \quad Q_4^*(x) = 1 + \frac{44}{3}x + \frac{176}{5}x^2 +$$

$$+ \frac{2288}{105}x^3 + \frac{16\,016}{4725}x^4,$$

$$P_3^*(x) = \frac{1287}{7168} + \frac{19\,001}{3584}x + \frac{56\,081}{5376}x^2 + \frac{42\,239}{13\,440}x^3, \quad Q_3^*(x) = 1 + 9x +$$

$$+ 12x^2 + \frac{22}{7}x^3,$$

$$P_4^*(x) = \frac{46\,189}{322\,560} + \frac{36\,109}{5040}x + \frac{5\,411\,039}{201\,600}x^2 +$$

$$+ \frac{1\,519\,249}{75\,600}x^3 + \frac{1\,025\,023}{302\,400}x^4, \quad Q_2^*(x) = 1 + \frac{14}{3}x + \frac{112}{45}x^2,$$

$$P_5^*(x) = \frac{96\,577}{811\,008} + \frac{10\,991\,179}{1\,216\,512}x + \frac{16\,644\,217}{304\,128}x^2 + \frac{18\,663\,287}{253\,440}x^3 +$$

$$+ \frac{466\,665\,671}{15\,966\,720}x^4 + \frac{25\,346\,047}{7\,983\,360}x^5, \quad Q_1^*(x) = 1 + \frac{5}{3}x.$$

Alkalmazott Matematikai Lapok **13** (1987—88) 127—135

KVADRATIKUS JÁTÉKOK STABILITÁSÁRÓL

SZIDAROVSZKY FERENC, OKUGUCHI KOJI

Budapest Tokió

Kvadratikus függvényekkel rendelkező n-személyes játékok stabilitására adunk feltételeket. Két modellt vizsgálunk. A diszkrét idejű modell esetén differenciaegyenletek, a folytonos idejű modell esetén pedig differenciálegyenletek stabilitására redukáljuk a játékok stabilitását. Konkrét alkalmazásként az oligopol problémát vizsgáljuk meg.

1. Bevezetés

Játékok stabilitásának vizsgálata az újabb játékelméleti kutatások egyik legfontosabb témakörét jelenti. Konkrét játékok stabilitásának feltételével számos szerző foglalkozott. SZÉP és FORGÓ (1985), THEOCHARIS (1959), SZIDAROVSZKY és OKUGUCHI (1985, 1986), OKUGUCHI és SZIDAROVSZKY (1986) különféle feltételeket ismertetnek konkrét játékok stabilitására.

Ebben a dolgozatunkban általános kvadratikus játékok stabilitásával foglalkozunk. Az általános eredmények bemutatásán kívül azok alkalmazását is megvizsgáljuk több speciális játék esetére.

Tekintsük a következő kvadratikus játékot: $\Gamma=\{n; S_1, \ldots, S_n; \varphi_1, \ldots, \varphi_n\}$, ahol n a játékosok száma, az $S_1, \ldots, S_n$ stratégiahalmazok véges dimenziós euklideszi terek korlátos, konvex, zárt részhalmazai, a $\varphi_1, \ldots, \varphi_n$ kifizetőfüggvények pedig a szimultán stratégiavektor kvadratikus függvényei:

$$\varphi_k(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \mathbf{x}^T \begin{pmatrix} \mathbf{A}_{11}^{(k)} & \ldots & \mathbf{A}_{1n}^{(k)} \\ \vdots & & \vdots \\ \mathbf{A}_{n1}^{(k)} & \ldots & \mathbf{A}_{nn}^{(k)} \end{pmatrix} \mathbf{x} + \mathbf{b}^{(k)T}\mathbf{x} + c^{(k)}, \tag{1.1}$$

ahol

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}, \qquad \mathbf{b}^{(k)} = \begin{pmatrix} \mathbf{b}_1^{(k)} \\ \mathbf{b}_2^{(k)} \\ \vdots \\ \mathbf{b}_n^{(k)} \end{pmatrix}, \tag{1.2}$$

$$\mathbf{x}_k \in S_k \qquad (k = 1, 2, \ldots, n).$$

Feltesszük, hogy az $\mathbf{A}_{ij}^{(k)}$ és $\mathbf{b}_i^{(k)}$ blokkok mérete a blokkonként elvégzendő mátrixműveleteknek megfelelő.

Tegyük fel, hogy $k=1, 2, \ldots, n$ esetén az $\mathbf{A}_{k,k}^{(k)}+\mathbf{A}_{k,k}^{(k)T}$ mátrixok negatív szemidefinitek. Ekkor φ_k konkáv $\mathbf{x}_k$-ban, így a *Nikaido—Isoda-tétel* (l. SZÉP és FORGÓ, 1985) értelmében a játéknak létezik egyensúlypontja.

Az egyensúlypontok stabilitásának vizsgálata érdekében tekintsük először a

$$\varphi_k(\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}, \mathbf{y}_k, \mathbf{x}_{k+1}, \ldots, \mathbf{x}_n) \tag{1.3}$$

függvényt, ahol az $\mathbf{x}_i$ $(i \neq k)$ stratégiákat rögzítjük, és csak az $\mathbf{y}_k$ stratégiavektort tekintjük változónak. Egyszerű számolással látható, hogy alakja:

$$\mathbf{y}_k^T \mathbf{A}_{kk}^{(k)} \mathbf{y}_k + \mathbf{y}_k^T \Big(\sum_{i \neq k} \mathbf{A}_{ki}^{(k)} \mathbf{x}_i\Big) + \Big(\sum_{i \neq k} \mathbf{x}_i^T \mathbf{A}_{ik}^{(k)}\Big) \mathbf{y}_k + \mathbf{b}_k^{(k)T} \mathbf{y}_k + \alpha^{(k)} \tag{1.4}$$

ahol $\alpha^{(k)}$ független az $\mathbf{y}_k$ vektortól.

Feltéve, hogy rögzített $\mathbf{x}_i$ $(i \neq k)$ stratégiák mellett e függvény maximumhelye S_k belsejébe esik, a maximumhelyet szolgáltató egyenletet $\mathbf{y}_k$ szerinti gradiensképpzéssel nyerjük:

(1.5)

$$\operatorname{grad}_{\mathbf{y}_k} \varphi_k(\mathbf{x}_1, \ldots, \mathbf{y}_k, \ldots, \mathbf{x}_n) = (\mathbf{A}_{kk}^{(k)} + \mathbf{A}_{kk}^{(k)T}) \mathbf{y}_k + \sum_{i \neq k} (\mathbf{A}_{ki}^{(k)} + \mathbf{A}_{ik}^{(k)T}) \mathbf{x}_i + \mathbf{b}_k^{(k)} = \mathbf{0}.$$

Ha még azt is feltesszük, hogy az $\mathbf{A}_{kk}^{(k)} + \mathbf{A}_{kk}^{(k)T}$ mátrix invertálható (ami az előzőekkel együtt azt jelenti, hogy negatív definit kell legyen), akkor $\mathbf{y}_k$ zárt alakban is kifejezhető:

$$\mathbf{y}_k = -(\mathbf{A}_{kk}^{(k)} - \mathbf{A}_{kk}^{(k)T})^{-1} \Big\{ \sum_{i \neq k} (\mathbf{A}_{ki}^{(k)} + \mathbf{A}_{ik}^{(k)T}) \mathbf{x}_i - \mathbf{b}_k^{(k)} \Big\}. \tag{1.6}$$

Ezzel bebizonyítottuk a következő lemmát:

1.1. LEMMA. Ha rögzített k index és $\mathbf{x}_i$ $(i \neq k)$ stratégiák esetén az (1.3) kifizetőfüggvény maximumhelye az S_k stratégiahalmaz belső pontja, akkor az szükségképpen az (1.6) kifejezéssel adható meg.

Igen gyakran az (1.3) függvény az S_k halmaz határán veszi fel maximumhelyét. Ilyenkor megfelelően definiált büntetőfüggvények bevezetésével, és a stratégiahalmazok kiterjesztésével visszavezethetjük a feladatot a lemmában tárgyalt esetre (l. SZÉP és FORGÓ, 1985).

A továbbiakban a lemma felhasználásával két konkrét dinamikus modellt vizsgálunk meg. A diszkrét és folytonos időskálával rendelkező dinamikus játékok stabilitásának feltételeivel foglalkozunk a következő két paragrafusban.

2. A diszkrét dinamikus modell

Tekintsük a következő dinamikus játékot. A $t=0$ időpillanatban valamennyi játékos egy-egy $\mathbf{x}_1^{(0)}, \ldots, \mathbf{x}_n^{(0)}$ induló stragéiát választ. Minden későbbi $t+1$ időpillanatban a játékosok egymástól függetlenül maximalizálják kifizetésüket feltételezve azt, hogy az összes többi játékos az előző, t időponthoz tartozó stratégiákat választja. Ez a dinamizmus matematikailag azt jelenti, hogy az egyes játékosok $t+1$ időpontbeli $\mathbf{x}_1^{(t+1)}, \ldots, \mathbf{x}_n^{(t+1)}$ stratégiája a

$$\varphi_k(\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_{k-1}^{(t)}, \mathbf{y}_k, \mathbf{x}_{k+1}^{(t)}, \ldots, \mathbf{x}_n^{(t)}) \quad (k = 1, 2, \ldots, n) \tag{2.1}$$

függvény maximumhelyeként adódik. Ha a lemma feltételei teljesülnek, akkor a

játék dinamikáját az

$$\mathbf{x}_k^{(t+1)} = -(\mathbf{A}_{kk}^{(k)} + \mathbf{A}_{kk}^{(k)T})^{-1}\Big\{\sum_{i \neq k} (\mathbf{A}_{ki}^{(k)} + \mathbf{A}_{ik}^{(k)T})\mathbf{x}_i^{(t)} + \mathbf{b}_k^{(k)}\Big\} \tag{2.2}$$

átmeneti egyenletek adják $k=1, 2, \ldots, n$ esetén.

A (2.2) összefüggések tömör alakban is felírhatók, ha bevezetjük az

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11}^{(1)} + \mathbf{A}_{11}^{(1)T} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22}^{(2)} + \mathbf{A}_{22}^{(2)T} & \ldots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{A}_{nn}^{(n)} + \mathbf{A}_{nn}^{(n)T} \end{pmatrix} \tag{2.3}$$

és a

$$\mathbf{B} = \begin{pmatrix} \mathbf{0} & \mathbf{A}_{12}^{(1)} + \mathbf{A}_{21}^{(1)T} & \ldots & \mathbf{A}_{1n}^{(1)} + \mathbf{A}_{n1}^{(1)T} \\ \mathbf{A}_{21}^{(2)} + \mathbf{A}_{12}^{(2)T} & \mathbf{0} & \ldots & \mathbf{A}_{2n}^{(2)} + \mathbf{A}_{n2}^{(2)T} \\ \vdots & \vdots & & \vdots \\ \mathbf{A}_{n1}^{(n)} + \mathbf{A}_{1n}^{(n)T} & \mathbf{A}_{n2}^{(n)} + \mathbf{A}_{2n}^{(n)T} & \ldots & \mathbf{0} \end{pmatrix} \tag{2.4}$$

hipermátrixokat, valamint az

$$\mathbf{x}^{(t)} = \begin{pmatrix} \mathbf{x}_1^{(t)} \\ \mathbf{x}_2^{(t)} \\ \vdots \\ \mathbf{x}_n^{(t)} \end{pmatrix} \quad \text{és az} \quad \mathbf{x}^{(t+1)} = \begin{pmatrix} \mathbf{x}_1^{(t+1)} \\ \mathbf{x}_2^{(t+1)} \\ \vdots \\ \mathbf{x}_n^{(t+1)} \end{pmatrix} \tag{2.5}$$

vektorokat. Ekkor ugyanis

$$\mathbf{x}^{(t+1)} = -\mathbf{A}^{-1}\mathbf{B}\mathbf{x}^{(t)} + \boldsymbol{\beta} \tag{2.6}$$

ahol $\boldsymbol{\beta}$ konstans vektor.

Ismeretes, hogy a (2.6) konstans együtthatós, lineáris differenciaegyenlet-rendszer aszimptotikus stabilitásának szükséges és elégséges feltétele az, hogy az $\mathbf{A}^{-1}\mathbf{B}$ mátrix sajátértékei a komplex egységkör belsejében legyenek. Ha az $\{\mathbf{x}^{(t)}\}$ vektorsorozat limeszének komponensei a stratégiahalmazok belső pontjai, akkor a limeszpont egyensúlypontot szolgáltat. Tehát fennáll a következő tétel:

2.1. TÉTEL. A diszkrét dinamikus játék stabilitásának[1] szükséges és elégséges feltétele az, hogy az $\mathbf{A}^{-1}\mathbf{B}$ mátrix sajátértékei a komplex egységkör belsejében legyenek.

Ez a feltétel közvetlenül nehezen ellenőrizhető, így különös jelentőséggel bír a következő tétel, amely már könnyebben ellenőrizhető elégséges feltételeket ad a rendszer stabilitására.

2.2. TÉTEL. A diszkrét dinamikus modell stabilitásának elégséges feltétele az, hogy az alábbi kritériumok valamelyike fennálljon:

(a) minden $\mathbf{x}$ vektor esetén

$$|\mathbf{x}^*\mathbf{B}\mathbf{x}| < |\mathbf{x}^*\mathbf{A}\mathbf{x}|,$$

ahol $*$ konjugált transzponáltat jelent;

[1] Egy dinamikus játékot stabilisnak mondunk, ha a dinamizmust leíró differencia-, vagy differenciálegyenlet-rendszer aszimptotikusan stabilis.

(b) tetszőleges mátrixnorma esetén

$$\|\mathbf{B}\| < \frac{1}{\|\mathbf{A}^{-1}\|};$$

(c) ha α_i jelöli az $\mathbf{A}$; γ_j pedig a $\mathbf{B}^*\mathbf{B}$ sajátértékeit, akkor tetszőleges i és j esetén

$$\gamma_j < \alpha_i^2.$$

Bizonyítás:

(a) Vegyük először észre, hogy az $\mathbf{A}^{-1}\mathbf{B}$ mátrix sajátértékegyenlete ekvivalens a

$$\mathbf{Bx} = \lambda\mathbf{Ax} \tag{2.7}$$

ún. általános sajátértékfeladattal. Szorozzuk be mindkét oldalt $\mathbf{x}^*$-gal, ekkor

$$\mathbf{x}^*\mathbf{Bx} = \lambda\mathbf{x}^*\mathbf{Ax},$$

azaz

$$\lambda = \frac{\mathbf{x}^*\mathbf{Bx}}{\mathbf{x}^*\mathbf{Ax}},$$

amelyből az (a) feltétel elégségessége már következik.

(b) Jelölje ismét λ az $\mathbf{A}^{-1}\mathbf{B}$ mátrix valamelyik sajátértékét. Ekkor tetszőleges mátrixnorma mellett

$$|\lambda| \leqq \|\mathbf{A}^{-1}\mathbf{B}\| \leqq \|\mathbf{A}^{-1}\| \cdot \|\mathbf{B}\|.$$

Ha a (b) feltétel teljesül, akkor $\|\mathbf{A}^{-1}\| \cdot \|\mathbf{B}\| < 1$, így $|\lambda| < 1$.

(c) Ismeretes, hogy az euklideszi vektornorma által indukált mátrixnorma:

$$\|\mathbf{C}\| = \sqrt{\max|\lambda_{\mathbf{C}^*\mathbf{C}}|},$$

ahol $\lambda_{\mathbf{C}^*\mathbf{C}}$ a $\mathbf{C}^*\mathbf{C}$ mátrix sajátértékeit jelöli. Ekkor pedig az előzőek alapján

$$|\lambda| \leqq \|\mathbf{A}^{-1}\| \cdot \|\mathbf{B}\| = \sqrt{\max|\lambda_{\mathbf{A}^{-1}\mathbf{A}^{-1}}|} \cdot \sqrt{\max|\lambda_{\mathbf{B}^*\mathbf{B}}|} = \frac{1}{\sqrt{\min|\lambda_{\mathbf{A}}^2|}} \cdot \sqrt{\max|\lambda_{\mathbf{B}^*\mathbf{B}}|}.$$

A (c) feltétel elégségessége azonnal adódik abból a tényből, hogy $\mathbf{A}$ szimmetrikus és $\mathbf{B}^*\mathbf{B}$ nemnegatív definit, így $\lambda_{\mathbf{A}}$ valós, valamint $\lambda_{\mathbf{B}^*\mathbf{B}}$ nemnegatív.

3. A folytonos dinamikus modell

Tekintsük most a következő dinamizmust. A $t=0$ időpillanatban az összes játékos választ egy-egy induló stratégiát, majd minden $t \geqq 0$ időpillanatban stratégiáikat az

$$\dot{\mathbf{x}}_k = \mathbf{R}_k(\mathbf{y}_k - \mathbf{x}_k^{(t)}) \tag{3.1}$$

differenciálegyenletnek megfelelően módosítják, ahol $\mathbf{y}_k$ a (2.1) függvény maximumhelye, $\mathbf{R}_k$ pedig adott konstans mátrix. Általában azt teszik fel, hogy $\mathbf{R}_k$ pozitív

[3] Egy $\mathbf{A}$ mátrixot S-stabilisnak mondunk, ha tetszőleges pozitív definit $\mathbf{S}$ mátrix esetén $\mathbf{S}\cdot\mathbf{A}$ sajátértékeinek valós része pozitív.

definit diagonális mátrix. Mi most ettől a speciális esettől eltekintünk. Ekkor behelyettesítve $\mathbf{y}_k$-nak az (1.6) egyenlettel adódó alakját, az

$$\dot{\mathbf{x}}_k = \mathbf{R}_k \cdot \left[-(\mathbf{A}_{kk}^{(k)} + \mathbf{A}_{kk}^{(k)T})^{-1} \Big\{ \sum_{i \neq k} (\mathbf{A}_{ki}^{(k)} + \mathbf{A}_{ik}^{(k)T}) \mathbf{x}_i + \mathbf{b}_k^{(k)} \Big\} - \mathbf{x}_k \right] =$$

$$= -\mathbf{R}_k (\mathbf{A}_{kk}^{(k)} + \mathbf{A}_{kk}^{(k)T})^{-1} \sum_{i=1}^{n} (\mathbf{A}_{ki}^{(k)} + \mathbf{A}_{ik}^{(k)T}) \mathbf{x}_i + \gamma_k$$

összefüggés adódik, ahol γ_k konstans vektor. Itt is feltesszük, hogy az $\mathbf{A}_{kk}^{(k)} + \mathbf{A}_{kk}^{(k)T}$ mátrixok negatív definitek. Ha ezt az összefüggést minden k index esetére felírjuk, és bevezetjük az

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & & & \\ & \mathbf{R}_2 & & \\ & & \ddots & \\ & & & \mathbf{R}_n \end{pmatrix} \tag{3.2}$$

blokkdiagonális mátrixot, ahol az $\mathbf{R}_k$ mátrixok pozitív definitek, ekkor az

$$\dot{\mathbf{x}} = -\mathbf{R}\mathbf{A}^{-1}(\mathbf{A}+\mathbf{B})\mathbf{x} + \gamma \tag{3.3}$$

állandó együtthatós, lineáris differenciálegyenlet-rendszert nyerjük, ahol γ konstans vektor.

A differenciálegyenletek elméletéből ismeretes a következő eredmény.

3.1. TÉTEL. A (3.3) rendszer akkor és csak akkor aszimptotikusan stabilis, ha az $\mathbf{R}\mathbf{A}^{-1}(\mathbf{A}+\mathbf{B})$ mátrix valamennyi sajátértékének a valós része pozitív.

Ez a feltétel a gyakorlatban csak nehezen ellenőrizhető. A továbbiakban néhány könnyebben ellenőrizhető elégséges feltételt adunk a rendszer aszimptotikus stabilitására.

3.2. TÉTEL. A (3.3) rendszer aszimptotikusan stabilis, ha a következő feltételek valamelyike fennáll:

(a) $k=1, \ldots, n$ esetén $\mathbf{R}_k$ felcserélhető az $\mathbf{A}_{kk}^{(k)} + \mathbf{A}_{kk}^{(k)T}$ mátrixszal, valamint $2\mathbf{A}+\mathbf{B}+\mathbf{B}^T$ negatív definit;

(b) $2\mathbf{I}+\mathbf{A}^{-1}\mathbf{B}+\mathbf{B}^T\mathbf{A}^{-1}$ pozitív definit.

Bizonyítás. (a) Belátjuk először, hogy a tett feltételek mellett $\mathbf{R}\mathbf{A}^{-1}$ negatív definit. Ekkor ugyanis $\mathbf{R}$ és $\mathbf{A}$ szimmetriájából adódóan

$$(\mathbf{R}\mathbf{A}^{-1})^T = (\mathbf{A}^{-1})^T\mathbf{R}^T = \mathbf{A}^{-1}\mathbf{R} = \mathbf{R}\mathbf{A}^{-1},$$

azaz $\mathbf{R}\mathbf{A}^{-1}$ szimmetrikus. Legyen ezután β az $\mathbf{R}\mathbf{A}^{-1}$ mátrix valamelyik sajátértéke. Ekkor az

$$\mathbf{R}\mathbf{A}^{-1}\mathbf{x} = \beta\mathbf{x}$$

egyenlet alapján

$$\mathbf{A}^{-1}\mathbf{x} = \beta\mathbf{R}^{-1}\mathbf{x},$$

azaz $\mathbf{x}^*$-gal balról szorozva

$$\beta = \frac{\mathbf{x}^*\mathbf{A}^{-1}\mathbf{x}}{\mathbf{x}^*\mathbf{R}^{-1}\mathbf{x}} < 0,$$

hiszen a számláló negatív, a nevező pedig pozitív. Tehát az

$$\mathbf{R}\cdot\mathbf{A}^{-1}(\mathbf{A}+\mathbf{B})$$

szorzatban $\mathbf{R}\cdot\mathbf{A}^{-1}$ negatív definit, így a rendszer aszimptotikus stabilitásának elégséges feltétele az, hogy $(\mathbf{A}+\mathbf{B})$ S-stabilis legyen. Ennek pedig elégséges feltétele az, hogy $\mathbf{A}+\mathbf{B}+(\mathbf{A}+\mathbf{B})^T$ negatív definit legyen (l. ARROW és MCMANUS, 1958).

(b) Minthogy $\mathbf{R}$ pozitív definit, a rendszer aszimptotikus stabilitásának elégséges feltétele az, hogy $\mathbf{A}^{-1}(\mathbf{A}+\mathbf{B})$ S-stabilis legyen. A (b) feltétel ezt biztosítja.

4. Alkalmazás az oligopol probléma esetén

A többtermékes oligopol probléma a következőképpen fogalmazható meg. Jelölje n a termelők, M pedig a termékek számát. Ekkor a termelőket játékosoknak tekintve, a k-adik játékos stratégiája az $\mathbf{x}_k=(x_k^{(1)}, \ldots, x_k^{(M)})^T$ vektorral írható le, ahol $x_k^{(m)}$ jelöli az m-edik termékféleségből előállított mennyiségét. Tegyük fel, hogy $k=1, \ldots, n$ esetén a k-adik játékos stratégiahalmaza a „nemnegatív térnyolcad" konvex, korlátos, zárt részhalmaza, amelyre fennáll, hogy tetszőleges $\mathbf{x}_k\in S_k$ és $\mathbf{0}\leqq\tilde{\mathbf{x}}_k\leqq\mathbf{x}_k$ esetén $\tilde{\mathbf{x}}_k\in S_k$. Jelölje $s^{(1)}, \ldots, s^{(M)}$ az egyes termékekből a játékosok által együttesen előállított mennyiségeket. Tegyük fel, hogy az egyes termékek egységára

$$f_m(s^{(1)}, \ldots, s^{(M)}) = \sum_{\mu=1}^{M} a_m^{(\mu)}s^{(\mu)} + A_m, \tag{4.1}$$

ahol $a_m^{(\mu)}(\forall m, \mu)$, $A_m(\forall m)$ adott konstansok. Tegyük fel továbbá, hogy a k-adik játékos költségfüggvénye is lineáris:

$$K_k(\mathbf{x}_k) = \sum_{m=1}^{M} b_k^{(m)}x_k^{(m)} + B_k. \tag{4.2}$$

Könnyen kimutathatjuk, hogy az imént definiált lineáris, többtermékes oligopol játék eleget tesz a dolgozat elejére tett általános feltételeknek, ugyanis a k-adik játékos kifizetőfüggvénye:

$$\varphi_k(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \mathbf{x}_k^T\cdot\mathbf{f}(\mathbf{s}) - K_k(\mathbf{x}_k), \tag{4.3}$$

ahol $\mathbf{s}=(s^{(1)}, \ldots, s^{(M)})^T$, $\mathbf{f}=(f_1, \ldots, f_n)^T$. A (4.3) kifejezés pedig a következő alakban is felírható:

$$\varphi_k(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \mathbf{x}_k^T\cdot\mathbf{E}\cdot(\mathbf{x}_1+\ldots+\mathbf{x}_n) - K_k(\mathbf{x}_k).$$

$$\mathbf{E} = \begin{pmatrix} a_1^{(1)} & \ldots & a_1^{(M)} \\ \vdots & & \vdots \\ a_M^{(1)} & \ldots & a_M^{(M)} \end{pmatrix},$$

Így az (1.1) jelöléseinek felhasználásával

$$\mathbf{A}_{ij}^{(k)} = \begin{cases} \mathbf{E}, & \text{ha} \quad i = k \\ \mathbf{0}, & \text{ha} \quad i \neq k. \end{cases}$$

Tehát a (2.3) és (2.4) jelöléseivel

$$\mathbf{A} = \begin{pmatrix} \mathbf{E}+\mathbf{E}^T & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{E}+\mathbf{E}^T \end{pmatrix}$$

valamint

$$\mathbf{B} = \begin{pmatrix} \mathbf{0} & \mathbf{E} & \dots & \mathbf{E} \\ \mathbf{E} & \mathbf{0} & \dots & \mathbf{E} \\ \dots & \dots & \dots & \dots \\ \mathbf{E} & \mathbf{E} & \dots & \mathbf{0} \end{pmatrix}.$$

Tegyük fel az előzőekhez hasonlóan, hogy $\mathbf{E}+\mathbf{E}^T$ negatív definit, valamint most azt is, hogy $\mathbf{E}^{-1}\mathbf{E}^T$ sajátértékei valósak.

Tekintsük először a diszkrét modellt. Ekkor

$$\mathbf{A}^{-1}\mathbf{B} = \begin{pmatrix} 0 & 1 & \dots & 1 \\ 1 & 0 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 0 \end{pmatrix} \times (\mathbf{E}+\mathbf{E}^T)^{-1}\mathbf{E}.$$

Ismeres, hogy az első tényező sajátértékei -1 és $n-1$ (l. SZIDAROVSZKY és OKUGUCHI, 1986). A második tényező pedig a következő alakú:

$$(\mathbf{E}+\mathbf{E}^T)^{-1}\mathbf{E} = (\mathbf{I}+\mathbf{E}^{-1}\mathbf{E}^T)^{-1}. \tag{4.4}$$

Könnyen belátható, hogy $\mathbf{E}^{-1}\mathbf{E}^T$ sajátértékei egységnyi értékűek. Tekintsük ennek igazolására $\mathbf{E}^{-1}\mathbf{E}^T$ sajátértékegyenletét:

$$\mathbf{E}^{-1}\mathbf{E}^T\mathbf{x} = \delta\mathbf{x},$$

amelyből $\mathbf{E}$-vel, majd $\mathbf{x}^*$-gal balról szorozva azonnal adódik, hogy

$$\delta = \frac{\mathbf{x}^*\mathbf{E}^T\mathbf{x}}{\mathbf{x}^*\mathbf{E}\mathbf{x}} = 1,$$

hiszen a számláló és a nevező megegyezik.

Így a (4.4) mátrix sajátértékei valamennyien $(1+1)^{-1}=1/2$ értékűek. Tehát $\mathbf{A}^{-1}\mathbf{B}$ sajátértékeinek értéke vagy $-\frac{1}{2}$, vagy $\frac{n-1}{2}$. Így a (2.6) diszkrét rendszer akkor és csak akkor aszimptotikusan stabilis, ha

$$\frac{n-1}{2} < 1,$$

azaz $n<3$. Tehát igaz a következő állítás.

4.1. TÉTEL. A fenti feltételek mellett a diszkrét többtermékes lineáris oligopol modell akkor és csak akkor aszimptotikusan stabilis, ha $n<3$.

Tekintsük ezután a folytonos modellt. Ekkor a (3.3) alapján

$$\mathbf{A}+\mathbf{B}=\begin{pmatrix}\mathbf{E}+\mathbf{E}^T & \mathbf{E} & \dots & \mathbf{E}\\ \mathbf{E} & \mathbf{E}+\mathbf{E}^T & \dots & \mathbf{E}\\ \dots & \dots & \dots & \dots\\ \mathbf{E} & \mathbf{E} & \dots & \mathbf{E}+\mathbf{E}^T\end{pmatrix},$$

így

$$2\mathbf{A}+\mathbf{B}+\mathbf{B}^T=\begin{pmatrix}2 & 1 & \dots & 1\\ 1 & 2 & \dots & 1\\ \dots & \dots & \dots & \dots\\ 1 & 1 & \dots & 2\end{pmatrix}\times(\mathbf{E}+\mathbf{E}^T). \tag{4.5}$$

Az első tényező

$$2\mathbf{I}+\begin{pmatrix}0 & 1 & \dots & 1\\ 1 & 0 & \dots & 1\\ \dots & \dots & \dots & \dots\\ 1 & 1 & \dots & 0\end{pmatrix},$$

így ennek sajátértékei 1 és $n+1$, azaz pozitívak. Minthogy $\mathbf{E}+\mathbf{E}^T$ negatív definit, a (4.5) mátrix is negatív definit. Ezért a 3.2. tétel (a) állítása alapján fennáll a következő:

4.2. TÉTEL. Tegyük fel, hogy $k=1, \dots, n$ esetén $\mathbf{R}_k$ és $\mathbf{E}$ felcserélhető. Ekkor a folytonos többtermékes lineáris oligopol játék aszimptotikusan stabilis.

A dolgozat befejezéseként a tett feltételeket elemezzük röviden.

1. Az S_k stratégiahalmazokra és a φ_k kifizetőfüggvényekre tett általános feltételek a játék konkavitását biztosítják. Ezzel az egyensúlypont létezését tudjuk biztosítani. Az $\mathbf{A}_{kk}^{(k)}+\mathbf{A}_{kk}^{(k)T}$ mátrixokra tett negatív definitségi feltétel a konkavitáshoz és az $\mathbf{A}$ mátrix invertálhatóságához szükséges. Az oligopol játék esetén azt kell feltennünk, hogy $\mathbf{E}+\mathbf{E}^T$ negatív definit. Egyetlen termék ($M=1$) esetén ez azt jelenti, hogy az f árfüggvény az együttesen előállított termékmennyiség szigorúan csökkenő függvénye.

2. Feltettük, hogy az (1.3) függvények maximumhelyei S_k belső pontjai. Ez a feltétel nyilvánvalóan fennáll, ha a modell asszimptotikusan stabilis, az egyensúlyi stratégiák a stratégiahalmazok belső pontjai, és a kezdeti stratégiavektorokat pedig alkalmasan választjuk meg.

IRODALOM

]1] ARROW, K. J. and MCMANUS, M., "A note on dynamic stability", *Econometrica* **26** (1958) 448—454.

[2] OKUGUCHI, K. és SZIDAROVSZKY, F., „A többtermékes oligopol probléma stabilitásáról", *Szigma* (1986) megjelenés alatt.

[3] SZÉP, J. and FORGÓ, F., *Introduction to the Theory of Games* (Akadémiai Kiadó, Budapest, 1985).

[4] SZIDAROVSZKY, F. and OKUGUCHI, K., "Linear and nonlinear oligopoly models", in: *Proceedings of the 12th IFIP Conference on Systems Modelling and Optimalization*, ed. *A. Prékopa, J. Szelezsán and B. Strazicky*, Springer Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, 1986, 949—956.

[5] SZIDAROVSZKY, F. and OKUGUCHI, K., "Stability of the Cournot equilibrium for an oligopoly with multi-product firms", in: *Proceedings of the 5th IFAC/IFORS Conference on Dynamic Modelling and Control of National Economics*, Budapest, 1986. június 17—20. (megjelenés alatt).
[6] THEOCHARIS, R. D., "On the stability of the Cournot solution on the oligopoly problem", *Review of Econometric Studies* **27** (1959) 133—134.

(Beérkezett: 1986. szeptember 8.)

SZIDAROVSZKY FERENC
MKKE MATEMATIKAI ÉS SZÁMÍTÁSTUDOMÁNYI INTÉZET
1093 BUDAPEST, DIMITROV TÉR 8.
OKUGUCHI KOJI
TOKYO METROPOLITAN UNIVERSITY
1-1-1 YAKUMO, MEGURO-KU, JAPÁN.

ON THE STABILITY OF QUADRATIC GAMES

F. SZIDAROVSZKY and K. OKUGUCHI

Sufficient conditions are given for the stability of n-person games with quadratic payoff functions. Two models are discussed. In the case of discrete time scale the stability of the game is reduced to the asymptotical stability of a system of linear difference equations, and in the case of continuous time scale the asymptotical stability of a system of linear differential equations is investigated. As an application the oligopoly problem is discussed.

Alkalmazott Matematikai Lapok **13** (1987—88) 137—143

EGY SZÓRÁSCSÖKKENTÉSI ELJÁRÁSRÓL

SZIDAROVSZKY FERENC

Budapest

Dolgozatunkban a korrelált változók módszerének egy új általánosítását vezetjük be, amikor az integrálandó függvény véges sor alakjában áll elő. Az egyes tagok integráljának becsléséhez szükséges mintaelem számot is meghatározzuk ket különböző feltételezés mellett. Az egyik esetben az adott becslési variancia eléréséhez szükséges műveletszámot minimalizáljuk, a másik modell esetében pedig adott műveletszám mellett minimalizáljuk a becslési varianciát. Eredményeink két tag feltételezésével FLICK és JONES (1986) módszerét és optimumát szolgáltatják, mint speciális esetet.

1. Bevezetés

Integrálok közelltő meghatározásának a gyakorlatban alkalmazott egyik legfontosabb módszercsaládját jelentik a szimulációs módszerek. A leggyakrabban alkalmazott *Monte-Carlo-módszerek* összefoglalását adja például SZIDAROVSZKY (1974). Az alapmódszer a következő. Tegyük fel, hogy egy f függvény

$$I = \int_D f(x)dx \tag{1.1}$$

integrálját kívánjuk meghatározni valamilyen D halmazon. Itt D általában R^k korlátos részhalmaza. Legyenek $X_1, \ldots, X_N$ a D halmazon egyenletes eloszlású független véletlen számok, ekkor I *Monte-Carlo becslése:*

$$I \cong \frac{1}{N}\sum_{j=1}^{N} f(X_j) \overset{\text{jel}}{=} I_1. \tag{1.2}$$

Az (1.2) becslés műveletigénye nyilvánvalóan NR, ahol R jelenti egy véletlen X_j előállításának és a hozzá tartozó $f(X_j)$ függvényérték kiszámításának együttes műveletigényét. Minthogy a műveleti igény N-nel egyenesen arányos, a gyakorlati alkalmazások során nagyon fontos probléma N értékének a csökkentése. Erre szolgálnak a szóráscsökkentési eljárások, amikor az (1.2) jobb oldalát úgy alakítják át, hogy annak varianciája csökkenjék, hiszen — mint ismeretes — ezzel együtt csökken az adott hibakorlát biztosításához szükséges mintaelemszám is.

Az egyik leggyakrabban alkalmazott szóráscsökkentő eljárás az, amikor az f függvényt egy könnyen integrálható g függvénnyel közelítjük, és (1.2) helyett az

$$I \simeq \frac{1}{N}\sum_{j=1}^{N}[f(X_j)-g(X_j)] + \int_D g(x)dx \overset{\text{jel}}{=} I_2 \tag{1.3}$$

becslést alkalmazzuk (RUBINSTEIN, 1981). Kimutatható, hogy

$$\text{(1.4)} \qquad \mathrm{Var}(I_1) = \frac{1}{N}\mathrm{Var}(f), \quad \mathrm{Var}(I_2) = \frac{1}{N}\mathrm{Var}(f-g),$$

így általában az I_2 becslés kisebb szórású, mint I_1.

Az (1.3) eljárás alkalmazhatóságának az a legfontosabb feltétele, hogy a g függvény integrálját pontosan ismerjük. Ennek a — gyakran kritikus — feltételnek az enyhítését javasolta FLICK és JONES (1986), amikor (1.3) helyett az

$$\text{(1.5)} \qquad I \simeq \frac{1}{N_f}\sum_{j=1}^{N_f}[f(X_j)-g(X_j)] + \frac{1}{N_g}\sum_{j=N_f+1}^{N_f+N_g} g(X_j) \overset{\text{jel}}{=} I_3$$

becslést bevezette. Az (1.3) és (1.5) módszer között az a különbség, hogy amíg (1.3) esetében g integráljainak pontos ismeretét tételezzük fel, addig (1.5) esetében azt is becsüljük. A módszer bevezetésén és illusztrálásán kívül megadták az optimális N_f, N_g értékeket is, amelyek adott műveletszám mellett minimalizálják a becslés szórását.

Dolgozatunkban az (1.5) becslési eljárás általánosításával foglalkozunk. Tegyük fel, hogy az f függvény az

$$\text{(1.6)} \qquad f = g_1 + g_2 + \ldots + g_n$$

véges összeg alakjában áll elő. Legyenek $N_1, N_2, \ldots, N_n$ pozitív egészek, és tekintsük az

$$\text{(1.7)} \quad I \simeq \frac{1}{N_1}\sum_{j=1}^{N_1} g_1(X_j) + \frac{1}{N_2}\sum_{j=N_1+1}^{N_1+N_2} g_2(X_j) + \ldots + \frac{1}{N_n}\sum_{j=N_1+\ldots+N_{n-1}}^{N_1+\ldots N_n} g_n(X_j) \overset{\text{jel}}{=} I_4$$

becslést. Nyilvánvaló, hogy I_4 is torzítatlan becslése I-nek, valamint

$$\mathrm{Var}(I_4) = \frac{1}{N_1}\mathrm{Var}(g_1) + \frac{1}{N_2}\mathrm{Var}(g_2) + \ldots + \frac{1}{N_n}\mathrm{Var}(g_n).$$

Tegyük fel, hogy egyetlen X_j mintaelem generálásának és a hozzá tartozó $g_i(X_j)$ függvényérték meghatározásának együttes műveletigénye R_i $(i=1, 2, \ldots, n)$, ekkor az I_4 becslés teljes várható műveletigénye

$$\text{(1.8)} \qquad T = N_1R_1 + N_2R_2 + \ldots + N_nR_n.$$

2. Optimális becslési stratégia kialakítása

Először az adott műveletigény melletti legkedvezőbb becslési stratégiát határozzuk meg. Tegyük fel, hogy $T=T_0$ (ahol T_0 adott), és minimalizáljuk e feltétel mellett a Var (I_4) becslési varianciát! A probléma *Lagrange-függvénye:*

$$\text{(2.1)} \qquad L_1 = \sum_{j=1}^{n}\frac{1}{N_j}\mathrm{Var}(g_j) + \lambda\Big(\sum_{j=1}^{n} N_jR_j - T_0\Big).$$

Egyszerű differenciálással adódik, hogy

$$\frac{\partial L_1}{\partial N_j} = -\frac{1}{N_j^2}\operatorname{Var}(g_j) + \lambda R_j = 0. \tag{2.2}$$

Ebből az összefüggésből azonnal leolvasható, hogy tetszőleges j és l esetén

$$\frac{N_l}{N_j} = \left\{\frac{R_j \operatorname{Var}(g_l)}{R_l \operatorname{Var}(g_j)}\right\}^{1/2}, \tag{2.3}$$

amely közvetlen általánosítása FLICK és JONES (1986) egyik eredményének ((1.8) egyenlőség). A (2.2) összefüggés egy másik következménye az, hogy az optimális N_j értékeket meghatározhatjuk. A (2.2) alapján

$$N_j = \left\{\frac{\operatorname{Var}(g_j)}{\lambda R_j}\right\}^{1/2}, \tag{2.4}$$

amelyet (1.8)-ba helyettesítve az

$$\frac{1}{\sqrt{\lambda}} \sum_{l=1}^{n} R_l \left\{\frac{\operatorname{Var}(g_l)}{R_l}\right\}^{1/2} = T_0 \tag{2.5}$$

összefüggést kapjuk. Azaz

$$\frac{1}{\sqrt{\lambda}} = \frac{T_0}{\sum_{l=1}^{n} \{R_l \operatorname{Var}(g_l)\}^{1/2}}, \tag{2.6}$$

így (2.4) alapján az optimális N_j érték:

$$N_j = \frac{T_0}{R_j + \sum_{l \neq j} \left\{R_l R_j \frac{\operatorname{Var}(g_l)}{\operatorname{Var}(g_j)}\right\}^{1/2}}, \tag{2.7}$$

amely közvetlen általánosítása FLICK és JONES (1986) eredményeinek ((2.3) és (2.4) relációk).

Másodszor az adott V_0 becslési variancia eléréséhez szükséges minimális műveletigényt határozzuk meg. Ebben az esetben a $\operatorname{Var}(I_4) = V_0$ feltétel mellett minimalizáljuk a T műveletigényt. A probléma *Lagrange-függvénye* most:

$$L_2 = \sum_{j=1}^{n} N_j R_j + \lambda\left(\sum_{j=1}^{n} \frac{1}{N_j} \operatorname{Var}(g_j) - V_0\right). \tag{2.8}$$

Az előző modellhez hasonlóan látható be, hogy (2.3) most is fennáll, valamint az optimális N_j értékeke:

$$N_j = \frac{\operatorname{Var}(g_j) + \sum_{l \neq j} \left\{\frac{R_l}{R_j} \cdot \operatorname{Var}(g_l) \cdot \operatorname{Var}(g_j)\right\}^{1/2}}{V_0}. \tag{2.9}$$

Az első modell esetén az optimális becslési variancia értéke:

$$\text{(2.10)} \qquad \mathrm{Var}_{\mathrm{opt}}(T_0) = \sum_{j=1}^{n} \frac{1}{N_j} \mathrm{Var}(g_j) =$$

$$= \sum_{j=1}^{n} \frac{\sum_{i=1}^{n} \left\{ R_i R_j \frac{\mathrm{Var}(g_i)}{\mathrm{Var}(g_j)} \right\}^{1/2}}{T_0} \mathrm{Var}(g_j) = \frac{1}{T_0} \left[\sum_{j=1}^{n} \{\mathrm{Var}(g_j)\}^{1/2} \right]^2,$$

a másik modell esetén pedig az optimális műveletigény:

$$\text{(2.11)} \qquad T_{\mathrm{opt}}(V_0) = \sum_{j=1}^{n} N_j R_j = \frac{1}{V_0} \sum_{j=1}^{n} R_j \cdot \sum_{i=1}^{n} \left\{ \frac{R_i}{R_j} \cdot \mathrm{Var}(g_i) \cdot \mathrm{Var}(g_j) \right\}^{1/2} =$$

$$= \frac{1}{V_0} \left[\sum_{j=1}^{n} \{R_j \mathrm{Var}(g_j)\}^{1/2} \right]^2.$$

A (2.10) és (2.11) összefüggésből a következő következtetések olvashatók le:

1. T_0 növelésével $\mathrm{Var}_{\mathrm{opt}}(T_0)$ úgy csökken, hogy szorzatuk konstans maradjon;
2. V_0 csökkentésével $T_{\mathrm{opt}}(V_0)$ úgy növekedik, hogy szorzatuk állandó maradjon;
3. A két modellt összeköti az a tény is, hogy

$$\text{(2.12)} \qquad T_0 \mathrm{Var}_{\mathrm{opt}}(T_0) = V_0 \cdot T_{\mathrm{opt}}(V_0) = \left[\sum_{j=1}^{n} \{R_j \cdot \mathrm{Var}(G_j)\}^{1/2} \right]^2.$$

3. Alkalmazás végtelen sorok esetére

Tegyük fel most, hogy $f = \sum_{i=1}^{\infty} g_i$, valamint

$$\text{(3.1)} \qquad \int_D f(x)\,dx = \sum_{i=1}^{\infty} \int_D g_i(x)\,dx.$$

Például egyenletesen konvergens függvénysorok esetén ez a helyzet. Ekkor elég nagy n esetén

$$\text{(3.2)} \qquad \int_D f(x)\,dy \simeq \sum_{i=1}^{n} \int_D g_i(x)\,dx \approx I_4.$$

Vegyük észre, hogy elég nagy N_j $(j=1, 2, \ldots, n)$ értékek mellett a centrális határeloszlástétel alapján I_4 minden tagja normális eloszlásúnak tekinthető. Függetlenségük következtében I_4 eloszlása is normális, ebből következően az

$$\text{(3.3)} \qquad \eta = \int_D f(x)\,dx - I_4$$

becslési hiba is normális eloszlású. Könnyen belátható, hogy

$$\text{(3.4)} \qquad E_n(\eta) = \int_D f(x)\,dx - \sum_{i=1}^{n} \int_D g_i(x)\,dx,$$

valamint

$$\text{Var}_n(\eta) = \text{Var}(I_4). \tag{3.5}$$

Legyen most $\varepsilon>0$ olyan rögzített hibakorlát, amelyre igaz, hogy

$$\varepsilon > |E_n(\eta)|. \tag{3.6}$$

Definiáljuk a (3.2) becslés „jóságát" a

$$P(|\eta| < \varepsilon) \tag{3.7}$$

valószínűségértékkel. Nyilvánvalóan annál jobb egy becslés, minél nagyobb a (3.7) valószínűségérték. Egyszerű számolás mutatja, hogy

$$P(|\eta| < \varepsilon) = \varphi\left(\frac{\varepsilon - E_n(\eta)}{\text{Var}_n(\eta)^{1/2}}\right) - \varphi\left(\frac{-\varepsilon - E_n(\eta)}{\text{Var}_n(\eta)^{1/2}}\right) = \tag{3.8}$$

$$= \varphi\left(\frac{\varepsilon - E_n(\eta)}{\text{Var}_n(\eta)^{1/2}}\right) + \varphi\left(\frac{\varepsilon + E_n(\eta)}{\text{Var}_n(\eta)^{1/2}}\right) - 1.$$

Minthogy $\varepsilon - E_n(\eta)$ és $\varepsilon + E_n(\eta)$ pozitívak és φ szigorúan növekedő, adott T_0 műveletigény mellett akkor maximális ez a valószínűségi érték, amikor $\text{Var}_n(\eta)^{1/2}$ minimális. Ez pedig akkor teljesül (3.5) alapján, amikor $\text{Var}(I_n)$ minimális. Ezzel beláttuk azt, hogy a (3.8) valószínűségi érték maximalizálása és az előző paragrafusban ismertetett első modell ekvivalensek rögzített n esetén. Végtelen sorok esetén az optimális n érték megkeresése is feladat lehet. Ilyenkor azt javasolhatjuk, hogy egyre növekedő n értékek mellett számítsuk ki a (3.8) valószínűségi érték optimális $\text{Var}_n(\eta)$ melletti értékeit, és ott álljunk meg, amikor ez a számított valószínűségi érték csökkenni kezd. Ezzel legalábbis lokális optimumot nyerünk. Megjegyezzük végül, hogy az $E_n(\eta)$ értékek pontosan általában nem ismertek, így a gyakorlatban valamilyen becslést alkalmazunk; valamint elég nagy n esetén a (3.6) feltétel biztosan teljesül.

4. Kockázati függvény minimalizálása

Mint láttuk, az (1.7) becslés alkalmazását két paraméter, nevezetesen T és $\text{Var}(I_4)$ jellemzi. A T műveletigény a számítások költségét jellemzi, míg a $\text{Var}(I_4)$ varianciaérték a számítások bizonytalanságát. Tegyük fel, hogy rendelkezésünkre áll egy $K(\text{Var}(I_4))$ kárfüggvény, amely a varianciaértékek függvényében megadja a bizonytalan integrálértékből eredő gazdasági kárt. Ha M jelenti egységnyi művelet költségét, akkor az együttes költség és kár összege:

$$MT + K(\text{Var}(I_4)) = M \cdot \left(\sum_{i=1}^{n} N_i R_i\right) + K\left(\sum_{i=1}^{n} \frac{1}{N_i} \text{Var}(g_i)\right). \tag{4.1}$$

Leggazdaságosabb becslési stratégiát nyilván akkor kapunk, amikor ez a függvény minimális értékű. Egyszerű deriválással látható, hogy ebből az

$$MR_i - K'\left(\sum_{i=1}^{n} \frac{1}{N_i} \text{Var}(g_i)\right) \cdot \frac{\text{Var}(g_i)}{N_i^2} = 0 \quad (i = 1, 2, \ldots, n) \tag{4.2}$$

egyenletrendszer adódik. Az egyszerűbb jelölések érdekében vezessük be a

$$\lambda = K'\left(\sum_{i=1}^{n} \frac{1}{N_i} \operatorname{Var}(g_i)\right) \tag{4.3}$$

változót. Ekkor (4.2) alapján $i=1, 2, \ldots, n$ esetén

$$N_i = \left\{\lambda \cdot \frac{\operatorname{Var}(g_i)}{NR_i}\right\}^{1/2}, \tag{4.4}$$

amelyet (4.3)-ba visszahelyettesítve a

$$\lambda = K'\left(\frac{1}{\sqrt{\lambda}} \sum_{i=1}^{n} \{MR_i \operatorname{Var}(g_i)\}^{1/2}\right) \tag{4.5}$$

egyenlőség adódik. Legyen

$$Q = \sum_{i=1}^{n} \{R_i \operatorname{Var}(g_i)\}^{1/2}, \tag{4.6}$$

ekkor (4.5) az egyszerűbb

$$\lambda = K'\left(\frac{Q\sqrt{M}}{\sqrt{\lambda}}\right) \tag{4.7}$$

alakba írható. Ennek megoldásai közül kerül ki az a λ érték, amelyből (4.4) alapján az optimális N_i értékek adódnak. Ha (4.7)-nek egyetlen megoldása van, akkor ezt kell elfogadnunk. Ha több megoldása létezik, akkor a (4.1) célfüggvényértékek összehasonlítása alapján kell az optimumot megkapnunk.

Például, lineáris K függvény esetén K' konstans. Ekkor pedig ez adja az optimumhoz tartozó λ értéket. Ha K' növekedik és $K'(\infty)>0$, akkor (4.7)-nek nyilvánvalóan egyetlen megoldása van, amely meghatározására standard módszerek állnak rendelkezésre (YAKOWITZ és SZIDAROVSZKY, 1986). Az általános esetben (4.7) megoldása bonyolult numerikus problémákat is felvethet, ezzel azonban most nem foglalkozunk.

IRODALOM

[1] FLICK, TH. E. and JONES, L. K., "A generalization of the method of correlated sampling for numerical integration", *SIAM Journal on Scientific Statistical Computations* 7 (1986) 1037—1040.
[2] RUBINSTEIN, R. Y., *Simulation and the Monte Carlo Methods* (John Wiley and Sons, New York, 1981).
[3] RUBINSTEIN, R. Y. and SZIDAROVSZKY, F., "Convergence of perturbation analysis estimates for queuing networks: A general approach", *IEEE Automatic Control* megjelenés alatt.
[4] RUBINSTEIN, R. Y. and SZIDAROVSZKY, F., "Convergence of perturbation analysis estimates for discontinuous sample functions: A general approach", Harvard University, kézirat.
[5] SZIDAROVSZKY, F., *Bevezetés a numerikus módszerebe* (Közgazdasági és Jogi Könyvkiadó, Budapest, 1974).
[6] YAKOWITZ, S. and SZIDAROVSZKY, F., *Introduction to Numerical Computations* (MacMillen, New York, 1986).

(Beérkezett: 1986. október 2.)

SZIDAROVSZKY FERENC
MARX KÁROLY KÖZGAZDASÁGTUDOMÁNYI EGYETEM MATEMATIKAI INTÉZET
1093 BUDAPEST, DIMITROV TÉR 9.

A VARIANCE REDUCTION TECHNIQUE

F. SZIDAROVSZKY

A new generalization of the method of correlated sampling is given when the integrand is in the form of an infinite sum. The necessary sample size is also given under two different assumptions. In the first case the number of operations is minimalized when the variance of the estimation is given and in the second model the variance of the estimation is minimized when the number of operations is given. The method by FLICK and JONES is a special case of our results.

Alkalmazott Matematikai Lapok **13** (1987—88) 145—161

MEGJEGYZÉS EGY TÖBB DIMENZIÓS CAUCHY ELOSZLÁSRÓL[1]

KLAFSZKY EMIL KAS PÉTER

Miskolc Budapest

A dolgozatban egy *több dimenziós Cauchy-eloszlás* sűrűségfüggvénynek paramétereit vizsgáljuk. Megadjuk az entrópiát és a kölcsönös információt a paraméterek függvényében. Megmutatjuk, hogy a paraméter mátrix egyfajta korreláció-mérő szerepet játszik.

1. Bevezetés

A dolgozat célja, hogy a *Cauchy-eloszlás* egy több dimenziós kiterjesztéséhez tartozó sűrűségfüggvény paramétereinek tartalmára rávilágítson. Minthogy a *Cauchy-eloszlásnak* általunk vizsgált több dimenziós általánosítása szoros kapcsolatban van a normális eloszlással, ezért gondolatmenetünket először a *több dimenziós Gauss-eloszlásra* alkalmazzuk.

A tárgyalás eszköze véletlen vektorok lineáris transzformációja. A vizsgált paraméterek erősen kapcsolódnak információelméleti fogalmakhoz. A felhasznált elemi eszközöket és fogalmakat a bevezetésben összefoglaljuk.

Véletlen vektor lineáris transzformáltja

Tárgyalásunkban fontos szerepet játszik véletlen vektorok lineáris transzformáltja. Az alábbi jól ismert eredményt (lásd [7]) többször fogjuk használni, ezért lemmaként előrevetítjük.

1.1. LEMMA. Legyen az m dimenziós ξ véletlen vektor sűrűségfüggvénye $f(\mathbf{x})$. Akkor az

$$\eta = \mathbf{S}\xi + \mathbf{c}$$

lineárisan transzformált véletlen vektor sűrűségfüggvénye

$$\frac{1}{|\det \mathbf{S}|} f(\mathbf{S}^{-1}(\mathbf{x}-\mathbf{c}))$$

ahol, $\mathbf{S}$ $m \times m$ méretű nem szinguláris mátrix és $\mathbf{c}$ m-dimenziós vektor.

[1] A dolgozat az OTKA—1049 kutatási pályázat keretében készült.

Információdivergencia, kölcsönös információ, entrópia

A dolgozatban csak pozitív értékű sűrűségfüggvényeket engedünk meg, ezért a fogalmakat is csak pozitív értékű sűrűségfüggvényre definiáljuk. Alapvető fogalom az információdivergencia, ebből származtatjuk a többi fogalmat is (lásd [1], [2], [4]).

1.1. *Definíció.* Legyenek ξ és η m-dimenziós véletlen vektorok $f_\xi(\mathbf{x})$ és $f_\eta(\mathbf{x})$ sűrűségfüggvénnyel. A ξ eltérése η véletlen vektortól (szokásos mondani, hogy $f_\xi(\mathbf{x})$ eltérése $f_\eta(\mathbf{x})$ sűrűségfüggvénytől):

$$D(\xi\|\eta) = \int_{R^m} f_\xi(\mathbf{x}) \log \frac{f_\xi(\mathbf{x})}{f_\eta(\mathbf{x})}\, d\mathbf{x},$$

amelyet információdivergenciának (röviden *I*-divergenciának) vagy diszkrimináló információnak nevezünk. Ismeretes, hogy $D(\xi\|\eta)\geqq 0$ és egyenlőség akkor és csak akkor, ha m.m. $f_\xi(\mathbf{x})=f_\eta(\mathbf{x})$.

A következő fogalom, az *I-divergenciára* támaszkodva, a véletlen vektor koordinátái között sztochasztikus kapcsolat mérését szolgálja. A $\xi=(\xi_1, \xi_2, \ldots, \xi_m)$ véletlen vektor sűrűségfüggvényét jelölje $f(x_1, x_2, \ldots, x_m)$, a peremeloszlások, azaz $\xi_1, \xi_2, \ldots, \xi_m$ sűrűségfüggvényeit $f_1(x_1), f_2(x_2), \ldots, f_m(x_m)$ amennyiben $\xi_1, \xi_2, \ldots, \xi_m$ koordináták sztochasztikusan függetlenek lennének úgy a ξ véletlen vektor sűrűségfüggvénye $f_1(x_1)\cdot f_2(x_2)\cdot \ldots \cdot f_m(x_m)$ lenne. A tényleges sűrűségfüggvénynek a sztochasztikus függetlenséget feltételező sürűségfüggvénytől *I-divergenciával* mért eltérését tekintjük a sztochasztikus kapcsolat mérőszámának. Ezt az alábbi definícióban adjuk meg.

1.2. *Definíció.* A $\xi=(\xi_1, \xi_2, \ldots, \xi_m)$ véletlen vektor koordinátái közötti kölcsönös információ

$$I(\xi) = \int_{R^m} f(\mathbf{x}) \log \frac{f(\mathbf{x})}{f_1(x_1) f_2(x_2) \ldots f_m(x_m)}\, d\mathbf{x}$$

ahol f az együttes, az $f_1, f_2, \ldots, f_m$ pedig a peremeloszlások sűrűségfüggvénye. Felhasználva az *I-divergenciára* tett megjegyzésünket nyilvánvaló, hogy $I(\xi)\geqq 0$ és egyenlőség akkor és csak akkor, ha $\xi_1, \xi_2, \ldots, \xi_m$ sztochasztikusan függetlenek. A kölcsönös információ egyszerűbb felírását és tartalmának más oldalú megvilágítását teszi lehetővé az entrópia fogalmának bevezetése.

1.3. *Definíció.* Legyen ξ m-dimenziós véletlen vektor $f(\mathbf{x})$ sűrűségfüggvénnyel. A $H(\xi)=\int_{R^m} f(\mathbf{x}) \log \frac{1}{f(\mathbf{x})}\, d\mathbf{x}$ értéket a ξ véletlen vektor (vagy az $f(\mathbf{x})$ sűrűségfüggvény) entrópiájának nevezzük, ha ez létezik. Amennyiben a kölcsönös információ véges és az entrópiák léteznek akkor az entrópia fogalmával a kölcsönös információ egyszerűen kifejezhető az alábbi módon:

$$I(\xi) = H(\xi_1) + H(\xi_2) + \ldots + H(\xi_m) - H(\xi)$$

A kölcsönös információnál tett megjegyzésünk szerint

$$H(\xi_1) + H(\xi_2) + \ldots + H(\xi_m) \geqq H(\xi)$$

és egyenlőség akkor és csak akkor áll, ha $\xi_1, \xi_2, \ldots, \xi_m$ sztochasztikusan függetlenek. A következő lemma a lineárisan transzformált véletlen vektor entrópiájának egyszerű kiszámolását teszi lehetővé.

1.2. LEMMA. Legyen ξ sűrűségfüggvénnyel bíró m-dimenziós véletlen vektor, $\mathbf{S}$ $m \times m$-es nem szinguláris mátrix és $\mathbf{c}$ m-dimenziós vektor. Akkor

$$H(\mathbf{S}\xi + \mathbf{c}) = H(\xi) + \log |\det \mathbf{S}|.$$

A lemma az entrópia definícióját felhasználva elemi számolással nyerhető.

2. Több dimenziós Gauss eloszlás

Ebben a fejezetben áttekintjük a *több dimenziós Gauss eloszlásra* vonatkozó azon eredményeket, melyek a *több dimenziós Cauchy eloszlás* tárgyalásához szükségesek. Ezek az eredmények több tankönyvben megtalálhatók (lásd pl.: TUSNÁDY [6]) itt a megismétlésük a *Cauchy eloszlás* tárgyalásmódjának motivációját szolgálja. A 2.1 tételre egyszerű mátrixaritmetikán alapuló bizonyítást is adunk. A *több dimenziós Gauss eloszlást* az egydimenziós sürüségfüggvény általánosításával definiáljuk. Ezért elöljáróban megismételjük az egydimenziós eloszlás fogalmát.

2.1. *Definíció.* A ξ véletlen skalárt *egyszerű Gauss eloszlásúnak* nevezzük, ha a sűrüségfüggvénye

$$f(x) = \frac{1}{\sqrt{\pi}} \cdot \exp(-x^2)$$

Az egyszerü véletlen skalár lineáris transzformáltjainak összességét eloszlás családnak nevezzük.

2.2. *Definíció.* A ξ *egyszerű Gauss eloszlású* véletlen skalár lineáris transzformáltját

$$\eta = s\xi + c \quad (s \neq 0,\ c \text{ tetszőleges})$$

Gauss eloszlású véletlen skalárnak nevezzük. Az 1.1 lemma szerint η sürüségfüggvénye:

$$f(x, c, s^2) = \frac{1}{\sqrt{\pi}} \cdot \frac{1}{|s|} \cdot \exp\left(-\left(\frac{x-c}{s}\right)^2\right)$$

A sürűségfüggvényből látható, hogy az csak c és s^2 paraméterektől függ, emiatt szokásos (c, s^2) paraméterű *Gauss eloszlásnak* is nevezi.

A több dimenziós Gauss eloszlás definíciója

A *több dimenziós egyszerű Gauss eloszlás* sűrűségfüggvényét az egydimenziós sűrűségfüggvény analógjaként definiáljuk.

2.3. *Definíció.* A $\xi = (\xi_1, \ldots, \xi_m)$ m-dimenziós véletlen vektor *egyszerű Gauss eloszlású*, ha koordinátái sztochasztikusan független *egyszerű Gauss eloszlású* véletlen skalárok. A definícióból azonnal adódik, hogy az *m-dimenziós egyszerű Gauss el-*

oszlású véletlen vektor sűrűségfüggvénye:

$$f(\mathbf{x}) = \frac{1}{\pi^{m/2}} \exp(-\mathbf{x}^2), \quad \mathbf{x} \in R^m.$$

A *több dimenziós általános Gauss eloszlást* az egydimenziós lineáris transzformált analógjaként definiáljuk.

2.4. *Definíció.* A ξ *m-dimenziós egyszerű Gauss eloszlású* véletlen vektor lineáris transzformáltját

$$\eta = \mathbf{S}\xi + \mathbf{c} \quad (\det \mathbf{S} \neq 0 \text{ és } \mathbf{c} \text{ tetszőlegesek})$$

Gauss eloszlású véletlen vektornak nevezzük. Az 1.1 lemma szerint nyilvánvalóan η sűrűségfüggvénye:

$$f(\mathbf{x}, \mathbf{c}, \mathbf{SS}^*) = \frac{1}{\pi^{m/2}} \frac{1}{|\det \mathbf{S}|} \exp\left(-(\mathbf{S}^{-1}(\mathbf{x}-\mathbf{c}))^* \mathbf{S}^{-1}(\mathbf{x}-\mathbf{c})\right)$$

($*$ a transzponáltat jelenti). Az η véletlen vektort szokásos $(\mathbf{c}, \mathbf{SS}^*)$ paraméterű *Gauss eloszlásúnak* nevezni. A $\mathbf{c}$, a centrum az $\mathbf{SS}^*$ pozitív definit mátrix, az úgynevezett koncentráció mátrix, szerepét jobban hangsúlyozandó a sűrűségfüggvényt az alábbi formában írjuk:

$$f(\mathbf{x}, \mathbf{c}, \mathbf{SS}^*) = \frac{1}{\pi^{m/2}} \frac{1}{\sqrt{\det \mathbf{SS}^*}} \exp\left(-(\mathbf{x}-\mathbf{c})^* (\mathbf{SS}^*)^{-1}(\mathbf{x}-\mathbf{c})\right).$$

A következő összefüggés a *Gauss eloszlással* kapcsolatos vizsgálatokban döntő szerepet játszik. A tételre több bizonyítás ismert, az általunk itt adott, csak a definícióra támaszkodó bizonyítás, úgy gondoljuk, érdeklődésre tarthat számot.

2.1. TÉTEL. Legyen a ξ *egyszerű Gauss eloszlású* n-dimenziós véletlen vektor és $\mathbf{T}$ $m \times n$-es, sorrangú, tetszőleges mátrix, valamint $\mathbf{c}$ tetszőleges m-dimenziós vektor. A lineárisan transzformált

$$\eta = \mathbf{T}\xi + \mathbf{c}$$

m-dimenziós véletlen vektor *Gauss eloszlású* $\mathbf{c}$ és $\mathbf{SS}^*$ paraméterekkel, ahol $\mathbf{S}$ olyan, hogy

$$\mathbf{SS}^* = \mathbf{TT}^*$$

Bizonyítás. A bizonyításban a mátrix aritmetikából jól ismert alábbi összefüggést használjuk fel:

Legyen $\mathbf{A}$ $m \times m$-es nem szinguláris mátrix és $\mathbf{b}$ m-dimenziós vektor, akkor

$$(\mathbf{A}+\mathbf{bb}^*)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{bb}^*\mathbf{A}^{-1}}{1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b}}, \tag{2.1}$$

ha $\mathbf{A}$, $\mathbf{b}$ olyanok, hogy $\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b} \neq -1$ (például, ha $\mathbf{A}$ pozitív definit akkor ez minden vektorra teljesül).

Továbbá

$$\det(\mathbf{A}+\mathbf{bb}^*) = (1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b}) \det(\mathbf{A}). \tag{2.2}$$

Rátérve a tétel bizonyítására megjegyezzük, hogy azt elég a $\mathbf{c}=\mathbf{0}$ esetre elvégezni, ebből egyszerű helyettesítéssel adódik a $\mathbf{c}\neq\mathbf{0}$ eset.

A bizonyítást n-re vonatkozó teljes indukcióval bizonyítjuk. Az $n=m$ indexre az 1.1 lemma alapján nyilvánvaló a tétel állítása. Tegyük fel, hogy valamely $n\geqq m$ indexre igaz az állítás, belátjuk, hogy $n+1$-re is igaz.

A bizonyítás áttekinthetősége érdekében az n indexhez tartozó transzformációs mátrixot $\mathbf{T}$-vel, az $n+1$-hez tartozót $\hat{\mathbf{T}}$-pal jellöjük; a $\hat{\mathbf{T}}$ mátrix utolsó oszlopát $\mathbf{b}$ vektorral jelöljük. Így

$$\hat{\mathbf{T}} = (\mathbf{T}, \mathbf{b})$$

és nyilvánvalóan

$$\hat{\mathbf{T}}\hat{\mathbf{T}}^* = \mathbf{T}\mathbf{T}^* + \mathbf{b}\mathbf{b}^*.$$

Az indukciós feltevés miatt a $\mathbf{T}$ transzformációs mátrix esetén a sűrűségfüggvény:

$$\frac{1}{\pi^{m/2}} \cdot \frac{1}{\sqrt{\det(\mathbf{T}\mathbf{T}^*)}} \cdot \exp\left(-\mathbf{x}^*(\mathbf{T}\mathbf{T}^*)^{-1}\mathbf{x}\right). \tag{2.3}$$

Be kell látnunk, hogy a $\hat{\mathbf{T}}$ transzformációs mátrix esetén a sűrűségfüggvény:

$$\frac{1}{\pi^{m/2}} \cdot \frac{1}{\sqrt{\det(\hat{\mathbf{T}}\hat{\mathbf{T}}^*)}} \cdot \exp\left(-\mathbf{x}^*(\hat{\mathbf{T}}\hat{\mathbf{T}}^*)^{-1}\mathbf{x}\right) \tag{2.4}$$

Használva a (2.1), (2.2) összefüggéseket és az $\mathbf{A}=\mathbf{T}\mathbf{T}^*$ jelölést bevezetve (2.3); (2.4) az alábbi formát ölti:

$$\frac{1}{\pi^{m/2}} \cdot \frac{1}{\sqrt{\det(\mathbf{A})}} \cdot \exp\left(\mathbf{x}^*\mathbf{A}^{-1}\mathbf{x}\right) \tag{2.5}$$

és

$$\frac{1}{\pi^{m/2}} \cdot \frac{1}{\sqrt{(1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b})\det(\mathbf{A})}} \cdot \exp\left(-\mathbf{x}^*\left(\mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{b}\mathbf{b}^*\mathbf{A}^{-1}}{1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b}}\right)\mathbf{x}\right).$$

Jelöljük

$$\zeta = \mathbf{T}\xi.$$

Így

$$\eta = \zeta + \xi_{n+1} \cdot \mathbf{b}$$

A teljes valószínűség tétel alapján

$$P(\eta < \mathbf{x}) = P(\zeta + \xi_{n+1}\mathbf{b} < \mathbf{x}) = \int_{-\infty}^{\infty} P(\zeta < \mathbf{x} - u\mathbf{b}) \cdot f(u)\, du = \int_{-\infty}^{\infty} F(\mathbf{x} - u\mathbf{b}) f(u)\, du, \tag{2.6}$$

ahol F a (2.5) alatti sűrűségfüggvényhez tartozó eloszlásfüggvény, f pedig az *egyszerű egydimenziós Gauss eloszlás* sűrűségfüggvénye. Ezeket helyettesítve (2.6)-ba kapjuk, hogy

$$\frac{\partial}{\partial x} P(\eta < \mathbf{x}) = \frac{1}{\pi^{m/2}} \cdot \frac{1}{\sqrt{\pi}} \cdot \frac{1}{\sqrt{\det(\mathbf{A})}} \cdot \int_{-\infty}^{+\infty} \exp\left(-(\mathbf{x} - u\mathbf{b})^*\mathbf{A}^{-1}(\mathbf{x} - u\mathbf{b}) - u^2\right) du. \tag{2.7}$$

Vizsgáljuk először a (2.7) exp függvényének argumentumát. Azonos átalakításokkal nyerjük, hogy

$$-(\mathbf{x}-u\mathbf{b})^*\mathbf{A}^{-1}(x-u\mathbf{b})-u^2 = -\left(u\sqrt{1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b}}-\frac{\mathbf{b}^*\mathbf{A}^{-1}\mathbf{x}}{\sqrt{1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b}}}\right)^2+ \tag{2.8}$$

$$+\frac{(\mathbf{b}^*\mathbf{A}^{-1}\mathbf{x})^2}{1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b}}-\mathbf{x}^*\mathbf{A}^{-1}\mathbf{x}.$$

Ezen (2.8) átalakítás felhasználásával a (2.7) integrál értékére kapjuk, hogy

$$\int_{-\infty}^{\infty}\exp\left(-(\mathbf{x}-u\mathbf{b})^*\mathbf{A}^{-1}(\mathbf{x}-u\mathbf{b})-u^2\right)du = \tag{2.9}$$

$$=\exp\left(\frac{(\mathbf{b}^*\mathbf{A}^{-1}\mathbf{x})^2}{1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b}}-\mathbf{x}^*\mathbf{A}^{-1}\mathbf{x}\right)\int_{-\infty}^{\infty}\exp\left(-\left(u\sqrt{1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b}}-\frac{\mathbf{b}^*\mathbf{A}^{-1}\mathbf{x}}{\sqrt{1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b}}}\right)^2\right)du=$$

$$=\frac{\sqrt{\pi}}{\sqrt{1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b}}}\exp\left(\frac{(\mathbf{b}^*\mathbf{A}^{-1}\mathbf{x})^2}{1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b}}-x^*\mathbf{A}^{-1}\mathbf{x}\right).$$

Az exp függvény argumentumát az alábbi formában írhatjuk:

$$\frac{(\mathbf{b}^*\mathbf{A}^{-1}\mathbf{x})^2}{1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b}}-\mathbf{x}^*\mathbf{A}^{-1}\mathbf{x}=\mathbf{x}^*\left(\frac{\mathbf{A}^{-1}\mathbf{b}\mathbf{b}^*\mathbf{A}^{-1}}{1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b}}-\mathbf{A}^{-1}\right)\mathbf{x} \tag{2.10}$$

A (2.10), (2.9), (2.3) átalakításokat visszaírva (2.7)-be kapjuk, hogy

$$\frac{\partial}{\partial\mathbf{x}}P(\boldsymbol{\eta}<\mathbf{x})=\frac{1}{\pi^{m/2}}\cdot\frac{1}{\sqrt{\det(\mathbf{A})}}\cdot\frac{1}{\sqrt{1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b}}}\times$$

$$\times\exp\left(-\mathbf{x}^*\left(\mathbf{A}^{-1}-\frac{\mathbf{A}^{-1}\mathbf{b}\mathbf{b}^*\mathbf{A}^{-1}}{1+\mathbf{b}^*\mathbf{A}^{-1}\mathbf{b}}\right)\mathbf{x}\right)$$

amit igazolni kellett. ∎

KÖVETKEZMÉNY. A peremeloszlások ismét *Gauss eloszlások*, melynek koncentráció mátrixa az eredeti koncentráció mátrix leszűkítése.

Bizonyítás. A tételből nyilvánvaló. ∎

Megjegyezzük, hogy kétdimenziós esetben a sűrűségfüggvényt gyakran a következő alakban használják:

$$f(x_1,x_2)=$$

$$=\frac{1}{\pi}\cdot\frac{1}{s_1s_2\sqrt{1-r^2}}\cdot\exp\left(-\frac{1}{1-r^2}\left(\frac{(x_1-c_1)}{s_1^2}-2r\frac{(x_1-c_1)(x_2-c_2)}{s_1s_2}+\frac{(x_2-c_2)^2}{s_2^2}\right)\right).$$

Ekkor a transzformációs mátrix:

$$\mathbf{S}=\begin{bmatrix}s_1 & 0\\ r\cdot s_2 & \sqrt{1-r^2}\cdot s_2\end{bmatrix}.$$

A koncentráció mátrix:

$$\mathbf{S}\cdot\mathbf{S}^*=\begin{bmatrix}s_1^2 & rs_1s_2\\ rs_1s_2 & s_2^2\end{bmatrix}.$$

A Gauss eloszlás entrópiája, kölcsönös információja

Az előzőekben láttuk, hogy a *Gauss eloszlást* a **c** centrum vektor és a SS* koncentráció mátrix jellemzi. Ebben a részben megmutatjuk, hogy a koncentráció mátrixot az entrópia és a kölcsönös információ karakterizálja.

2.1. LEMMA. A ξ m-dimenziós *egyszerű Gauss eloszlású* véletlen vektor entrópiája

$$H(\xi) = \frac{m}{2}(1+\log \pi)$$

Bizonyítás. A függetlenséget kihasználva elemi integrálással adódik. ∎

A következő lemma az *általános Gauss eloszlású* véletlen vektor entrópiáját adja.

2.2. LEMMA. Legyen ξ *egyszerű Gauss eloszlású* n-dimenziós véletlen vektor és $\mathbf{T}$ $m\times n$-es, sorrangú, tetszőleges mátrix, valamint $\mathbf{c}$ tetszőleges m-dimenziós vektor. A lineárisan transzformált

$$\eta = \mathbf{T}\xi + \mathbf{c}$$

m-dimenziós véletlen vektor entrópiája:

$$H(\eta) = \frac{m}{2}(1+\log \pi) + \frac{1}{2}\log \det(\mathbf{TT}^*).$$

Bizonyítás. A 2.1 tétel szerint

$$\mathbf{T}\xi^{(n)} = \mathbf{S}\xi^{(m)}$$

ahol $\mathbf{S}$ $m\times m$-es reguláris mátrix olyan, hogy

$$\mathbf{SS}^* = \mathbf{TT}^* \tag{2.11}$$

és $\xi^{(n)}$, valamint $\xi^{(m)}$ *egyszerű Gauss eloszlású* véletlen vektorok (a felső index a vektor dimenzióját jelenti). Az 1.2 lemma szerint η entrópiája:

$$H(\eta) = \frac{m}{2}(1+\log \pi) + \log|\det(\mathbf{S})|$$

azonban $\log|\det(\mathbf{S})| = \frac{1}{2}\log\det(\mathbf{SS}^*)$ és (2.11) fennállása alapján kapjuk a lemma állítását. ∎

KÖVETKEZMÉNY. Jelöljük a lemmában szereplő $\mathbf{T}$ mátrix sorait $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_m$ vektorokkal, akkor

$$H(\eta_i) = \frac{1}{2}(1+\log \pi) + \frac{1}{2}\log(\mathbf{t}_i^2) \tag{2.12}$$

Bizonyítás. A lemmából nyilvánvaló. ∎

A lemma alapján azonnal kapjuk a *Gauss eloszlású* véletlen vektor koordinátái közötti kölcsönös információt. Ezt fontos szerepe miatt tételként mondjuk ki.

2.2. TÉTEL. Legyen ξ *egyszerű Gauss eloszlású* n-dimenziós véletlen vektor és $\mathbf{T}$ $m \times n$-es, sorrangú, tetszőleges mátrix ($m \geqq 2$); valamint $\mathbf{c}$ tetszőleges m-dimenziós vektor. A lineárisan transzformált

$$\eta = \mathbf{T}\xi + \mathbf{c}$$

m-dimenziós véletlen vektor koordinátái közötti kölcsönös információ

$$I(\eta) = -\frac{1}{2} \log \frac{\det(\mathbf{TT}^*)}{\mathbf{t}_1^2 \mathbf{t}_2^2 \ldots \mathbf{t}_m^2}$$

ahol $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_m$ a $\mathbf{T}$ mátrix sorvektorai.

Bizonyítás. Az η véletlen vektor koordinátái közötti kölcsönös információ:

$$I(\eta) = H(\eta_1) + \ldots + H(\eta_m) - H(\eta)$$

felhasználva a 2.2 lemma és a következmény állítását kapjuk a tétel állítását. ∎

Megjegyezzük, hogy a *Hadamard egyenlőtlenség* szerint

$$\det(\mathbf{TT}^*) \leqq \mathbf{t}_1^2 \mathbf{t}_2^2 \ldots \mathbf{t}_m^2$$

és egyenlőség akkor és csak akkor áll, ha $\mathbf{t}_i \cdot \mathbf{t}_j^* = 0$ minden $i \neq j$ indexre, azaz ha $\mathbf{t}_i$ merőleges $\mathbf{t}_j$ vektorra. Mivel $\eta_1, \ldots, \eta_m$ komponensek akkor és csak akkor függetlenek, ha $I(\eta) = 0$ így $\eta_1, \eta_2, \ldots, \eta_m$ sztochasztikusan függetlenek akkor és csak akkor, ha $\mathbf{t}_i \cdot \mathbf{t}_j^* = 0$ minden $i \neq j$ indexpárra.

KÖVETKEZMÉNY. Ha az η vektornak egy kétkoordinátás részvektorát tekintjük, akkor

$$I(\eta_i, \eta_j) = -\frac{1}{2} \log \left(1 - \frac{(\mathbf{t}_i \mathbf{t}_j^*)^2}{\mathbf{t}_i^2 \cdot \mathbf{t}_j^2}\right) \tag{2.13}$$

Bizonyítás. A tételből nyilvánvaló. ∎

Megjegyezzük, hogy a következményből a *Cauchy—Schwartz—Bunyakovszkij egyenlőtlenséget* használva azonnal adódik, hogy η_i, η_j, $i \neq j$ akkor és csak akkor sztochasztikusan függetlenek, ha $\mathbf{t}_i \cdot \mathbf{t}_j^* = 0$.

Érdekességként megemlítjük, hogy a tétel és a következménye után tett megjegyzések összefűzéséből a *Hadamard egyenlőtlenségnek* a *Cauchy—Schwartz—Bunyakovszkij egyenlőtlenségből* egy gyors levezetését kapjuk. (Ugyanis ha $I(\eta) = 0$, akkor $\eta_1, \ldots, \eta_m$ sztochasztikusan függetlenek, de akkor η_i, η_j is sztochasztikusan függetlenek, ekkor viszont $\mathbf{t}_i \cdot \mathbf{t}_j^* = 0$.

A lemma és a tétel következményei (2.12) és (2.13) mutatják a *Gauss eloszlás* koncentráció mátrixának tartalmát. Azonban mint (2.13)-ból látszik a koordináták páronkénti kölcsönös információja a koncentráció mátrix elemeinek csak az abszolút értékét adja meg. Az előjel meghatározásához további vizsgálatok szükségesek.

3. Több dimenziós Cauchy eloszlás

Az alábbiakban egy *több dimenziós Cauchy eloszlás* definícióját adjuk meg ([5] 57. old.).

A több dimenziós Cauchy eloszlás definíciója

A *Cauchy eloszlást* a *Gauss eloszlásból* származtatjuk.

3.1. Definíció. Legyen ζ véletlen skalár és ζ m-dimenziós véletlen vektor, amelyek sztochasztikusan független *egyszerű Gauss eloszlásúak*. A

$$\xi = \frac{\zeta}{\zeta}$$

véletlen vektort *egyszerű m-dimenziós Cauchy eloszlásúnak* nevezzük.

3.1. LEMMA. Az *egyszerű m-dimenziós Cauchy eloszlás* sűrűségfüggvénye

$$f(\mathbf{x}) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\pi^{(m+1)/2}} \cdot \frac{1}{(1+\mathbf{x}^2)^{(m+1)/2}}, \quad \mathbf{x} \in R^m.$$

Bizonyítás. Minthogy az *egyszerű Gauss eloszlás* centrálszimmetrikus, ezért a

$$\frac{\zeta}{\zeta} \quad \text{és} \quad \frac{\zeta}{|\zeta|}$$

véletlen vektorok egyforma eloszlásúak. A ξ eloszlásfüggvényét tehát a következőképpen írhatjuk:

$$P(\xi < \mathbf{x}) = P(\zeta < |\zeta|\mathbf{x}) = 2\int_0^{+\infty} P(\zeta < u\mathbf{x}) f(u)\, du = 2\int_0^{+\infty} F(u\mathbf{x}) f(u)\, du,$$

ahol F az *egyszerű több dimenziós Gauss eloszlás*, f az *egyszerű Gauss eloszlás* sűrűségfüggvénye.

Áttérve a sűrűségfüggvényre

$$\frac{\partial}{\partial \mathbf{x}} P(\xi < \mathbf{x}) = 2\int_0^{+\infty} f(u\mathbf{x}) \cdot u^m f(u)\, du = 2\int_0^{+\infty} \frac{1}{\pi^{m/2}} \cdot e^{-u^2\mathbf{x}^2} \frac{1}{\pi^{1/2}} e^{-u^2} u^m\, du =$$

$$= \frac{2}{\pi^{(m+1)/2}} \int_0^{+\infty} e^{-u^2(\mathbf{x}^2+1)} u^m\, du = \frac{\Gamma\left(\frac{m+1}{2}\right)}{(1+\mathbf{x}^2)^{(m+1)/2}} \cdot \frac{1}{\pi^{(m+1)/2}},$$

amit bizonyítanunk kellett. Az utolsó lépésben felhasznált

$$2\int_0^{+\infty} u^m e^{-au^2}\, du = \frac{\Gamma\left(\frac{m+1}{2}\right)}{a^{(m+1)/2}}, \quad a > 0, \quad m = 0, 1, 2, \ldots$$

összefüggés a gamma függvény definíciójából

$$\Gamma(p) = \int_0^{+\infty} x^{p-1}e^{-x}\,dx, \quad p > 0$$

egyszerű transzformációval adódik. ∎

A több dimenziós *általános Cauchy eloszlást* az egyszerű eloszlás lineáris transzformáltjaként definiáljuk.

3.2. Definíció. A ξ m-dimenziós *egyszerű Cauchy eloszlású* véletlen vektor lineáris transzformáltját

$$\eta = S\xi + c, \quad \det S \neq 0 \text{ és } c \text{ tetszőlegesek}$$

Cauchy eloszlású véletlen vektornak nevezzük. Az 1.1 lemma szerint nyilvánvalóan η sűrűségfüggvénye:

$$f(x, c, SS^*) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\pi^{(m+1)/2}} \cdot \frac{1}{|\det S|} \cdot \frac{1}{\left(1+(S^{-1}(x-c))^2\right)^{(m+1)/2}}.$$

Az η véletlen vektort szokásos (c, SS^*) paraméterű *Cauchy eloszlásúnak* nevezni. A c, a centrum vektor az SS^* pozitív definit mátrix, az úgynevezett koncentráció mátrix szerepét jobban hangsúlyozandó a sűrűségfüggvényt az alábbi formában írjuk:

$$f(x, c, SS^*) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\pi^{(m+1)/2}} \cdot \frac{1}{\sqrt{\det SS^*}} \cdot \frac{1}{\left(1+(x-c)^*(SS^*)^{-1}(x-c)\right)^{(m+1)/2}}.$$

A 3.1 lemma segítségével a *Gauss eloszlás* alapvető 2.1. transzformációs tételével teljesen analóg transzformációs összefüggés állítható.

3.1. TÉTEL. Legyen a ξ *egyszerű Chauchy eloszlású n-dimenziós véletlen vektor* és T $(m \times n)$ méretű, sorrangú mátrix, valamint c tetszőleges m-dimenziós vektor. A lineárisan transzformált

$$\eta = T\xi + c$$

m-dimenziós véletlen vektor *Cauchy eloszlású* c és SS^* paraméterekkel, ahol S olyan, hogy

$$SS^* = TT^*.$$

Bizonyítás. A bizonyításban különböző dimenziójú *egyszerű Gauss*, illetve *Cauchy eloszlású* véletlen vektorok fordulnak elő, ezért dimenziójukat felső indexben jelöljük és a ζ szimbólumot *Gauss*, a ξ szimbólumot *Cauchy eloszlású* véletlen vektorra használjuk. Feltevésünk szerint

$$\eta = T\xi^{(n)} + c$$

A *Cauchy eloszlás* definíciója alapján

$$\zeta^{(n)} = \frac{\zeta^{(n)}}{\xi}$$

Így

$$\eta = \mathbf{T}\left(\frac{\zeta^{(n)}}{\xi}\right) + \mathbf{c} = \frac{\mathbf{T}\zeta^{(n)}}{\xi} + \mathbf{c} \tag{3.1}$$

A 2.1 tétel alapján

$$\mathbf{T}\zeta^{(n)} = \mathbf{S}\zeta^{(m)} \tag{3.2}$$

ha S olyan, hogy $\mathbf{SS}^* = \mathbf{TT}^*$. A (3.1)-be (3.2)-t helyettesítve kapjuk, hogy

$$\eta = \frac{\mathbf{S}\zeta^{(m)}}{\zeta} + \mathbf{c} = \mathbf{S}\left(\frac{\zeta^{(m)}}{\zeta}\right) + \mathbf{c}$$

ami a *Cauchy eloszlás* definíciója szerint

$$\eta = \mathbf{S}\xi^{(m)} + \mathbf{c}$$

amit bizonyítani kellett. ∎

KÖVETKEZMÉNY. A peremleoszlások ismét *Cauchy eloszlások*, amelyeknek koncentráció mátrixa az eredeti koncentráció mátrix leszűkítése.

Bizonyítás. A tételből nyilvánvaló. ∎

Az egydimenziós *Cauchy eloszlás* nyilvánvalóan az alábbi alakot ölti:

$$\frac{1}{\pi} \cdot \frac{1}{|s|} \cdot \frac{1}{1 + \left(\frac{x-c}{s}\right)^2}.$$

A kétdimenziós esetben a sűrűségfüggvényt gyakran a következő alakban írják fel:

$$\frac{1}{2\pi} \cdot \frac{1}{s_1 s_2 \sqrt{1-r^2}} \cdot \frac{1}{\left(1 + \frac{1}{1-r^2}\left(\frac{(x_1-c_1)^2}{s_1^2} - 2r\frac{(x_1-c_1)}{s_1} \cdot \frac{(x_2-c_2)}{s_2} + \frac{(x_2-c_2)^2}{s_2^2}\right)\right)^{3/2}}.$$

Ekkor a transzformáció mátrixa:

$$\mathbf{S} = \begin{bmatrix} s_1 & 0 \\ rs_2 & \sqrt{1-r^2} \cdot s_2 \end{bmatrix}.$$

A koncentráció mátrix pedig:

$$\mathbf{SS}^* = \begin{bmatrix} s_1^2 & rs_1 s_2 \\ rs_1 s_2 & s_2^2 \end{bmatrix}.$$

A Cauchy eloszlás entrópiája, kölcsönös információja

Az előzőekben láttuk, hogy a *Cauchy eloszlást* a **c** centrum és a $\mathbf{TT}^*$ koncentráció mátrix jellemzi. Ebben a részben megmutatjuk, hogy az entrópia és a kölcsönös információ fogalmával hogyan karakterizálhatjuk a koncentráció mátrix elemeit.

3.2. LEMMA. A ξ *m-dimenziós egyszerű Cauchy eloszlású véletlen vektor* entró-

piája

$$H(\xi) = \log \frac{\pi^{(m+1)/2}}{\Gamma\left(\frac{m+1}{2}\right)} + \frac{m+1}{2}\left(\psi\left(\frac{m+1}{2}\right) - \psi\left(\frac{1}{2}\right)\right), \tag{3.3}$$

ahol

$$\psi(x) = \frac{d}{dx}\log\Gamma(x).$$

Mielőtt a lemma bizonyítását elvégeznénk a Γ és ψ függvények speciális helyeken felvett értékeit felhasználva a (3.3) formulát számolásra alkalmasabb formába írjuk. Ismeretes, hogy

$$\Gamma\left(\frac{m+1}{2}\right) = \begin{cases} \frac{\sqrt{\pi}}{2^k}\prod_{i=0}^{k-1}(2i+1), & \text{ha } m = 2k \\ k! & \text{ha } m = 2k+1 \end{cases}$$

(az üres indexhalmazon $\prod$ értéke legyen 1). Tudjuk még, hogy

$$\psi\left(\frac{m+1}{2}\right) = \begin{cases} -C + 2\left(\sum_{i=1}^{k}\frac{1}{2i-1} - \log 2\right), & \text{ha } m = 2k \\ -C + \sum_{i=1}^{k}\frac{1}{i}, & \text{ha } m = 2k+1 \end{cases} \tag{3.4}$$

(az üres indexhalmazon Σ értéke legyen 0), C pedig az *Euler konstans*.

Ezeket felhasználva (3.3) az alábbi formát ölti:

$$H(\xi) = \begin{cases} k\log 2\pi + \sum_{i=0}^{k-1}\left(\frac{2k+1}{2i+1} - \log(2i+1)\right), & \text{ha } m = 2k \\ (k+1)\log 4\pi + \sum_{i=0}^{k-1}\left(\frac{k+1}{i+1} - \log(i+1)\right), & \text{ha } m = 2k+1 \end{cases}$$

Speciálisan az egy- és kétdimenziós esetben:

$$H(\xi_i) = \log 4\pi$$
$$H(\xi_i, \xi_j) = \log 2\pi + 3 \tag{3.5}$$

A (3.4) formulából egyszerű számolással kapjuk, hogy

$$\psi\left(\frac{m+1}{2}\right) - \psi\left(\frac{m-1}{2}\right) = \frac{2}{m-1}, \quad \text{ha } m > 2,$$

amit szintén felhasználunk a lemma bizonyításánál. (Megjegyezzük, hogy az általánosabb

$$\psi(x+1) - \psi(x) = \frac{1}{x}, \quad x > 0$$

összefüggés is fenn áll).

A lemma bizonyítása. Az entrópia integrál definícióját felírva:

$$H = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{\infty} \frac{\Gamma\left(\frac{m+1}{2}\right)}{\pi^{(m+1)/2}} \cdot \frac{1}{\left(1+\sum_{i=1}^{m} x_i^2\right)^{(m+1)/2}} \times$$

$$\times \log\left(\frac{\pi^{(m+1)/2}}{\Gamma\left(\frac{m+1}{2}\right)} \cdot \left(1+\sum_{i=1}^{m} x_i^2\right)^{(m+1)/2}\right) dx_1 \dots dx_m =$$

$$= \log \frac{\pi^{(m+1)/2}}{\Gamma\left(\frac{m+1}{2}\right)} + \frac{\Gamma\left(\frac{m+1}{2}\right)}{\pi^{(m+1)/2}} -$$

$$- \frac{m+1}{2} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\log\left(1+\sum_{i+1}^{m} x_i^2\right)}{\left(1+\sum_{i=1}^{m} x_i^2\right)^{(m+1)/2}} dx_1 \dots dx_m.$$

Tehát a lemma igazolásához azt kell megmutatnunk, hogy

(3.6)

$$\frac{\Gamma\left(\frac{m+1}{2}\right)}{\pi^{(m+1)/2}} \int_{-\infty}^{+\infty} \dots \int_{+\infty}^{+\infty} \frac{\log\left(1+x_1^2+\dots+x_m^2\right)}{\left(1+x_1^2+\dots+x_m^2\right)^{(m+1)/2}} dx_1 \dots dx_m = \psi\left(\frac{m+1}{2}\right) - \psi\left(\frac{1}{2}\right).$$

Ezt a teljes indukcióval bizonyítjuk. Először $m=1$ és $m=2$ értékére, majd $m-2$ indexről m-re. Az $m=1$ esetben (3.6) azt jelenti, hogy

$$\int_0^{\infty} \frac{\log\left(1+x^2\right)}{1+x^2} dx = \pi \log 2.$$

Ezt pedig az ismert (lásd pl. [3])

$$\int_0^{\pi/2} \log \cos t dt = -\frac{\pi}{2} \log 2$$

integrál felhasználásával,

$$x = \operatorname{tg} t$$

helyettesítéssel kapjuk:

$$\int_0^{\infty} \frac{\log\left(1+x^2\right)}{1+x^2} dx = \int_0^{\pi/2} \log\left(1+\operatorname{tg}^2 t\right) dt = -2 \int_0^{\pi/2} \log \cos t \, dt = \pi \log 2.$$

Az $m=2$ esetben (3.6) azt jelenti, hogy

$$\int_0^\infty \int_0^\infty \frac{\log(1+x^2+y^2)}{(1+x^2+y^2)^{3/2}}\,dx\,dy = \pi.$$

Az

$$x = r\cos\varphi, \quad y = r\sin\varphi$$

helyettesítéssel

$$\int_0^\infty \int_0^\infty \frac{\log(1+x^2+y^2)}{(1+x^2+y^2)^{3/2}}\,dx\,dy = \int_0^{\pi/2} \int_0^\infty \frac{\log(1+r^2)}{(1+r^2)^{3/2}}\,r\,dr d\varphi =$$

$$= \frac{\pi}{2} \int_0^\infty \frac{\log(1+r^2)}{(1+r^2)^{3/2}}\,r\,dr =$$

$$t = (1+r^2)^{-1/2}$$

helyettesítéssel

$$= -\frac{\pi}{2} \int_1^0 \log t^{-2}\,dt = \pi.$$

Tegyük fel, hogy (3.6) igaz $m-2$ indexre, azaz

$$2^{m-2} \frac{\Gamma\left(\frac{m-1}{2}\right)}{\pi^{(m-1)/2}} \int_0^\infty \dots \int_0^\infty \frac{\log(1+x_1^2+\dots+x_{m-2}^2)}{(1+x_1^2+\dots+x_{m-2}^2)^{(m-1)/2}}\,dx_1 \dots dx_{m-2} =$$

$$= \psi\left(\frac{m-1}{2}\right) - \psi\left(\frac{1}{2}\right)$$

be kell látnunk, hogy m-re is fenn áll. A

$$J_m = \underbrace{\int_0^\infty \dots \int_0^\infty}_{m} \frac{\log(1+x_1^2+\dots+x_m^2)}{(1+x_1^2+\dots+x_m^2)^{(m+1)/2}}\,dx_1 \dots dx_m$$

integrált az

$$x_{m-1} = r\cos\varphi, \quad x_m = r\sin\varphi$$

helyettesítéssel az alábbi formára hozzuk:

$$(3.7) \qquad J_m = \frac{\pi}{2} \underbrace{\int_0^\infty \dots \int_0^\infty}_{m-2} \left(\int_0^\infty \frac{\log(1+x_1^2+\dots+x_{m-2}^2+r^2)}{(1+1^2+\dots+x_{m-2}^2+r^2)^{(m+1/2)}}\,r\,dr \right) dx_1 \dots dx_{m-2}$$

Jelöljük

$$a = 1+x_1^2+\dots+x_{m-2}^2$$

és a

$$t = \frac{1}{(a+r^2)^{(m-1)/2}}$$

helyettesítéssel számoljuk ki a belső integrált.

$$\int_0^\infty \frac{\log(a+r^2)}{(a+r^2)^{(m+1)/2}} r dr = \frac{2}{(m-1)^2} \cdot \int_{1/a^{(m-1)/2}}^0 \log t \, dt =$$

$$= \frac{2}{(m-1)^2} \cdot \frac{1}{a^{(m-1)/2}} + \frac{1}{m-1} \cdot \frac{\log a}{a^{(m-1)/2}}.$$

Ezt (3.7)-be visszahelyettesítve:

$$J_m = \frac{\pi}{2} \underbrace{\int_0^\infty \dots \int_0^\infty}_{m-2} \left(\frac{2}{(m-1)^2} \cdot \frac{1}{a^{(m-1)/2}} + \frac{1}{m-1} \cdot \frac{\log a}{a^{(m-1)/2}} \right) dx_1 \dots dx_{m-2} =$$

$$= \frac{\pi^{(m+1)/2}}{(m-1)^2 2^{m-2} \Gamma\left(\frac{m-1}{2}\right)} + \frac{\pi^{(m+1)/2}}{(m-1) \cdot 2^{m-1} \Gamma\left(\frac{m-1}{2}\right)} \left(\psi\left(\frac{m-1}{2}\right) - \psi\left(\frac{1}{2}\right) \right) =$$

$$= \frac{\pi^{(m+1)/2}}{2^{m-1}} \cdot \frac{1}{\Gamma\left(\frac{m-1}{2}\right)} \cdot \frac{1}{(m-1)} \left(\frac{2}{m-1} + \psi\left(\frac{m-1}{2}\right) - \psi\left(\frac{1}{2}\right) \right).$$

Térjünk vissza (3.6) igazolásához, J_m értékét visszahelyettesítve:

$$2^m \frac{\Gamma\left(\frac{m-1}{2}\right)}{\pi^{(m+1)/2}} \int_0^\infty \dots \int_0^\infty \frac{\log(1+x_1^2+\dots+x_m^2)}{(1+x_1^2+\dots+x_m^2)^{(m+1)/2}} dx_1 \dots dx_m =$$

$$= 2^m \frac{\Gamma\left(\frac{m+1}{2}\right)}{\pi^{(m+1)/2}} \cdot \frac{\pi^{(m+1)/2}}{2^{m-1} \Gamma\left(\frac{m-1}{2}\right)} \cdot \frac{1}{m-1} \left(\frac{2}{m-1} + \psi\left(\frac{m-1}{2}\right) - \psi\left(\frac{1}{2}\right) \right)$$

felhasználva, hogy

$$\Gamma\left(\frac{m+1}{2}\right) = \Gamma\left(\frac{m-1}{2}\right) \cdot \frac{m-1}{2}$$

és hogy

$$\frac{2}{m-1} + \psi\left(\frac{m-1}{2}\right) = \psi\left(\frac{m+1}{2}\right)$$

kapjuk, hogy

$$2^m \frac{\Gamma\left(\frac{m+1}{2}\right)}{\pi^{(m+1)/2}} \int_0^\infty \dots \int_0^\infty \frac{\log(1+x_1^2+\dots+x_m^2)}{(1+x_1^2+\dots+x_m^2)^{(m+1)/2}} dx_1 \dots dx_m = \psi\left(\frac{m+1}{2}\right) - \psi\left(\frac{1}{2}\right),$$

amit igazolnunk kellett. ∎

A következő lemma az *általános Cauchy eloszlású* véletlen vektor entrópiáját adja.

3.3. LEMMA. Legyen ξ *egyszerű Cauchy eloszlású* véletlen n-dimenziós vektor és $\mathbf{T}$ $m\times n$-es sorrangú tetszőleges mátrix, valamint $\mathbf{c}$ tetszőleges m-dimenziós vektor. A lineárisan transzformált

$$\eta = \mathbf{T}\xi + \mathbf{c}$$

m-dimenziós véletlen vektor entrópiája:

$$H(\eta) = \frac{m+1}{2}\left(\psi\left(\frac{m+1}{2}\right) - \psi\left(\frac{1}{2}\right) + \log \pi\right) - \log \Gamma\left(\frac{m+1}{2}\right) + \frac{1}{2}\log \det (\mathbf{TT}^*)$$

Bizonyítás. A 3.1 tétel szerint, ha az $\mathbf{S}$ $m\times m$-es reguláris mátrix olyan, hogy

$$\mathbf{SS}^* = \mathbf{TT}^* \tag{3.8}$$

akkor a $\mathbf{T}\xi^{(n)}$ és az $\mathbf{S}\xi^{(m)}$ véletlen vektorok sürűségfüggvényei megegyeznek ($\xi^{(n)}$ és $\xi^{(m)}$ n, illetve m dimenziós vektorok *egyszerű Cauchy eloszlásúak*). Az 1.2 lemma szerint így η entrópiája:

$$H(\eta) = H(\xi^{(m)}) + \log |\det \mathbf{S}|,$$

azonban $\log |\det \mathbf{S}| = \frac{1}{2}\log \det (\mathbf{SS}^*)$ és (3.8) miatt ez a lemma állítását jelenti ∎

KÖVETKEZMÉNY. Jelöljük a $\mathbf{T}$ mátrix sorait $\mathbf{t}_1, \ldots, \mathbf{t}_m$ vektorokkal, akkor

$$H(\eta_i) = \log 4\pi + \log \mathbf{t}_i^2 \tag{3.9}$$

Bizonyítás. A lemmából nyilvánvaló. ∎

A lemma alapján azonnal kapjuk a *Cauchy eloszlású* véletlen vektor koordinátái közötti kölcsönös információt. Ezt fontos szerepe miatt tételként mondjuk ki.

3.2. TÉTEL. Legyen ξ *egyszerű Cauchy eloszlású* n-dimenziós véletlen vektor és $\mathbf{T}$ $m\times n$-es, sorrangú, tetszőleges mátrix ($m \geqq 2$), valamint $\mathbf{c}$ tetszőleges m-dimenziós vektor. A lineárisan transzformált

$$\eta = \mathbf{T}\xi + \mathbf{c}$$

m-dimenziós véletlen vektor koordinátái közötti kölcsönös információ

$$I(\eta) = m\log 4 + \frac{m-1}{2}\log \pi + \log \Gamma\left(\frac{m+1}{2}\right) - \frac{m+1}{2}\left(\psi\left(\frac{m+1}{2}\right) - \psi\left(\frac{1}{2}\right)\right) -$$

$$- \frac{1}{2}\log \frac{\det (\mathbf{TT}^*)}{\mathbf{t}_1^2 \ldots \mathbf{t}_m^2},$$

ahol $\mathbf{t}_1, \ldots, \mathbf{t}_m$ a $\mathbf{T}$ mátrix sorvektorai és

$$\psi(x) = \frac{d}{dx}\log \Gamma(x) \quad (x > 0)$$

Bizonyítás. Az η véletlen vektor koordinátái közötti kölcsönös információ:

$$I(\eta) = H(\eta_1) + \ldots + H(\eta_m) - H(\eta)$$

Használva (3.5) és (3.6)-ot kapjuk a tétel állítását. ∎

Megjegyezzük, hogy a *Hadamard egyenlőtlenség* szerint $I(\eta)$ akkor és csak akkor minimális, ha $\mathbf{t}_i\mathbf{t}_j^* = 0$ minden $i \neq j$ indexpárra.

KÖVETKEZMÉNY. Ha az η vektornak csak egy kétkoordinátás részvektorát tekintjük, akkor

$$I(\eta_i, \eta_j) = \log\left(\frac{8\pi}{3}\right) - \frac{1}{2}\log\left(1 - \frac{(\mathbf{t}_i\mathbf{t}_j^*)^2}{\mathbf{t}_i^2\mathbf{t}_j^2}\right) \tag{3.10}$$

Bizonyítás. A tételből, valamint (3.5)-ből nyilvánvaló. ∎

A 3.3 lemma és a 3.2 tétel következménye (3.9) és (3.10) mutatják a *Cauchy eloszlás* koncentráció mátrixának tartalmát. Azonban mint (3.10)-ből látható a koordináták páronkénti kölcsönös információja a koncentráció mátrix elemeinek csak abszolút értékét adja meg. Az előjel meghatározásához további vizsgálatok szükségesek.

IRODALOM

[1] CSISZÁR, I., „Eloszlások eltérésének információ-típusú mértékszámai I.", *MTA III. Osztály Közleményei* **17** (1967) 123—149.
[2] CSISZÁR, I. és KÖRNER, J., *Information Theory* (Academic Press, 1981).
[3] KORN, G. A. és KORN, T. M., *Matematikai Kézikönyv Műszakiaknak* (Műszaki Könyvkiadó, 1975).
[4] KULLBACK, S., *Information Theory and Statistics* (Wiley, New York, 1959).
[5] MARDIA, K. V. and KENT, J. T. and BIBBY, J. M., *Multivariate Analysis* (Academic Press, 1979).
[6] TUSNÁDY, G., A több dimenziós normális eloszlás. Megjelent: *Több változós statisztikai analízis* I. fejezet, szerk.: *Móri F. T. és Székely J. G.* (Műszaki Könyvkiadó, 1986).
[7] PRÉKOPA, A., *Valószínűségelmélet műszaki alkalmazásokkal* (Műszaki Könyvkiadó, 1972).

(Beérkezett: 1986. december 2.)

KLAFSZKY EMIL
NEHÉZIPARI MŰSZAKI EGYETEM
MISKOLC

KAS PÉTER
ELTE TTK OPERÁCIÓKUTATÁSI TSZ
1088 MÚZEUM KRT. 6—8.
BUDAPEST

REMARKS ON THE MULTIVARIATE CAUCHY DISTRIBUTION

E. KLAFSZKY and P. KAS

In the present paper a multivariate generalization of the *Cauchy distribution* is studied. Entropy and mutual informations are given in term of the parameters of the density function, which enables us to attribute a kind of correlation — measure role of these parameters.

Alkalmazott Matematikai Lapok **13** (1987—88) 163—170

SZTOCHASZTIKUS VEZÉRLÉS RÉSZLEGES MEGFIGYELÉS ALAPJÁN (SZÁMÍTÓGÉPES SZIMULÁCIÓ)

MÁRKUS LÁSZLÓ

Budapest

A dolgozatban a (2.1), (2.2) állandó együtthatós lineáris rendszerrel foglalkozunk. Számítógéppel előállítjuk diszkrét pontokban az ismert optimális vezérlést, vizsgáljuk a költséget és az állapotbecslés viszonyt. Megállapítjuk, hogy minél stabilabb a rendszer, annál kisebb a költségmegtakarítás, továbbá, hogy sem a felosztás finomítása, sem a megfigyelésben jelenlevő zaj csökkentése nem teszi lehetővé az állapot nagy pontosságú becslését, s ez utóbbi meglepő megállapítás nem tükrözi az előzetes megfontolások alapján várható viselkedést.

1. Bevezetés

A sztochasztikus differenciálegyenletek vezérléselméletének egyik alapfeladata a minimális költségű irányítás az állapot részleges ismeretében. Az optimális megoldást — mint az jól ismert [2], [3], [6], [7] lineáris rendszerre a szétválasztási elv adja, mely szerint *Kálmán szűréssel* becsüljük az állapotot, s így teljesen ismertnek tekintve a folyamatot — vezéreljük. Mi számítógéppel előállított fehér zajból generáltuk a rendszert, számítottuk a vezérlést és vizsgáltuk statisztikai tulajdonságait.

2. A feladat

Az egydimenziós esetben vizsgáljuk a következő rendszert:

$$d\theta(t) = \alpha\theta(t)\,dt + u(t)\,dt + \sigma_1\,dw_1(t), \tag{2.1}$$

$$d\xi(t) = \theta(t)\,dt + \sigma_2\,dw_2(t), \tag{2.2}$$

$$\theta(0) = \theta_0 \qquad \xi(0) = \xi_0,$$

a következő költség mellett

$$V(u) = E\left\{\chi\cdot\theta(1) + \int_0^1 \left(\varrho\cdot\theta^2(t) + u^2(t)\right)dt\right\}. \tag{2.3}$$

Itt $t\in[0, 1]$, a $\theta(t)$ állapot a $\xi(t)$ megfigyelés, az $u(t)$ vezérlés egydimenziós csak a múlttól függő véletlen folyamatok, $w_1(t)$, $w_2(t)$ független standard *Wiener-folyamatok*, θ_0, ξ_0 valószínűségi változók, σ_1, σ_2, ϱ, $\varkappa$, $\alpha\in\mathbf{R}$. A valószínűségi változó várható értékét E-vel jelöljük. $\alpha<0$, vagyis az állapot stabil, továbbá $\sigma_1, \sigma_2, \varrho>0$, $\varkappa\geqq 0$. A $V(u)$ költség- (veszteség) függvényt kívánjuk minimalizálni. Ehhez először $\theta(t)$-t

becsüljük az $m(t)$ feltételes várható értékkel $\xi(t)$ szerint, azaz ha $\mathcal{F}_t^\xi$ a ξ által generált σ-algebra, akkor

$$m(t) = E(\theta(t)|\mathcal{F}_t^\xi).$$

Így $m(t)$ kielégíti a következő sztochasztikus differenciálegyenletet [7]:

$$(2.5) \qquad dm(t) = \left(\alpha - \frac{\gamma(t)}{\sigma_2^2}\right) m(t)\,dt + \left[u(t) + \frac{\gamma(t)}{\sigma_2^2}\,\theta(t)\right] dt + \frac{\gamma(t)}{\sigma_2}\,dw_2(t),$$

$$m_0 = E(\theta_0|\xi_0),$$

ahol $\gamma(t) = E[\theta(t) - m(t)]^2$ és $\gamma(t)$ megoldása a következő *Riccati-típusú differenciálegyenletnek:*

$$(2.6) \qquad \gamma'(t) = -\frac{1}{\sigma_2^2}\gamma^2(t) + 2\alpha\gamma(t) + \sigma_1^2,$$

$$\gamma(0) = E(m_0 - \theta_0)^2.$$

A továbbiakban csak a stacionárius, ill. a 0-ból indított rendszert vizsgáljuk. $\gamma(0) = 0$ esetén:

$$\gamma(t) = \frac{\sigma_1^2}{\delta_1 \operatorname{cth}(\delta_1 t) - \alpha}, \quad \text{ahol} \quad \delta_1 = \sqrt{\alpha^2 - \frac{\sigma_1^2}{\sigma_2^2}},$$

míg a stacionárius esetben $\gamma(t)$ konstans, így

$$\gamma(t) = (\alpha + \delta_1)\sigma_2^2.$$

(A *Riccati-egyenlet* megoldását illetően l. [1], [4], [5]).

A számítógépes vizsgálathoz a folytonos idejű rendszert diszkretizálni kell, ezért oldjuk meg (2.5)-öt. A

$$\Phi(t) = e^{\varphi(t)} = \exp \int_0^t \left(\alpha - \frac{\gamma(s)}{\sigma_2^2}\right) ds$$

jelölés mellett

$$(2.7) \qquad m(t) = e^{\varphi(t)} \left\{ m_0 + \int_0^t e^{-\varphi(s)} \left[\left(u(s) + \frac{\gamma(s)}{\sigma_2^2}\,\theta(s) \right) ds + \frac{\gamma(s)}{\sigma_2}\,dw_2(s) \right] \right\}$$

(lásd [7], (4.158)). Ez alapján diszkretizálva:

$$m(t_2) = e^{\varphi(t_2) - \varphi(t_1)} m(t_1) + e^{\varphi(t_2)} \int_{t_1}^{t_2} e^{-\varphi(s)} \left[u(s) + \frac{\gamma(s)}{\sigma_2^2} \left(\theta(s)\,ds + \sigma_2\,dw_2(s) \right) \right].$$

Az integrált becsülve és (2.2)-t felhasználva:

$$(2.8) \qquad m(t_2) \approx e^{\varphi(t_2) - \varphi(t_1)} \left[m(t_1) + u(t_1)(t_2 - t_1) + \frac{\gamma(t_1)}{\sigma_2^2} \left(\xi(t_2) - \xi(t_1) \right) \right].$$

A 0-ból indított rendszerre

$$\varphi(t_2) - \varphi(t_1) \approx \left(\alpha - \frac{\gamma(t_1)}{\sigma_2^2} \right) (t_2 - t_1).$$

Míg a stacionárius rendszerre

$$m(t_2) \approx e^{-\delta_1(t_2-t_1)}[m(t_1)+u(t_1)(t_2-t_1)+(\alpha+\delta_1)(\xi(t_2)-\xi(t_1))]. \tag{2.9}$$

A becslés alapján a következő vezérlés minimalizálja a költséget ([6]):

$$u(t) = -P(t)\cdot m(t). \tag{2.10}$$

Ahol $P(t)$ a $P(1)=\varkappa$ végfeltétel melletti megoldása a

$$-P(t) = -P^2(t)+2\alpha P(t)+\varrho$$

Riccati-egyenletnek. Ez az ún. *algebrai Riccati-egyenlet* megoldását adja $\varkappa = \alpha+\sqrt{\alpha^2+\varrho}$ esetén. Ekkor

$$P(t) = P = \alpha+\sqrt{\alpha^2+\varrho}.$$

3. A számítógépes szimuláció

A munkát IBM PC XT személyi számítógépen végeztük.

Elsőként az állapotot és a megfigyelést kellett előállítanunk a következő rekurzív egyenletek alapján.

$$t_i = e^{\frac{\alpha}{n}}\left(t_{i-1}+u_{i-1}\cdot\frac{1}{n}+g_{1,i}\right) \tag{3.1}$$

$$k_i = t_{i-1}\cdot\frac{1}{n}+g_{2,i}. \tag{3.2}$$

Itt t_i, k_i, u_i az n egyenlő részre osztott [0, 1] intervallum i-edik osztópontjához tartozó szimulált állapot, megfigyelés és vezérlésérték. $g_{j,i}$-k egymástól független $N\left(0, \frac{\sigma_i}{\sqrt{n}}\right)$ eloszlású mintaelemek. A (3.1) egyenlet (2.1) pontos megfelelője a diszkrét esetben (lásd [1], [7]). A $g_{j,i}$-k előállítására a gép véletlen szám generátorát használtuk, mely [0, 1]-beli értékeket állít elő. Ezekből 12-t összeadva és normalizálva a kapott mintákat független standard normális eloszlásúnak tekintettük. Ezekből állítottuk elő $g_{j,i}$ értékeit, illetve stacionárius indítás esetén az $N\left(0, \frac{\sigma_1}{\sqrt{2\alpha}}\right)$ eloszlású t_0 változót [1]. Az indításnak megfelelően (2.8), ill. (2.9) alapján becsültük az állapotot és (2.10)-ből számítottuk a vezérlést.

A kiértékeléshez a $g_{j,i}$ mintaelemeket tárolva ugyanazon zajból a 0 vezérlés mellett is generáltuk a rendszert, számítottuk a költséget és összehasonlítottuk. Egy másik lehetőség volt az állapotbecslés eltérés kiszámítása. Ezt megadtuk a $\frac{t_i-m_i}{t_i}\cdot 100\%$ hibaszázalékkal, továbbá összehasonlítottuk a $(t_i-m_i)^2$ négyzetes hibát az elméleti $\gamma(t)$ értékkel.

(3.1) (3.2)-t felfoghatjuk diszkrét idejű rendszerként is, és így vezérelhetjük a [7], (14.64)—(14.66), (14.59), (14.60) egyenletei alapján is. Ezek esetünkre némi

számolás után a következő rekurzív egyenleteket adják az előző pont jelöléseivel:

$$m_{i+1} = e^{\frac{\alpha}{n}}\left[\frac{u_i}{n} + \left(1 + \frac{\gamma_i}{n\sigma_2^2}\right)^{-1}\left(m_i + \frac{\gamma_i}{\sigma_2^2}\cdot k_{i+1}\right)\right], \tag{3.3}$$

$$u_{i+1} = \frac{\varrho - P_i}{n}\cdot m_{i+1}, \tag{3.4}$$

$$\gamma_{i+1} = \left[e^{-\frac{2\alpha}{n}} + \frac{\gamma_i}{n\sigma_2^2}\right]^{-1}\cdot\gamma_i + \frac{\sigma_1^2}{n}, \tag{3.5}$$

$$P_{i-1} = \left[e^{-\frac{2\alpha}{n}} + \frac{P_i}{n^2}\right]^{-1}\cdot P_i + \varrho. \tag{3.6}$$

Ez alapján lehetőség nyílt diszkrét rendszerként vezérelni a folyamatokat, és a kapott becsléseket és költségeket összehasonlítani a folytonos idejű rendszerre kapottakkal.

A rendszer paramétereinek az α, ϱ, $\frac{\sigma_1}{\sigma_2}$, n értékeit tekintettük és ezeket 10-es nagyságrendenként változtattuk.

4. A tapasztalatok

Rendszerünk paraméterei (α, $\frac{\sigma_1}{\sigma_2}$, ϱ, n) egyenként is 4-5 olyan nagyságrendileg különböző értéket vehetnek fel, amelyek gyakorlati szempontból érdekesek. Így összefüggések megállapítása még a nagyszámú (több száz) futtatás ellenére sem könnyű, és a tévedés lehetőségét is magában rejti.

A rendszert mindvégig a [0, 1] időintervallumban vizsgáltuk. Tíz program-

1.a TÁBLÁZAT

A (2.3) költség értéke és aránya a vezérlés nélküli költség százalékában folytonos vezérlés mellett

σ_1/σ_2	n	150					1500				
	ϱ α	−0,01	−0,1	−1	−10	−100	−0,01	−0,1	−1	−10	−100
10	1	284,5 88%	361,3 86%	26,2 95%	5,29 100%	0,55 100%	83,4 85,7%	397,1 92,2%	68,3 92,8%	3,09 100%	0,80 100%
	100	2779 38%	2768 38%	1634 45%	873 94%	47,26 100%	5085 16%	2442 45%	3764 55%	299,1 101%	103,1 100%
100	1	5042 81%	339,7 84%	21,9 92%	4,69 99,8%	1,24 100%	187,0 87,4%	48,44 96,5%	102,14 90,2%	17,92 99,8%	0,486 100%
	100	1964 38%	4536 26%	1449 131%	405,2 87%	79,2 100%	3228 53,7%	1151 60,4%	1507 81,8%	1693 16,3%	53,3 100%

futtatás közül kiválasztottuk a legjellemzőbbet és ennek adatait közöljük táblázatainkban.

Az 1.a táblázat mutatja, hogy α értékét csökkentve jelentősen csökken a költség (felső sor), összhangban azzal, hogy kisebb α-hoz kisebb szórású, stabilabb állapot tartozik. Ugyanakkor a vezérléssel elérhető költségmegtakarítás aránya is kisebb, ezt mutatják az alsó sorban álló értékek. Ezek százalékban adják meg a vezérelt rendszer költségének arányát a vezérlés nélküli $(u(t)\equiv 0)$ rendszeréhez képest (mindkét rendszert ugyanazon $g_{j,i}$ zajból generáltuk!). Ugyanez a jelenség figyelhető meg a diszkrét elven vezérelt rendszerben is. (1b táblázat.) Ugyancsak leolvashatjuk, hogy ha a költségben az állapot nagyobb súllyal szerepel ($\varrho=100$), akkor jelentősebb költségmegtakarítás érhető el.

1.b TÁBLÁZAT

A (2.3) költség értéke és aránya a vezérlés nélküli költség százalékában diszkrét vezérlés mellett

σ_1/σ_2	n		150									
	ϱ	α	−0,01		−0,1		−1		−10		−100	
10	1		41,96	96%	53,07	254%	23,93	365%	11,24	99,9%	5,42	100%
	100		2342	39%	2298	24%	1387	42%	569,4	97%	511,4	100%
100	1		15,7	94%	64,3	84%	18,8	96%	4,91	99,9%	0,85	100%
	100		969	14%	635	33%	584	47%	316	94%	108	99,9%

2. TÁBLÁZAT

A költség, a megtakarítás és a becslés pontosság változása a felosztás finomításával folytonos rendszere

σ_1/σ_2	α		−0,1				−1			
	n		15	150	1500	15000	15	150	1500	15000
10	Költség kts. megtak.		117,9 108%	207,9 96%	632,7 92%	65,9 89%	61,9 95%	5,05 113%	28,42 107%	4,15 99%
	Elm. szórás[2]		9,90				9,05			
	Stat. szórás[2] Hiba %		27,0 64%	9,83 20%	29,07 6%	4,021 5%	18,9 29%	10,05 203%	16,5 65%	2,25 55%
100	Költség		18,92* 99%	45,9 113%	309,7 89%	67,32 87%	147* 99,8%	66,7 91%	102,1 90,2%	101,1 90,7%
	Elm. szórás[2]		0,999				0,999			
	Stat. szórás[2] Hiba %		18,59* 99,1%	1,17 11%	0,872 5%	0,951 9%	144,56* 99%	6,59 30%	0,76 5%	0,68 1%

A 2. táblázat szerint a felosztás finomításával a becslés átlagos hibaszázaléka (amelyet a

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|t_i-m_i|}{t_i}\cdot 100\% \tag{4.1}$$

képlet szerint számítottunk) jelentősen javul, de ezt csak jóval szerényebb mértékben követi a költségmegtakarítás javulása. Figyelemre méltó azonban, hogy a program futási ideje n növelésével egyenes arányban nő.

A $\gamma(t)$ függvény és a $(t_i-m_i)^2$ tapasztalati szórásnégyzet osztáspontonkénti kiszámítása során úgy találtuk, hogy viszonyukat a gyors stacionarizálódás miatt jól jellemzi az időátlag

$$\frac{1}{n}\sum_{i=1}^{n}(t_i-m_i)^2 \tag{4.2}$$

és $\gamma(1)$ viszonya ($\gamma(t)$ közel konstans). Ezért a függvények teljes menete helyett e két számot és az ehhez járuló (4.1) átlagos hibaszázalékot adjuk meg a becslés pontosságának jellemzésére a 3. és 4. táblázatban.

Ezzel szándékoztunk „ellenőrizni" a szimulációt, és kedvezően értékelhettük, hogy általában a tapasztalat jól egyezik az elmélettel. Ugyanakkor érdekes, bár nem meglepő, hogy néha kis szórásértékhez is nagy hibaszázalék járul. Magyarázatra szorul viszont az állapot rossz becslése, ha $\sigma_1/\sigma_2 \approx 10n$. Vegyük észre, hogy ekkor a (2.9)-ben szereplő kitevő -10 körüli, így az integrálközelítésből adódó hibára

3. TÁBLÁZAT

A statisztikai és az elméleti szórásnégyzet a hozzájáruló átlagos hibaszázalékkal a folytonos vezérlés mellett

n	σ_1/σ_2 / α	10					100				
		−0,01	−0,1	−1	−10	−100	−0,01	−0,1	−1	−10	−100
Elméleti szórás²		9,99	9,90	9,05	4,14	0,499	1,000	0,999	0,990	0,905	0,414
15	Stat. szórás²	28,92	11,14	9,76	21,69	0,65	18,9*	14,3	48,1	1,51	1,81
	Átlagos hiba %	40%	31%	87%	84%	99,3%	99,2%*	99,2%	99,1%	98,9%	97,2%
150	Stat. szórás²	10,17	8,81	9,03	4,85	0,32	5,26	0,919	3,35	1,09	0,38
	Átlagos hiba %	18%	44%	21%	87%	94%	38%	29%	24%	61%	95%
1500	Stat. szórás²	11,99	6,76	8,14	1,83	0,58	1,31	1,16	0,90	1,62	0,35
	Átlagos hiba %	67%	16%	23%	54%	102%	5%	11%	20%	47%	44%

4. TÁBLÁZAT

A költség és aránya a vezérlés nélkülihez, valamint a becslés pontossága a jel-zaj viszony függvényében

α	$n=1500$ σ_1/σ_2	Folytonos vezérlés 1	10	100	1000	Diszkrét vezérlés 1	10	100	1000
	Költség	225,4 116%	405 93,2%	210 89,3%	179 81,0%	58,1 99,9%	8,68 93%	112 85%	127 80%
−0,01	Elm. sz.	99,0	9,99	1,00	0,100	61,3	10,3	1,38	0,681
	Stat. szórás[a] Hiba	178 79%	31,0 53%	1,42 14%	10,7 29%	75,4 134%	6,57 81%	0,39 3%	0,011 1%
	Költség Megt.	13,3 99,1%	301 94%	45,2 95%	16,1 93,5%	29,68 98%	21,2 93%	12,6 91%	20,6 94%
−1	Elm. sz.[a]	41,42	9,0	0,99	0,100	38,8	9,36	1,37	0,681
	Stat. szórás[a] Hiba	14,3 85%	38,1 23%	1,56 28,7%	1,507 30%	22,3 83%	4,92 16%	0,516 13%	0,014 2%

rendkívül érzékeny a rendszer. A fentiekből érhetően $\sigma_1/\sigma_2 \approx 100n$ teljesen irreális eredményt ad, sőt itt már a gép véges pontosságát is figyelembe kell venni. Megállapíthatjuk tehát, hogy a jó vezérléshez a felosztást a jel-zaj viszonynál $\left(\frac{\sigma_1}{\sigma_2}\right)$ nem választhatjuk durvábbra. (*-gal megjelöltük azokat az eseteket, ahol mégis ez áll.)

A 4. táblázatban összehasonlítjuk a folytonos és a diszkrét vezérlést a jel-zaj viszony növelésének függvényében. Meglepve tapasztaltuk, hogy bár a vezérlés körülbelül ugyanolyan arányban csökkenti a költséget mindkét esetben, ez a diszkrét rendszer állapotának jóval pontosabb becslése mellett valósul meg. Míg az elméleti hibaszórások közel azonosak, addig a tapasztalati szórásnégyzet jóval kisebb a diszkrét esetben, és jóval nagyobb a folytonos rendszerre. Ez azt mutatja, hogy a folytonos rendszerben fellép valamilyen ki nem szűrt zaj is, melynek eredete valószínűleg a rendszer diszkretizálása. A jelenség pontos elméleti magyarázatát azonban nem ismerjük, ez későbbi vizsgálat tárgya lehet.

Végezetül ezúton szeretnék köszönetet mondani ARATÓ MÁTYÁS professzornak, aki a probléma ilyen irányú megközelítésére felhívta figyelmemet, munkámat mindvégig figyelemmel kísérte és segítette hasznos megjegyzéseivel.

IRODALOM

[1] ARATÓ, M., *Linear Stochastic Systems with Constant Coefficients* (Springer-Verlag, 1982).
[2] ARNOLD, L., *Sztochasztikus differenciálegyenletek* (Műszaki Könyvkiadó, 1984).
[3] FLEMING, W. H., RISHEL, R. W., *Deterministic and Stochastic Optimal Control* (Springer-Verlag, 1975).
[4] KUCERA, V., "A review of the matrix Riccati equation", *Kybernetika* 9 (1973) 43—61.

[5] REID, W., *Riccati Differential Equations* (Academic Press, New York, 1972).
[6] KISS, A., „A szeparációs elv alkalmazása lineáris optimalizálási feladatban színes zaj esetén *Alkalmazott Matematikai Lapok*;
[7] Липцер, Р. Ш., Ширяев, А. Н., Статистика случайных процессов (Наука, Москва, 1974.)

(Beérkezett: 1986. augusztus 21.)
(Átdolgozva beérkezett: 1986. szeptember 24.)

MÁRKUS LÁSZLÓ
SZÁMALK
1115 BUDAPEST, SZAKASITS Á. U. 68.

PARTIALLY OBSERVABLE SYSTEM UNDER STOCHASTIC CONTROL
(Simulation on a personal computer)

L. MÁRKUS

At discrete points we compute the state estimation, the control and the cost function of the continuous, partially observable system (2.1), (2.2). We have found that the more stable the system is, the less cost saving can be achieved and, surprisingly, lower noise during observation does not result in much better state estimation.

Alkalmazott Matematikai Lapok 13 (1987—89) 171—185

A KÜLFÖLDI SZAKIRODALOMBÓL

ZAJOS RENDSZEREK IDENTIFIKÁCIÓJA[1]

R. E. KALMAN

(Swiss Federal Institute of Technology, Zürich University of Florida, Gainesville, Florida)

Az elmúlt huszonöt évben számos tudományos és személyes kapcsolatom volt a *Sztyeklov Intézettel.* Nagy öröm és megtiszteltetes számomra, hogy itt lehetek és közreműködhetek az Önök 50. évfordulós ünnepélyén. Hozzá szeretném tenni, hogy előadásom inkább a következő ötven évvel fog foglalkozni, mint az elmúltakkal.

Mivel a *Szovjetunióban, Moszkvában* a *Sztyeklov Intézetben* beszélek, bevallom, hogy mostani előadásom motivációja — éppúgy, mint az utóbbi időben gondolataim többsége — visszanyúlik PONTRJAGIN professzor úr egy véletlen megjegyzéséhez, amit 1969 októberében, stanfordi látogatásakor mondott. Azt hiszem, annyira jól emlékszem az alkalomra, hogy szó szerint tudom őt idézni:

„A matematikusok nem hisznek a véletlenben."

Bár akkor nem értettem pontosan, sohasem felejtettem el ezt a mondást. Azonnal meg kellett volna kérdeznem magamtól, hogy ez igaz-e, és észre kellett volna vennem, hogy ez egy kutatási tervet sugall. Kétségtelenül fontos lenne tudni, hogy PONTRJAGIN személyes véleményét mondta-e, vagy pedig úgy kell tekinteni az állítását, mint egy nyitott tudományos kérdést.

Sohasem beszéltem erről PONTRJAGINnal, de szeretném itt megköszönni Neki, hogy — bár nem tudatosan —, de ösztönzést adott arra, hogy megfogalmazhassuk napjaink talán legfontosabb tudományos problémáját, amely magában foglalja a (tiszta) matematika és a való világ viszonyát.

1. Rendszerelméleti tények rövid áttekintése

Mivel főleg új problémákkal szeretnék foglalkozni, ezért nem fogom Önöket a rendszerelmélet matematikai apparátusának technika részleteivel untatni. Megpróbálom az eredményeket és ötleteket úgy megfogalmazni, hogy a megmaradó pontatlanságok ne legyenek túlságosan kellemetlenek. A referenciákban meg lehet találni a szükséges technikai részleteket.

Bizonyára meg fog lepni néhány matematikust, hogy a következő probléma vizsgálata,

adatok → rendszer (az adatokat magyarázó modell),

amelyet úgy kell tekintenünk, mint szinte minden tudományos terület alapvető feladatát; komoly és érdekes matematikai problémákhoz vezet. A fenti probléma vizsgálatának nyilván fontos kapcsolata van a klasszikus *(Kolmogorov-féle)* valószínűségszámítással is. Itt a „valószínűség" szót úgy szereném használni, mint az ideális vagy közel-ideális mechanizmusok (kártyajáték, kockadobás, rulett) szokásos axiomatizálása.

[1] Ez a dolgozat a szerzőnek a *Szovjet Tudományos Akadémia Sztyeklov Matematikai Intézetének* 1984. szeptember 26-án megtartott 50 éves jubileumi szimpóziumán elhangzott előadásának módosított szövegét tartalmazza. A dolgozat eredetije oroszul az Успехи Математических Наук **40** (1985) 27—41. oldalain jelent meg, a fordítás közléséhez a szerző hozzájárult.

Ennek a problémának klasszikus rendszerelméleti megközelítése a következőképpen fejezhető ki.

(1.1) *EGYÉRTELMŰSÉGI ELV.* Ha az adatok

(i) pontosak,
(ii) teljesek,

akkor pontosan egy olyan minimális rendszer (matematikai modell) létezik, amely reprodukálja az adatokat. Más szóval, az adatok minimális magyarázatai izomorf modelleknek felelnek meg.

Ez az elv annyira általános, hogy pillanatnyilag nem lehet tételként kimondani, hiszen akkor definiálnunk kellene az „adat", „rendszer" stb. általános kifejezéseket. Ha ezeket a fogalmakat már pontosan definiáltuk (l. később), valóban olyan tételek sokaságát kapjuk, amelyek megfogalmazása pontosan definiálni fogja azt is, mit értünk azon, hogy „minimális", „izomorf" stb. Semmilyen érdemleges ellenpélda nem ismeretes jelenleg, amely korlátozni látszana az *Egyértelműségi Elv* egyetemességét.

A rendszert (modellt), amely reprodukálja az adatokat, gyakran *realizációnak* hívjuk. Egy ilyen rendszert úgy kell tekintenünk, mint konkrét, explicit leírása annak a lehetséges mechanizmusnak, amely véleményünk szerint az adatokat generálta. A realizációs *tétel* egy olyan matematikai eredmény, amely megmutatja, hogy hogyan is kell tulajdonképpen megkonstruálni egy ilyen realizációt az adatokból.

1. Példa: Valószínűleg ez a legrégebbi nemtriviális, történelmileg érdekes realizációs probléma. Adatok: valós számpárok (x_t, y_t), $t=1, \ldots, N$. A gyanított mechanizmus, amely ezeket az adatokat generálja: π polinom, amelyre

$$\pi(x_t) = y_t \tag{1.2}$$

minden t-re.

A polinom pontosan akkor minimális, ha a legegyszerűbb polinom ezzel a tulajdonsággal, azaz pontosan akkor, ha $\deg \pi =$ minimális. Erre a problémára a klasszikus realizációs tétel *Lagrange-féle interpolációs formula* néven ismeretes, nevezetesen

$$\pi_L(z) := \sum_t y_t \prod_{s \neq t} \frac{z - x_s}{x_t - x_s}. \tag{1.3}$$

Mivel π_L eleget tesz (1.2)-nek, a formula nyilván az adatok egy realizációját adja. Az a tény, hogy ez minimális realizáció, nincs kihangsúlyozva numerikus analízis könyvekben. Valószínűleg azért nincs, mert mindeddig igen csekély hatással volt a rendszerelmélet erre a területre. (A bizonyítás megtalálható KALMAN [1982]). Megjegyzem, hogy ebben az esetben a „teljesség" lényegtelen, hiszen tetszőleges N-re az adatok teljesnek tekinthetők. Az adatok pontossága azonban fontos (l. később). Az „izomorfizmus" szintén triviális, hiszen ha a probléma követelményeiben lerögzítjük π_L főegyütthatóját; pontosan egy minimális π létezik. Tehát ebben az esetben az *Egyértelműségi Elv* egy világos tétel.

2. Példa. Egy Σ állandó együtthatós dinamikus rendszert tökéletesen meg lehet adni $Z(z)$ átmeneti függvénye segítségével, amely egy valódi racionális mátrix. Ha Z-t „adat"-ként tekintjük, akkor az *Egyértelműségi Elv* újra egy tétel. A realizáció egy Σ rendszer lesz, amely n állandó együtthatós, elsőrendű, lineáris differenciálegyenlet *(= konstans; véges dimenziós lineáris rendszer)*. „Minimális" azt jelenti,

hogy n minimális, vagy ami ugyanaz: az állapottér dimenziója minimális. Ebben a formában a tétel először KALMAN [1962, 63]-ban jelent meg. Szorosan kapcsolódó eredményeket közölt véges automatákkal kapcsolatban SCHUTZENBERGER legalább egy évvel korábban.

Érdemes megjegyezni azt a tényt, hogy a következő fejezetbeli absztrakt (ellenőrizhetetlen) „minimális" feltétel ekvivalensnek bizonyult a következő konkrét és igazolható feltétellel: *egy Σ rendszer pontosan akkor minimális, ha teljesen megfigyelhető és teljesen irányítható.* (Hozzá szeretném tenni, hogy a megfigyelhetőségi feltétel fontosságát PONTRJAGIN és GAMKRELIDZE fedezték fel az optimális kontrol területén végzett kutatásaik során, egymástól függetlenül. A megfigyelhetőségi feltételt — amelyet dualitási megfontolások sugalltak az irányíthatóság megfelelőjeként —, Moszkvában, 1960-ban én vezettem be elsőként. (L. KALMAN [1960].) *Mindkét* feltétel — a teljes irányíthatóság és a teljes megfigyelhetőség — szükséges a minimalitás igazolásához. Körülbelül két évbe telt, 1960-tól 1962-ig, mire tisztáztuk ezeket a ma már klasszikusnak számító összefüggéseket.)

2. A Példa. Semmi lényeges változás nem történik, ha az előző példában megengedjük, hogy a (véges-dimenziós, lineáris) rendszernek időben változó együtthatói legyenek. Tény, hogy az egyértelműségi tételt ezen az általánosabb szinten bizonyítottam be első ízben. L. KALMAN, FARB,, ARBIB [1969. 10. Fej., c. Függelék].

Sem a „lineáris", sem a „véges-dimenziós" nem korlátai az *Egyértelműségi Elvnek.* Az algebrai geometria klasszikus fogalmait használva SONTAG és ROUCHALEAU [1976] bebizonyították, hogy az *Elv* polinomiális, diszkrét idejű nemlineáris rendszerek esetében is tétel. YAMAMOTO [1981] bebizonyította, hogy az „irányíthatóság" és „megfigyelhetőség" fogalmak alkalmas újradefiniálásával az *Elv* ismét tétel végtelen dimenziós lineáris rendszerek nagy osztálya esetén. Ezeknek az eredményeknek az általánossága bizonyára meglepő azoknak, akik alkalmazott matematikai területen dolgoznak. Kevésbé meglepő ez a (tiszta) matematika területén dolgozóknak, hiszen az egyértelműségi tétel majdnem igaz a kategóriaelméletben. A téma részleteiről l. KALMAN [1976].

Mivel a rendszerelméleti meggondolások egyformán jól alkalmazhatók a fizikai, biológiai, gazdasági vagy absztrakt területeken, természetes dolog kapcsolatot keresni a rendszerelmélet és a klasszikus fizika között. Ezeket a vizsgálatokat WILLEMS [1979] kezdte. Következő példánk az ő dolgozatából való.

3. Példa. A gravitációról szóló *Newton-törvény* igen jól ismert, de ennek születését általában rosszul mondják el bevezető fizika könyvek. A következő okoskodás segítségével az olvasó megértheti NEWTON nagyszerű felfedézésének matematikai hátterét:

Tekintsük adatként a *Kepler-„törvényeket":* minden bolygó elliptikus pályán kering a *Nap* körül, melynek egyik fókuszában a *Nap* van; továbbá teljesülnek a keringési sebességre vonatkozó jól ismert addíciós szabályok.

A *Kepler-törvények,* természetesen, *nem* törvények, hiszen nem pontosan igazak. Feltehetjük azonban, hogy pontosak, hiszen NEWTON is ezt tette. Ezzel a realizációs problémában definiáltuk a pontos adatokat.

NEWTON fő eredménye abból áll, hogy a gravitációs erőről szóló fordítottan-négyzetes törvénye olyan ellpitikus mozgásokat eredményeznek, amelyek kielégítik a *Kepler-törvényeket.* Más szóval, a kéttest probléma megoldásával NEWTON megmutatta, hogy gravitációs törvénye egy olyan modell, amely a *Kepler-törvényekben* összefoglalt tapasztalati tények egy realizációja.

Mit mondhatunk az egyértelműségről? Nyilvánvalóan NEWTONT is érdekelte ez a kérdés, mert úgy tűnik, hogy foglalkozott azzal a gondolattal, hogy a „fordítottan-négyzetes"-beli 2-t $2+\varepsilon$-ra kicserélje. Arra az eredményre jutott, hogy ε-nak nullának kell lennie. NEWTON szerencsés volt (valamint okos), hiszen fordítottan négyzetes törvénye a *Kepler-ellipszisnek* minimális magyarázatát adta, és így az *Egyértelműségi Elv* szerint várhatóan ez az egyetlen lehetséges magyarázat. WILLEMS [1979] bebizonyította, hogy itt az *Egyértelműségi Elv* ismét tétel.

NEWTON gravitációs formulája azonban csak azért válhatott törvénnyé, mert a minimális realizáció *egyértelmű* (így nyilván elegendő volt csak ezt megkeresni). Tehát ez a törvény egy egyszerű matematikai ténynek — nevezetesen az *Egyértelműségi Tételnek* — a következménye. (Persze itt az (1.1)-ben említett „izomorfizmus" tetszőleges fizikai objektum stb. választását megengedi). Ez a kijelentés nem annyira triviális, mint aminek látszik; emlékeztetjük az olvasót arra, hogy NEWTON nem örült túlságosan annak, hogy a gravitáció a távolságtól függ, mivel ezt nem tudta az általános fizikai szemlélettel összeegyeztetni. Ez még nekünk is problémát okoz. Az igazat megvallva, NEWTON felfedezése csak matematikai oldalról nézve világos és átfogó érvényű, az alapul szolgáló fizikai probléma (ti. mi okozza a gravitációt) még mindig jórészt megoldatlan.

NEWTON másképp is szerencsés volt (és nem csak okos). Azért volt szerencsés, mert KEPLER szerencsés volt. KEPLER nem tudta volna megfogalmazni „törvényeit", ha lett volna egy tucat *Nap* körül keringő, *Jupiter méretű bolygó,* mert akkor a pályák már nem lettek volna ellipszisek. Tény, hogy sok szempontból még mindig megoldatlan $n>2$-re az n-test probléma. KEPLER szerencsés lépése (és zsenialitása) volt a pontatlan mérési eredményeket pontos ellipszisekkel kicserélni (legalábbis, ő úgy hitte, hogy pontosak), és ezzel elindította a nyugati tudomány egyik legnagyobb sikertörténetét.

A mostani előadás keretében úgy tekintjük KEPLER felfedezését, melyben a pontatlan (zajos) adatokat pontosakkal helyettesítette, mint az a lépés, amellyel a matematika hatalmát használatba vettük az *Egyértelműségi Elven* keresztül. Most általában arra vagyunk kíváncsiak: Mi a teendő, ha az *Egyértelműségi Elv* nem alkalmazható, mert az adatok pontatlanok?

2. Zajos adatok

A II. Világháború óta a fizikusok, mérnökök és a gyakorlati tudományok területén dolgozók nagy része között általános megállapodás jött létre: a pontatlan adatokat zajosnak tekintjük. Zaj alatt sok mindent érthetünk: bizonytalanságot, mérési pontatlanságot, ismeretlen hatásokat, véletlen hatásokat, nemlineáris hatásokat (ha lineáris problémákkal foglalkozunk), és általában bármilyen eltérést a pontos esettől. Fontos észben tartani, hogy a zajt úgy kell felfogni, mint egy nem modellezhető mennyiséget. Ha valami modellezhető, és így előrejelezhető, akkor az adatokból felépített rendszer részének tekintjük és nem zajnak.

Zajos adatokról csak egy igazán általános állítást mondhatunk:

(2.1) *BIZONYTALANSÁGI ELV:* Pontatlan (bizonytalan) adatok$\Rightarrow$

$\Rightarrow$ nem egyértelmű (bizonytalan) rendszer.

Azt hiszem nem kell részletezni ezt a kijelentést. (Azért használtam a „bizonytalansági elv" kifejezést, mert hasonló helyzetben vagyunk, mint a fizikában.)

A zajos-realizációelmélet feladata szintén nyilvánvaló: hogyan lehet az adatokból származó bizonytalanságot „átvinni" az adatokból felépített rendszer bizonytalanságába. Semmiféle olyan számítást nem tudunk elvégezni, amellyel az adatok bizonytalanságának számszerű leírását kapnánk.

Sajnos ez a valóságos nehézség gyakran egy egyszerű pszichologiai nehézséggel találkozik, nevezetesen az emberi idegenkedés azoktól a problémáktól, amelyek nem adnak egyértelmű megoldást.

Kialakult valahogy egy filozófia a bizonytalan adatok kezelésére. Úgy tűnik, ez a filozófia mindkét nehézséggel megbirkózik. Statisztikus körökben általánosan elfogadott ez a megközelítés, de kevésbé elfogadott azon komoly emberek számára, akik ezen a területen kívül dolgoznak. Lényege a következő:

(i) Tegyük fel, hogy egy mereven lefixált absztrakt valószínűségi mechanizmus (gyakran igen egyszerű típusú) generálja az összes bizonytalanságot.

(ii) Tegyük fel, hogy az adatok egy absztrakt valószínűségi mechanizmus által irányított, fix egyensúlyi állapotban levő populáció független mintái.

Még konvencionálisabb megfogalmazása ennek a következő: az adatok egy fix, valószínűségi törvényeket teljesítő, végtelen populáció véges, független megfigyelései. Ezt úgy fogjuk hívni, hogy általános statisztikai előfeltevés.

Klasszikus statisztikusok az adatokat mintának tekintik (anélkül, hogy gondolkoznának ezen). Úgy tekintik az adatok bizonytalanságát, mint a véges N mintaelemszám következményét. A standard statisztikai előfeltevést különösen erőteljesen támogatta az 1920-as években R. A. FISCHER, akit úgy tekintenek, mint a „kis mintás statisztika legelső szakértője a világon". FISCHER „mintavételi modell'-je jónéhány statisztikai problémára alkalmazható, de más területek kutatói számára nagyrészt elfogadhatatlan. Ez abból a tényből látható, hogy manapság az N elemű minta fogalma a zaj-mérő apparátusok többségében egyáltalán nem játszik szerepet. Mérési hibák bárhol előfordulhatnak, de ezek a hatások valószínűleg kicsik. Durva hiba azt állítani, hogy minden adat mintavétel, vagy hogy minden bizonytalanság a mintavételezésből fakad.

Tehát a klasszikus statisztikus előfeltevéssel dolgozik (olyan a priori feltevést tesz, aminek nincs vagy csak igen kicsi a tudományos alapja), mikor az adatokat a „mintavételezési modell" alapján vizsgálja. Minden deduktív következtetés ennek a munkának irreleváns, ha az előfeltevés hamis abban a szituációban, amire ezt alkalmazzák.

Még tágabb értelemben az előfeltevés a zajos identifikáció összes jelenlegi eljárásának átható problémája. Például a „legkisebb négyzetes" eljárás is beleesik ebbe a csapdába még egy hozzávett (és így előzetes) statisztikai feltevés nélkül is. Ez abból a tényből is látható, hogy a legkisebb négyzetes módszer egyértelmű megoldást ad arra a feladatra, hogy a zajos adatokra egyenest illesztünk. Ez a megoldás nem lehet korrekt, mert megsérti az *Egyértelműségi Elvet*. A pontos adatok esetében minden pontnak pontosan az egyenesre kell esnie. Ekkor az egyenes egyértelműsége — amely az egyetemesen érvényes *Egyértelműségi Elv* következménye — nyilvánvalóan egy igaz, bár elég triviális matematikai tény. Hogy az előfeltevés szerepét jobban megérthessük, a következő fejezetben egy speciális helyzetet fogunk megvizsgálni. A már ismert identifikációs eljárásokban olyan feltételeket fogunk keresni, amelyeket előfeltevéseknek tekinthetünk. Ez gyakran messze nem triviális. Általában, az előfeltevés nélküli identifikáció lehetőségét fogjuk vizsgálni.

Megjegyezzük még, hogy a klasszikus *(Kolmogorov-féle)* valószínűségi meg-

közelítés valóságos adatokon *képtelen* a bizonytalansági problémával foglalkozni. A gyakorlati problémák többségében nem lehet ilyen információt identifikálni az elérhető adatokból. Jól ismert, hogy eldöntetlen kérdés, vajon egy hosszú számsorozat tekinthető-e véletlen sorozatnak. Érdekes terület a véletlennel önmagában foglalkozni, de túl szegény tudományos eszköz a zajos adatok kezelésére. Végül is, főleg a rendszerek érdekelnek bennünket — amely bizonytalan lehet, de remélhetőleg nem véletlen — nem pedig a zaj. Manapság a statisztikában épp ellentétes a tendencia. Egyetértek KOLMOGOROVval, hogy „valami nincs rendben a statisztikában".

3. Altér identifikációja

Nézzünk még egy problémát, ami valóban igen triviális pontos adatok esetén.

4. Példa. Tekintsünk egy fix r-dimenziós q kodimenziós H alteret $\mathbf{R}^n$-ben, ahol $r+q=n$. Tegyük fel, hogy az adatok olyan $\mathbf{R}^n$-beli pontok, $\{x_t\}$, $t=1, \ldots, N$, amelyek (pontosan) H-ban vannak. Ha ennek a halmaznak a rangja r, akkor az adatok teljesek. Az *Egyértelműségi Elv* alapján azt várjuk, hogy H-t egyértelműen identifikálni lehet x_t-ből. Természetesen ez egy tétel, de egy kicsit többet fogunk állítani.

Egy alteret kétféleképpen adhatunk meg: (1) generátorok vagy (2) lineáris relációk segítségével. Ha kiválasztunk r független H-beli x_t pontot mint generátorokat, akkor az adatokat — megfelelő koordinátarendszerben — a következő mátrix írja le:

$$\begin{pmatrix} I_r \\ G \end{pmatrix}, \tag{3.1}$$

ahol I_r $r\times r$-dimenziós egységmátrix, és az $m\times r$-dimenziós G mátrix elemeit úgy tekinthetjük, mint H *Plücker* vagy *Grassmann koordinátái*. Szokásos megközelítése az identifikációs problémának, hogy olyan kovektorokat kell az ezekből az adatokból konstruálni, amelyek H-t a sokaságban lineáris relációk segítségével írják le. Erre a konstrukcióra szokás úgy hivatkozni, mint *„duális Grassmann koordináták"*. (L. HODGE és PEDOE [1968. VII. 3. Fej. 292. old.].) A (3.1) kanonikus alakból ez a konstrukció triviális, hiszen ezt a következőképpen adhatjuk meg:

$$(G-I_m). \tag{3.2}$$

Tehát a G mátrix elemeit szintén meg lehet kapni a *duális Grassmann koordináták* segítségével. Magától értetődik, hogy az „egyértelműség"-et itt olyan értelemben kell tekintenünk, hogy H a *Grassmann sokaság* pontjain fut végig.

A példában szereplő feladat az egyetlen, melyre részletes irodalmat találhatunk a zajos identifikációban. Most kritikusan megvizsgáljuk az elérhető eredményeket.

Csak additív zajjal fognak foglalkozni. Tegyük fel az egyszerűség kedvéért, hogy az adatok átlaga $\sum_t x_t=0$. Tehát a zaj fogalmának matematikai jelentést adunk, ha közvetetten így definiáljuk:

$$x_t =: \hat{x}_t + \tilde{x}_t, \tag{3.3}$$

ahol $\hat{x}_t$ a *pontos* vagy *igazi* komponense x_t-nek, míg $\tilde{x}_t$ a zaj. Pontos adatok esetén $\tilde{x}_t=0$ minden t-re és $\hat{x}_t$ H-ban fekszik. Megjegyezzük, hogy (3.3)-ban a jobb oldalt (ami nem egyértelmű), a bal oldal segítségével definiáltuk.

Adott, fix H esetén az $\{x_t\}$ zajos adatokat úgy tekinthetjük, mint egy H köré csoportosuló felhőt. Ha $\tilde{x}_t$ minden t-re elég kicsi (ez a *gyenge*-zaj eset), nincs rá intuitive magyarázat, hogy miért ne lehetne ez igaz. Az elméleti probléma azonban nem könnyű, mert „ad hoc" szabályok helyett általános eredmények érdekelnek bennünket. Minden ilyen eredmény maga után vonja az adatok (3.3)-nak megfelelő felbontását úgy, hogy minden $\hat{x}_t$ abban a H altérben van, amelyet a felhőből identifikáltunk. Azonkívül az *Egyértelműségi Elv* szerint H *nem* lehet egyértelmű.

Alapvető folytonossági meggondolásokból következnie kell — legalábbis matematikailag szép esetekben —, hogy ha a zaj 0-hoz tart, határértékként megkapjuk az *Egyértelműségi Elvet*. Tehát a zajos adatok problémáját a pontos esettel való kapcsolatában kell tekintenünk, legalábbis akkor, ha a feladat *zaj-mentes határértékként* folytonos. A mai statisztikából hiányzik a zaj-mentes határérték fogalma, mivel itt belefeledkeznek mintavételezésbe. Nyilvánvalóan a végtelen minta határeset nem elég érdekes az ortodox statisztikusoknak.

Az ismert zajos identifikációs sémák mind legalább egy feltevéssel élnek (3.3)-ra vonatkozóan:

(3.4) $$\hat{x} \text{ ortogonális } \tilde{x}\text{-ra}$$

„ortogonalitás" itt azt jelenti, hogy $\hat{x}_t$ és $\tilde{x}_t$ átlaga zérus, továbbá az adatok kereszt-kovariancia mátrixa

(3.5) $$\sum_t \hat{x}_t \tilde{x}_t' = 0.$$

(Ha a szokásos statisztikai feltevéssel élünk, miszerint a populációban $\hat{x}$ és $\tilde{x}$ (mint valószínűségi változók) ortogonálisak vagy függetlenek, még *nem* következik, hogy a (3.5) kovariancia nulla, hiszen mérési hibák előfordulhatnak. Ha $\hat{x}_t$ és $\tilde{x}_t$ különkülön ismertek, akkor (3.5) csak a ∞-mintavételi határértékben igaz. Ennek ellenére a legkisebb négyzetek elvét vagy a (3.3)-ra vonatkozó identifikáció más módszereit is úgy használják, mintha (3.5) pontosan teljesülne az adott adatokra (mintákra).)

Megszokott dolog, hogy nem közvetlenül az $\{x_t\}$ adatokkal foglalkozunk, hanem a felhő pontjait az adatok kovarianciamátrixával helyettseítjük, melyet a következőképpen definiálunk:

(3.6) $$\Sigma := \sum_t x_t x_t'.$$

Σ nyilván szimmetrikus mátrix. A definíció alapján könnyen látható, hogy nemnegatív definit.

Tegyük fel, hogy (3.5) teljesül. Ekkor az adatok (3.3)-nak megfelelő feltételezett felbontása az adatok kovarianciamátrixának következő felbontását eredményezi:

(3.7) $$\Sigma =: \hat{\Sigma} + \tilde{\Sigma}$$

ahol $\hat{\Sigma}$, $\tilde{\Sigma}$ szimmetrikus, nemnegatív definit. Ha az adatokat úgy generáljuk, hogy a fix q kodimenziós H altéren választjuk az $\hat{x}_t$ pontokat, akkor corank $(\hat{\Sigma})=q$ és $\hat{\Sigma}$ oszlopai H generátorhalmazát adják. (3.3)-nak és (3.4)-nek megfelelően általános zajt hozzáadva, $\tilde{\Sigma}$ pozitív definit lesz. Így feltehetjük, hogy a (3.6)-tal definiált Σ mátrix (általában) pozitív definit.

A fordított irányú felbontás; nevezetesen (3.7) alapján (3.3) igazolása kicsit kényesebb, s így később tárgyaljuk. A zajos identifikáció három standard sémáját fogjuk most áttekinteni.

(1) *Legkisebb négyzetek.* Tekintsük az x_t pontok koordinátáinak egy fix (tetszőleges) felbontását két halmazra. Legyenek ezek a halmazok x_{*t} és $x_{\#t}$. Mivel ez összeegyeztethető azzal a feltevéssel, hogy az adatok átlaga zéró, egy ilyen felbontás Σ egy felbontását vonja maga után, éspedig

$$\Sigma = \begin{pmatrix} \Sigma_{**} & \Sigma_{*\#} \\ \Sigma_{\#*} & \Sigma_{\#\#} \end{pmatrix}. \tag{3.8}$$

Definiáljuk S-t: $S := \Sigma^{-1}$. Ekkor S-nek értelmezhetjük a fentinek megfelelő felbontását. Vezessünk be néhány további definíciót:

$$\begin{gathered} A := S_{**}^{-1}(S_{**}S_{*\#}) \\ \tilde{\Sigma}_{**} := S_{**}^{-1}, \quad \tilde{\Sigma}_{*\#} = \tilde{\Sigma}_{\#*} = 0, \\ \hat{\Sigma} := \Sigma - \tilde{\Sigma}. \end{gathered} \tag{3.9}$$

Könnyen ellenőrizhető, hogy

$$A\hat{\Sigma} = 0, \quad \hat{\Sigma} \text{ és } \tilde{\Sigma} \text{ nemnegatív definit mátrixok} \tag{3.10}$$

és (3.7) teljesül. A (3.9—3.10) összefüggések a *legkisebb négyzetes zajos identifikációs tételt* alkotják, mivel egy olyan $\hat{H}$ alteret identifikálnak, melynek kodimenziója $q = \text{rank}\,(A)$.

Kapcsoljuk össze ezt az eredményt a legkisebb négyzetes eljárás megszokottabb jelöléseivel. Először is vegyük észre, hogy ebben az esetben a (3.3)-ban megkövetelt felbontást a következő választással érhetjük el:

$$x_t := \hat{\Sigma}\Sigma^{-1}x_t. \tag{3.11}$$

(3.11)-ben $P := \hat{\Sigma}\Sigma^{-1}$ egy vetítés operátor. Nyilván $AP = 0$. Így $P\,x_t$-t egy $\hat{H}$-ban fekvő $\hat{x}_t$-ra vetíti. Ezekből a tényekből $\{\hat{x}_t\}$ és $\hat{H}$ legkisebb-négyzetes tulajdonsága könnyen bebizonyítható. Ebből következik, hogy a legkisebb-négyzetes hiba tr $\tilde{\Sigma}$. Végül: (3.4—3.5) a legkisebb négyzetek klasszikus tulajdonságát adják.

Mostani mószerünk, ami semmi esetre sem nevezhető standardnak, a legkisebb négyzetek fontos új arculatát tárja fel; GAUSS óta ezek az első új tételek a témakörben. Az eredmények matematikailag triviálisak, igazak és így (ANDRÉ WEIL szerint) érdekesek!

Nézzük a legfontosabb tényeket:

(a) A legkisebb négyzetes realizációs tétel x_t tetszőleges $(x_{*t}, x_{\#t})$ felbontása esetén működik, és egy $q = \text{rank}\,(A) = \{* \text{ halmaz elemszáma}\}$ dimenziós $\hat{H}$ alteret identifikál. Ez azt jelenti, hogy a legkisebb négyzetek módszere teljesen használhatatlan identifikációra még kis zaj esetén is. A konstrukció eredménye kizárólag azon az előfeltevésen múlik, hogy a $*$ halmaz a változók mely részhalmazából áll. Az eredmény egyáltalán nem függ az adatoktól, hiszen az eljárás tetszőleges pozitív definit Σ esetén működik.

(b) A $\#$ részhalmazban levő változóknak igen egyszerű az értelmezése; zaj-mentesek. Ez a konstrukcióból látható, miszerint $\tilde{\Sigma}_{*\#} = \tilde{\Sigma}_{\#\#} = 0$. Ez nyilván nagyon szigorú feltevés. Általában semmi okunk nincs feltenni, hogy bármelyik változó is teljesen zaj mentes. A legkisebb négyzetek általánossága miatt nincs arra feltétel, hogy melyik változó kevésbé zajos, mint a többi — így bármelyik változót zajmentesnek mondhatunk.

(c) Ha az összes változót, egyet kivéve, zajmentesnek tekintünk, a szokásos legkisebb négyzetes esetet kapjuk (egyszerű regresszió). Ekkor $\tilde{\Sigma}$-nak csak egy nem nulla eleme van, σ_{ii}, ami annak felel meg, hogy az i-edik változót választjuk regresszornak. Ebből következik, hogy $\tilde{\sigma}_{ii}:=1/S_{ii}$, tehát a problémát triviálisan megoldottuk. Semmi tényleges minimalizálás nincsen!
(d) Mivel $\hat{\Sigma}$ nemnegatív definit, teljesülnie kell, hogy $\sigma_{ii}\cdot\geqq\tilde{\sigma}_{ii}\geqq 1/S_{ii}$ amiből az következik, hogy

$$\sigma_{ii}s_{ii}\geqq 1. \tag{3.12}$$

Ez a fizikában szereplő *Bizonytalansági Elv* szokásos alakja. Látjuk tehát, hogy szoros kapcsolat lehet a *Bizonytalansági Elv* fizikai értelmezése és modern legkisebb négyzetes eljárásunk bizonytalanságának elemi vizsgálata között.

(2) *Főkomponensek.* A statisztikai fogalmaknak gyakran sokkal régebbi matematikai forrása van. Mivel Σ szimmetrikus mátrix, egyszerű spektrálelőállítása van. Így $\Sigma=U\Lambda U'$, ahol Λ diagonális (a diagonálisban pozitív elemekkel) és U egy ortogonális mátrix. Sokszor megjegyezték már, hogy ez a felbontás egy zajos realizációs tételt ad, a következőképpen.

Jelölje Λ^* azt a diagonális mátrixot, melynek q nem nulla diagonális eleme Λ q sajátértékével egyezik meg. (Például a q legkisebb sajátértékkel), és az összes többi eleme 0. Hasonlóan, legyen $\Lambda_\#$ diagonálisában Λ maradék r sajátértéke, mint nem nulla diagonálisbeli elemek. Megköveteljük, hogy $\Lambda=\Lambda_*+\Lambda_\#$ és $\Lambda_\#\Lambda_*=0$. Ismét, $*$ jelentése „zajos", $\#$ jelentése „zajmentes". Legyenek továbbá

$$\begin{cases} A := \Lambda_* U' \\ \hat{\Sigma} := U\Lambda_\# U' \\ \tilde{\Sigma} := U\Lambda_* U', \end{cases} \tag{3.13}$$

ekkor nyilván teljesülnek, hogy

$$\begin{aligned} \Sigma &= \hat{\Sigma}+\tilde{\Sigma} \\ A\hat{\Sigma} &= 0 \end{aligned} \tag{3.14}$$

ami ekvivalens azzal, hogy egy q kodimenziós H alteret identifikáltunk! A $\hat{\Sigma}$, $\tilde{\Sigma}$ mátrixok nyilván pozitív definitek. Tanulságos bebizonyítani, hogy $\hat{\Sigma}\Sigma^{-1}$ most is vetítés operátor.

PEARSON (1901) vezette be ezt a sémát, miután — helyesen — arra a következtetésre jutott, hogy a legkisebb négyzetek előfeltevése lényegében benne van a zajmentes változók feltételezésében.
Itt minden változó zajos. Sajnos PEARSON nem különítette el a legkisebb négyzetek összes előfeltevését. Az ő sémája is tele van előfeltevésekkel, mivel az adatokból nem lehet q természetes (vagy valódi) értékét meghatározni. Továbbá, a *Pearson sémában* a $\tilde{\Sigma}$ zaj-mátrix tetszőleges korrelációkat tartalmazhat; ezek a korrelációk a spektráltétel alkalmazásából származnak. A spektrál tétel egy egyszerű matematikai eredmény, de semmi köze a zajos identifikációhoz. További részletekről, l. KALMAN [1982, 8.b fejezet].

4. A Frish-séma

Az altér zajos identifikációjára vonatkozó előfeltevésmentes realizációs tétel lehetősége először FRISH [1934] fontos monográfiájában vetődött fel, legalábbis hallgatólagosan. Nem követte a *Fisher-féle statisztikai dogmákat.* Be akarta bizonyítani, hogy a legkisebb négyzetes előfeltevés (szükséges) velejárója q tetszőleges választása. Nem sikerült ezt teljesen bebizonyítania; bár tulajdonképpen nincs is szükség semmiféle bizonyításra, hiszen q tetszőlegessége azonnal következik abból a fent említett tényből, hogy a legkisebb négyzetes formulák tetszőleges q esetén igazak. Mindazonáltal a *Frish-séma* egy fontos séma, mert ennek mélyebb analízise elvezet a zajos identifikáció modern munkáihoz, l. KALMAN [1981].

(3.4—3.5) mellett FRISH bevezetett egy második axiómát:

(4.1) $\tilde{x}$ minden komponense merőleges $\tilde{x}$ összes többi komponensére,

vagy a kovarianciák nyelvén megfogalmazva

(4.2) $\hat{\Sigma}$ = diagonális, nemnegatív definit.

Ez az axióma elkerülhetetlen, ha a lineáris relációk identifikálásának problémáját pontosan akarjuk definiálni. Sajnos (4.1—4.2)-ből az is következik, hogy nincs egyáltalán megfigyelési hiba a kovarianciák meghatározásakor, mivel feltevésünk szerint (3.4—3.5) és (4.1—4.2) az adat-kovarianciákra pontosan teljesülnek, hiszen várhatóan teljesülnek a populációs kovarianciákra.

FRISH azt is észrevette — bár nem teljesen expliciten —, hogy q nincs jól definiálva, hacsak nem maximális. Valóban, ha a Σ adatok q független lineáris relációt tartalmaznak, akkor mindig tartalmazzák $q'<q$ lineáris relációk végtelen halmazát is, így q' nem lehet egy identifikálható invariánsa az adatoknak. Ezekkel a megjegyzésekkel:

(3) *Frish-séma.* Legyen Σ egy fix szimmetrikus, pozitívdefinit mátrix. Határozzuk meg az összes olyan $\hat{\Sigma}$ és $\tilde{\Sigma}$ mátrixot, melyekre

$\Sigma = \hat{\Sigma} + \tilde{\Sigma}$

$\hat{\Sigma}$, $\tilde{\Sigma}$ = szimmetrikus, nemnegatív definit,

$\tilde{\Sigma}$ diagonális,

corank $(\hat{\Sigma})$ = maximális.

A *Frish-séma* matematikai elmélete eredményes lenne, ha közvetlenül meg tudnánk határozni corank $\hat{\Sigma}$ maximumát. Akkor az összes lehetséges $\hat{\Sigma}$ explicit meghatározása várhatóan egy világos feladathoz vezetne. (A *Frish megoldások* általában nem egyértelműek, hiszen az *Egyértelműségi Elv* ezt állítja.) A megoldáshalmaz dimenziója könnyen látható algebrai-geometriai megfontolásokkal (konstansok számolása), miszerint

$$d(n, q) = n - q/2 - q^2/2. \tag{4.4}$$

Ezt a formulát az 1920-as évek óta ismerik és vitatják. Sajnos az általános esetre még ma sincs bizonyítás.

A (4.4) formulát úgy kell értelmezni, melyben n fix és q változhat. Ez az adatoktól függetlenül jónéhány feltételt kiró q-ra, hiszen nyilvánvalóan, $d \geqq 0$.

Így, sajnos a *Frish-séma* is használ előfeltevést egy elég rejtett módon.

Tekintsünk R^n-ben egy L egyenest (kodimenziója $n-1$), amelynek nagy része körül egy szűk felhő csoportosul. Intuitíven világos, hogy L identifikálható egy ilyen felhőből, de a *Frish-sémával* nem lehet identifikálni a Σ-ra vonatkozó $-d(n, n-1)=$

$=n^2/2-3n/2$ megszorítás nélkül. Egy ilyen megkötés bármilyen értelmes L-körül csoportosuló zaj-definícióval ellentmondana, függetlenül attól, hogy milyen kicsi.

Így nincs remény arra, hogy egy egyszerű és általános úton meghatározzuk corank $\hat{\Sigma}$ maximumát. Mégis, van egy általános tétel, melyet már FRISH is megsejtett és KALMAN [1981] fogalmazta meg és bizonyította be precízen először.

(4.5) FŐ TÉTEL. *A Frish-sémában* max corank $\hat{\Sigma}=1$ *pontosan akkor teljesül, ha* Σ^{-1} *egy Frobenius mátrix (minden eleme szigorúan pozitív), vagy ha* Σ^{-1} *Frobenius mátrixba transzformálható a változók (az* x_t *adatvektor elemei) előjelének alkalmas újradefiniálásával.*

Ha $q=1$, akkor a H alteret egy egyszerű a' kovektorra vonatkozó vetítés definiálja. A *Frobenius mátrix* definíciója szerint ebből az következik, hogy a' minden komponense nem nulla; tulajdonképpen pozitív, ha az előjeleket alkalmasan újradefiniáljuk. Így a' *tetszőleges eleme 1-gyé normalizálható.* Az összes a' vektorok halmaza expliciten meghatározható. Ez a számítás meghatározza az identifikált rendszer bizonytalanságát is, mely az adatok rejtett bizonytalanságának következménye.

(4.6) TÉTEL. *Ha* max corank $\hat{\Sigma}=1$, *akkor tetszőleges fix normalizálás esetén normalizált* a' *kovektorok az összes olyan* $\hat{H}$ *alteret meghatározzák, amelyek Frish-kompatibilisek egy n-szimplex* Σ *formával, amelynek csúcspontjai az összes lehetséges elemi regressziós probléma (legkisebb négyzetek* $q=1$ *esetén) megoldásai.*

Így a legkisebb négyzeteknek azt az előfeltevését, mellyel éppen $q=1$ *zajos változót választ ki, a Frish-séma egyszerűen úgy kerüli ki, hogy a Frish megoldás-halmaz magába foglalja az összes ilyen lehetőséget.* Más szóval $q=1$ esetén a *Frish-megoldás* összeegyeztethető a *Bizonytalansági Elv* intuitív követelményeivel, ezáltal elkerülve az előfeltevést.

(4.5—4.6) összekombinálásával kielégítő tételt kapunk a $q=1$ kodimenziós alterek identifikálására vonatkozóan. Megjegyezzük, hogy $d(n, 1)=n-1$, amely természetesen a (4.6)-ban említett szimplex dimenziójával egyezik meg.

Még precízebben megfogalmazhatjuk azt a követelményt, hogy egy $q=1$ kodimenziós fix altér körül csoportosuló tetszőleges felhőt tekintsünk, megfelelően kicsi zajjal. Még arra a feltételre sincs szükségünk, hogy az adatok eleget tegyenek a (3.4—3.5) első axiómának.

(4.7) TÉTEL. *Tekintsünk egy általános, 1-kodimenziós, fix,* H^* *alteret. Legyen* x_t^*, $t=1, \ldots, N$ H^* *pontjainak egy megfelelően teljes halmaza, melynek átlaga nulla, és* corank $\Sigma^*=1$. *Legyenek* $x_t:=x_t^*+\tilde{x}_t$, *ahol* $\tilde{x}_t$ *tetszőleges, általános zajvektorok, melyek átlaga nulla, és szintén elegendően kicsik ahhoz, hogy* Σ *pozitív definit legyen, megfelelően közel* Σ^**-hoz.*

Ekkor a Frish-séma maximális korangja 1, *továbbá létezik egy* $\hat{H}$ *Frish-altér, amely egybeesik* H^**-gal.*

Ez az eredmény azt mutatja, hogy $q=1$ típusú tetszőleges (kicsi) zaj a felhőben mindig helyettesíthető *„Frish-zajjal"* amely mind (3.4—3.5)-nek, mint (4.1—4.2)-nek eleget tesz. Így még abban az esetben sem egyértelműen a zaj identifikálása, mikor H pontosan identifikált. Ez természetesen jól összeegyeztethető a *Bizonytalansági Elvvel.* Azt is látjuk, hogy a zajra vonatkozó matematikai eredményeket a valószínűségre vonatkozó bármilyen előfeltevés nélkül is megkaphatjuk.

Egy fontos technikai kérdés van még hátra. Indukál-e egy (3.3) felbontást a *Frish probléma* valamelyik $\hat{\Sigma}$ megoldása (azaz valamely $\hat{\Sigma}$, amely (4.3)-nak eleget tesz)?

Ha adott az adatok kovarianciája, akkor létezik egy normális valószínűségi változó ezzel a kovarianciával. Valószínűségi tényeket tekintve x és $\hat{x}$-t normális valószínűségi változóknak tekinthetjük, melyeknek kovarianciáját a *Frish-probléma* egy (fix) megoldása adja és úgy tekinthetjük x_t és $\hat{x}_t$-t, mint ezeknek a valószínűségi változóknak a megfigyelései. (3.4) miatt x és $\hat{x}$ keresztkovarianciáit $\hat{\Sigma}$ tartalmazza. Ekkor jól ismert tény, hogy a feltételes várható érték

$$E(\hat{x}_t|x_t) = \hat{\Sigma}\Sigma^{-1}x_t, \tag{4.8}$$

a feltételes kovariancia pedig

$$\operatorname{cov}(\hat{x}_t|x_t) = \hat{\Sigma} - \hat{\Sigma}\Sigma^{-1}\hat{\Sigma}. \tag{4.9}$$

Ha a $P=\hat{\Sigma}\Sigma^{-1}$ operátor egy vetítés, mint ez a legkisebb négyzetek és a főkomponensek esetében volt, ekkor az $\hat{x}_t$ x_t-re vonatkozó feltételes kovarianciája nulla; más szóval $\hat{x}_t$ egyértelműen meghatározható az x_t adatok és Σ adat-statisztika segítségével. Ez azt jelenti, hogy az x_t-beli $\tilde{x}_t$ zaj pontosan „kiszűrhető" egyszerűen Σ, majd P kiszámításával. Egy ilyen lehetőség eléggé ellene mond ösztönös érzéseinknek, ráadásul nyilván megsérti a *Bizonytalansági Elvet*. Így megint leszűrhetjük azt a tanulságot, hogy legkisebb négyzetek és főkomponensek előfeltevéseket tartalmaznak, *kivéve* mikor *tudjuk*, hogy x és $\hat{x}$ normális valószínűségi változók (fix populáció), és elegendően sok megfigyelésünk van, amiből pontosan becsülhetjük Σ-t és $\hat{\Sigma}$-t.

A *Frsih-probléma* esetén $\hat{\Sigma}\Sigma^{-1}$ általában *nem* vetítés. Emiatt a (4.9)-beli kovariancia nem nulla, de ez a kovariancia méri az $\hat{x}_t$ meghatározásában levő belső bizonytalanságot, ha adott x_t, Σ és egy $\hat{\Sigma}$ *Frish-megoldás*. Ez épp az, aminek lennie kell, tökéletes szűrés nem lehetséges. Megjegyezzük, hogy ez a következtetés nem függ a valószínűségelmélettől, a (4.8) formula igaz anélkül is, hogy bármiféle feltételt tennénk valószínűségi változókra vonatkozóan.

(4.10) TÉTEL. *A Frish-probléma tetszőleges fix* $\hat{\Sigma}$, Σ *megoldására létezik (végtelen sok)* (3.3)*-nak megfelelő felbontás, melyre* $\hat{x}_t$ *és* $\tilde{x}_t$ *a megkövetelt kovarianciájú. Még pontosabban,* $\hat{x}_t$*-nak magának is van egy felbontása*

$$\hat{x}_t = \hat{\hat{x}}_t + \hat{\tilde{x}}_t,$$

ahol (4.8) *egyértelműen meghatározza* $\hat{\hat{x}}_t$*-t és a második tag kovarianciáját* (4.8) *adja.*

Ez a helyzet nagyon hasonló ahhoz, amikor adott kovarianciájú számsorozatot kell generálni. Egy ilyen problémának nagyon sok megoldása van. Hasonlóan, nagyon sok $\hat{\tilde{x}}_1, \ldots, \hat{\tilde{x}}_N$ sorozat van, aminek a (4.9)-ben megkövetelt kovarianciája van. Ha ezt a sokaságot absztrakt eseménytérnek tekintjük, nagyon sok különböző valószínűségeloszlást definiálhatunk ezen az eseménytéren. Ezek a valószínűségeloszlások nyilván nem identifikálhatóak, hiszen a problémában csak az az egy követelmény szerepel, hogy $\hat{\tilde{x}}_t$ kovarianciáját (4.9) adja. Mivel a *Frish-megoldás* általában nem egyértelmű, ezt a kovarianciát sem lehet identifikálni. (De lehet, hogy korlátos!)

Attól félek, hogy nem hagyhatjuk el azt a következtetést, hogy az identifikációs probléma bizonytalanságának leírását szolgáló valószínűségi struktúra feltétele (előfeltétele) lényegileg hiábavaló, mivel ilyen struktúrát nem tudunk az adatokból identifikálni.

Egyetértek azzal a mai fizikai filozófiával, hogy amit nem lehet identifikálni, azaz megmérni, annak nincs konkrét léte. Így tagadnunk kell a klasszikus valószínűségi struktúra tudományos megalapozottságát, azaz a klasszikus valószínűség-

elméletet egészében, mint a bizonytalanság leírását. Ez azonban nem zárja ki, hogy más értelemben foglalkozzunk a bizonytalansággal.

Az itt bemutatott korábbi eredmények erősen függtek különféle pozitivitásokon; például azon a nyilvánvaló, bár érdekes tényen, hogy minden változó pozitív. A poztitivitás (valószínűségé, sűrűségfüggvényé stb.) természetesen szintén alapvető a valószínűségelméletben. Lehetséges-e, hogy van egy alternatív mód egy újfajta (nem klasszikus) valószínűségi elmélet felépítésére ugyanarról a pontról — nevezetesen a pozitivitásból — kiindulva?

5. Instabilitás

Sajnos a *Frish-séma*, amely $q=1$ esetén megfelelő, elromlik $q>1$ esetén. Mivel FRISH kutatásának főleg a $q>1$ volt a célja, ez eléggé kiábrándító, bár elkerülhetetlen kritika.

(5.1). *ELLENPÉLDA. Adott fix $q^*>1$ és megfelelően általános q^* kodimenziós H^* altér esetén létezik egy teljes felhő tetszőlegesen kicsi zajjal, úgy hogy Σ^{-1} Frobenius mátrix, más szóval a Frish-séma alapján identifikált $q=1$ hamis.*

Így $q>1$ esetén a *Frish-séma* nem megbízható identifikációs séma, még a gyengezaj határérték közelében sem. Más szóval a *Frish-séma* a pontos esetben nem folytonos a zajperturbációra vonatkozóan.

Az itt tárgyalt általános jelenséget a következőképpen lehet elképzelni. Emlékeztetünk arra, hogy a *realizációs tétel* egy olyan matematikai eredmény (formula vagy algoritmus), amely lehetőséget ad a pontos adatokból a minimális realizáció megkonstruálására. Az erre vonatkozó klasszikus példa, mint már említettük, a *Lagrange interpolációs probléma.* Tegyük fel, hogy kiválasztottunk 10^6 pontot, amelyek pontosan egy gömb felületén helyezkednek el (de nem egy alacsonyabb fokú görbén). Ekkor $\deg \pi_L=3$. Ha a 10^6 pontot tetszőlegesen kicsi zajjal perturbáljuk, akkor általában azt kapjuk, hogy $\deg \pi_L=10^6$. Más szóval, a *Lagrange formula instabil* kis perturbáció esetén. Nyilvánvaló, hogy az identifikált rendszer komplexitása nem függhet az adatok számától, de a *Lagrange formulából* éppen ez következik a zajos esetben. A gyakorlatias emberek azt mondanák „LAGRANGE egyszerűen beilleszti a zajt". Így a *Lagrange formula* tévedés: az egzakt esetben teljesíti a minimális realizáció egyértelműségét, de haszontalan zajos esetben.

A *Frish-sémában* is előfordul instabilitás, bár sokkal körmönfontabb módon. Fix n és q, de ha $d(n, q)$ negatív, akkor (legalábbis heurisztikusan) az adat mátrixban $-d(n, q)$ algebrai feltételnek teljesülnie kell ahhoz, hogy a *Frish megoldás* létezhessen. Egy tetszőlegesen kicsi perturbáció (például a megfigyelési hibából eredően) elrontja ezt a megoldást. Ezt mostanában bizonyította be SHAPIRO [1982], aki megmutatta, hogy $d(n, q^*)<0$ esetén az adott Σ^* mátrixhoz — mely megenged egy ilyen megoldást — tetszőlegesen közel léteznek olyan Σ^+ mátrixok, melyekre $d(n, q^+)\geqq 0$. Más szóval, a *Frish megoldások* maximális korangjai q^*-ról q^+-re ugranak.

A dimenzionális megfontolásokból jelentkező nehézségek csak nagy q esetén lépnek fel. Pillanatnyilag az előfeltevésen kívül nem ismeretes olyan séma, amely kezelni tudná a nagy korang esetet.

Ez egy komoly dolog, különösen azért, mert a probléma ma már több, mint nyolcvan éves!

Tekintettel ezekre a megjegyzésekre, érdemes visszatérni a *Kepler—Newton*

problémához. Vajon ez is instabil? Mostani elméletünkkel a választ a következőképpen adhatjuk meg.

Naprendszerünkben egy *E* bolygó *Nap* körüli mozgása megközelítőleg meghatározható *E* és a *Nap* segítségével. A többi bolygó hatását zajnak tekinthetjük. (Egy kellemes példa nem valószínűségi zajra!) KEPLER, mint kiváló beleérző zseni, továbblépett egyet. A számára elérhető adatokat értelmezve a zaj-mentes határértékig ment, azaz a csillagászati adatokat pontos ellipsziseknek tekintette, melyet *E* és a *Nap* teljesen meghatároznak.

NEWTON a pontos realizációs problémát oldotta meg gravitációs formulája kifejlesztésével. Bebizonyította matematikailag, hogy a két-test probléma megoldása mindig *Kepler-elipsziseket* eredményez. Ezt hívjuk ma egy realizációs tételnek.

A *Newton-formula* bármely két testre alkalmazható, így a *P* bolygó *Nap* körüli pályájából NEWTON (és követői) ki tudták számolni a többi bolygótól és testtől eredő zaj nagyságrendjét. Azokat a testeket tekintve, amik ténylegesen jelen vannak naprendszerünkben, a zajt általában igen kicsinek találták. Ily módon NEWTON realizációs tétele — gravitációs törvényének formulája — alkalmazható volt a zajos esetre is. Így itt nem volt instabilitás.

Hasonlóan lehet megvizsgálni MENDEL genetikai törvényeinek sikerét. Részletesen l. KALMAN [1982].

6. Következtetések

Ennek a dolgozatnak fő gondolata, hogy a zajos adatokon alapuló identifikáció vizsgálata eltávolodást kíván a klasszikus matematikától — nem logikájában vagy eljárásaiban, hanem a tanulmányozás tárgyát tekintve — és ez az eltávolodás nemtriviális problémákat okoz a legegyszerűbb esetben is, mikor lineáris rendszert identifikálunk. A legtöbb esetben hibás volt GAUSSNAK az ötlete, hogy a zajt legkisebb négyzetekkel kezeli, a zajmentes változók feltételezéséből adódó erős előfeltevés miatt. Még inkább kifogásolható az a próbálkozás, mely megpróbálja GAUSS ötletét úgy megjavítani, hogy a legkisebb négyzeteket beágyazza egy valószínűségi modellbe, hiszen ez még erősebb előfeltevést követel. Más szóval a problémának fontos matematikai kapcsolata van, amely független minden előfeltevéstől. A legkisebb négyzeteket egy alapvető matematikai segédeszköznek kell tekinteni, nem pedig a zajos identifikáció elméletének. Így tehát nagy örömmel elfogadom PONTRJAGIN álláspontját. Jobban szeretem a matematikát bármiféle valószínűségi hiedelem (azaz előfeltevés) nélkül.

Kritikus vitapont az előfeltevés. Mindamellett KOPERNIKUSZ fő gondolata volt azt hangsúlyozni, hogy előfeltevés, és nem logikai, matematikai, vagy fizikai szükségszerűség a Föld középpontját választani a csillagászati koordinátarendszer középpontjának. Ennek az előfeltevésnek a megszüntetése óriási pszichológiai következményeket okozna.

Matematikusoknak nem kell megmagyarázni, hogy mi az egyértelműség. De a gyakorlati tudományokban, speciálisan a soft tudományokban dolgozó kutatóknak az *Egyértelműségi Elv* kihangsúlyozza azt a tényt, hogy a tudományos eredményeket az adatok objektív vizsgálatával kell megkapni, nem pedig úgy, hogy egoista módon „okos" egyéni modelleknek kedvezünk.

Természetesen minden alkalmazott matematikus számára alapvető a zajos probléma vizsgálata. Ezt matematikusoknak kell megoldani, nem pedig előfeltevéseknek.

Gyakran megemlítik EINSTEIN szavait, „*Gott würfelt nicht*". (Isten nem kockajátékos.) Ezt a kijelentést ma úgy magyarázhatjuk, hogy a *Természet* nem a kockajátékok, a rulett vagy a kártyajátékok szabályai alapján épült fel, azaz a *Természet* nem a klasszikus *(Kolmogorov-féle)* valószínűségelmélet szabályai szerint épül fel. EINSTEIN nem volt annyira naiv, hogy azzal érvelt volna, a *Természet* kizárja a bizonytalanságot.

IRODALOM

R. FRISCH
[1934] "Statistical confidence analysis by means of complete regression systems", Publication no. 5, University of Oslo Economic Institute, 192 pages.
W. V. D. HODGE and D. PEDOE
[1968] *Methods of Algebraic Geometry*, Volume 1, Cambridge University Press.
R. E. KALMAN
[1960] "On the general theory of control systems", in *Proceedings First IFAC Congress on Automatic Control*, Moscow, vol. 1, (Butterworths, London) pages 481—492.
[1962] "Canonical structure of linear dynamical systems", Proc. National Academy of Sciences, **48**: 596—600.
[1963] "Mathematical description of linear dynamical systems", *SIAM J. Control*, **1**: 152—192.
[1976] "Realization theory of linear dynamical systems" in *Control Theory and Functional Analysis*, Vol. II, International Atomic Energy Agency, Vienna, pages 235—256.
[1981] "System identification from noisy data", in *Dynamical Systems II*, (edited by A. R. Bednarek and L. Cesari), (Academic Press) pages 331—342.
[1982] "Identification from real data", in *Current Developments in the Interface: Economics, Econometrics, Mathematics*, (edited by M. Hazewinkel and A. H. G. Rinnooy Kan), D. Riedel, Dordrecht, pages 161—196.
R. E. KALMAN, P. FALB, and M. A. ARBIB
[1969] *Topics in Mathematical System Theory* (McGraw-Hill) 358 pages.
K. PEARSON
[1901] "On lines and planes of closest fit to systems of points in space", Philosophical Magazine, *VI:* 559—572.
A. SHAPIRO
[1982] "Rank reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis", *Psychometrica*, **47**: 187—199.
E. SONTAG and Y. ROUCHALEAU
[1976] "On discrete-time polynomial systems", *J. Nonlinear Analysis*, **1**: 55—64.
J. C. WILLEMS
[1979] "System-theoretic models of physical systems", *Richerche di Automatica*, **10**: 71—106.
Y. YAMAMOTO
[1981] "Realization theory of finite-dimensional linear systems". *Mathematical System Theory*, **15**: 55—77 and 169—190.

FORDÍTOTTA:
G. VÁGÓ ZSUZSA
BME GÉPÉSZMERNÖKI KAR MATEMATIKA TANSZÉK
1521 BUDAPEST, STOCZEK U. H ÉP. IV. E.

A kiadásért felel az Akadémiai Kiadó és Nyomda Vállalat főigazgatója
Műszaki szerkesztő: Sándor István
A kézirat nyomdába érkezett: 1986. június 10. — Terjedelem: 16,45 (A/5 ív)
87-3016 — Szegedi Nyomda, Szeged. — Felelős vezető: Surányi Tibor igazgató

MTA KÖNYVTÁRA
INFORMATIKAI IGAZGATÓSÁG

EGY SZAKTERÜLET, AMELYNEK LEGÚJABB EREDMÉNYEI MEGVÁLTOZTATHATJÁK FIZIKAI ISMERETEINKET

SZUPRAVEZETÉS

Az 1986. októberi, ill. 1987. márciusi fontáttörés óta
hetente táguló keretekkel

KÖVESSE A TÉMA SZAKIRODALMÁT
HETI SZÁMÍTÓGÉPES SZAKIRODALMI SZOLGÁLTATÁSUNK IGÉNYBEVÉTELÉVEL

Beküldendő: **MTA Könyvtára Informatikai Igazgatóság 1361 Bp. Pf. 7.**
Kérem, hogy a gépi szakirodalmi információszolgáltatás díjmentes ismertetőjét az alábbi címre szíveskedjék megküldeni:

Név:.....

Munkahely: :.....

Postai cím:.....

.....

Kelt:.....198.....hó.....nap

.....
aláírás

MATEMATIKAI LAPOK

A Bolyai János Matematikai Társulat lapja

Felelős szerkesztő: Császár Ákos

Önálló kutatások eredményeit közli a matematika területéről, valamint a matematika egyes részterületeinek összefoglaló, monografikus feldolgozásait. Rendszeresen közöl feladatokat, és azok megoldását is megadja. Ismerteti az egyetemisták számára évenként megrendezésre kerülő matematikai versenyek feladatait és feladatmegoldásait is.

Alapítva: 1892
Magyar nyelven, angol és orosz összefoglalókkal
Megjelenik évente 1 kötet, 4 füzetben
Évi előfizetési díja: 48,-Ft
Előfizethető a Hírlapelőfizetési és Lapellátási Irodánál (HELIR)
Budapest, József nádor tér 1. 1900
Pénzforgalmi jelzőszám: 215-96162

AKADÉMIAI KIADÓ, BUDAPEST

ÚTMUTATÁS A SZERZŐKNEK

Az Alkalmazott Matematikai Lapok csak magyar nyelvű dolgozatokat közöl. A kéziratok gépelését olyan formában kérjük, hogy minden gépelt oldal 25, egyenként átlag 50 betűhelyes sort tartalmazzon. A közlésre szánt dolgozatokat három példányban kell beküldeni.

A kéziratok szerkezeti felépítésének a következő követelményeket kell kielégíteni. A fejlécnek tartalmaznia kell a dolgozat címét, a szerző teljes nevét, valamint annak a városnak a nevét, ahol a szerző dolgozik. A fejléc után egy, képletet nem tartalmazó, legfeljebb 200 szóból álló kivonatot kell minden esetben megadni. A dolgozatot címmel ellátott szakaszokra kell bontani, és az egyes szakaszokat arab sorszámmal kell ellátni. Az esetleges bevezetésnek mindig az első szakaszt kell alkotnia. Az irodalomjegyzék mindig az utolsó szakasz kell hogy legyen, és azt nem kell sorszámmal ellátni. Az irodalomjegyzék után, a kézirat befejezéseképpen fel kell tüntetni a szerző teljes nevét és a munkahelye (illetve lakása) pontos postai címét. A dolgozatban előforduló képleteket szakaszonként újrakezdődően, a képlet előtt két zárójel közé írt kettős számozással kell azonosítani. Természetesen nem szükséges minden képletet számozással ellátni. Az esetleges definíciókat és tételeket (segédtételeket és lemmákat) ugyancsak szakaszonként újrakezdődő, kettős számozással kell ellátni. Kérjük a szerzőket, hogy ezeket, valamint a tételek bizonyítását a szövegben kellő módon emeljék ki. Minden dolgozathoz csatolni kell egy angol, német, francia vagy orosz nyelvű, külön oldalra gépelt összefoglalót. Amennyiben lehetséges, kérjük a nyomtatás számára különösen nehézkes matematikai jelölések használatának az elkerülését.

A dolgozat ábráit és az esetleges lábjegyzeteket a dolgozat végén, különálló lapokon kérjük beküldeni. Mind az ábrákat, mind a lábjegyzeteket a dolgozat szakaszokra bontásától független, folytatólagos arab sorszámozással kell ellátni. Az ábrák elhelyezését a dolgozat megfelelő helyén, széljegyzetként feltüntetett, ábraazonosító sorszámokkal kell megadni. A lábjegyzetekre a dolgozaton belül az azonosító sorszám felső indexkénti használatával lehet hivatkozni.

Az irodalmi hivatkozások formája a következő. Minden hivatkozást fel kell sorolni a dolgozat végén található irodalomjegyzékben, a szerzők, illetve társszerzők esetén az első szerző neve szerint alfabetikus sorrendben úgy, hogy külön, de folytatólagos sorszámozású listát alkossanak a latin és a cirill betűs nevű szerzők műveire vonatkozó hivatkozások, és mindkét részben a megfelelő alfabetikus sorrend legyen kialakítva. A folyóiratban megjelent cikkekre [1], a könyvekre [5], a kötetben megjelent dolgozatokra [4], a disszertációkra [3] és a gépi program leírásokra [2] a következő minta szerint kell hivatkozni:

[1] Farkas, J., Über die Theorie der einfachen Ungleichungen , *Journal für die reine und angewandte Mathematik* **124** (1902) 1—27.

[2] Kéri, G., „DUALSIMP", rutin a CDC 3300-as gépekre (Magyar Tudományos Akadémia Számítástechnikai és Automatizálási Kutató Intézete, CDC 3300 felhasználói ismertetők 2. 1973. május) 19—20.

[3] Prékopa, A., „Sztohasztikus rendszerek optimalizálási problémáiról", doktori értekezés. Magyar Tudományos Akadémia, Budapest, 1970.

[4] Prabhu, N. U., "Recent research on the ruin problem of collective risk theory", in: *Inventory Control and Water Storage* Ed. A. Prékopa (János Bolyai Mathematical Society and North-Holland Publishing Company, Amsterdam—London, 1973) 221—228.

[5] Zoutendijk, G., *Methods of Feasible Directions* (Elsevier Publishing Company, Amsterdam and New York, 1960).

A dolgozatok szövegében az irodalmi hivatkozás számait szögletes zárójelben kell megadni, mint például [5] vagy [4, 76—78]. A szerzők a dolgozatukról 100 darab különlenyomatot kapnak ezek költsége — nyomtatott oldalanként 25 forint — a szerzői díjat terheli.

Ára: 64 Ft

ISSN 0133—3399

TARTALOMJEGYZÉK

INDEX

Alkalmazott matematikai lapok

1987–88/3–4

A MAGYAR TUDOMÁNYOS AKADÉMIA
MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK
OSZTÁLYÁNAK KÖZLEMÉNYEI

13.

KÖTET

AKADÉMIAI KIADÓ, BUDAPEST

ALKALMAZOTT MATEMATIKAI LAPOK

A MAGYAR TUDOMÁNYOS AKADÉMIA
MATEMATIKAI ÉS FIZIKAI
TUDOMÁNYOK OSZTÁLYÁNAK KÖZLEMÉNYEI

XIII. kötet 3—4. szám

Szerkesztőség: 1502 Budapest XII., Kende u. 13—17.
Kiadóhivatal: 1055 Budapest V., Alkotmány u. 21.

Az Alkalmazott Matematikai Lapok változó terjedelmű füzetekben jelenik meg, és olyan eredeti tudományos cikkeket publikál, amelyek a gyakorlatban, vagy más tudományokban közvetlenül felhasználható új matematikai eredményt tartalmaznak, illetve már ismert, de színvonalas matematikai apparátus újszerű és jelentős alkalmazását mutatják be. A folyóirat közül cikk formájában megírt, új tudományos eredménynek számító programokat, és olyan, külföldi folyóiratban már publikált dolgozatokat, amelyek magyar nyelven történő megjelentetése elősegítheti az elért eredmények minél előbbi, széles körű hazai felhasználását.

A folyóirat feladata a Magyar Tudományos Akadémia III. (Matematikai és Fizikai) Osztályának munkájára vonatkozó közlemények, könyvismertetések stb. publikálása is.

A kéziratok a főszerkesztőhöz, vagy a szerkesztő bizottság bármely tagjához beküldhetők. A főszerkesztő címe:

Prékopa András, főszerkesztő
1502 Budapest, Kende u. 13—17.

Közlésre el nem fogadott kéziratokat a szerkesztőség lehetőleg visszajuttat a szerzőhöz, de a beküldött kéziratok megőrzéséért vagy továbbításáért felelősséget nem vállal.

Az Alkalmazott Matematikai Lapok előfizetési ára kötetenként 128 forint. Belföldi megrendelések az Akadémiai Kiadó, 1055 Budapest V., Alkotmány u. 21. címen (pénzforgalmi jelzőszám 215—11 488), külföldi megrendelések a Kultúra Külkereskedelmi Vállalat, H-1389 Budapest, Pf. 149. címen (pénzforgalmi jelzőszám 218—10 990) lehetségesek.

A Magyar Tudományos Akadémia III. (Matematikai és Fizikai) Osztálya a következő idegen nyelvű folyóiratokat adja ki:

1. Acta Mathematica Hungaricae,
2. Acta Physica Hungaricae,
3. Studia Scientiarum Mathematicarum Hungarica.

Alkalmazott Matematikai Lapok 13 (1987—88) 189—206

HILBERT-TEREK LINEÁRIS OPERÁTORAINAK SZINGULÁRIS FELBONTÁSA
(Optimumtulajdonságok statisztikai alkalmazásai és numerikus módszerek)

BOLLA MARIANNA

Budapest

Általános *Hilbert-terek* lineáris operátorainak elméletéből jól ismertek a kompakt operátorok spektrál-, illetve szinguláris felbontására vonatkozó tételek (pl. [13]).

Ennek ellenére a továbbiak megértése céljából teljesen általánosan kimondjuk ezeket. A felbontás optimumtulajdonsagait a szinguláris értékek szeparálására vonatkozó mini-max elv alapján bizonyítjuk.

Ezután speciális valószínűségi mezőkön értelmezett L^2-terek közti integráloperátorokra alkalmazzuk az elmondottakat, így kapjuk a magfüggvény sorfejtését. Ennek alapján sok, bonyolult módon bevezetett statisztikai probléma (pl. korrespondancia-analízis, maximál-korreláció, regressziós feladatok) egységesen tárgyalható.

Szó lesz továbbá a hatványiteráció módszerének a sajátértékek és sajátfüggvények meghatározására történő általánosításáról. (A feltételes várhatóértékképzés operátorára ez a módszer az [5] cikkben kerül bevezetésre, itt az iteráció konvergenciáját tetszőleges korlátos operátorra bizonyítjuk.)

1. Spektrálfelbontással és szinguláris felbontással rendelkező legáltalánosabb operátorok

Legyen $\mathbf{H}$ és $\mathbf{H}'$ *Hilbert-tér* és $\mathbf{A}: \mathbf{H} \to \mathbf{H}'$ lineáris operátor.

1.1. Definíció. Az $\mathbf{A}$ operátort kompaktnak nevezzük, ha minden $\mathbf{H}$-beli korlátos halmazt $\mathbf{H}'$-beli prekompaktba (olyan halmaz, melynek lezárása kompakt) visz. Ezzel ekvivalens, hogy tetszőleges $\mathbf{H}$-beli korlátos sorozat képéből kiválasztható konvergens részsorozat. Szintén ekvivalens definíció, hogy $\mathbf{A}$ előáll véges rangú operátorok sorozatának operátornormában vett határértékeként.

Könnyen látható, hogy egy kompakt operátor egyben korlátos is, és ha $\mathbf{H}$ végtelen dimenziós *Hilbert-tér,* akkor az $\mathbf{I}: \mathbf{H} \to \mathbf{H}$ identitásoperátor nem kompakt (tekintsük ugyanis az egységgömbön ortogonális egységvektoroknak egy sorozatát, ez korlátos, de nem választható ki belőle konvergens részsorozat).

A *Hilbert-terek* elméletéből jól ismert tény, hogy egy kompakt operátor spektruma diszkrét, és a sajátértékek, ill. szinguláris értékek (ezek az 1.1., ill. 1.2. tételekben kerülnek definiálásra) — amennyiben nem csak véges sok különbözik nullától — nullához torlódnak. Ehhez a kompaktság nemcsak elégséges, hanem szükséges feltétel is (ez heurisztikusan onnan is látható, hogy a nullához közeli sajátértékekhez tartozó projekciók leválasztásával véges rangú operátorokat kapunk, amelyek operátornormában vett határértéke épp az eredeti operátor). Ezt rögtön precízen is kifejtjük, részletesebben l. KATO [10] könyvében (III. 6.26. tétel).

Így, ha egy $\mathbf{A}$ operátor szinguláris értékei nullától különböző valós számhoz (c-hez) torlódnak, akkor az $\mathbf{A} - c\mathbf{I}$ operátor szinguláris értékei nullához torlódnak,

így az kompakt lesz. Tehát tárgyalásunk alapját azok az operátorok képezik, amelyek előállíthatók egy kompakt operátor és az identitás-operátor konstansszorosának összegeként. Mivel ezek szinguláris felbontása a kompakt operátorokéra visszavezethető, a tételeket elég kompaktakra kimondani. Jelölje $\langle \cdot, \cdot \rangle$ a skaláris szorzást, $\| \cdot \|$ pedig a **H**-beli normát.

1.1. TÉTEL. (spektrálfelbontási tétel). Legyen $\mathbf{B}: \mathbf{H} \to \mathbf{H}$ kompakt, önadjungált, pozitív szemidefinit lineáris operátor. Ekkor van olyan

$$\{\varphi_n: n = 1, 2, \ldots\} \subset \mathbf{H}$$

ortonormált (o. n.) rendszer és λ_n nem-negatív, monoton nem növő számsorozat, hogy tetszőleges $\mathbf{u} \in \mathbf{H}$ elemre:

$$\mathbf{Bu} = \sum_{i=1}^{\infty} \lambda_i \langle \mathbf{u}, \varphi_i \rangle \varphi_i. \tag{1.1}$$

(A konvergencia **H**-beli normában, pontonként értendő), és

$$\lim_{n \to \infty} \lambda_n = 0, \tag{1.2}$$

amennyiben a különböző λ_i-k száma megszámlálhatóan végtelen.

Ebből következik az operátornormában való konvergencia is. Legyen ui. $\mathbf{B}_n \mathbf{u} = \sum_{i=1}^{n} \lambda_i \langle \mathbf{u}, \varphi_i \rangle \varphi_i$. Ekkor

$$\lim_{n \to \infty} \|\mathbf{B} - \mathbf{B}_n\| = \lim_{n \to \infty} \sup_{\|\mathbf{u}\| = 1} \|(\mathbf{B} - \mathbf{B}_n)\mathbf{u}\| = \lim_{n \to \infty} \lambda_1(\mathbf{B} - \mathbf{B}_n) = \lim_{n \to \infty} \lambda_{n+1} = 0,$$

ahol $\lambda_i(\,.\,)$ jelenti az argumentumban álló mátrix nagysága szerint i-edik sajátértékét.

Az (1.1)-beli erős konvergenciából következik az ún. gyenge konvergencia is, azaz tetszőleges $\mathbf{u} \in \mathbf{H}$ elemre:

$$\lim_{n \to \infty} \langle (\mathbf{B} - \mathbf{B}_n)\mathbf{u}, \mathbf{u} \rangle = 0.$$

A felbontásban szereplő számok (sajátértékek) egyértelműek, és az azonos sajátértékekhez tartozó sajátfüggvények által kifeszített alterek (sajátalterek) is egyértelműek. Könnyen látható, hogy a **B** operátor előáll az ezekre a sajátalterekre való vetítéseknek a sajátértékekkel súlyozott összegeként, vagyis a $\mathbf{B}_n$ véges rangú operátorok operátornormában vett határértékeként Ha a $\{\varphi_i\}_{i=1}^{\infty}$ rendszerhez még hozzáveszünk egy, az operátor magterében választott o. n. bázist (amennyiben ez megszámlálható), teljes o. n. rendszert kapunk. Így vizsgálatainkat elég szeparábilis *Hilbert-terekre* elvégezni

1.1. ÁLLÍTÁS (sajátértékek maximum-tulajdonsága).

$$\lambda_1 = \max_{\|\mathbf{u}\| = 1} \langle \mathbf{Bu}, \mathbf{u} \rangle = \max_{\|\mathbf{u}\| = 1} \|\mathbf{Bu}\| = \|\mathbf{B}\|, \quad \text{és eléretik} \quad \mathbf{u} = \varphi_1\text{-re},$$

$$\lambda_n = \max_{\substack{\|\mathbf{u}\| = 1 \\ \langle \mathbf{u}, \varphi_i \rangle = 0 \\ (i = 1, \ldots, n-1)}} \langle \mathbf{Bu}, \mathbf{u} \rangle = \max_{\substack{\|\mathbf{u}\| = 1 \\ \langle \mathbf{u}, \varphi_i \rangle = 0 \\ (i = 1, \ldots, n-1)}} \|\mathbf{Bu}\|, \quad n = 2, 3, \ldots \quad \text{és eléretik} \quad \mathbf{u} = \varphi_n\text{-re}. \tag{1.3}$$

A következő tétel úgy nyerhető az elsőből, mint a mátrixokra vonatkozó szinguláris felbontási tétel a spektrálfelbontásiból (l. [2]).

1.2. TÉTEL (szinguláris felbontási tétel). Legyen $\mathbf{A}: \mathbf{H} \to \mathbf{H}'$ kompakt lineáris operátor. Akkor létezik olyan s_n nem-negatív, monoton nem növő számsorozat és léteznek olyan $\{\psi_n\} \subset \mathbf{H}$, ill. $\{\varphi_n\} \subset \mathbf{H}'$ o. n. rendszerek, hogy tetszőleges $\mathbf{u} \in \mathbf{H}$-ra:

$$\mathbf{Au} = \sum_{i=1}^{\infty} s_i \langle \mathbf{u}, \psi_i \rangle \varphi_i \tag{1.4}$$

és

$$\lim_{n \to \infty} s_n = 0, \tag{1.5}$$

amennyiben a különböző szinguláris értékek száma megszámlálhatóan végtelen.

Könnyen látható, hogy $\mathbf{A}$ szinguláris felbontása az $\mathbf{AA}^*: \mathbf{H}' \to \mathbf{H}'$ ill. $\mathbf{A}^*\mathbf{A}: \mathbf{H} \to \mathbf{H}$ kompakt, önadjungált, pozitív szemidefinit operátorok spektrálfelbontásából nyerhető, ahol

$$\mathbf{AA}^* = \sum_{i=1}^{\infty} s_i^2 \langle \cdot, \varphi_i \rangle \varphi_i, \quad \mathbf{A}^*\mathbf{A} = \sum_{i=1}^{\infty} s_i^2 \langle \cdot, \psi_i \rangle \psi_i,$$

továbbá $\mathbf{A}^*$ szinguláris felbontása tetszőleges $\mathbf{v} \in \mathbf{H}'$ elemre:

$$\mathbf{A}^*\mathbf{v} = \sum_{i=1}^{\infty} s_i \langle \mathbf{v}, \varphi_i \rangle \psi_i,$$

ahol $\mathbf{A}^*$ jelöli az $\mathbf{A}$ operátor adjungáltját. Így az (1.5) állítás az (1.3) állításból következik és szintén következik a szinguláris értékek és az azonos szinguláris értékekhez tartozó sajátelemek által kifeszített ún. *izotróp alterek* egyértelműsége is. Az azonos s_i-khez tartozó izotróp alterek elemei közt a következő összefüggés áll fenn:

$$\mathbf{A}\psi_i = s_i \varphi_i, \quad \mathbf{A}^*\varphi_i = s_i \psi_i \quad (i = 1, 2, \ldots). \tag{1.6}$$

Azaz $\mathbf{A}$ előállítható ezen izotróp alterek közti affinitások összegeként, vagyis a véges rangú $\mathbf{A}_n = \sum_{i=j}^{n} s_i \langle \cdot, \psi_i \rangle \varphi_i$ operátorok operátornormában vett határértékeként. Ezt az bizonyítja, hogy $\mathbf{A} - \mathbf{A}_n$ legnagyobb szinguláris értéke $s_{n+1} \to 0$ $(n \to \infty)$. Hogy a legnagyobb szinguláris érték az operátor normája, azt a következő állítás biztosítja, amely az 1.1. állítás következménye:

1.2. ÁLLÍTÁS (szinguláris értékek maximumtulajdonsága).

$$s_1 = \max_{\|\mathbf{u}\|=1} \|\mathbf{Au}\| = \max_{\|\mathbf{v}\|=1} \|\mathbf{A}^*\mathbf{v}\|, \quad \text{és eléretik, ha} \quad \mathbf{u} = \psi_1, \ \mathbf{v} = \varphi_1,$$

$$s_n = \max_{\substack{\|\mathbf{u}\|=1 \\ \langle \mathbf{u}, \psi_i \rangle = 0 \\ (i=1,\ldots,n-1)}} \|\mathbf{Au}\| = \max_{\substack{\|\mathbf{v}\|=1 \\ \langle \mathbf{v}, \varphi_i \rangle = 0 \\ (i=1,\ldots,n-1)}} \|\mathbf{Av}\| \tag{1.7}$$

és a maximum az $\mathbf{u} = \psi_n$, $\mathbf{v} = \varphi_n$ értékpáron vétetik fel.

A szinguláris értékek egy másfajta, a skaláris szorzatok segítségével megfogalmazható maximumtulajdonságnak is eleget tesznek:

$$s_1 = \max_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} \langle \mathbf{Au}, \mathbf{v}\rangle = \max_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} \langle \mathbf{u}, \mathbf{A}^*\mathbf{v}\rangle, \quad \text{és eléretik, ha} \quad \mathbf{u} = \psi_1,\ \mathbf{v} = \varphi_1,$$

$$(1.8) \qquad s_n = \max_{\substack{\|\mathbf{u}\|=\|\mathbf{v}\|=1 \\ \langle \mathbf{u}, \psi_i\rangle = 0 \\ \langle \mathbf{v}, \varphi_i\rangle = 0 \\ (i=1,\ldots,n-1)}} \langle \mathbf{Au}, \mathbf{v}\rangle = \max_{\substack{\|\mathbf{u}\|=\|\mathbf{v}\|=1 \\ \langle \mathbf{u}, \psi_i\rangle = 0 \\ \langle \mathbf{v}, \varphi_i\rangle = 0 \\ (i=1,\ldots,n-1)}} \langle \mathbf{u}, \mathbf{A}^*\mathbf{v}\rangle,$$

és a maximum az $\mathbf{u}=\psi_n$, $\mathbf{v}=\varphi_n$ értékpárokra vétetik fel.

2. Szinguláris értékekre vonatkozó szeparációs tételek

Az 1.1. ill. 1.2. állítás szeparációs tételnek is tekinthető, ezek azonban a most következő tételek igen speciális esetei. Az első, ún. mini-max tételt be is bizonyítjuk, a bizonyítás lényegében véges dimenziós megfontolásokon alapszik. A további tételek is a mátrixok szinguláris értékeire, ill. nyomára vonatkozó állítások (l. [14]) bizonyításaihoz hasonlóan nyerhetők.

2.1. TÉTEL (mini-max elv). Legyen az $\mathbf{A}: \mathbf{H} \to \mathbf{H}'$ kompakt lineáris operátor szinguláris felbontása $\sum_{i=1}^{\infty} s_i \langle \cdot, \psi_i\rangle \varphi_i$ és k tetszőleges pozitív egész olyan, hogy $s_{k+1} < s_k$. Akkor

$$(2.1) \qquad s_{k+1} = \inf_{\substack{\mathscr{A}_k \subset \mathbf{H} \\ k\text{-dim. altér}}} \sup_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u} \perp \mathscr{A}_k}} \|\mathbf{Au}\|,$$

és a fenti nyeregpont tetszőleges olyan elem lehet, amely benne van az s_{k+1}-hez tartozó izotróp altérben, és merőleges az s_k-hoz tartozóra.

Bizonyítás. Jelölje $U_i = \{\psi_1, \ldots, \psi_i\}$ az első i $\mathbf{H}$-beli sajátelem által kifeszített alteret és z_{k+1} (2.1) jobb oldalát.

Mivel $\mathscr{A}_k = U_k$ választással $\sup_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u} \perp \mathscr{A}_k}} \|\mathbf{Au}\| = s_{k+1}$, $s_{k+1} \geqq z_{k+1}$. Másrészt belátjuk, hogy $s_{k+1} \leqq z_{k+1}$, amiből $s_{k+1} = z_{k+1}$ már következik és ezzel a tételt már be is bizonyítottuk.

Először belátjuk, hogy tetszőleges $\mathscr{A}_k \subset \mathbf{H}$ k-dimenziós altérhez van olyan $\mathbf{x} \in U_{k+1}$, hogy $\|\mathbf{x}\| = 1$ és $\mathbf{x} \perp \mathscr{A}_k$. Keressük ui. $\mathbf{x}$-et $\sum_{i=1}^{k+1} a_i \psi_i$, $\sum_{i=1}^{k+1} a_i^2 = 1$ alakban. Legyen $\mathscr{A}_k = \{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ tetszőleges, ahol $\mathbf{e}_1, \ldots, \mathbf{e}_k$ o. n. rendszer. Olyan $\mathbf{x}$-et keresünk, hogy

$$\langle \mathbf{x}, \mathbf{e}_j\rangle = 0 \quad (j = 1, \ldots, k)$$

teljesüljön (ekkor könnyen látható, hogy $\langle \mathbf{x}, \mathbf{y}\rangle = 0$ tetszőleges $\mathbf{y} \in \mathscr{A}_k$ esetén). Az a_i $(i=1, \ldots, k+1)$ együtthatókat a

$$\sum_{i=1}^{k+1} a_i \langle \psi_i, \mathbf{e}_j\rangle = 0 \quad (j = 1, \ldots, k)$$

homogén lineáris egyenletrendszer megoldása adja. Mivel az ismeretlenek száma $(k+1)$ nagyobb, mint az egyenletek száma (k), így mindig létezik nemtriviális megoldás, sőt minden megoldással együtt annak konstansszorosa is megoldás. Így még az a_i együtthatókra tett $\sum_{i=1}^{k+1} a_i^2 = 1$ mellékfeltétel is kielégíthető.

Egy fenti $\mathbf{x}$-et választva, az (1.6) összefüggésekkel

$$\|\mathbf{Ax}\|^2 = \sum_{i=1}^{k+1} a_i^2 \|\mathbf{A}\psi_i\|^2 = \sum_{i=1}^{k+1} a_i^2 \|s_i \varphi_i\|^2 = \sum_{i=1}^{k+1} a_i^2 s_i^2 \geqq \sum_{i=1}^{k+1} a_i^2 s_{k+1}^2 = s_{k+1}^2,$$

mivel $s_1 \geqq s_2 \geqq \ldots \geqq s_k > s_{k+1}$. Így ilyen $\mathbf{x}$-ekre

$$\sup \|\mathbf{Ax}\|^2 \geqq s_{k+1}^2,$$

ahonnan

$$z_{k+1} \geqq s_{k+1}.$$

A bizonyítás elején azt is láttuk, hogy az infimum $\mathscr{A}_k = U_k$ választással éretik el, és tetszőleges olyan elemen felvétetik, amely benne van az s_{k+1}-hez tartozó izotróp altérben, és merőleges az s_k-hoz tartozóra.

2.1. KÖVETKEZMÉNY. $\mathscr{P}_k$-val jelölve a k-adrangú, $\mathbf{H}$-beli vetítések halmazát (azaz azon vetítésekét, melyek $\mathbf{H}$-nak valamely k-dimenziós alterére vetítenek), könnyen látható, hogy a 2.1. tétel állítása a következő alakban is kimondható:

$$s_{k+1} = \inf_{\mathbf{P} \in \mathscr{P}_k} \|\mathbf{A}(\mathbf{I}-\mathbf{P})\|, \tag{2.2}$$

hiszen tetszőleges $\mathbf{P} \in \mathscr{P}_k$-nak megfeleltethető egy $\mathscr{A}_k \subset \mathbf{H}$ k-dimenziós altér és

$$\|\mathbf{A}(\mathbf{I}-\mathbf{P})\| = \sup_{\|\mathbf{u}\|=1} \|\mathbf{A}(\mathbf{I}-\mathbf{P})\mathbf{u}\| = \sup_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u} \perp \mathscr{A}_k}} \|\mathbf{Au}\|,$$

ugyanis $\mathbf{I}-\mathbf{P}$ éppen az $\mathscr{A}_k$ altér ortogonális kiegészítőjére vetít.

2.2. KÖVETKEZMÉNY.

$$\sum_{i=1}^{k} s_i = \max_{\substack{\{\mathbf{u}_1, \ldots, \mathbf{u}_k\} \subset \mathbf{H} \\ \{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subset \mathbf{H}' \\ \text{o. n. rendszerek}}} \sum_{i=1}^{k} \langle \mathbf{Au}_i, \mathbf{v}_i \rangle = \max_{\substack{\{\mathbf{u}_1, \ldots, \mathbf{u}_k\} \subset \mathbf{H} \\ \{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subset \mathbf{H}' \\ \text{o. n. rendszerek}}} \sum_{i=1}^{k} \langle \mathbf{A}^* \mathbf{v}_i, \mathbf{u}_i \rangle, \tag{2.3}$$

és a maximum eléretik, ha $\mathbf{u}_i = \psi_i$, $\mathbf{v} = \varphi_i (i = 1, \ldots, k)$.

Ebből az 1.1. és 1.2. szekvenciális állítások is következnek. A bizonyítás analóg az állítás véges dimenziós megfelelőjének (l. [14]) bizonyításával.

2.2. TÉTEL. Legyen $\mathbf{A}: \mathbf{H} \to \mathbf{H}'$ tetszőleges, $\mathbf{A}_k: \mathbf{H} \to \mathbf{H}'$ pedig k-rangú operátor. Akkor

$$s_i(\mathbf{A}-\mathbf{A}_k) \geqq s_{k+i}(\mathbf{A}) \quad (i = 1, 2, \ldots), \tag{2.4}$$

ahol $s_i(\cdot)$ jelenti az argumentumban álló operátor nagyság szerint i-edik szinguláris értékét.

Bizonyítás. Jelölje rang $(\cdot)$ egy leképezés rangját, dim $(\cdot)$ pedig egy altér dimenzióját. Először belátjuk, hogy

$$s_j(\mathbf{A}) = \inf_{\operatorname{rang}(\mathbf{A}_{j-1}) \leqq j-1} \|\mathbf{A}-\mathbf{A}_{j-1}\|, \quad j = 1, 2, \ldots \quad \mathbf{A}_0 = 0. \tag{2.5}$$

Ui. $\mathbf{A}_{j-1} = \sum_{i=1}^{j-1} s_i \langle \cdot, \psi_i \rangle \varphi_i$ választással $\|\mathbf{A}-\mathbf{A}_{j-1}\| = s_j(\mathbf{A})$, így

$$\inf_{\operatorname{rang}(\mathbf{A}_{j-1}) \leqq j-1} \|\mathbf{A}-\mathbf{A}_{j-1}\| \leqq s_j(\mathbf{A}).$$

Most belátjuk a fordított irányú egyenlőtlenséget is. Legyen $\mathbf{A}_{j-1}$ tetszőleges, legfeljebb $j-1$ rangú operátor. Jelölje $\mathcal{G}$ ennek magterét. Akkor

$$\|\mathbf{A}-\mathbf{A}_{j-1}\| = \sup_{\|\mathbf{u}\|=1} \|(\mathbf{A}-\mathbf{A}_{j-1})\mathbf{u}\| \geqq \sup_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u} \in \mathcal{G}}} \|\mathbf{A}\mathbf{u}\| \geqq \sup_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u} \in \mathcal{G} \\ \tilde{\mathcal{G}} \subset \mathcal{G} \\ \dim(\tilde{\mathcal{G}}^{\perp}) = j-1}} \|\mathbf{A}\mathbf{u}\| \geqq$$

$$\geqq \inf_{\substack{\mathcal{M} \subset \mathbf{H} \\ \dim(\mathcal{M}) = j-1}} \sup_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u} \perp \mathcal{M}}} \|\mathbf{A}\mathbf{u}\| = s_j(\mathbf{A}),$$

ahol $^{\perp}$ jelöli az ortogonális kiegészítő alteret, és a fenti $\tilde{\mathcal{G}}$ létezik, hisz $\mathcal{G}$ dimenziója legalább $j-1$. Az utolsó egyenlőség nem más, mint a mini-max tétel. Ezek után a (2.5) segédállítást az $\mathbf{A}-\mathbf{A}_k$ operátorra alkalmazva:

$$\inf_{\operatorname{rang}(\mathbf{A}_k) \leqq k} s_i(\mathbf{A}-\mathbf{A}_k) = \inf_{\operatorname{rang}(\mathbf{A}_k) \leqq k} \inf_{\operatorname{rang}(\mathbf{B}_{i-1}) \leqq i-1} \|\mathbf{A}-\mathbf{A}_k-\mathbf{B}_{i-1}\| =$$

$$= \inf_{\operatorname{rang}(\mathbf{C}_{i+k-1}) \leqq i+k-1} \|\mathbf{A}-\mathbf{C}_{i+k-1}\| = s_{i+k}(\mathbf{A}),$$

ahol a (2.5) összefüggést az első és az utolsó egyenlőségben is felhasználtuk. A második egyenlőség pedig abból adódik, hogy az $\mathbf{A}_k + \mathbf{B}_{i-1} = \mathbf{C}_{i+k-1}$ leképezés rangja $\leqq i+k-1$. Az is látható, hogy az $\mathbf{A}_k = \sum_{i=1}^{k} s_i \langle \cdot, \psi_i \rangle \varphi_i$ operátorra az egyenlőség mindenütt eléretik.

2.1. Definíció. Egy kompakt lineáris operátor $\|\cdot\|_{\text{un}}$ normáját *unitér invariáns normának* nevezzük, ha a szokásos normatulajdonságok mellett még igaz rá az is, hogy tetszőleges $\mathbf{A}: \mathbf{H} \to \mathbf{H}'$ kompakt, és $\mathbf{U}: \mathbf{H} \to \mathbf{H}$ $\mathbf{V}: \mathbf{H}' \to \mathbf{H}'$ unitér ($\mathbf{U}\mathbf{U}^* = \mathbf{I}$, $\mathbf{V}\mathbf{V}^* = \mathbf{I}$) lineáris operátor esetén

$$\|\mathbf{A}\|_{\text{un}} = \|\mathbf{U}\mathbf{A}\mathbf{V}\|_{\text{un}}.$$

Azaz egy kompakt lineáris operátor unitér invariáns normája csak az operátor szinguláris értékeitől függ, hisz válasszuk speciálisan $\mathbf{U}$-nak és $\mathbf{V}$-nek az $\mathbf{A}$ szinguláris felbontásában fellépő unitér leképezések (melyeket mindkét irányban végtelen mátrix alakjában felírva az oszlopokban a $\mathbf{H}$-, ill. $\mathbf{H}'$-beli sajátelemek állnak) adjungáltját.

Mivel a szinguláris értékek pozitívak, a háromszög-egyenlőtlenség miatt az unitér invariáns norma a szinguláris értékek monoton függvénye (ez úgy értendő, hogy bármely szinguláris értéket növelve az unitér invariáns norma is nő). Így a 2.2. tételből a következő tétel bizonyítható:

2.3. TÉTEL. Tetszőleges unitér invariáns normára adott $\mathbf{A}=\sum_{i=1}^{\infty} s_i\langle\cdot, \psi_i\rangle\varphi_i$ kompakt lineáris operátor esetén $\|\mathbf{A}-\mathbf{A}_k\|_{un}$ akkor minimális $\mathbf{A}_k$-ban (ahol $\mathbf{A}_k$ legfeljebb k-rangú operátor), ha $\mathbf{A}_k=\sum_{i=1}^{k} s_i\langle\cdot, \psi_i\rangle\varphi_i$.

Pl. a szokásos operátornormára

$$\|\mathbf{A}-\mathbf{A}_k\| = s_1(\mathbf{A}-\mathbf{A}_k) \geqq s_{k+1}(\mathbf{A}),$$

és az egyenlőség eléretik, ha $\mathbf{A}_k$ $\mathbf{A}$ szinguláris felbontásának k-adik szelete, a minimum értéke pedig $s_{k+1}(\mathbf{A})$.

2.2. Definíció. Legyen $\mathbf{A}: \mathbf{H}\to\mathbf{H}'$ olyan kompakt lineáris operátor, melyre $\sum_{i=1}^{\infty} s_i^2 < +\infty$. Ekkor a $\sqrt{\sum_{i=1}^{\infty} s_i^2}$ számot az $\mathbf{A}$ operátor *kettős normájának* nevezzük és $|||\mathbf{A}|||$-val jelöljük. Természetesen $\|\mathbf{A}\| \leqq |||\mathbf{A}|||$.

Definíció szerint $|||\cdot|||$ is unitér invariáns norma. Ezzel tetszőleges $\mathbf{A}_k$ k-rangú operátorra

$$|||\mathbf{A}-\mathbf{A}_k|||^2 = \sum_{i=1}^{\infty} s_i^2(\mathbf{A}-\mathbf{A}_k) \geqq \sum_{i=k+1}^{\infty} s_i^2(\mathbf{A}),$$

és a minimum megint csak akkor éretik el, ha $\mathbf{A}_k$ az $\mathbf{A}$ operátor szinguláris felbontásának k-adik szelete.

2.3. Definíció. Ha az $\mathbf{A}$ kompakt lineáris operátor szinguláris értékeinek négyzetösszege konvergens, akkor $\mathbf{A}$-t *Hilbert—Schmidt-operátornak* nevezzük.

Megjegyezzük, hogy ha $\mathbf{A}$ *Hilbert—Schmidt-operátor,* akkor az $\mathbf{A}_n = \sum_{i=1}^{n} s_i^2\langle\cdot, \psi_i\rangle\varphi_i$ sorozat kettős normában is konvergál $\mathbf{A}$-hoz ($n\to\infty$).

3. Integráloperátorok

Legyenek $(\mathbf{X}, \mathscr{A}, P)$ és $(\mathbf{Y}, \mathscr{B}, Q)$ valószínűségi mezők és jelölje $L^2(\mathbf{X}, \mathscr{A}, P)$ és $L^2(\mathbf{Y}, \mathscr{B}, Q)$ az ezeken értelmezett nulla várható értékű, véges szórású, valós értékű valószínűségi változókat. Ezek *Hilbert-teret* alkotnak a kovarianciával mint skaláris szorzattal. Ezentúl ilyen *Hilbert-terek* közti lineáris operátorokkal foglalkozunk. (Valójában a tételeket kimondhatnánk tetszőleges mértékterek feletti L^2-terekre is.)

Jelölje $(\mathbf{X}*\mathbf{Y}, \sigma(\mathscr{A}*\mathscr{B}), P*Q)$ a szorzatteret és legyen $\mathbf{g}\in L^2(\mathbf{X}*\mathbf{Y}, \sigma(\mathscr{A}*\mathscr{B}), P*Q)$, a szorzatmérték szerint véges szórású valószínűségi változó. Ez a $\mathbf{g}$ mint magfüggvény (1 valószínűséggel) egyértelműen meghatároz egy

$$\mathbf{A}: L^2(\mathbf{Y}, \mathscr{B}, Q) \to L^2(\mathbf{X}, \mathscr{A}, P)$$

lineáris operátort a következőképpen: tetszőleges $\varphi\in L^2(\mathbf{Y}, \mathscr{B}, Q)$ valószínűségi változóhoz hozzárendeljük azt a ψ valószínűségi változót, melyre

$$\psi(x) = \int_{\mathbf{Y}} \mathbf{g}(x, y)\varphi(y)Q(dy). \tag{3.1}$$

3.1. ÁLLÍTÁS. Ez a ψ eleme az $L^2(\mathbf{X}, \mathscr{A}, P)$ térnek.

Bizonyítás. Fubini tételét és a *Cauchy—Schwarz egyenlőtlenséget* alkalmazva

$$\int_X |\psi(x)|^2 P(dx) = \int_X \left| \int_Y \mathbf{g}(x, y)\varphi(y) Q(dy)\right|^2 P(dx) \leqq$$

$$\leqq \int_X \int_Y |\mathbf{g}(x, y)|^2 P(dx) Q(dy) \cdot \int_X \int_Y |\varphi(y)|^2 Q(dy) P(dx) = \|\mathbf{g}\|^2 \cdot \|\varphi\|^2 < +\infty,$$

ahol $\|\cdot\|$ a megfelelő téren vett L^2-normát jelenti.

A bizonyításból az is kijön, hogy

$$\|\mathbf{A}\| \leqq \|\mathbf{g}\|. \tag{3.2}$$

Az is látható, hogy a $\mathbf{g}$ magfüggvény lényegében véve egyértelműen meghatározza az $\mathbf{A}$ operátort.

3.1. TÉTEL. A fenti $\mathbf{g} \in L^2(\mathbf{X} * \mathbf{Y}, \sigma(\mathscr{A} * \mathscr{B}), P * Q)$ magfüggvény által generált integráloperátor *Hilbert—Schmidt operátor*. Továbbá, ha $\sum_{i=1}^{\infty} s_i \langle \cdot, \psi_i \rangle \varphi_i$ jelöli $\mathscr{A}$ szinguláris felbontását, akkor

$$|||\mathbf{A}|||^2 = \sum_{i=1}^{\infty} s_i^2 = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |\langle \mathbf{A}\psi_i, \varphi_j \rangle|^2. \tag{3.3}$$

(A tétel bizonyítását l. [9]-ben.)

Az $\mathbf{A}$ integráloperátor szinguláris felbontása alapján könnyen látható, hogy $\mathbf{A}$ magfüggvénye is a következő, L^2-normában konvergens sorba fejthető:

3.2. TÉTEL *(Karhunen—Loève sorfejtés)*. Legyen $\mathbf{g} \in L^2(\mathbf{X} * \mathbf{Y}, \sigma(\mathscr{A} * \mathscr{B}), P * Q)$. Akkor létezik olyan s_i nem-negatív, monoton nem növő számsorozat és olyan $\{\psi_i\} \subset L^2(\mathbf{X}, \mathscr{A}, P)$, ill. $\{\varphi_i\} \subset L^2(\mathbf{Y}, \mathscr{B}, Q)$ korrelálatlan, egységszórású valószínűségi változó sorozatok, hogy ezekkel

$$\mathbf{g}(x, y) = \sum_{i=1}^{\infty} s_i \psi_i(x) \varphi_i(y),$$

azaz $\mathbf{g}$ végtelen diádösszegként állítható elő, és a sor a szorzattéren értelmezett L^2-normában konvergens.

Bizonyítás. Legyen a $\mathbf{g}$ által generált $\mathbf{A}$ integráloperátor szinguláris felbontása $\sum_{i=1}^{\infty} s_i \langle \cdot, \psi_i \rangle \varphi_i$. Könnyen látható, hogy az ezekkel a mennyiségekkel definiált $\mathbf{g}'(x, y) = \sum_{i=1}^{\infty} s_i \psi_i(x) \varphi_i(y)$ valószínűségi változóra $\mathbf{g}' \in L^2(\mathbf{X} * \mathbf{Y}, \sigma(\mathscr{A} * \mathscr{B}), P * Q)$, mivel $\sum_{i=1}^{\infty} s_i^2 < +\infty$. Az is látható, hogy a $\mathbf{g}'$ által generált $\mathbf{A}'$ integráloperátor hatása a báziselemeken ugyanaz, mint $\mathbf{A}$ hatása, tehát $\mathbf{A} = \mathbf{A}'$. Mivel a magfüggvény az

integráloperátort lényegében egyértelműen meghatározza, $\mathbf{g}=\mathbf{g}'$ 1 valószínűséggel. Az L^2-normában való konvergencia abból következik, hogy

$$\|\mathbf{g}-\mathbf{g}_n\|^2 = |||\mathbf{A}-\mathbf{A}_n|||^2 = \sum_{i=n+1}^{\infty} s_i^2 \to 0, \quad \text{ha} \quad n \to \infty, \tag{3.4}$$

ahol $\mathbf{A}_n$ a $\mathbf{g}_n = \sum_{i=1}^{n} s_i \psi_i(x)\varphi_i(y)$ magfüggvény által meghatározott integráloperátor.

Megjegyezzük, hogy bizonyos feltételek mellett (pl. $\mathbf{H}=\mathbf{H}'$ *szeparábilis Hilbert-terek* és $\mathbf{A}$ önadjungált) 1 valószínűségű konvergencia is igaz, ez *Mercer tétele* (l. [13]). Erről és a *Karhunen—Loève sorfejtés* idősorokra vonatkozó alakjáról részletesebben ír FRITZ JÓZSEF [7], GULYÁS OTTÓ [8] és VARGA LÁSZLÓ [15] dolgozatában.

A 2.3. tétel alapján a *Karhunen—Loève sorfejtésnek* egy fontos optimumtulajdonsága adódik:

Legyen $\mathbf{g} \in L^2(\mathbf{X} * \mathbf{Y}, \sigma(\mathscr{A} * \mathscr{B}), P * Q)$ és $k>0$ adott egész. Keresendők olyan $\alpha_i \in L^2(\mathbf{X}, \mathscr{A}, P)$, $\beta_i \in L^2(\mathbf{Y}, \mathscr{B}, Q)$ egységszórású valószínűségi változók és z_i pozitív számok $(i=1, 2, \ldots, k)$, hogy az ezekből képzett diádösszegre:

$$E_{P*Q}[\mathbf{g} - \sum_{i=1}^{k} z_i \alpha_i \beta_i]^2 \to \min \quad \text{legyen.} \tag{3.5}$$

3.3. TÉTEL. A (3.5) minimumfeladat megoldását $\mathbf{g}$ *Karhunen—Loeve sorfejtésének* első k tagja adja, azaz

$$z_i = s_i, \quad \alpha_i = \psi_i, \quad \beta_i = \varphi_i \quad (i = 1, 2, \ldots, k)$$

1 valószínűséggel. Ha $s_k > s_{k+1}$, akkor a $\{\psi_1, \ldots, \psi_k\} \subset L^2(\mathbf{X}, \mathscr{A}, P)$, ill. $\{\varphi_1, \ldots, \varphi_k\} \subset L^2(\mathbf{Y}, \mathscr{B}, Q)$ alterek egyértelműek, a minimum értéke pedig mindig $\sum_{i=k+1}^{\infty} s_i^2$.

Bizonyítás. Jelölje a $\mathbf{g}_k = \sum_{i=1}^{k} z_i \alpha_i \beta_i$ magfüggvény által meghatározott integráloperátort $\mathbf{A}_k$. Ez k-rangú lineáris operátor. A 2.3. tétel értelmében

$$E_{P*Q}[\mathbf{g}-\mathbf{g}_k]^2 = \|\mathbf{g}-\mathbf{g}_k\|^2 = |||\mathbf{A}-\mathbf{A}_k|||^2,$$

és ez minimális, ha $\mathbf{A}_k$ az $\mathbf{A}$ operátor szinguláris felbontásának k-adik szelete. Mivel $\mathbf{A}_k$ $\mathbf{g}_k$-t egyértelműen meghatározza, az állítás következik. Az is adódik, hogy a minimumot adó valószínűségi változók mindkét térben páronként ortogonálisak.

Megjegyezzük, hogy amennyiben $\mathbf{g}$ szimmetrikus, *Karhunen—Loève sorfejtése* 1 valószínűséggel is konvergens (ez *Mercer tételének* a magfüggvényre vonatkozó alakja, l. [13]).

A 3.3. tétel a *Karhunen—Loève sorfejtés* optimumtulajdonságát mondja ki.

4. A Karhunen—Loève sorfejtés optimumtulajdonságának alkalmazásai

Legyenek $L^2(\mathbf{X}, \mathscr{A}, P)$ és $L^2(\mathbf{Y}, \mathscr{B}, Q)$ az előbbi *Hilbert-terek*, és tekintsük az alapterek szorzatát az R valószínűségi mértékkel. Tegyük fel, hogy R abszolút folytonos a $P*Q$ szorzatmértékre. Ekkor a *Radon—Nikodym-tétel* alapján lényegében egyértelműen (0 mértékű halmazon felvett értékektől eltekintve) létezik egy

$$\mathbf{r}: \mathbf{X}*\mathbf{Y} \to \mathbf{R}$$

Borel-mérhető függvény, hogy minden $\mathbf{C}\in\sigma(\mathscr{A}*\mathscr{B})$ *Borel-halmazra:*

$$R(\mathbf{C}) = \int_{\mathbf{C}} \mathbf{r}(x, y)\, P(dx)\, Q(dy).$$

Ezt az $\mathbf{r}$ valószínűségi változót nevezzük R $P*Q$ szerinti *Radon—Nikodym deriváltjának*.

Tegyük fel továbbá azt is, hogy R megfelelő marginális eloszlásai P és Q, azaz tetszőleges $\mathbf{C}_X\in\mathscr{A}$ és $\mathbf{C}_Y\in\mathscr{B}$ *Borel-halmazokra:*

$$R(\mathbf{C}_X * \mathbf{Y}) = \int_{\mathbf{C}_X}\int_{\mathbf{Y}} \mathbf{r}(x, y)\, Q(dy)\, P(dx) = P(\mathbf{C}_X)$$

és

$$R(\mathbf{X} * \mathbf{C}_Y) = \int_{\mathbf{X}}\int_{\mathbf{C}_Y} \mathbf{r}(x, y)\, P(dx)\, Q(dy) = Q(\mathbf{C}_Y).$$

Értelmezzük most a

$$\mathbf{P}_X: L^2(\mathbf{Y}, \mathscr{B}, Q) \to L^2(\mathbf{X}, \mathscr{A}, P)$$

feltételes várható érték képzés operátorát a szokásos módon, azaz tetszőleges $\varphi\in L^2(\mathbf{Y}, \mathscr{B}, Q)$ valószínűségi változóhoz hozzárendeljük a

$$\psi = E_R(\varphi|\mathscr{A})\in L^2(\mathbf{X}, \mathscr{A}, P) \tag{4.1}$$

valószínűségi változót, ahol $E_R(\,.|.\,)$ az R együttes eloszlás szerint vett feltételes várható értéket jelenti.

Regresszióanalízisből jól ismert a feltételes várható érték optimumtulajdonsága, amely szerint rögzített $\varphi\in L^2(\mathbf{Y}, \mathscr{B}, Q)$ esetén a

$$\|\varphi - \psi\|^2 = E_R|\varphi - \psi|^2 \to \min \tag{4.2}$$

feladat megoldását a $\psi = E_R(\varphi|\mathscr{A})$ feltételes várható érték adja. Következésképpen, a feltételes várható érték képzés operátora úgy fogható fel, mint a szorzattérnek az $L^2(\mathbf{X}, \mathscr{A}, P)$ altérre való vetítése (ezért is jelöltük $\mathbf{P}_X$-szel). Hasonlóan értelmezhető a

$$\mathbf{P}_Y: L^2(\mathbf{X}, \mathscr{A}, P) \to L^2(\mathbf{Y}, \mathscr{B}, Q)$$

feltételes várható érték képzés operátora is, amiről látni fogjuk, hogy $\mathbf{P}_X$-nek az adjungáltja. Mivel ezek vetítések, $\|\mathbf{P}_X\| \leqq 1$ és $\|\mathbf{P}_Y\| \leqq 1$, azaz $\mathbf{P}_X$ és $\mathbf{P}_Y$ szinguláris értékei 1-nél nem nagyobbak.

A szinguláris felbontás létezéséhez azonban még fel kell tenni, hogy $\mathbf{P}_X$ és $\mathbf{P}_Y$ kompakt. Erre elégséges feltétel pl. az $\mathbf{r}$ magfüggvény négyzetintegrálhatósága (l. [12]),

azaz hogy

$$\int_X \int_Y |\mathbf{r}(x, y)|^2 P(dx) Q(dy) < +\infty. \tag{4.3}$$

Könnyen látható, hogy az $\mathbf{r}$ magfüggvény által meghatározott Hilbert—Schmidt-operátor éppen $\mathbf{P}_X$, ui.:

$$\psi(x) = \int_Y \mathbf{r}(x, y)\varphi(y) Q(dy) = E_R(\varphi|\mathscr{A})(x) \tag{4.4}$$

1 valószínűséggel, hiszen a feltételek miatt $\int_Y \mathbf{r}(x, y)\ Q(dy) = P(\mathbf{X}) = 1$. (Vigyázat, félreértést ne okozzon: $\mathbf{r}$ nem az együttes sűrűségfüggvény, hanem $\mathbf{r}(x, y) = f(x, y)/f_1(x) f_2(y)$, ahol f jelöli az együttes, f_1 és f_2 pedig a marginális sűrűségeket a *Lebesgue-mérték* szerint.)

4.1. ÁLLÍTÁS. Ha $\mathbf{H}$ és $\mathbf{H}'$ szeparábilis *Hilbert-terek,* akkor a $\mathbf{H} \to \mathbf{H}'$ korlátos lineáris operátorok és a $\mathbf{H} * \mathbf{H}' \to \mathbf{R}$ bilineáris függvények között egy-egy értelmű megfeleltetés létesíthető.

Bizonyítás. Legyen $\mathbf{A}: \mathbf{H} \to \mathbf{H}'$ korlátos lineáris operátor, $\{\mathbf{e}_i\} \subset \mathbf{H}$ és $\{\mathbf{f}_j\} \subset \mathbf{H}'$ pedig teljes o. n. bázisok. Az $(\mathbf{e}_i, \mathbf{f}_j)$ párok teljes o. n. rendszert alkotnak $\mathbf{H} \otimes \mathbf{H}'$-ben ahol $\otimes$ a két tér tenzorszorzatát jelöli. Elég tehát ezeken megadni egy B bilineáris függvényt. Legyen

$$B(\mathbf{e}_i, \mathbf{f}_j) = \langle \mathbf{A}\mathbf{e}_i, \mathbf{f}_j \rangle.$$

Megfordítva, ha B adott, akkor legyen

$$\mathbf{A}\mathbf{e}_i = \sum_{k=1}^{\infty} B(\mathbf{e}_i, \mathbf{f}_k)\mathbf{f}_k.$$

Ekkor

$$\langle \mathbf{A}\mathbf{e}_i, \mathbf{f}_j \rangle = \Big\langle \sum_{k=1}^{\infty} B(\mathbf{e}_i, \mathbf{f}_k)\mathbf{f}_k,\ \mathbf{f}_j \Big\rangle = B(\mathbf{e}_i, \mathbf{f}_j),$$

tehát egy-egy értelmű megfeleltetést létesítettünk. A $B(\mathbf{e}_i, \mathbf{f}_j)$ számokat egy mindkét irányban végtelen mátrixba rendezve a véges mátrixok analógiájára az $\mathbf{A}$ (szeparábilis *Hilbert-terek* közt közvetítő) korlátos, lineáris operátor mátrixalakját kapjuk az $\{\mathbf{e}_i\}$ ill. $\{\mathbf{f}_j\}$ bázisokban felírva.

4.1. *Definíció.* A $C: L^2(\mathbf{X}, \mathscr{A}, P) * L^2(\mathbf{Y}, \mathscr{B}, Q) \to \mathbf{R}$ bilineáris függvényt kovarianciafüggvénynek nevezzük, ha tetszőleges $\psi \in L^2(\mathbf{X}, \mathscr{A}, P)$, $\varphi \in L^2(\mathbf{Y}, \mathscr{B}, Q)$ valószínűségi változó párhoz azok kovarianciáját rendeli hozzá, azaz

$$C(\psi, \varphi) = \int_{\mathbf{X}*\mathbf{Y}} \psi(x)\varphi(y) R(dx\, dy) = \int_X \int_Y \psi(x)\varphi(y)\mathbf{r}(x, y) P(dx) Q(dy),$$

ahol $\mathbf{r}$ az R mérték $P * Q$ szerinti *Randon—Nikodym deriváltja.*

A definícióból az is látható, hogy a kovarianciafüggvény szimmetrikus.

4.2. ÁLLÍTÁS. A fenti terekben (amennyiben azok szeparábilisek, vagy szeparábilis altereikben, ha nem azok) a kovarianciafüggvénynek, mint bilineáris függvénynek megfelelő lineáris operátor a feltételes várható érték képzés operátora.

Bizonyítás. Legyen $\psi \in L^2(\mathbf{X}, \mathscr{A}, P)$ és $\varphi \in L^2(\mathbf{Y}, \mathscr{B}, Q)$. Akkor a skaláris szorzást mint kovarianciát felírva (hogy mely térben tekintjük a skaláris szorzást, alsó index jelöli) és *Fubini tételét* alkalmazva a $\mathbf{P}_Y$ lineáris operátor által meghatározott bilineáris függvény:

$$\langle \mathbf{P}_Y \psi, \varphi \rangle_Y = \int_Y \Big(\int_X \psi(x) \mathbf{r}(x, y) P(dx) \Big) \varphi(y) Q(dy) =$$

$$= \int_X \int_Y \psi(x) \varphi(y) \mathbf{r}(x, y) P(dx) Q(dy),$$

ami éppen a kovarianciafüggvény. A megfeleltetés a 4.1. állítás alapján egy-egy értelmű.

4.3. ÁLLÍTÁS. $P_X^* = P_Y$ és $P_Y^* = P_X$.

Bizonyítás. Elég csak az egyik állítást belátni. Legyen ismét $\psi \in L^2(\mathbf{X}, \mathscr{A}, P)$ és $\varphi \in L^2(\mathbf{Y}, \mathscr{B}, Q)$.

A (4.4) összefüggést és *Fubini tételét* felhasználva

$$\langle \mathbf{P}_Y \psi, \varphi \rangle_Y = \int_Y \Big(\int_X \psi(x) \mathbf{r}(x, y) P(dx) \Big) \varphi(y) Q(dy) =$$

$$= \int_X \psi(x) \Big(\int_Y \varphi(y) \mathbf{r}(x, y) Q(dy) \Big) P(dx) = \langle \psi, \mathbf{P}_X \varphi \rangle_X,$$

ami bizonyítandó volt.

Most három olyan statisztikai alkalmazást mutatunk, amelyben a feltételes várvárhatóértékképzés operátorának szinguláris felbontására van szükség.

a) *A maximálkorreláció feladata*

Keresendők olyan $\psi \in L^2(\mathbf{X}, \mathscr{A}, P)$ és $\varphi \in L^2(\mathbf{Y}, \mathscr{B}, Q)$ egységszórású valószínűségi változók, hogy

$$C(\psi, \varphi) \to \max \tag{4.5}$$

legyen. (A feladatot gyakran úgy fogalmazzák meg, hogy keresendők egy ξ és η valószínűségi változónak olyan f és g *Borel-mérhető* függvényei, hogy $f(\xi)$ és $g(\eta)$ korrelációja maximális legyen. Tekintsük ui. az $L^2(\mathbf{X}, \mathscr{A}, P)$ teret egy ξ, az $L^2(\mathbf{Y}, \mathscr{B}, Q)$ teret pedig egy η valószínűségi változó **0** várható értékű, véges szórású, *Borel-mérhető* függvényei által generált tereknek, l. [6]). Esetünkben a korreláció a kovariancia, mivel egységszórású valószínűségi változókat keresünk. A kovarianciafüggvényt a feltételes várhatóértékképzés operátorával felírva keresendő

$$\max_{\|\psi\| = \|\varphi\| = 1} C(\psi, \varphi) = \max_{\|\psi\| = \|\varphi\| = 1} \langle \psi, \mathbf{P}_X \varphi \rangle_X = \max_{\|\psi\| = \|\varphi\| = 1} \langle \mathbf{P}_Y \psi, \varphi \rangle_Y.$$

Az (1.8) összefüggés alapján a feladat megoldását a ψ_1, φ_1 pár adja, a maximum értéke pedig s_1, ahol

$$\mathbf{P}_X = \sum_{i=1}^{\infty} s_i \langle \cdot, \varphi_i \rangle \psi_i$$

szinguláris felbontás. Láttuk, hogy $\mathbf{P}_X$ vetítés lévén $s_1 \leqq 1$, így valóban korrelációt kaptunk.

b) *Regressziós feladat*

Jól ismert, hogy rögzített φ esetén $\psi = \mathbf{P}_X \varphi$ adja a

$$\|\varphi - \psi\|^2 = E_R |\varphi - \psi|^2$$

kifejezés minimumát. Most ennél általánosabban, keressük azt a $\psi \in L^2(\mathbf{X}, \mathcal{A}, P)$ és $\varphi \in L^2(\mathbf{Y}, \mathcal{B}, Q)$ valószínűségi változópárt, melyre a fenti kifejezés minimális a $\|\varphi\|^2 = E_Q |\varphi|^2 = 1$ mellékfeltétellel.
Alakítsuk át a kifejezést:

$$\|\varphi - \psi\|^2 = \|\varphi\|^2 - 2C(\varphi, \psi) + \|\psi\|^2 = 1 - 2C(\varphi, \psi) + \|\psi\|^2 >$$
$$> 1 - 2C(\varphi, \psi)\|\psi\| + \|\psi\|^2,$$

ahol $\hat{\psi} = \psi / \|\psi\|$. Mivel φ és $\hat{\psi}$ egységszórásúak, az előző feladat megoldása alapján $C(\varphi, \hat{\psi}) \leqq s_1$ és egyenlőség pontosan akkor áll fenn, ha $\varphi = \varphi_1$ és $\hat{\psi} = \psi_1$, ahol $\mathbf{P}_X = \sum_{i=1}^{\infty} s_i \langle \cdot, \varphi_i \rangle \psi_i$ szinguláris felbontás. Így

$$\|\varphi - \psi\|^2 \geqq 1 - 2s_1 \|\psi\|^2.$$

A jobb oldalt $\|\psi\|$ szerint minimalizálva azt kapjuk, hogy a minimumot adó ψ-re $\|\psi\| = s_1$ és a minimum értéke $1 - s_1^2$. Tehát a minimumot adó valószínűségi változópár

$$\varphi = \varphi_1 \quad \text{és} \quad \psi = s_1 \psi_1.$$

Ezt a problémát kicsit más megfogalmazásban BREIMAN és FRIEDMAN [5] veti fel, csak ők adott valószínűségi változók függvényeiként keresik φ-t és ψ-t, tehát esetükben az L^2-terek ismét csak ezen valószínűségi változók véges szórású, *Borel-mérhető függvényei* által generált terek lesznek.

c) *Korrespondanciaanalízis*

Legyen most $\mathbf{I} = \{1, 2, \ldots, n\}$ és $\mathbf{J} = \{1, 2, \ldots, m\}$ véges halmazok és tekintsünk az $\mathbf{I} * \mathbf{J}$ szorzattéren valamely

$$R = \{r_{ij} : i = 1, \ldots, n; j = 1, ;\ldots, m\}$$

együttes eloszlást. Legyenek

$$P = \{r_{i\cdot} = \sum_{j=1}^{m} r_{ij} : i = 1, \ldots, n\}$$

ill.

$$Q = \{r_{\cdot j} = \sum_{i=1}^{n} r_{ij} \colon j = 1, \ldots, m\}$$

a marginális eloszlások. Tekintsük az $L^2(\mathbf{I}, \mathscr{A}, P)$, ill. az $L^2(\mathbf{J}, \mathscr{B}, Q)$ *Hilbert-tereket* és ezek közt a

$$\mathbf{P}_I \colon L^2(\mathbf{J}, \mathscr{B}, Q) \to L^2(\mathbf{I}, \mathscr{A}, P)$$

feltételes várhatóértékképzés operátorát. Ez tehát tetszőleges $\varphi \in L^2(\mathbf{J}, \mathscr{B}, Q)$ valószínűségi változóhoz azt a $\psi \in L^2(\mathbf{I}, \mathscr{A}, P)$ valószínűségi változót rendeli hozzá, melyre

$$\psi(i) = \frac{1}{r_{i\cdot}} \sum_{j=1}^{m} r_{ij} \varphi(j) = \sum_{j=1}^{m} \frac{r_{ij}}{r_{i\cdot} r_{\cdot j}} \varphi(j) r_{\cdot j}, \tag{4.6}$$

ahol $\psi(i)$ jelöli a ψ diszkrét valószínűségi változó $i \in \mathbf{I}$ helyen felvett értékét. $\mathbf{P}_I$-t integráloperátornak tekintve, a magfüggvény tehát $\mathbf{g}(i,j) = \frac{r_{ij}}{r_{i\cdot} r_{\cdot j}}$. A feladat $\mathbf{P}_I$-nek olyan k-rangú $\mathbf{A}_k$ (vagy ami ezzel ekvivalens: $\mathbf{g}$-nek olyan k tagú diádból álló $\mathbf{g}_k$) közelítését keresni, melyre

(4.7)

$$|||\mathbf{P}_I - \mathbf{A}_k|||^2 = E_{P*Q}|\mathbf{g} - \mathbf{g}_k|^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\frac{r_{ij}}{r_{i\cdot} r_{\cdot j}} - \sum_{l=1}^{k} z_l \alpha_l(i) \beta_l(j) \right]^2 r_{i\cdot} r_{\cdot j} \to \min,$$

ahol $k (1 < k \leqq r \leqq \min\{n, m\})$ adott pozitív egész, r a $\mathbf{P}_I$ operátor rangja, z_l-ek valós számok, $\alpha_l \in L^2(\mathbf{I}, \mathscr{A}, P)$ $\beta_l \in L^2(\mathbf{J}, \mathscr{B}, Q)$ $(l = 1, \ldots, k)$.

A feladat megoldását a 3.3. tétel alapján $\mathbf{P}_I$ szinguláris felbontásának k-adik szelete (vagy $\mathbf{g}_k$ előállításában egy k-tagú diádösszeg) adja. Legyen $\mathbf{P}_I = \sum_{l=1}^{r} s_l \langle \cdot, \varphi_l \rangle \psi_l$ szinguláris felbontás, ahol az s_l, ψ_l, φ_l $(l = 1, \ldots, r)$ mennyiségekre:

$$1 = s_1 \geqq s_2 \geqq \ldots \geqq s_r, \tag{4.8}$$

$$E_P \psi_l \psi_{l'} = \sum_{i=1}^{n} \psi_l(i) \psi_{l'}(i) r_{i\cdot} = \delta_{ll'}, \tag{4.9}$$

$$E_Q \varphi_l \varphi_{l'} = \sum_{j=1}^{m} \varphi_l(j) \varphi_{l'}(j) r_{\cdot j} = \delta_{ll'}, \tag{4.10}$$

$$E_R \psi_l \varphi_{l'} = \sum_{i=1}^{n} \sum_{j=1}^{m} \psi_l(i) \varphi_{l'}(j) r_{ij} = s_l \delta_{ll'}, \tag{4.11}$$

ahol $\delta_{ll'}$ a *Kronecker-delta*.

Mivel (l. [1]) $s_1 = 1$, $\psi_1 \equiv 1$ és $\varphi_1 \equiv 1$, ezért a (4.9) és (4.10) összefüggésekből az is következik, hogy

$$E_P \psi_l = E_Q \varphi_l = 0 \quad (l = 2, \ldots, k).$$

A ψ_1, φ_1 valószínűségi változókat triviális faktoroknak, a ψ_l, φ_l $(l=2, \ldots, k)$ változókat pedig korrespondancia-faktoroknak nevezzük. Az 1.2. állítás (1.8) összefüggései és a (4.9)—(4.11) összefüggések miatt ezek olyan, mindkét térben páronként ortogonális faktorok, melyekre az azonos indexűek maximálisan korreláltak ($s_l \leqq 1$ korrelációval), a különböző indexűek pedig korrelálatlanok.

Az r_{ij} számokat egy kontingenciatábla (ahol az elemek összegével már leosztottunk) elemeinek felfogva, a magfüggvény *Karhunen—Loève sorfejtéséből* következik a cellaelemek k-tagú diádösszeggel való közelítése (L^2-normában):

$$r_{ij} \sim r_{i.} r_{.j} \left(1 + \sum_{l=2}^{k} s_l \psi_l(i) \varphi_l(j)\right), \quad (i = 1, \ldots, n;\ j = 1, \ldots, m). \tag{4.12}$$

(Ha a $g'(i,j) = \frac{r_{ij}}{r_{i.} r_{.j}} - 1$ magfüggvényből indulnánk ki, akkor csak a valódi, L^2-beli, nem-triviális faktorokat kapnánk.)

5. A szinguláris felbontás numerikus módszerei

L. BREIMANtól és J. H. FRIEDMANtól (l. [5]) származik az az ötlet, hogy a véges mátrixokra vonatkozó hatványiteráció módszerét természetes módon általánosítani lehet egy kompakt lineáris operátor legnagyobb szinguláris értékének és a hozzá tartozó sajátaltér valamely elemének meghatározására. Ők ezt ACE (*a*lternating *c*onditional *e*xpectation) algoritmusnak nevezik és közvetlenül a 4.b. regressziós feladat megoldására alkalmazzák, vagyis a feltételes várhatóértékképzés operátorát bontják fel vele. Az algoritmus konvergenciájának bizonyításában azonban nem kell kihasználni, hogy projekcióról van szó, a tétel tetszőleges kompakt lineáris operátorra általánosítható.

5.1. TÉTEL. Legyen $\mathbf{A}: \mathbf{H} \to \mathbf{H}'$ kompakt lineáris operátor. Jelölje s_1 $\mathbf{A}$ legnagyobb szinguláris értékét, E a hozzá tartozó sajátalteret, továbbá $\mathbf{A}^*$ az $\mathbf{A}$ operátor adjungáltját. Tetszőleges $\psi^{(0)} \in \mathbf{H}$ $(\|\psi^0\| = 1)$ kezdeti elemből kiindulva (melyre $\|\mathbf{P}_E \psi^{(0)}\| \neq 0$, azaz $\psi^{(0)}$ nem merőleges az E altérre) konstruáljuk a következő sorozatot:

$$\varphi^{(n+1)} = \mathbf{A}\psi^{(n)}, \quad \psi^{(n+1)} = \frac{\mathbf{A}^* \varphi^{(n+1)}}{\|\mathbf{A}^* \varphi^{(n+1)}\|} \quad (n = 0, 1, \ldots). \tag{5.1}$$

Jelölje $\psi^* = \frac{\mathbf{P}_E \psi^{(0)}}{\|\mathbf{P}_E \psi^{(0)}\|}$ $\psi^{(0)}$-nak az E altérre való 1-re normált vetületét és legyen $\varphi^* = \mathbf{A}\psi^*$. Akkor

$$\lim_{n \to \infty} \|\psi^{(n)} - \psi^*\| = 0, \quad \lim_{n \to \infty} \|\varphi^{(n)} - \varphi^*\| = 0$$

$\mathbf{H}$-, ill. $\mathbf{H}'$-beli normában és

$$\lim_{n \to \infty} \|\mathbf{A}^* \varphi^{(n)}\| = s_1.$$

Bizonyítás. Az $\mathbf{U}=\mathbf{A}^*\mathbf{A}:\mathbf{H}\to\mathbf{H}$ és $\mathbf{V}=\mathbf{A}\mathbf{A}^*$ $\mathbf{H}'\to\mathbf{H}'$ operátorok kompakt, pozitív szemidefinit, önadjungált operátorok és spektrálfelbontásuk

$$\mathbf{U}=\sum_{i=1}^{\infty} s_i^2\langle\cdot,\psi_i\rangle\psi_i \quad \text{ill.} \quad \mathbf{V}=\sum_{i=1}^{\infty} s_i^2\langle\cdot,\varphi_i\rangle\varphi_i,$$

továbbá

$$\mathbf{A}=\sum_{i=1}^{\infty} s_i\langle\cdot,\psi_i\rangle\varphi_i, \quad \text{ill.} \quad \mathbf{A}^*=\sum_{i=1}^{\infty} s_i\langle\cdot,\varphi_i\rangle\psi_i$$

az $\mathbf{A}$, ill. $\mathbf{A}^*$ operátorok szinguláris felbontása. (5.1) miatt ezzel

$$\psi^{(n+1)}=\frac{\mathbf{A}^*\mathbf{A}\psi^{(n)}}{\|\mathbf{A}^*\mathbf{A}\psi^{(n)}\|}=\frac{\mathbf{U}\psi^{(n)}}{\|\mathbf{U}\psi^{(n)}\|}. \tag{5.2}$$

$\psi^{(n)}$ egyértelműen felírható, mint egy E-beli és egy erre merőleges elem összege:

$$\psi^{(n)}=a^{(n)}\psi^*+\mathbf{g}^{(n)}, \tag{5.3}$$

ahol $\psi^*\in E$, $a^{(n)}\in\mathbf{R}$ és $\mathbf{g}^{(n)}\perp E$.

Azt kell belátni, hogy $a^{(n)}\to 1$ és $\|\mathbf{g}^{(n)}\|\to 0$, ha $n\to\infty$. Ehhez az (5.2) és (5.3) összefüggésekből kiindulva írjunk fel $a^{(n)}$-ekre és $\mathbf{g}^{(n)}$-ekre rekurziót:

$$\psi^{(n+1)}=\frac{\mathbf{U}\psi^{(n)}}{\|\mathbf{U}\psi^{(n)}\|}=\frac{a^{(n)}s_1^2\psi^*+\mathbf{U}\mathbf{g}^{(n)}}{\|a^{(n)}s_1^2\psi^*+\mathbf{U}\mathbf{g}^{(n)}\|},$$

ahonnan

$$a^{(n+1)}=\frac{s_1^2a^{(n)}}{\|\mathbf{U}\psi^{(n)}\|}, \quad \mathbf{g}^{(n+1)}=\frac{\mathbf{U}\mathbf{g}^{(n)}}{\|\mathbf{U}\psi^{(n)}\|} \quad (n=0,1,\ldots). \tag{5.4}$$

Szükségünk van arra a segédállításra, hogy tetszőleges $\mathbf{g}\perp E$ elemre $\|\mathbf{U}\mathbf{g}\|\leqq\leqq r\|\mathbf{g}\|$, ahol $s_i^2\leqq r<s_1^2$ $(i: s_i<s_1)$. Ugyanis egy E-re merőleges $\mathbf{g}$ felírható az s_1-nél kisebb szinguláris értékekhez tartozó sajátalterek elemeinek lineáris kombinációiként: $\mathbf{g}=\sum_{\substack{i\\ s_i<s_1}} b_i\psi_i$.

A háromszög-egyenlőtlenséget és a *Parseval-formulát* használva

$$\|\mathbf{U}\mathbf{g}\|^2\leqq\sum_{\substack{i\\ s_i<s_1}} b_i^2\|\mathbf{U}\psi_i\|^2=\sum_{\substack{i\\ s_i<s_1}} b_i^2s_i^4\|\psi_i\|^2\leqq r^2\sum_{\substack{i\\ s_i<s_1}} b_i^2\|\psi_i\|^2=r^2\|\mathbf{g}\|^2,$$

ami a segédállítást bizonyítja.

Ebből az $\|\mathbf{U}\mathbf{g}^{(n)}\|\leqq r\|\mathbf{g}^{(n)}\|$ összefüggést az előbbi r-rel felhasználva:

$$\frac{\|\mathbf{g}^{(n+1)}\|}{a^{(n+1)}}=\frac{\|\mathbf{U}\mathbf{g}^{(n)}\|}{s_1^2a^{(n)}}\leqq\frac{r}{s_1^2}\frac{\|\mathbf{g}^{(n)}\|}{a^{(n)}}.$$

Ebből

$$\frac{\|\mathbf{g}^{(n)}\|}{a^{(n)}}<c\left(\frac{r}{s_1^2}\right)^n\to 0, \quad n\to\infty, \tag{5.5}$$

mert $r<s_1^2$ pozitív szám volt (c pedig kiszámolható konstans).

Mivel azonban $\|\psi^{(n)}\|=1$, az (5.3) összefüggés miatt $a^{(n)2}+\|\mathbf{g}^{(n)}\|^2=1$. Másrészt (5.5) miatt $\lim_{n\to\infty}\frac{\|\mathbf{g}^{(n)}\|}{a^{(n)}}=0$. Ez csak úgy lehetséges, ha $\|\mathbf{g}^{(n)}\|\to 0$ és $a^{(n)2}\to 1$ $(n\to\infty)$. Utóbbi alapján $a^{(n)}\to 1$ $(n\to\infty)$ is igaz, mivel $a^{(0)}>0$ miatt az (5.4) rekurzióból látszik, hogy az összes $a^{(n)}$ pozitív.

A tétel állítása ezek után már könnyen adódik:

$$\|\psi^{(n)}-\psi^*\|^2=\|a^{(n)}\psi^*+\mathbf{g}^{(n)}-\psi^*\|^2=(1-a^{(n)})^2\|\psi^*\|^2+\|\mathbf{g}^{(n)}\|^2\to 0 \quad (n\to\infty).$$

Másrészt

$$\|\varphi^{(n+1)}-\varphi^*\|=\|\mathbf{A}\psi^{(n)}-\mathbf{A}\psi^*\|\leqq\|\mathbf{A}\|\cdot\|\psi^{(n)}-\psi^*\|<K\cdot\|\psi^{(n)}-\psi^*\|\to 0 \quad (n\to\infty),$$

ahol $\|\mathbf{A}\|<K$ (ui. $\mathbf{A}$ operátor a kompaktság miatt korlátos is).

Már csak a legnagyobb szinguláris érték meghatározására vonatkozó állítást kell belátni: az (5.1) és (5.3) összefüggésekkel

$$\mathbf{A}^*\varphi^{(n+1)}=\mathbf{U}\psi^{(n)}=a^{(n)}s_1^2\psi^*+\mathbf{U}\mathbf{g}^{(n)},$$

amiből az (5.4) összefüggéssel

$$\|\mathbf{A}^*\varphi^{(n+1)}\|^2=a^{(n)2}s_1^2+\|\mathbf{U}\psi^{(n)}\|^2\cdot\|\mathbf{g}^{(n+1)}\|^2\to s_1^2 \quad n\to\infty,$$

mert $a^{(n)}\to 1$, $\|\mathbf{g}^{(n)}\|\to 0$, $\mathbf{U}$ korlátossága miatt pedig van olyan $K_1>0$ konstans, hogy $\|\mathbf{U}\psi^{(n)}\|\leqq\|\mathbf{U}\|\cdot\|\psi^{(n)}\|<K_1$. Innen $\|\mathbf{A}^*\varphi^{(n+1)}\|\to s_1$ $(n\to\infty)$, azaz, ha az (5.1) formulában a még normálatlan $\psi^{(n+1)}$-et tekintjük, épp a norma — amivel leosztottunk — adja ki határértékként a legnagyobb szinguláris értéket.

Megjegyezzük, hogy amennyiben az $\mathbf{A}$ operátort most az E-re merőleges $\mathbf{H}$-beli altérre szorítjuk meg, hasonló eljárással megkapható a második legnagyobb szinguláris érték és a hozzá tartozó sajátaltér egy eleme tetszőleges, erre nem merőleges, 1-re normált elemből kiindulva ... s.í.t. Ezzel az eljárással „fokozatos lehámozással" egymás után, nagyság szerint csökkenő sorrendben elvileg $\mathbf{A}$ véges sok szinguláris értékét meghatározhatjuk. Mivel gyakorlati feladatokban általában csak az első néhány szinguláris értékre van szükség, ez az iteráció alkalmazható, de már mátrixoknál is láttuk (l. [4]), hogy elég hosszadalmas.

Köszönetnyilvánítás

Végül köszönetet mondok TUSNÁDY GÁBORnak, hogy ezekre az alkalmazásokra, és CSÁKI PÉTERnek, hogy a maximálkorreláció feladatára felhívták figyelmemet.

IRODALOM

[1] BENZÉCRI, J. P. ET AL., L'Analyse dés Données. Tome 2. L'Analyse dés Correspondances. (Dunod, Paris, 1973).

[2] BOLLA, M., "A QRPS-transzformáció: a QR-algoritmus általánosítása valós téglalapmátrixok szinguláris felbontására", *Alk. Mat. Lapok* **8** (1982) 125—139.

[3] BOLLA, M., TUSNÁDY, G., "The QRPS algorithm: A generalization of the QR algorithm for the singular values decomposition of rectangular mtrices", *Periodica Math. Hung.*, **16** (1985) 201—207.

[4] BOLLA, M., „Mátrixok spektrálfelbontásának és szinguláris felbontásának módszerei", *MTA SZTAKI Tanulmányok*, **74** /1985.
[5] BREIMAN, L., FRIEDMAN, H. J., "Estimating Optimal Transformations for Multiple Regression and Correlation", *Journal of the American Statistical Association* **80** (1985) 580—619.
[6] CSÁKI, P., FISCHER, J., "On the general notion of maximal correlation", az *MTA Mat. Kut. Int. Közleményei*, VIII. évfolyam, A. sorozat, 1—2. füzet (1963).
[7] FRITZ, J., „Az alakfelismerés statisztikus módszerei", Az *MTA Mat. Kut. Int. ,A számítástechnika matematikai alapjai"* című tanfolyamának jegyzete (Révész Pál és Fritz József előadásai alapján), Budapest,1974.
[8] GULYÁS, O., „Folytonos paraméterű folyamatok diszkretizálásának módszerei", az *Idősorok Analízise* című könyv (szerk. Tusnády, G. és Ziermann, M.) fejezete, Műszaki Könyvkiadó, Budapest, 1986., pp. 119—140.
[9] HALMOS, P. R., SUNDER, V. S., *Bounded Integral Operators on L^2 Spaces* (Springer-Verlag, Berlin—Heidelberg—New York, 1978).
[10] KATO, T., *Perturbation Theory for Linear Operators* (Springer-Verlag, 1966).
[11] KOLMOGOROV, A. N., FOMIN, Sz. V., *A függvényelmélet és funkcionálanalízis elemei* (Műszaki Könyvkiadó, Budapest, 1981).
[12] RÉNYI, A., "On measures of dependence", *Acta Math. Acad. Sci. Hung.* **10** (1959) 441—451.
[13] RIESZ, F., Sz. NAGY, B., *Lecons d'Analyse Fonctionnelle* (Akadémiai Kiadó, Budapest—Szeged, 1953).
[14] TUSNÁDY, G., „Mátrixok szinguláris felbontása", *Alk. Mat. Lapok* **5** (1979) 375—384.
[15] VARGA, L., "On the error-minimizing property of the Karhunen—Loeve expansion", *Proc. of the Ninth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, 1982, 257—261.
[16] ZAANEN, A. C., *Linear Analysis* (North Holland, Amsterdam, 1956).

(Beérkezett: 1987. március 13.)

BOLLA MARIANNA
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI
KUTATÓ INTÉZET
1111 BUDAPEST, KENDE U. 13—17.

SINGULAR VALUES DECOMPOSITION OF LINEAR OPERATORS ON HILBERT SPACES (THE APPLICATIONS OF THE OPTIMAL FEATURES OF THE DECOMPOSITION AND NUMERICAL METHODS FOR THE DECOMPOSITION)

M. BOLLA

The theorems for the spectral and singular values decomposition of the compact linear operators on *Hilbert spaces* are well known. For the sake of better understanding these theorems are stated as generally, as possible and we prove the optimal features of the decomposition on the basis of the mini-max principle for the separation of the different singular values.

These theorems are applied for special integral operators of L^2 spaces, which is resulted in the *Karhunen—Loève expansion* of the kernel. On this basis a lot of statistical problems can be solved very easily (e.g. correspondence analysis, maximal correlation, regression).

Eventually we prove the convergence of the power iteration for the decomposition (suggested by BREIMAN and FRIEDMAN in a special problem).

Alkalmazott Matematikai Lapok **13** (1987—88) 207—230

KORRESPONDANCIAANALÍZIS

BOLLA MARIANNA

Budapest

A korrespondanciaanalízis az utóbbi évtizedben elterjedt többváltozós statisztikai módszer kontingenciatáblák vizsgálatára. Bár maga az eljárás már régóta ismert (l. FISCHER [4], GUTTMAN [13], LANCASTER [19], KENDALL és STUART [17], MCKEON [22], akik ún. szkórokat rendelnek a sorokhoz és oszlopokhoz, hogy a kontingenciatáblát a kanonikus korrelációanalízisből ismert módon alacsonyabb rangú táblával közelítsék), a korrespondanciaanalízis jelölésrendszerét és apparátusát a francia iskola dolgozta ki (BENZECRI et al. [1]) 1973-ban. Ezek után HILL 1974-ben írott [16] cikke hívja fel a módszer hasznosságára a figyelmet. Sajnos, a franciák leírásmódja nehezen illeszthető a többváltozós statisztikai módszerek szokásos tárgyalásába, az összeegyeztetésre JAN de LEUW [21], GOODMAN [7], [8], HABERMAN [14], GORDON [11] és munkatársaik tesznek kísérletet.

Jelen cikkben megpróbáljuk a korrespondanciaanalízis feladatát a lehető legáltalánosabb módon (a feltételes várható érték operátorának szinguláris felbontásával) leírni. Ebbe a trágyalásmódba a fentiek is beilleszthetők és a megoldás egyszerű lineáris algebrai segédeszközökkel adódik. Megmutatjuk, hogy a korrespondanciaanalízis nem más, mint kvalitatív változók kanonikus analízise, és a kapott faktorok legkisebb négyzetes becslést adnak.

Ezenkívül foglalkozunk speciális többdimenziós kontingenciatáblák ún. többszörös korrespondanciaanalízisével, és ezt összehasonlítjuk a JAN de LEUW-féle homogenitásvizsgálattal [21], továbbá a TUSNÁDY GÁBOR és általam kidolgozott többdimenziós beágyazással.

1. A korrespondanciaanalízis feladata

Adott egy $n*m$-es kontigenciatábla az f_{ij} cellagyakoriságokkal. Legyen N ezek összege, N-nel cellánként leosztva az r_{ij} relatív gyakoriságokhoz jutunk. Ezeket tekinthetjük két diszkrét eloszlású valószínűségi változó (amelyek közül az egyik n, a másik m lehetséges értéket vesz fel) együttes eloszlásának (jelölje ezt R), és szintén **R**-rel jelöljük a relatív gyakoriságok $n*m$-es tábláját. Jelölje $r_{i.}$ $(i=1, \ldots, n)$, ill. $r_{.j}$ $(j=1, \ldots, m)$ a megfelelő marginálisokat. Ezek az $\mathbf{I}=\{1, \ldots, n\}$ ill. $\mathbf{J}=\{1, \ldots, m\}$ alaphalmazokon a P, ill. Q valószínűségeloszlást alkotják. Az általánosság megszorítása nélkül feltehető, hogy a marginális gyakoriságok határozottan pozitívak (ellenkező esetben ui. az azonosan nulla sorok és/vagy oszlopok elhagyhatók, azaz **I** és/vagy **J** redukálható).

Célunk a kontingenciatáblának valamely alacsonyabb rangú táblával való közelítése. Ehhez a kanonikus korrelációanalízishez hasonlóan keresünk mindkét fenti diszkrét valószínűségi változóhoz olyan, egymással páronként korrelálatlan faktorokat, hogy a megegyező indexű faktorok korrelációja maximális legyen. Ilyen módon a kontingenciatábla előáll a faktor-valószínűségi változók értékei (ún. *szkórok*) diádszorzatainak súlyozott összegeként. A legnagyobb súlyok közül bizonyos számút megtartva a kontingenciatábla egy alacsonyabb rangú közelítését kapjuk.

A feladat pontos leírására és megoldására az integráloperátorok elméletét használjuk fel (l. [2]). Mivel itt nagyon speciális *Hilbert-terek:* az $L^2(\mathbf{I}, \mathscr{A}, P)$, ill. $L^2(\mathbf{J}, \mathscr{B}, Q)$ terek (melyeket az **I**-n, ill, **J**-n értelmezett P, ill. Q eloszlású valószínűségi változók alkotnak) közti lineáris operátorokról van szó, ezek természetesen véges rangúak, így a rájuk vonatkozó spektrál-, ill. szinguláris felbontási tételek, továbbá a felbontások optimumtulajdonságára vonatkozó tételek alkalmazhatók.

Vizsgálódásaink középpontjában az

$$\mathbf{A}: L^2(\mathbf{J}, \mathscr{B}, Q) \to L^2(\mathbf{I}, \mathscr{A}, P)$$

feltételes várható érték operátora áll, ennek szinguláris felbontását végezzük el. Nézzük meg, milyen algebrai segédeszközök kellenek ehhez.

1.1. ÁLLÍTÁS. Az **A** feltételes várható érték operátora integráloperátor, melynek magfüggvénye a $\mathbf{g}: \mathbf{I} * \mathbf{J} \to \mathbf{R}$ függvény, ahol

$$\mathbf{g}(i, j) = r_{ij}/r_{i\cdot} r_{\cdot j} \quad (i = 1, \ldots, n;\ j = 1, \ldots, m).$$

Bizonyítás. Legyen $\varphi \in L^2(\mathbf{J}, \mathscr{B}, Q)$ és

$$\psi(i) = \sum_{j=1}^{m} \frac{r_{ij}}{r_{i\cdot} r_{\cdot j}} \varphi(j) r_{\cdot j} = \frac{1}{r_{i\cdot}} \sum_{j=1}^{m} r_{ij} \varphi(j) \quad (i = 1, \ldots, n). \tag{1.1}$$

Ekkor $\psi = E_R(\varphi|\mathscr{A}) = \mathbf{A}\varphi$, amivel készen is vagyunk, hiszen egy integráloperátor és magfüggvénye kölcsönösen egyértelműen meghatározzák egymást (l. pl. [2]).

Nézzük meg, mi lesz a feltételes várhatóértékképzés operátorának mátrixalakja R^n-, ill. R^m-beli euklideszi normában ortonormált (o.n.) bázisokban felírva. Az

$$\mathbf{x}(i) = \sqrt{r_{i\cdot}}\, \psi(i) \quad \text{és} \quad \mathbf{y}(j) = \sqrt{r_{\cdot j}}\, \varphi(j) \tag{1.2}$$

transzformációkat bevezetve könnyen látható, hogy ha $E_P \psi^2 = 1$, akkor $\sum_{i=1}^{n} \mathbf{x}^2(i) = 1$ és ha $E_P \psi \psi' = 0$, akkor $\sum_{i=1}^{n} \mathbf{x}(i)\mathbf{x}'(i) = 0$, azaz egy P mérték szerint o.n. rendszerből a fenti transzformációval euklideszi normában o.n. rendszert kapunk. Hasonló igaz φ és **y** viszonyára.

Az (1.2) összefüggésekkel, amennyiben $\psi = \mathbf{A}\varphi$, az ezeknek megfeleltetett **x** és **y** vektorok között a

$$\sum_{j=1}^{m} \frac{r_{ij}}{\sqrt{r_{i\cdot}}\sqrt{r_{\cdot j}}} \mathbf{y}(j) = \mathbf{x}(i) \quad (i = 1, \ldots, n)$$

összefüggés közvetít, tehát az **A** operátor mátrixalakja (az előbbi, euklideszi normában o.n. bázisokban felírva):

$$\mathbf{B} = \mathbf{P}^{-\frac{1}{2}} \mathbf{R} \mathbf{Q}^{-\frac{1}{2}}, \tag{1.3}$$

ahol **P** és **Q** jelöli a megfelelő marginálisokat fődiagonálisában tartalmazó $n * m$-es, ill. $m * m$-es diagonális mátrixokat.

Ezek után a korrespondanciaanalízis feladata az **A** feltételes várható érték operátor, azaz a **B** mátrix szinguláris felbontása. A felbontás optimumtulajdonságai az

integráloperátorok szinguláris felbontásának optimumtulajdonságai alapján (l. [2]) a következők lesznek:

Adott $k(1 \leqq k \leqq \min\{n, m\})$ egészhez keressük a kontingenciatábla k tagú diádösszeggel való közelítését, melyre

$$a) \quad |||\mathbf{A}-\mathbf{A}_k|||^2 = E_{P*Q}\|\mathbf{g}-\mathbf{g}_k\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\frac{r_{ij}}{r_{i\cdot}r_{\cdot j}} - \sum_{l=1}^{k} z_l\alpha_l(i)\beta_l(j)\right]^2 r_{i\cdot}r_{\cdot j} =$$

$$= \operatorname{tr}(\mathbf{B}-\mathbf{B}_k)^T(\mathbf{B}-\mathbf{B}_k)$$

minimális, ahol $\mathbf{A}_k: L^2(\mathbf{J}, \mathscr{B}, Q) \to L^2(\mathbf{I}, \mathscr{A}, P)$ legfeljebb k-adrangú operátor, $g_k = \sum_{l=1}^{k} z_l\alpha_l\beta_l$ k-tagú diádösszeg, $\mathbf{B}_k$ legfeljebb k-rangú mátrix, $|||\cdot|||$ pedig egy lineáris operátor ún. kettős normája, amely nem más, mint szinguláris értékeinek négyzetösszege (ez integráloperátorok esetén mindig konvergens, l. [2]).

Az integráloperátorok elméletéből jól ismert tény (l. [2]), hogy a feladat megoldását az $\mathbf{A}$ operátor szinguláris felbontásának k legnagyobb szinguláris értékéhez tartozó sajátelemei adják, azaz

$$z_l = s_l, \quad \alpha_l = \psi_l, \quad \beta_l = \varphi_l \quad (l = 1, \ldots, k),$$

ahol

$$\mathbf{A} = \sum_{l=1}^{r} s_l\langle\cdot, \varphi_l\rangle\psi_l \tag{1.4}$$

az $\mathbf{A}$ operátor szinguláris felbontása, és a keresett minimum értéke $\sum_{l=k+1}^{r} s_l^2$. Itt $0 < r \leqq \min\{n, m\}$ az $\mathbf{A}$ operátor rangja.

$$1 = s_1 \geqq s_2 \geqq \ldots \geqq s_r > 0, \tag{1.5}$$

$$E_P\psi_l\psi_{l'} = \sum_{i=1}^{n}\psi_l(i)\psi_{l'}(i)r_{i\cdot} = \delta_{ll'} \quad (1 \leqq l, l' \leqq r) \tag{1.6}$$

$$E_Q\varphi_l\varphi_{l'} = \sum_{j=1}^{m}\varphi_l(j)\varphi_{l'}(j)r_{\cdot j} = \delta_{ll'} \quad (1 \leqq l, l' \leqq r), \tag{1.7}$$

ahol $\delta_{ll'}$ a *Kronecker-delta*.

Az (1.5) összefüggésre még visszatérünk, az (1.6), (1.7) összefüggések pedig a szinguláris felbontás tulajdonságaiból következnek. Azt is megmutatjuk, hogy az 1 szinguláris értékhez mindig van egy azonosan 1 bázispár, jelölje ezeket ψ_1, φ_1. Az ezekre való ortogonalitás egyben azt is jelenti, hogy

$$E_P\psi_l = \sum_{i=1}^{n}\psi_l(i)r_{i\cdot} = 0, \quad E_Q\varphi_l = \sum_{j=1}^{m}\varphi_l(j)r_{\cdot j} = 0 \quad (l = 2, \ldots, r). \tag{1.8}$$

A gyakorlatban természetesen a $\mathbf{B}$ mátrix szinguláris felbontását hajtjuk végre: $\mathbf{B} = \sum_{l=1}^{r} s_l\mathbf{u}_l\mathbf{v}_l^T$, ahol $\{\mathbf{u}_l\}_{l=1}^{r} \subset R^n$, ill. $\{\mathbf{v}_l\}_{l=1}^{r} \subset R^m$ euklideszi normában o.n. rendszer, ^{T} pedig a transzponálást jelöli. A megfelelő $L^2(\mathbf{I}, \mathscr{A}, P)$, ill. $L^2(\mathbf{J}, \mathscr{B}, Q)$ terekben o.n. rendszerek a következő összefüggéssel kaphatók:

$$\psi_l = \mathbf{P}^{-\frac{1}{2}}\mathbf{u}_l, \quad \varphi_l = \mathbf{Q}^{-\frac{1}{2}}\mathbf{v}_l \quad (l = 1, \ldots, r). \tag{1.9}$$

Szintén jól ismert a szinguláris felbontás következő optimumtulajdonsága:

$$b) \max_{\substack{E_P\alpha_l\alpha_{l'}=\delta_{ll'} \\ E_Q\beta_l\beta_{l'}=\delta_{ll'}}} E_R\alpha_l\beta_l = \max_{\substack{\sum_{i=1}^{n}\psi_l(i)\psi_{l'}(i)r_{i\cdot}=\delta_{ll'} \quad (1\leq l'\leq l) \\ \sum_{j=1}^{m}\varphi_l(j)\varphi_{l'}(j)r_{\cdot j}=\delta_{ll'} \quad (1\leq l'\leq l)}} \sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_l(i)\beta_l(j)r_{ij} = s_l,$$

$$(l = 1, \ldots, r)$$

és a maximum a ψ_l, φ_l páron vétetik fel. Azaz $\psi_1 \equiv 1$, $\varphi_1 \equiv 1$ és $l=2, \ldots, r$-re az l-edik faktorpár (ψ_l, φ_l) korrelációja maximális az n, ill. m diszkrét értéket felvevő, egységszórású, az előző faktorokra mindkét térben ortogonális valószínűségi változók körében, és ez a korreláció éppen s_l. Ez is mutatja, hogy a szinguláris értékek 1-nél nem nagyobbak. Ez a tulajdonság éppen a kanonikus korrelációanalízis feladatának megfogalmazása diszkrét valószínűségi változókra.

1.1. Definíció. A fenti **A** feltételes várható értékképzés operátorának szinguláris felbontásában szereplő ψ_l, $\varphi_l (l=2, \ldots, r)$ elemeket *korrespondanciafaktorok*nak, a ψ_1, φ_1 párt pedig *triviális faktorok*nak nevezzük.

A kiindulásul vett diszkrét változóknak egyértelműen megfeleltethető egy ξ n-, ill. η m-komponensű bináris vektorváltozó, azaz a kontingenciatáblának egyértelműen megfeleltethető egy n, ill. m kérdést tartalmazó kérdőív, ahol mindkét esetben ugyanazt az N egyént kérdezik meg, és mindenki mindkét kérdéscsoportból pontosan egy kérdésre válaszolhat igennel, a többire nemmel válaszol (igen =: 1, nem =: 0).

Jelölje **X**, ill. **Y** ezeket az $N*n$-es, ill. $N*m$-es megfigyelésmátrixokat, amelyeknek tehát minden sora pontosan egy 1-est tartalmaz. Könnyen látható, hogy

$$\text{(1.10)} \qquad \mathbf{R} = \frac{1}{N}\mathbf{X}^T\mathbf{Y}, \quad \mathbf{P} = \frac{1}{N}\mathbf{X}^T\mathbf{X}, \quad \mathbf{Q} = \frac{1}{N}\mathbf{Y}^T\mathbf{Y}.$$

Így **R**, **P** és **Q** tekinthető a két bináris vektorváltozó empirikus keresztkovariancia-, ill. kovarianciamátrixának és így formálisan a $\mathbf{B} = \mathbf{P}^{-\frac{1}{2}}\mathbf{R}\mathbf{Q}^{-\frac{1}{2}}$ mátrix éppen az, amelynek szinguláris felbontását normális vektorváltozók kanonikus korrelációanalízisénél elvégezzük.

Mivel ξ, ill. η kovarianciamátrixa diagonális, komponenseik korrelálatlanok, ami bináris változóknál függetlenséget is jelent. Ezzel a korrespondanciaanalízis célját tekintve is megfelel a normális változókon végrehajtott kanonikus korrelációanalízisnek, hiszen a keresett ψ, ill. φ faktorok (maguk 1-dimenziós diszkrét valószínűségi változók) ξ, ill. η komponenseinek lineáris kombinációi, hiszen $\psi = \sum_{i=1}^{n}\psi(i)\xi_i$, ill. $\varphi = \sum_{j=1}^{m}\varphi(j)\eta_j$ (ahol ξ_i és η_j ξ és η komponenseit jelöli) ui. $P(\psi = \psi(i)) = P(\xi_i = 1) = r_{i\cdot}$, ill. $P(\varphi = \varphi(j)) = P(\eta_j) = 1 = r_{\cdot j}$ a komponensek függetlensége miatt. Mivel

$$E\xi = (r_{1\cdot}, \ldots, r_{n\cdot})^T \quad \text{és} \quad E\eta = (r_{\cdot 1}, \ldots, r_{\cdot m})^T,$$

ezért a kanonikus analízist valójában a $\xi' = \xi - E\xi$, ill. $\eta' = \eta - E\eta$ bináris vektorváltozókon hajtjuk végre, melyek empirikus keresztkovarianciája R', ahol

$$r'_{ij} = r_{ij} - r_{i\cdot}r_{\cdot j} \quad (i = 1, \ldots, n;\ j = 1, \ldots, m).$$

A $g'(i,j)=\frac{r'_{ij}}{r_{i.}r_{.j}}$ magfüggvény sorfejtésével a ψ_l, $\varphi_l(l=2,\ldots,r)$ korrespondanciafaktorokat kapjuk, melyek tehát maximálisan korreláltak (s_l korrelációval) ξ' ill. η' komponenseinek lineáris kombinációi körében, akárcsak a normális változók kanonikus faktorai.

Összefoglalva tehát, az a. és b. pontok alapján láthatjuk, hogy a szinguláris felbontás báziselemei a korrespondanciafaktorokra legkisebb négyzetes becslést adnak, és egyben a két tér változói közti korrelációt is maximalizálják a kanonikus korrelációanalízis értelmében.

2. A korrespondanciafaktorok tulajdonságai

2.1. Definíció. Az első fejezetben leírt szinguláris felbontással kapott korrespondenciafaktorok koordinátáit *kanonikus szkóroknak* nevezzük. Ezek segítségével adott $k(1\leqq k\leqq r$, ahol r a kontingenciatábla rangja) egészre a gyakoriságtábla k-tagú diádösszeggel való közelítését $(k-1)$-*edrendű közelítésnek* nevezzük.

Egy $(k-1)$-edrendű közelítés a kontingenciatáblának egy k-rangú táblával való közelítése.

A kanonikus szkórokat felhasználva a gyakoriságtábla $(k-1)$-edrendű közelítésének elemeire:

$$r_{ij}^{(k)}=r_{i.}r_{.j}\Big[1+\sum_{l=2}^{k}s_l\psi_l(i)\varphi_l(j)\Big], \tag{2.1}$$

ahol $k=1$ esetén a fenti szumma nulla.

Az (1.8) összefüggések alapján könnyen látható, hogy

$$\sum_{i=1}^{n}\sum_{j=1}^{m}r_{ij}^{(k)}=\underbrace{\sum_{i=1}^{n}r_{i.}\sum_{j=1}^{m}r_{.j}}_{1}+\sum_{l=2}^{k}s_l\sum_{i=1}^{n}\psi_l(i)r_{i.}\underbrace{\sum_{j=1}^{m}\varphi_l(j)r_{.j}}_{0}=1.$$

A 0-adrendű közelítés ($k=1$ eset) feltételezése éppen a függetlenség hipotézisét jelenti, vagyis hogy a kontingenciatábla elemei a marginálisok diádszorzataként állnak elő:

$$H_0\colon\ \varrho_{ij}=\varrho_{i.}\varrho_{.j},$$

ahol ϱ az r relatív gyakoriságnak megfelelő valódi valószínűséget jelöli.

Az ezt tesztelő χ^2-statisztika, amely a H_0 hipotézis fennállása esetén $(n-1)*(m-1)$ szabadságfokú χ^2-eloszlást követ:

$$\chi^2=N\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(r_{ij}-r_{i.}r_{.j})^2}{r_{i.}r_{.j}}=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(f_{ij}-\frac{f_{i.}f_{.j}}{N}\right)^2}{\frac{f_{i.}f_{.j}}{N}}= \tag{2.2}$$

$$=N\Big[\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{f_{ij}^2}{f_{i.}f_{.j}}-1\Big]=N\Big[\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{r_{ij}^2}{r_{i.}r_{.j}}-1\Big],$$

ahol $f_{ij}=Nr_{ij}$ az (i,j) cella megfigyelt gyakorisága.

2.2. Definíció. A (2.2) összefüggésben a szögletes zárójelben álló kifejezést *négyzetes kontingenciának* nevezzük (l. RÉNYI [27]).

(Abban az esetben, mikor **I** és **J** megszámlálhatóan végtelen, a négyzetes kontingencia végessége elégséges feltételt ad a feltételes várhatóérték operátorának kompaktságára, l. [27]. Véges **I**, **J** halmazok esetén természetesen ez is véges.) Könnyen látható, hogy

$$\chi^2 = N[\operatorname{tr} \mathbf{B}^T\mathbf{B} - 1] = N \operatorname{tr} (\mathbf{B}-\mathbf{B}_1)^T(\mathbf{B}-\mathbf{B}_1) = \mathbf{N} \sum_{l=1}^{r} s_l^2, \tag{2.3}$$

azaz a függetlenség hipotézisét vizsgáló χ^2 éppen a nem-triviális korrespondanciafaktorokhoz tartozó szinguláris értékek négyzetösszege.

2.3. Definíció. A fenti χ^2-értéket a *korrespondancia mértékének* (a külföldi irodalomban *trace* =nyom), az $\dfrac{N \sum_{l=2}^{k} s_l^2}{\chi^2}$ számot pedig *az első* $k-1$ *nem-triviális korrespondancia-faktor által magyarázott korrespondanciának* (a francia irodalomban *taux* =hányad) nevezzük.

A gyakorlatban, ha a függetlenség hipotézisét már elvetettük, $k=2, 3, \ldots, r$-re ennek a magyarázott korrespondanciának a %-os nagysága határozza meg, hogy hány korrespondanciafaktor „magyarázza elég jól" a kontingenciatáblát. Itt azonban a normális eloszlású változókra kidolgozott kanonikus korrelációanalíziséhez hasonló *Bartlett-féle likelihood-hányados próba* nem hajtható végre. Ennek az az oka, hogy a szinguláris felbontással kapott kanonikus szkórok nem maximum lekelihood becslések a

$$H_{k-1}\colon \varrho_{ij} = \varrho_{i\cdot}\varrho_{\cdot j}\Big[1 + \sum_{l=2}^{k} s_l \psi_l(i)\varphi_l(j)\Big]$$

modellben (GOODMAN [7]—[10] cikkeiben ezt $(k-1)$-edrendű *asszociációs modell*nek nevezi), amely azon a hipotézisen alapul, hogy létezik $(k-1)$-edrendű közelítés $(k=2, \ldots, r)$.

A kanonikus szkórok márcsak azért sem lehetnek a fenti modellben maximum likelihood becslések, mert az $r_{ij}^{(k)}$ elemek között negatív is előfordulhat. A legegyszerűbb ellenpéldák 3×3-as kontingenciatáblák körében adhatók a másodrangú közelítésre (az 1- és r-rangú közelítések ui. mindig pozitívak).

Legyen kontingenciatáblánk a 3×3-as identitás. Mivel $s_1=s_2=s_3=1$, $\psi_i=\varphi_i$ $(i=1, \ldots, 3)$, a ψ_2, ψ_3 pár tetszőlegesen választható egy, a $\psi_1=1$ faktorra merőleges altérben. Ha

$$\psi_2(1) = -\sqrt{\frac{3}{2}}, \quad \psi_2(2) = \sqrt{\frac{3}{2}}, \quad \psi_2(3) = 0,$$

$$\psi_3(1) = -\frac{1}{\sqrt{2}}, \quad \psi_3(2) = -\frac{1}{\sqrt{2}}, \quad \psi_3(3) = \sqrt{2},$$

akkor $r_{12}^{(2)}=r_{21}^{(2)}=-\dfrac{1}{18}<0.$

Ha ψ_2 és ψ_3 szerepét felcseréljük, akkor a másodrangú közelítés csupa pozitív elemekből áll.

De a negativitás előfordulhat egyszeres multiplicitású szinguláris értékek (egyértelmű korrespondanciafaktorok) esetén is. Legyen pl. kontingenciatáblánk

$$\begin{pmatrix} 126 & 0 & 1 \\ 1 & 622 & 0 \\ 1 & 1 & 248 \end{pmatrix}.$$

Ekkor $s_1=1$, $s_2=\sqrt{0{,}989\,75}$, $s_3=\sqrt{0{,}969\,76}$, és a másodrangú közelítés:

$$\begin{pmatrix} 33{,}839 & 15{,}38 & 77{,}781 \\ 15{,}918 & 619{,}51 & -12{,}429 \\ 78{,}243 & -11{,}89 & 183{,}647 \end{pmatrix}.$$

$k=2$ esetén HABERMAN és GILULA [6] felírják a likelihood-egyenleteket és numerikus eljárást adnak az egyenletrendszer megoldására. KENDALL [17] ugyan a (2.3) felbontást χ^2-partíciónak nevezi, de mint LANCASTER [19] rámutat, az $N\sum_{l=k+1}^{r} s_l^2$ mennyiségek már nem χ^2-eloszlásúak akkor sem, ha a H_{k-1} hipotézis fennáll. O'NEILL [24] cikkében bebizonyítja, hogy H_0 fennállása esetén Ns_1^2 határeloszlása $(N\to\infty)$ olyan, mint egy centrális *Wishart-mátrix* legnagyobb sajátértékéé. HABERMAN [14] pedig H_1 fennállása esetén ad meg határeloszlásokat.

Az első részben leírt szinguláris felbontás alapján a **B** mátrix sajátelemei között fennállnak a

$$\mathbf{B}\mathbf{v}_l = s_l\mathbf{u}_l, \quad \mathbf{B}^T\mathbf{u}_l = s_l\mathbf{v}_l \quad (l=1,\ldots,r)$$

összefüggések (l. [29]). Megjegyezzük, hogy **B** rangja megegyezik **A** rangjával, mivel a marginálisok pozitívak. Így az (1.2) összefüggések miatt a korrespondanciafaktorokra

$$\mathbf{F}\varphi_l = s_l\mathbf{P}\psi_l, \quad \mathbf{F}^T\psi_l = s_l\mathbf{Q}\varphi_l \quad (l=1,\ldots,r), \tag{2.4}$$

ahonnan a kanonikus szkórokra a következő, ún. *átváltási formulák* érvényesek:

$$\sum_{j=1}^{m} r_{ij}\varphi_l(j) = s_l r_{i\cdot}\psi_l(i), \quad (l=1,\ldots,r) \tag{2.5}$$

$$\sum_{i=1}^{n} r_{ij}\psi_l(i) = s_l r_{\cdot j}\varphi_l(j), \quad (l=1,\ldots,r). \tag{2.6}$$

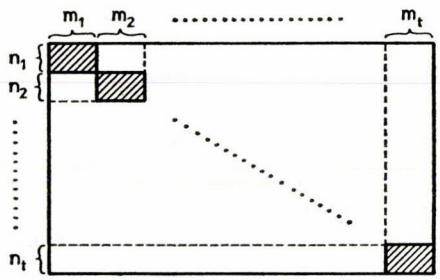
2.1. ÁLLÍTÁS. Tetszőleges kontingenciatáblából kiindulva

$$s_1 = 1, \quad \psi_1 \equiv 1, \quad \varphi_1 \equiv 1.$$

Bizonyítás. Ezekkel a mennyiségekkel a (2.5), (2.6) átváltási formulák teljesülnek, melyek egyértelműen leírják az összetartozó báziselemeket (l. [29], mivel 1 a maximális szinguláris érték és s_l-ek korrelációk.

Az 1 azonban többszörös szinguláris érték is lehet (ilyenkor ψ_1, φ_1 nem egyértelmű, de választható a fentienek), erről szól a 2.1. tétel. Ennek megértésére szolgál a

2.4. Definíció. Azt mondjuk, hogy egy $n*m$-es kontingenciatábla maximálisan $t(1 \leqq t \leqq \min\{n, m\})$ számú diagonális blokkból épül fel, ha a sorok és oszlopok permutációjával elérhető, hogy van benne t rész-kontingenciatábla, melyek diszjunktak és sorba rendezhetők olyan módon, hogy az első bal felső sarka a nagy tábla bal felső sarka, egy rákövetkező bal felső sarkában álló elem mindkét indexe eggyel nagyobb, mint az őt megelőző jobb alsó sarkában levő elemé (a blokkok sarkosan érintkeznek), és az utolsó jobb alsó sarka a nagy tábla jobb alsó sarka. Ezeken a rész-kontingenciatáblákon kívül mindenütt nulla elemek állnak. A maximalitás azt jelenti, hogy a rész-kontingenciatáblák egyike sem bontható tovább a fenti módon diagonális blokkokra. (Az itt látható ábra sraffozott részei t maximális diagonális blokkot ábrázolnak, az üresen hagyott részeken nullák állnak, ahol $\sum_{i=1}^{t} n_i = n$ és $\sum_{i=1}^{t} m_i = m$).



1. ábra

A következő tételből az is kijön, hogy a diagonális blokkok maximális száma egyértelmű.

2.1. TÉTEL. A fenti feltételes várható érték operátorának az 1 pontosan annyiszoros szinguláris értéke, ahány maximális diagonális blokkból a kontingenciatábla felépül.

Bizonyítás. Tegyük fel, hogy kontingenciatáblánk t maximális diagonális blokkból épül fel. Láttuk, hogy $s_1 = 1$ és ψ_1, φ_1 választható azonosan **1**-nek. Most keresünk olyan ψ_l, φ_l $(l = 2, \ldots, t)$ faktorokat, hogy

$$E_P \psi_l \psi_{l'} = \delta_{ll'} \quad (2 \leqq l \leqq t,\ 1 \leqq l' \leqq t)$$

és

$$E_Q \varphi_l \varphi_{l'} = \delta_{ll'} \quad (2 \leqq l \leqq t,\ 1 \leqq l' \leqq t).$$

Legyen ψ_2, ill. φ_2 a következő:

$$\psi_2(i) = \begin{cases} c_1^{(2)}, & i \leqq n_1 \\ c_2^{(2)}, & i > n_1 \end{cases}, \qquad \varphi_2(j) = \begin{cases} d_1^{(2)}, & j \leqq m_1 \\ d_2^{(2)}, & j > m_1 \end{cases},$$

azaz két lehetséges értéket vehetnek fel. Ekkor a $c_1^{(2)}$, $c_2^{(2)}$, $d_1^{(2)}$, $d_2^{(2)}$ konstansok megválaszthatók úgy, hogy

$$E_P \psi_1 \psi_2 = \sum_{i=1}^{n} \psi_1(i) r_{i\cdot} = c_1^{(2)} \sum_{i=1}^{n_1} r_{i\cdot} + c_2^{(2)} \sum_{i=n_1+1}^{n} r_{i\cdot} = 0$$

és

$$E_Q\varphi_1\varphi_2 = \sum_{j=1}^{m}\varphi_2(j)r_{\cdot j} = d_1^{(2)}\sum_{j=1}^{m_1}r_{\cdot j}+d_2^{(2)}\sum_{j=m_1+1}^{m}r_{\cdot j} = 0$$

legyen. Mivel

$$\sum_{i=1}^{n_1}r_{i\cdot} = \sum_{j=1}^{m_1}r_{\cdot j} = F_1^{(2)} \quad \text{és} \quad \sum_{i=n_1+1}^{n}r_{i\cdot} = \sum_{j=m_1+1}^{m}r_{\cdot j} = F_2^{(2)}$$

(ahol $F_2^{(2)} = 1 - F_1^{(2)}$), ezért $c_1^{(2)}$ és $c_2^{(2)}$, ill. $d_1^{(2)}$ és $d_2^{(2)}$ ellenkező előjelűek:

$$c_2^{(2)} = \frac{-c_1^{(2)}F_1^{(2)}}{F_2^{(2)}}, \quad d_2^{(2)} = \frac{-d_1^{(2)}F_1^{(2)}}{F_2^{(2)}}. \tag{2.7}$$

Ezenkívül teljesülniük kell még a következő egyenlőségeknek is:

$$E_P\psi_2^2 = (c_1^{(2)})^2F_1^{(2)}+(c_2^{(2)})^2F_2^{(2)} = 1,$$

$$E_Q\psi_2^2 = (d_1^{(2)})^2F_1^{(2)}+(d_2^{(2)})^2F_2^{(2)} = 1.$$

Ezekbe a (2.7) összefüggéseket behelyettesítve azt kapjuk, hogy

$$c_1^{(2)} = d_1^{(2)} = \sqrt{\frac{F_2^{(2)}}{F_1^{(2)}}}, \quad c_2^2 = d_2^2 = -\sqrt{\frac{F_1^{(2)}}{F_2^{(2)}}}.$$

Ezekkel

$$E_R\psi_2\varphi_2 = c_1^{(2)}d_1^{(2)}F_1^{(2)}+c_2^{(2)}d_2^{(2)}F_2^{(2)} = (c_1^{(2)})^2F_1^{(2)}+(c_2^{(2)})^2F_2^{(2)} =$$

$$= \frac{F_2^{(2)}}{F_1^{(2)}}F_1^{(2)}+\frac{F_1^{(2)}}{F_2^{(2)}}F_2^{(2)} = F_2^{(2)}+F_1^{(2)} = 1,$$

azaz 1 legalább 2-szeres szinguláris érték.

$l=3, 4, \ldots, t$-re hasonlóan konstruálható meg a ψ_l, φ_l pár úgy, hogy ezek pontosan l különböző értéket vesznek fel az $\{1, \ldots, n_1\}$, $\{n_1+1, \ldots, n_1+n_2\}$, ..., ..., $\{\sum_{i=1}^{l-1}n_i+1, \ldots, n\}$, ill. $\{1, \ldots, m_1\}$, $\{m_1+1, \ldots, m_1+m_2\}$, ..., $\{\sum_{i=1}^{l-1}m_i+1, \ldots, m\}$ lépcsőkön, szórásuk 1 legyen, és a náluk kisebb indexű faktorokra ortogonálisak legyenek a P, ill. Q mértékek szerint. Mivel

$$E_R\psi_l\varphi_l = \sum_{i=1}^{n}\sum_{j=1}^{m}\psi_l(i)\varphi_l(j)r_{ij} = \sum_{u=1}^{l}\sum_{\substack{i=\sum_{v=1}^{u-1}n_v+1 \\ j=\sum_{v=1}^{u-1}m_v+1}}^{\sum_{v=1}^{u}n_v \quad \sum_{v=1}^{u}m_v}\psi_l(i)\varphi_l(j)r_{ij} =$$

$$= \sum_{u=1}^{l}c_u^{(l)}d_u^{(l)}\sum_{\substack{i=\sum_{v=1}^{u-1}n_v+1 \\ j=\sum_{v=1}^{u-1}m_v+1}}^{\sum_{v=1}^{u}n_v \quad \sum_{v=1}^{u}m_v}r_{ij} = \sum_{u=1}^{l}c_u^{(l)}d_u^{(l)}F_u^{(l)} = \sum_{u=1}^{l}(c_u^{(l)})^2F_u^{(l)} = 1$$

$$(l = 2, \ldots, t),$$

ahol F_u^l az r_{ij} gyakoriságok összege az u-adik sraffozott részen $(u=1, \ldots, l-1)$, $F_l^{(l)}=1-\sum_{u=1}^{l-1} F_u^{(l)}$, $c_u^{(l)}$, $d_u^{(l)}$ pedig ψ_l, ill. φ_l értéke az u-adik lépcsőn $(u=1, \ldots, l)$, és ha a szumma jel felett kisebb szám áll, mint alatta, az összeget 0-nak tekintjük. Az $l=2$ esethez hasonlóan tetszőleges l-re $(l=1, \ldots, t)$ $d_u^{(l)}=c_u^{(l)}$ $(u=1, \ldots, l)$ és $\sum_{u=1}^{l}(c_u^{(l)})^2 F_u^{(l)}=1$, tehát az 1 legalább t-szeres szinguláris érték.

Megfordítva tegyük fel, hogy az 1 pontosan t-szeres szinguláris érték. Jelölje a hozzá tartozó sajátbázis elemeket ψ_l, φ_l $(l=1, \ldots, t)$. Mivel $E_R \psi_l \varphi_l=1$, ezért a normáltsági feltételek miatt $\psi_l \equiv \varphi_l$. Ez csak úgy lehetséges, ha

$$P\{\psi_l = c_u^{(l)}\} = P\{\varphi_l = d_u^{(l)}\}, \quad u = 1, \ldots, l,$$

azaz

$$\sum_{\psi_l(i)=c_u^{(l)}} r_{i\cdot} = \sum_{\varphi_l(j)=d_u^{(l)}} r_{\cdot j} \quad (u = 1, \ldots, l).$$

Ez pontosan azt jelenti, hogy a kontingenciatábla legalább t diagonális blokkból áll. Összevetve az előző irányú állítással, látható, hogy a maximális diagonális blokkok száma megegyezik az 1-es szinguláris érték multiplicitásával.

2.2. ÁLLÍTÁS. A korrespondanciafaktorokra

$$E_R \psi_l \varphi_{l'} = s_l \delta_{ll'} \quad (l, l' = 1, \ldots, r). \tag{2.8}$$

Bizonyítás. Az $l=l'$ esetet már láttuk. Ha $l \neq l'$, akkor pedig a (2.5), (2.6) átváltási formulákkal

$$E_R \psi_l \varphi_{l'} = \sum_{i=1}^{n} \sum_{j=1}^{m} \psi_l(i) \varphi_{l'}(j) r_{ij} = \sum_{i=1}^{n} \psi_l(i) \sum_{j=1}^{m} \varphi_l(j) r_{ij} =$$

$$= \sum_{i=1}^{n} \psi_l(i) s_{l'} r_{i\cdot} \psi_{l'}(i) = s_{l'} E_P \psi_l \psi_{l'} = 0.$$

2.3. ÁLLÍTÁS.

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{[r_{ij} - r_{ij}^{(k)}]^2}{r_{i\cdot} r_{\cdot j}} = \sum_{l=k+1}^{r} s_l^2 \quad (k = 2, \ldots, r).$$

Bizonyítás. A bal oldalon álló négyzetösszeg éppen

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \left[\frac{r_{ij}}{r_{i\cdot} r_{\cdot j}} - \sum_{l=1}^{k} s_l \psi_l(i) \varphi_l(j) \right]^2 r_{i\cdot} r_{\cdot j} = \|\mathbf{g} - \mathbf{g}_k\|^2 = \operatorname{tr} (\mathbf{B} - \mathbf{B}_k)^T (\mathbf{B} - \mathbf{B}_k) = \sum_{l=k+1}^{k} s_l^2.$$

Most röviden ismertetjük a korrespondanciaanalízis [1]-beli bevezetését és ennek kapcsolatát a mi levezetésünkkel.

Legyen $F=\{f_{ij}: i=1, \ldots, n;\ j=1, \ldots, m\}$ gyakoriságtábla és $\mathbf{f}_j^i$ jelölje ennek i-edik sorát leosztva a sorösszeggel, azaz

$$\mathbf{f}_j^i = \left\{ \frac{f_{ij}}{f_{i\cdot}} : j \in \mathbf{J} \right\},$$

ennek elemei tehát **J**-n eloszlást alkotnak. Jelölje $\mathcal{N}(\mathbf{I})$ az ezekből alkotott pontfelhőt (a francia szakirodalomban *nuage*) a **J**-dimenziós térben:

$$\mathcal{N}(\mathbf{I}) = \{\mathbf{f}_J^i:\ i \in \mathbf{I}\}.$$

Jelölje $\mathbf{f}_I = \{f_{i.}: i \in I\}$, ill. $\mathbf{f}_J = \{f_{.j}: j \in \mathbf{J}\}$ a marginálisokból álló vektorokat, ekkor

$$E_P \mathcal{N}(\mathbf{I}) = \mathbf{f}_J,$$

azaz $\mathbf{f}_J$ az $\mathcal{N}(\mathbf{I})$ pontfelhő súlypontja. $\mathbf{f}_J$-t mint első (triviális) faktort tekintjük, és e köré akarunk további $k-1$ faktort keresni úgy, hogy az $\mathcal{N}(\mathbf{I})$ pontfelhő összvarianciájából az első k faktor a lehető legtöbbet magyarázza, ahol k rögzített egész ($k=1, \ldots, r$). Az $\mathcal{N}(\mathbf{I})$ pontfelhő összvarianciája:

$$\sum_{i \in I} \|\mathbf{f}_J^i - \mathbf{f}_J\|^2_{\mathbf{f}_J} f_{i.}, \tag{2.9}$$

ahol $\|\,.\,-\,.\,\|_{\mathbf{f}_J}$ jelöli az argumentumban álló két eloszlás $\mathbf{f}_J$-re centrált χ^2-távolságát, amit [1]-ben a következőképpen definiálnak.

2.5. Definíció. A $\mathbf{p}^{(1)}$ és $\mathbf{p}^{(2)}$ eloszlások közti, $\mathbf{f}_J$-re centrált χ^2-távolság:

$$\sum_{j \in \mathbf{I}} \frac{[p_j^{(1)} - p_j^{(2)}]^2}{\mathbf{f}_{.j}}.$$

(2.9)-et tovább kifejtve az $\mathcal{N}(\mathbf{I})$ pontfelhő összvarianciája:

$$\sum_{i \in \mathbf{I}} \sum_{j \in \mathbf{J}} \frac{\left[\frac{f_{ij}}{f_{i.}} - f_{.j}\right]^2}{f_{.j}} f_{i.} = \sum_{i \in \mathbf{I}} \sum_{j \in \mathbf{J}} \frac{[f_{ij} - f_{i.} f_{.j}]^2}{f_{i.}^2 f_{.j}} f_{i.} =$$

$$= \sum_{i \in \mathbf{I}} \sum_{j \in \mathbf{J}} \frac{[f_{ij} - f_{i.} f_{.j}]^2}{f_{.i} f_{.j}} = \frac{\chi^2}{N} \quad \text{(l. (2.2)-t)},$$

ami nem más, mint az f_{ij}, ill. $f_{ij}^{(1)}$ eloszlások közti, $f_{i.} f_{.j}$-re centrált χ^2-távolság.

A kanonikus szkóroknál kapott s_l, ψ_l, φ_l ($l=2, \ldots, k$) mennyiségekkel bevezetve az

$$\mathbf{a}_l(i) = s_l \psi(i), \quad i \in \mathbf{I}$$

$$\mathbf{u}_l(j) = f_{.j} \varphi_l(j), \quad j \in \mathbf{J} \tag{2.10}$$

és

jelöléseket és $\mathbf{u}_l$-et tekintve az $\mathcal{N}(\mathbf{I})$ pontfelhő l-edik faktortengelyének ($l=2, \ldots, k$), ez a rendszer adott k-ra a pontfelhő összvarianciájából éppen a legtöbbet magyarázza. Ui. a nem magyarázott rész összvarianciája:

$$\sum_{i \in \mathbf{I}} \|\mathbf{f}_J^i - \mathbf{f}_J - \sum_{l=2}^{k} \mathbf{a}_l(i)\mathbf{u}_l\|^2_{\mathbf{f}_J} f_{i.} = \sum_{i \in \mathbf{I}} \sum_{j \in \mathbf{J}} \frac{[f_{ij} - f_{i.} f_{.j} - f_{i.} \sum_{l=2}^{k} \mathbf{a}_l(i)\mathbf{u}_l(j)]^2}{f_{i.} f_{.j}} =$$

$$= \sum_{i \in \mathbf{I}} \sum_{j \in \mathbf{J}} \frac{[f_{ij} - f_{i.} f_{.j}(1 + \sum_{l=2}^{k} s_l \psi_l(i) \varphi_l(j))]^2}{f_{i.} f_{.j}} = \sum_{i \in \mathbf{I}} \sum_{j \in \mathbf{J}} \frac{[f_{ij} - f_{ij}^{(k)}]^2}{f_{i.} f_{.j}},$$

ami a 2.3. állítás alapján az $\mathcal{N}(\mathbf{I})$ pontfelhő összvarianciájának az első k faktor által nem magyarázott részét minimalizálja, azaz a magyarázott részt maximalizálja.

Ugyanezt az eredményt kapjuk, ha az I-n értelmezett eloszlásokból álló $\mathcal{N}(\mathbf{J})$ pontfelhőből indulunk ki. Ekkor a nem-triviális faktorok a $\mathbf{v}_l=\{\mathbf{v}_l(i)=f_{i.}\psi_l(i): i\in\mathbf{I}\}$ vektorok lesznek $(l=2, \ldots, k)$ és ezeknek a $\mathbf{b}_l(j)=s_l\varphi_l(j)$, $(j\in\mathbf{J})$ együtthatókkal képezzük a fenti értelemben optimális lineáris kombinációját.

Az elmondottak egyben módszert is adnak adott k esetén az $\mathcal{N}(\mathbf{I})$, $\mathcal{N}(\mathbf{J})$ pontfelhő elemeinek az első k faktor terében való ábrázolására. Az $(\mathbf{f}_I, \mathbf{f}_J)$ párt képzelve az origóba, a pontok a $k-1$ dimenziós térben ábrázolhatók az

$$(\mathbf{a}_2(i), \ldots, \mathbf{a}_k(i)): \quad i\in\mathbf{I},$$

ill.

$$(\mathbf{b}_2(j), \ldots, \mathbf{b}_k(j)): \quad j\in\mathbf{J}$$

koordináták alapján.

Az $\mathcal{N}(\mathbf{I})$ pontfelhő i_1-edik, ill. i_2-edik pontja közötti χ^2-távolság éppen a megfelelő pontok euklideszi távolsága lesz, ui.:

$$\sum_{j\in\mathbf{J}} \frac{\left[\frac{f_{i_1j}^{(k)}}{f_{i_1.}} - \frac{f_{i_2j}^{(k)}}{f_{i_2.}}\right]^2}{f_{.j}} =$$

$$= \sum_{j\in\mathbf{J}} \frac{\left[\frac{f_{i_1.}f_{.j}\left(1+\sum_{l=2}^{k}\mathbf{a}_l(i_1)\varphi_l(j)\right)}{f_{i_1.}} - \frac{f_{i_2.}f_{.j}\left(1+\sum_{l=2}^{k}\mathbf{a}_l(i_2)\varphi_l(j)\right)}{f_{i_2.}}\right]^2}{f_{.j}} =$$

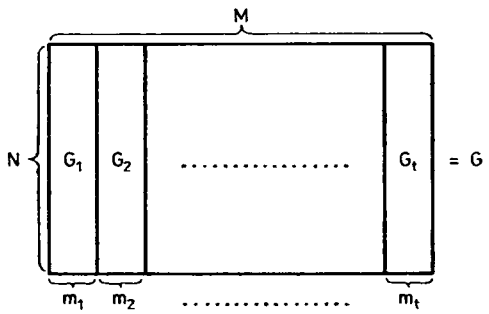
$$= \sum_{j\in\mathbf{J}} f_{.j}\left[\sum_{l=2}^{k}(\mathbf{a}_l(i_1)-\mathbf{a}_l(i_2))\varphi_l(j)\right]^2 = \sum_{l=2}^{k}(\mathbf{a}_l(i_1)-\mathbf{a}_l(i_2))^2 \underbrace{\sum_{j\in\mathbf{J}}\varphi_l(j)^2 f_{.j}}_{1},$$

felhasználva, hogy $E_Q\varphi_l\varphi_{l'}=0$. Ez éppen az i_1-edik, ill. i_2-edik pontok euklideszi távolsága a $(k-1)$-dimenziós faktortérben. Hasonló mondható az $\mathcal{N}(\mathbf{J})$ pontfelhő pontjairól is. Sőt a két pontfelhőt ugyanabban a $(k-1)$-dimenziós térben ábrázolva heurisztikusan látható (részletesebben l. [1]-ben), hogy az $\mathcal{N}(\mathbf{I})$ pontfelhő i-edik és az $\mathcal{N}(\mathbf{J})$ pontfelhő j-edik pontja annál közelebb lesz egymáshoz, minél nagyobb az f_{ij} cellagyakoriság. A kérdőíves megfogalmazásban ez azt interpretálja, hogy az egyik kérdőív i-edik és a másik kérdőív j-edik kérdésére sokan válaszoltak egyszerre igennel.

Szokás a kontingenciatábla helyett speciális 0—1 adatmátrixból kiindulni (l. kérdőíves megközelítés). Ekkor a korrespondanciaanalízis lehetővé teszi az objektumok és a dichotóm változók ugyanazon térben való ábrázolását. Ezzel foglalkozik még általánosabban (2-nél több változó esetén) az ún. többszörös korrespondanciaanalízis.

3. Többszörös korrespondanciaanalízis

Legyen **G** a következő speciális, 0—1 elemekből álló mátrix: **G** sorai N objektumra vonatkozó megfigyeléseket tartalmaznak úgy, hogy a **G** mátrix t számú blokkból áll $(\mathbf{G}_1, \ldots, \mathbf{G}_t)$ és minden blokkon belül minden sorban pontosan egy 1-es áll, a többi helyen pedig 0. Képzeljünk el pl. egy olyan szituációt, mikor N személynek t számú kérdőívet kell kitöltenie, és a j-edik kérdőívben m_j-féle lehetőség közül kell az egyiket kiválasztani. Legyen $M=\sum_{j=1}^{t} m_j$, így az $N*M$-es **G** mátrix a következő alakú:



2. ábra

ahol minden blokkon belül minden sorban pontosan egy 1-es áll.

3.1. Definíció. A fenti alakú **G** adatmátrixot *teljes diszjunkt táblának* nevezzük (a francia szakirodalomban: *tableau sous forme disjonctif complète*).

Könnyen látható, hogy a $\mathbf{B}=\mathbf{G}^T\mathbf{G}$ mátrix szimmetrikus és t^2 számú blokkból áll. A diagonális $\mathbf{G}_j^T\mathbf{G}_j=\mathbf{D}_j$ blokkok maguk is diagonálisak és a

$$\mathbf{D}_j = \begin{pmatrix} d_j^1 & & 0 \\ & \ddots & \\ 0 & & d_j^{m_j} \end{pmatrix}$$

mátrix fődiagonálisában a j-edik változó egyes *kategóriái*nak *(modalitè)* gyakoriságai állnak. A nem diagonális blokkok pedig közönséges kontingenciatáblák, pl. $\mathbf{G}_i^T\mathbf{G}_j$ $(i\neq j)$ az i-edik és j-edik változó kategóriáinak együttes gyakoriságait tartalmazza, mint azt már $t=2$ esetén az 1. fejezetben megmutattuk.

3.2. Definíció. A fenti **B** blokk-mátrixot *Burt-táblának (tableaux de Burt)* nevezzük.

Az is könnyen látható, hogy a *Burt-tábla* valamennyi blokkján belül az elemek összege N, sor- és oszlopösszegei pedig a $t\mathbf{D}$ diagonális mátrix fődiagonálisában állnak, ahol **D** a $\mathbf{D}_1, \ldots, \mathbf{D}_t$ diagonális mátrixokat tartalmazó szintén csak diagonális mátrix.

Hajtsunk most végre a **G** mátrixon az 1. fejezetben leírt módon korrespondanciaanalízist. Mivel **G** elemeinek összege tN, sorrösszegei $t-k$, oszlopösszegeit pedig

a **D** diagonális mátrix fődiagonálisa tartalmazza, a következő mátrix szinguláris felbontását kell elvégezni:

$$\left(\frac{t}{tN}\right)^{-1/2} \frac{\mathbf{G}}{tN} \left(\frac{\mathbf{D}}{tN}\right)^{-1/2} = \frac{1}{\sqrt{t}} \mathbf{G}\mathbf{D}^{-1/2} = \sum_{l=1}^{r} s_l \mathbf{u}_l \mathbf{v}_l^T,$$

ahol $r = \operatorname{rang}(\mathbf{G}) \leqq \min\{N, M-t\}$, mivel mind a t blokkon belül az oszlopok között legalább egy lineáris összefüggés van és

$$1 = s_1 \geqq s_2 \geqq \ldots \geqq s_r > 0,$$

$\mathbf{u}_l \in R^N$, $\mathbf{v}_l \in R^M$ $(l = 1, \ldots, r)$ euklideszi normában ortormált rendszereket alkotnak. A korrespondancia-faktorokhoz (jelölje ezeket most $\mathbf{X}_l$, $\mathbf{Y}_l$) az (1.9) összefüggések vezetnek:

$$\mathbf{X}_l = \left(\frac{t}{tN}\right)^{-1/2} \mathbf{u}_l = \sqrt{N}\mathbf{u}_l \quad (l = 1, \ldots, r), \tag{3.1}$$

$$\mathbf{Y}_l = \left(\frac{\mathbf{D}}{tN}\right)^{-1/2} \mathbf{v}_l = \sqrt{tN}\mathbf{v}_l \quad (l = 1, \ldots, r).$$

Itt is igaz, hogy $\mathbf{X}_1 \equiv \mathbf{1} \in R^N$, $\mathbf{Y}_1 \equiv \mathbf{1} \in R^M$, és az ortogonalitási feltételek miatt

$$\sum_{i=1}^{N} \mathbf{X}_l(i) \frac{t}{tN} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_l(i) = 0, \quad \text{azaz} \quad \sum_{i=1}^{N} \mathbf{X}_l(i) = 0, \tag{3.2}$$

$$\sum_{i=1}^{N} \mathbf{X}_l^2(i) \frac{t}{tN} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_l^2(i) = 1, \quad \text{azaz} \quad \sum_{i=1}^{N} \mathbf{X}_l^2(i) = N, \tag{3.3}$$

$$\frac{1}{tN} \sum_{j=1}^{t} \sum_{i=1}^{m_j} \mathbf{Y}_l^j(i) d_j^i = 0, \quad \text{azaz} \quad \sum_{j=1}^{t} \sum_{i=1}^{m_j} \mathbf{Y}_l^j(i) d_j^i = 0, \tag{3.4}$$

$$\frac{1}{tN} \sum_{j=1}^{t} \sum_{i=1}^{m_j} [\mathbf{Y}_l^j(i)]^2 d_j^i = 1, \quad \text{azaz} \quad \sum_{j=1}^{t} \sum_{i=1}^{m_j} [\mathbf{Y}_l^j(i)]^2 d_j^i = tN. \tag{3.5}$$

ahol

$$\mathbf{Y}_l = (\mathbf{Y}_l^1, \ldots, \mathbf{Y}_l^t)^T.$$

A (2.4) átváltási formulák alapján pedig:

$$\frac{\mathbf{G}}{tN} \mathbf{Y}_l = s_l \frac{t}{tN} \mathbf{X}_l, \qquad \frac{\mathbf{G}^T}{tN} \mathbf{X}_l = s_l \frac{\mathbf{D}}{tN} \mathbf{Y}_l, \quad (l = 1, \ldots, r)$$

azaz

$$\mathbf{X}_l = \frac{1}{s_l t} \mathbf{G}\mathbf{Y}_l, \quad \mathbf{Y}_l = \frac{1}{s_l} \mathbf{D}^{-1}\mathbf{G}^T\mathbf{X}_l. \tag{3.6}$$

3.1. TÉTEL. A **B** *Burt-táblán* végrehajtott korrespondanciaanalízis faktorai megegyeznek a **G** teljes diszjunkt táblán végrehajtott korrespondanciaanalízis $\mathbf{Y}_l$ faktoraival $(l = 1, \ldots, r)$.

Bizonyítás. Mivel $\mathbf{B}$ szimmetrikus, elemeinek összege t^2N, $\frac{\mathbf{B}}{t^2\mathbf{N}}$ sor- és oszlopösszegeit pedig a $\frac{t\mathbf{D}}{t^2N}=\frac{1}{tN}\mathbf{D}$ diagonális mátrix tartalm azza, ezért a következő mátrix szinguláris felbontását kell elvégezni:

$$\left(\frac{\mathbf{D}}{tN}\right)^{-1/2}\frac{\mathbf{B}}{t^2N}\left(\frac{\mathbf{D}}{tN}\right)^{-1/2}=\frac{1}{t}\mathbf{D}^{-1/2}\mathbf{B}\mathbf{D}^{-1/2}.$$

Ez a mátrix szintén szimmetrikus, így szinguláris felbontásának bázisvektorai a sajátvektorok, szinguláris értékei pedig a sajátértékek abszolútértékei.

Mint láttuk a $\mathbf{G}$-n végrehajtott korrespondanciaanalízisnél az $\frac{1}{\sqrt{t}}\mathbf{G}\mathbf{D}^{-\frac{1}{2}}=$ $=\sum_{l=1}^{r}s_l\mathbf{u}_l\mathbf{v}_l^T$ szinguláris felbontást kellett elvégezni. Innen az $\left(\frac{1}{\sqrt{t}}\mathbf{G}\mathbf{D}^{-\frac{1}{2}}\right)^T\times$ $\times\left(\frac{1}{\sqrt{t}}\mathbf{G}\mathbf{D}^{-\frac{1}{2}}\right)$ mátrix spektrálfelbontása:

$$\frac{1}{t}\mathbf{D}^{-1/2}\mathbf{G}^T\mathbf{G}\mathbf{D}^{-1/2}=\sum_{l=1}^{r}s_l^2\mathbf{v}_l\mathbf{v}_l^T.$$

Mivel $\mathbf{G}^T\mathbf{G}=\mathbf{B}$ definíció szerint, ezért a *Burt-táblán* végrehajtott korrespondnaciaanalízisnél a szinguláris értékek: $1=s_1^2\geqq s_2^2>\ldots>s_r^2>0$ lesznek, a korrespondanciafaktorok pedig:

$$\left(\frac{\mathbf{D}}{tN}\right)^{-1/2}\mathbf{v}_l=\sqrt{tN}\,\mathbf{D}^{-1/2}\mathbf{v}_l\quad(l=1,\ldots,r),$$

ami definíció szerint épp az $\mathbf{Y}_l$ vektor.

Természetes módon felvetődik a kérdés, hogy az $\mathbf{Y}_l=(\mathbf{Y}_l^1,\ldots,\mathbf{Y}_l^t)^T$ alakban particionált korrespondancia-faktorokra igaz-e, hogy az $\mathbf{Y}_l^i$, $\mathbf{Y}_l^j$ pár a *Burt-tábla* (i,j)-edik blokkjának (mint közönséges kontingenciatáblának) az összetartozó korrespondancia-faktor pártjait adja. A válasz a következő:

3.2. TÉTEL. *$t=2$ esetén a Burt-tábla többszörös korrespondanciaanalízisének $\mathbf{Y}_l$ faktorai az $(\mathbf{Y}_l^1,\mathbf{Y}_l^2)$ párból állnak, ahol $\mathbf{Y}_l^1$ és $\mathbf{Y}_l^2$ a $\mathbf{G}_1^T\mathbf{G}_2$ kontingenciatábla korrespondancia-faktorai. $t>2$ esetén a Burt-tábla korrespondancia-faktorai általában nem állnak így elő az egyes kontingenciatáblák faktoraiból.*

Bizonyítás. Legyenek a $\mathbf{G}_1^T\mathbf{G}_2$ kontingenciatáblán végrehajtott közönséges korrespondanciaanalízis szinguláris értékei $1=z_1\geqq z_2\geqq\ldots\geqq z_{r_1}>0$ és faktorai $\mathbf{Y}_l^1$, ill. $\mathbf{Y}_l^2$ $(l=1,\ldots,r_1)$, ahol $r_1=\operatorname{rang}(\mathbf{G}_1^T\mathbf{G}_2)\leqq\min\{m_1-1,m_2-1\}$. A (2.4) formula alapján

$$\mathbf{G}_1\mathbf{G}_2\mathbf{Y}_l^2=z_l\mathbf{D}_1\mathbf{Y}_l^1,\quad\mathbf{G}_2^T\mathbf{G}_1\mathbf{Y}_l^1=z_l\mathbf{D}_2\mathbf{Y}_l^2\quad(l=1,\ldots,r_1).$$

Ezek felhasználásával

$$\mathbf{B}\begin{pmatrix}\mathbf{Y}_l^1\\ \mathbf{Y}_l^2\end{pmatrix}=\begin{pmatrix}\mathbf{D}_1 & \mathbf{G}_1^T\mathbf{G}_2\\ \mathbf{G}_2^T\mathbf{G}_1 & \mathbf{D}_2\end{pmatrix}\begin{pmatrix}\mathbf{Y}_l^1\\ \mathbf{Y}_l^2\end{pmatrix}=\begin{pmatrix}\mathbf{D}_1\mathbf{Y}_l^1+\mathbf{G}_1^T\mathbf{G}_2\mathbf{Y}_l^2\\ \mathbf{G}_2^T\mathbf{G}_1\mathbf{Y}_l^1+\mathbf{D}_2\mathbf{Y}_l^2\end{pmatrix}=$$

$$=\begin{pmatrix}\mathbf{D}_1\mathbf{Y}_l^1+z_l\mathbf{D}_1\mathbf{Y}_l^1\\ z_l\mathbf{D}_2\mathbf{Y}_l^2+\mathbf{D}_2\mathbf{Y}_l^2\end{pmatrix}=(1+z_l)\begin{pmatrix}\mathbf{D}_1 & 0\\ 0 & \mathbf{D}_2\end{pmatrix}\begin{pmatrix}\mathbf{Y}_l^1\\ \mathbf{Y}_l^2\end{pmatrix}=\frac{1+z_l}{2}(2\mathbf{D})\begin{pmatrix}\mathbf{Y}_l^1\\ \mathbf{Y}_l^2\end{pmatrix}\quad(l=1,\ldots,r_1).$$

Mivel $t=2$ esetén $\mathbf{B}$ sor-, ill. oszlopösszegeit a $2\mathbf{D}$ diagonális mátrix tartalmazza láthatjuk, hogy a *Burt-tábla* első r_1 korrespondanciafaktorára és szinguláris értékére:

$$\mathbf{Y}_l = (\mathbf{Y}_l^1, \mathbf{Y}_l^2)^T \quad \text{és} \quad s_l^2 = \frac{1+z_l}{2} \quad (l = 1, \ldots, r_1).$$

A $\mathbf{G}_1^T\mathbf{G}_2$ kontingenciatábla $z_1=1$ szinguláris értékéhez $s_1=1$ tartozik és az $\frac{1+x}{2}$ függvény a $(0, 1]$ intervallumot az $\left(\frac{1}{2}, 1\right]$ intervallumra képezi le. Tehát a *Burt-tábla* első r_1 szinguláris értéke $\frac{1}{2}$-nél nagyobb és $r_1 < r = \text{rang}\,(\mathbf{B})$.

Az is látható, hogy a bizonyítás menete $t>2$ esetén csak akkor lenne alkalmazható, ha a $\mathbf{G}_i^T\mathbf{G}_j$ kontingenciatáblák bal oldali kanonikus faktorai rögzített i mellett tetszőleges $j \neq i$ indexre megegyeznének (ill. a jobb oldali faktorok rögzített j mellett tetszőleges $i \neq j$-re ugyanazok lennének). Ez általánosságban azonban nincs így, mert a rész-kontingenciatáblák szinguláris felbontása különböző.

Annak ellenére tehát, hogy — mint azt előbb láttuk — a *Burt-tábla* kanonikus felbontása a rész-kontingenciatáblákra nem adja ugyanazokat a kanonikus szkórokat, mint ha azokon külön-külön korrespondanciaanalízist hajtottunk volna végre, a Burt-tábla fenti felbontása lehetőséget ad egy teljes diszjunkt tábla struktúrájának leírására, pontosabban az objektumoknak és az egyes változók kategóriáinak ugyanabban az alacsony dimenziós térben való ábrázolására a következőképpen:

Legyen k $(1 < k \leqq r = \text{rang}\,(\mathbf{G}))$ tetszőleges egész. Akkor az n-edik objektum koordinátái a $(k-1)$-dimenziós faktortérben a 2. fejezet alapján a következők lesznek:

$$(s_2\mathbf{X}_2(n), s_3\mathbf{X}_3(n), \ldots, s_k\mathbf{X}_k(n)), \quad (n = 1, \ldots, N).$$

A j-edik változó i-edik kategóriájának koordinátái pedig:

$$(s_2\mathbf{Y}_2^j(i), s_3\mathbf{Y}_3^j(i), \ldots, s_k\mathbf{Y}_k^j(i)), \quad (i = 1, \ldots, m_j; \quad j = 1, \ldots, t).$$

Az előző fejezetben leírtak alapján ez lehetőséget ad az objektumok, ill. a változó-kategóriák szimultán clusteresítésére, ui.:

— azok az objektumok lesznek közel a faktortérben, melyek sok változó esetén ugyanazt a kategóriát választották;

— azok a változó-kategóriák lesznek közel, melyeket sokan választottak egyszerre ki;

— és végül minden objektum az az általa kiválasztott változó-kategóriáknak megfelelő pontok súlypontjához lesz közel (ezt a következő fejezetben indokoljuk meg precízen).

(Megjegyezzük, hogy a (3.2), (3.4) összefüggések miatt a $k-1$ dimenziós faktortér mind a 2^{k-1} részterében lesznek pontok, és hogy ez a faktortér k-dimenziósnak is képzelhető, azonban a triviális faktorok létezése miatt a pontok valójában egy $(k-1)$-dimenziós altérben helyezkednek el.

4. Homogenitásvizsgálat

A módszer a [21] cikkben kerül bevezetésre, itt most megmutatjuk, hogy nem más, mint a többszörös korrespondanciaanalízis feladatának optimumkeresésként való, megfogalmazása. Legyen **G** továbbra is az előbbi teljes diszjunkt tábla. Átveszünk néhány jelölést és definíciót.

Itt úgy tekintjük, hogy minden változónak ugyanannyi kategóriája van: $m_1 = m_2 = \ldots = m_t = m$. Ez nem megszorítás, hisz az előző fejezet jelöléseivel legyen $m = \max\{m_1, m_2, \ldots, m_t\}$ és üres kategóriákat is megengedünk. Így a $\mathbf{D}_j = \mathbf{G}_j^T\mathbf{G}_j$ mátrix lehet szinguláris is, de a (3.1), (3.6) összefüggések akkor is érvényben maradnak, ha általánosított inverzet veszünk (esetünkben a pozitív diagonális elemeknek a reciprokát vesszük, a 0-kat pedig meghagyjuk). Jelölje $\mathbf{D}_j^+$ a $\mathbf{D}_j$ mátrix *Moore—Penrose-féle általánosított inverzét!* Ezzel **G** rangja és szinguláris felbontása ugyanaz marad.

4.1. Definíció. Ha adott k $(1 \leqq k \leqq \operatorname{rang}(\mathbf{G}))$ egész mellett az objektumokat és a változó-kategóriákat is a k-dimenziós euklideszi térbe képezzük le, akkor a koordinátákat (szkórokat) tartalmazó $N*k$-s **X**, ill. $M*k$-s **Y** mátrixokat k-dimenziós *kvantifikációknak* nevezzük, ahol $M = tm$.

Itt az **Y** mátrix **G** partíciójának megfelelően szintén partícionálható az $m*k$-s $\mathbf{Y}^j$ részmátrixokra $(j = 1, \ldots, t)$, de fordított irányban, mint **G**: $\mathbf{Y} = (\mathbf{Y}^1 \ldots \mathbf{Y}^t)^T$.

4.2. Definíció. Az **X** objektum- és az **Y** kategóriakvantifikációk teljesen konzisztensek, ha

$$\mathbf{X} = \mathbf{G}_1\mathbf{Y}^1 = \mathbf{G}_2\mathbf{Y}^2 = \ldots = \mathbf{G}_t\mathbf{Y}^t. \tag{4.1}$$

Ez azt jelenti, hogy bármely változót tekintve, ha két objektum ugyanazt a változókategóriát választja, akkor a szkórjaik is megegyeznek. Természetesen a (4.1) összefüggést kielégítő szkórozást a legritkább esetben lehet találni. Ehelyett bevezetjük a következő *konzisztenciaveszteség függvényt,* amit minimalizálni akarunk **X**-ben és **Y**-ban:

$$f(\mathbf{X}, \mathbf{Y}) = f(\mathbf{X}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t) = \frac{1}{t}\sum_{j=1}^{t} \|\mathbf{X} - \mathbf{G}_j\mathbf{Y}^j\|^2 = \tag{4.2}$$

$$= \frac{1}{t}\sum_{j=1}^{t} \operatorname{tr}(\mathbf{X} - \mathbf{G}_j\mathbf{Y}^j)^T(\mathbf{X} - \mathbf{G}_j\mathbf{Y}^j).$$

Először **X**-et rögzítjük és (4.2)-t **Y**-ban minimalizáljuk. Az **Y** szerinti deriváltakat 0-val egyenlővé téve:

$$-\frac{2}{t}\mathbf{G}_j^T(\mathbf{X} - \mathbf{G}_j\mathbf{Y}^j) = 0 \quad (j = 1, \ldots, t),$$

ahonnan

$$\mathbf{G}_j^T\mathbf{X} = \mathbf{G}_j^T\mathbf{G}_j\mathbf{Y}^j = \mathbf{D}_j\mathbf{Y}^j \quad (j = 1, \ldots, t).$$

Innen **D** általánosított inverzével

$$\mathbf{Y}^j = \mathbf{D}_j^+\mathbf{G}_j^T\mathbf{X} \quad (j = 1, \ldots, t).$$

Ezt (4.2)-be beírva:

$$\min_{\mathbf{Y}^1,\ldots,\mathbf{Y}^t} f(\mathbf{X}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t) = \frac{1}{t}\sum_{j=1}^{t} \operatorname{tr}(\mathbf{X}-\mathbf{G}_j\mathbf{D}_j^+\mathbf{G}_j^T\mathbf{X})^T(\mathbf{X}-\mathbf{G}_j\mathbf{D}_j^+\mathbf{G}_j^T\mathbf{X}). \tag{4.3}$$

Vegyük észre, hogy a $\mathbf{G}_j\mathbf{D}_j^+\mathbf{G}_j^T$ mátrix szimmetrikus (ez triviális) és idempotens, mivel

$$\mathbf{G}_j\mathbf{D}_j^+\underbrace{\mathbf{G}_j^T\mathbf{G}_j}_{\mathbf{D}_j}\mathbf{D}_j^+\mathbf{G}_j^T = \mathbf{G}_j\mathbf{D}_j^+\mathbf{G}_j^T,$$

így $\mathbf{G}_j\mathbf{D}_j^+\mathbf{G}_j^T$ projekció, jelölje ezt $\mathbf{P}_j$, és legyen

$$P_* = \frac{1}{t}\sum_{j=1}^{t} P_j.$$

Ezt felhasználva (4.3) tovább kibontható:

$$\frac{1}{t}\sum_{j=1}^{t} \operatorname{tr}(\mathbf{X}-\mathbf{P}_j\mathbf{X})^T(\mathbf{X}-\mathbf{P}_j\mathbf{X}) = \operatorname{tr}\mathbf{X}^T\left[\frac{1}{t}\sum_{j=1}^{t}(\mathbf{I}-\mathbf{P}_j)\right]\mathbf{X} = \operatorname{tr}\mathbf{X}^T(\mathbf{I}-\mathbf{P}_*)\mathbf{X}. \tag{4.4}$$

Ezt kell $\mathbf{X}$-ben minimalizálni. Ehhez mellékfeltételt is szokás tenni, legyen pl. $\mathbf{X}^T\mathbf{X} = \mathbf{I}_k$. Ismert, hogy a megoldást az $\mathbf{I}-\mathbf{P}_*$ szimmetrikus mátrix spektrálfelbontása adja. A $\mathbf{P}_*$ $N*N$-es mátrix saját értékeit $\lambda_1 \geqq \lambda_2 \geqq \ldots \geqq \lambda_N$-nel jelölve a minimum értéke $\mathbf{I}-\mathbf{P}_*$ k legkisebb, azaz $\mathbf{P}_*$ k legnagyobb sajátértékének összege: $\sum_{l=1}^{k}\lambda_l$, a minimumot adó $\mathbf{X}$ mátrix oszlopaiban pedig a $\lambda_1, \ldots, \lambda_k$ sajátértékekhez tartozó, euklideszi normában normált sajátvektorok állnak. Mik lesznek ezek?

$$\mathbf{P}_* = \frac{1}{t}\sum_{j=1}^{t}\mathbf{G}_j\mathbf{D}_j^+\mathbf{G}_j^T = \frac{1}{t}\mathbf{G}\mathbf{D}^+\mathbf{G}^T. \tag{4.5}$$

Mivel a $\mathbf{G}$ teljes diszjunk tábla korrespondanciaanalízisénél a 3. fejezetben láttuk, hogy

$$\frac{1}{\sqrt{t}}\mathbf{G}(\mathbf{D}^+)^{1/2} = \sum_{l=1}^{r} s_l\mathbf{u}_l\mathbf{v}_l^T$$

szinguláris felbontás (itt most $\mathbf{D}^{-\frac{1}{2}}$ helyett az általánosított inverz gyökét írtuk), ezért

$$\left(\frac{1}{\sqrt{t}}\mathbf{G}(\mathbf{D}^+)^{1/2}\right)\left(\frac{1}{\sqrt{t}}\mathbf{G}(\mathbf{D}^+)^{1/2}\right)^T = \frac{1}{t}\mathbf{G}\mathbf{D}^+\mathbf{G}^T = \sum_{l=1}^{r} s_l^2\mathbf{u}_l\mathbf{u}_l^T$$

spektrálfelbontás lesz, ahol $r = \operatorname{rang}(G)$. Ezt összevetve (4.5)-tel:

$$1 = \lambda_1 = s_1^2 \geqq \lambda_2 = s_2^2 \geqq \ldots \geqq \lambda_r = s_r^2 > 0 \tag{4.6}$$

és

$$\mathbf{X}_l = \mathbf{u}_l \quad (l = 1, \ldots, k), \tag{4.7}$$

ahol $\mathbf{X}_l$ jelöli a minimumot adó $\mathbf{X}$ mátrix l-edik oszlopát (az $\mathbf{X}^T\mathbf{X}=\mathbf{I}_k$ mellékfeltétel miatt). A konzisztencia-veszteség függvény minimuma pedig $\sum_{l=1}^{k} s_l^2$ lesz. Innen

$$\mathbf{Y}^j = \mathbf{D}_j^+ \mathbf{G}_j^T \mathbf{X} \quad (j = 1, \ldots, t) \tag{4.8}$$

és így

$$\mathbf{Y} = (\mathbf{Y}^1 \ldots \mathbf{Y}^t)^T = \mathbf{D}^+ \mathbf{G}^T \mathbf{X}. \tag{4.9}$$

A (4.7) és (4.9) eredményeket összevetve a (3.1) és (3.6) összefüggésekkel láthatjuk, hogy az $\mathbf{X}$ objektum-kvantifikációk az objektumokra vonatkozó korrespondancia-szkórok N-nel leosztva (hisz itt $\mathbf{X}^T\mathbf{X}=\mathbf{I}_k$, ott pedig $\mathbf{X}^T\mathbf{X}=\left(\frac{t}{tN}\right)^{-1}\mathbf{I}_k=N\cdot\mathbf{I}_k$ volt a mellékfeltétel).

Az $\mathbf{Y}^j$ kategória-kvantifikációkból pedig a többszörös korrespondanciaanalízis szkórjait úgy kapjuk, hogy N-nel szorzunk és $\mathbf{Y}_l^j$ l-edik oszlopát leosztjuk s_l-lel $(l=1, \ldots, k)$.

Végül nézzük meg még, hogyan fejezhetők ki a minimumot adó $\mathbf{X}$ objektum-kvantifikációk a kategória-kvantifikációkkal. Ehhez (4.2)-t most $\mathbf{X}$ szerint deriváljuk. A deriváltat 0-vá téve:

$$\frac{2}{t}\sum_{j=1}^{t}(\mathbf{X}-\mathbf{G}_j\mathbf{Y}^j) = 0,$$

ahonnan

$$\mathbf{X} = \frac{1}{t}\sum_{j=1}^{t}\mathbf{G}_j\mathbf{Y}^j = \frac{1}{t}\mathbf{G}\mathbf{Y}. \tag{4.10}$$

Tehát egy objektum kvantifikációjának szkórjai (minden $l=1, 2, \ldots, k$-ra) azon kategória-szkórok változókra vett átlagai, amely kategóriákat ő kiválasztotta (igennel válaszolta meg). A (3.6) összefüggésben a kategória-szkórokat még oszloponként s_l-lel leosztjuk $(l=1, \ldots, k)$. A (4.9) összefüggés miatt

$$\mathbf{Y}^T\mathbf{D}\mathbf{Y} = \mathbf{X}^T\mathbf{G}\mathbf{D}^+\mathbf{D}\mathbf{D}^+\mathbf{G}^T\mathbf{X} = \mathbf{X}^T\mathbf{G}\mathbf{D}^+\mathbf{G}^T\mathbf{X} = t\mathbf{X}^T\mathbf{P}_*\mathbf{X} =$$

$$= t(\mathbf{u}_1 \ldots \mathbf{u}_k)^T\Big(\sum_{l=1}^{r} s_l^2\mathbf{u}_l\mathbf{u}_l^T\Big)(\mathbf{u}_1 \ldots \mathbf{u}_k) = t\begin{pmatrix} s_1^2 & & 0 \\ & \ddots & \\ 0 & & s_k^2 \end{pmatrix} = t\mathbf{\Lambda},$$

ahol $\mathbf{\Lambda}$ a $\mathbf{P}_*$ mátrix k legnagyobb sajátértékét fődiagonálisában tartalmazó diagonális mátrix.

(Megjegyezzük, hogy az $\mathbf{X}$ szerinti deriválás után kapott (4.10) eredményt (4.2)-be visszahelyettesítve $\mathbf{Y}$-t kapnánk meg, mint egy sajátértékfeladat megoldását az $\mathbf{Y}^T\mathbf{D}\mathbf{Y}=t\mathbf{I}_k$ mellékfeltétellel. Ekkor azonban $\mathbf{X}$ oszlopait kellene az s_l szinguláris értékekkel átosztani, és $\mathbf{X}$-re az $\mathbf{X}^T\mathbf{X}=\mathbf{\Lambda}^{-1}$ mellékfeltételt kapnánk.)

Tekintsük most a következő, beágyazásnak nevezett feladatot: Legyen $\mathbf{A}$ egy $N*t$-s 0–1 mátrix, amely N objektumra és t bináris változóra vonatkozó megfigyelést tartalmaz ($\mathbf{A}$-ból természetesen készíthető egy $N*2t$-s teljes diszjunkt tábla). Adott k $(1\leqq k\leqq r=\text{rang}\,(\mathbf{A}))$ egészhez keresendők a k-dimenziós tér azon $\mathbf{x}_i$ $(i=1, \ldots, N)$ és $\mathbf{y}_j$ $(j=1, \ldots, t)$ pontjai (ezek az objektumok, ill. változók szkórjait

tartalmazzák koordinátáikban), hogy a következő kvadratikus célfüggvény minimális legyen:

$$C = \sum_{i=1}^{N} \sum_{j=1}^{t} a_{ij} \|\mathbf{x}_i - \mathbf{y}_j\|^2.$$

A szkórokat az

$$\mathbf{X} = (\mathbf{x}_1 \dots \mathbf{x}_N)^T, \quad \text{ill.} \quad \mathbf{Y} = (\mathbf{y}_1 \dots \mathbf{y}_t)^T$$

$N*k$-s, ill. $t*k$-s mátrixok tartalmazzák. $\mathbf{Y}$ szerinti deriválással adódik, hogy a minimumot adó $\mathbf{X}, \mathbf{Y}$ párra

$$\mathbf{X} = \mathbf{P}^{-1}\mathbf{A}\mathbf{Y}, \tag{4.11}$$

ahol $\mathbf{P}$ az $\mathbf{A}$ mátrix sorösszegeit fődiagonálisában tartalmazó diagonális mátrix (az általánosság megszorítása nélkül feltehető, hogy ez nem szinguláris). Ezt C-be visszahelyettesítve és a szummákat kibontva keresendő a következő kifejezés minimuma $\mathbf{X}$-ben:

$$\operatorname{tr}\{\mathbf{X}^T(\mathbf{P} - \mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T)\mathbf{X}\}, \tag{4.12}$$

ahol a $t*t$-s $\mathbf{Q}$ diagonális mátrix az $\mathbf{A}$ mátrix oszlopösszegeit tartalmazza diagonálisában. A megoldást azzal a mellékfeltétellel keressük, hogy $\mathbf{X}^T\mathbf{P}\mathbf{X} = \mathbf{I}_k$. A (4.12) kifejezést átalakítva

$$\begin{aligned} \operatorname{tr}\{\mathbf{X}^T(P - \mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T)\mathbf{X}\} = \tag{4.13} \\ = \operatorname{tr}\{(\mathbf{X}^T\mathbf{P}^{1/2})[\mathbf{P}^{-1/2}(\mathbf{P} - \mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T)\mathbf{P}^{-1/2}](\mathbf{P}^{1/2}\mathbf{X})\} = \\ = \operatorname{tr}\{(\mathbf{P}^{1/2}\mathbf{X})^T[\mathbf{I} - \mathbf{P}^{-1/2}\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T\mathbf{P}^{-1/2}](\mathbf{P}^{1/2}\mathbf{X})\}. \end{aligned}$$

Ha $\mathbf{A}$-n, mint kontingenciatáblán korrespondanciaanalízist hajtunk végre, akkor a

$$\mathbf{P}^{-1/2}\mathbf{A}\mathbf{Q}^{-1/2} = \sum_{l=1}^{r} s_l \mathbf{u}_l \mathbf{v}_l^T \tag{4.14}$$

szinguláris felbontást kell elvégezni. Ha $\mathbf{A}$-ban az 1-esek összegét M jelöli, akkor ebből a korrespondancia-faktorokra a

$$\psi_l = \left(\frac{\mathbf{P}}{M}\right)^{-1/2} \mathbf{u}_l = \sqrt{M}\,\mathbf{P}^{-1/2}\mathbf{u}_l \quad (l = 1, \dots, r),$$

ill.

$$\varphi_l = \left(\frac{\mathbf{Q}}{M}\right)^{-1/2} \mathbf{v}_l = \sqrt{M}\,\mathbf{Q}^{-1/2}\mathbf{v}_l \quad (l = 1, \dots, r)$$

összefüggéseket kapjuk, ahol $r = \operatorname{rang}(\mathbf{A})$ és az 1. fejezet szerint $\psi_1 \equiv 1$, $\varphi_1 \equiv 1$, $1 = s_1 \geqq s_2 \geqq \dots \geqq s_r > 0$.

A (4.14) felbontás alapján

$$(\mathbf{P}^{-1/2}\mathbf{A}\mathbf{Q}^{-1/2})(\mathbf{P}^{-1/2}\mathbf{A}\mathbf{Q}^{-1/2})^T = \mathbf{P}^{-1/2}\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T\mathbf{P}^{-1/2} = \sum_{l=1}^{r} s_l^2 \mathbf{u}_l \mathbf{u}_l^T.$$

Ebből

$$\mathbf{I}-\mathbf{P}^{-1/2}\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T\mathbf{P}^{-1/2}=\sum_{l=1}^{r}(1-s_l^2)\mathbf{u}_l\mathbf{u}_l^T$$

spektrálfelbontás, és C minimuma $\sum_{l=1}^{k}(1-s_l^2)$ lesz. Mivel $s_1=1$, a legkisebb sajátérték mindig 0, a hozzá tartozó sajátvektor pedig azonos koordinátákból áll, így az ábrázolás adott k esetén itt is a $(k-1)$-dimenziós térben történik. A mellékfeltétel miatt

$$\mathbf{P}^{1/2}\mathbf{X}=(\mathbf{u}_1\ \dots\ \mathbf{u}_k),$$

azaz $\mathbf{X}$ oszlopaiban a korrespondanciafaktorok konstansszorosa $\left(\frac{1}{\sqrt{M}}\text{-szerese}\right)$ áll.

Az $\mathbf{Y}=\mathbf{Q}^{-1}\mathbf{A}^T\mathbf{X}$ összefüggést a (2.4) átváltási formulákkal összehasonlítva látható, hogy $\mathbf{Y}$ l-edik oszlopa úgy keletkezik, hogy a φ_l korrespondanciafaktort s_l-lel leosztjuk $(l=1, \dots, k)$ és még az összes szkór konstansszorosát $\left(\frac{1}{\sqrt{M}}\text{-szeresét}\right)$ vesszük. Tehát az összefüggés a beágyazás szkórjai és a korrespondanciaanalízissel kapott kanonikus szkórok között hasonló, mint a homogenitásvizsgálat kvantifikációi és a kanonikus szkórok között. Így tehát a beágyazás, a homogenitásvizsgálat és a többszörös korrespondanciaanalízis lényegében ugyanazokból a szinguláris felbontásokból kiindulva határoz meg szkórokat, csak mindegyik másképp normálja őket.

Nemcsak az eredményekben, a feladatok célkitűzésében is látjuk a hasonlóságot, ha azokat a következő, absztrakt módon fogalmazzuk meg:

Az absztrakt korrespondanciaanalízis feladata

Legyenek $\mathbf{I}$ és $\mathbf{J}$ megszámlálható halmazok. Jelölje $L^2(\mathbf{I}, \mathscr{A}, P)$, ill. $L^2(\mathbf{J}, \mathscr{B}, Q)$ az ezekkel, mint alapterekkel vett valószínűségi mezőkön értelmezett 0 várható értékű, véges szórású valószínűségi változók terét. A szorzat valószínűségi mező az R együttes eloszlással van ellátva (melynek megfelelő marginálisai P és Q). Adott $k\geqq 1$ egészre keresendő az a ξ és η valószínűségi vektorváltozó, melyek komponensei az $L^2(\mathbf{I}, \mathscr{A}, P)$, ill. $L^2(\mathbf{J}, \mathscr{B}, Q)$ terekből kerülnek ki és melyekre

$$E_R\|\xi-\eta\|^2 \tag{4.15}$$

minimális azokkal a mellékfeltételekkel, hogy

$$E_R\xi\xi^T=E_P\xi\xi^T=\mathbf{I}_k \tag{4.16}$$

és

$$E_R\eta\eta^T=E_Q\eta\eta^T=\mathbf{I}_k. \tag{4.17}$$

Mivel

$$E_R\|\xi-\eta\|^2=E_R\|\xi\|^2+E_R\|\eta\|^2-2\operatorname{tr}E_R\xi\eta^T$$

és mellékfeltételek miatt $E_R\|\xi\|^2=E_R\|\eta\|^2=k$, ezért a feladat ekvivalens tr $E_R\xi\eta^T$ maximalizálásával. A lineáris operátorok elméletéből jól ismert (l. [2]), hogy a megoldást az

$\mathbf{A}: L^2(\mathbf{J}, \mathscr{B}, Q)\to L^2(\mathbf{I}, \mathscr{A}, P)$ feltételes várható érték operátor (tegyük fel, hogy ez kompakt) szinguláris felbontásának k legnagyobb szinguláris értékéhez tartozó saját elemei adják, vagyis

$$\xi=(\psi_1, \ldots, \psi_k)^T, \quad \eta=(\varphi_1, \ldots, \varphi_k)^T,$$

ahol ψ_l, φ_l az 1. fejezetben bevezetett korrespondancia-faktorok, $(l=1, \ldots, k)$.

Az absztrakt beágyazás feladata

Szintén adottak az előbbi valószínűségi mezők és a szorzattér az R mértékkel. Ekkor is rögzített $k\geqq 1$ egészre keresendő a ξ, η k-dimenziós valószínűségi vektorváltozó pár (komponenseik megint csak az előbbi terekből kerülnek ki), melyre

$$E_R\|\xi-\eta\|^2 \tag{4.18}$$

minimális, mellékfeltételt azonban csak az egyik változóra teszünk, pl.

$$E_R\xi\xi^T=E_P\xi\xi^T=\mathbf{I}_k. \tag{4.19}$$

Vegye fel ξ és η az $\mathbf{x}_i$, ill. $\mathbf{y}_j$ értékeket (ezek k-dimenziós vektorok) $\mathbf{I}$ és $\mathbf{J}$ elemein. Alkossuk ezekből, mint sorvektorokból az $|\mathbf{I}|*k$-s $\mathbf{X}$, ill. $|\mathbf{J}|*k$-s $\mathbf{Y}$ mátrixot, ahol $|\,.\,|$ a tartalmazott halmaz számosságát jelöli és pl. $\mathbf{X}$ i-edik sorának l-edik eleme ξ l-edik komponensének az $i\in\mathbf{I}$ helyen felvett értéke. Akkor a (4.18) célfüggvényt ezekkel a „szkórokkal" felírva, a minimalizálandó várható érték:

$$\sum_{i\in\mathbf{I}}\sum_{j\in\mathbf{J}} r_{ij}\|\mathbf{x}_i-\mathbf{y}_j\|^2.$$

Amennyiben $\mathbf{I}$ és $\mathbf{J}$ véges halmazok és r_{ij} $(i\in\mathbf{I},\ j\in\mathbf{J})$ vagy 0, vagy 1, célfüggvényünk éppen olyan alakú, mint a beágyazás célfüggvénye. A (4.19) mellékfeltételnek az $\mathbf{X}^T\mathbf{P}\mathbf{X}=\mathbf{I}_k$ összefüggés felel meg. Amennyiben az $\mathbf{A}$ 0—1 mátrixot az 1-esek összegével már leosztottuk, látható, hogy $\mathbf{X}$ oszlopaiban a korrespondancia-faktorok állnak és igaz az $\mathbf{Y}=\mathbf{Q}^{-1}\mathbf{R}^T\mathbf{X}$ összefüggés, ahol a $\mathbf{P}$, ill. $\mathbf{Q}$ diagonális mátrixok a p_i: $i\in\mathbf{I}$, ill. $q_j: j\in\mathbf{J}$ marginális valószínűségeket tartalmazzák fődiagonálisukban (tegyük fel, hogy ezek nem 0-k, és természetesen lehetnek mindkét irányban végtelen mátrixok is). Azaz a megoldást adó ξ ugyanaz, mint a (4.15) feladatban és

$$\eta=E(\xi|\mathscr{B})$$

az $\mathbf{Y}=\mathbf{Q}^{-1}\mathbf{R}^T\mathbf{X}$ összefüggés miatt. Az absztrakt korrespondanciaanalízisnél η l-edik komponense $\varphi_l=\frac{1}{s_l}\mathbf{Q}^{-1}\mathbf{R}^T\psi_l$, tehát a szinguláris értékekkel még komponenseként le kell osztani.

Így az absztrakt beágyazás feladatában a minimum kisebb lesz, mint az absztrakt korrespondanciáéban, hiszen itt η éppen ξ-nek a szorzattér $\mathscr{B}$ alterére vett vetülete, míg az absztrakt korrespondancianál a vetület affin képe.

Köszönetnyilvánítás

Végül köszönetet mondok TELEGDI LÁSZLÓnak és CSÁKI PÉTERnek, hogy a korrespondanciaanalízisre felhívták figyelmemet, továbbá TUSNÁDY GÁBORnak és TÓTH JÁNOSnak értékes tanácsaikért.

IRODALOM

[1] BENZÉCRI, J. P. et al., *L'Analyse dés Données. Tome 1. La Taxinomie, Tome 2. L'Analyse dés Correspondances* (Dunod, Paris, 1973).
[2] BOLLA, M., „Hilbert-terek lineáris operátorainak szinguláris felbontása (optimumtulajdonságok statisztikai alkalmazásai és numerikus módszerek)", *Alk. Mat. Lapok*, **13** (1987—88) 189—206.
[3] BREIMAN, L. FRIEDMAN, H. J., "Estimating Optimal Transformations for Multiple Regression and Correlation", *Journal of the American Statistical Association* **80** (1985) 580—619.
[4] FISCHER, R. A., "The Precision of Discriminant Functions", *Annals of Eugenics* 10 (1940) 422—429.
[5] GILULA, Z., "Singular Values Decomposition of Probability Matrices: Probabilistic Aspects of Latent Dichotomous Variables", *Biometrika* **66** (1979) 339—344.
[6] GILULA, Z., HABERMAN, S., "Canonical Analysis of Contingency Tables by Maximum Likelihood", *J. Amer. Stat. Assoc.* **80** (1985) 780—788.
[7] GOODMAN, L. A., KRUSKALL, W. H., "Measures of association for cross-classifications III: approximate sampling theory", *J. Amer. Stat. Assoc.* **58** (1963) 310—364.
[8] GOODMAN, L. A., KRUSKALL, W. H., "Measures of association for cross-classifications IV: simplification of asymptotic variances", *J. Amer. Stat. Assoc.* **67** (1972) 415—421.
[9] GOODMAN, L. A., "Association Models and Canonical Correlation in the Analysis of Cross—Classifications Having Ordered Categories", *J. Amer. Stat. Assoc.* **76** (1981) 320—334.
[10] GOODMAN, L. A., "Discussion in the Round Table on 'The Comparison of Different Approaches to the Analysis of Contingency Tables, in the Light of Applications' ", *Metron* **40** (1982) 344—352.
[11] GORDON, A. D., *Classification Methods for the Exploratory Analysis of Multivariate Data*, (Chapman and Hall, London—New York, 1981).
[12] GREENACRE, M. J., *Correspondence Analysis*, (Academic Press, New York, 1984).
[13] GUTTMAN, L., "The quantification of a class of attributes: a theory and method of scale construction" in: *Horst, P.* (ed.), *The prediction of personal adjustment* (Social Science Research Countil), New York, 1941.
[14] HABERMAN, S. J., "Tests for Independence in two-way Contingency Tables Based on Canonical Correlation and on Linear, by-Linear Interaction", *Ann. Stat.* 9 (1981) 1178—1186.
[15] HEISER, J. W., "Unfolding analysis of proximity data", Ph. D. Thesis, Leiden University, The Netherlands, 1981.
[16] HILL, M. O., "Correspondence Analysis: A Neglected Multivariate Method", *Applied Statistics* 23 (1974) 340—354.
[17] KENDALL, M. G., STUART, A., *The Advanced Theory of Statistics, Vol. 2. Inference and Relationship*, (London, Griffin, 1967) 574—575.
[18] KRUSKAL, J. B., "Three—Way Analysis: Rank and Uniqueness of Tri-Linear Decompositions, with Application to Arithmetic Complexity and Statistics", *Linear Algebra and its Applications* **18** (1977) 95—138.
[19] LANCASTER, H. O., "Canonical Correlations and Partitions of χ^2", *Quarterly Journal of Mathematics* **14** (1963) 220—244.
[20] LENGYEL, T., "A cluster analízis néhány kombinatorikai és valószínűségszámítási problémája", *MTA SZTAKI Tanulmányok*, 188/1986 és kandidátusi értekezés.
[21] De LEUW, J., "The Gifi System of Nonlinear Multivariate Analysis", In: *Proc. of the Data Analysis and Informatics III., Versailles* (ed. *Diday, E. et al.*), (Elsevier Science Publishers B. V. (North Holland), 1984).
[22] MC.KEON, J. J., "Canonical Analysis: Some Relations between Canonical Correlation, Factor Analysis, Discriminant Function Analysis and Scaling Theory", *Psychometric Monograph no. 13 of the Psychometric Society*, 1966.

[23] MÓRI, T. F., SZÉKELY, J. G. *„Többváltozós matematikai statisztika"* (Műszaki könyvkiadó, Budapest, 1986.)
[24] O'NEILL, M. E., "Asymptotic Distributions of the Canonical Correlations from Contingency Tables", *Australian J. Stat.* **20** (1978) 75—82.
[25] O'NEILL, M. E., "Distributional Expansions for Canonical Correlations from Contingency Tables", *J. Royal. Stat. Soc. Ser. B.* **40** (1978) 303—312.
[26] RAO, C. R., "Separation Theorems for Singular Values of Matrices and their Applications in Multivariate Analysis", *J. of. Multivariate Analysis* **9** (1979) 362—377.
[27] RÉNYI, A., "On measures of dependence", *Acta Math. Acad. Sci. Hung.* **10** (1959) 441—451.
[28] SILVEY, D. D., "The Lagrangian Multiplier Test", *Ann. Math. Stat.* **30** (1959) 389—407.
[29] TUSNÁDY, G., „Mátrixok szinguláris felbontása", *Alk. Mat. Lapok* **5** (1979) 375—384.

(Beérkezett: 1987. április 4).

BOLLA MARIANNA
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI
KUTATÓ INTÉZET
1111 BUDAPEST, KENDE U. 13—17.

CORRESPONDENCE ANALYSIS

M. BOLLA

The correspondence analysis is a multivariate statistical method widely spread during the last ten years for the analysis of contingency tables. Though the method itself was well known (e.g. FISCHER, GUTTMAN, LANCASTER, KENDALL and STUART assigned so-called scores to the rows and columns respectively so that the contingency table be approximated by a table of lower rank, like the canonical correlation analysis), the taxonomy of the correspondence analysis was elaborated by the French school (BENZECRI et al.) in 1973. Unfortunately the way of their description can hardly be fitted to the usual discussion of the multivariate statistical methods. GREENACRE. JAN de LEUW, GOODMAN, HABERMAN and GORDON have attempted to reconcile them.

In the present paper we try to describe the task of the correspondence analysis in as a general way as possible (by the singular values decomposition of the conditional expectation operator). In this way the optimal properties of the decomposition follow inmediately: e.g. the so-called canonical factors are maximally correlated under the given constraints (the orthogonality is in terms of the marginals) and give a least squares solution.

Eventually the multiple correspondence analysis is introduced for the description of questionnaires and will be compared with the so-called homogenity analysis elaborated by JAN de LEUW and with a natural embedding method for bivariate indicator matrices.

Alkalmazott Matematikai Lapok **13** (1987—88) 231—284

TÖBBDIMENZIÓS IDŐSOROK ÁLLAPOTTERES LEÍRÁSA

MICHALETZKY GYÖRGY ÉS TUSNÁDY GÁBOR

Budapest

A többdimenziós stacionáris *Gauss folyamatok* széles osztálya állítható elő magasabb dimenziós *Markov folyamat* vetületeként, ez a KÁLMÁN RUDOLF által kidolgozott ún. állapotteres leírás. Noha az elmélet alapjai közel harmincévesek, az utóbbi tíz évben sok jelentős eredmény született, melyek közül a legfontosabbak KEITH GLOVER modell redukciós tételei. Dolgozatunkban ezeket az eredményeket ismertetjük saját felépítésünkben.

TARTALOMJEGYZÉK

1. Bevezetés

Stacionárius *Gauss folyamatok* egy osztályáról lesz szó ebben a dolgozatban. Ez az osztály azokból a q-dimenziós

$$\{\mathbf{y}_t \in R^q,\ t = 0, \pm 1, \ldots\} \tag{1.1}$$

idősorokból áll, amelyekhez található olyan p-dimenziós $\mathbf{x}_t$ stacionárius folyamat, és találhatók olyan

$$\mathbf{A}\colon p\times p,\ \mathbf{\Sigma}\colon p\times p,\ \mathbf{C}\colon q\times p \tag{1.2}$$

mátrixok, ahol $\mathbf{\Sigma}$ szimmetrikus és pozitív szemidefinit, hogy

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_t, \tag{1.3}$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t, \quad t = 0, \pm 1, \ldots. \tag{1.4}$$

Itt $\mathbf{w}_t$ $\mathbf{0}$ várható értékű, Σ kovariancia mátrixú normális eloszlású változó, és $\mathbf{w}_t$ független a

$$\xi = \sum_{j=1}^{n} \mathbf{a}_j^T \mathbf{x}_{t-j} \tag{1.5}$$

alakú valószínűségi változók által generált X_{t-1}^- lineáris altértől.

Feltesszük, hogy $E\mathbf{x}_t = \mathbf{0}$, és $\mathbf{x}_t$ kovariancia mátrixát $\mathbf{P}$-vel jelöljük. Ekkor $\mathbf{x}_t$ stacionaritása az

$$\mathbf{P} = \mathbf{APA}^T + \Sigma \tag{1.6}$$

ún. *Ljapunov egyenletet* jelenti. Legyen

$$\pi_A = \{\mathbf{P}: \mathbf{P} \text{ szimmetrikus és } \mathbf{P} \geqq \mathbf{APA}^T\} \tag{1.7}$$

(ahol az egyenlőtlenség azt jelenti, hogy a két mátrix különbsége pozitív szemidefinit).

1.1. Kérdés. Melyek azok az $\mathbf{A}$ mátrixok, amelyekre π_A nem üres? Ha π_A nem üres, mik az elemei?

Könnyen látható, hogy ha az $\mathbf{A}$, $\mathbf{P}$, Σ mátrixokra teljesül (1.6), akkor (1.3)-nak van stacionárius megoldása. Ennek a folyamatnak a kovariancia függvénye

$$E\mathbf{x}_t \mathbf{x}_0^T = \mathbf{A}^t \mathbf{P}, \tag{1.8}$$

és az $\mathbf{x}_t$ által generált (1.4) folyamat $\mathbf{C}_t = E\mathbf{y}_t \mathbf{y}_0^T$ kovariancia függvénye

$$\mathbf{C}_t = \mathbf{CA}^t \mathbf{B} \tag{1.9}$$

alakú, ahol $t \geqq 0$ és

$$\mathbf{B} = \mathbf{PC}^T. \tag{1.10}$$

1.2. Kérdés. Melyek azok a $\mathbf{C}_t$ mátrix sorozatok, amelyek előállíthatóak (1.9) alakban?

1.3. Kérdés. Adott $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ mátrixok mellett melyek π_A-nak azok a $\mathbf{P}$ elemei, amelyekre teljesül (1.10)?

Jelöljük az

$$\eta = \sum_{j=0}^{n} \mathbf{b}_j^T \mathbf{y}_{t-j} \tag{1.11}$$

alakú valószínűségi változók által generált lineáris alteret Y_t^--szal.

1.1. Definíció. Azt mondjuk, hogy az (1.1) folyamat (1.3)—(1.4) előállítása prediktív, ha X_t^- és Y_t^- azonosak.

1.4. Kérdés. Melyek azok az (1.2) alatti mátrix-hármasok, amelyek az általuk generált (1.3)—(1.4) folyamat prediktív előállítását adják?

1.2. Definíció. Azt mondjuk, hogy az (1.1) folyamat reguláris, ha

$$\bigcap_{t=0}^{\infty} Y_{-t}^- = \emptyset. \tag{1.12}$$

1.5. Kérdés. Hogyan lehet az (1.1) folyamat kovarianciáinak (1.9) előállításából kiolvasni, hogy a folyamat reguláris-e?

Ebben a dolgozatban ezekre a kérdésekre keressük a választ. A közölt tételek KÁLMÁN, FAURRE, DESAI és GLOVER eredményeinek az átfogalmazásai lesznek, de a skaláris szorzatok kereszt reprezentációja eddig nem szerepelt az irodalomban, és a közölt bizonyítások többsége tőlünk származik.

A dolgozat anyaga egy évek óta működő szemínárium munkájának a terméke, amelynek tagjai rajtunk kívül ARATÓ MIKLÓS, BÁRTFAI PÁL, BENCSIK ISTVÁN, FEHÉR ZOLTÁN, GERENCSÉRNÉ VÁGÓ ZSUZSA, GERENCSÉR LÁSZLÓ, GYÖNGY ISTVÁN, LÁSZLÓ ZOLTÁN, LOVÁSZ D. LÁSZLÓ, PRÖHLE TAMÁS, REJTŐ LÍDIA, VERES SÁNDOR és ZIERMANN MARGIT voltak.

A teljesség kedvéért a dolgozat néhány közismert és nyilvánvaló állítás bizonyítását is tartalmazza, elsősorban annak érdekében, hogy a kérdéssel foglalkozók számára önmagában használható bevezetéssé válhasson.

2. Mátrixok faktorizációja

2.1. LEMMA. Jelöljük a $\mathbf{H}$ mátrix sorai által kifeszített alteret $S(\mathbf{H})$-val. Az $n\times m$, illetve $p\times m$ méretű $\mathbf{H}$ és $\mathbf{V}$ mátrixokra akkor és csakis akkor teljesül

$$S(\mathbf{H}) \subseteq S(\mathbf{V}), \tag{2.1}$$

ha van olyan $n\times p$ méretű $\mathbf{M}$ mátrix, hogy

$$\mathbf{H} = \mathbf{MV}. \tag{2.2}$$

Ha itt rang $\mathbf{M}=p$, akkor $S(\mathbf{H})=S(\mathbf{V})$. Az $n\times m$ méretű $\mathbf{H}$ mátrixhoz akkor és csakis akkor találhatóak olyan $n\times p$, $p\times m$ méretű $\mathbf{M}$, $\mathbf{V}$ mátrixok, amelyekre teljesül (2.2), ha rang $\mathbf{H}\leqq p$.

Bizonyítás. Jelöljük $\mathbf{H}$ egy tetszőleges sorát $\mathbf{h}^T$-vel, $\mathbf{M}$ megfelelő sorát $\mathbf{m}^T$-vel. Akkor (2.2) szerint

$$\mathbf{h} = \mathbf{V}^T\mathbf{m},$$

vagyis

$$\mathbf{h} = \sum_{i=1}^{p} m_i \cdot \mathbf{v}_i,$$

ahol az m_i-k $\mathbf{m}$ koordinátái, a $\mathbf{v}_i^T$-k a $\mathbf{V}$ sorai. Tehát (2.2) pontosan azt jelenti, hogy $\mathbf{H}$ sorai előállíthatóak $\mathbf{V}$ sorainak a lineáris kombinációiként, és (2.1) is ezt állítja.

Ha rang $\mathbf{M}=p$, akkor $\mathbf{M}$ soraiból összeállítható egy $p\times p$ méretű reguláris $\mathbf{M}_0$ mátrix, és ha a $\mathbf{H}$ megfelelő soraiból álló mátrixot $\mathbf{H}_0$-lal jelöljük, akkor (2.2) szerint

$$\mathbf{H}_0 = \mathbf{M}_0\mathbf{V}.$$

Ebből $\mathbf{V}=\mathbf{M}_0^{-1}\mathbf{H}_0$ következik, amiből a lemma már bizonyított állítása szerint kapjuk, hogy

$$S(\mathbf{V}) \subseteq S(H_0) \subseteq S(\mathbf{H}),$$

amiből (2.1) miatt $S(\mathbf{V})=S(\mathbf{H})$ következik.

Ha $\mathbf{H}$ előállítható (2.2) alakban, akkor (2.1) miatt rang $\mathbf{H} \leqq$ rang $\mathbf{V} \leqq p$. Megfordítva, ha rang $\mathbf{H} \leqq p$, akkor van olyan $p \times m$ méretű $\mathbf{V}$ mátrix, amelyre eljesül (2.1), és amint láttuk, emiatt $\mathbf{H}$ előállítható (2.2) alakban.

2.1. Definíció. A végtelen

$$\mathbf{H} = (h_{ij},\ 1 \leqq i,\ j < \infty) \tag{2.3}$$

mátrix szeleteinek a

$$\mathbf{H}_{nm} = (h_{ij},\ 1 \leqq i \leqq n,\ 1 \leqq j \leqq m) \tag{2.4}$$

mátrixokat nevezzük $(n, m \geqq 1)$. Azt mondjuk, hogy $\mathbf{H}$ véges rangú, ha szeleteinek rangja korlátos. Ilyenkor $\mathbf{H}$ rangja a legkisebb felső korlát.

2.2. LEMMA. A vételen $\mathbf{H}$ mátrix (2.4) szeleteire akkor és csakis akkor teljesül, hogy

$$\text{rang } \mathbf{H}_{nm} \leqq p, \tag{2.5}$$

ha vannak olyan

$$\mathbf{M} = (m_{ij},\ 1 \leqq i < \infty,\ 1 \leqq j \leqq p) \tag{2.6}$$

és

$$\mathbf{V} = (v_{ij},\ 1 \leqq i \leqq p,\ 1 \leqq j < \infty) \tag{2.7}$$

mátrixok, amelyek

$$\mathbf{M}_n = (m_{ij},\ 1 \leqq i \leqq n,\ 1 \leqq j \leqq p) \tag{2.8}$$

és

$$\mathbf{V}_m = (v_{ij},\ 1 \leqq i \leqq p,\ 1 \leqq j \leqq m) \tag{2.9}$$

szeletére minden $n \geqq 1$ és $m \geqq 1$ mellett teljesül, hogy

$$\mathbf{H}_{nm} = \mathbf{M}_n \mathbf{V}_m. \tag{2.10}$$

Bizonyítás. Az egyik irány nyilvánvaló, hiszen (2.10)-ből következik (2.5). Megfordítva, legyen k olyan nagy, hogy tetszőleges $k \leqq n$, $k \leqq m$ mellett

$$\text{rang } \mathbf{H}_{kk} = \text{rang } \mathbf{H}_{nm}.$$

A 2.1. lemma szerint (2.5)-ből következik, hogy $\mathbf{H}_{kk}$ előállítható

$$\mathbf{H}_{kk} = \mathbf{M}_k \mathbf{V}_k$$

alakban, ahol $\mathbf{M}_k$, $\mathbf{V}_k$ rendre $k \times p$ és $p \times k$ méretűek. Mivel tetszőleges $n \geqq k$ mellett rang $\mathbf{H}_{nk} =$ rang $\mathbf{H}_{kk}$, (2.1) szerint

$$S(\mathbf{H}_{nk}) \subseteq S(\mathbf{H}_{kk}) \subseteq S(\mathbf{V}_k),$$

tehát $\mathbf{H}_{nk}$ új sorai is előállíthatók $\mathbf{V}_k$ sorainak a lineáris kombinációiként. Ez azt jelenti, hogy $\mathbf{M}_k$ kiegészíthető olyan $\mathbf{M}_n$ mátrixra, amelyre $\mathbf{H}_{nk} = \mathbf{M}_n \mathbf{V}_k$ teljesül. Ezután hasonlóan kiegészíthetjük $\mathbf{V}_k$-t olyan $\mathbf{V}_m$ mátrixra, amelyre $\mathbf{H}_{nm} = \mathbf{M}_n \mathbf{V}_m$ teljesül. Ezt az eljárást vég nélkül folytatva kapjuk a (2.6), (2.7) mátrixokat.

3. Alterek skaláris szorzatának kereszt reprezentációja

3.1. Definíció. Egy *Hilbert-tér* A és B altereinek skaláris szorzat kereszt reprezentációja (SSzKR) az

$$\mathscr{A}\colon A \to R^p, \quad \mathscr{B}\colon B \to R^p \tag{3.1}$$

lineáris leképezés pár, ha minden $\mathbf{a}\in A$, $\mathbf{b}\in B$ párra teljesül, hogy

$$(\mathscr{A}\mathbf{a}, \mathscr{B}\mathbf{b}) = (\mathbf{a}, \mathbf{b}), \tag{3.2}$$

ahol $(\mathbf{u}, \mathbf{v})$ az $\mathbf{u}$, $\mathbf{v}$ elemek skaláris szorzatát jelöli. Ha az A, B altereknek van SSzKR-ja, akkor azt mondjuk, hogy a kereszt ranguk véges. Ilyenkor a legkisebb p-t, amelyre teljesül (3.1), a két altér kereszt rangjának nevezzük, és corank (A, B)-vel jelöljük.

3.1. LEMMA. Ha $p = \text{corank}\,(A, B)$, és a (3.1) leképezés pár (A, B)-nek SSzKR-ja, akkor $\mathscr{A}$ és $\mathscr{B}$ képtere a teljes R^p.

Bizonyítás. Jelöljük S-sel $\mathscr{A}$ képterét, és $\mathscr{P}_S$-sel az S-re való vetítést. Akkor

$$(\mathscr{P}_S\mathscr{A}\mathbf{a}, \mathscr{P}_S\mathscr{B}\mathbf{b}) = (\mathscr{P}_S^2\mathscr{A}\mathbf{a}, \mathscr{B}\mathbf{b}) = (\mathscr{A}\mathbf{a}, \mathscr{B}\mathbf{b})$$

miatt $(\mathscr{P}_S\mathscr{A}, \mathscr{P}_S\mathscr{B})$ is SSzKR, mivel p minimális, $p = \dim S$. Hasonlóan látható be, hogy $\mathscr{B}$ képtere is a teljes R^p.

3.2. Definíció. Legyen $\mathscr{A}$, $\mathscr{B}$ tetszőleges SSzKR. Jelöljük A_1-gyel, B_1-gyel az

$$A_1 = \{\mathbf{a}\in A,\ \mathscr{A}\mathbf{a} = \mathbf{0}\}, \tag{3.3}$$

$$B_1 = \{\mathbf{b}\in B,\ \mathscr{B}\mathbf{b} = \mathbf{0}\} \tag{3.4}$$

altereket, és legyen A_0, B_0 az A_1, B_1 alterek A-beli illetve B-beli ortogonális komplementere. Az A_0, B_0 altereket az $\mathscr{A}$, $\mathscr{B}$ SSzKR saját altereinek nevezzük.

3.2. LEMMA. Ha $p = \text{corank}\,(A, B)$, és az $(\mathscr{A}, \mathscr{B})$ leképezéspár (A, B)-nek SSzKR-ja, akkor $\mathscr{A}$ és $\mathscr{B}$ saját altereire

$$A_0 = \mathscr{P}_A B, \quad B_0 = \mathscr{P}_B A \tag{3.5}$$

teljesül, ahol $\mathscr{P}_A$, $\mathscr{P}_B$ az A-ra, illetve B-re vetítő operátor.

Bizonyítás. A (3.3) altér merőleges B-re, ezért $\mathscr{P}_A B \subset A_0$. Legyen

$$\mathscr{L}\colon A_0 \to R^p$$

tetszőleges skaláris szorzat tartó lineáris leképezés. Akkor $(\mathscr{L}\mathscr{P}_A, \mathscr{L}\mathscr{P}_A)$ SSzKR, emiatt $\mathscr{L}\mathscr{P}_A B$ dimenziója nem lehet p-nél kisebb. Tehát $\mathscr{P}_A B$ azonos A_0-lal.

KÖVETKEZMÉNY. A $\mathscr{P}_A B$, $\mathscr{P}_B A$ alterek dimenziója egyenlő. (Ez akkor is igaz, ha az alterek dimenziója végtelen.)

3.3. Definíció. Az A, B alterek $(\mathbf{a}_1, \mathbf{a}_2, \ldots)$ $(\mathbf{b}_1, \mathbf{b}_2, \ldots)$ bázisának skaláris szorzat kereszt mátrixa a

$$\mathbf{C} = \left(C_{ij} = (\mathbf{a}_i, \mathbf{b}_j),\ i, j \geqq 1\right) \tag{3.6}$$

mátrix.

3.3. LEMMA. Az A, B alterek kereszt rangja akkor és csakis akkor véges, ha tetszőleges bázisaik $\mathbf{C}$ skaláris szorzat kereszt mátrixának rangja véges. A $\mathbf{C}$ mátrix tetszőleges,

$$\mathbf{C} = \mathbf{MV} \tag{3.7}$$

faktorizációjának megfelel az (A, B)-nek egy SSzKR-ja, melyben $\mathscr{A}\mathbf{a}_i$ az $\mathbf{M}$ i-edik sora és $B\mathbf{b}_j$ a $\mathbf{V}$ j-edik oszlopa. Megfordítva, (A, B) minden (3.1) SSzKR-jához hozzárendelhető $\mathbf{C}$-nek (3.7) alatti faktorizációja.

Bizonyítás. A (3.7) faktorizáció létezése a 2.2. lemma szerint ekvivalens $\mathbf{C}$ rangjának végességével. Ez a faktorizáció ugyanakkor azzal ekvivalens, hogy a lemmában mondott megfeleltetés a bázisok elemein SSzKR, ami nyilván kiterjeszthető a teljes alterekre.

3.4. Definíció. Az A, B alterek $(\mathbf{a}_1, \mathbf{a}_2, \ldots)$, $(\mathbf{b}_1, \mathbf{b}_2, \ldots)$ bázisait konjugált bázispárnak nevezzük, ha $1 \leqq i \leqq p$ mellett $\mathbf{a}_i \in A_0$, $\mathbf{b}_i \in B_0$, és

$$(\mathbf{a}_i, \mathbf{b}_j) = \delta_{ij}, \tag{3.8}$$

ahol δ_{ij} a *Kronecker függvény*, továbbá $i > p$ mellett $\mathbf{a}_i \in A_1$, $\mathbf{b}_i \in B_1$, ahol $A_0 = \mathscr{P}_A B$, $B_0 = \mathscr{P}_B A$, és A_1, B_1 az A_0, B_0 ortogonális komplementerei.

3.4. LEMMA. Legyen $p = \operatorname{corank}(A, B)$, és legyen a (3.1) leképezéspár (A, B)-nek SSzKR-ja, továbbá legyen $(\mathbf{e}_1, \ldots, \mathbf{e}_p)$ az R^p ortonormált bázisa.

Jelöljük $\mathbf{a}_i$-vel, illetve $\mathbf{b}_i$-vel az $\mathbf{e}_i A_0$-beli, illetve B_0-beli inverz képét és egészítsük ki bázissá ezeket az elemeket testzés szerinti A_1, B_1-beli elemekkel. Így konjugált bázispárt kapunk, és megfordítva, ha $(\mathbf{e}_1, \ldots, \mathbf{e}_p)$ az R^p ortonormált bázisa és $(\mathbf{a}_1, \mathbf{a}_2, \ldots)$, $(\mathbf{b}_1, \mathbf{b}_2, \ldots)$ tetszőleges konjugált bázispár, akkor az

$$\mathscr{A}\colon \mathbf{a}_i \to \mathbf{e}_i, \quad \mathscr{B}\colon \mathbf{b}_i \to \mathbf{e}_i, \quad 1 \leqq i \leqq p \tag{3.9}$$

leképezés SSzKR.

Bizonyítás. Ha $i > p$ vagy $j > p$, akkor egy konjugált bázispárra

$$(\mathbf{a}_i, \mathbf{b}_j) = 0,$$

tehát (3.8) pontosan azt jelenti, hogy (3.9) SSzKR. (Megjegyezzük, hogy a 3.2. lemma szerint a 3.4. definícióban szereplő A_0, B_0 alterek tetszőleges minimális dimenziójú SSzKR-ra saját alterek, és ezekre megszorítva a SSzKR leképezései egyértelműek.)

3.5. Definíció. Az A, B alterek konjugált bázispárját kanonikus bázispárnak nevezzük, ha ortogonálisak.

3.6. Definíció. Az $\mathbf{x}_i \in B_0$, $\mathbf{z}_i \in A_0$, $1 \leqq i \leqq p$ bázispárt vetített konjugált bázispárnak nevezzük, ha van olyan $(\mathbf{a}_1, \mathbf{a}_2, \ldots,)$, $(\mathbf{b}_1, \mathbf{b}_2, \ldots)$ konjugált bázispár, melyre

$$\mathbf{x}_i = \mathscr{P}_B \mathbf{a}_i, \quad \mathbf{z}_i = \mathscr{P}_A \mathbf{b}_i, \quad 1 \leqq i \leqq p. \tag{3.10}$$

3.5. LEMMA. Ha $p = \operatorname{corank}(A, B)$, és a (3.1) leképezéspár (A, B)-nek SSzKR-ja, $(\mathbf{e}_1, \ldots, \mathbf{e}_p)$ ortogonált bázis R^p-ben, $(\mathbf{a}_1, \mathbf{a}_2, \ldots)$, $(\mathbf{b}_1, \mathbf{b}_2, \ldots)$ a megfelelő konjugált bázispár, amelyből $\mathbf{x}_i, \mathbf{z}_i$ a (3.10) szerint kapható, továbbá tetszőleges $\mathbf{a} \in A$ mellett $\mathscr{A}\mathbf{a}$ i-edik koordinátája α_i, akkor

$$\mathscr{P}_B \mathbf{a} = \sum_{i=1}^{p} \alpha_i \mathbf{x}_i. \tag{3.11}$$

Bizonyítás. Legyen $\mathscr{P}_{A_0}\mathbf{a} = \sum_{i=1}^{p} \alpha_i \mathbf{a}_i$, akkor $\mathscr{A}\mathbf{a}$ i-edik koordinátája α_i. Ebből (3.11) $\mathscr{P}_B$-vel való szorzással kapható.

3.6. LEMMA. Legyen $\mathbf{a}_i$, $\mathbf{b}_i$ egy konjugált bázispár, és $\mathbf{x}_i$, $\mathbf{z}_i$ a belőle származó vetített konjugált bázispár. Legyen továbbá

$$a_{ij} = (\mathbf{a}_i, \mathbf{a}_j),$$
$$b_{ij} = (\mathbf{b}_i, \mathbf{b}_j),$$
$$x_{ij} = (\mathbf{x}_i, \mathbf{x}_j),$$
$$y_{ij} = (\mathbf{x}_i, \mathbf{z}_j),$$
$$z_{ij} = (\mathbf{z}_i, \mathbf{z}_j),$$

és jelöljük a megfelelő $p \times p$ méretű mátrixokat rendre **A**-val, **B**-vel, **X**-szel, **Y**-nal, **Z**-vel. Akkor

$$\mathbf{X} = \mathbf{B}^{-1}, \quad \mathbf{Z} = \mathbf{A}^{-1}, \quad \mathbf{Y} = \mathbf{XZ}. \tag{3.12}$$

Bizonyítás. Jelöljük az $\mathbf{a}_i$, $\mathbf{b}_i$, $\mathbf{x}_i$, $\mathbf{z}_i$ elemekből álló p-dimenziós vektorokat **a**-val, **b**-vel, **x**-szel, **z**-vel. Akkor

$$\mathbf{z} = \mathbf{A}^{-1}\mathbf{a}, \quad \mathbf{x} = \mathbf{B}^{-1}\mathbf{b},$$

hiszen ezekkel az elemekkel

$$(\mathbf{b}-\mathbf{z})\mathbf{a}^T = 0, \quad (\mathbf{a}-\mathbf{x})\mathbf{b}^T = 0.$$

3.1. TÉTEL. Ha $p = \operatorname{corank}(A, B)$, és a (3.1) leképezéspár (A, B)-nek SSzKR-ja, továbbá

$$\tilde{\mathscr{A}}: A \to R^m, \quad \tilde{\mathscr{B}}: B \to R^m \tag{3.13}$$

tetszőleges SSzKR-ja (A, B)-nek C, D képterekkel, akkor $R^m(C, D)$ altereinek van olyan

$$\mathscr{C}: C \to R^p, \quad \mathscr{D}: D \to R^p \tag{3.14}$$

SSzKR-ja, amelyre

$$\mathscr{A} = \mathscr{C}\tilde{\mathscr{A}}, \quad \mathscr{B} = \mathscr{D}\tilde{\mathscr{B}} \tag{3.15}$$

Bizonyítás. Legyen $(\mathbf{e}_1, \ldots, \mathbf{e}_p)$ az R^p ortogonált bázisa, $(\mathbf{a}_1, \mathbf{a}_2, \ldots)$, $(\mathbf{b}_1, \mathbf{b}_2, \ldots)$ a megfelelő konjugált bázispár, és legyen

$$\mathbf{c}_i = \tilde{\mathscr{A}}\mathbf{a}_i, \quad \mathbf{d}_i = \tilde{\mathscr{B}}\mathbf{b}_i.$$

Az nem lehet, hogy itt $1 \leqq i \leqq p$ mellett $\mathbf{c}_i$ vagy $\mathbf{d}_i$ $\mathbf{0}$ legyen, hiszen abból $\mathbf{c}_i \perp B$, illetve $\mathbf{d}_i \perp A$, és végül $\mathbf{a}_i \perp B$, illetve $\mathbf{b}_i \perp A$ következne. Legyen

$$\mathscr{C}: \mathbf{c}_i \to \mathbf{e}_i, \quad \mathscr{D}: \mathbf{d}_i \to \mathbf{e}_i,$$

ez pontosan az a SSzKR, amire szükségünk van.

KÖVETKEZMÉNY. Ha $m=p$, akkor a $(\mathscr{C}, \mathscr{D})$ párra R -ben

$$(\mathscr{C}\mathbf{x}, \mathscr{D}\mathbf{y}) = (\mathbf{x}, \mathbf{y})$$

teljesül minden $\mathbf{x}, \mathbf{y} \in R^p$ mellett. Ebből

$$\mathscr{D} = \mathscr{C}^{-T} \tag{3.16}$$

következik, tehát $\mathscr{C}$ nem szinguláris. Megfordítva, ha $\mathscr{C}$ nem szinguláris, és (3.1) minimális dimenziójú SSzKR, akkor

$$(\mathscr{C}\mathscr{A}, \mathscr{C}^{-T}\mathscr{B}) \tag{3.17}$$

nyilván minimális dimenziójú SSzKR, a tétel szerint tehát (3.17) adja meg az összes minimális dimenziójú SSzKR-t, ha $(\mathscr{A}, \mathscr{B})$ rögzített, és $\mathscr{C}$ befutja az R^p tér nem szinguláris lineáris leképezéseit. Ha $(\mathbf{a}_1, \ldots, \mathbf{a}_p)$, $(\mathbf{b}_1, \ldots, \mathbf{b}_p)$ az $(\mathscr{A}, \mathscr{B})$ párhoz tartozó konjugált bázispár A_0, B_0-beli része (nyilván elegendő ezt megadni a konjugált bázispárból), és $(\tilde{\mathbf{a}}_1, \ldots, \tilde{\mathbf{a}}_p)$ az A_0 tetszőleges bázisa, amelyet $\mathscr{A}$ az R^p $(\mathbf{f}_1, \ldots, \mathbf{f}_p)$ bázisába visz, legyen $\mathscr{C}$ az a leképezés, amely $(\mathbf{f}_1, \ldots, \mathbf{f}_p)$-t $(\mathbf{e}_1, \ldots, \mathbf{e}_p)$-be viszi. Az így kapott (3.17) SSzKR-hoz tartozó konjugált bázispár első tagja nyilván $(\tilde{\mathbf{a}}_1, \ldots, \tilde{\mathbf{a}}_p)$ lesz, tehát a konjugált bázispár egyik tagját tetszés szerint megválaszthatjuk, ez a másikat egyértelműen meghatározza.

Azt, hogy egyáltalán van SSzKR, ha $\mathscr{P}_A B$ véges dimenziójú, többféleképpen is beláthatjuk. Egyrészt a 3.2. lemma bizonyításában explicit megadtunk egyet, másrészt belátható, hogy a $\mathscr{P}_A\mathscr{P}_B\mathscr{P}_A$, $\mathscr{P}_B\mathscr{P}_A\mathscr{P}_B$ operátorok saját vektorai megfelelő sorrendben kanonikus bázispárt adnak. Előállítható egy konjugált bázispár szekvenciálisan is, lépésenként választva meg az $(\mathbf{a}_i, \mathbf{b}_i)$ párokat. További lehetőség, hogy kiindulunk egy tetszőleges bázispárból, és annak a skaláris szorzat kereszt mátrixában (3.3. definíció) keresünk a (3.7) faktorizációhoz használható bázist. Például úgy, hogy $\mathbf{C}$ maximális rangú szeletében (2.1. Definíció) a sorok vagy oszlopok között keresünk lineárisan függetleneket. (Láttuk, hogy a faktorizáció SSzKR-t ad. Ilyenkor persze a konjugált bázispárt még valamilyen eszközzel elő kell állítani.)

3.7. Definíció. A *Hilbert-tér* elemű p-dimenziós

$$\mathbf{z}^T = (z_1, \ldots, z_p), \quad \mathbf{x}^T = (x_1, \ldots, x_p)$$

vektorpárt az A, B alterek *reprezentánsának* nevezzük, ha

$$\mathbf{z}_i \in A, \quad \mathbf{x}_i \in B, \quad 1 \leqq i \leqq p,$$

és tetszőleges $\mathbf{u} \in A$, $\mathbf{v} \in B$ mellett teljesül, hogy

$$(\mathbf{u}, \mathbf{v}) = ((\mathbf{u}, \mathbf{z}), (\mathbf{v}, \mathbf{x})). \tag{3.18}$$

Megjegyezzük, hogy (3.18) jobb oldalán $(\mathbf{u}, \mathbf{z})$ valós együtthatós p-dimenziós vektor, amelynek a k-adik koordinátája az $(\mathbf{u}, \mathbf{z}_k)$ skaláris szorzat $((\mathbf{v}, \mathbf{x})$ értelmezése hasonló).

3.8. Definíció. A lineárisan független elemekből álló

$$\mathbf{z}^T = (\mathbf{z}_1, \ldots, \mathbf{z}_p)$$

bázis *duálisa* az

$$\mathbf{a}^T = (\mathbf{a}_1, \ldots, \mathbf{a}_p)$$

bázis, ha $1 \leqq i \leqq p$ mellett $\mathbf{a}_i$ benne van a $\mathbf{z}_1, \ldots, \mathbf{z}_p$ elemek által kifeszített altérben, és

$$(\mathbf{a}_i, \mathbf{z}_j) = \delta_{ij}, \quad 1 \leqq i, j \leqq p. \tag{3.19}$$

3.7. LEMMA. Jelölje **A**, **Z** ugyanazokat a mátrixokat, mint a 3.6. lemmában, akkor **z** bázis duálja az

$$\mathbf{a} = \mathbf{Z}^{-1}\mathbf{z} \tag{3.20}$$

bázis, és itt $\mathbf{Z}^{-1} = \mathbf{A}$.

Bizonyítás.

$$\mathbf{a}\mathbf{z}^T = \mathbf{Z}^{-1}\mathbf{z}\mathbf{z}^T = \mathbf{I}.$$

3.8. LEMMA. Ha a **z**, **x** vektorpár az A, B alterek reprezentánsa, akkor az

$$\mathscr{A}\mathbf{u} = (\mathbf{u}, \mathbf{z}), \quad \mathscr{B}\mathbf{v} = (\mathbf{v}, \mathbf{x}) \tag{3.21}$$

lineáris leképezések az A, B alterek SSzKR-t adják. Megfordítva, ha $\mathscr{A}$, $\mathscr{B}$ SSzKR, akkor (3.21) egyértelműen meghatároz egy (**z**, **x**) vektorpárt, és ez reprezentáns. Ha (3.21) minimális dimenziójú, akkor

$$\mathbf{x} = \mathscr{P}_B\mathbf{a}, \quad \mathbf{z} = \mathscr{P}_A\mathbf{b}, \tag{3.22}$$

ahol az **a** vektor a **z** és **b** az **x** duálja, és itt **a**, **b** konjugált bázispár.

Bizonyítás. A 3.7. definícióban szereplő (3.18) feltétel pontosan azt jelenti, hogy (3.21) SSzKR. *Riesz tétele* szerint egy *Hilbert-tér* feletti funkcionál mindig előállítható skaláris szorzatként, így (3.21) alapján a **z**, **x** vektorok koordinátánként megépíthetőek. Egy SSzKR nyilván akkor minimális dimenziójú, ha reprezentánsainak a koordinátái lineárisan függetlenek. Ha az **a** bázis a **z** duálja, akkor minden $\mathbf{v} \in B$ elemre teljesül, hogy

$$(\mathbf{v}, \mathbf{a}_i) = \big((\mathbf{v}, \mathbf{x}), (\mathbf{a}_i, \mathbf{z})\big) = (\mathbf{v}, \mathbf{x})_i,$$

ahol az utolsó forma a (**v**, **x**) vektor i-edik koordinátáját jelöli. Tehát a (**v**, **a**), (**v**, **x**) vektorok megegyeznek, emiatt **x** az **a** vektor B-re eső vetülete. Hasonlóan látható be (3.22) másik állítása is. Végül (3.18) szerint $\mathbf{Z}\mathbf{X} = \mathbf{z}\mathbf{x}^T$, tehát

$$\mathbf{a}\mathbf{b}^T = \mathbf{Z}^{-1}\mathbf{z}\mathbf{x}^T\mathbf{X}^{-1} = \mathbf{I},$$

vagyis **a**, **b** valóban konjugált bázispár.

3.7. és 3.8. lemma alapján úgy kaphatunk a legegyszerűbben konjugált bázispárt, hogy A_0, B_0 közül az egyikben felveszünk egy tetszőleges bázist, és annak duálját levetítjük a másik altérre. Maga a dualizálás nyilván egyszerűbb ha ortogonális vagy ortonormált bázisból indulunk ki (ez utóbbiak duálja önmaguk). Látható, hogy a bázispár struktúrája akkor a legegyszerűbb, ha kanonikus, vagyis olyan ortogonált bázisból indulunk ki, amelynek a vetülete ortogonális. Ez a feladat ekvivalens a vetítés operátorának szinguláris felbontásával, ez utóbbinak a kanonikus korrelációval való kapcsolatáról bővebben TUSNÁDY (1979) és BOLLA (1988) dolgozatában, valamint a MÓRI—SZÉKELY (1986) könyvben van szó.

4. Hankel-sorozatok

4.1. Definíció. A $q\times r$ méretű $(\mathbf{\Phi}_0, \mathbf{\Phi}_1, \ldots)$ mátrixokból álló végtelen sorozat *Hankel mátrixa* a

$$\mathbf{H} = \begin{pmatrix} \mathbf{\Phi}_0 & \mathbf{\Phi}_1 & \ldots & \mathbf{\Phi}_t & \ldots \\ \mathbf{\Phi}_1 & \mathbf{\Phi}_2 & \ldots & \mathbf{\Phi}_{t+1} & \ldots \\ \vdots & \vdots & & \vdots & \\ \mathbf{\Phi}_t & \mathbf{\Phi}_{t+1} & \ldots & \mathbf{\Phi}_{2t} & \ldots \\ \vdots & \vdots & & \vdots & \end{pmatrix} \tag{4.1}$$

végtelen hipermátrix. Ennek $\mathbf{H}_{nm}$ szelete az első $n+1$ hiper sorból és első $m+1$ hiper oszlopból álló mátrix.

4.1. TÉTEL. A $(\mathbf{\Phi}_0, \mathbf{\Phi}_1, \ldots)$ sorozat elemeihez akkor és csakis akkor találhatók olyan

$$\mathbf{A}: p\times p, \quad \mathbf{B}: p\times r, \quad \mathbf{C}: q\times p \tag{4.2}$$

mátrixok, hogy

$$\mathbf{\Phi}_t = \mathbf{C}\mathbf{A}^t\mathbf{B}, \quad t = 0, 1, \ldots, \tag{4.3}$$

ha a sorozat $\mathbf{H}$ *Hankel mátrixának* a rangja legfeljebb p. Ha $\mathbf{H}$ rangja p, és $\mathbf{\Phi}_t$ (4.3)-on kívül a

$$\tilde{\mathbf{A}}: p\times p, \quad \tilde{\mathbf{B}}: p\times r, \quad \tilde{\mathbf{C}}: q\times p \tag{4.4}$$

mátrixokkal is előállítható

$$\mathbf{\Phi}_t = \tilde{\mathbf{C}}\tilde{\mathbf{A}}^t\tilde{\mathbf{B}}, \quad t = 0, 1, \ldots \tag{4.5}$$

alakban, akkor van olyan $p\times p$ méretű $\mathbf{S}$ nem szinguláris mátrix, melyre

$$\tilde{\mathbf{C}} = \mathbf{C}\mathbf{S}, \quad \tilde{\mathbf{A}} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}, \quad \tilde{\mathbf{B}} = \mathbf{S}^{-1}\mathbf{B}. \tag{4.6}$$

Bizonyítás. Ha $\mathbf{\Phi}_t$ előállítható (4.3) alakban, akkor legyen

$$\mathbf{M}_n = \begin{pmatrix} \mathbf{C} \\ \mathbf{C}\mathbf{A} \\ \vdots \\ \mathbf{C}\mathbf{A}^n \end{pmatrix}, \quad \mathbf{V}_m = (\mathbf{B}, \mathbf{A}\mathbf{B}, \ldots, \mathbf{A}^m\mathbf{B}). \tag{4.7}$$

Ezekkel (4.3) szerint $\mathbf{H}_{nm} = \mathbf{M}_n\mathbf{V}_m$. A 2.2. lemma szerint $\mathbf{H}$ akkor és csakis akkor véges rangú, ha faktorizálható, tehát (4.3)-ból következik, hogy $\mathbf{H}$ rangja legfeljebb p. Megfordítva annyi minden esetre következik a 2.2. lemmából, hogy ha $\mathbf{H}$ rangja legfeljebb p, akkor van a $q\times p$ méretű $\mathbf{M}_0, \mathbf{M}_1, \ldots$, illetve $p\times r$ méretű $\mathbf{V}_0, \mathbf{V}_1, \ldots$ mátrixoknak olyan végtelen sorozata, hogy az

$$\mathbf{M}_n = \begin{pmatrix} \mathbf{N}_0 \\ \mathbf{N}_1 \\ \vdots \\ \mathbf{N}_n \end{pmatrix}, \quad \mathbf{V}_m = (\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_m) \tag{4.8}$$

szeletekre $\mathbf{H}_{nm} = \mathbf{M}_n\mathbf{V}_m$ teljesül.

Megmutatjuk, hogy a (4.8) mátrixok (4.7) alakúak. Legyen $\mathbf{H}_{nm}$ maximális rangú, és tegyük fel, hogy a rangja p. Vezessük be a

$$\mathbf{H}_{nm}^{\mathrm{LE}} = \begin{pmatrix} \mathbf{\Phi}_1 & \cdots & \mathbf{\Phi}_{m+1} \\ \vdots & & \vdots \\ \mathbf{\Phi}_{n+1} & \cdots & \mathbf{\Phi}_{n+m+1} \end{pmatrix} \quad \mathbf{M}_n^{\mathrm{LE}} = \begin{pmatrix} \mathbf{N}_1 \\ \vdots \\ \mathbf{N}_{n+1} \end{pmatrix}$$

$$\mathbf{V}_m^{\mathrm{JOBB}} = (\mathbf{W}_1, \ldots, \mathbf{W}_{m+1})$$

jelöléseket. Ezekre $\mathbf{H}$ *Hankel-struktúrája* miatt

$$\mathbf{H}_{nm}^{\mathrm{LE}} = \mathbf{M}_n^{\mathrm{LE}} \mathbf{V}_m = \mathbf{M}_n \mathbf{V}_m^{\mathrm{JOBB}} \tag{4.9}$$

teljesül. Így rang $\mathbf{M}_n = p$ miatt a 2.1. lemma szerint

$$S(\mathbf{V}_m^{\mathrm{JOBB}}) = S(\mathbf{H}_{nm}^{\mathrm{LE}}) \subseteq S(\mathbf{V}_m),$$

tehát van olyan $p \times p$ méretű $\mathbf{A}$ mátrix (ugyancsak a 2.1. lemma szerint), hogy

$$\mathbf{V}_m^{\mathrm{JOBB}} = \mathbf{A}\mathbf{V}_m,$$

vagyis

$$\mathbf{W}_t = \mathbf{A}\mathbf{W}_{t-1}, \quad \text{ha} \quad 1 \leqq t \leqq m. \tag{4.10}$$

Ezt (4.9)-be helyettesítve kapjuk, hogy

$$\mathbf{M}_n^{\mathrm{LE}} \mathbf{V}_m = \mathbf{M}_n \mathbf{A} \mathbf{V}_m,$$

amiből rang $\mathbf{V}_m = p$ miatt $\mathbf{M}_n^{\mathrm{LE}} = \mathbf{M}_n \mathbf{A}$, vagyis

$$\mathbf{N}_t = \mathbf{N}_{t-1}\mathbf{A}, \quad \text{ha} \quad 1 \leqq t \leqq n. \tag{4.11}$$

Ha ezután a fenti meggondolást valamilyen $\tilde{n} \geqq n$ vagy $\tilde{m} \geqq m$ mellett megismételjük, kapjuk, hogy az (n, m) pár mellett kapott $\mathbf{A}$ mátrixra (4.10), illetve (4.11) m helyett $\tilde{m}$-re, illetve n helyett $\tilde{n}$-re is igaz. Legyen $\mathbf{N}_0 = \mathbf{C}$, $\mathbf{W}_0 = \mathbf{B}$, akkor ezek szerint tetszőleges $t \geqq 0$ mellett igaz, hogy

$$N_t = \mathbf{C}\mathbf{A}^t, \quad \mathbf{W}_t = \mathbf{A}^t\mathbf{B},$$

amiből (4.3) következik.

A 3.1. tétel szerint $\mathbf{H}$ minimális dimenziójú $\mathbf{M}_n\mathbf{V}_m$, $\tilde{\mathbf{M}}_n\tilde{\mathbf{V}}_m$ faktorizációhoz mindig van olyan $\mathbf{S}$ nem szinguláris mátrix, melyre

$$\tilde{\mathbf{M}}_n = \mathbf{M}_n\mathbf{S}, \quad \tilde{\mathbf{V}}_m = \mathbf{S}^{-1}\mathbf{V}_m,$$

amit a fenti meggondolásba helyettesítva kapjuk (4.6)-ot.

Ezt a tételt alkalmazva $\mathbf{0}$ várható értékű q-dimenziós $\mathbf{y}_t$ folyamat $\mathbf{C}_t = E\mathbf{y}_t\mathbf{y}_0^T$ kovariancia függvényére kapjuk az 1.2. kérdésre a választ. $\mathbf{C}_t$ pontosan akkor (1.9) alakú, ha corank $(Y_t^+, Y_t^-) \leqq p$, ahol Y_t^- az (1.11), és Y_t^+ az

$$\eta = \sum_{j=0}^{n} \mathbf{a}_j^T \mathbf{y}_{t+j} \tag{4.12}$$

alakú valószínűségi változók által generált lineáris altér. A $\mathbf{C}_t$ sorozat által generált *Hankel-mátrix* rangja ugyanis a két altér skaláris szorzatainak a kereszt mátrixa, ha

bázisnak az $\mathbf{y}_t$ folyamat koordinátáit választjuk. Tehát $\mathbf{C}_t$ (1.9) faktorizációjához minden t mellett tartozik az Y_t^+, Y_t^- altereknek egy SSzKR-ja. Mivel az Y_s^+, Y_t^+, illetve Y_s^-, Y_t^- terek között az idő eltolása egy skaláris szorzat tartó kölcsönösen egyértelmű megfeleltetést létesít, az (1.9) faktorizáció egyidejűleg minden t-re megad az (Y_t^+, Y_t^-) alterekben egy $\mathbf{z}_t$, $\mathbf{x}_t$ reprezentációt. Mivel $\mathbf{x}_t$ koordinátái Y_t^--beliek, illetve $\mathbf{z}_t$-éi Y_t^+-beliek, továbbá a mondott idő-eltolás miatt az $(\mathbf{x}_t, \mathbf{z}_t)$ pár *stacionárius Gauss-folyamatot* alkot.

A 3.5. lemma alapján $\mathbf{CA}^j\mathbf{x}_t$ az $\mathbf{y}_{t+j}$ vektornak Y_t^--re eső vetülete, ami normális eloszlású változókról lévén szó egyben feltételes várható értéket is jelent. Mivel ugyanakkor ha $t>0$, akkor $\mathbf{y}_{t+j}$-nek Y_0^--ra eső vetülete $\mathbf{CA}^{j+t}\mathbf{x}_0$, és $Y_0^- \subset Y_t^-$, emiatt $\mathbf{CA}^j\mathbf{x}_t$-nek Y_0^--ra eső vetülete tetszőleges $j \geqq 0$ mellett $\mathbf{A}^t\mathbf{x}_0$-nak $\mathbf{CA}^j$-szerese. A $\mathbf{CA}^j$ mátrixok sorai $j \geqq 0$ mellett kifeszítik a p-dimenziós alteret, ebből következik, hogy

$$\mathbf{x}_t = \mathbf{A}^t\mathbf{x}_0 + \mathbf{w}_{0t}$$

tetszőleges t mellett, ahol $\mathbf{w}_{0t}$ független Y_t^--tól.

A stacionaritás miatt tetszőleges $t=0, \pm 1, \ldots$ mellett

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_t,$$

ahol $\mathbf{w}_t$ független Y_{t-1}^--tól, és így $\mathbf{x}_{t-1}$-től is. Így kapjuk az (1.3)—(1.4) állapotteres leírást, és vele egyidejűleg a fordított idejű

$$\mathbf{z}_t = \mathbf{A}^T\mathbf{z}_{t+1} + \boldsymbol{\omega}_t, \tag{4.13}$$

$$\mathbf{y}_t = \mathbf{B}^T\mathbf{z}_t \tag{4.14}$$

előállítást, ahol $\omega_t\,\mathbf{0}$ várható értékű, $\mathbf{M}$ kovarianciájú, Y_{t+1}^+-tól független normális eloszlású sorozat, A $\mathbf{z}_t$ folyamat $\mathbf{Q}$ kovarianciájára (1.6) helyett

$$\mathbf{Q} = \mathbf{A}^T\mathbf{Q}\mathbf{A} + \mathbf{M} \tag{4.15}$$

teljesül, és (3.12) alapján

$$E\mathbf{x}_t\mathbf{z}_t^T = \mathbf{P}\mathbf{Q}. \tag{4.16}$$

A p-dimenziós $\mathbf{x}_t$, $\mathbf{z}_t$ vektorokat (3.21) szerint az

$$E\mathbf{x}_t\mathbf{y}_{t-s}^T = \mathbf{A}^s\mathbf{B}, \quad E\mathbf{y}_{t+s}\mathbf{z}_t^T = \mathbf{C}\mathbf{A}^s, \quad s \geqq 0 \tag{4.17}$$

összefüggések definiálják magukat a $\boldsymbol{\Sigma}$, $\mathbf{M}$, illetve $\mathbf{P}$, $\mathbf{Q}$ mátrixokat azonban az $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ mátrixok ismeretében nem lehet véges eljárással meghatározni, szükség van az ún. *Kálmán szűrésre.*

5. Kálmán-szűrés

Tegyük fel, hogy a q-dimenziós stacionárius $\mathbf{y}_t$ *Gauss folyamatra*

$$E\mathbf{y}_t = \mathbf{0}, \quad \mathbf{C}_t = E\mathbf{y}_t\mathbf{y}_0^T = \mathbf{C}\mathbf{A}^t\mathbf{B}, \quad t \geqq 0$$

teljesül, ahol $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ rendre $p\times p$, $p\times q$ és $q\times p$ méretű mátrixok, és legyen ez az előállítás minimális dimenziójú. Jelöljük tetszőleges $t>1$ mellett $\tilde{\mathbf{y}}_t$-mal az

$$\tilde{\mathbf{y}}_t = E(\mathbf{y}_t|\mathbf{y}_1, \ldots, \mathbf{y}_{t-1})$$

feltételes várható értéket, és $\mathbf{D}_t$-vel a

$$\mathbf{D}_t = E(\mathbf{y}_t - \tilde{\mathbf{y}}_t)(\mathbf{y}_t - \tilde{\mathbf{y}}_t)^T$$

feltételes kovarianciát. Legyen $t=1$ mellett $\tilde{\mathbf{y}}_t=\mathbf{0}$, $\mathbf{D}_t=\mathbf{C}_0$.

5.1. Definíció. Azt mondjuk, hogy az $\mathbf{y}_t$ folyamat nem elfajuló, ha a $\mathbf{D}_t$ mátrixok minden $t\geqq 1$ mellett pozitív definitek.

5.2. Definíció. Az $\mathbf{y}_t$ folyamat $\mathbf{C}_t$ kovarianciájának $\mathbf{CA}^t\mathbf{B}$ alakú faktorizációjához tartozó háttérfolyamata az az $\mathbf{x}_t$ folyamat, melyre tetszőleges $t>0$ mellett teljesül, hogy $\mathbf{x}_t$ benne van az

$$\mathbf{y}_1, \ldots, \mathbf{y}_t$$

vektorok által kifeszített lineáris altérben, és amelyre $E\mathbf{x}_t=\mathbf{0}$,

$$E\mathbf{x}_t\mathbf{y}_{t-s}^T = \mathbf{A}^s\mathbf{B}, \quad 0 \leqq s \leqq t$$

teljesül.

Ha az $\mathbf{y}_t$ folyamat nem elfajuló, ennek a vektornak az egyértelmű létezését a következő lemma biztosítja.

5.1. LEMMA. Legyen $\mathbf{y}$ tetszőleges d dimenziós, $\mathbf{0}$ várható értékű normális eloszlású vektor, amelynek a kovariancia mátrixa reguláris, és legyen $\mathbf{D}$ tetszőleges $p\times d$ méretű mátrix. Akkor van olyan $p\times d$ méretű $\mathbf{F}$ mátrix, hogy az $\mathbf{x}=\mathbf{Fy}$ vektorra $E\mathbf{xy}^T=\mathbf{D}$ teljesül. Ha a p-dimenziós $\tilde{\mathbf{x}}$ vektorhoz található olyan $\tilde{\mathbf{F}}$ mátrix, hogy $\tilde{\mathbf{x}}=\tilde{\mathbf{F}}\mathbf{y}$, és $E\tilde{\mathbf{x}}\mathbf{y}^T=\mathbf{D}$, akkor $\tilde{\mathbf{x}}$ egyenlő $\mathbf{x}$-szel.

Bizonyítás. Ha $\mathbf{y}$ kovariancia mátrixa Σ, $E\mathbf{xy}^T=\mathbf{D}$ akkor teljesül, ha $\mathbf{D}=\mathbf{F}\Sigma$, vagyis $\mathbf{F}=\mathbf{D}\Sigma^{-1}$ egyértelműen meg van határozva.

5.2. LEMMA. Legyen $\mathbf{x}_t$ az $\mathbf{y}_t$ folyamat $\mathbf{C}_t$ kovarianciájának $\mathbf{CA}^t\mathbf{B}$ faktorizációjához tartozó háttérfolyamata. Akkor tetszőleges $s\geqq 0$ mellett

$$E(\mathbf{x}_{t+s}|\mathbf{y}_1, \ldots, \mathbf{y}_t) = \mathbf{A}^s\mathbf{x}_t,$$

$$E(\mathbf{y}_{t+s}|\mathbf{y}_1, \ldots, \mathbf{y}_t) = \mathbf{CA}^s\mathbf{x}_t.$$

Bizonyítás. Az $\mathbf{y}_{t+s}-\mathbf{CA}^s\mathbf{x}_t$ különbség független az $\mathbf{y}_1, \ldots, \mathbf{y}_t$ vektoroktól, hiszen $1\leqq k\leqq t$ mellett

$$E(\mathbf{y}_{t+s}-\mathbf{CA}^s\mathbf{x}_t)\mathbf{y}_k^T = \mathbf{CA}^{t+s-k}\mathbf{B}-\mathbf{CA}^s\mathbf{A}^{t-k}\mathbf{B} = \mathbf{O}.$$

Az első összefüggés hasonlóan látható be.

5.1. TÉTEL. Legyen $\mathbf{x}_0=\mathbf{0}$, $\mathbf{P}_0=\mathbf{0}$. A nem elfajuló $\mathbf{y}_t$ folyamat $\mathbf{C}_t$ kovarianciájának $\mathbf{CA}^t\mathbf{B}$ faktorizációjához tartozó $\mathbf{x}_t$ háttérfolyamata a következő rekurzióval

állítható elő.

$$\tilde{\mathbf{P}}_t = \mathbf{A}\mathbf{P}_{t-1}\mathbf{A}^T \tag{5.1}$$

$$\mathbf{B}_t = \mathbf{B} - \tilde{\mathbf{P}}_t\mathbf{C}^T \tag{5.2}$$

$$\mathbf{D}_t = \mathbf{C}\mathbf{B}_t \tag{5.3}$$

$$\mathbf{K}_t = \mathbf{B}_t\mathbf{D}_t^{-1/2} \tag{5.4}$$

$$\mathbf{P}_t = \tilde{\mathbf{P}}_t + \mathbf{K}_t\mathbf{K}_t^T \tag{5.5}$$

$$\mathbf{u}_t = \mathbf{D}_t^{-1/2}(\mathbf{y}_t - \mathbf{C}\mathbf{A}\mathbf{x}_{t-1}) \tag{5.6}$$

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{K}_t\mathbf{u}_t. \tag{5.7}$$

Bizonyítás. Az 5.2 lemma alapján az

$$\tilde{\mathbf{u}}_t = \mathbf{y}_t - \mathbf{C}\mathbf{A}\mathbf{x}_{t-1}$$

különbség független az $\mathbf{y}_1, \ldots, \mathbf{y}_{t-1}$ vektoroktól, így $\mathbf{x}_{t-1}$-től is. Jelöljük $\mathbf{D}_t$-vel ennek kovarianciamátrixát (ez a jelölés nyilván összhangban van az 5.1 definíció jelölésével). Akkor

$$\mathbf{C}_0 = \mathbf{C}\mathbf{B} = \mathbf{C}\tilde{\mathbf{P}}_t\mathbf{C}^T + \mathbf{D}_t$$

és

$$\mathbf{B} = E\mathbf{x}_t\mathbf{y}_t^T = \tilde{\mathbf{P}}_t\mathbf{C}^T + E(\mathbf{x}_t - \mathbf{A}\mathbf{x}_{t-1})\tilde{\mathbf{u}}_t.$$

Ez utóbbit $\mathbf{B}_t$-vel jelölve kapjuk az (5.2), (5.3) összefüggéseket. Mivel $\mathbf{u}_t = \mathbf{D}_t^{-1/2}\tilde{\mathbf{u}}_t$ kapjuk, hogy az (5.7)-ben szereplő $\mathbf{K}_t$ együttható alakja (5.4), amiből kapjuk (5.5)-öt.

Ha a rekurzió kezdőpontját áthelyezzük az origóból $(-k)$-ba $(k>0)$, nyilván lényegében minden változatlan marad, sőt, ha az új háttérfolyamatot $\mathbf{x}_t(k)$-val jelöljük, annak az $(\mathbf{y}_1, \ldots, \mathbf{y}_t)$ vektorokra vett feltételes várható értéke nyilván $\mathbf{x}_t$ lesz. Emiatt $\mathbf{x}_t(k)$ $\mathbf{P}_t(k)$ kovarianciája nő, és mivel $k \to \infty$ mellett az eredeti $\mathbf{x}_t$ reprezentánst kapjuk határértékként, $\mathbf{P}_t(k)$ konvergál a keresett $\mathbf{P}$ mátrixhoz. Így kapjuk az

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{K}\mathbf{u}_t, \quad \mathbf{y}_t = \mathbf{C}\mathbf{x}_t \tag{5.8}$$

állapotteres leírást ($\mathbf{x}_t$ itt már újra a $\lim_{k\to\infty} \mathbf{x}_t(k)$ határértéket jelöli), és az egész eljárást fordított idő mellett megismételve kapjuk a

$$\mathbf{z}_t = \mathbf{A}^T\mathbf{z}_{t+1} + \mathbf{L}^T\mathbf{v}_t, \quad \mathbf{y}_t = \mathbf{B}^T\mathbf{z}_t \tag{5.9}$$

egyenleteket. Ezekből

$$E\mathbf{x}_t\mathbf{u}_t^T = \mathbf{K}, \quad E\mathbf{z}_t\mathbf{v}_t^T = \mathbf{L}^T$$

következik, amiből viszont (3.18) alapján azt kapjuk, hogy

$$E\mathbf{v}_{t+s}\mathbf{u}_t^T = \mathbf{L}\mathbf{A}^s\mathbf{K}, \quad s \geqq 0. \tag{5.10}$$

6. Forgatások állapotteres leírása

Tekintsünk egy $\{\mathbf{y}_t, t=0, \pm 1, 2, \ldots\}$ normális eloszlású q-dimenziós valószínűségi változókból álló stacionárius, 0 várható értékű valószínűségi változó sorozatot, melyre fennáll, hogy

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{e}_t \tag{6.1}$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t,$$

ahol $\{\mathbf{e}_t, t=0, \pm 1, \ldots\}$ független $N_q(\mathbf{0}, \mathbf{I})$ eloszlású sorozat, az $\mathbf{A}$ mátrixról feltesszük, hogy nem esik saját értéke az egységkör kerületére, és az $\{\mathbf{x}_t, t=0, \pm 1, \ldots\}$ p-dimenziós stacionárius sorozat kovarianciamátrixa nem elfajuló. (Vegyük észre, hogy nemcsak $\mathbf{A}$-ról teszünk fel most kevesebbet, mint eddig, hanem $\mathbf{e}_t$ és $\mathbf{x}_t$ függetlenségét sem tettük fel.)

Mivel az $\mathbf{A}$ mátrix egységkörön belül és egységkörön kívüli sajátértékeihez tartozó invariáns alterek metszete nyilván a $\mathbf{0}$ vektor és ezek a terek együttesen kifeszítik R^p-t, így alkalmas bázisban

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix} \tag{6.2}$$

alakú, ahol $\mathbf{A}_1$ sajátértékei az egységkörön belülre, $\mathbf{A}_2$-éi pedig kívülre esnek, így $\mathbf{A}_2^{-1}$ létezik.

Ha eredetileg tehát a (6.1) leírásban $\mathbf{A}$ nem (6.2) alakú lenne, akkor a megfelelő bázistranszformációt elvégezve — azaz az $\{\mathbf{x}_t, t=0, \pm 1, \ldots\}$ sorozat elemeit a $\{\mathbf{T}\mathbf{x}_t, t=0, \pm 1, \ldots\}$ sorozattal helyettesítve elérhetjük, hogy a $\mathbf{T}\mathbf{A}\mathbf{T}^{-1}$ mátrixra már teljesüljön (6.2).

Tegyük fel tehát eleve, hogy $\mathbf{A}$ (6.2) alakú. Az $\mathbf{x}_t, \mathbf{B}, \mathbf{C}$ mennyiségeket ennek megfelelően partícionálva az

$$\begin{bmatrix} \mathbf{x}_{t+1}^1 \\ \mathbf{x}_{t+1}^2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_t^1 \\ \mathbf{x}_t^2 \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} \mathbf{e}_t \tag{6.3}$$

$$\mathbf{y}_t = [\mathbf{C}_1, \mathbf{C}_2] \begin{bmatrix} \mathbf{x}_t^1 \\ \mathbf{x}_t^2 \end{bmatrix}$$

egyenletekhez jutunk. Kiírva blokkonként:

$$\mathbf{x}_{t+1}^1 = \mathbf{A}_1 \mathbf{x}_t^1 + \mathbf{B}_1 \mathbf{e}_t, \tag{6.4}$$

$$\mathbf{x}_{t+1}^2 = \mathbf{A}_2 \mathbf{x}_t^2 + \mathbf{B}_2 \mathbf{e}_t, \tag{6.5}$$

$$\mathbf{y}_t = \mathbf{C}_1 \mathbf{x}_t^1 + \mathbf{C}_2 \mathbf{x}_t^2. \tag{6.6}$$

Vegyük szemügyre először a (6.4) egyenletet. A jobb oldalába $\mathbf{x}_t^1$, majd $\mathbf{x}_{t-1}^1$, helyébe rendre behelyettesítve az egyenletből adódó értékeket a következő összefüggést kapjuk:

$$\mathbf{x}_{t+1}^1 = \mathbf{A}_1^{k+1} \mathbf{x}_{t-k}^1 + \sum_{j=0}^{k} \mathbf{A}_1^j \mathbf{B}_1 \mathbf{e}_{t-j}.$$

Mivel $\mathbf{A}_1$ spektrálsugara — $\varrho(\mathbf{A}_1)$ — egynél kisebb, így $\mathbf{x}_t^1$ stacionaritása miatt az $\mathbf{A}_1^k \mathbf{x}_{t-k}^1$ vektorváltozó sorozat koordinátái L_2-ben nullához tartanak, így

$$\mathbf{x}_{t+1}^1 = \sum_{j=0}^{\infty} \mathbf{A}_1^j \mathbf{B}_1 \mathbf{e}_{t-j}, \quad t = 0, \pm 1, \ldots. \tag{6.7}$$

Vegyük észre, hogy (6.7) közvetlen következménye, hogy tetszőleges t esetén $\mathbf{x}_t^1$ független az $\mathbf{e}_t, \mathbf{e}_{t+1}, \ldots$ valószínűségi változóktól.
Érdemes megjegyezni, hogy ez utóbbi állítás megfordítása is igaz. Ezt tartalmazza a

6.1. LEMMA. Legyen $\{\mathbf{e}_t, t=, \pm 1, \ldots\}$ független, $N_q(0, 1)$ eloszlású valószínűségi változókból álló sorozat, $\{\mathbf{x}_t, t=0, \pm 1, \ldots\}$ stacionárius p-dimenziós sorozat, melynek kovariancia mátrixa nem elfajuló, és tegyük fel, hogy teljesül az

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{e}_t$$

összefüggés, ahol $\mathbf{A}$-nak nincs egységnyi abszolút értékű sajátértéke. Ekkor a következő két állítás ekvivalens:

(i) $\varrho(\mathbf{A}) < 1$.

(ii) $\mathbf{x}_t$ független az $\mathbf{e}_t, \mathbf{e}_{t+1} \ldots$ változóktól.

Bizonyítás. Az (i)⇒(ii) állítást már igazoltuk. A fordított irányú lépés kedvéért írjuk tovább rekurzívan az $\mathbf{x}_t$ sorozatra vonatkozó egyenletet.

$$\mathbf{x}_{t+1} = \mathbf{A}^{k+1}\mathbf{x}_{t-k} + \sum_{j=0}^{k} \mathbf{A}^j \mathbf{B}\mathbf{e}_{t-j}. \tag{6.8}$$

Legyen

$$\mathbf{P} = E\mathbf{x}_t \mathbf{x}_t^T.$$

Ekkor (6.8) és (ii) alapján

$$\mathbf{P} = \mathbf{A}^{k+1}\mathbf{P}(\mathbf{A}^T)^{k+1} + \sum_{j=0}^{k} \mathbf{A}^j \mathbf{B}\mathbf{B}^T(\mathbf{A}^T)^j.$$

Tegyük fel, hogy $\varrho(\mathbf{A}) > 1$. Legyen λ sajátértéke $\mathbf{A}^T$-nek, ξ egy hozzá tartozó saját vektor (esetleg komplex koordinátákkal), és legyen $|\lambda| > 1$. Ekkor

$$\xi^* \mathbf{P} \xi = |\lambda|^{2(k+1)} \xi^* \mathbf{P} \xi + \sum_{j=0}^{k} |\lambda|^{2j} \xi^* \mathbf{B}\mathbf{B}^T \xi.$$

A jobb oldalon minden tag nem negatív, sőt $\xi^*\mathbf{P}\xi > 0$, így $|\lambda| > 1$ esetén a

$$\lim_{k \to \infty} |\lambda|^{2(k+1)} = \infty$$

összefüggés ellentmondáshoz vezet. A $\varrho(\mathbf{A}) = 1$ eset kizárása bonyolultabb, azt nem részletezzük.

Térjünk vissza a (6.4)—(6.6) egyenlethez (6.5)-ben, az $\mathbf{A}_2$ mátrixra nem teljesül, hogy $\varrho(\mathbf{A}_2) < 1$, viszont a (6.2) konstrukció alapján $\varrho(\mathbf{A}_2^{-1}) < 1$. Rendezzük át (6.5)-t:

$$\mathbf{x}_t^2 = \mathbf{A}_2^{-1}\mathbf{x}_{t+1}^2 - \mathbf{A}_2^{-1}\mathbf{B}_2\mathbf{e}_t.$$

Ekkor nyilván

$$\mathbf{x}_t^2 = -\sum_{k=0}^{\infty} \mathbf{A}_2^{-(k+1)} \mathbf{B}_2 \mathbf{e}_{t+k}, \tag{6.9}$$

és $\mathbf{x}_t^2$ független az $\mathbf{e}_{t-1}, \mathbf{e}_{t-2}, \ldots$ változóktól. Ezek alapján $\mathbf{y}_t$-re a következő állítás adódik.

$$\mathbf{y}_t = \sum_{k=1}^{\infty} \mathbf{C}_1 \mathbf{A}_1^{k-1} \mathbf{B}_1 \mathbf{e}_{t-k} - \sum_{k=0}^{\infty} \mathbf{C}_2 \mathbf{A}_2^{-(k+1)} \mathbf{B}_2 \mathbf{e}_{t+k}. \tag{6.10}$$

Ebben a fejezetben a következő tulajdonságokat próbáljuk jellemezni.

6.1. Tulajdonság. A (6.10) képlettel definiált $\{\mathbf{y}_t, t=0, \pm 1, \ldots\}$ stacionárius folyamat független, $N_q(\mathbf{0}, \sigma^2\mathbf{I})$ eloszlású valószínűségi változókból álló sorozat.

6.2. Tulajdonság. Létezik olyan Γ $q \times q$-as mátrix, hogy az $\bar{\mathbf{y}}_t = \mathbf{y}_t + \Gamma \mathbf{e}_t$ sorozat független $N_q(\mathbf{0}, \sigma^2\mathbf{I})$ eloszlású változókból áll.

Tegyük fel, hogy fennáll a (6.2) tulajdonság:

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{e}_t, \tag{6.11}$$

$$\bar{\mathbf{y}}_t = \mathbf{C}\mathbf{x}_t + \Gamma \mathbf{e}_t.$$

Legyen $\mathbf{E}$ az

$$\eta = \sum_{i=1}^{k} \mathbf{a}_i^T \mathbf{e}_{t-i}$$

alakú valószínűségi változók által generált zárt $L_2(P)$-beli altér. (Itt $L_2(P)$ a véges második momentumú valószínűségi változók terét jelöli), és $\mathbf{Y}$ jelölje az

$$\eta = \sum_{i=1}^{k} \mathbf{a}_i^T \mathbf{y}_{t-i}$$

alakú valószínűségi változók által kifeszített zárt alteret. ($\mathbf{E}$ és $\mathbf{Y}$ definíciójában k és t is szabadon változik.) Nyilván

$$\mathbf{Y} \subset \mathbf{E}.$$

(Ezek *Hilbert-terek*. A $\langle , \rangle$ skalárszorzat a valószínűségi változók szorzatának a várható értéke.) Mivel a 6.2 tulajdonság alapján $\frac{1}{\sigma} \bar{\mathbf{y}}_t$ is független, $N_q(\mathbf{0}, \mathbf{I})$ sorozat, ezért az a

$$\tau\colon \mathbf{E} \to \mathbf{E}$$

leképezés, melyre tetszőleges $\mathbf{a} \in R^q$ esetén

$$\tau \mathbf{a}^T \mathbf{e}_t = \frac{1}{\sigma} \mathbf{a}^T \mathbf{y}_t, \quad t = 0, \pm 1, \ldots,$$

izometria, pontosabban forgatás.
(A jelölések egyszerűsítése céljából röviden azt írjuk majd, hogy

$$\tau \mathbf{e}_t = \frac{1}{\sigma} \bar{\mathbf{y}}_t. \tag{6.12}$$

Ugyanezzel a konvencióval fogunk élni tetszőleges $E \to E$ operátor esetén.) Mivel τ forgatás, így a τ^* adjungált operátor egyben inverze is. Vajon hogyan lehet $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, Γ segítségével felírni a τ^* leképezést?

6.2. LEMMA. Legyen $S: E \to E$ az időeltolás operátora, tehát $Se_t = e_{t+1}$. Ekkor tetszőleges $\xi \in E$ esetén

$$\tau^*\xi = \sum_{k=1}^{\infty} \mathbf{B}_1^T (\mathbf{A}_1^T)^{k-1} \mathbf{C}_1^T S^k \xi - \sum_{k=0}^{\infty} \mathbf{B}_2^T (\mathbf{A}_2^T)^{-(k+1)} \mathbf{C}_2^T S^{-k} \xi + \Gamma^T \xi. \tag{6.13}$$

Bizonyítás. τ^* linearitása miatt elegendő (6.13)-at egy alkalmas generátorrendszeren ellenőrizni. Válasszuk erre a célra az $\{e_t, t=0, \pm 1, \pm 2, \ldots\}$ koordinátáiból álló ortonormált bázist. Legyen $\varphi, \psi \in R^q$. Számoljuk ki a

$$\langle \tau^* \psi^T e_s, \varphi^T e_t \rangle$$

skaláris szorzat értékét.

$$\langle \tau^* \psi^T e_s, \varphi^T e_t \rangle = \langle \psi^T e_s, \tau \varphi^T e_t \rangle =$$

$$= E\big((\psi^T e_s)(\tau \varphi^T e_t)\big) = E(\psi^T e_s \bar{y}_t^T \varphi) \frac{1}{\sigma} = \begin{cases} \mathbf{B}_1^T \mathbf{A}_1^{T^{(t-s-1)}} \mathbf{C}_1^T \dfrac{1}{\sigma}, & s < t \\ (-\mathbf{B}_2^T \mathbf{A}_2^{T^{-1}} \mathbf{C}_2^T + \Gamma^T) \dfrac{1}{\sigma}, & s = t \\ -\mathbf{B}_2^T \mathbf{A}_2^{T^{-(t-s+1)}} \mathbf{C}_2^T \dfrac{1}{\sigma}, & s > t. \end{cases}$$

Tehát

$$\tau^* e_s = \Big(\sum_{k=1}^{\infty} \mathbf{B}_1^T \mathbf{A}_1^{T^{k-1}} \mathbf{C}_1^T e_{s+k} - \sum_{k=0}^{\infty} \mathbf{B}_2^T \mathbf{A}_2^{T^{-(k+1)}} \mathbf{C}_2^T e_{s-k} + \Gamma^T e_s \Big) \frac{1}{\sigma}.$$

Mivel

$$e_{s+k} = S^k e_s,$$

így a 6.2. lemmát igazoltuk.

Alkalmazzuk ezt a $\xi = \bar{y}_t$ választással. Ekkor

$$S^k \bar{y}_t = \bar{y}_{t+k},$$

így

$$\tau^* \bar{y}_t = \Big(\sum_{k=1}^{\infty} \mathbf{B}_1^T (\mathbf{A}_1^T)^{(k-1)} \mathbf{C}_1^T \bar{y}_{t+k} - \sum_{k=0}^{\infty} \mathbf{B}_2^T (\mathbf{A}_2^T)^{-(k+1)} \mathbf{C}_2^T \bar{y}_{t-k} + \Gamma^T \bar{y}_t \Big) \frac{1}{\sigma}. \tag{6.14}$$

A (6.2) tulajdonság alapján $\tau^* \tau$ az identitás operátor, így

$$\tau^* \frac{1}{\sigma} \bar{y}_t = e_t.$$

A (6.14) alapján $\sigma^2 \tau^* \bar{y}_t$ előállítása hasonlít a (6.10) előállításra. (6.10) viszont szoros kapcsolatban van a (6.1) egyenletekkel. Keressük meg ennek megfelelőit.
Legyen

$$\zeta_t^1 = \sum_{k=1}^{\infty} (\mathbf{A}_1^T)^{k-1} \mathbf{C}_1^T \bar{y}_{t+k}, \tag{6.15}$$

$$\zeta_t^2 = -\sum_{k=0}^{\infty} (\mathbf{A}_2^T)^{-(k+1)} \mathbf{C}_2^T \bar{y}_{t-k},$$

és legyen

$$\zeta_t = \begin{bmatrix} \zeta_t^1 \\ \zeta_t^2 \end{bmatrix}.$$

Ekkor

(6.16)
$$\zeta_{t-1}^1 = \mathbf{A}_1^T \zeta_t^1 + \mathbf{C}_1^T \bar{\mathbf{y}}_t,$$

$$\zeta_t^2 = \mathbf{A}_2^T \zeta_{t-1}^2 - (\mathbf{A}_2^T)^{-1} \mathbf{C}_2^T \bar{\mathbf{y}}_t.$$

Tömörebben

(6.17)
$$\zeta_{t-1} = \mathbf{A}^T \zeta_t + \mathbf{C}^T \bar{\mathbf{y}}_t,$$

illetve

$$\sigma^2 \mathbf{e}_t = \mathbf{B}_1^T \zeta_t^1 + \mathbf{B}_2^T + \mathbf{\Gamma}^T \bar{\mathbf{y}}_t,$$

azaz

(6.18)
$$\sigma^2 \mathbf{e}_t = \mathbf{B}^T \zeta_t + \mathbf{\Gamma}^T \bar{\mathbf{y}}_t.$$

A (6.18) egyenlet jobb oldalán szereplő mennyiségek felírhatók $\{\mathbf{e}_s, s=0, \pm 1, \pm 2, \ldots\}$ segítségével. Mivel az összeg értéke $\sigma^2 \mathbf{e}_t$, így $s \neq t$ esetín $\mathbf{e}_s$ együtthatója csak 0 lehet. Próbáljuk meg az

$$E\big((\mathbf{B}^T \zeta_t + \mathbf{\Gamma}^T \bar{\mathbf{y}}_t)\mathbf{e}_s^T\big) = 0 \quad s \neq t$$

feltételt valahogy tömörebben megfogalmazni. Ne számoljuk ki tetszőleges $\mathbf{e}_s$ együtthatóját (6.18)-ban, hanem bontsuk fel a jobb oldalt három részre: az $\mathbf{e}_t$ konstanssszorosa, az $\mathbf{e}_{t-1}, \mathbf{e}_{t-2}, \ldots$ változókból képzett mennyiség, végül az $\mathbf{e}_{t+1}, \mathbf{e}_{t-2}, \ldots$ változók függvénye.

Az egyszerűség kedvéért tegyük fel egyelőre, hogy $\varrho(\mathbf{A}) < 1$. Ekkor

$$\mathbf{x}_t = \mathbf{x}_t^1,$$

$$\zeta_t = \zeta_t^1.$$

Ezért elhagyjuk az 1, 2 indexeket. Alakítsuk át (6.18) jobb oldalát (6.11) és (6.17) segítségével.

$$\mathbf{B}^T \zeta_t + \mathbf{\Gamma}^T \bar{\mathbf{y}}_t = \mathbf{B}^T \zeta_t + \mathbf{\Gamma}^T \mathbf{C} \mathbf{x}_t + \mathbf{\Gamma}^T \mathbf{\Gamma} \mathbf{e}_t.$$

$\mathbf{x}_t$, $\mathbf{e}_t$ már a kívánt mennyiségek. ζ_t értékét kellene kettévágni.
Számoljuk ki

$$E(\zeta_t \mathbf{e}_s^T)$$

$s \leqq t$ értékét.

$$E(\zeta_t \mathbf{e}_s^T) = E\Big(\sum_{k=1}^{\infty} (\mathbf{A}^T)^{k-1} \mathbf{C}^T \bar{\mathbf{y}}_{t+k} \mathbf{e}_s^T\Big) =$$

$$= \sum_{k=1}^{\infty} (\mathbf{A}^T)^{k-1} \mathbf{C}^T \mathbf{C} \mathbf{A}^{t+k-s-1} \mathbf{B} = \Big(\sum_{k=1}^{\infty} (\mathbf{A}^T)^{k-1} \mathbf{C}^T \mathbf{C} \mathbf{A}^{k-1}\Big) \mathbf{A}^{t-s} \mathbf{B}.$$

Legyen

$$\mathbf{Q} = \sum_{k=1}^{\infty} \mathbf{A}^{T(k-1)} \mathbf{C}^T \mathbf{C} \mathbf{A}^{k-1}.$$

Ekkor egyfelől

$$E(\zeta_t \mathbf{e}_s^T) = \mathbf{Q}\mathbf{A}^{t-s}\mathbf{B}, \tag{6.19}$$

másfelől $\mathbf{Q}$ a

$$\mathbf{Q} = \mathbf{A}^T\mathbf{Q}\mathbf{A} + \mathbf{C}^T\mathbf{C} \tag{6.20}$$

egyenlet egyértelmű megoldása (vö.: (4.15), az egyértelműség és az egzisztencia most a $\varrho(\mathbf{A}) < 1$ feltevés következménye). Vegyük észre, hogy (6.17) alapján, ha igaz a (6.2) tulajdonság, akkor

$$\sigma^2\mathbf{Q} = E(\zeta_t \zeta_t^T),$$

és

$$E(\mathbf{x}_{t+1}\mathbf{e}_s^T) = \mathbf{A}^{t-s}\mathbf{B}.$$

(6.19)-cel ezt összevetve kapjuk, hogy

$$\zeta_t - \mathbf{Q}\mathbf{x}_{t+1}$$

független az $\mathbf{e}_t, \mathbf{e}_{t-1}, \ldots$ valószínűségi változóktól. A (6.18) jobb oldalának keresett alakja a következő lesz

$$\mathbf{B}^T\zeta_t + \mathbf{\Gamma}^T\bar{\mathbf{y}}_t = \mathbf{B}^T(\zeta_t - \mathbf{Q}\mathbf{x}_{t+1}) + \mathbf{B}^T\mathbf{Q}\mathbf{x}_{t+1} + \mathbf{\Gamma}^T\mathbf{C}\mathbf{x}_t + \mathbf{\Gamma}^T\mathbf{\Gamma}\mathbf{e}_t = \tag{6.21}$$

$$= \mathbf{B}^T(\zeta_t - \mathbf{Q}\mathbf{x}_{t+1}) + (\mathbf{B}^T\mathbf{Q}\mathbf{A} + \mathbf{\Gamma}^T\mathbf{C})\mathbf{x}_t + (\mathbf{B}^T\mathbf{Q}\mathbf{B} + \mathbf{\Gamma}^T\mathbf{\Gamma})\mathbf{e}_t.$$

Nem meglepő ezek után a következő lemma.

6.3. LEMMA. Legyen

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{e}_t,$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t.$$

Jelölje $\mathbf{P}$ az $\mathbf{x}_t$ kovarianciamátrixát. Tegyük fel, hogy $\varrho(\mathbf{A}) < 1$, és $\mathbf{P}^{-1}$ létezik. Ha teljesül a 6.2 tulajdonság, akkor $\mathbf{Q}\mathbf{P}\mathbf{Q} = \sigma^2\mathbf{Q}$.

Bizonyítás. (6.18)-at és a (6.21) átalakítást felhasználva, mivel

$$\zeta_t - \mathbf{Q}\mathbf{x}_{t+1}, \ \mathbf{x}_t, \ \mathbf{e}_t$$

függetlenek, így

$$\mathbf{B}^T\mathbf{Q}\mathbf{B} + \mathbf{\Gamma}^T\mathbf{\Gamma} = \sigma^2\mathbf{I}, \tag{6.22}$$

$$\mathbf{B}^T\mathbf{Q}\mathbf{A} + \mathbf{\Gamma}^T\mathbf{C} = \mathbf{0}. \tag{6.23}$$

Megmutatjuk, hogy

$$\zeta_t = \mathbf{Q}\mathbf{x}_{t+1}. \tag{6.24}$$

(6.17) és (6.20) alapján

$$\zeta_t - \mathbf{Q}\mathbf{x}_{t+1} = \mathbf{A}^T\zeta_{t+1} + \mathbf{C}^T\bar{\mathbf{y}}_{t+1} - \mathbf{A}^T\mathbf{Q}\mathbf{A}\mathbf{x}_{t+1} - \mathbf{C}^T\mathbf{C}\mathbf{x}_{t+1},$$

(6.11) szerint továbbalakítva

$$\zeta_t - \mathbf{Q}\mathbf{x}_{t+1} = \mathbf{A}^T(\zeta_{t+1} - \mathbf{Q}\mathbf{x}_{t+2}) + (\mathbf{C}^T\mathbf{\Gamma} + \mathbf{A}^T\mathbf{Q}\mathbf{B})\mathbf{e}_{t+1}.$$

Rekurzíven kifejtve a

$$\zeta_t - \mathbf{Q}\mathbf{x}_{t+1} = \sum_{k=1}^{\infty} (\mathbf{A}^T)^{(k-1)}(\mathbf{C}^T\Gamma + \mathbf{A}^T\mathbf{Q}\mathbf{B})\mathbf{e}_{t+k}$$

előállítást kapjuk. Használjuk ki (6.23)-at, (6.24) adódik . Tehát

$$\zeta_t = \mathbf{Q}\mathbf{x}_{t+1}.$$

Mindkét oldal kovarianciáját véve:

$$\sigma^2\mathbf{Q} = \mathbf{Q}\mathbf{P}\mathbf{Q} \tag{6.25}$$

és ez az, amit bizonyítani akartunk.

Megjegyzés. Ha feltesszük, hogy $\mathbf{Q}^{-1}$ létezik, akkor a

$$\mathbf{P}\mathbf{Q} = \sigma^2\mathbf{I}$$

összefüggés is teljesül.

Megjegyzés. A (6.1) tulajdonság létezésének egy kritériumát kaphatjuk a (6.22), (6.23) és (6.25) egyenletekből. A $\Gamma = \mathbf{0}$ értéknek gyöknek kell lennie. Azaz

$$\mathbf{B}^T\mathbf{Q}\mathbf{B} = \sigma^2\mathbf{I}, \tag{6.26}$$

$$\mathbf{B}^T\mathbf{Q}\mathbf{A} = \mathbf{0}, \tag{6.27}$$

$$\mathbf{P}\mathbf{Q} = \sigma^2\mathbf{I} \tag{6.28}$$

a szükséges feltételek $\varrho(\mathbf{A}) < 1$ esetén, ha tudjuk, hogy $\mathbf{P}^{-1}$ és $\mathbf{Q}^{-1}$ létezik.

6.1. KÖVETKEZMÉNY. Legyen

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{e}_t$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t$$

Tegyük fel, hogy $\varrho(\mathbf{A}) < 1$; $\mathbf{P}^{-1}$ és $\mathbf{Q}^{-1}$ létezik. Ekkor, ha teljesül a 6.2 tulajdonság, akkor létezik olyan $\mathbf{T}\ p \times p$-s, invertálható mátrix, hogy

$$\mathbf{T}\mathbf{A}\mathbf{T}^{-1} = \begin{bmatrix} \mathbf{0} \\ \tilde{\mathbf{A}} \end{bmatrix} \begin{matrix} (q \\ ((p-q) \end{matrix} = \tilde{\mathbf{A}} \tag{6.29}$$

$$\mathbf{T}\mathbf{B} = \begin{bmatrix} \sigma & & \\ & \ddots & \\ & & \sigma \\ & \mathbf{0} & \end{bmatrix} \begin{matrix} \\ \\ \\ (q \end{matrix} = \tilde{\mathbf{B}} \tag{6.30}$$

$$\mathbf{C}\mathbf{T}^{-1} = \tilde{\mathbf{C}} \tag{6.31}$$

és a $\begin{bmatrix} \tilde{\mathbf{C}} \\ \tilde{\mathbf{A}}_0 \end{bmatrix}$ mátrix forgatás, azaz

$$\begin{bmatrix} \tilde{\mathbf{C}} \\ \tilde{\mathbf{A}}_0 \end{bmatrix} [\tilde{\mathbf{C}}^T\tilde{\mathbf{A}}^T] = \mathbf{I}.$$

Másképpen az

$$\tilde{\mathbf{x}}_t = \mathbf{T}\mathbf{x}_t \tag{6.32}$$

állapotvektort bevezetve

$$\tilde{\mathbf{x}}_{t+1} = \tilde{\mathbf{A}}\tilde{\mathbf{x}}_t + \tilde{\mathbf{B}}\mathbf{e}_t$$

$$\mathbf{y}_t = \tilde{\mathbf{C}}\tilde{\mathbf{x}}_t.$$

Bizonyítás. Vegyük észre, hogy a ζ_t-re vonatkozó egyenleteket kissé átalakítva — az $\eta_t = \frac{1}{\sigma}\zeta_t$ változót bevezetve — az

$$\eta_{t-1} = \mathbf{A}^T\eta_t + \mathbf{C}^T\frac{1}{\sigma}\mathbf{y}_t$$

$$\sigma\mathbf{e}_t = \mathbf{B}^T\eta_t$$

egyenletekre jutunk, melyek az $\left\{\frac{1}{\sigma}\mathbf{y}_t,\ t=0,\ \pm 1,\ \pm 2,\ \ldots\right\}$ $\mathbf{I}$ kovariankiájú független sorozathoz a $\{\sigma\mathbf{e}_t,\ t=0,\ \pm 1,\ \pm 2,\ \ldots\}$ $\sigma^2\mathbf{I}$ kovarianciájú független sorozatot rendelik hozzá, így igaz η_t-re a 6.3 lemma és az azt követő megjegyzés.
Speciálisan

(6.33) $$\mathbf{CPC}^T = \sigma^2\mathbf{I}$$

(6.34) $$\mathbf{CPA}^T = \mathbf{0}.$$

Tudjuk már, hogy

$$\mathbf{B}^T\mathbf{QB} = \sigma^2\mathbf{I}$$

$$\mathbf{B}^T\mathbf{QA} = \mathbf{0}$$

$$\mathbf{PQ} = \sigma^2\mathbf{I}.$$

Legyen $\mathbf{T}_1$ tetszőleges olyan $p\times p$-s mátrix, melyre

$$\mathbf{T}_1^T\mathbf{T}_1 = \mathbf{Q}.$$

A $\mathbf{PQ}=\sigma^2\mathbf{I}$ egyenletet balról $\mathbf{T}_1$-gyel, jobbról $\mathbf{T}_1^{-1}$-gyel szorozva, a

$$\mathbf{T}_1\mathbf{P}\mathbf{T}_1^T = \sigma^2\mathbf{I}$$

összefüggést kapjuk. Mivel (6.26) alapján

$$(\mathbf{T}_1\mathbf{B})^T(\mathbf{T}_1\mathbf{B}) = \sigma^2\mathbf{I},$$

ezért létezik olyan $\mathbf{T}_2$ $p\times p$-s mátrix, melyre $\mathbf{T}_2^T\mathbf{T}_2=\mathbf{I}$, és

$$\mathbf{T}_2\mathbf{T}_1\mathbf{B} = \begin{bmatrix} \sigma & & \\ & \ddots & \\ & & \sigma \\ & \mathbf{0} & \end{bmatrix}.$$

(6.27) szerint

$$(\mathbf{T}_2\mathbf{T}_1\mathbf{B})^T(\mathbf{T}_2\mathbf{T}_1\mathbf{A}) = \mathbf{0},$$

ill.

$$(\mathbf{T}_2\mathbf{T}_1\mathbf{B})^T(\mathbf{T}_2\mathbf{T}_1\mathbf{A}\mathbf{T}_1^{-1}\mathbf{T}_2) = \mathbf{0},$$

így

$$\mathbf{T}_2\mathbf{T}_1\mathbf{A}\mathbf{T}_1^{-1}\mathbf{T}_2^{-1} = \begin{bmatrix} \mathbf{0} \\ \mathbf{A}_0 \end{bmatrix} \begin{matrix} (q \\ ((p-q)\cdot \end{matrix}$$

Alakítsuk át a

$$\mathbf{P} = \mathbf{A}\mathbf{P}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T$$

egyenlet úgy, hogy $\mathbf{T}_2\mathbf{T}_1\mathbf{A}\mathbf{T}_1^{-1}\mathbf{T}_2^{-1}$ és $\mathbf{T}_2\mathbf{T}_1\mathbf{B}$ szerepeljenek benne:

$$\mathbf{T}_2\mathbf{T}_1\mathbf{P}\mathbf{T}_1^T\mathbf{T}_2^T = (\mathbf{T}_2\mathbf{T}_1\mathbf{A}\mathbf{T}_1^{-1}\mathbf{T}_2^{-1})(\mathbf{T}_2\mathbf{T}_1\mathbf{P}\mathbf{T}_1^T\mathbf{T}_2^T)((\mathbf{T}_2^T)^{-1}(\mathbf{T}_1^T)^{-1}\mathbf{A}^T\mathbf{T}_1^T\mathbf{T}_2^T) +$$

$$+(\mathbf{T}_2\mathbf{T}_1\mathbf{B})(\mathbf{T}_2\mathbf{T}_1\mathbf{B})^T.$$

Azaz

$$\begin{bmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{bmatrix} = \begin{bmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & \sigma^2\tilde{\mathbf{A}}_0\tilde{\mathbf{A}}_0^T \end{bmatrix} + \begin{bmatrix} \sigma^2 & & & \\ & \ddots & & \\ & & \sigma^2 & \\ & & & \ddots \\ & & & & 0 \end{bmatrix}.$$

Tehát

$$\tilde{\mathbf{A}}_0\tilde{\mathbf{A}}_0^T = \mathbf{I}.$$

A

$$\mathbf{Q} = \mathbf{A}^T\mathbf{Q}\mathbf{A} + \mathbf{C}^T\mathbf{C}$$

egyenlet megfelelő transzformációjával, a $\tilde{\mathbf{C}} = \mathbf{C}\mathbf{T}_1^{-1}\mathbf{T}_2^{-1}$ jelöléssel, az

$$\tilde{\mathbf{A}}_0^T\tilde{\mathbf{A}}_0 + \tilde{\mathbf{C}}^T\tilde{\mathbf{C}} = \mathbf{I}$$

összefüggést kapjuk tehát a $\begin{bmatrix} \tilde{\mathbf{C}} \\ \tilde{\mathbf{A}}_0 \end{bmatrix}$ mátrix forgatás. Látjuk tehát, hogy a

$$\mathbf{T} = \mathbf{T}_2\mathbf{T}_1$$

választás jó.

Megjegyzés. A 6.2. következmény állítását szemléletesen így lehet megfogalmazni: Az $\mathbf{x}_t$ értékéből $\mathbf{y}_t$ és $\mathbf{x}_{t+1}$ úgy olvasható ki, hogy először elvégzünk egy forgatást $\mathbf{x}_t$-n; az első q koordináta adja $\mathbf{y}_t$ értékét, $\mathbf{x}_{t+1}$ pedig úgy adódik, hogy az első q koordináta helyébe a $\sigma\mathbf{e}_t$ mennyiséget írjuk.

Megjegyzés. A 6.1. következmény bizonyításának elején alkalmazott gondolatmenet a $\mathbf{\Gamma} \neq \mathbf{0}$ esetre a

$$\mathbf{C}\mathbf{P}\mathbf{C}^T + \mathbf{\Gamma}\mathbf{\Gamma}^T = \sigma^2\mathbf{I} \tag{6.35}$$

$$\mathbf{C}\mathbf{P}\mathbf{A}^T + \mathbf{\Gamma}\mathbf{B}^T = \mathbf{0} \tag{6.36}$$

egyenleteket eredményezi.
(6.36) és (6.11) alapján

$$E(\bar{\mathbf{y}}_t\mathbf{x}_{t+1}^T) = E\big((\mathbf{C}\mathbf{x}_t + \mathbf{\Gamma}\mathbf{e}_t)(\mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{e}_t)^T\big) = \mathbf{C}\mathbf{P}\mathbf{A}^T + \mathbf{\Gamma}\mathbf{B}^T = \mathbf{0}.$$

Azaz $\mathbf{y}_t$ független az $\mathbf{x}_{t+1}$ valószínűségi változótól.

Térjünk most vissza az általános esetre. A 6.3. lemma a $\varrho(\mathbf{A})<1$ esetben a 6.2. tulajdonságot a $\mathbf{P}$, ill. $\mathbf{Q}$ mátrix segítségével jellemzi. Ekkor $\mathbf{P}$ egyfelől az $\mathbf{x}_t$ kovariancia mátrixa, másfelől a

$$\mathbf{P} = \mathbf{APA}^T + \mathbf{BB}^T$$

egyenlet egyetlen gyöke. Az általános esetben — mikor tehát csak annyit tudunk, hogy $\mathbf{A}$-nak nincs egységnyi abszolút értékű sajátértéke, az $\mathbf{x}_t$ kovarianciamátrixa nem elégíti ki a fenti egyenletet, sőt az sem biztos, hogy egyáltalán létezik gyöke az egyenletnek. Ezért a következő lemma megfogalmazása kicsit más, mint a 6.3. lemmáé.

6.4. LEMMA. Legyen

(6.37)
$$\mathbf{x}_{t+1}^1 = \mathbf{A}_1\mathbf{x}_t^1 + \mathbf{B}_1\mathbf{e}_t,$$
$$\mathbf{x}_t^2 = \mathbf{A}_2^{-1}\mathbf{x}_{t+1}^2 - \mathbf{A}_2^{-1}\mathbf{B}_2\mathbf{e}_t,$$
$$\mathbf{y}_t = \mathbf{C}_1\mathbf{x}_t^1 + \mathbf{C}_2\mathbf{x}_t^2,$$

ahol $\varrho(\mathbf{A}_1)<1$, $\varrho(\mathbf{A}_2^{-1})<1$. Legyen

$$\mathbf{P}_1 = \sum_{k=1}^{\infty} \mathbf{A}_1^{k-1}\mathbf{B}_1\mathbf{B}_1^T(\mathbf{A}_1^T)^{k-1}$$

$$\mathbf{P}_2 = -\sum_{k=1}^{\infty} \mathbf{A}_2^{-k}\mathbf{B}_2\mathbf{B}_2^T(\mathbf{A}_2^T)^{-k}$$

$$\mathbf{Q}_1 = \sum_{k=1}^{\infty} (\mathbf{A}_1^T)^{k-1}\mathbf{C}_1^T\mathbf{C}_1\mathbf{A}_1^{k-1}$$

$$\mathbf{Q}_2 = -\sum_{k=1}^{\infty} (\mathbf{A}_2^T)^{-k}\mathbf{C}_2^T\mathbf{C}_2\mathbf{A}_2^{-k}.$$

Tegyük fel, hogy $\mathbf{P}_1, \mathbf{P}_2, \mathbf{Q}_1, \mathbf{Q}_2$ invertálhatók és, hogy teljesül a (6.2) tulajdonság. Ekkor létezik a

(6.38)
$$\mathbf{P} = \mathbf{APA}^T + \mathbf{BB}^T$$

$$\mathbf{Q} = \mathbf{A}^T\mathbf{QA} + \mathbf{C}^T\mathbf{C}$$

egyenletnek olyan $\mathbf{P}$, $\mathbf{Q}$ gyökök ($\mathbf{P}=\mathbf{P}^T$, $\mathbf{Q}=\mathbf{Q}^T$), melyekre

(6.39)
$$\mathbf{PQ} = \sigma^2\mathbf{I}$$

Bizonyítás. Vegyük észre, hogy a tételben definiált $\mathbf{P}_1, \mathbf{P}_2, \mathbf{Q}_1, \mathbf{Q}_2$ a

$$\mathbf{P}_1 = \mathbf{A}_1\mathbf{P}_1\mathbf{A}_1^T + \mathbf{B}_1\mathbf{B}_1^T$$
$$\mathbf{P}_2 = \mathbf{A}_2\mathbf{P}_2\mathbf{A}_2^T + \mathbf{B}_2\mathbf{B}_2^T$$
$$\mathbf{Q}_1 = \mathbf{A}_1^T\mathbf{Q}_1\mathbf{A}_1 + \mathbf{C}_1^T\mathbf{C}_1$$
$$\mathbf{Q}_2 = \mathbf{A}_2^T\mathbf{Q}_2\mathbf{A}_2 + \mathbf{C}_2^T\mathbf{C}_2$$

egyenletek egyértelmű gyöke, és $\mathbf{P}_1 = E(\mathbf{x}_t^1(\mathbf{x}_t^1)^T)$, $\mathbf{P}_2 = -E(\mathbf{x}_t^2(\mathbf{x}_t^2)^T)$.

Írjuk fel az $\mathbf{e}_t \to \frac{1}{\sigma}\bar{\mathbf{y}}_t$ leképezés duálisát, ahogy (6.15)—(6.18) során tettük.

(6.40)
$$\zeta_{t-1}^1 = \mathbf{A}_1^T \zeta_t^1 + \mathbf{C}_1^T \bar{\mathbf{y}}_t$$
$$\zeta_t^2 = \mathbf{A}_2^{T-1} \zeta_{t-1}^2 - \mathbf{A}_2^{T-1} \mathbf{C}_2^T \bar{\mathbf{y}}_t$$

(6.41)
$$\sigma^2 \mathbf{e}_t = \mathbf{B}^T \zeta_t + \boldsymbol{\Gamma}^T \bar{\mathbf{y}}_t.$$

Alakítsuk tovább (6.41) jobb oldalát (6.37) és (6.40) felhasználásával:

$$\sigma^2 \mathbf{e}_t = \mathbf{B}_1^T \zeta_t^1 + \mathbf{B}_2^T (\mathbf{A}_2^T)^{-1} \zeta_{t-1}^2 + (\boldsymbol{\Gamma}^T - \mathbf{B}_2^T (\mathbf{A}_2^T)^{-1} \mathbf{C}_2^T)(\mathbf{C}_1 \mathbf{x}_t^1 + \mathbf{C}_2 \mathbf{x}_t^2 + \boldsymbol{\Gamma} \mathbf{e}_t),$$

$\mathbf{x}_t^1$ már független $\mathbf{e}_t$-től, $\mathbf{x}_t^2$ azonban még nem. Használjuk a (6.37) egyenletet:

$$\sigma^2 \mathbf{e}_t = \mathbf{B}_1^T \zeta_t^1 + \mathbf{B}_2^T (\mathbf{A}_2^T)^{-1} \zeta_{t-1}^2 + (\boldsymbol{\Gamma}^T \mathbf{C}_1 - \mathbf{B}_2^T (\mathbf{A}_2^T)^{-1} \mathbf{C}_2^T \mathbf{C}_1) \mathbf{x}_t^1 +$$
$$+ (\boldsymbol{\Gamma}^T \mathbf{C}_2 - \mathbf{B}_2^T (\mathbf{A}_2^T)^{-1} \mathbf{C}_2^T \mathbf{C}_2) \mathbf{A}_2^{-1} \mathbf{x}_{t+1}^2 +$$
$$+ (\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} - \mathbf{B}_2^T \mathbf{A}_2^{T-1} \mathbf{C}_2^T \boldsymbol{\Gamma} - \boldsymbol{\Gamma}^T \mathbf{C}_2 \mathbf{A}_2^{-1} \mathbf{B}_2 + \mathbf{B}_2^T (\mathbf{A}_2^T)^{-1} \mathbf{C}_2^T \mathbf{C}_2 \mathbf{A}_2^{-1} \mathbf{B}_2) \mathbf{e}_t.$$

Próbáljuk meg a ζ_t^1 és a ζ_{t-1}^2 mennyiségeket felbontani úgy, hogy a jobb oldal független valószínűségi változók összege legyen. Először is $k \geqq 0$ esetén

$$E(\zeta_t^1 \mathbf{e}_{t-k}^T) = E\Big(\sum_{j=1}^{\infty} (\mathbf{A}_1^T)^{j-1} \mathbf{C}_1^T \mathbf{y}_{t+j} \mathbf{e}_{t-k}^T\Big) =$$

$$= \sum_{j=1}^{\infty} (\mathbf{A}_1^T)^{j-1} \mathbf{C}_1^T \mathbf{C}_1 E(\mathbf{x}_{t+j}^1 \mathbf{e}_{t-k}^T) = \sum_{j=1}^{\infty} (\mathbf{A}_1^T)^{j-1} \mathbf{C}_1^T \mathbf{C}_1 \mathbf{A}_1^{j+k-1} \mathbf{B}_1.$$

Így
$$E(\zeta_t^1 \mathbf{e}_{t-k}^T) = \mathbf{Q}_1 \mathbf{A}_1^k \mathbf{B}_1.$$

Rögtön adódik, hogy
$$E(\zeta_t^1 - \mathbf{Q}_1 \mathbf{x}_{t+1}^1) \mathbf{e}_{t-k}^T = \mathbf{O} \quad k \geqq 0,$$

ugyanígy megmutatható, hogy

$$E\big((\zeta_{t-1}^2 - \mathbf{Q}_2 \mathbf{x}_t^2) \mathbf{e}_{t+k}^T\big) = \mathbf{O} \quad k \geqq 0.$$

Alakítsuk tovább tehát az egyenletünk jobb oldalát (az újonnan fellépő $\mathbf{x}_t^2$ tagot ismét felírva $\mathbf{x}_{t+1}^2$ és $\mathbf{e}_t$ segítségével $\mathbf{x}_{t+1}^1$-et pedig $\mathbf{x}_t^1$ és $\mathbf{e}_t$ segítségével). Összevonások után a következő egyenletet kapjuk:

$$\sigma^2 \mathbf{e}_t = \mathbf{B}_1^T (\zeta_t^1 - \mathbf{Q}_1 \mathbf{x}_{t+1}^1) + \mathbf{B}_2^T (\mathbf{A}_2^T)^{-1} (\zeta_{t-1}^2 - \mathbf{Q}_2 \mathbf{x}_t^2) +$$
$$+ (\mathbf{B}_1^T \mathbf{Q}_1 \mathbf{A}_1 - \mathbf{B}_2^T (\mathbf{A}_2^T)^{-1} \mathbf{C}_2^T \mathbf{C}_1 + \boldsymbol{\Gamma}^T \mathbf{C}_1) \mathbf{x}_t^1 +$$
$$+ (\mathbf{B}_2^T \mathbf{Q}_2 + \boldsymbol{\Gamma}^T \mathbf{C}_2 \mathbf{x}_2^{-1}) \mathbf{x}_{t+1}^2 +$$
$$+ (\mathbf{B}_1^T \mathbf{Q}_1 \mathbf{B}_1 - \mathbf{B}_2^T \mathbf{Q}_2 \mathbf{B}_2 - \mathbf{B}_2^T (\mathbf{A}_2^T)^{-1} \mathbf{C}_2^T \boldsymbol{\Gamma} - \boldsymbol{\Gamma}^T \mathbf{C}_2 \mathbf{A}_2^{-1} \mathbf{B}_2 + \boldsymbol{\Gamma}^T \boldsymbol{\Gamma}) \mathbf{e}_t.$$

Már csak a $\zeta_t^1 - \mathbf{A}_1 \mathbf{x}_{t+1}^1$ és $\mathbf{x}_{t+1}^2$, ill. $\zeta_{t-1}^2 - \mathbf{Q}_2 \mathbf{x}_t^2$ és $\mathbf{x}_t^1$ korreláltak.
Legyen

(6.42)
$$\boldsymbol{\Phi} = E(\zeta_t^1 \mathbf{x}_{t+1}^{2T}),$$

és

(6.43)
$$\boldsymbol{\psi} = E(\zeta_{t-1}^2 \mathbf{x}_t^{1T}).$$

Ennek alapján a

$$\sigma^2\mathbf{e}_t = \mathbf{B}_1^T(\zeta_t^1 - \mathbf{Q}_1\mathbf{x}_{t+1}^1 + \Phi\mathbf{P}_2^{-1}\mathbf{x}_{t+1}^2) + (\mathbf{B}_2^T\mathbf{Q}_2 + \Gamma^T\mathbf{C}_2\mathbf{A}_2^{-1} - \mathbf{B}_1^T\Phi\mathbf{P}_2^{-1})\mathbf{x}_{t+1}^2 +$$
$$+ \mathbf{B}_2^T(\mathbf{A}_2^T)^{-1}(\zeta_{t-1}^2 - \mathbf{Q}_2\mathbf{x}_t^2 - \psi\mathbf{P}_1^{-1}\mathbf{x}_t^1) +$$
$$+ (\mathbf{B}_1^T\mathbf{Q}_1\mathbf{A}_1 + \Gamma^T\mathbf{C}_1 - \mathbf{B}_2^T(\mathbf{A}_2^T)^{-1}\mathbf{C}_2^T\mathbf{C}_1 + \mathbf{B}_2^T(\mathbf{A}_2^T)^{-1}\psi\mathbf{P}_1^{-1})\mathbf{x}_t^1 +$$
$$+ (\mathbf{B}_1^T\mathbf{Q}_1\mathbf{B}_1 - \mathbf{B}_2^T\mathbf{Q}_2\mathbf{B}_2 - \mathbf{B}_2^T(\mathbf{A}_2^T)^{-1}\mathbf{C}_2^T\Gamma - \Gamma^T\mathbf{C}_2\mathbf{A}_2^{-1}\mathbf{B}_2 + \Gamma^T\Gamma)\mathbf{e}_t$$

egyenlet jobb oldalán levő tagok már függetlenek egymástól. Tehát

$$\mathbf{B}_1^T(\zeta_t^1 - \mathbf{Q}_1\mathbf{x}_{t+1}^1 + \Phi\mathbf{P}_2^{-1}\mathbf{x}_{t+1}^2) = \mathbf{O}, \tag{6.44}$$

$$\mathbf{B}_2^T Q_2 + \Gamma^T\mathbf{C}_2\mathbf{A}_2^{-1} - \mathbf{B}_1^T\Phi\mathbf{P}_2^{-1} = \mathbf{O}, \tag{6.45}$$

$$\mathbf{B}_2^T(\mathbf{A}_2^T)^{-1}(\zeta_t^2 - \mathbf{Q}_2\mathbf{x}_t^2 - \psi\mathbf{P}_1^{-1}\mathbf{x}_t^1) = \mathbf{O}, \tag{6.46}$$

$$\mathbf{B}_1^T\mathbf{Q}_1\mathbf{A}_1 + \Gamma^T\mathbf{C}_1 - \mathbf{B}_2^T(\mathbf{A}_2^T)^{-1}\mathbf{C}_2^T\mathbf{C}_1 + \mathbf{B}_2^T(\mathbf{A}_2^T)^{-1}\psi\mathbf{P}_1^{-1} = \mathbf{O}, \tag{6.47}$$

$$\mathbf{B}_1^T\mathbf{Q}_1\mathbf{B}_1 - \mathbf{B}_2^T\mathbf{Q}_2\mathbf{B}_2 - \mathbf{B}_2^T(\mathbf{A}_2^T)^{-1}\mathbf{C}_2^T\Gamma - \Gamma^T\mathbf{C}_2\mathbf{A}_2^{-1}\mathbf{B}_2 + \Gamma^T\Gamma = \sigma^2\mathbf{I}. \tag{6.48}$$

Azt kellene megmutatni, a 6.3. lemma bizonyításához hasonlóan, hogy

$$\zeta_t^1 - \mathbf{Q}_1\mathbf{x}_{t+1}^1 + \Phi\mathbf{P}_2^{-1}\mathbf{x}_{t+1}^2 = \mathbf{O},$$

és

$$\zeta_{t-1}^2 - \mathbf{Q}_2\mathbf{x}_t^2 - \psi\mathbf{P}_1^{-1}\mathbf{x}_t^1 = \mathbf{O}.$$

Próbáljuk felhasználni a (6.37) és (6.40) egyenleteket:

$$\zeta_{t-1}^1 = \mathbf{A}_1^T\zeta_t^1 + \mathbf{C}_1^T\bar{\mathbf{y}}_t = \mathbf{A}_1^T\zeta_t^1 + (\mathbf{Q}_1 - \mathbf{A}_1^T\mathbf{Q}_1\mathbf{A}_1)\mathbf{x}_t^1 + \mathbf{C}_1^T\mathbf{C}_2\mathbf{x}_t^2 + \mathbf{C}_1^T\Gamma\mathbf{e}_t.$$

Átrendezve

$$\zeta_{t-1}^1 - \mathbf{Q}_1\mathbf{x}_t^1 = \mathbf{A}_1^T(\zeta_t^1 - \mathbf{Q}_1\mathbf{x}_{t+1}^1) + \mathbf{C}_1^T\mathbf{C}_2\mathbf{A}_2^{-1}\mathbf{x}_{t+1}^2 + (\mathbf{C}_1^T\Gamma + \mathbf{A}_1^T\mathbf{Q}_1\mathbf{B}_1 - \mathbf{C}_1^T\mathbf{C}_2\mathbf{A}_2^{-1}\mathbf{B}_2)\mathbf{e}_t.$$

Írjuk fel $\mathbf{e}_t$ együtthatóját (6.47) alapján és adjuk hozzá az egyenlethez a

$$\Phi\mathbf{P}_2^{-1}\mathbf{x}_t^2 = \mathbf{A}_1^T\Phi\mathbf{P}_2^{-1}\mathbf{x}_{t+1}^2 + (\Phi\mathbf{P}_2^{-1}\mathbf{A}_2^{-1} - \mathbf{A}_1^T\Phi\mathbf{P}_2^{-1})\mathbf{x}_{t+1}^2 - \Phi\mathbf{P}_2^{-1}\mathbf{A}_2^{-1}\mathbf{B}_2\mathbf{e}_t$$

egyenlet megfelelő oldalait:

$$\zeta_{t+1}^1 - \mathbf{Q}_1\mathbf{x}_t^1 + \Phi\mathbf{P}_2^{-1}\mathbf{x}_t^2 = \mathbf{A}_1^T(\zeta_t^1 - \mathbf{Q}_1\mathbf{x}_{t+1}^1 + \Phi\mathbf{P}_2^{-1}\mathbf{x}_{t+1}^2) + \tag{6.49}$$
$$+ (\Phi\mathbf{P}_2^{-1}\mathbf{A}_2^{-1} - \mathbf{A}_1^T\Phi\mathbf{P}_2^{-1} + \mathbf{C}_1^T\mathbf{C}_2\mathbf{A}_2^{-1})\mathbf{x}_{t+1}^2 -$$
$$- (\Phi\mathbf{P}_2^{-1}\mathbf{A}_2^{-1}\mathbf{B}_2 + \mathbf{P}_1^{-1}\psi^T\mathbf{A}_2^{-1}\mathbf{B}_2)\mathbf{e}_t.$$

A bal oldal $\mathbf{B}_1^T$-szorosa éppen nulla. Mivel a jobb oldal tagjai függetlenek egymástól, így azok $\mathbf{B}_1^T$-szeresei külön-külön is nullák. Tehát speciálisan

$$\mathbf{B}_1^T\mathbf{A}_1^T(\zeta_t^1 - \mathbf{Q}_1\mathbf{x}_{t+1}^1 + \Phi\mathbf{P}_2^{-1}\mathbf{x}_{t+1}^2) = \mathbf{O}.$$

Ezt az eljárást iterálva — a $\mathbf{B}_1^T\mathbf{A}_1\mathbf{T}^k$, $k = 1, 2, \ldots$ mátrixszal szorozva — a

$$\mathbf{B}_1^T\mathbf{A}_1\mathbf{T}^k(\zeta_t^1 - \mathbf{Q}_1\mathbf{x}_{t+1}^1 + \Phi\mathbf{P}_2^{-1}\mathbf{x}_{t+1}^1) = \mathbf{O}$$

adódik. Mivel

$$\mathbf{P}_1 = \sum_{k=0}^{\infty} \mathbf{A}_1^k,\ \mathbf{B}_1\mathbf{B}_1^T\mathbf{A}_1\mathbf{T}^k,$$

így

$$(\zeta_t^1 - \mathbf{Q}_1\mathbf{x}_{t+1}^1 + \Phi\mathbf{P}_2^{-1}\mathbf{x}_{t+1}^2)^T\mathbf{P}_1(\zeta_t^1 - \mathbf{Q}\mathbf{x}_{t+1}^1 + \Phi\mathbf{P}_2^{-1}\mathbf{x}_{t+1}^2) = \mathbf{O}.$$

$\mathbf{P}_1$ pozitív definit mátrix, így

$$\zeta_t^1 = \mathbf{Q}_1\mathbf{x}_{t+1}^1 - \Phi\mathbf{P}_2^{-1}\mathbf{x}_{t+1}^2. \tag{6.50}$$

Ezt visszaírva (6.49)-be a

$$\Phi\mathbf{P}_2^{-1}\mathbf{A}_2^{-1} - \mathbf{A}_1^T\Phi\mathbf{P}_2^{-1} + \mathbf{C}_1^T\mathbf{C}_2\mathbf{A}_2^{-1} = \mathbf{O}, \tag{6.51}$$

$$(\Phi\mathbf{P}_2^{-1} + \mathbf{P}_1^{-1}\psi^T)\mathbf{A}_2^{-1}\mathbf{B}_2 = \mathbf{O} \tag{6.52}$$

összefüggéseket kapjuk.
A $\xi_{t-1}^2 - \mathbf{Q}_2\mathbf{x}_t^2 - \psi\mathbf{P}_1^{-1}\mathbf{x}_t^1$ mennyiségre a következő összefüggés teljesül:

$$\zeta_t^2 = \mathbf{Q}_2\mathbf{x}_{t+1}^2 - \psi\mathbf{P}_1^{-1}\mathbf{x}_{t+1}^1 = (\mathbf{A}_2^T)^{-1}(\zeta_{t-1}^2 - \mathbf{Q}_2\mathbf{x}_t^2 - \psi\mathbf{P}_2^{-1}\mathbf{x}_t^1) +$$

$$+ ((\mathbf{A}_2^T)^{-1}\psi\mathbf{P}_1^{-1} - \psi\mathbf{P}_1^{-1}\mathbf{A}_1 - (\mathbf{A}_2^T)^{-1}\mathbf{C}_1^T\mathbf{C}_1)\mathbf{x}_t^1 - (\mathbf{P}_2^{-1}\Phi^T\mathbf{B}_1 + \psi\mathbf{P}_1^{-1}\mathbf{B}_1)\mathbf{e}_t,$$

ebből a

$$\zeta_t^2 = \mathbf{Q}_2\mathbf{x}_{t+1}^2 + \psi\mathbf{P}_1^{-1}\mathbf{x}_{t+1}^1, \tag{6.53}$$

$$(\mathbf{A}_2^T)^{-1}\psi\mathbf{P}_1^{-1} - \psi\mathbf{P}_1^{-1}\mathbf{A}_1 - (\mathbf{A}_2^T)^{-1}\mathbf{C}_2^T\mathbf{C}_1 = \mathbf{O}, \tag{6.54}$$

$$(\mathbf{P}_2^{-1}\Phi^T + \psi\mathbf{P}_1^{-1})\mathbf{B}_1 = \mathbf{O} \tag{6.55}$$

összefüggések adódnak. (6.51) transzponáltját (6.54)-hez hozzáadva, átrendezés után az alábbi egyenlőséghez jutunk:

$$(\mathbf{A}_2^T)^{-1}(\mathbf{P}_2^{-1}\Phi^T + \psi\mathbf{P}_1^{-1}) = (\mathbf{P}_2^{-1}\Phi^T + \psi\mathbf{P}_1^{-1})\mathbf{A}_1.$$

Így (6.55)-öt az $(\mathbf{A}_2^T)^{-1}$ mátrixszal szorozva a

$$(\mathbf{P}_2^{-1}\Phi^T + \psi\mathbf{P}_1^{-1})\mathbf{A}_1\mathbf{B}_1 = \mathbf{O}$$

egyenlethez jutunk. Ezt iterálva a

$$(\mathbf{P}_2^{-1}\Phi^T + \psi\mathbf{P}_1^{-1})\mathbf{A}_1^k\mathbf{B}_1 = \mathbf{O}$$

egyenletekre, s végül a

$$\mathbf{P}_2^{-1}\Phi^T + \psi\mathbf{P}_1^{-1} = \mathbf{O} \tag{6.56}$$

összefüggésre jutunk.
Legyen

$$\mathbf{M} = \mathbf{P}_1^{-1}\psi^T = -\Phi\mathbf{P}_2^{-1}. \tag{6.57}$$

Ekkor (6.51) alapján

$$\mathbf{M} = \mathbf{A}_1^T\mathbf{M}\mathbf{A}_2 + \mathbf{C}_1^T\mathbf{C}_2.$$

Így a

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_1 \mathbf{M} \\ \mathbf{M}^T \mathbf{Q}_2 \end{bmatrix} \tag{6.58}$$

mátrix megoldása a

$$\mathbf{Q} = \mathbf{A}^T \mathbf{Q} \mathbf{A} + \mathbf{C}^T \mathbf{C} \tag{6.59}$$

egyenletnek.

A (6.50) és (6.53) azonosságok mindkét oldalának kovarlanciáit véve a

$$\sigma^2 \mathbf{Q}_1 = \mathbf{Q}_1 \mathbf{P}_1 \mathbf{Q}_1 - \Phi \mathbf{P}_2^{-1} \Phi^T, \tag{6.60}$$

$$-\sigma^2 \mathbf{Q}_2 = -\mathbf{Q}_2 \mathbf{P}_2 \mathbf{Q}_2 + \psi \mathbf{P}_1^{-1} \psi^T \tag{6.61}$$

$$\mathbf{O} = \mathbf{Q}_1 \mathbf{P}_1 \mathbf{P}_1^{-1} \psi^T + \Phi \mathbf{P}_2^{-1} \mathbf{P}_2 \mathbf{Q}_2$$

összefüggések adódnak. A legutolsó egyenlet alapján

$$\psi^T \mathbf{Q}_2^{-1} + \mathbf{Q}_1^{-1} \Phi = \mathbf{O}.$$

Legyen

$$\mathbf{R} = \mathbf{Q}_1^{-1} \Phi = -\psi^T \mathbf{Q}_2^{-1}.$$

Ekkor könnyen ellenőrizhető, hogy

$$\mathbf{R}^T \mathbf{Q}_1 + \mathbf{P}_2 \mathbf{M}^T = \mathbf{O},$$

$$\mathbf{P}_1 \mathbf{M} + \mathbf{R} \mathbf{Q}_2 = \mathbf{O},$$

és (6.60)—(6.61) alapján

$$\mathbf{P}_1 \mathbf{Q}_1 + \mathbf{R} \mathbf{M}^T = \sigma^2 \mathbf{I},$$

$$\mathbf{P}_2 \mathbf{Q}_2 + \mathbf{R}^T \mathbf{M} = \sigma^2 \mathbf{I}.$$

Bevezetve a

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 & \mathbf{R} \\ \mathbf{R}^T & \mathbf{P}_2 \end{bmatrix}$$

mátrixot, az előző négy egyenlet röviden így írható;

$$\mathbf{P}\mathbf{Q} = \sigma^2 \mathbf{I}. \tag{6.62}$$

(6.47) és (6.54) alapján

$$\Gamma^T \mathbf{C}_1 + \mathbf{B}_1^T \mathbf{Q}_1 \mathbf{A}_1 + \mathbf{B}_2^T \psi \mathbf{P}_1^{-1} \mathbf{A}_1 = \mathbf{O}.$$

Így

$$\Gamma^T \mathbf{C}_1 + \mathbf{B}_1^T \mathbf{Q}_1 \mathbf{A}_1 + \mathbf{B}_2^T \mathbf{M}^T \mathbf{A}_1 = \mathbf{O}. \tag{6.63}$$

(6.45) (6.57) felhasználásával így írható:

$$\Gamma^T \mathbf{C}_2 + \mathbf{B}_2^T \mathbf{Q}_2 \mathbf{A}_2 + \mathbf{B}_1^T \mathbf{M} \mathbf{A}_2 = \mathbf{O}. \tag{6.64}$$

E két utóbbi egyenlet összevontan a

$$\mathbf{B}^T \mathbf{Q} \mathbf{A} + \Gamma^T \mathbf{C} = \mathbf{O} \tag{6.65}$$

egyenletet jelenti. Mivel (6.64) alapján

$$\mathbf{Q}_2\mathbf{B}_2+(\mathbf{A}_2^T)^{-1}\mathbf{C}_2^T\boldsymbol{\Gamma}=\mathbf{M}^T\mathbf{B}_1,$$

ezért (6.48) összevontan így írható

$$\mathbf{B}^T\mathbf{Q}\mathbf{B}+\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}=\sigma^2\mathbf{I}. \tag{6.66}$$

Már csak annak igazolása maradt hátra, hogy teljesül a

$$\mathbf{P}=\mathbf{A}\mathbf{P}\mathbf{A}^T+\mathbf{B}\mathbf{B}^T \tag{6.67}$$

egyenlőség. Azonban, ha nem a

$$\sigma^2\mathbf{e}_t=\mathbf{B}^T\boldsymbol{\zeta}_t+\boldsymbol{\Gamma}^T\mathbf{y}_t$$

egyenlet jobb oldalán levő tagok független összeadandókra bontásából indulunk ki, hanem az

$$\bar{\mathbf{y}}_t=\mathbf{C}^T\mathbf{x}_t+\boldsymbol{\Gamma}\mathbf{e}_t$$

egyenlet jobb oldalába írjuk be $\mathbf{e}_t$ helyébe (6.41) jobb oldalát és így próbáljuk független összeadandókra bontani — azaz $\mathbf{u}_t$ és $\bar{\mathbf{y}}_t$ szerepét megcseréljük —, akkor $\mathbf{x}_t$ és $\boldsymbol{\zeta}_t$ is szerepet cserélnek, és az előző gondolatmenet végigkövetésével megkaphatjuk a bizonyítandó (6.67) egyenletet.

Láttuk, hogy a $\mathbf{P}\mathbf{Q}=\sigma^2\mathbf{I}$ feltétel szükséges a 6.2 tulajdonság teljesüléséhez. Vizsgáljuk meg, hogy elégséges-e! Tekintsük tehát a (6.37) állapotegyenleteket, a hozzá tartozó (6.38) mátrix-egyenleteket és tegyük fel, hogy (6.38)-nak létezik a (6.39) feltételének eleget tevő megoldása. Azzaz

$$\mathbf{x}_{t+1}=\mathbf{A}\mathbf{x}_t+\mathbf{B}\mathbf{e}_t$$

$$\mathbf{y}_t=\mathbf{C}\mathbf{x}_t \tag{6.37'}$$

$$\varrho(\mathbf{A}_1)<1,\quad \varrho(\mathbf{A}_2^{-1})<1,$$

$$\mathbf{P}=\mathbf{A}\mathbf{P}\mathbf{A}^T+\mathbf{B}\mathbf{B}^T \tag{6.38'}$$

$$\mathbf{Q}=\mathbf{A}^T\mathbf{Q}\mathbf{A}+\mathbf{C}^T\mathbf{C}$$

$$\mathbf{P}\mathbf{Q}=\sigma^2\mathbf{I}. \tag{6.39'}$$

Definiáljuk a $\{\xi_t,\ t=0,\pm1,\pm2,\ldots\}$ valószínűségi változókat a következőképpen:

$$\xi_t^1=\sum_{k=1}^{\infty}(\mathbf{A}_1^T)^{k-1}\mathbf{C}_1^T\mathbf{e}_{t+k}, \tag{6.68}$$

$$\xi_t^2=-\sum_{k=0}^{\infty}(\mathbf{A}_2^T)^{-(k+1)}\mathbf{C}_2^T\mathbf{e}_{t-k}.$$

Ekkor

$$\xi_t^1=\mathbf{A}_1^T\xi_{t+1}^1+\mathbf{C}_1^T\mathbf{e}_{t+1}, \tag{6.69}$$

$$\xi_{t+1}^2=(\mathbf{A}_2^T)^{-1}\xi_t^2-(\mathbf{A}_1^T)^{-1}\mathbf{C}_2^T\mathbf{e}_{t+1}.$$

Legyen

$$\xi_t = \begin{bmatrix} \xi_t^1 \\ \xi_t^2 \end{bmatrix},$$

$$\xi_t = \mathbf{A}^T \xi_{t+1} + \mathbf{C}^T \mathbf{e}_{t+1}. \tag{6.70}$$

Tekintsük az $\{\mathbf{e}_t, t=0, \pm 1, \ldots\}$ valószínűségi változók koordinátái által generált E zárt lineáris alteret, és legyen X az $\{\mathbf{x}_t, t=0, \pm 1, \ldots\}$, Ξ a $\{\xi_t, t=0, \pm 1, \ldots\}$ változók koordinátái által kifeszített zárt altér. Ekkor $X \subset E$, $\Xi \subset X$.

Vegyük észre, hogy X-et éppen a $\{\mathbf{B}\mathbf{e}_t, t=0, \pm 1, \ldots\}$ változók koordinátái generálják, Ξ-t pedig $\{\mathbf{C}^T\mathbf{e}_t, t=0, \pm 1, \ldots\}$ koordinátái. Nézzük meg, hogy $\mathbf{e}_t$ „mennyivel lóg ki" az X, ill. Ξ alterekből!

Mivel $s \neq t$ esetén $\mathbf{e}_t$ és $\mathbf{B}\mathbf{e}_s$ függetlenek, $\mathbf{e}_t$ és $\mathbf{C}^T\mathbf{e}_s$ függetlenek, így az

$$\mathbf{e}_t - E(\mathbf{e}_t|\mathbf{B}\mathbf{e}_t) \quad \text{ill.} \quad \mathbf{e}_t - E(\mathbf{e}_t|\mathbf{C}^T\mathbf{e}_t)$$

különbségeket kell meghatározni.

6.5. LEMMA. Tegyük fel, hogy a (6.38) egyenleteknek létezik (6.39)-nek eleget tevő megoldása. Ekkor létezik olyan $\Delta q \times q$-s invertálható mátrix, melyre

$$\Delta\Delta^T = \mathbf{I},$$

és

$$\left(\mathbf{e}_t - E(\mathbf{e}_t|\mathbf{C}^T\mathbf{e}_t)\right) \quad \text{és} \quad \Delta\left(\mathbf{e}_t - E(\mathbf{e}_t|\mathbf{B}\mathbf{e}_t)\right) \tag{6.71}$$

kovariancia mátrix megegyeznek.

Bizonyítás. Mivel

$$E(\mathbf{e}_t|\mathbf{C}^T\mathbf{e}_t) = \mathbf{C}(\mathbf{C}^T\mathbf{C})^-\mathbf{C}^T\mathbf{e}_t,$$

$$E(\mathbf{e}_t|\mathbf{B}\mathbf{e}_t) = \mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^-\mathbf{B}\mathbf{e}_t,$$

ezért azt kell mutatni, hogy létezik olyan Δ mátrix ($\Delta\Delta^T = \mathbf{I}$), melyre

$$\left(\mathbf{I} - \mathbf{C}(\mathbf{C}^T\mathbf{C})^-\mathbf{C}^T\right) = \Delta\left(\mathbf{I} - \mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^-\mathbf{B}\right)\Delta^T.$$

Induljunk ki a

$$\mathbf{Q} = \mathbf{A}^T\mathbf{Q}\mathbf{A} + \mathbf{C}^T\mathbf{C},$$

$$\mathbf{P} = \mathbf{A}\mathbf{P}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T$$

egyenletekből.

A második egyenlet alapján

$$\mathbf{P}^{-1} = \mathbf{A}^T\mathbf{P}^{-1}\mathbf{A} + \mathbf{K}\mathbf{K}^T,$$

ahol

$$\mathbf{K} = -(\mathbf{I} + \mathbf{A}^T)\mathbf{P}^{-1}(\mathbf{I} + \mathbf{A})^{-1}\mathbf{B}.$$

(6.38) és (6.39) összevetésével a

$$\mathbf{C}^T\mathbf{C} = \sigma^2\mathbf{K}\mathbf{K}^T$$

adódik. Mivel $\sigma^{-1}\mathbf{C}^T$ és $\mathbf{K}$ is $p \times q$-s mátrixok, így létezik Δ ortogonális mátrix ($\Delta\Delta^T = \mathbf{I}$), hogy

$$\mathbf{C} = \Delta\sigma\mathbf{K}^T.$$

Ekkor

$$\mathbf{C}(\mathbf{C}^T\mathbf{C})^-\mathbf{C}^T = \Delta\sigma\mathbf{K}^T(\sigma\mathbf{K}\Delta^T\Delta\sigma\mathbf{K}^T)^-\sigma\mathbf{K}\Delta^T =$$

$$= \Delta(\mathbf{K}^T(\mathbf{K}\mathbf{K}^T)^-\mathbf{K})\Delta^T = \Delta(\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^-\mathbf{B})\Delta^T$$

igazolja a (6.71) állítást.

6.6. LEMMA. Legyen

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{e}_t,$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t.$$

Tegyük fel, hogy $\varrho(\mathbf{A}_1)<1$, $\varrho(\mathbf{A}_2^{-1})<1$, és hogy a

$$\mathbf{P} = \mathbf{A}\mathbf{P}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T,$$

$$\mathbf{Q} = \mathbf{A}^T\mathbf{Q}\mathbf{A} + \mathbf{C}^T\mathbf{C}$$

egyenleteknek létezik a

$$\mathbf{P}\mathbf{Q} = \sigma^2\mathbf{I}$$

feltételnek eleget tevő megoldása.
Ekkor teljesül a 6.2 tulajdonság, azaz létezik olyan Γ mátrix, hogy az

$$\bar{\mathbf{y}}_t = \mathbf{C}\mathbf{x}_t + \Gamma\mathbf{e}_t$$

sorozat független, $\sigma^2\mathbf{I}$ kovarianciájú valószínűségi változókból áll.

Bizonyítás. Számítsuk ki az $\{\mathbf{x}_t, t=0, \pm1, \ldots\}$ és a $\{\xi_t, t=0, \pm1, \ldots\}$ sorozat kovariancia függvényét. Könnyen látható, hogy

$$E\mathbf{x}_t^1\mathbf{x}_{t-k}^1 = \mathbf{A}_1^k\mathbf{P}_1 \quad k \geqq 0.$$

$$E\mathbf{x}_t^2\mathbf{x}_{t+k}^2 = (\mathbf{A}_2)^{-k}(-\mathbf{P}_2) \quad k \geqq 0,$$

$$E\mathbf{x}_t^1\mathbf{x}_{t+k}^2 = \mathbf{Q} \quad k \geqq 0.$$

Ugyanakkor

$$E\mathbf{x}_t^1\mathbf{x}_{t-k}^2 = -\sum_{j=1}^{k}\mathbf{A}_1^{j-1}\mathbf{B}_1\mathbf{B}_2^T(\mathbf{A}_2^T)^{-(k-j+1)}, \quad k \geqq 1.$$

Használjuk ki a (6.38) összefüggést. Eszerint

$$\mathbf{R} - \mathbf{A}_1\mathbf{R}\mathbf{A}_2^T = \mathbf{B}_1\mathbf{B}_2^T,$$

illetve

$$\mathbf{R}(\mathbf{A}_2^T)^{-1} - \mathbf{A}_1\mathbf{R} = \mathbf{B}_1\mathbf{B}_2^T(\mathbf{A}_2^T)^{-1}.$$

Így

$$E\mathbf{x}_t^1\mathbf{x}_{t-k}^2 = -\Big(\sum_{j=1}^{k}\mathbf{A}_1^{j-1}\mathbf{R}(\mathbf{A}_2^T)^{-(k-j+1)} - \sum_{j=1}^{k}\mathbf{A}_1^{j-1}\mathbf{R}(\mathbf{A}_2^T)^{-(k-j)}\Big) =$$

$$= \mathbf{A}_1^k\mathbf{R} - \mathbf{R}(\mathbf{A}_2^T)^{-k}.$$

Tehát

$$E\mathbf{x}_t\mathbf{x}_{t-k}^T = \begin{bmatrix} \mathbf{A}_1^k\mathbf{P}_1 & \mathbf{A}_1^k\mathbf{R} - \mathbf{R}(\mathbf{A}_2^T)^{-k} \\ \mathbf{O} & -\mathbf{P}_2(\mathbf{A}_2^T)^{-k} \end{bmatrix}, \quad k \geqq 0. \tag{6.72}$$

Hasonlóképpen

$$E\xi_t\xi_{t+k}^T = \begin{bmatrix} (\mathbf{A}_1^T)^k\mathbf{Q}_1 & (\mathbf{A}_1^T)^k\mathbf{M}-\mathbf{M}\mathbf{A}_2^{-k} \\ \mathbf{O} & -\mathbf{Q}_2\mathbf{A}_2^{-k} \end{bmatrix}, \quad k \geqq 0. \tag{6.73}$$

Vegyük észre, hogy $\mathbf{PQ}=\sigma^2\mathbf{I}$ esetén

$$E(\xi_t\xi_{t+k}^T) = \sigma^{-2}\mathbf{Q}E(x_t x_{t+k}^T)\mathbf{Q}, \quad k = 0, \pm 1, \ldots; t = 0, \pm 1, \ldots \tag{6.74}$$

Definiálunk egy

$$\tau\colon E \to E$$

izometriát, és megmutatjuk, hogy az

$$\bar{\mathbf{y}}_t = \sigma^2\tau e_t$$

sorozat kielégíti a tétel feltételeit.
Először definiáljuk τ-t a Ξ altéren a következőképpen:

$$\tau\xi_t = \sigma\mathbf{P}^{-1}\mathbf{x}_{t+1}.$$

(6.74) miatt τ értelmesen definiált és τ izometria. Terjesszük ki τ-t E-re az alábbi módon. Legyen

$$\tau\left[(\mathbf{I}-\mathbf{C}(\mathbf{C}^T\mathbf{C})^-\mathbf{C}^T)\mathbf{e}_t\right] = \Delta(\mathbf{I}-\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^-\mathbf{B})\mathbf{e}_t.$$

A 6.5. lemma alapján

$$\tau\colon E \to E$$

is izometria. Számoljuk ki $\tau\mathbf{e}_t$ értékét:

$$\tau\mathbf{e}_t = \tau\big(\mathbf{C}(\mathbf{C}^T\mathbf{C})^-\mathbf{C}^T\mathbf{e}_t\big)+\Delta\big(\mathbf{I}-\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^-\mathbf{B}\big)\mathbf{e}_t.$$

Próbáljuk meg az első tagot kifejezni az $\{\mathbf{x}_t, t=0, \pm 1, \ldots\}$ sorozat elemeivel. Mivel

$$\tau\xi_t = \sigma\mathbf{P}^{-1}\mathbf{x}_{t+1},$$

és

$$\mathbf{C}^T\mathbf{e}_t = \xi_{t-1}-\mathbf{A}^T\xi_t,$$

így

$$\tau\mathbf{C}^T\mathbf{e}_t = \sigma(\mathbf{P}^{-1}\mathbf{x}_1-\mathbf{A}^T\mathbf{P}^{-1}\mathbf{x}_{t+1}).$$

Számoljuk ki az

$$E(\mathbf{P}^{-1}\mathbf{x}_t-\mathbf{A}^T\mathbf{P}^{-1}\mathbf{x}_{t+1}|\mathbf{x}_t^1\mathbf{x}_{t+1}^2)$$

mennyiséget. (6.38), (6.39) és (6.7) alapján ennek értéke

$$\sigma^{-2}\mathbf{C}^T\mathbf{C}\begin{bmatrix}\mathbf{I} & 0\\ 0 & \mathbf{A}_2^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{x}_t^1\\ \mathbf{x}_{t+1}^2\end{bmatrix}.$$

Ugyanakkor közvetlen számolás mutatja, hogy

$$\sigma(\mathbf{P}^{-1}\mathbf{x}_t-\mathbf{A}^T\mathbf{P}^{-1}\mathbf{x}_{t+1})-\sigma^{-1}\mathbf{C}^T\mathbf{C}\begin{bmatrix}\mathbf{x}_t^1\\ \mathbf{A}_2^{-1}\mathbf{x}_{t+1}^2\end{bmatrix}$$

független az $\{\mathbf{e}_s, s \neq t\}$ valószínűségi változóktól, így ez $\mathbf{e}_t$ konstansszorosa. Mivel

$$\mathbf{A}_2^{-1}\mathbf{x}_{t+1}^2 = \mathbf{x}_t^2 + \mathbf{A}_2^{-1}\mathbf{B}_2\mathbf{e}_t,$$

így létezik olyan Γ' mátrix, hogy

$$\tau\mathbf{e}_t - \mathbf{C}(\mathbf{C}^T\mathbf{C})^- \sigma^{-1}\mathbf{C}^T\mathbf{C}\mathbf{x}_t = \Gamma'\mathbf{e}_t.$$

Azaz

$$\tau\mathbf{e}_t = \sigma^{-1}\mathbf{C}\mathbf{x}_t + \Gamma'\mathbf{e}_t.$$

Azonban τ izometria, így $\tau\mathbf{e}_t$ független, $\mathbf{I}$ kovarianciájú sorozat, tehát a $\Gamma = \sigma\Gamma'$ választással az

$$\bar{\mathbf{y}}_t = \mathbf{C}\mathbf{x}_t + \Gamma\mathbf{e}_t$$

sorozat megfelel a tétel követelményeinek.

Megjegyzés. Vegyük észre, hogy a $\mathbf{P}^{-1}\mathbf{x}_t$ folyamat valamilyen értelemben az $\mathbf{x}_t$ „időbeli megfordítottja", pontosabban az

$$\eta_t = \begin{bmatrix} \eta_t^1 \\ \eta_t^2 \end{bmatrix} = \mathbf{P}^{-1}\mathbf{x}_t$$

jelöléssel

$$\eta_t^1 = \mathbf{A}_1^T\eta_{t+1}^1 + \mathbf{K}_1\mathbf{f}_{t+1},$$

$$\eta_{t+1}^2 = (\mathbf{A}_2^T)^{-1}\eta_t^2 - (\mathbf{A}_2^T)^{-1}\mathbf{K}_2\mathbf{f}_{t+1}$$

ahol $\{\mathbf{f}_t, t = 0, \pm 1, \ldots\}$ független, $\mathbf{I}$ kovarianciájú sorozat, $\mathbf{K}_1, \mathbf{K}_2$ pedig alkalmas mátrixok. A $\mathbf{K} = \begin{bmatrix} \mathbf{K}_1 \\ \mathbf{K}_2 \end{bmatrix}$ mátrix megválasztható pl. a

$$\mathbf{K} = -(\mathbf{I} + \mathbf{A}^T)\mathbf{P}^{-1}(\mathbf{I} + \mathbf{A})^{-1}\mathbf{B}$$

mátrixnak.

Megjegyzés. Ha $X = E = \Xi$, akkor

$$\tau\mathbf{e}_t = \mathbf{C}(\mathbf{C}^T\mathbf{C})^- \sigma(\mathbf{P}^{-1}\mathbf{x}_t - \mathbf{A}^T\mathbf{P}^{-1}\mathbf{x}_{t+1}),$$

tehát

$$\bar{\mathbf{y}}_t = \sigma^2\mathbf{C}(\mathbf{C}^T\mathbf{C})^- \mathbf{K}\mathbf{f}_{t+1}.$$

Mivel $\mathbf{C}^T\mathbf{C} = \sigma^2\mathbf{K}\mathbf{K}^T$, és $\Xi = E$ esetben $\mathbf{C}(\mathbf{C}^T\mathbf{C})^-\mathbf{C}^T = \mathbf{I}$, így $\mathbf{C}(\mathbf{C}^T\mathbf{C})^-\mathbf{K}$ ortogonális mátrix, tehát $\mathbf{y}^t$ lényegében az $\mathbf{x}^t$ időbeli megfordított folyamatában fellépő független sorozat

7. A forgatások progresszív része

Az előző fejezetben azt a kérdést vizsgáltuk meg, hogy mikor lehet egy

(7.1)
$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{e}_t,$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t$$

összefüggésekkel megadott $\{\mathbf{y}_t, t = 0, \pm 1, \ldots\}$ folyamatból az $\mathbf{e}_t$ hiba konstansszorosának hozzáadásával független, $N(0, \sigma^2\mathbf{I})$ eloszlású valószínűségi változó sorozatot

csinálni. Az $\mathbf{A}$ mátrixról 6.3. lemmában előbb feltettük, hogy $\varrho(\mathbf{A})<1$, majd a 6.4. és 6.6. lemmában csak annyi kellett, hogy $\mathbf{A}$-nak ne essék saját értéke a komplex egységkörre — s így ekkor feltehetjük, hogy

$$\mathbf{A}=\begin{bmatrix}\mathbf{A}_1 & 0\\ 0 & \mathbf{A}_2\end{bmatrix}, \tag{7.2}$$

ahol $\varrho(\mathbf{A}_1)<1$, $\varrho(\mathbf{A}_2^{-1})<1$. Ekkor az egyenleteinket az $\mathbf{x}_t$ változó, és a $\mathbf{B}$, $\mathbf{C}$ mátrix partícionálásával így írhattuk:

$$\begin{aligned}\mathbf{x}_{t+1}^1&=\mathbf{A}_1\mathbf{x}_t^1+\mathbf{B}_1\mathbf{e}_t,\\ \mathbf{x}_{t+1}^2&=\mathbf{A}_2\mathbf{x}_t^2+\mathbf{B}_2\mathbf{e}_t,\\ \mathbf{y}_t&=\mathbf{C}_1\mathbf{x}_t^1+\mathbf{C}_2\mathbf{x}_t^2,\end{aligned} \tag{7.3}$$

(vö.: (6.4), (6.5), (6.6)).

Most azzal az általánosabb kérdéssel fogunk foglalkozni, mikor $\mathbf{A}_1\mathbf{B}_1\mathbf{C}_1$ adottak és keresendő $\mathbf{A}_2$, $\mathbf{B}_2$, $\mathbf{C}_2$, Γ úgy, hogy $\varrho(\mathbf{A}_2^{-1})<1$, és az

$$\bar{\mathbf{y}}_t=\mathbf{y}_t+\Gamma\mathbf{e}_t \tag{7.4}$$

folyamat, független $N(0,\sigma^2\mathbf{I})$ eloszlású változókból áll. (Az $\mathbf{y}_t$ folyamat előállításában szereplő $\mathbf{C}_1\mathbf{x}_t^1$ tagot progressziónak, a $\mathbf{C}_2\mathbf{x}_t^2$ tagot regressziónak nevezzük.) További nehézséget jelent majd, hogy az $\bar{\mathbf{y}}_t$ kovariaciájában szereplő σ^2 értéke sem adott előre, ennek értékét minél kisebbre szeretnénk megválasztani.

Az $(\mathbf{A}_1,\mathbf{B}_1,\mathbf{C}_1)$ mátrixok segítségével definiáljunk egy $\chi: l_2\to l_2$ leképezést a következőképpen: χ mátrixa legyen a

$$\mathbf{H}=\begin{bmatrix}\mathbf{C}_1\mathbf{B}_1 & \mathbf{C}_1\mathbf{A}_1\mathbf{B}_1 & \mathbf{C}_1\mathbf{A}_1^2\mathbf{B}\ldots\\ \mathbf{C}_1\mathbf{A}_1\mathbf{B}_1 & \mathbf{C}_1\mathbf{A}_1^2\mathbf{B}_1 & \ldots\\ \ldots & & \end{bmatrix} \tag{7.5}$$

úgynevezett végtelen *Hankel-mátrix*. A $\varrho(\mathbf{A}_1)<1$ feltétel biztosítja, hogy valóban $l_2\to l_2$ leképezést definiáltunk. Sőt a 4. fejezet alapján az is nyilvánvaló, hogy $R(\chi)$ véges dimenziós, így speciálisan korlátos operátor. Legyen

$$\sigma_H^2=\|\chi\chi^T\|, \tag{7.6}$$

ahol $\|\,.\,\|$ az operátornormát jelöli.

7.1. LEMMA. Adott $\mathbf{A}_1$, $\mathbf{B}_1$, $\mathbf{C}_1$ esetén $(\varrho(\mathbf{A}_1)<1)$ a (7.3), ill. (7.4) egyenletekkel megadott $\bar{\mathbf{y}}$ folyamatra tetszőleges $\mathbf{A}_2$, $\mathbf{B}_2$, $\mathbf{C}_2$, Γ mellett fennáll a

$$\sigma\geqq\sigma_H \tag{7.7}$$

egyenlőtlenség.

Bizonyítás. Tetszőleges $\mathbf{A}_2$, $\mathbf{B}_2$, $\mathbf{C}_2$, Γ választás esetén (ha $\varrho(\mathbf{A}_2^{-1})<1$)

$$\bar{\mathbf{y}}_t=\sum_{k=1}^{\infty}\mathbf{C}_1\mathbf{A}_1^{k-1}\mathbf{B}_1\mathbf{e}_{t-k}-\sum_{k=0}\mathbf{C}_2\mathbf{A}_2^{-(k+1)}\mathbf{B}_2\mathbf{e}_{t+k}+\Gamma\mathbf{e}_t.$$

Legyen $\zeta \in l_2$ olyan végtelen számsorozat, melynek normája 1. Képezzünk ebből R^q-beli vektorok egy $\eta_1, \eta_2, \ldots$ sorozatát, úgy, hogy η_1 koordinátáit ζ első q eleme adja, η_2-ét a következő q, stb... Ha $\{\bar{y}_t, t=0, \pm 1, \ldots\}$ független, $N(0, \sigma^2\mathbf{I})$ eloszlású sorozat, akkor a

$$\xi = \sum_{t=1}^{\infty} \eta_t^T \bar{y}_t$$

valószínűségi változó $N(0, \sigma^2)$ eloszlású, azaz

$$E(\xi^2) = \sigma^2.$$

A ξ valószínűségi változó felírható az $\{e_s, s=0, \pm 1, \ldots\}$ sorozat segítségével. Vegyük ξ felírásából azonban csak azokat a tagokat, melyekben az $e_0, e_{-1}, \ldots$ változók szerepelnek. Ezen tagok összege legyen ξ^-. Ekkor

$$\xi^- = \sum_{t=1}^{\infty} \eta_t^T \sum_{s=0}^{\infty} \mathbf{C}_1 \mathbf{A}_1^{t-s-1} \mathbf{B}_1 e_{-s}, \tag{7.8}$$

és nyilván

$$E(\xi^{-2}) \leqq E(\xi^2) = \sigma^2. \tag{7.9}$$

Másfelől azonban (7.8) alapján

$$E(\xi^{-2}) = \zeta^T \mathbf{H}\mathbf{H}^T \zeta.$$

Tehát tetszőleges $\zeta \in l_2$ egységvektor esetén

$$\sigma^2 \geqq \zeta^T \mathbf{H}\mathbf{H}^T \zeta,$$

így $\sigma^2 \geqq \sigma_H^2$, ami a bizonyítandó (7.7) egyenlőtlenség.

Megjegyzés. Vegyük észre, hogy σ_H értékét közvetlenül is — csak a χ operátor segítségével — definiálhattuk volna. σ_H éppen χ normája, vagy másképpen, a jelen esetben, maximális szinguláris értéke. A szinguláris értékek definiálásához azt kell csak kihasználnunk, hogy mivel a (7.5) mátrix segítségével definiált χ *Hankel-operátor* képtere véges dimenziós, így ha megszorítjuk az operátort a nulltere ortogonális kiegészítő alterére, az $l_2 \ominus N(\chi)$ altérre, akkor már véges dimenziós terek között leképező operátort kapunk és ezek szinguláris értékeit tekinthetjük (lásd TUSNÁDY (1979)). A 4. fejezetből tudjuk, hogy $l_2 \ominus N(\varkappa)$ dimenziója éppen p. Legyenek a nem nulla szinguláris értékek sorban

$$\sigma_1 \geqq \sigma_2 \geqq \ldots \geqq \sigma_p > 0, \tag{7.10}$$

ekkor

$$\sigma_H = \sigma_1.$$

Legyen

$$\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_p).$$

Mostantól fogva feltesszük, hogy a fejezetben szereplő folyamat esetén $q=1$, azaz $\mathbf{e}_t$, ill. $\mathbf{y}_t$ skalár folyamatok. Megmutatjuk, hogy ekkor a $\sigma=\sigma_1$ érték el is érhető. (Lásd ADAMJAN—AROV—KREIN 1(971), ill. a $q \geqq 1$ esetre GLOVER (1984), KUNG—LIU (1981), BALL—HELTON (1983).)

Legyen $\mathbf{P}_1$ $\mathbf{x}_t^1$ kovarianciamátrixa, vagy másképpen a

$$\mathbf{P}_1 - \mathbf{A}_1 \mathbf{P}_1 \mathbf{A}_1^T = \mathbf{B}_1 \mathbf{B}_1^T \tag{7.11}$$

egyenlet megoldása, és $\mathbf{Q}_1$ a

$$\mathbf{Q}_1 - \mathbf{A}_1^T \mathbf{Q}_1 \mathbf{A}_1 = \mathbf{C}_1^T \mathbf{C}_1 \tag{7.12}$$

egyenlet gyöke. Tegyük fel, hogy $\mathbf{P}_1$, $\mathbf{Q}_1$ pozitív definit. Ha az $\mathbf{A}_1$, $\mathbf{B}_1$, $\mathbf{C}_1$ mátrixokat kicseréljük a

$$\mathbf{T}\mathbf{A}_1\mathbf{T}^{-1}, \quad \mathbf{T}\mathbf{B}_1, \quad \mathbf{C}_1\mathbf{T}^{-1}$$

mátrixokra, ahol $\mathbf{T}: R^P \to R^P$ nem elfajuló transzformáció, akkor könnyen látható, hogy $\mathbf{P}_1$ és $\mathbf{Q}_1$ helyett

$$\mathbf{T}\mathbf{P}_1\mathbf{T}^T \quad \text{és} \quad (\mathbf{T}^T)^{-1}\mathbf{Q}_1\mathbf{T}^{-1}$$

lesznek. Így $\mathbf{P}_1\mathbf{Q}_1$ helyébe $\mathbf{T}\mathbf{P}_1\mathbf{Q}_1\mathbf{T}^{-1}$ lép, azaz a $\mathbf{P}_1\mathbf{Q}_1$ szorzat sajátértékei invariánsak a $\mathbf{T}$ transzformációra.
Írjuk fel $\mathbf{Q}_1$-et $\mathbf{Q}_1 = \mathbf{\Phi}\mathbf{\Phi}^T$ alakban.
A $\mathbf{\Phi}^T P_1 \mathbf{\Phi}$ mátrixon végezzünk főtengely transzformációt

$$\mathbf{\Phi}^T \mathbf{P}_1 \mathbf{\Phi} = \psi \Delta^2 \psi^T,$$

és legyen

$$\mathbf{T} = \Delta^{-1/2} \psi^T \mathbf{\Phi}^T.$$

Ekkor

$$\mathbf{T}\mathbf{P}_1\mathbf{T}^T = \Delta,$$

és

$$(\mathbf{T}^T)^{-1}\mathbf{Q}_1\mathbf{T}^{-1} = \Delta.$$

Azaz feltehetjük, hogy

$$\mathbf{P}_1 = \mathbf{Q}_1 = \Delta \tag{7.12}$$

diagonális mátrix. Azonban vegyük észre, hogy az

$$\mathbf{M} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_1\mathbf{A}_1 \\ \mathbf{C}_1\mathbf{A}_1^2 \\ \vdots \end{bmatrix}, \quad \mathbf{V} = [\mathbf{B}_1, \mathbf{A}_1\mathbf{B}_1, \ldots]$$

jelöléseket felhasználva

$$\mathbf{P}_1 = \mathbf{V}\mathbf{V}^T, \quad \mathbf{Q}_1 = \mathbf{M}^T\mathbf{M},$$

és ugyanakkor

$$\mathbf{H} = \mathbf{M}\mathbf{V}.$$

Így

$$\mathbf{H}\mathbf{H}^T = \mathbf{M}\mathbf{V}\mathbf{V}^T\mathbf{M}^T,$$

ezért $\mathbf{H}\mathbf{H}^T$ sajátértékei megegyeznek a $\mathbf{V}\mathbf{V}^T\mathbf{M}^T\mathbf{M}$ mátrix sajátértékeivel, tehát $\mathbf{P}_1\mathbf{Q}_1$ sajátértékeivel. Másképpen $\mathbf{P}_1\mathbf{Q}_1$ sajátértékei $\mathbf{H}$ szinguláris értékeinek négyzetei. Tehát

$$\Delta = \Sigma. \tag{7.13}$$

Definiáljuk megint a $\{\xi_t^1, t=0, \pm 1, \ldots\}$ folyamatot:

$$\xi_t^1 = \mathbf{A}_1^T \xi_{t+1}^1 + \mathbf{C}_1^T \mathbf{e}_t.$$

Ekkor

$$\mathbf{Q}_1 = \Sigma = E(\xi_t^1 \xi_t^{1T}),$$

és

$$\mathbf{P}_1 = \Sigma = E(\mathbf{x}_t^1 \mathbf{x}_t^{1T}).$$

Jelölje E az

$$\eta = \Sigma \mathbf{a}_t \mathbf{e}_t$$

alakú változók által generált zárt lineáris alteret, vagy másképpen az $\{\mathbf{e}_t, t=0, \pm 1, \ldots\}$ változók által kifeszített zárt alteret, és legyen E_s^-, ill. E_s^+ az $\{\mathbf{e}_t, t \leqq s\}$, ill. $\{\mathbf{e}_t, t \geqq s\}$ változók által generált zárt altér. Definiálunk egy

(7.14) $$\chi\colon E_0^+ \to E_{-1}^-$$

operátort. Legyen

(7.15) $$\chi \mathbf{e}_t = \sum_{s=1}^{\infty} \mathbf{C}_1 \mathbf{A}_1^{t+s-1} \mathbf{B}_1 \mathbf{e}_{-s}, \quad t = 0, 1, \ldots$$

(Bár a χ jelölést korábban már használtuk, ott $\chi\colon l_2 \to l_2$ operátor volt, azonban a két operátor mátrixa megfelelő bázisban megegyezik, és az operátor argumentuma alapján — hogy milyen objektumra alkalmazzuk — majd mindig lehet azonosítani, hogy melyik χ operátorról van éppen szó.) Legyen

$$S\colon E \to E$$

a léptető operátor, azaz

(7.16) $$S\mathbf{e}_t = \mathbf{e}_{t+1}, \quad t = 0, \pm 1, \pm 2, \ldots .$$

Vegyük észre, hogy tetszőleges $\eta \in E_0^+$ esetén $S\eta \in E_0^+$ és

(7.17) $$\chi S \eta = \mathrm{Pr}_{E_{-1}^-} S \chi \eta.$$

Ezt az összefüggést közvetlenül ellenőrizhetjük E_0^+ generátor elemein, az $\{\mathbf{e}_t, t=0, \pm 1, \ldots\}$ vektorokon. A (7.17) összefüggésben a merőleges vetítés operátorát Pr. jelöli. A továbbiakban helyette a feltételes várható értékre megszokott jelölést fogjuk használni — normális eloszlású változók esetén ez valóban a generált zárt lineáris altérre való vetítéssel esik egybe. Tehát (7.17) így írható

(7.17′) $$\chi S \eta = E(S \chi \eta | E_{-1}^-).$$

A ξ_t^1, ill. $\mathbf{x}_t^1$ p-dimenziós változók koordináta változóit jelölje $\xi_t^1(k)$, $\mathbf{x}_t(k)$, $k=1, \ldots, p$. Vajon mivel egyenlő $\chi \xi_0^1(k)$? Vagy a szokásos, kissé pongyola jelöléssel, mivel egyenlő $\chi \xi_0^1$?
Mivel

$$\xi_0^1 = \sum_{t=0}^{\infty} (\mathbf{A}_1^T)^t \mathbf{C}_1^T \mathbf{e}_t,$$

így

$$\chi \xi_0^1 = \sum_{t=0}^{\infty} (\mathbf{A}_1^T)^t \mathbf{C}_1^T \chi \mathbf{e}_t = \sum_{t=0}^{\infty} (\mathbf{A}_1^T)^t \mathbf{C}_1^T \sum_{s=1}^{\infty} \mathbf{C}_1 \mathbf{A}_1^{t+s-1} \mathbf{B}_1 \mathbf{e}_{-s} = \mathbf{Q}_1 \mathbf{x}_0^1 = \Sigma \mathbf{x}_0^1.$$

Tehát

(7.18) $$\chi \xi_0^1(k) = \sigma_k \mathbf{x}_0^1(k), \quad k = 1, \ldots, p.$$

Ugyanígy

$$\chi^T \mathbf{x}_0^1(k) = \sigma_k \xi_0^1(k), \quad k = 1, \ldots, p.$$

(Ebből egyébként látszik, hogy ha a $\xi_0^1(t)$, $k=1, \ldots, p$, $\mathbf{x}_0^1(k)$, $k=1, \ldots, p$ változókat normáljuk, akkor éppen ezen χ operátor szinguláris vektorait kapjuk, tehát $\frac{1}{\sigma_k}\xi_0^1(k)$, $k=1, \ldots, p$, és $\frac{1}{\sigma_k}\mathbf{x}_0^1(k)$, $k=1, \ldots, p$ a szinguláris vektorrendszer.)

Mivel $q=1$ esetben a χ operátor $\mathbf{H}$ mátrixa szimmetrikus, ezért a

$$V\colon E \to E$$

$$V\mathbf{e}_t = \mathbf{e}_{1-t}, \quad t = 0, \pm 1, \pm 2, \ldots \tag{7.19}$$

operátor bevezetésével a

$$V\xi_0^1(k) = \mathbf{x}_0^1(k), \quad k = 1, \ldots, p$$

adódik.

Újabb alterekre van szükségünk. A $\Xi_s^+(k)$ alteret generálják a $\{\xi_t^1(k), t \geqq s\}$ változók, a $\Xi_s^-(k)$ alteret a $\{\xi_t^1(k), t \leqq s\}$, az $X_s^+(k)$ alteret az $\{\mathbf{x}_t^1(k), t \geqq s\}$, az $X_s^-(k)$ alteret pedig az $\{\mathbf{x}_t^1(k), t \leqq s\}$, valószínűségi változók. Legyenek ezek is zárt alterek. Nyilván

$$\begin{aligned} &\Xi_0^+(k) \subset E_0^+, \quad k = 1, \ldots, p, \\ &X_0^-(k) \subset E_{-1}^-, \quad k = 1, \ldots, p. \end{aligned} \tag{7.20}$$

Vizsgáljuk először a $k=1$ esetet. Tegyük fel, hogy

$$\Xi_0^+(1) \neq E_0^+.$$

Válasszunk orotonormált bázist a $\Xi_0^+(1)$ altérben, és pedig éppen azt, amit az

$$\eta_t = \xi_t^1(1) - \mathrm{Pr}_{\Xi_{t+1}^+(1)} \xi_t^1(1), \quad t = 0, 1, \ldots$$

vektorok normálásával nyerünk. Jelölje ezen bázis elemeit

$$\zeta_0, \zeta_1, \ldots,$$

(azaz η_t és ζ_t kollineárisak). A $\xi_0^1(1)$ nyilván felírható $\zeta_0, \zeta_1, \ldots$ segítségével és a $\Xi_1^+(1) \subset E_0^+$ reláció alapján, ζ_0 kifejezhető az $\mathbf{e}_0, \mathbf{e}_1, \ldots$ mennyiségekkel. Legyen

$$\xi_0^1(1) = \sum_{t=0}^{\infty} \alpha_t \zeta_t, \tag{7.21}$$

$$\zeta_0 = \sum_{t=0}^{\infty} \beta_t \mathbf{e}_t. \tag{7.22}$$

Ekkor nyilván a

$$\xi_0^1(s) = \sum_{t=0}^{\infty} \alpha_t \zeta_{t+s}, \tag{7.23}$$

és a

$$\zeta_s = \sum_{t=0}^{\infty} \beta_t \mathbf{e}_{t+s} \tag{7.24}$$

összefüggések is teljesülnek.

Legyen

$$\psi = \sum_{t=0}^{\infty} \alpha_t \mathbf{e}_t, \tag{7.25}$$

ill.

$$\psi_s = \sum_{t=0}^{\infty} \alpha_t \mathbf{e}_{t+s}, \quad s \geqq 0 \quad (\psi = \psi_0). \tag{7.26}$$

Nyilván $E(\psi^2)=E([\xi_0^1(1)]^2=\sigma_1^2$ és próbáljuk meghatározni $\chi\psi$ értékét. (7.23), (7.24) és (7.25) alapján

$$\xi_0^1(1) = \sum_{t=0}^{\infty} \alpha_t \sum_{s=0}^{\infty} \beta_s \mathbf{e}_{t+s} = \sum_{s=0}^{\infty} \beta_s \sum_{t=0}^{\infty} \alpha_t \mathbf{e}_{t+s} = \sum_{s=0} \beta_s \psi_s. \tag{7.27}$$

Legyen

$$\varphi = \chi\psi,$$

(7.27) miatt

$$\chi\xi_0^1(1) = \sum_{s=0}^{\infty} \beta_s \chi\psi_s = \sum_{s=0}^{\infty} \beta_s \operatorname{Pr}_{E_{-1}} S^s\varphi = \operatorname{Pr}_{E_{-1}} \sum_{s=0} \beta_s S^s\varphi.$$

Számoljuk ki a $\sum_{s=0}^{\infty} \beta_s S^s\varphi$ vektorok normáját! Legyen $\varrho^2=E(\varphi^2)$, akkor

$$E\Big(\Big(\sum_{s=0}^{\infty} \beta_s S^s\varphi\Big)^2\Big) = \sum_{s=0}^{\infty} \beta_s^2 E\big((S^s\varphi)^2\big) + 2\sum_{t=1}^{\infty}\sum_{s=0}^{\infty} \beta_s\beta_{s+t} E\big((S^s\varphi)(S^{s+t}\varphi)\big).$$

Azonban

$$E\big((S^s\varphi)(S^{s+t}\varphi)\big) = E\big(\varphi(S^t\varphi)\big),$$

és (7.24) alapján

$$\sum_{s=0}^{\infty} \beta_s\beta_{s+t} = 0, \quad \text{ha} \quad t \geqq 1,$$

ill.

$$\sum_{s=0}^{\infty} \beta_s^2 = 1.$$

Így

$$E\Big(\Big(\sum_{s=0}^{\infty} \beta_0 S^s\varphi\Big)^2\Big) = \varrho^2.$$

Ugyanakkor

$$\sigma_1 \mathbf{x}_0^1(1) = \chi\xi_0^1(1).$$

Így

$$\varrho^2 \geqq \sigma_1^4.$$

De az $E(\psi^2)=\sigma_1^2$, $\varphi=\chi\psi$, $E(\varphi^2)=\varrho^2$ összefüggések miatt, mivel $\|\chi\|=\sigma_1$, így ezért a $\varrho=\sigma_1^2$ összefüggésnek kell teljesülnie. Ez azt jelenti, hogy ψ benne van a σ_1 szinguláris értékhez tartozó szinguláris vektorok által kifeszített altérben és így alkalmas bázis transzformációval elérhetjük, hogy ψ a ξ_0^1 vektor transzformáltjának első koordinátája legyen. Mivel az $\{\alpha_t,\ t=0, 1, \ldots\}$ sorozat definíciója alapján a $\psi_0, \psi_1, \ldots$ sorozat generálja az E_0^+ alteret, így feltehetjük eleve, hogy

$$\Xi_0^+(1) = E_0^+. \tag{7.28}$$

Most már bebizonyíthatjuk a következő tételt.

7.1. TÉTEL. Legyen

$$\mathbf{x}_{t+1}^1 = \mathbf{A}_1\mathbf{x}_t^1 + \mathbf{B}_1\mathbf{e}_t,$$

$$\mathbf{y}_t = \mathbf{C}_1\mathbf{x}_t^1,$$

tegyük fel, hogy $\mathbf{P}_1 = \mathbf{Q}_1 = \mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_p)$, $\sigma_1 \geqq \sigma_2 \geqq \ldots \geqq \sigma_p > 0$, és legyen $\mathbf{e}_t$, $\mathbf{y}_t$ skalárértékű. Ekkor léteznek olyan $\mathbf{A}_2$, $\mathbf{B}_2$, $\mathbf{C}_2$, $\mathbf{\Gamma}$ mátrixok, hogy az

$$\mathbf{x}_t^2 = \mathbf{A}_2^{-1}\mathbf{x}_{t+1}^2 - \mathbf{A}_2^{-1}\mathbf{B}_2\mathbf{e}_t,$$

$$\bar{\mathbf{y}}_t = \mathbf{C}_1\mathbf{x}_t^1 + \mathbf{C}_2\mathbf{x}_t^2 + \mathbf{\Gamma}\mathbf{e}_t$$

egyenletekkel meghatározott $\{\bar{\mathbf{y}}_t, t = 0, \pm 1, \ldots\}$ sorozat független, $N(0, \sigma_1^2)$ eloszlású valószínűségi változókból áll.

Bizonyítás. Számoljuk ki a

$$\xi_t^1 = \mathbf{A}_1^T\xi_{t+1}^1 + \mathbf{C}_1^T\mathbf{e}_t,$$

és az $\mathbf{x}_t^1$ folyamatok kovariancia függvényeit.

$$E(\mathbf{x}_{t+k}^1(\mathbf{x}_t^1)^T) = \mathbf{A}_1^k\mathbf{\Sigma}, \quad k \geqq 0, \tag{7.29}$$

$$E(\xi_{t+k}^1(\xi_t^1)^T) = \mathbf{\Sigma}\mathbf{A}_1^k, \quad k \geqq 0. \tag{7.30}$$

Vegyük észre, hogy $A_1^k\mathbf{\Sigma}$ és $\mathbf{\Sigma}A_1^k$ diagonális elemei megegyeznek, sőt

$$\sigma_1 \geqq \ldots \geqq \sigma_j > \sigma_{j+1} = \ldots = \sigma_{j+r} > \sigma_{j+r+1} \geqq \ldots \geqq \sigma_p > 0$$

esetén a $[(j+1):(j+r)]\times[(j+1):(j+r)]$-es blokkok is. Tehát speciálisan

$$E(\xi_{t+k}^1(1)\xi_t^1(1)) = E(\mathbf{x}_{t+k}^1(1)\mathbf{x}_t^1(1)). \tag{7.31}$$

Ezért az a

$$T: E \to E$$

leképezés, melyre

$$T\xi_t^1(1) = \mathbf{x}_t^1(1), \quad t = 0, \pm 1, \ldots,$$

értelmesen definiált és izometria. (Mivel $\Xi_0^+(1) = E_0^+$, így T az egész téren definiált.) Legyen

$$\bar{\mathbf{y}}_t = \sigma_1 T\mathbf{e}_t, \quad t = 0, \pm 1, \ldots.$$

Megmutatjuk, hogy $\bar{\mathbf{y}}_t$ rendelkezik a kívánt tulajdonságokkal.

T izometria, így $\{\bar{\mathbf{y}}_t, t = 0, \pm 1, \ldots\}$ független $N(0, \sigma_1^2)$ eloszlású valószínűségi változókból álló sorozat. Mivel $\bar{\mathbf{y}}_t \in E$ így léteznek $\{\gamma_s, s = 0, \pm 1, \ldots\}$ számok, hogy

$$\bar{\mathbf{y}}_0 = \sum_{s=-\infty}^{\infty} \gamma_s\mathbf{e}_s.$$

(T definíciója alapján ekkor $\bar{\mathbf{y}}_t = \sum_{s=-\infty}^{\infty} \gamma_s\mathbf{e}_{s+t}$.) Vegyük észre, hogy tetszőleges $\eta \in E$ esetén

$$\mathbf{TS}\eta = \mathbf{ST}\eta,$$

és ugyanakkor

$$\mathbf{T}\xi_0^1(1) = \mathbf{x}_0^1(1) = \frac{1}{\sigma_1}\chi\xi_0^1(1).$$

Tehát

$$\mathbf{T}\xi_0^1(t) = \frac{1}{\sigma_1}S^t\chi\xi_0^1(1), \quad t \geqq 0.$$

Így

$$E(\mathbf{T}\xi_0^1(t)|E_{-1}^-) = \frac{1}{\sigma_1}\chi\xi_0^1(t).$$

Azaz az E altéren $\chi = \sigma_1 \Pr_{E_{-1}^-} T$.
Speciálisan az $e_0 \in E_0^+$ vektor esetén

$$E(\mathbf{y}_t|E_{-1}^-) = \sum_{s=1}^{\infty} \mathbf{C}_1\mathbf{A}_1^{s-1}\mathbf{B}_1\mathbf{e}_{-s}.$$

Tehát

$$E(\mathbf{y}_0|E_{-1}^-) = \mathbf{C}_1\mathbf{x}_0^1 = \mathbf{y}_0.$$

Annak igazolása van hátra, hogy az $\{\bar{\mathbf{y}}_t - \mathbf{y}_t,\ t = 0, \pm 1, \ldots\}$ folyamatnak van véges dimenziós realizációja. De ez nyilvánvaló, hiszen az $\bar{\mathbf{y}}_t$ és $\mathbf{y}_t$ folyamatoknak külön-külön van véges dimenziós realizációjuk, így a különbségüknek is van.

Megjegyzés. Vegyük észre, hogy mivel $\mathbf{e}_t$ a $\xi_t^1(1) - E(\xi_t^1(1)|\Xi_{t+1}^+(1))$ változó konstansszorosa, így $\bar{\mathbf{y}}_t$ az $\mathbf{x}_t^1(1) - E(\mathbf{x}_t^1(1)|X_{t+1}^+(1))$ konstansszorosa lesz, azaz ha időben megfordítva vesszük az $\{\mathbf{x}_t^1(1),\ t = 0, \pm 1, \pm 2, \ldots\}$ folyamatot, és az így kapott folyamat innovációját tekintjük, akkor $\mathbf{y}_t$ konstansszorosához jutunk.

8. Az állapottér dimenziójának a csökkentése

Ebben a fejezetben a 7. fejezetben definiált fogalmakat fogjuk felhasználni annak a problémának a megoldására, hogyan lehet egy χ p-rangú *Hankel-operátort* (ill. a $\mathbf{H}$ *Hankel-mátrixot*) operátornormában közelíteni legfeljebb k-rangú *Hankel-operátorral (Hankel-mátrixszal)*.

8.1. LEMMA. Tetszőleges $\mathcal{G}: E^+ \to E_{-1}^+$ legfeljebb k-rangú operátor esetén

$$\|\chi - \mathcal{G}\| \geqq \sigma_{k+1}. \tag{8.1}$$

Bizonyítás. (SONNEVEND (1987).) Mivel $\mathcal{G}$ legfeljebb k-rangú, így

$$\dim\left(E_0^+ \ominus \mathcal{N}(\mathcal{G})\right) \leqq k,$$

tehát létezik olyan

$$\eta = \sum_{i=1}^{k+1} \alpha_i \xi_0^1(i)$$

egységnormájú vektor, melyre

$$\mathcal{G}\eta = \mathbf{O}.$$

Ekkor

$$(\chi - \mathcal{G})\eta = \chi\eta = \sum_{i=1}^{k+1} \alpha_i\sigma_i\mathbf{x}_0^1(i).$$

Így

$$\|(\chi-\mathcal{G})\eta\|^2 = \sum_{i=1}^{k+1} \alpha_i^2\sigma_i^2 \leqq \sigma_{k+1}^2.$$

Sőt, itt egyenlőség csak akkor lehet, ha $\sigma_i > \sigma_{k+1}$ esetén $\alpha_i = 0$, $1 \leqq i \leqq k$. Így speciálisan $\sigma_k > \sigma_{k+1}$ esetén, ahhoz, hogy a $\|\chi-\mathcal{G}\| = \sigma_{k+1}$ egyenlőség teljesüljön, szükséges a $\mathcal{G}\xi_0^1(k+1) = 0$ feltétel.

8.1. TÉTEL. Tegyük fel, hogy $\sigma_k > \sigma_{k+1} > \sigma_{k+2}$. Ekkor létezik olyan

$$\chi_k\colon E_0^+ \to E_{-1}^+$$

Hankel-operátor, melyre

$$\dim R(\chi_k) \leqq k,$$

és

$$\|\chi-\chi_k\| = \sigma_{k+1}.$$

Bizonyítás: (7.29) és (7.30) alapján

$$E\big(\mathbf{x}_{t+j}^1(k+1)\mathbf{x}_t^1(k+1)\big) = E\big(\xi_{t+j}^1(k+1)\xi_t^1(k+1)\big), \quad j \geqq 0. \tag{8.2}$$

Definiáljuk a

$$T_k\colon E \to E$$

operátort a következőképpen:

$$T_k\xi_t^1(k+1) = \mathbf{x}_t^1(k+1), \quad t = 0, \pm 1, \pm 2, \ldots \tag{8.3}$$

T_k ismét izometria (és megmutatható, hogy az egész E-n definiált). A χ_k *Hankel-operátor* legyen a következő:

$$\chi_k\eta = \chi\eta - \sigma_{k+1}E(T_k\eta|E_{-1}^-), \quad \eta \in E_0^+. \tag{8.4}$$

Ekkor

$$\chi_k S\eta = E(S\varkappa_k\eta|E_{-1}^-), \tag{8.5}$$

így χ_k tényleg *Hankel-operátor*. Ugyanakkor

$$\chi-\chi_k = \sigma_{k+1}E(T_k|E_{-1}),$$

így

$$\|\chi-\chi_k\| \leqq \sigma_{k+1}. \tag{8.6}$$

Vegyük észre, hogy

$$\chi\xi_0^1(k+1) = \sigma_{k+1}\mathbf{x}_0^1(k+1) = \sigma_{k+1}E(T_k\xi_0^1(k+1)|E_{-1}^-),$$

tehát

$$\chi_k\xi_0^1(k+1) = 0. \tag{8.7}$$

Így azonban (8.5) alapján $\chi_k\xi_t^1(k+1) = 0$, $t \geqq 0$ esetén, azaz

$$\Xi_0^+(k+1) \subset \mathcal{N}(\chi_k).$$

Ugyanakkor a 8.1. lemma alapján $\dim R(\chi_k) \geqq k$. Sőt, vegyük észre, hogy tetszőleges

$$\eta = \sum_{i=1}^{k} \alpha_i\xi_0^1(i), \quad \eta \neq 0$$

esetén, $\eta \notin \mathcal{N}(\chi_k)$.

Megmutatjuk, hogy

$$\dim R(\chi_k) = k, \tag{8.8}$$

pontosabban azt fogjuk igazolni, hogy

$$\dim \left(E_0^+ \ominus \mathcal{N}(\chi_k)\right) = k.$$

Ehhez néhány apró lemmára lesz szükségünk.

8.2. LEMMA. (ADAMJAN—AROV—KREIN (1971)). Ha

$$\Phi\colon E \to E$$

izometria, melyre

$$\Phi(E_0^+) \subset E_0^+,$$

és

$$\Phi S = S\Phi,$$

akkor a $\varkappa\Phi$ operátor *Hankel-operátor,* melynek szinguláris értékei rendre legfeljebb akkorák mint H megfelelő szinguláris értékei.

Bizonyítás. Mivel

$$\varkappa\Phi S = \varkappa S\Phi = E(S\varkappa\Phi|E_{-1}^-),$$

így $\varkappa\Phi$ *Hankel-operátor.* Legyen

$$\Phi\mathbf{e}_0 = \sum_{i=0}^{\infty} \alpha_i \mathbf{e}_i.$$

Mivel $S\Phi = \Phi S$, így

$$\mathbf{e}_t = \sum_{i=0}^{\infty} \alpha_i \mathbf{e}_{t+i}, \quad t = 0, \pm 1, \ldots$$

s ezért $\eta \in E_0^+$ esetén

$$\Phi\eta = \sum_{i=0}^{\infty} \alpha_i S^i \eta.$$

Legyen most $\eta \in E_0^+$. Ekkor

$$\chi\Phi\eta = \chi \sum_{i=0}^{\infty} \alpha_i S^i \eta = \sum_{i=0}^{\infty} \alpha_i E(S^i \chi\eta|E_{-1}^-) = \tag{8.9}$$

$$= E\Big(\sum_{i=0}^{\infty} \alpha_i S^i \chi\eta|E_{-1}^-\Big) = E(\Phi H\eta|E_{-1}^-).$$

Mivel Φ izometria, így

$$\|\Phi\chi\eta\| = \|\chi\eta\|,$$

ezért

$$\|\chi\Phi\eta\| \leqq \|\chi\eta\|, \quad \eta \in E_0^+.$$

Ebből adódik, hogy $\chi\Phi$ szinguláris értékei rendre legfeljebb akkorák, mint χ szinguláris értékei.

8.3. LEMMA. Legyen

$$\zeta_0 \in \Xi_0^+(k+1) \ominus \Xi_1^+(k+1), \quad \|\zeta_0\| = 1,$$

és

$$\zeta_t = S^t \zeta_0. \quad t = 0, \pm 1, \ldots$$

Definiáljuk a $\Phi: E \to E$ leképezést a következőképpen:

$$\Phi \mathbf{e}_t = \zeta_t.$$

Ekkor

$$\|\chi\Phi\| = \sigma_{k+1}. \tag{8.10}$$

Bizonyítás. A Φ operátor eleget tesz a 8.2. lemma követelményeinek, így speciálisan $\chi\Phi$ *Hankel-operátor*, és Φ izometria. Mivel

$$\Phi(E_0^+) = \Xi_0^+(k+1),$$

és a $\Xi_0^+(k+1)$ alteret a $\xi_t(k+1)$, $t=0, 1, \ldots$ vektorok generálják, így

$$\|\chi\Phi\| \leqq \sigma_{k+1},$$

ugyanakkor

$$\chi\Phi\Phi^{-1}\xi_0^1(k+1) = \sigma_{k+1}\mathbf{x}_0^1(k+1),$$

tehát

$$\|\chi\Phi\| = \sigma_{k+1}.$$

8.4. LEMMA. Legyen $m = \dim\left(E_0^+ \ominus \Xi_0^+(k+1)\right)$. Ekkor a $\varkappa\Phi$ operátornak σ_{k+1} legalább $(m+1)$-szeres szinguláris értéke.

Bizonyítás. Legyen $\eta \in E_0^+$.
Legyen Λ_0 az $\eta, S\eta, \ldots$ és $\Xi_0^+(k+1)$ elemei által generált zárt altér, ill. jelölje Λ_t az $S^t\Lambda_0$ alteret. Ekkor nyilván

$$\Xi_0^+(k+1) \subset \Lambda_0 \subset E_0^+, \tag{8.11}$$

és

$$S\Lambda_0 \subset \Lambda_0.$$

Legyen

$$\varphi_0 \in \Lambda_0 \ominus \Lambda_1, \quad \|\varphi_0\| = 1,$$

és

$$\varphi_t = S^t \varphi_0.$$

Definiáljuk a

$$\psi: E \to E,$$

operátort a következőképpen

$$\psi \mathbf{e}_t = \varphi_t. \tag{8.12}$$

Nyilván ψ izometria, és

$$\psi E_0^+ = \Lambda_0 \subset E_0^+. \tag{8.13}$$

Végezetül legyen

$$\tilde{\psi} = V\psi^{-1}V, \tag{8.14}$$

ahol V a (7.19)-ben definiált operátor.

Azaz speciálisan

(8.15) $$\tilde{\psi}(V\zeta_t) = Ve_t = \mathbf{e}_{1-t},$$

(8.16) $$\tilde{\psi}(V\Lambda_0) = VE_0^+ = E_{-1}^-.$$

Azonban $\Lambda_0 \subset E_0^+$ miatt $V\Lambda_0$ elemei merőlegesek E_0^+ elemeire, így

$$\tilde{\psi}(V\Lambda_0) \perp \tilde{\psi}(E_0^+).$$

Ezért (8.16) alapján

$$\tilde{\psi}(E_0^+) \subset E_0^+,$$

(8.14) alapján

$$\tilde{\psi}S = S\tilde{\psi},$$

ezért (8.9)-ből adódik, hogy

(8.17) $$\chi\Phi\tilde{\psi} = E(\tilde{\psi}\chi\Phi|E_{-1}^-).$$

Határozzuk meg a

$$\chi\Phi\left(\tilde{\psi}\Phi^{-1}\xi_0^1(k+1)\right)$$

vektor normáját

$$H\Phi\tilde{\psi}\Phi^{-1}\xi_0^1(k+1) = E\left(\tilde{\psi}\chi\Phi\Phi^{-1}\xi_0^1(k+1)|E_{-1}^-\right) =$$

$$= E\left(\tilde{\psi}\chi\xi_0^1(k+1)|E_{-1}^-\right) = E\left(\tilde{\psi}\sigma_{k+1}\mathbf{x}_0^1(k+1)|E_{-1}^-\right).$$

Másfelől

$$\tilde{\psi}\mathbf{x}_0^1(k+1) = V\psi^{-1}V\mathbf{x}_0^1(k+1) = V\psi^{-1}\xi_0^1(k+1).$$

De

$$\psi^{-1}\xi_0^1(k+1) \in E_0^+,$$

így

$$\tilde{\psi}\mathbf{x}_0^1(k+1) \in E_{-1}^-,$$

tehát

$$\chi\Phi\left(\tilde{\psi}\Phi^{-1}\xi_0^1(k+1)\right) = \sigma_{k+1}\tilde{\psi}\mathbf{x}_0^1(k+1).$$

Azaz a $\tilde{\psi}\Phi^{-1}\xi_0^1(k+1)$ vektoron $\chi\Phi$ felveszi a normáját. Ezért ha $m = \dim[E_0^+ \ominus \Xi_0^+(k+1)]$, akkor tudunk $m+1$ darab lineárisan független vektort konstruálni, melyeken a $\chi\Phi$ operátor felveszi a normáját, tehát σ_{k+1} a $\chi\Phi$ legalább $m+1$-szeres szinguláris értéke.

Vegyük észre, hogy ekkor a 8.2. lemma alapján $\sigma_{k+1} \leqq \sigma_{m+1}$, azaz $k \geqq m$.

Térjünk vissza a 8.1 tétel bizonyításához.

Már tudjuk, hogy

$$\Xi_0^+(k+1) \subset \mathcal{N}(\chi_k).$$

Így

$$\dim\left(E_0^+ \ominus \mathcal{N}(\chi_k)\right) \leqq \dim\left(E_0^+ \ominus \Xi^+(k+1) = m \leqq k.\right.$$

Tehát H_k rangja legfeljebb k.

Összevetve ezt a 8.1. lemmával, adódik, hogy

(8.18) $$\dim R(\chi_k) = k.$$

Végezetül vegyük észre, hogy a ξ_t^1 folyamat definíciójában szereplő $\mathbf{e}_t$ valószínűségi változó sorozatot tetszőleges más független $N(0, 1)$ eloszlású sorozatra kicserélhetnénk, hiszen a 6., 7., 8. fejezetben követett gondolatmeneteinkben sohasem játszottak szerepet az $E(\xi_t^1 \mathbf{x}_s^{1T})$ kovarianciák. Azaz definiálhattuk volna a ξ_t folyamatot a

(8.19) $$\xi_{t+1}^1 = \mathbf{A}^T\xi_t^1 + \mathbf{C}^T\mathbf{e}_t$$

egyenlettel, vagy más $\{\mathbf{f}_t, t=0, \pm 1, \ldots\}$ sorozatot használva pl. a

$$\xi_t^1 = \mathbf{A}^T \xi_{t+1}^1 + \mathbf{C}^T \mathbf{f}_t \tag{8.20}$$

összefüggéssel. Pusztán kényelmi okokból haszháltuk $\mathbf{f}_t$ helyett az $\mathbf{e}_t$ sorozatot és a

$$\xi_t^1 = \mathbf{A}^T \xi_{t+1}^1 + \mathbf{C}^T \mathbf{e}_t$$

egyenletet.

Vegyük észre, hogy a (7.19) választással $q=1$, $\mathbf{P}_1 = \mathbf{Q}_1 = \Xi$ esetén a $\sigma_1 > \sigma_2 > \ldots > \ldots > \sigma_n > 0$ feltétel mellett a

$$\xi_0^1 = \pm \mathbf{x}_0^1$$

összefüggés adódnék, hiszen ekkor

$$\chi\colon E_{-1}^- \to E_{-1}^-,$$

és $\varkappa$-nak sajátvektorai a $\xi_0^1(k)$, $k=1, \ldots, p$ vektorok.

9. Kanonikus korrelációk

Ebben a fejezetben *Hankel-mátrixok* szinguláris értékeit kanonikus korrelációkkal hozzuk kapcsolatba.

Legyen $\{y_t, t=0, \pm 1, \ldots\}$ normális eloszlású skalár értékű valószínűségi változókból álló stacionárius folyamat. Tegyük fel, hogy létezik véges dimenziós realizációja. Készítsük el a 6. fejezet eredményei alapján előre, ill. hátrafelé haladó (progresszív és regresszív) leírását, azonban kiinduló pontként az

$$E(y_{t+k} y_t), \quad k \geqq 1$$

kovarianciákból építsük fel a faktorizálandó *Hankel-mátrixot*, az

$$E(y_t y_t)$$

kovarianciát ne vegyük bele. Ekkor a következő folyamathoz jutunk

$$\begin{aligned} x_t &= \mathbf{A} x_{t-1} + \mathbf{K}_+ e_t, \\ y_t &= \mathbf{C} x_{t-1} + \mathbf{D}_+ e_t, \\ z_t &= \mathbf{A}^T z_{t+1} + \mathbf{K}_- f_t, \\ y_t &= \mathbf{B}^T z_{t+1} + \mathbf{D}_- f_t, \end{aligned} \tag{9.1}$$

ahol $\{e_t, t=0, \pm 1, \ldots\}$ és $\{f_t, t=0, \pm 1, \ldots\}$ független, $N(0, 1)$ eloszlású változókból álló folyamat, melyek az $\{y_t\}$ folyamat előre, ill. hátrafelé haladó innovációjának egységnyire normált értékei, azaz pontosabban, ha y_t^- jelöli az $y_t, y_{t-1}, \ldots$ változók által generált zárt alteret, y_t^+ pedig az $y_t y_{t+1}, \ldots$ változók által generáltat, akkor az

$$y_t - E(y_t | \mathbf{Y}_{t-1}^-), \quad t = 0, \pm 1, \ldots$$

és az

$$y_t - E(y_t | Y_{t+1}), \quad t = 0, \pm 1, \ldots$$

valószínűségi változók konstansszorosa e_t, ill. f_t. Legyen

$$\mathbf{P} = E(x_t x_t^T),$$

$$Q = E(z_t z_t^T),$$

P, **Q** pozitív definit mátrixok. A fenti képletekben szereplő mátrixokra a következő összefüggések teljesülnek:

$$\varrho(\mathbf{A}) < 1,$$

$$\mathbf{P} = \mathbf{APA}^T + \mathbf{K}_+ \mathbf{K}_+^T, \tag{9.2}$$

$$\mathbf{Q} = \mathbf{A}^T\mathbf{QA} + \mathbf{K}_- \mathbf{K}_-^T,$$

$$E(y_{t+k} y_t) = \mathbf{CA}^{k-1}\mathbf{B}, \quad k \geqq 1,$$

$$\mathbf{K}_+ \mathbf{D}_+ = \mathbf{B} - \mathbf{APC}^T,$$

$$\mathbf{K}_- \mathbf{D}_- = \mathbf{C}^T - \mathbf{A}^T\mathbf{QB},$$

$$E(x_t z_{t+1}^T) = PQ.$$

Könnyű kiszámolni, hogy ekkor

$$E(f_t x_s^T) = \mathbf{K}^T\mathbf{A}^{t-s-1}\mathbf{P}, \quad t > s, \tag{9.3}$$

$$E(f_t z_s^T) = \mathbf{K}^T\mathbf{A}^{t-s}, \quad t \geqq s,$$

ill.

$$E(x_t z_{t+k}^T) = \mathbf{PA}^{Tk-1}\mathbf{Q}, \quad k \geqq 1. \tag{9.4}$$

Vegyük az Y_0^- és az Y_1^+ altereket.

9.1. LEMMA. Az Y_0^- és az Y_1^+ altérbe tartozó valószínűségi változók kanonikus korreláció a $\mathbf{K}_-^T\mathbf{A}^{t-1}\mathbf{K}_+$, $t \geqq 1$ mátrixokból felépített *Hankel-mátrix* szinguláris értékeivel esnek egybe.

Bizonyítás. A kanonikus korrelációk a vetítés operátor szinguláris értékei. Azt kell tehát megmutatnunk, hogy az Y_1^+-ből Y_0^--ra való vetítés operátora alkalmas ortonormált bázisban éppen a kívánt alakú.

Az $\{e_t, t=0, -1, \ldots\}$ és az $\{f_t, t=1, 2, \ldots\}$ vektorok bázist alkotnak az Y_0^-, ill. Y_1^+ altérben. Írjuk fel ebben a bázisban a vetítés operátorát, azaz számoljuk ki

$$E(f_t | Y_0^-), \quad t \geqq 1$$

értékét. Mivel

$$f_t = \mathbf{D}^{-1}(y_t - \mathbf{B}^T z_{t-1}),$$

így

$$E(f_t|Y_0^-) = E\big(\mathbf{D}_-^{-1}(y_t - \mathbf{B}^T z_{t+1})|Y_0^-\big) = \mathbf{D}_-^{-1}[E(y_t|x_0) - \mathbf{B}^T E(z_{t+1}|x_0)] =$$

$$= \mathbf{D}_-^{-1}(\mathbf{CA}^{t-1}x_0 - \mathbf{B}^T\mathbf{QA}^t x_0) = \mathbf{K}_-^T\mathbf{A}^{t-1}x_0.$$

Ugyanakkor

$$x_0 = \sum_{s=0}^{\infty} \mathbf{A}^s \mathbf{K}_+ \mathbf{e}_s,$$

ezért

$$E(f_t|Y_0^-) = \sum_{s=0}^{\infty} \mathbf{K}_-^T \mathbf{A}^{t+s-1} \mathbf{K}_+ e_{-s} \quad t \geqq 1.$$

Tehát az együttható mátrix éppen a kívánt *Hankel-mátrix.*

Nézzük meg, mi történik a kanonikus korrelációkkal, ha az így megkapott p-rangú *Hankel-mátrixot* közelítjük alacsonyabb rangúval.

Hajtsuk először végre az x_t, ill. z_t állapottéren azt a transzformációt, melynek eredményeként

$$P = Q = \sum = \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \tag{9.5}$$

mátrixokat kapjuk, ahol $\sigma_1^2, \ldots \sigma_p^2$ a **PQ** saját értékei. De a 7. fejezetből tudjuk, hogy **PQ** saját értékei éppen a megfelelő *Hankel-mátrix* szinguláris értékeinek négyzetei, de ez a *Hankel-mátrix* a $\mathbf{K}_-^T \mathbf{A}^{t-1} \mathbf{K}_+$, $t \geqq 1$ mátrixokból kell felépüljön, azaz egybeesik a vetítés operátor *Hankel-mátrixával,* tehát $\sigma_1, \ldots, \sigma_p$ éppen a kanonikus korrelációk, és

$$z_1(1), \ldots, z_1(p), \quad \text{ill.} \quad x_0(1), \ldots, x_0(p)$$

a kanonikus vektorok (ahol $z_1(i)$ a z_1, $x_0(i)$ az x_0 i-edik koordinátája).

Tegyük fel, hogy $\sigma_1 > \sigma_2 \ldots > \sigma_p > 0$. Az operátornormában legjobban közelítő k-rangú *Hankel-operátort* úgy kapjuk, ha vesszük a

$$T_k z_{t+1}(k+1) = x_t(k+1)$$

egyenlőséggel definiált izometriát, és képezzük a

$$\chi_k = \Pr_{Y_0^-} - \sigma_{k+1} \Pr_{Y_0^-} T_k$$

operátort.

Legyen $Z_t^+(k+1)$ a $z_t(k+1)$, $z_{t+1}(k+1) \ldots$; $Z_t^-(k+1)$ a $z_t(k+1)$, $z_{t-1}(k+1) \ldots$ $\ldots X_t^+(k+1)$ az $x_t(k+1)$, $x_{t+1}(k+1), \ldots$ $X_t^-(k+1)$ az $x_t(k+1)$, $x_{t+1}(k+1), \ldots$ változók által generált zárt altér.

Tudjuk, hogy H_k eltűnik a $Z_1^+(k+1)$ altéren. Tetszőleges $\eta \in Y_1^+$ felírható

$$\eta = \eta^+ + \eta^-$$

alakban, ahol

$$\eta^+ \in Z_1^+(k+1),$$

$$\eta^- \in Z_0^-(k+1) \cap Y_1^+.$$

9.1. TÉTEL.

$$\chi_k \eta = E(\eta^- | Y_0^- \ominus X^-(k+1)).$$

Bizonyítás. Legyen

$$\eta \in Z_0^-(k+1) \cap Y_1^+.$$

Ekkor

$$T_k \eta \in X_{-1}^-(k+1) \subset Y_0^-.$$

Ugyanakkor $\eta \in Y_1^+$ miatt

$$\eta = \sum_{t=1} \alpha_t f_t.$$

Legyen

$$\xi = T_k \eta.$$

Megmutatjuk, hogy $\eta - \sigma_{k-1}\xi \perp X_1^-(k+1)$. Mivel $\xi \in X_{-1}^-(k+1)$, ezért ξ értékének meghatározásához elég az

$$E(\xi x_s(k+1)), \quad s \leqq -1$$

kovarianciákat vizsgálni. Mivel T_k izometria, ezért

$$E(\xi x_s(k+1)) = E(\eta z_{s+1}(k+1)).$$

(9.3) alapján $s \leqq 0$ esetén

$$E(\eta z_{s+1}(k+1)) = \sum_{t=1}^{\infty} \alpha_t E(f_t z_{s+1}(k+1)) = \sum_{t=1}^{\infty} \alpha_t \mathbf{K}^T A^{t-s-1} \varepsilon_{k+1}^T,$$

ahol

$$\varepsilon_{k+1}^T = (\overset{1}{\overbrace{0}}, \ldots, \overset{k}{\overbrace{0}}, \overset{k+1}{\overbrace{1}}, \overset{k+2}{\overbrace{0}}, \ldots, \overset{p}{\overbrace{0}}).$$

Ugyanakkor

$$E(\eta x_s(k+1)) = \sum_{t=1}^{\infty} \alpha_t E(f_t x_s(k+1)) = \sum_{t=1}^{\infty} \alpha_t \mathbf{K}^T A^{t-s-1} \sum \varepsilon_{k+1}^T.$$

Tehát

$$E(\eta x_s(k+1)) = \sigma_{k+1} E(\xi x_s(k+1)).$$

Ezért

$$E((\eta - \sigma_{k+1}\xi) x_s(k+1)) = 0.$$

Más szóval $\eta \in Z_0^-(k+1) \cap Y_1^+$ esetén

$$\sigma_{k+1} \mathbf{T}_k \eta = E(\eta | X_{-1}^-(k+1)).$$

Tehát

$$\chi_k \eta = E(\eta^- | Y_0 \ominus X_{-1}^-(k+1)).$$

Így ξ_k szinguláris értékei az $Y_1^+ \cap Z_0^-(k+1)$ és az $Y_0^- \ominus X_{-1}^-(k+1)$ altérbe eső valószínűségi változók kanonikus korrelációival esnek egybe.

10. Az átviteli függvény

Legyen **A** tetszőleges $p \times p$ méretű mátrix, amelynek nincs saját értéke az egységkörön, és legyenek **B**, **C**, **D** tetszőleges $p \times q$, $q \times p$, illetve $q \times q$ méretű mátrixok. Tekintsük az általunk meghatározott

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t, \tag{10.1}$$

$$\mathbf{y}_{t+1} = \mathbf{C}\mathbf{x}_t + \mathbf{D}\mathbf{u}_t, \tag{10.2}$$

$$\mathbf{z}_{t-1} = \mathbf{A}^T\mathbf{z}_t + \mathbf{C}^T\mathbf{y}_t, \tag{10.3}$$

$$\mathbf{v}_{t-1} = \mathbf{B}^T\mathbf{z}_t + \mathbf{D}^T\mathbf{y}_t \tag{10.4}$$

lineáris leképezéseket. A 6. fejezetben láttuk, hogy a bázis alkalmas megválasztásával elérhető, hogy például a (10.1)—(10.2) egyenletek (6.4)—(6.5)—(6.6) alaknak legyenek (ebben a fejezetben a folyamatokat kissé másképp indexeljünk, mint a korábbiakban). Azt is láttuk, hogy $\mathbf{x}_t$

$$\mathbf{x}_t = \sum_{k=-\infty}^{\infty} \Lambda_k \mathbf{u}_{t-k}$$

alakban is előállítható, ahol a Λ_k együtthatók legegyszerűbben a

$$\Lambda(s) = \sum_{k=-\infty}^{\infty} \Lambda_k s^k$$

generátorfüggvénnyel jellemezhetőek, mely

$$G(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$$

alakban is előállítható, és hasonlóan állítható elő $\mathbf{y}_t$ is:

$$\mathbf{y}_t = \sum_{k=-\infty}^{\infty} \Phi_k \mathbf{u}_{t-k},$$

ahol

$$\Phi(s) = \sum_{k=-\infty}^{\infty} \Phi_k s^k = \sigma\big(C\Lambda(s) + D\big) = \sigma\big(\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}\big),$$

és $\sigma = s^{-1}$ (vö.: (6.10)).

Ezeknek a generátorfüggvényeknek a duáljai a (10.3)—(10.4) egyenleteknek megfelelő

$$\Delta(s) = \sum_{k=-\infty}^{\infty} \Delta_k \sigma^k = (\sigma\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{C}^T,$$

$$\psi(s) = \sum_{k=-\infty}^{\infty} \psi_k \sigma^k = s\big(\mathbf{B}^T\Delta(s) + \mathbf{D}^T\big) = s\big(\mathbf{B}^T(\sigma\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{C}^T + \mathbf{D}^T\big)$$

függvények, amelyek együtthatóinak a segítségével $\mathbf{z}_t$, $\mathbf{v}_t$

$$\mathbf{z}_t = \sum_{k=-\infty}^{\infty} \Delta_k \mathbf{y}_{t+k},$$

$$\mathbf{v}_t = \sum_{k=-\infty}^{\infty} \psi_k \mathbf{y}_{t+k}$$

alakban állíthatók elő.

A 6. fejezetben azt a kérdést vizsgáltuk, hogy mikor lesz az **A**, **B**, **C**, **D** mátrixon által indukált lineáris leképezés adjungáltja és inverze identikus, vagyis mikor teljesül, hogy $\mathbf{v}_t = \mathbf{u}_t$, vagy ami ezzel ekvivalens:

$$\Phi(s)\psi(s) = \mathbf{I}. \tag{10.5}$$

Láttuk, hogy ez viszont ekvivalens azzal, hogy

$$\mathbf{x}_t = \mathbf{P}\mathbf{z}_t \tag{10.6}$$

legyen alkalmas, nem szinguláris **P** mátrix mellett. (Korábbi jelöléseink mellett $\mathbf{v}_t$ az $\mathbf{u}_t$ konstansszorosa volt, ez nyilván elhanyagolható körülmény.)

Könnyen látható, hogy ekkor elegendő feltenni, hogy

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}\begin{pmatrix} \mathbf{P} & \mathbf{O} \\ \mathbf{O} & \mathbf{I} \end{pmatrix}\begin{pmatrix} \mathbf{A}^T & \mathbf{C}^T \\ \mathbf{B}^T & \mathbf{D}^T \end{pmatrix} = \begin{pmatrix} \mathbf{P} & \mathbf{O} \\ \mathbf{O} & \mathbf{I} \end{pmatrix}, \tag{10.7}$$

hiszen ha a bal oldalon álló szorzat tagjait a $\begin{pmatrix}\mathbf{z}_t\\ \mathbf{y}_t\end{pmatrix}$ vektorra alkalmazzuk, akkor előbb a

$$\begin{pmatrix}\mathbf{P}\mathbf{z}_{t-1}\\ \mathbf{v}_{t-1}\end{pmatrix} = \begin{pmatrix}\mathbf{x}_{t-1}\\ \mathbf{u}_{t-1}\end{pmatrix}$$

vektort, majd ebből az $\begin{pmatrix}\mathbf{x}_t\\ \mathbf{y}_t\end{pmatrix}$ vektort kapjuk, ami végül is ugyanaz, mintha a jobb oldalt közvetlenül alkalmaztuk volna (vö. 6.1. következmény).

Vezessük be az

$$\mathbf{U} = \mathbf{A}\mathbf{P}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T - \mathbf{P}, \tag{10.8}$$

$$\mathbf{V} = \mathbf{A}\mathbf{P}\mathbf{C}^T + \mathbf{B}\mathbf{D}^T, \tag{10.9}$$

$$\mathbf{W} = \mathbf{C}\mathbf{P}\mathbf{C}^T + \mathbf{D}\mathbf{D}^T \tag{10.10}$$

jelöléseket. Ezek mellett

$$\mathbf{Q}(s) = \mathbf{\Lambda}(s)\boldsymbol{\psi}(s) - \mathbf{P}\mathbf{\Delta}(s) = s(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}(\mathbf{B}^T\mathbf{\Delta}+\mathbf{D}^T) - \mathbf{P}\mathbf{\Delta} =$$

$$= s(s\mathbf{I}-\mathbf{A})^{-1}\{(\mathbf{B}\mathbf{B}^T - (\mathbf{I}-\sigma\mathbf{A})\mathbf{P})\mathbf{\Delta} + \mathbf{V} - \mathbf{A}\mathbf{P}\mathbf{C}^T\} = s(s\mathbf{I}-\mathbf{A})^{-1}(\mathbf{U}\mathbf{\Delta}+\mathbf{V}),$$

$$\mathbf{R}(s) = \mathbf{\Phi}(s)\boldsymbol{\psi}(s) - \mathbf{I} = \sigma(\mathbf{C}\mathbf{\Lambda}+\mathbf{D})\boldsymbol{\psi} - \mathbf{I} =$$

$$= \sigma\mathbf{C}(Q+\mathbf{P}\mathbf{\Delta}) + \sigma\mathbf{D}\boldsymbol{\psi} - \mathbf{I} = \mathbf{T}\mathbf{U}\mathbf{\Delta} + \mathbf{T}\mathbf{V} + \mathbf{V}^T\mathbf{\Delta} + \mathbf{W},$$

ahol

$$\mathbf{T}(s) = \mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1} = \mathbf{\Delta}^T(\sigma).$$

10.1. TÉTEL. Ha **A**-nak nincs saját értéke az egységkörön, és az **A**, **B**, **C**, **D** mátrixokhoz található olyan **P** mátrix, amelyre teljesül (10.7), akkor tetszőleges q-dimenziós ortonormált $\{\mathbf{u}_t, t=0, \pm 1, \ldots\}$ kétirányban végtelen sorozathoz találhatóak olyan p-dimenziós $\mathbf{x}_t$, illetve q-dimenziós $\mathbf{y}_t$, $\mathbf{v}_t$ sorozatok, amelyekre teljesülnek a (10.1), (10.2), (10.3), (10.4) egyenletek. Ezek az egyenletek az $\mathbf{x}_t$ $\mathbf{z}_t$, $\mathbf{y}_t$, $\mathbf{v}_t$ sorozatokat egyértelműen meghatározzák, és $\mathbf{y}_t$ szintén ortonormált, $\mathbf{x}_t = \mathbf{P}\mathbf{z}_t$, $\mathbf{v}_t$ pedig azonos $\mathbf{u}_t$-vel.

Bizonyítás. A (10.1)—(10.4) egyenletek egyértelmű megoldhatóságát a 6. fejezetben beláttuk (a leképezések ott használt sztochasztikus köntöse nyilván lényegtelen). A tétel kimondása előtt levezetett algebrai azonosságok alapján a (10.8)—(10.10) jelölések mellett, ha $\mathbf{U}=\mathbf{O}$, $\mathbf{V}=\mathbf{O}$, $\mathbf{W}=\mathbf{O}$ (ezek különböző méretű mátrixok), akkor

$$\mathbf{\Lambda}(s)\boldsymbol{\psi}(s) = \mathbf{P}\mathbf{\Delta}(s),$$

vagyis $\mathbf{x}_t = \mathbf{P}\mathbf{z}_t$, és

$$\mathbf{\Phi}(s)\boldsymbol{\psi}(s) = \mathbf{I},$$

vagyis $\mathbf{u}_t = \mathbf{v}_t$, tehát a (10.1), (10.2) egyenletekkel meghatározott leképezés duálja egyben az inverze, és így $\mathbf{y}_t$ szintén ortonormált.

10.2. TÉTEL. Ha **A**-nak nincs saját értéke az egységkörön, ha az **A**, **B**, **C**, **D** mátrixokkal létesített (10.1), (10.2) egyenletekkel meghatározott lineáris leképezés ortonormált sorozatokat ortonormált sorozatokba visz, és ez a reprezentáció minimális dimenziójú (tehát a q-dimenziós $\mathbf{u}_t$, $\mathbf{y}_t$ sorozatok közti leképezés nem állítható elő valamilyen $p' < p$ dimenziós állapottérrel), akkor van olyan nem szinguláris **P** mátrix, amelyre teljesül (10.7).

Bizonyítás. Ismeretes, hogy ha a négyzetes **F**, **G**, és

$$\mathbf{H} = \mathbf{F}+\mathbf{X}\mathbf{G}^{-1}\mathbf{Y}$$

mátrixok invertálhatóak (**F** és **G** nem szükségképpen azonos méretű, ennek megfelelően **X**, **Y** nem szükségképpen négyzetesek), akkor

$$\mathbf{E} = \mathbf{G}+\mathbf{Y}\mathbf{F}^{-1}\mathbf{X}$$

is invertálható, és

$$\mathbf{H}^{-1} = \mathbf{F}^{-1}-\mathbf{F}^{-1}\mathbf{X}\mathbf{E}^{-1}\mathbf{Y}\mathbf{F}^{-1}.$$

Ha ugyanis állításunk igaz, a betűk szerepének a felcserélésével **E** inverzét is kiolvashatjuk belőle:

$$\mathbf{E}^{-1} = \mathbf{G}^{-1}-\mathbf{G}^{-1}\mathbf{Y}\mathbf{H}^{-1}\mathbf{X}\mathbf{G}^{-1},$$

és nyilván elegendő igazolni, hogy ez valóban **E** inverze, ami egyszerű visszaszorzással látható.

Mivel a (10.1), (10.2) leképezés ortonormált sorozatot ortonormált sorozatba visz, teljesül (10.5), nevezetesen a

$$\mathbf{K} = -\Phi(-1) = \mathbf{D}-\mathbf{C}(\mathbf{I}+\mathbf{A})^{-1}\mathbf{B}$$

mátrix ortonormált:

$$\mathbf{K}\mathbf{K}^T = \mathbf{I}.$$

(A **K** előállításában szereplő inverz azért létezik, mert **A**-nak nincs sajátértéke az egységkörön.)

Alakítsuk át Φ-t, és invertáljuk előrebocsátott észrevétel segítségével:

$$\begin{aligned}(s\Phi(s))^{-1} &= \left(\mathbf{K}+(1+s)\mathbf{C}(\mathbf{I}+\mathbf{A})^{-1}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\right)^{-1} = \\ &= \mathbf{K}^T-(1+s)\mathbf{K}^T\mathbf{C}(\mathbf{I}+\mathbf{A})^{-1}\{(s\mathbf{I}-\mathbf{A})+ \\ &+\mathbf{B}\mathbf{K}^T(1+s)\mathbf{C}(\mathbf{I}+\mathbf{A})^{-1}\}^{-1}\mathbf{B}\mathbf{K}^T = \\ &= \mathbf{K}^T+(1+s)\mathbf{C}_1(\mathbf{M}-s\mathbf{N})^{-1}\mathbf{B}_1,\end{aligned}$$

ahol

$$\begin{aligned}\mathbf{C}_1 &= -\mathbf{K}^T\mathbf{C}(\mathbf{I}+\mathbf{A})^{-1}, \\ \mathbf{B}_1 &= -\mathbf{B}\mathbf{K}^T, \\ \mathbf{N} &= \mathbf{I}-\mathbf{B}\mathbf{C}_1, \\ \mathbf{M} &= \mathbf{A}+\mathbf{B}\mathbf{C}_1.\end{aligned}$$

Feltevésünk szerint ez azonos $\sigma\psi(s)$-sel, amelyik az egységkör alkalmas környezetében folytonos, tehát **M** invertálható, és

$$\mathbf{D}^T = \mathbf{K}^T+\mathbf{C}_1\mathbf{M}^{-1}\mathbf{B}_1. \tag{10.11}$$

Ennek alapján

$$(s\Phi(s))^{-1} = \mathbf{D}^T+\mathbf{H}(\sigma\mathbf{I}-\mathbf{F})^{-1}\mathbf{G},$$

ahol

$$\begin{aligned}\mathbf{F} &= \mathbf{M}^{-1}\mathbf{N}, \\ \mathbf{G} &= \mathbf{M}^{-1}\mathbf{B}_1, \\ \mathbf{H} &= \mathbf{C}_1(\mathbf{I}+\mathbf{M}^{-1}\mathbf{N}).\end{aligned}$$

Kálmán tétele a (vö.: 4.1. tétel) 6. fejezet módszerével általánosítható, tehát található olyan nem szinguláris **P**, melyre

$$\mathbf{F} = \mathbf{P}\mathbf{A}^T\mathbf{P}^{-1}, \quad \mathbf{G} = \mathbf{P}\mathbf{C}^T, \quad \mathbf{H} = \mathbf{B}^T\mathbf{P}^{-1},$$

vagyis

$$\mathbf{M}^{-1}\mathbf{N} = \mathbf{P}\mathbf{A}^T\mathbf{P}^{-1}, \tag{10.12}$$

$$M^{-1}\mathbf{B}_1 = \mathbf{P}\mathbf{C}^T, \tag{10.13}$$

$$\mathbf{C}_1(\mathbf{I}+\mathbf{M}^{-1}\mathbf{N}) = \mathbf{B}^T\mathbf{P}^{-1}. \tag{10.14}$$

Megmutatjuk, hogy ezzel a **P**-vel teljesül (10.7). A (10.12), (10.14) egyenletek alapján

$$\mathbf{A}\mathbf{P}\mathbf{A}^T+\mathbf{B}\mathbf{B}^T = \mathbf{A}\mathbf{F}\mathbf{P}+\mathbf{B}\mathbf{C}_1(\mathbf{I}+\mathbf{F})\mathbf{P} =$$

$$= \{(\mathbf{A}-\mathbf{B}\mathbf{C}_1)\mathbf{F}-\mathbf{B}\mathbf{C}_1\}\mathbf{P} = (\mathbf{M}\mathbf{F}+\mathbf{I}-\mathbf{N})P = \mathbf{P}.$$

A (10.11), (10.13) egyenletek alapján

$$\mathbf{A}\mathbf{P}\mathbf{C}^T+\mathbf{B}\mathbf{D}^T = \mathbf{A}\mathbf{P}\mathbf{C}^T+\mathbf{B}(\mathbf{K}^T+\mathbf{C}_1\mathbf{P}\mathbf{C}^T) = \mathbf{M}\mathbf{P}\mathbf{C}^T-\mathbf{B}_1 = \mathbf{O}.$$

Végül a (10.11), (10.13) egyenletek alapján

$$\mathbf{D}\mathbf{D}^T = (\mathbf{K}-\mathbf{K}\mathbf{C}_1\mathbf{B})(\mathbf{K}^T+\mathbf{C}_1\mathbf{P}\mathbf{C}^T) =$$

$$= \mathbf{I}-\mathbf{K}\mathbf{C}_1\mathbf{B}\mathbf{K}^T+\mathbf{K}\mathbf{C}_1\mathbf{P}\mathbf{C}^T-\mathbf{K}\mathbf{C}_1(\mathbf{M}-\mathbf{A})\mathbf{P}\mathbf{C}^T =$$

$$= \mathbf{I}-\mathbf{K}\mathbf{C}_1(\mathbf{B}\mathbf{K}^T+\mathbf{M}\mathbf{P}\mathbf{C}^T)+\mathbf{K}\mathbf{C}_1(\mathbf{I}+\mathbf{A})\mathbf{P}\mathbf{C}^T =$$

$$= \mathbf{I}-\mathbf{C}\mathbf{P}\mathbf{C}^T.$$

Ezzel a 10.2. tétel bizonyítását befejeztük. Megjegyezzük, hogy (10.11) a (10.12) (10.13), (10.14) egyenletek következménye:

$$\mathbf{K}^T = \mathbf{D}^T+\mathbf{B}^T(\mathbf{I}+\mathbf{A}^T)^{-1}\mathbf{C}^T = \mathbf{K}^T+\mathbf{C}_1(\mathbf{I}+\mathbf{F})\mathbf{P}(\mathbf{I}+\mathbf{A}^T)^{-1}\mathbf{C}^T =$$

$$= \mathbf{K}^T+\mathbf{C}_1\mathbf{P}\mathbf{C}^T = \mathbf{K}^T+\mathbf{C}_1\mathbf{M}^{-1}\mathbf{B}_1.$$

Ez a bizonyítás lényegében GLOVERtől származik. Azért ismételjük itt el, mert cikkünk megírása közben hosszasan kerestünk helyette közvetlenebb bizonyítást. Így keletkezett a 6. fejezet anyaga, mégis úgy gondoljuk, létezik egyszerűbb bizonyítás is.

IRODALOM

ADAMJAN, V. M., AROV, D. Z. and KREIN M. G., "Analytic properties of Schmidt pairs of Hankel operator and the generated Schur—Takagi problem", *Math. USSR—Sb.* **15** (1971) 31—73.

ARATÓ, M., *Linear Stochastic Systems with Constant Coefficients* (Springer, Berlin, 1982).

BALL, J. A. and HELTON, J. W., "A Beurling-Lax theorem for the Lie group $U(m, n)$ which contains most classical interpolation theory", *J. Operator Theory* **9** (1982) 107—142.

BALL, J. A. and RAN A. C. M., "Optimal Hankel norm model reductions and Wiener—Hopf faktorization. I: the canonical case", *SIAM J. Control and Optimization* **25** (1987) 362—382.

BOLLA, M., „Hilbert-terek lineáris operátorainak szinguláris felbontása (optimális tulajdonságok statisztikai alkalmazásai és numerikus módszerek)", *Alk. Mat. Lapok* **13** (1988).

BYRNES, C. I. and LINDQUIST, A., *Modelling, Identification and Robust Control* (Elsevier, Nort Holland, 1986).

BYRNES, C. I. and MARTIN, C. F., *Geometrical Methods for the Theory of Linear Systems* (D. Reidel, Boston, 1980).
DESAI, V. B. and PAL, D. A., "A transition approach to stochastic model reduction", *IEEE Trans. on Automatic Control* **AC—29** (1984) 1097—1100.
FAURRE, P., "Stochastic realization algorithms", In *System Identification, advances and case studies* (ed. *R. K. Mehra, D. G. Lainotis*) (Acad. Press, 1976).
GLOVER, K., "All optimal Hankel-norm approximations of linear multivariable systems and their L^{∞}-error bounds", *Int. J. Control* **39** (1984) 1115—1193.
FUHRMANN, P. A., *Linear Systems and Operators in Hilbert Space* (McGraw-Hill, New York 1981).
HANNAN, E. J. and KAVALIERS, L., "Multivariate linear time series models", *Adv. Appl. Prob.* **16** (1984) 492—561.
IBRAGIMOV, I. A. and ROZANOV, Y. A., *Gaussian Random Processes* (Springer, Berlin, 1978).
KALMAN, R. E., *Lectures on Controllability and Observability* (Cremonose, Roma, 1969).
KALMAN, R. E., FALB, P. L. and ARBIB, M. A., *Topics in Mathematical System Theory* (McGraw-Hill, New York, 1969).
KATO, T., *Perturbation Theory for Linear Operators* (Springer, Berlin, 1984).
KUNG, S. Y. and LIN, D. W., "Optimal Hankel norm reductions multivariable systems", *IEEE Trans. Automat. Control* **26** (1981) 232—252.
MÓRI, T. és SZÉKELY, J. G., *Többdimenziós statisztikai vizsgálatok* (Műszaki Könyvkiadó, Bp., 1986).
NEHARI, Z., "On bounded bilinear forms", Ann. of Math **65** (1957) 153—162.
TUSNÁDY, G., "Mátrixok szinguláris felbontása", *Alk. Mat. Lapok* **5** (1979) 375—384.
TUSNÁDY, G. és ZIERMANN, M., *Idősorok analízise* (Műszaki Könyvkiadó, Budapest, 1986).
WHITTLE, P., *Optimization over Time* (Wiley, New York, 1982).

(Beérkezett: 1987. december 29.)

MICHALETZKY GYÖRGY
ELTE TTK VALÓSZÍNŰSÉGSZÁMÍTÁS TANSZÉK
1088 BUDAPEST, MÚZEUM KRT. 6—8.

TUSNÁDY GÁBOR
MTA MATEMATIKAI KUTATÓINTÉZET
1053 BUDAPEST, REÁLTANODA U. 13—15.

ON THE STATE SPACE REPRESENTATION OF MULTIDIMENSIONAL TIME SERIES

GY. MICHALETZKY, G. TUSNÁDY

The state space representation of time series introduced by R. KALMAN gives a very effective tool for investigation of wide class of *multidimensional Gauss processes*. The first results of this theory were developed about 1960 but in the last decade many intrinsically new theorem was proven, among which the model-reducing construction of KEIT GLOVER was the most inportant one. We give a conscious introduction into this theory in the present paper.

Alkalmazott Matematikai Lapok **13** (1987—88) 285—289

ADATBÁZIS-KEZELŐ RENDSZEREK HATÉKONYSÁGÁNAK JELLEMZÉSE KOLMOGOROV ALGORITMIKUS INFORMÁCIÓMENNYISÉGE ALAPJÁN

BENCZÚR ANDRÁS

Budapest

A dolgozatban relációs adatbázisok információtartalmának mérésére KOLMOGOROV algoritmikus információmértékét alkalmazzuk. A módszer lehetőséget ad adatbázis-kezelő rendszerek hatékonyságának modellezésére.

1. Bevezetés

Az adatbázis-kezelő rendszerek működésének alapvető feladata az, hogy előre adott feltételek mellett, amit sémaként adunk meg, az adatbázis lehetséges állapotait kódolja és dekódolja.

Ez lényegében úgy történik, hogy a séma kódja alapján az adatbázis-kezelő rendszer generál egy speciális algoritmust, amely a lehetséges állapotokat kódolja-dekódolja. Az így kapott speciális algoritmus jóságának egyik jellemzője, hogy mennyire rövid az állapot bemenő kódjához képest a tárolási kód. Az adott séma mellett elérhető legrövidebb kód hossza jellemzi az állapot feltételes információmennyiségét. Ebben a heurisztikus mérési módszerben két önkényes komponens van, az egyik a „bemenő kód", a másik pedig a séma melletti legrövidebb kód használata.

A „bemenő kódok" halmazaként az adatbázis-kezelő rendszer által kezelhető összes lehetséges állapothoz rendelt olyan kódhalmazt kell értenünk, amely külön megadott sémától függetlenül, „önmagában érthető" jelsorozatokból áll, amelyeket az állapotok kanonikus alakjának fogunk majd nevezni.

A második önkényes komponenstől, a sémához választott algoritmusok szerinti méréstől a KOLMOGOROV által [1]-ben bevezetett algoritmikus információmérték alaptulajdonsága (lásd 1. tétel) alapján tudunk elszakadni. Az említett tulajdonság szemléletesen azt jelenti, hogy létezik olyan univerzális kódoló-dekódoló eljárás, hogy bármilyen sémához választott kódoló-dekódoló eljáráshoz képest az univerzális eljárás legfeljebb egy konstanssal hosszabb kódot használ minden állapotra. Ez a konstans hossz nem más, mint a séma kódjának hossza. Végső soron az adatbázis-kezelő rendszer, mint univerzális dekódoló, úgy viselkedik, mint egy kétváltozós algoritmus, melynek egyik változója a séma kódja, a másik pedig az állapot kódja. A séma kódja alapján kerül kiválasztásra az a speciális algoritmus, amely az adott állapotkódból az állapotot kanonikus alakra dekódolja.

A következőkben a fenti gondolatmenet precíz kifejtését adjuk meg a relációs adatbázis kezelő rendszerekre. A relációs adatbázis-kezelő rendszerekről részletesebb leírás található a [2], [5] könyvekben és az [1] dolgozatban. Itt csak röviden ismertetjük a leglényegesebb alapfogalmakat.

A relációs adatbázisok adatmodelljének alapegysége a reláció. A reláció nem más, mint véges sok véges értéktartomány direktszorzataként adott halmaznak egy részhalmaza. A relációra névvel hivatkozunk, és típussal írjuk le, ami megadja dimenzióját, az egyes értéktartományokat, valamint a reláció elemeinek komponenseire, mint tulajdonságokra való hivatkozásokat, amelyeket attributum névnek nevezünk. Az adott típusú relációk megengedett előfordulásaira a típuson belül önmagában, vagy több típusra összefüggésben feltételek adhatók meg, amit általában függőségeknek neveznek. A gyakorlatban előforduló feltételek a relációnevekkel, mint predikátumszimbólumokkal felírt elsőrendű predikátumkalkulus formulákat jelentenek. Hasonlóan jellemezhetők a lekérdezések, mint elsőrendű predikátumkalkulus kifejezések.

A relációs adatbázis sémája a relációtípusok és a feltételek megadásából áll.

2. A relációk kanonikus alakja

Abból indulunk ki, hogy adatbázis-kezelő rendszerünk véges dimenziós, véges relációk kezelését biztosítja. Az „összes reláció" halmazát következőképpen adjuk meg:

Legyenek adottak a pozitív egészekből álló $D_1, \ldots, D_n$ halmazok, $D_i = \{1, 2, \ldots, m_i\}$.

A $D_1, \ldots, D_n$ értéktartományok feletti r reláción a $D = D_1 \times \ldots \times D_n$ szorzathalmaz részhalmazát értjük, s r típusát $R(n; m_1, \ldots, m_n)$-nel jelöljük. Az „összes reláció" halmazán a lehetséges $R(n; m_1, \ldots, m_n)$ típusok előfordulásainak halmazát értjük.

Megjegyzés. A modell a tetszőleges véges tartományok feletti relációknak az egész számok feletti relációkká való átkódolásával nem foglalkozik.

Az $r \in R(n; m_1, \ldots, m_n)$ reláció indikátor-vektorán a $\varrho(r) = (a_1, \ldots, a_M)$ bináris vektort értjük, ahol a_i 1 vagy 0 értéket vesz fel aszerint, hogy a D halmaz lexikografikus sorrendben vett i-edik elemét tartalmazza-e r vagy sem.

A pozitív egész számok és a nem üres bináris sorozatok között a $0 \leftrightarrow 1$, $1 \leftrightarrow 2$, $00 \leftrightarrow 3$, $01 \leftrightarrow 4$, ..., azaz a hossz szerinti és azon belül a lexikografikus megfeleltetést használjuk. Az x pozitív egész számnak megfeleltetett bináris sorozatot jelölésben nem különböztetjük meg, szintén x-szel jelöljük.

Szükségünk van még az $x = x_1 \ldots x_k$ bináris sorozatra definiált $\bar{x} = (x_1 x_1 x_2 x_2 \ldots x_k x_k 0\ 1)$ transzformációra.

1. *Definíció.* Az r reláció kanonikus alakján az $x(r) = \bar{n}\bar{m}_1 \ldots \bar{m}_n \varrho(r)$ bináris sorozatot értjük. Nyilvánvaló, hogy az $x(r)$ leképezés kölcsönösen egyértelmű, és a lehetséges kanonikus alakok halmaza rekurzíven felsorolható.

3. A relációk algoritmikus információmennyisége

Az 1. definíció szerint bevezetett kanonikus alak alkalmas arra, hogy a KOLMOGOROV által [7]-ben bevezetett algoritmikus információértéket alkalmazzuk. Idézzük fel az alapdefiníciókat és tételeket.

Legyen $K(p)$ a bináris pozitív egészeken értelmezett és bináris pozitív egész értéket előállító függvény.

2. *Definíció.* Egy x bináris pozitív egész K szerinti bonyolultságán a $C_K(x)=\min_{K(p)=x} l(p)$ értéket értjük, ahol $l(p)$ a p bináris sorozat hossza. Amennyiben nem létezik p, amelyre $K(p)=x$, úgy $C_K(x)=\infty$.

1. TÉTEL (KOLMOGOROV [7]). Létezik optimális $K_0(p)$ rekurzív függvény, olyan, hogy minden $K(p)$ parciális rekurzív függvényre

$$C_{K_0}(x) \leqq C_K(x)+L_K \quad \text{minden } x\text{-re,}$$

ahol az L_K konstans csak K-tól függ.

A $K_0(p)$ optimális rekurzív függvény nem egyértelmű, de ha $K_1(p)$ egy másik optimális rekurzív függvény, akkor

$$|C_{K_0}(x)-C_{K_1}(x)| \leqq L_{K_0,K_1} \quad \text{minden } x\text{-re.}$$

3. *Definíció* Legyen K_0 egy mostantól adott optimális rekurzív függvény. Az r reláció algoritmikus információmennyiségén a

$$C(r) = C_{K_0}(x(r)) \quad \text{mennyiséget értjük.}$$

Az így definiált információmennyiség szemléletesen azt mutatja, hogy nagyméretű relációkat $C(r)$-nél rövidebb kóddal nem lehet ábrázolni. Egy r reláció egyszerűségét az $S(r)=\dfrac{C(r)}{l(x(r))}$ hányados mutatja, ahol

$$l(x(r)) = 2(\log_2(n))+\sum_{i=1}^{n}\log_2 m_i+\sum_{i=1}^{n} m_i+2n+2.$$

Az egyszerűségi mutatót adott feltételeket kielégítő relációk halmazának jellemzésére is felhasználjuk.

4. *Definíció.* Legyen G relációk egy rekurzívan felsorolható halmaza, amelyet egy K parciális rekurzív függvény sorol fel, pontosabban $K(i)=x(r_i)$, $r_i \in G$, $i=1, 2, \ldots$ és $r_i \neq r_j$ $i \neq j$ esetén. A G halmaz egyszerűségén az

$$S(G) = \limsup_{i\to\infty} \frac{C(r_i)}{l(x(r_i))} \quad \text{értéket értjük.}$$

Nyilvánvaló, hogy $S(G)$ nem függ a G-t felsoroló K parciális rekurzív függvénytől. Kihasználhatjuk, hogy $C_K(X(r_i))=\log_2 i$, s így

$$C(r_i) \leqq \log_2(i)+L_K. \quad \text{Ezzel az alábbi}$$

$$S(G) \leqq \limsup_{i\to\infty} \frac{\log_2(i)}{l(x(r_i))} \tag{1}$$

egyenlőtlenséget kapjuk.

Az 1. tétel alapján a G halmaz elemeinek kódolására $S(G)$-nél hatékonyabb kódolás nem létezik. Ezt használjuk ki a relációs függőségek, megszorítások (lásd [1], [2] és [5] jellemzésére.

Legyen F egy olyan feltétel, amely minden relációra algoritmikusan eldönthető. Jelöljük SAT (F)-fel az F feltételt kielégítő relációk halmazát. SAT (F) ekkor nyilván

rekurzívan felsorolható, tehát $S(\mathrm{SAT}(F))$ létezik, s jellemzi az F feltétel egyszerűségét.

$S(\mathrm{SAT}(F))$ becslésére (1)-hez hasonló egyenlőtlenséget kaphatunk, ha ismerjük az F feltételt kielégítő $R(n; m_1, \ldots, m_n)$ típusú relációk $N_F(n; m_1, \ldots, m_n)$ számát. Ugyanis $r\ R(n; m_1, \ldots, m_n)$ esetén $\varrho(r)$ helyett egy $\log_2 N_F(n; m_1, \ldots, m_n)$ hosszú $\delta(r)$ kódot használva az $y(r)=01\ \bar{\varphi}\bar{n}\bar{m}_1 \ldots \bar{m}_n\delta(r)$ algoritmikusan előállítható és dekódolható kódot kapjuk, amelyben φ az F feltételt megadó kód, például a $\mathrm{SAT}(F)$ elemeit felsoroló algoritmus sorszáma. Az $y(r)$ kód alapján

$$S(\mathrm{SAT}(F)) \leqq \limsup \frac{l(y(r))}{l(x(r))}, \quad \text{ahol} \tag{2}$$

a határérték az $n, m_1, \ldots, m_n$ paraméterek valamilyen szabály szerinti végtelenhez tartása mellett történik.

Az $\dfrac{l(y(r))}{l(x(r))}$ hányados viselkedését nagy m_i értékekre a

$$\frac{\log_2 N_F(n;\ m_1, \ldots, m_n)}{\prod_{i=1}^{n} m_i}$$

hányados adja meg.

A $C_{K_0}(x)$ bonyolultsági mérték ismert tulajdonságai (lásd [6]) alapján a $\log_2 N_F(n; m_1, \ldots, m_n)$ érték $C(r)$ alsó becslésére is használható bizonyos feltételek esetén.

Egyetlen példával illusztráljuk, hogy a $X(r)$ értékét megközelítő kód használata nem mindig ad hatékony megoldást, ha a tárolási kód hossza mellett a dekódoló algoritmus bonyolultságát is figyelembe kell vennünk.

Példa. Legyen $n=3$, és $m_1=m_2=m_3=m$. Az F feltétel jelentse azt, hogy az r relációban nincs két olyan elem, amelynek két komponense megegyezik, de a harmadik komponens eltér, s legyen r elemeinek száma m^2.

Belátható (lásd [1]), hogy rögzített m érték mellett $\mathrm{SAT}(F)$ kölcsönösen egyértelműen leképezhető az $m\times m$-es latin négyzetek halmazára. A latin négyzetek algoritmikus előállításának nehézsége közismert (lásd [3], [4],) $N_F(3; m, m, m)$ pontos értékét is csak viszonylag kis m-re határozták meg, s éles aszimptotikus becslés sem ismeretes.

4. Alkalmazás karbantartási és lekérdezési eljárásokra

Egy elemi karbantartási vagy lekérdezési feladatot (tranzakciót) úgy tekinthetünk, mint egy leképezést a megengedett relációk halmazáról a relációk egy, esetleg ugyanazon halmazára. Amennyiben minden relációt kanonikus alakban adunk meg, a feladat algoritmikus bonyolultsága egyértelmű.

Formálisan, legyen F_1 és F_2 két feltételrendszer és h egy leképezés $\mathrm{SAT}(F_1)$-ből $\mathrm{SAT}(F_2)$-be. Legyen $\Pi_1(k)$ a h bonyolultsága, ha argumentuma egy k hosszúságú kanonikus forma. Hasonlóan $\Pi_2(k')$ legyen h bonyolultsága, ha argumentuma egy K algoritmus szerint dekódolható k' hosszú relációkód. Végül legyen $\Pi_3(k')$ a K bonyolultsága és $\Pi_4(k)$ a K^{-1}, azaz a kódoló eljárás bonyolultsága.

A $\Pi_1(k)$ és a $\Pi_2(k')$ bonyolultsági mérőszámok egymással becsülhetők, a kódoló és dekódoló eljárások bonyolultsági mérőszámával térhetnek el egymástól. Kicsit egyszerűsítve, azaz a k és k' közötti összefüggést kölcsönösen egyértelműnek feltételezve, az alábbi egyenlőtlenségpár írható fel:

$$\Pi_1(k) \leqq \Pi_2(k') + \Pi_4(k) \quad \text{és}$$

$$\Pi_2(k') \leqq \Pi_1(k) + \Pi_3(k')$$

Ezeket az egyenlőtlenségeket felhasználhatjuk az adatbázis kezelő rendszer által megvalósított megoldás hatékonyságának vizsgálatára.

Befejezésként megjegyezzük még, hogy a több relációt tartalmazó adatbázisok vizsgálatára a modell egyszerűen kiterjeszthető, hiszen az $r_1, r_2, \ldots, r_k$ relációk együttesét például az

$$x(r_1, \ldots, r_k) = 10\bar{k}x(r_1)\ldots x(r_k)$$

kóddal, mint kanonikus alakkal adhatjuk meg. Az adatbázis sémáján az $r_1, \ldots, r_k$ relációk típusára adott feltételeket és a lehetséges relációegyüttesekre vonatkozó megszorításokat értjük.

IRODALOM

[1] BENCZÚR, A., „Az információ mérése relációs adatbázisokban", *Alk. Mat. Lapok* **12** (1986).

[2] DEMETROVICS, J., DENEV, J. és PAVLOV, R., *A számítástudomány matematikai alapjai* (Tankönyvkiadó, Budapest, 1985).

[3] MARSHALL HALL, JR., *Combinatorial Theory*, (John Wiley & Sons, Inc., New York, 1967).

[4] RYSER, H. J., *Combinatorial Mathematics* (The Mathematical Association of America, 1963).

[5] ULLMAN, J. D., *Principles of Database Systems*, (Computer Science Press, Rockville, 1982). (2nd Edition.)

[6] Эвонкин, А. К.—Левин, Л. А.: Сложность конечных обьектов и обоснованние понятий информации и случайности с помощью теории алгоритмов. Успехи Мат. Наук, т. XXV. 1970. Москва.

[7] Колмогоров, А. Н.: Три подхода к определению понятия „количество нформацни". Проблемы передачи информации, 1965. т. 1, № 1, 3—11.

(Beérkezett: 1987. május 18.)

BENCZÚR ANDRÁS
ELTE SZÁMÍTÓKÖZPONT
1117 BUDAPEST, BOGDÁNFY ÚT 10/B.

ON THE PERFORMANCE EVALUATION OF DATABASE SYSTEMS USING KOLMOGOROV'S ALGORITHMIC INFORMATION MEASURE

A. BENCZÚR

In the paper KOLMOGOROV's algorithmic information measure is applied for the measurement of the information quantity of relational databases. The approach gives a chance to model the performance of a database management system.

Alkalmazott Matematikai Lapok **13** (1987—88) 291—312

RELÁCIÓS ADATMODELL FÜGGŐSÉGEINEK EGY ÁLTALÁNOS OSZTÁLYA ÉS IMPLIKÁCIÓS PROBLÉMÁINAK VIZSGÁLATA

BENCZÚR ANDRÁS, KISS ATTILA, MÁRKUS TIBOR

Budapest

Jelen cikk a relációs adatbázis elmélet egyik fontos problémájával, a függőségek vizsgálatával foglalkozik. Megmutatjuk, hogy a függőségeknek egy elég tág osztálya („feltétel-következmény" típusú függőségek) megadható az elsőrendű predikátumkalkulus segítségével (HCD-formulák). A cikk további részében foglalkozunk a függőségi rendszerek implikációs problémáival, nevezetesen a BEERI—VARDI 1984-ben megjelent, s a [2] cikkben bevezetett TTGD, TGD, EGD típusú függőségek implikációs problémáinak általánosításával, valamint ezekhez tartozó általánosított *„chase"* bizonyító eljárással. Bevezetjük az XGD típusú függőségeket, s vizsgáljuk azon relációk struktúráját, melyek XGD függőségeket elégítenek ki.

1. Bevezetés

Az adatbázis rendszerek közül a kutatásokban, alkalmazásokban fontos szerepet töltenek be a relációs adatbázisok és kezelő rendszereik. A relációs adatbázisokban szereplő adatok között általában bizonyos kapcsolatok találhatók. E kapcsolatokat függőségeknek hívjuk. E függőségek egyrészt előzetesen előírhatók, s vizsgálható, hogy egy adatbázis ezen előírt függőségeknek, mint megszorításoknak, eleget tesz-e. Ilyen típusú vizsgálatok fontos szerepet játszanak létező relációs adatbázisokban történő adatmódosítások (sor, sorok törlése, bevitele, egyedi adatok változtatása) függőségmegőrző tulajdonságainak vizsgálatában. A függőségi vizsgálatok másik problémaköre nem konkrétan megadott adatbázishoz kötődik. A kérdés úgy vethető fel, hogy ha adott függőségeknek egy rendszere, akkor vajon egy adott függőség a rendszernek következménye-e (implikálja-e). Számos kutatási eredmény ismert a funkcionális-, több értékű függőségek esetén a fenti problémák vizsgálatára. (Lásd [5, 7, 9].)

Jelen cikkben a függőségeknek az előbbieknél általánosabb típusaival, a sorgeneráló-, az egyenlőséggeneráló-, a felcserélhetőség generáló függőségek relációs adatbázisokra való alkalmazásával és az ezekre vonatkozó implikációs problémáival foglalkozunk. E függőségek speciális esetként többek között tartalmazzák mind a funkcionális-, mind pedig a több értékű függőségeket. E függőségek bevezetése, tulajdonságainak vizsgálata BEERI—VARDI 1984-ben megjelent cikkeiben találhatók ([1, 2]).

Jelen cikk a BEERI—VARDI [2] munkájában szereplő eredmények általánosításait, új függőség — a felcserélhetőség generáló függőség — bevezetését és implikációs problémákra való alkalmazását, valamint a problémáknak az elsőrendű predikátumkalkulus segítségével történő megfogalmazását adja. A második részben foglalkozunk a kompozíció/dekompozíció, valamint az implikációs probléma formális

rendszereivel. A cikk 3. és 4. része az általánosított függőségekkel, valamint ezekre vonatkozó „chase" bizonyító eljárással foglalkozik. Az 5. rész a felcserélhetőség generáló függőségeket kielégítő relációk osztályozását adja. A 6. rész az implikációra vonatkozó *„chase"* bizonyító eljárás alkalmazását tartalmazza a totális sorgeneráló függőségek (TTGD-k), valamint a TTGD-k és az egyenlőséggeneráló függőségek (EGD-k) esetére.

Az utolsó, 7. részben a kiértékelésnek nevezett homomorf leképezéseknek az általánosítását, a szimbólumleképezéseket tárgyaljuk.

2. Fogalmak és jelölések

Jelölje $R(A_1, A_2, \ldots, A_n)$ a $D_1, D_2, \ldots, D_n$ halmazokhoz tartozó n-változós relációk halmazát. A relációs adatmodellezésben $R(A_1, A_2, \ldots, A_n)$-t a reláció típusának, $A_1, A_2, \ldots, A_n$-t attributumoknak, $D_1, D_2, \ldots, D_n$-t pedig értelmezési tartományoknak nevezzük. $R(A_1, A_2, \ldots, A_n)$ elemeit relációknak vagy reláció tábláknak nevezzük, vagyis egy R reláció a $D = D_1 \times D_2 \times \ldots \times D_n$ *Descartes-szorzatnak* egy részhalmaza. Az R reláció bármely elemét reprezentálhatjuk egy olyan sorral (n-essel), amely rendre a $D_1, D_2, \ldots, D_n$-nek megfelelő komponensekből áll, vagyis R egy sorának A_i attributumhoz tartozó értéke D_i-beli érték. Jelöljük $\mathrm{VAL}_R(A_i)$-vel az R reláció sorai A_i attributumhoz tartozó értékeinek halmazát. Világos, hogy $\mathrm{VAL}_R(A_i) \subseteq D_i$.

Egy véges vagy megszámlálhatóan végtelen két dimenziós T táblát, melynek n oszlopa van $(A_1, A_2, \ldots, A_n)$, tekinthetünk relációnak is a $\mathrm{VAL}_T(A_1), \ldots, \mathrm{VAL}_T(A_n)$ értelmezési tartományok felett. Ebből a megjegyzésből következik, hogy nem szükséges feltennünk az értelmezési tartományok halmazának apriori létezését.

Az egyszerűség kedvéért vizsgálódásainkat csak egy reláció típusra korlátozzuk, melynek attributum halmaza $U = \{A_1, A_2, \ldots, A_n\}$.

Adott típusú lehetséges relációknak egy $\mathscr{F}$ részhalmazát felfoghatjuk a relációkra tett megszorításnak is. A gyakorlati alkalmazásokban ezek a korlátozások a megengedett relációkra vonatkozó konzisztencia megszorításokat jelentenek.

Az $\mathscr{F}$ korlátozó halmaz előállításához különféle eszközöket használhatunk. Ezek közül számos az U részhalmazain hat úgy, hogy különböző függőségeket állít a soroknak adott attributum halmazhoz tartozó értékei között. Az ilyen típusú korlátozások közül a legfontosabbak a funkcionális, a több értékű, az összekapcsolási, a tartalmazási, a gyenge, a duális és az erős függőségek. (Lásd [1, 2, 3, 4, 5, 8, 9].) A megszorítást általánosabb formában egy elsőrendű logikában megfogalmazott kifejezéssel adhatjuk meg, mely vagy igaz, vagy hamis egy adott relációra nézve.

Legyen F egy megszorításnak a formális leírása, és nevezzük F-et függőségnek.

A relációknak azt a korlátozott halmazát, amely csak azokból a relációkból áll, melyek kielégítik az F függőséget, SAT (F)-fel fogjuk jelölni.

A függőségek elméletében három fő probléma merül fel: az implikáció problémája, a függőségek formális (axiomatizált) rendszereinek problémája, és a relációknak kompozíciója/dekompozíciója egyszerűbb függőségi struktúrája relációkba, illetve relációkból. (Ezt az utóbbi problémát normalizációnak is hívják.)

Függőségek implikációs problémája

Legyen $DF=\{F_1, \ldots, F_n\}$ függőségeknek egy halmaza és F egy függőség. Azt mondjuk, hogy DF implikálja F-et, jelölésben $DF \models F$, ha

$$\mathrm{SAT}(F) \supseteq \mathrm{SAT}(DF) := \bigcap_{i=1}^{n} \mathrm{SAT}(F_i).$$

A kérdés az, hogy miképp tudjuk egy $DF \models F$ implikációról eldönteni, vagy bizonyítani, hogy igaz vagy nem igaz.

Formális rendszerek problémája

Legyen L kifejezéseknek egy halmaza egy adott rendszerben, amely függőségeknek valamilyen részhalmazát írja le. (L egy speciális formális nyelv.) Egy $DF \subseteq L$ részhalmaz lezártja az a maximális függőségi halmaz L-ből, melynek elemeit DF implikálja. DF lezártját DF^+-szal jelöljük. L egy G részhalmazáról azt mondjuk, hogy zárt, ha $G=G^+$. L-beli függőségek egy zárt rendszerének olyan típusú szabályokat kell kielégítenie, hogy „függőségeknek ilyen és ilyen típusa hozzátartozik a rendszerhez" vagy „ha valamilyen függőség hozzátartozik a rendszerhez, akkor egy másiknak is hozzá kell tartoznia". Bizonyos szabályokat axiómáknak választva kérdezhetjük, hogy vajon ez az axiómarendszer sértetlen és teljes-e az L-ben (*„sound and complete"*). Ez annyit jelent, hogy L-nek egy tetszőleges részhalmaza, amely az axiómákat kielégíti, zárt és minden zárt részhalmaz kielégíti az axiómákat.

Kompozíció és dekompozíció problémája

Legyen $U=\{A_1, \ldots, A_n\}$ az attributumokból álló univerzum, és $X_1, \ldots, X_k$ legyenek U olyan részhalmazai, melyek lefedik U-t, azaz $U=\bigcup_{i=1}^{k} X_i$. Tekintsük az $R(U)$, $R(X_1)$, ..., $R(X_k)$ reláció típusokat és legyen Π egy függvény, mely $R(U)$-ból képez $R(X_1)\times\ldots\times R(X_k)$-ba, és σ egy olyan függvény, amely $R(X_1)\times\ldots\times R(X_k)$-ból képez $R(U)$-ba. A Π leképezést dekompozíciónak, a σ leképezést kompozíciónak hívjuk.

Legyen L a fenti relációtípusokon adott függőségeknek egy halmaza és L_1 egy speciális részhalmaza L-nek. A kompozíció és dekompozíció problémája a Π és σ leképezések inverz tulajdonságaival kapcsolatos azzal a megkötéssel, hogy mindegyik komponens reláció elégítsen ki egy függőséget az L_1-ből. (Másképp fogalmazva a komponensek az L_1 által meghatározott normál formában vannak.)

A függőségek tág osztálya adható meg egy speciális relációs komponens kalkulus segítségével, melyet feltételkövetkezmény (*„hypothesis-conclusion"*, rövidítve HC) formulának hívunk. Először a relációs komponens kalkulus egy szűkebb változatát ismertetjük. (Lásd [9].)

A relációs komponens kalkulusban egy formula atomokból és operátorokból épül fel.

Az atomok kétfélék lehetnek:

1. $R(x_1, \ldots, x_n)$, ahol $x_1, \ldots, x_n$ a komponens változók és R egy relációnév. Rendeljünk R-hez egy T relációt, melynek n oszlopa van. Ennél a hozzárendelésnél az $R(x_1, \ldots, x_n)$ atom igaz az $x_1, \ldots, x_n$ változók aktuális értékeire akkor és csak akkor, ha az értékekből álló sor T-nek valamelyik sorával egyezik meg.
2. $u=v$, ahol u, v komponens változók és $=$ az egyenlőség operátor. Az atom akkor és csak akkor igaz az u és v aktuális értékeire, ha teljesül rájuk az egyenlőség.

Az operátorok a szokásos $\wedge$, $\vee$, $\neg$ logikai operátorok és a $\exists$, $\forall$ kvantorok. Ezeket rendszerint szabad vagy kötött változókkal használjuk, és általában rögzítünk valamilyen kiértékelési sorrendet közöttük és a formulákat elválasztó zárójelek között.

Legyen φ egy formula, melynek a szabad komponens változói az $x_1, x_2, \ldots, x_n$ (jelölésben $\varphi(x_1, x_2, \ldots, x_n)$). Legyenek $R_1, \ldots, R_k$ a φ formulában előforduló reláció nevek. Minden olyan leképezésnél, mely az $R_1, R_2, \ldots, R_k$ reláció nevekhez $T_1, T_2, \ldots, T_k$ relációkat rendel, hozzárendelhetünk a φ-hez egy univerzális értelmezési tartományt, mégpedig a $T_1, T_2, \ldots, T_k$ értelmezési tartományainak unióját. Ezt a hozzárendelést DOM (φ)-vel jelölve látható, hogy

$$\mathrm{DOM}(\varphi) = \{u \mid \exists (x_1, x_2, \ldots, x_n)(\sum_{i,j} R_i(x_1, x_2, \ldots, x_{j-1}, u, x_{j+1}, \ldots, x_n))\}.$$

Egy φ formulát biztonságos *(„safe")* formulának hívunk, ha az alábbi három feltétel teljesül rá:

(i) Minden olyan hozzárendelésnél, mikor a φ-ben szereplő reláció nevekhez relációkat rendelünk, teljesül, hogy valahányszor a φ igaz az értékek egy halmazán, akkor ezek az aktuális értékek valamennyien a DOM (φ)-nek elemei.

(ii) A φ minden $(\exists u)$ $(\omega(u, v_1, \ldots, v_m))$ alakú részformulájára teljesül, hogy ha ω-t kielégíti az u-nak megfelelő x érték és az ω többi szabad változójának, $v_1, v_2, \ldots, v_m$-nek megfelelő valamilyen érték, akkor $x \in \mathrm{DOM}(\omega)$.

(iii) A φ minden $(\forall u)$ $(\omega(u, v_1, \ldots, v_m))$ alakú részformulájára teljesül, hogy bármely $x \notin \mathrm{DOM}(\omega)$ értékre $u=x$ kielégíti az ω-t a többi szabad változójának tetszőleges értéke mellett.

Egy relációs kalkulus kifejezés $\{x_1, \ldots, x_n \mid \varphi(x_1, \ldots, x_n)\}$ alakú, mely egy relációt határoz meg abban az esetben, ha a φ-ben szereplő reláció nevekhez relációkat rendelünk; méghozzá azt a relációt, amely azokból az n-esekből áll, amelyekre φ igaz.

A feltétel-következmény formula (HC-formula) két olyan $\varphi(x_1, \ldots, x_N)$ és $\psi(x_1, x_2, \ldots, x_N)$ biztonságos részformulát tartalmaz, melyeknek ugyanazok a szabad változói. A φ és ψ csak egy reláció nevet tartalmazhat, mondjuk R-et. Legyen R változóinak száma n.

A HC-formula a fenti feltevésekkel élve

$$\forall (x_1, \ldots, x_N)(\neg \varphi(x_1, \ldots, x_N) \vee \psi(x_1, \ldots, x_N))$$

alakú.

Ha T egy n-változós reláció, amelyet az R reláció névhez rendeltünk, akkor a formula vagy igaz, vagy hamis a T-re. Ha igaz, akkor a T kielégíti a φ és ψ által megadott függőséget. Könnyű megadni egy olyan kalkulus kifejezést, amely a T relációt adja

vissza abban az esetben, ha T kielégíti a formulát, és az üres relációt adja meg különben:

$$\{u_1, \ldots, u_n | R(u_1, u_2, \ldots, u_n) \wedge (\forall (x_1, \ldots, x_N)(\neg \varphi(x_1, \ldots, x_N) \vee \psi(x_1, \ldots, x_N)))\}.$$

Az ekképp megadott függőséget HCD (φ, ψ)-vel jelöljük és feltétel-következmény függőségnek, (HCD)-nek nevezzük.

Gyakorlati szempontból nézve egy függőségtől elvárjuk, hogy ne legyen érzékeny az értelmezési tartományok leképezéseire, más szóval az értelmezési tartományok elemeinek kódolására. Ahhoz, hogy ezt a tulajdonságot vizsgálhassuk, be kell vezetnünk a relációk érték transzformáló leképezéseinek fogalmát, nevezetesen a szimbólum leképezést és a kiértékelést *(„symbol mapping", „valuation")*.

Legyen R, S két reláció, melyek ugyanolyan $U=\{A_1, \ldots, A_n\}$ típusúak, de különbözik az értelmezési tartományuk, azaz $R \subseteq D_1 \times D_2 \times \ldots \times D_n$ és $S \subseteq E_1 \times E_2 \times \ldots \times \ldots \times E_n$. Legyen h_i $(i=1, 2, \ldots, n)$ D_i-ből E_i-be képező függvény. Abban az esetben, mikor adott az egyenlőség reláció a különböző értelmezési tartományok között, elég olyan $h_i (i=1, \ldots, n)$ függvényekre szorítkoznunk, amelyek egy $\bigcup_{i=1}^{n} D_i$-ből $\bigcup_{i=1}^{n} E_i$-be képező h függvénynek a megfelelő D_i-halmazokra vett megszorításai. Ebben az esetben a h-t az $\bigcup_{i=1}^{n} D_i$-ből $\bigcup_{i=1}^{n} E_i$-be képező szimbólum leképezésnek hívjuk. (Ezt a terminológiát használja ULLMAN [9]).

Ha nem akarjuk összehasonlítani a különböző attributumok értékeit, akkor a $h_1, \ldots, h_n$ függvények egy független leképezést definiálnak az attributum értékekre, vagyis definiálhatunk egy h függvényt úgy, hogy $h(x) := h_i(x)$, ha $x \in \mathrm{VAL}_R(A_i)$. Ebben az esetben a h-t kiértékelésnek hívjuk. (Ezt a terminológiát használja BEERI, VARDI [1], [2]-ben.)

Mindkét esetben a h függvény az R egy $t=(a_1, \ldots, a_n)$ sorát egy $h(t)=(h(a_1), \ldots, h(a_n))$ sorra képezi le. Ha $h(t) \in S$ minden $t \in R$ esetén, akkor azt mondjuk, hogy a h az R relációt az S relációba képezi. (Jelölésben $h(R) \subseteq S$.) Hasonlóan használjuk a $h(R) \supseteq S$ jelölést. A $h(R)=S$ jelentse a $h(R) \subseteq S$ és a $h(R) \supseteq S$ egyidejű fennállását. Azt mondjuk, hogy R és S gyengén ekvivalensek, ha létezik olyan h és g kiértékelés, melyre $h(R)=S$ és $g(S)=R$.

(Ebben az esetben a h_i és g_i függvények egy-egy értelmű leképezések $(i=1, 2, \ldots, n)$ lesznek $\mathrm{VAL}_R(A_i)$ és $\mathrm{VAL}_S(A_i)$ között, ha az értelmezési tartományok végességét is feltesszük, továbbá h és g is egy-egy értelmű leképezés lesz R és S között.)

Azt mondjuk, hogy R és S ekvivalensek, ha léteznek olyan h és g szimbólum leképezések, melyekre $h(R)=S$ és $g(S)=R$. (Véges értelmezési tartományok esetén h és g egy-egy értelmű leképezések R és S között.) Világos, hogy ha R és S ekvivalensek, akkor gyengén is ekvivalensek.

Az S reláció h leképezésnél vett inverz képét definiáljuk úgy, hogy legyen R az a maximális reláció, amelyre $h(R)=S$. (Jelölésben $h^{-1}(S) := R$.)

Az ekvivalencia definíciójából rögtön következik, hogy véges relációk esetén az ekvivalencia megőrzi a HCD-ket.

Egy atom értéke nyilván nem változik, ha a benne szereplő komponensek eredetileg $\bigcup_{i=1}^{n} \mathrm{VAL}_R(A_i)$-ből vették fel értékeiket, és megváltoztatjuk a hozzárendelt

R relációt S-re, és a komponensek eredeti értékei helyett a h, illetőleg a g^{-1} leképezésnél vett képüket vesszük. Mivel a HC formula biztonságos, ezért a komponensek számára más értékek úgysem jönnek számításba.

A gyenge ekvivalenciára hasonló állítás azonban nem igaz, de megadható a HCD-knek olyan részhalmaza, amely invariáns a gyenge ekvivalenciára nézve.

A φ-re és ψ-re kirótt megszorítások a HC formulában annak a ténynek az egyenes következményei, hogy a gyenge ekvivalenciában a különböző attributumokhoz tartozó értékek között nincs összehasonlítás.

Ebből a megjegyzésből következik, hogy egy komponens változó csak egy attributumhoz tartozhat. Azt mondjuk, hogy x egy A_i-változó, ha x egy első típusú atom A_i attributumnak megfelelő helyén áll. A második természetes megszorítás az, hogy egy második típusú atomban a változók ugyanahhoz az attributumhoz tartozzanak Az ilyen típusú formulát attributum korlátozottnak hívjuk, és a megfelelő HCD-k osztályát AHCD-vel jelöljük.

Az ekvivalencia eseténél látott meggondoláshoz hasonlóan meg lehet mutatni, hogy az AHCD-k osztálya invariáns a gyenge ekvivalenciára nézve.

3. Általánosított függőségek

Ebben a részben csak olyan speciális típusú HCD-ket vizsgálunk, amelyekre az teljesül, hogy a φ és a ψ a logikai operátorok közül csak a $\vee$ diszjunkciót tartalmazza, továbbá φ nem tartalmazza $\exists$ egzisztenciális kvantort. Ez azt jelenti, hogy a φ formula első típusú atomok uniója, ugyanis ebben az esetben a második típusú atom annak felel meg, hogy a két változó ekvivalens, és így valamelyiküket a másikkal lehet helyettesíteni. A φ-ben szereplő $\exists$ egzisztenciális kvantort ki lehet vinni $\forall$ univerzális kvantorként a φ elé, ugyanis ha $\varphi = \varphi_1 \wedge (\exists v)\varphi_2$, akkor

$$(\forall x_1, \ldots, x_n)\big(\neg(\varphi_1 x_1, \ldots, x_n) \wedge (\exists v)(\varphi_2(x_1, \ldots, x_n, v)) \vee \psi(x_1, \ldots, x_n)\big) \equiv$$

$$\equiv (\forall x_1, \ldots, x_n, v)\big(\neg(\varphi_1(x_1, \ldots, x_n) \wedge \varphi_2(x_1, \ldots, x_n, v)) \vee \psi(x_1, \ldots, x_n)\big).$$

A ψ következmény formulában a második típusú atomok egyenlőség generátorok. A ψ-nek azokat a változóit, amelyek egy $\exists$ kvantorral vannak lekötve, egyedi (*„unique"*) változóknak hívjuk. A második típusú atomok egyedi változói helyettesíthetők φ-beli változókkal, és így elérhető, hogy a második típusú atomokban csak a φ változói szerepeljenek.

Könnyen látható, hogy ha $\psi_2 = (x_i = x_j)$, akkor

$$\mathrm{SAT}(\mathrm{HCD}(\varphi, \psi)) \cap \mathrm{SAT}(\mathrm{HCD}(\varphi, \psi_2)) = \mathrm{SAT}(\mathrm{HCD}(\varphi, \psi \wedge \psi_2)).$$

Az egyszerűség kedvéért feltesszük, hogy a ψ formula vagy csak első típusú atomokat tartalmaz, vagy csak egy második típusú atomot. Így a φ formula általános alakja a következő:

$$\varphi(x_1, \ldots, x_N) = R(u_{11}, \ldots, u_{1n}) \wedge R(u_{21}, \ldots, u_{2n}) \wedge \ldots \wedge R(u_{k1}, \ldots, u_{kn}),$$

ahol $u_{ij} = x_l$ valamilyen l-re.

A ψ általános alakja pedig a következő:

$$\psi(x_1, \ldots, x_N) = (\exists y_1, \ldots, y_M)\psi_1(x_1, \ldots, x_N, y_1, \ldots, y_M) =$$
$$= (\exists y_1, \ldots, y_M)R(v_{11}, \ldots, v_{1n}) \wedge \ldots \wedge R(v_{m1}, \ldots, v_{mn}) \wedge$$
$$\wedge (z_{11} = z_{12}) \wedge (z_{21} = z_{22}) \wedge \ldots \wedge (z_{p1} = z_{p2})$$

ahol a v-k és z-k valamilyen indexű x-ek és y-ok. Az ilyen korlátozott HCD-knek egy ekvivalens formáját adhatjuk meg az úgynevezett általánosított függőségek táblázatos reprezentációjának segítségével.

Legyen I a φ-nek megfelelő táblázat (vagy reláció):

$$I = \begin{bmatrix} u_{11} \ldots u_{1n} \\ \vdots \quad \vdots \\ u_{k1} \ldots u_{kn} \end{bmatrix}.$$

Hasonlóan kapjuk a J relációt a ψ-ben szereplő első típusú atomokból:

$$J = \begin{bmatrix} v_{11} \ldots v_{1n} \\ \vdots \quad \vdots \\ v_{m1} \ldots v_{mn} \end{bmatrix}.$$

Egy T reláció esetén $T \in \mathrm{SAT}(\mathrm{HCD}(\varphi, \psi))$ akkor és csak akkor teljesül, ha minden olyan h szimbólum leképezésre, amelyre $h(I) \subseteq T$, létezik a h-nak olyan h' kiterjesztése a J-re (vagyis az $y_1, \ldots, y_M$ egyedi változókra), amelyre $h'(J) \subseteq T$ és $h'(z_{j1}) = h'(z_{j2})$ teljesül $(j = 1, \ldots, p)$.

Az általánosított függőségek elméletében a szimbólum leképezés ugyanazt a szerepet játssza, mint a kalkulus formulák esetében a komponens változók értékadása. Abban az esetben, mikor a ψ nem tartalmaz egyenlőség generátort, a megfelelő HCD-t sor-generáló függőségnek nevezzük és TGD-vel jelöljük.

Abban az esetben, mikor a ψ csak egy egyenlőség generátorból áll, a megfelelő HCD-t egyenlőség generáló függőségnek nevezzük és EGD-vel jelöljük.

Az egyenlőség generátorokat szét lehet választani, ugyanis ha $\psi = \psi_1 \wedge (x_i = x_j)$, akkor

$$\mathrm{SAT}(\mathrm{HCD}(\varphi, \psi)) = \mathrm{SAT}(\mathrm{HCD}(\varphi, \psi_1)) \cap \mathrm{SAT}(\mathrm{HCD}(\varphi, (x_i = x_j))).$$

Hasonlóan, ha $R(u_1, \ldots, u_n)$ nem tartalmaz egyedi változókat, és $\psi = \psi_1 \wedge \wedge (R(u_1, \ldots, u_n))$, akkor

$$\mathrm{SAT}(\mathrm{HCD}(\varphi, \psi)) = \mathrm{SAT}(\mathrm{HCD}(\varphi, \psi_1)) \cap \mathrm{SAT}(\mathrm{HCD}(\varphi, R(u_1, \ldots, u_n))).$$

Abban az esetben, mikor a J két sorában illetőleg a ψ két első típusú atomjában közös egyedi szimbólumok vannak, akkor az egyedi változókat nem lehet szétválasztani. Ez a tulajdonság egy ekvivalencia relációt definiál a J táblázat sorain, és a sorokból álló ekvivalencia osztályokat már szét lehet választani a fenti értelemben. (Ezt AHCD-kre bizonyította BEERI és VARDI [2]-ben.)

A bizonyítás az alábbi relációs komponens kalkulus kifejezések ekvivalenciáján alapul:

$$\{u_1, \ldots, u_n | R(u_1, \ldots, u_n) \wedge (\forall (x_1, \ldots, x_N)(\neg\varphi \vee \psi_1)) \wedge (\forall (x_1, \ldots, x_N)(\neg\varphi \vee \psi_2))\} \equiv$$
$$\equiv \{u_1, \ldots, u_n | R(u_1, \ldots, u_n) \wedge (\forall (x_1, \ldots, x_n)(\neg\varphi \vee (\psi_1 \wedge \psi_2)))\}.$$

Az első kifejezés a SAT (HCD $(\varphi, \psi_1)) \cap$ SAT (HCD (φ, ψ_2))-t helyettesíti , a második kifejezés a SAT $(\varphi, \psi_1 \wedge \psi_2)$-t. Fontos megjegyezni, hogy $x_1, \ldots, x_N$ a ψ_1 és ψ_2 összes szabad változója, így a ψ_1 é s ψ_2-ben szereplő többi változó mind kötött változó.

Egy

$$\exists(y_1, \ldots, y_M)\psi(x_1, \ldots, x_N, y_1, \ldots, y_M) = \exists(y_1, \ldots, y_k)\psi_1(x_1, \ldots, x_N, y_1, \ldots, y_k) \wedge$$

$$\wedge \exists(y_{k+1}, \ldots, y_M)\psi_2(x_1, \ldots, x_N, y_{k+1}, \ldots, y_M)$$

típusú formulát felbonthatunk két részformula uniójára, és ez pont a J sorainak ekvivalencia osztályozását jelenti. Az AHCD-k esetében az AHCD (φ, ψ)-nek megfeleltetett táblázat olyan, hogy a különböző attributumokhoz tartozó értelmezési tartományok diszjunkt halmazok, és az egyenlőség generátorok csak ugyanannak az attributumnak a szimbólumait tehetik egyenlővé.

4. Attributum korlátozott általánosított függőségek (AGD-k):

— sor generáló függőségek (TGD)
— totális sor generáló függőségek (TTGD)
— egyenlőség generáló függőségek (EGD)
— felcserélhetőség generáló függőségek (XGD)

Az attributum korlátozott függőségekre vonatkozó implikációs problémát számos cikk vizsgálja és ULLMAN könyvében ([9]-ben) is található említés róla. Ebben a részben a függőségek implikációjához használt legismertebb technikát, az üldöző *(„chase")* eljárást fogjuk leírni és részletesen elemezni. A célunk az, hogy minimálisra csökkentsük az EGD-k szerepét, és hogy megvizsgáljuk azoknak a relációknak a struktúráját, amelyeknek közös a magjuk, és ez a mag EGD-knek egy halmazát elégítik ki.

Az AGD-k alapvető típusainak definiálásakor BEERI és VARDI [2]-ben használt terminológiáját követjük. Ezek közül az első három függőséget az előző rész szerint elemi függőségnek hívhatjuk, ami azt jelenti, hogy tetszőleges AGD-t ilyen típusú függőségekre lehet szétbontani, és az ilyen típusú függőségek már tovább nem bonthatók a φ feltétel mellett.

A *sor generáló függőség* (TGD) formailag egy $\langle I, J \rangle$ reláció pár úgy, hogy I bármely sorához létezik egy másik sor J-ben, melyeknek legalább egy közös egyedi értékük van. Egy R reláció akkor és csak akkor elégíti ki az $\langle I, J \rangle$TGD-t, ha minden olyan h kiértékelést, melyre $h(I) \subseteq R$ teljesül, ki lehet terjeszteni a J-re úgy, hogy $h'(J) \subseteq R$ is teljesüljön.

A *totális sor generáló függőség* (TTGD) a TGD-nek olyan speciális esete, melyben a J csak egyetlen w sorból áll és $w[A_i] \in \mathrm{VAL}_I(A_i)$ $i = 1, 2, \ldots, n$.

Az *egyenlőség generáló függőség* (EGD) formálisan egy $\langle I, (a, b) \rangle$ pár, ahol I egy reláció és a, b a $\mathrm{VAL}_I(A_i)$ elemei, valamely rögzített A_i attributumra. Ha hangsúlyozni akarjuk, hogy egy EGD az A_i oszlopon hat, akkor az A_i-EGD jelölést használjuk. Egy R reláció akkor és csak akkor elégíti ki az $\langle I, (a, b) \rangle$ EGD-t, ha minden olyan h kiértékelésre, amelyre $h(I) \subseteq R$, teljesül a $h(a) = h(b)$ egyenlőség.

A *felcserélhetőség generáló függőség* (XGD) a TTGD-nek egy olyan speciális esete, amely az EGD-k általánosításának tekinthető. Egy X felcserélhetőség generáló függőség formálisan egy $\langle I \cup (x_1, \ldots, x_{i-1}, a, x_{i+1}, \ldots, x_n), (x_1, \ldots, x_{i-1}, b, x_{i+1}, \ldots, \ldots, x_n)\rangle$ alakú pár. Látható, hogy minden XGD-nek formálisan megfeleltethető egy $\langle I, (a, b)\rangle$EGD. Ha egy E EGD és egy X XGD között ilyen kapcsolat áll fenn, akkor az E_X, illetve az X_E jelöléssel fogjuk ezt kifejezni.

Példák függőségekre

a) Egy R relációs adatbázisban az A, B, C attributumok jelentsék rendre a vállalati alkalmazottak, vezetők azonosítóit, telefonszámokat.

Írjuk le formális eszközökkel az alábbi függőségeket:

1. Minden vezető alkalmazott;
2. Minden vezető csak egy telefonszámon érhető el;
3. Minden alkalmazott csak egy vezetőhöz van beosztva.

Megoldás:

Az $R[B] \subseteq R[A]$ azt fejezi ki, hogy az R relációban a B attributumon felvett értékek részhalmazát képezik az A attributumon felvett értékeknek, vagyis a feladat 1. részét írja le.

A feladat 2. része olyan feltételes funkcionális kapcsolatot tartalmaz, mely azt mondja, hogy ha egy alkalmazott egyúttal vezető is (azaz $t \in R$ és $t[A] \in R[B]$), akkor az R reláció ilyen t soraiból alkotott R' részrelációján fennáll az $A \to C$ funkcionális függőség, hiszen ha t_1, $t_2 \in R'$ és $t_1[A] = t_2[A]$ (azonos vezető alkalmazott esete), akkor $t_1[C] = t_2[C]$ kell hogy teljesüljön. A feladat 2. része a relációs komponens kalkulus segítségével az alábbi formulával írható le:

$$\varphi\colon \forall(a, a_1, b_1, b_2, c_1, c_2, c_3)([R(a, b_1, c_1) \wedge R(a, b_2, c_2) \wedge R(a_1, a, c_3)] \to c_1 = c_2).$$

A feladat 3. részét az $A \to B$ funkcionális függőséggel tudjuk megfogalmazni.

Ha tehát egy R reláció eleget tesz a példa mindhárom feltételének, akkor ez azzal ekvivalens, hogy az R kielégíti az $\mathcal{F} = \{R[B] \subseteq R[A], \varphi, A \to B\}$ függőségi rendszert. A feladatban szereplő $A \to B$ funkcionális függőséget az alábbi EGD-vel helyettesíthetjük:

A	B
a_1	b_1
a_1	b_2
$b_1 = b_2$.	

A feladat 2. része pedig a φ formula helyett az alábbi EGD-vel írható le:

A	B	C
a	b_1	c_1
a	b_2	c_2
a_1	a	c_3
$c_1 = c_2$.		

b) Az A, B, C attributumok jelentése legyen azonos az a) feladatban adottakkal.

Írjuk le formális eszközökkel az alábbi megszorítást: minden alkalmazott csak olyan telefonszámon érhető el, melyen vezető is elérhető.

A megoldás TGD alkalmazásával a következő:

A	B	C	
a	b_1	c_1	$\}I$
a_1	b_2	c_1	$\}J$.
a_2	a_1	c_2	

Az I feltétel rész azt jelenti, hogy ha az a azonosítóhoz rendelt alkalmazott c_1-hez rendelt telefonszámon elérhető, akkor a J következmény szerint létezik olyan a_1 azonosítóhoz rendelt alkalmazott, mely vezető, s ezen vezető elérhető a c_1-hez rendelt telefonszámon.

$\langle I, (a, b)\rangle$ EGD alkalmazása egy R relációra

Legyen $\mathscr{H}$ az I-ből R-be képező kiértékelések halmaza. Ha R végtelen reláció, akkor fel kell tenni, hogy megszámlálhatóan végtelen sok sora van, és hogy az értelmezési tartományok teljesen rendezett jólrendezett halmazok. Mivel ekkor a $\mathscr{H}$ is megszámlálható halmaz, ezért megindexelhetjük az elemeit valamilyen felsorolásnak megfelelő sorrendben. Legyen h_n az első olyan kiértékelés, amelyre $x := h_n(a) \neq h_n(b) := y$ teljesül, legyen továbbá g_1 az a kiértékelés, amelyre $g_1(x) = g_1(y) = \min(x, y)$, különben a g_1 az identitással egyezik meg az R-en.

Hasonlóan folytatva, ha már adott a g_n kiértékelés, akkor legyen m az az első olyan index, amelyre $u := g_n(h_m(a)) \neq g_n(h_m(b)) := v$. Ekkor definiáljuk g_{n+1}-et úgy, hogy legyen azonos g_n-nel az R-en, kivéve az u és v értékeket, ahol legyen $g_{n+1}(u) = g_{n+1}(v) = \min(u, v)$. Ezzel az eljárással egy olyan g_n sorozatot nyerünk, amely figyelembe véve az értelmezési tartományok jólrendezettségét és ennek azt a következményét, hogy egy értéket csak véges sokszor lehet kisebbre cserélni, egyértelműen meghatároz egy g kiértékelést az R-en.

A fent leírt eljárást hajtsuk végre most úgy, hogy az R reláció helyett a $g(R)$ relációt vesszük, és így tovább addig, amíg az eredmény reláció már kielégíti az $\langle I, (a, b)\rangle$ függőséget.

A minimalizálások végessége miatt az eredmény egy egyértelműen meghatározott kiértékelés lesz az R-en.

$\langle I, w\rangle$ TTGD alkalmazása egy R relációra

Jelöljük $\langle I, w\rangle$-t E-vel. Legyen $\mathscr{H}$ az előbbiekben definiált halmaz. Minden egyes $h_n \in \mathscr{H}$ kiértékelésre vegyük hozzá R-hez a $h_n(w)$ sort, ha ez a sor nem szerepelt eddig az R-ben. Ha ezt minden $\mathscr{H}$-beli kiértékeléssel elvégeztük, akkor az R relációnak egy R_1 bővítését kapjuk.

Most hajtsuk végre ugyanezt az eljárást az R helyett az R_1 relációval és így tovább mindaddig, amíg egy olyan $E(R)$ relációt nem kapunk, amely már kielégíti az E függőséget.

Az $E(R)$ reláció egyértelműsége abból a tényből fakad, hogy igaz az alábbi felírás:

$$E(R) = \bigcap_{\substack{R \subseteq T \\ T \in \mathrm{SAT}(E)}} T.$$

Ennek a bizonyítása nyilvánvaló, hiszen ha $T_1 \in \mathrm{SAT}(E)$ és $T_2 \in \mathrm{SAT}(E)$, akkor $T_1 \cap T_2 \in \mathrm{SAT}(E)$, valamint, ha $T \in \mathrm{SAT}(E)$, és $R \subseteq T$, akkor $R_1 \subseteq T$.

Az $E(R)$ reláció létezéséhez ezért elég egy R-et tartalmazó E-t kielégítő relációt mutatni, nevezetesen a $\mathrm{VAL}_R(A_1) \times \ldots \times \mathrm{VAL}_R(A_n)$ reláció ilyen.

$\langle I, J\rangle$ TGD alkalmazása egy R relációra

Jelöljük $\langle I, J\rangle$-t E-vel és legyen $\mathscr{H}$ ugyanaz, mint az előbbiekben. Először ellenőrizzük, hogy h_1-nek van-e olyan J-re történő h_1' kiterjesztése, amelyre $h_1'(J) \subseteq R$ teljesül. Ha nincs ilyen h_1' kiterjesztés, akkor válasszunk olyan új, különböző értéket a J egyedi változóinak, amelyek nagyobbak a változónak megfelelő attributum bármely R-beli értékénél. Legyen h_1' az a kiterjesztése h_1-nek, amely a J egyedi változóihoz ezeket az új értékeket rendeli. Legyen R_{h_1} az a reláció, amelyet úgy kapunk, hogy az R relációhoz hozzávesszük a $h_1'(J)$ reláció sorait. Hasonlóan folytatva az eljárást, ha már az R_{h_n} relációt megkaptuk, akkor először ellenőrizzük, hogy van-e a h_{n+1}-nek olyan J-re történő h_{n+1}' kiterjesztése, amelyre $h_{n+1}'(J) \subseteq R_{h_n}$ teljesül. Ha nincs ilyen kiterjesztés, akkor válasszunk olyan új, különböző értékeket a J egyedi változóihoz, amelyek nagyobbak a változónak megfelelő attributum bármely R_{h_n}-beli értékénél. Az így kapott kiterjesztést h_{n+1}'-gyel jelölve vegyük hozzá R_{h_n}-hez a $h_{n+1}'(J)$ reláció sorait. Ha ezt az eljárást a $\mathscr{H}$ minden elemére elvégezzük, akkor az R-nek egy véges vagy végtelen R_1 bővítését kapjuk. (Itt jegyezzük meg, hogy az értelmezési tartományok ily módon történő kiterjesztése megőrzi a jólrendezettséget.) Ezután hajtsuk végre az eljárást az R helyett az R_1 relációra és így tovább. Legyen $E(R) = \bigcup_{i=1}^{\infty} R_i$. $E(R)$ kielégíti az E függőséget, ugyanis legyen h egy I-ből $E(R)$-be ható leképezés. Ekkor $h(I) \subseteq R_k$ valamilyen k-ra. Így $h(I) \subseteq R_s$ teljesül minden $s \geq k$-ra. Mivel h olyan leképezés, amilyeneket az R_{k+1} reláció konstruálásánál használunk, ezért van olyan J-re történő h' kiterjesztése, amelyre $h'(J) \subseteq R_{k+1}$ teljesül. Tehát $E(R) \in \mathrm{SAT}(E)$.

Az üldözéses eljárás („chase")

A következőkben leírjuk azt az általános eszközt, amelyet az általánosított függőségek implikációs problémájához döntési eljárásként használhatunk.

Legyen $\mathscr{F}=\{F_1, \ldots, F_k\}$ függőségek halmaza, ahol az F_i-k EGD-k, TGD-k és TTGD-k is lehetnek. Továbbá legyen F egy I feltételhez tartozó EGD, TGD vagy TTGD. Feltesszük, hogy $\mathrm{VAL}_I(A_i)$ teljesen rendezett halmaz, és ha F egy $\langle I, (x, y)\rangle A_i$–EGD, akkor x a $\mathrm{VAL}_I(A_i)$ legkisebb értéke. Alkalmazzuk I-re az $F_1, \ldots, F_k$ függőségeket tetszőleges sorrendben, ciklikus módon. A TGD-k alkalmazására azonban egy megszorítást kell tennünk: nem használhatunk fel újra egy olyan értéket, amelyet valamely EGD alkalmazásakor már egyenlővé tettünk egy kisebb értékkel. Jelöljük az eljárás során kapott relációkat $I_1=F_{i_1}(I)$, $I_2=F_{i_2}(I_1)$, $I_3=F_{i_3}(I_2)$, ..., $I_m=F_{im}(I_{m-1})$, ...-gyel.

Definiáljuk a $\mathrm{Chase}_{\mathscr{F}}(I)$ relációt az alábbi módon: egy t sor akkor és csak akkor tartozzon a $\mathrm{Chase}_{\mathscr{F}}(I)$-hez, ha van olyan $m(t)$ index, hogy $t\in \bigcap_{i=m(t)}^{\infty} I_i$, vagyis a t sor a függőségek alkalmazásának során egy bizonyos lépéstől kezdve már nem változik. Ha $\mathscr{F}$-ben nincs EGD, akkor $I\subseteq \mathrm{Chase}_{\mathscr{F}}(I)$. Ha $\mathscr{F}$-ben EGD is van, akkor az EGD-k alkalmazása miatt az I reláció megváltozhat, így $\mathrm{Chase}_{\mathscr{F}}(I)$-ben az I helyett egy $g(I)$ reláció szerepel valamilyen d kiértékelés hatására.

A TGD-k alkalmazásánál mutatott módon be lehet bizonyítani, hogy $\mathrm{Chase}_{\mathscr{F}}(I)\in \in \mathrm{SAT}(\mathscr{F})$. Ehhez csak azt kell még megjegyezni, hogy ha egy lépésben egy F_i TGD-t vagy TTGD-t alkalmazunk és ezáltal a $h'(J_i)$ vagy a $h(w_i)$ sorokat kell hozzávennünk az eddigi relációhoz, akkor ezek a sorok a $\mathrm{Chase}_{\mathscr{F}}(I)$-ben $g(h'(J_i))$, illetve $g(h(w))$ formában fognak megjelenni, ami azt mutatja, hogy a $g(h(I_i))$ nem sérti meg az F_i-t.

4.1. TÉTEL. Az üldözéses eljárás a következő értelemben bizonyítja az $\mathscr{F}\models F$ implikációt:

(i) Ha F egy $\langle I, (x, y)\rangle$EGD, akkor $\mathscr{F}\models F$ akkor és csak akkor teljesül, ha minden $\mathrm{Chase}_{\mathscr{F}}(I)$-re a végső g kiértékelést véve $g(x)=g(y)$ teljesül.

(ii) Ha F egy $\langle I, w\rangle$TTGD, akkor $\mathscr{F}\models F$ akkor és csak akkor teljesül, ha minden $\mathrm{Chase}_{\mathscr{F}}(I)$-re a végső g kiértékelést véve $g(w)\in \mathrm{Chase}_{\mathscr{F}}(I)$ teljesül.

(iii) Ha F egy $\langle I, J\rangle$TGD, akkor $\mathscr{F}\models F$ akkor és csak akkor teljesül, ha minden $\mathrm{Chase}_{\mathscr{F}}(I)$-re létezik a végső g kiértékelésnek olyan g' kiterjesztése a J-re, hogy $g'(J)\subseteq \mathrm{Chase}_{\mathscr{F}}(I)$ teljesül.

Az állításokban a szükségesség bizonyítása nyilvánvaló, mert a $\mathrm{Chase}_{\mathscr{F}}(I)$ ellenpéldaként szolgálna az indirekt feltevéshez. Az elégségességet a következő részben fogjuk megmutatni.

5. XGD-ket kielégítő relációk osztályozása

5.1. TÉTEL. Legyen $\mathscr{F}=\{E_1, \ldots, E_N\}$ EGD-knek egy halmaza és legyen $\mathscr{F}^*=\{X_{E_1}, X_{E_2}, \ldots, X_{E_N}\}$ az E_i EGD-knek megfelelő X_{E_i} XGD-ből álló halmaz. Az $\mathscr{F}^*$-ot kielégítő relációk $\mathrm{SAT}(\mathscr{F}^*)$ halmaza felbontható olyan diszjunk osztályokra, hogy minden osztály a gyenge ekvivalencia értelmében pontosan egy $\mathrm{SAT}(\mathscr{F})$-beli relációt

tartalmaz, amelyet az osztály magjának *(„kernel")* hívunk és az osztályozásra a következő tulajdonságok teljesülnek:

(i) $\mathrm{SAT}(\mathscr{F}^*)=\bigcup_{R\in\mathrm{SAT}(\mathscr{F})}\mathrm{Class}(R)$ ahol $R_1\not\equiv R_2$ esetén $\mathrm{Class}(R_1)\cap\mathrm{Class}(R_2)=\emptyset$.

(ii) Ha $S\in\mathrm{Class}(R)$, akkor van olyan h kiértékelés, amelyre $h(S)=R$ és $h^{-1}(R)=S$.

(iii) Ha $S\in\mathrm{Class}(R)$ és $S\in\mathrm{SAT}(G)$, ahol G egy TGD, akkor $\mathrm{Class}(R)\subseteq\mathrm{SAT}(G)$.

(iv) Ha E egy A_i-EGD és van $\mathscr{F}$-ben A_i-EGD, akkor tetszőleges $R\in\mathrm{SAT}(\mathscr{F})$ relációra az $R\in\mathrm{SAT}(E)$ akkor és csak akkor teljesül, ha $\mathrm{Class}(R)\subseteq\mathrm{SAT}(X_E)$.

(v) Ha E egy olyan A_i-EGD, hogy $\mathscr{F}$-ben nincsen A_i-EGD, akkor ha $S\in\mathrm{Class}(R)$ és $S\in\mathrm{SAT}(E)$, akkor $\mathrm{Class}(R)\subseteq\mathrm{SAT}(E)$.

Az 5.1. tétel lényegében azt fejezi ki, hogy a függőségek osztálytulajdonságok, így elég a függőséget az osztály egy tetszőleges elemére bizonyítani, kivéve a (iv) esetet, ahol ezt az elemet az osztály magjának kell választani.

Az 5.1. tétel következményeként adódik az alábbi tétel.

5.2. TÉTEL. Legyen $\mathscr{G}$ egy TGD-kből és EGD-kből álló halmaz és $\mathscr{G}^*$ legyen az a halmaz, amelyet a $\mathscr{G}$-ből úgy kapunk, hogy az összes $\mathscr{G}$-ben szereplő EGD-t kicseréljük a megfelelő XDG-re. Ekkor egy G TGD-re a $\mathscr{G}\models G$ akkor és csak akkor teljesül, ha a $\mathscr{G}^*\models G$ implikáció teljesül. Továbbá, ha E egy olyan A_i-EGD, hogy $\mathscr{G}$-ben van A_i-EGD, akkor $\mathscr{G}\models E$ implikáció akkor és csak akkor teljesül, ha $\mathscr{G}^*\models X_E$ teljesül.

Megjegyezzük, hogy az EGD-kre tett feltétel valójában nem jelent megszorítást, mivel a $\mathscr{G}\models E$ nem teljesülhet, ha E egy olyan A_i-EGD, hogy $\mathscr{G}$ nem tartalmaz semmilyen A_i-EGD-t. Ugyanis az

$$S_i=\begin{Bmatrix}x_1,\ldots,x_{i-1},a,x_{i+1},\ldots,x_n\\ x_1,\ldots,x_{i-1},b,x_{i+1},\ldots,x_n\end{Bmatrix}$$

két soros reláció minden TGD-t és EGD-t kielégít, kivéve az A_i-EGD-ket.

Az 5.1. tétel bizonyítása. A bizonyítást a $\mathrm{SAT}(E)$ és $\mathrm{SAT}(X_E)$ halmazok vizsgálatával kezdjük, ahol E egy $\langle I,(a,b)\rangle$ alakú A_i-EGD, és X_E a megfelelő $\langle I\cup\{t_a\},t_b\rangle$ alakú XGD, ahol $t_a=(x_1,\ldots,x_{i-1},a,x_{i+1},\ldots,x_n)$, illetve $t_b=(x_1,\ldots,x_{i-1},b,x_{i+1},\ldots,x_n)$.

Vegyük észre, hogy $\mathrm{SAT}(E)\subseteq\mathrm{SAT}(X_E)$, mivel ha $R\in\mathrm{SAT}(E)$, akkor minden olyan h kiértékelésre, amelyre $h(I)\subseteq R$ teljesül, $h(a)=h(b)$ is fennáll, így ha $h(I\cup\{t_a\})\subseteq R$ teljesül, akkor $h(t_a)=h(t_b)$-nek is igaznak kell lenni.

5.1. LEMMA. Legyen $R\in\mathrm{SAT}(X_E)$, ahol X_E az $\langle I\cup\{t_a\},t_b\rangle$ alakú XGD és $x,y\in\mathrm{VAL}_R(A_i)$. Azt mondjuk, hogy x és y felcserélhető, ha van olyan t_1, t_2 sor az R-ben, hogy $t_1[A_i]=x$, $t_2[A_i]=y$ és $t_1[U-A_i]=t_2[U-A_i]$. A felcserélhetőséget $x\sim y$-nal jelölve az alábbi tulajdonságok igazak:

(i) Ha $x \sim y$, akkor minden olyan $t \in R$ sorhoz, amelyre $t[A_i]=x$, létezik R-nek olyan s sora, amelyre $s[A_i]=y$ és $s[U-A_i]=t[U-A_i]$.
(ii) $x \sim y$ akkor és csak akkor, ha van olyan h kiértékelés, amelyre $h(I) \subseteq R$ és $h(a)=x$, $h(b)=y$ teljesül.
(iii) A $\sim$ reláció ekvivalencia reláció.

Az 5.1. lemma bizonyítása

(i) Mivel $x \sim y$, ezért létezik R-ben a $\sim$ relációnak megfelelő t_1, t_2 sor. Tekintsük azt a h kiértékelést, amely az I relációt a t_1, t_2 sorokra képezi úgy, hogy $h(a)=x$ és $h(b)=y$ teljesüljön. Terjesszük ki h-t a t_a-ra úgy, hogy t_a-t az adott t sorba képezze. Mivel $R \in \mathrm{SAT}(X_E)$, ezért az $s:=h(t_b)$ sor is R-ben van.
(ii) Ha $h(I) \subseteq R$ és $h(a)=x$, $h(b)=y$, akkor van az I-ben olyan t sor, amelyre $t[A_i]=a$ és $h(t) \in R$. Terjesszük ki a h kiértékelést a t_a sorra a $h(t_a):=h(t)$ definícióval. Ekkor a $h(t_a)$, $h(t_b)$ két olyan sor, mely az $x \sim y$ relációhoz szükséges. Az ellenkező irány az (i) tulajdonság bizonyításából egyszerűen adódik.
(iii) A reflexivitás és szimmetria triviálisan teljesül. A tranzitivitás pedig következik az (i) tulajdonságból. □

Megjegyzés. BEERI és VARDI [2]-ben az $\langle I \cup \{t_a\}, t_b\rangle$, $\langle I \cup \{t_b\}, t_a\rangle$ XGD párokat vezeti be. Erre valójában nincs szükség, mivel az 5.1. lemmából látható, hogy ez a két XGD ekvivalens, s így elég az egyiket használni.

5.2. LEMMA. Legyen $R \in \mathrm{SAT}(X_E)$, ahol X_E az $\langle I \cup \{t_a\}, t_b\rangle$ alakú XGD, és jelöljük $Z_1, \ldots, Z_M$-mel a felcserélhetőség ekvivalencia osztályait a $\mathrm{VAL}_R(A_i)$ halmazon, azaz $\mathrm{VAL}_R(A)_i = \bigcup_{k=1}^{M} Z_k$. Legyen $z_j \in Z_j$ tetszőlegesen kiválasztott elem ($j=1, \ldots, M$). Tekintsük azt a $g=(g_1, g_2, \ldots, g_n)$ kiértékelést, ahol a komponens függvények közül csak a g_i nem az identitás, és $g_i(x):=z_j$, ha $x \in Z_j$. Ez a g kiértékelés a következő tulajdonságokkal bír:

(i) $g(R) \in \mathrm{SAT}(E)$
(ii) $g^{-1}(g(R))=R$
(iii) Ha f olyan kiértékelés, amelyre (i) és (ii) teljesül, továbbá az f egy-egy értelmű leképezés az $U-A_i$-n, akkor $g(R) \equiv f(R)$.

Az 5.2. lemma bizonyítása

(i) Mivel $g(R) \subseteq R$, így ha egy h kiértékelésre a $h(I) \subseteq g(R)$ teljesül, akkor 5.1. lemma (ii) tulajdonságából következik, hogy $h(a) \sim h(b)$, ez viszont azt jelenti, hogy $g(h(a))=g(h(b))$, másrészt $g(h(a))=h(a)$ és $g(h(b))=h(b)$ miatt a $h(a)=h(b)$ egyenlőség is teljesül.
(ii) Az 5.1. lemma (i) tulajdonságából következik, hogy ha egy $t=(y_1, \ldots, y_{i-1}, z_j, y_{i+1}, \ldots, y_n)$ sorra $t \in g(R)$, akkor az összes $(y_1, \ldots, y_{i-1}, z, y_{i+1}, \ldots, y_n)$ alakú sor, ahol $z \in Z_j$, az R relációnak is sora, amiatt, hogy a $t \in R$ is fennáll.

(iii) Mivel f(i) tulajdonságú, ezért az f minden Z_j-n ($j=1, \ldots, M$) állandó, különben lenne az $f(R)$-ben két olyan sor, amelyek az A_i kivételével minden oszlopon megegyeznek, mely viszont ellentmond az E függőségnek. Mivel f(ii) tulajdonsággal is rendelkezik, ezért $f(z)=f(y)$ csak akkor teljesülhet a $\mathrm{VAL}_R(A_i)$-n, ha $z \sim y$, mert különben lenne az R-ben két olyan $t_z = (x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n)$, illetve $t_y=(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n)$ sor, amelyekre $z \sim y$ nem teljesül, viszont mindkét sor az $f(t_z)$ sornak egy inverze.

Mindez azt jelenti, hogy megadható egy egy-egy értelmű leképezés a $g(\mathrm{VAL}_R(A_i))$-ról az $f(\mathrm{VAL}_R(A_i))$-ra. □

5.1. KÖVETKEZMÉNY. Az 5.2. lemmából következik, hogy ha $R \in \mathrm{SAT}(X_E)$ és ha egy (i)—(iii) tulajdonságú g az R-t egy $R_E := g(R)$ relációra képezi, akkor az R_E reláció a gyenge ekvivalencia értelmében egyértelmű, továbbá $R_E \in \mathrm{SAT}(E)$ és $g^{-1}(R_E)=R$. Ily módon tetszőleges $R \in \mathrm{SAT}(X_E)$ relációhoz definiálhatjuk a reláció magját úgy, hogy legyen az R magja egy olyan R_E reláció, melyhez van olyan (i)—(iii) tulajdonságú g leképezés, hogy $R_E=g(R)$.

5.3. LEMMA. Azt mondjuk, hogy az $R, S \in \mathrm{SAT}(X_E)$ relációk hasonlóak, ha megegyezik a magjuk a gyenge ekvivalencia értelmében. Ekkor ez a hasonlósági reláció ekvivalencia reláció. Jelöljük a $\mathrm{SAT}(X_E)$ osztályait $\mathrm{Class}(R_E)$-vel, ahol R_E az osztály magja.

Ekkor

$$\mathrm{SAT}(X_E) = \bigcup_{Q \in \mathrm{SAT}(E)} \mathrm{Class}(Q)$$

és

$$\mathrm{Class}(Q_1) \cap \mathrm{Class}(Q_2) = \emptyset, \quad \text{ha} \quad Q_1 \not\equiv Q_2$$

és

$$\mathrm{Class}(Q_1) = \mathrm{Class}(Q_2), \quad \text{ha} \quad Q_1 \equiv Q_2.$$

Az 5.3. lemma bizonyítása az 5.2. lemmának és 5.1. következménynek egyenes következményeként adódik.

5.4. LEMMA. Legyen $G=\langle K, L\rangle$ egy TGD és $R \in \mathrm{SAT}(X_E)$, $R \in \mathrm{Class}(Q)$. Ekkor vagy $\mathrm{Class}(Q) \subseteq \mathrm{SAT}(G)$ vagy $\mathrm{Class}(Q) \cap \mathrm{SAT}(G)=\emptyset$.

Az 5.4. lemma bizonyítása

1. Először azt bizonyítjuk be, hogy ha egy osztály magja kielégíti a G-t, akkor az osztály minden eleme kielégíti a G-t. Tekintsünk egy $\mathrm{Class}(Q)$ osztályt és legyen $R \in \mathrm{Class}(Q)$. Legyen h egy olyan kiértékelés, amelyre $h(K) \subseteq R$ teljesül. Ekkor az 5.2. lemma alapján tekintsük azt a g leképezést, amelyre $g(R)=Q$. Legyen f a g és h kompozíciója, azaz $f(K)=g(h(K)) \subseteq Q$. Mivel $Q \in \mathrm{SAT}(G)$, ezért az f kiterjeszthető L-re úgy, hogy $f(L) \subseteq Q$. (A jelölések egyszerűsítése végett a kiterjesztést ugyanazzal a függvény szimbólummal jelöljük.) Az 5.2. lemmából következik, hogy $g^{-1}(f(L)) \subseteq R$ és $g^{-1}(f(K))=h(K)$. Így az inverz kiértékelés elemeinek segítségével megadhatjuk a h-nak egy olyan kiterjesztését az L-re, hogy $h(L) \subseteq g^{-1}(f(L))$, ami azt jelenti, hogy $R \in \mathrm{SAT}(G)$.

2. Ezután belátjuk azt, hogy ha egy $R \in \mathrm{SAT}(X_E)$ reláció kielégíti a G-t, akkor a reláció magja is kielégíti. Legyen Q az R reláció magja. Vegyünk egy olyan h kiértékelést, amire $h(K) \subseteq Q$. Tekintsük az 5.2. lemmában megadott g leképezést, melyre $g(R) = Q$. Ekkor $Q \subseteq R$, így $h(K) \subseteq R$ is teljesül.
Mivel $R \in \mathrm{SAT}(G)$, ezért a h-t ki lehet terjeszteni az L-re úgy, hogy $h(L) \subseteq R$ teljesüljön. Másrészt $g(h(K)) = h(K)$ és $g(h(L)) \subseteq Q$, így a $g \circ h$ kompozíció a h-nak egy kiterjesztése az L-re, mely azt jelenti, hogy $Q \in \mathrm{SAT}(G)$. □

5.5. LEMMA. Legyen X_E egy XGD, ahol E egy $A_i - \mathrm{EGD}$ és legyen $\mathrm{Class}(Q)$ a $\mathrm{SAT}(X_E)$-nek egy ekvivalencia osztálya. Legyen továbbá $G = \langle K, (c, d) \rangle$ egy $A_i - \mathrm{EGD}$.

1. vagy $\mathrm{Class}(Q) \subseteq \mathrm{SAT}(X_G)$ és $Q \in \mathrm{SAT}(G)$,
2. vagy $\mathrm{Class}(Q) \cap \mathrm{SAT}(X_G) = 0$.

Az 5.5. lemma bizonyítása

1. Először azt bizonyítjuk be, hogy ha $Q \in \mathrm{SAT}(G)$, akkor egy tetszőleges $R \in \mathrm{Class}(Q)$-ra $R \in \mathrm{SAT}(X_G)$ teljesül. Tekintsük az 5.2. lemmában szereplő g leképezést, amire $Q = g(R)$ és $g^{-1}(Q) = R$. Vegyünk egy olyan h kiértékelést, amelyre $h(K) \subseteq R$ és $h(c) \neq h(d)$. Ekkor $g(h(K)) \subseteq Q$, és mivel $Q \in \mathrm{SAT}(G)$, ezért a $g(h(c)) = g(h(d))$ is teljesül. Ez viszont azt jelenti, hogy a $h(c)$ és a $h(d)$ felcserélhető, így ha az X_G $\mathrm{XGD} \langle \{K \cup \{t_c\}, t_d\rangle$ alakú, akkor a $h(t_c) \in R$ feltételből következik, hogy $h(t_d) \in R$.
2. Ezután azt látjuk be, hogy ha $R \in \mathrm{Class}(Q)$ és $R \in \mathrm{SAT}(X_G)$, akkor $Q \in \mathrm{SAT}(G)$. Legyen h egy olyan kiértékelés, amelyre $h(K) \subseteq Q$. A fenti g leképezést használva $Q = g(R) \subseteq R$, így $h(K) \subseteq R$. Tegyük fel, hogy $h(c) \neq h(d)$. Mivel $R \in \mathrm{SAT}(X_G)$, ezért a h tetszőleges kiterjesztését véve a t_c-re, a $h(t_c)$ és $h(t_d)$ egyszerre sorai vagy nem sorai az R-nek. Így tetszőleges olyan $v \in K$ sorra, amelyre $v[A_i] = c$, a h-t ki tudjuk úgy terjeszteni a t_c-re, hogy $h(t_c) = h(v)$ teljesüljön. Ha $h(t_c)$ és $h(t_d)$ az R-nek sorai, akkor ez azt jelenti, hogy $h(c) \sim h(d)$, tehát $g(h(c)) = g(h(d))$, viszont $g(h(c)) = h(c)$ és $g(h(d)) = h(d)$, ami ellentmondás. □

5.6. LEMMA. Legyen X_E egy XGD, ahol E egy $A_i - \mathrm{EGD}$, és legyen $\mathrm{Class}(Q)$ a $\mathrm{SAT}(X_E)$-nek egy ekvivalencia osztálya. Legyen továbbá $G = \langle K, (c, d) \rangle$ egy $A_j - \mathrm{EGD}$, ahol $j \neq i$. Ekkor vagy $\mathrm{Class}(Q) \subseteq \mathrm{SAT}(G)$ vagy $\mathrm{Class}(Q) \cap \mathrm{SAT}(G) = \emptyset$.

Az 5.6. lemma bizonyítása az 5.4. lemma bizonyításához hasonló módon történhet.

A fenti előkészítések után az 5.1. tételt a következőképpen bizonyíthatjuk be.

Legyen $R \in \mathrm{SAT}(\mathscr{F}^*)$. Ekkor $R \in \mathrm{SAT}(X_{E_i})$ minden $i = 1, 2, \ldots, N$-re. Legyen $g_i(R) \subseteq R$ az R reláció magja a $\mathrm{SAT}(X_{E_i})$-ben.

Az 5.2. lemmából és a g_i leképezések konstrukciójából következik, hogy az $S = g_1(g_2(\ldots g_N(R)\ldots))$ relációt véve $S \in \mathrm{SAT}(\mathscr{F})$, és ez az S a gyenge ekvivalencia értelmében egyértelmű. Azt mondjuk, hogy S az R reláció magja a $\mathrm{SAT}(\mathscr{F}^*)$-ban, továbbá az R_1, $R_2 \in \mathrm{SAT}(\mathscr{F}^*)$ relációkat hasonlóknak hívjuk, ha a gyenge ekvivalencia értelmében megegyezik a magjuk a $\mathrm{SAT}(\mathscr{F}^*)$-ban.

Az 5.3. lemma értelmében ez a hasonlóság ekvivalencia reláció, és minden $\mathrm{Class}_{\mathscr{F}^*}(S)$ ekvivalencia osztályra teljesül, hogy $\mathrm{Class}_{\mathscr{F}^*}(S)=\{R \mid R \in \mathrm{SAT}(\mathscr{F}^*)$ és R magja SAT $(\mathscr{F}^*)$-ban gyengén ekvivalens S-sel$\}$.

A tételben szereplő (i) és (ii) tulajdonság az 5.2. lemma és az 5.3. lemma közvetlen következményeként kapható. A többi tulajdonság az 5.4.—5.6. lemmák segítségével egyszerűen bizonyítható.

6. Az üldözéses algoritmus *(„Chase")*

Az általánosság megszorítása nélkül feltehetjük, hogy a relációkban szereplő szimbólumok természetes számok, mivel a gyenge ekvivalencia nem változtatja meg a függőségeket és nem hasonlítunk össze különböző oszlopban álló értékeket a kiértékelések definíciója miatt.

A TTGD-k esete

Legyen $\mathscr{F}=\{T_1, \ldots, T_N\}$ TTGD-knek egy halmaza és $T=\langle I, w\rangle$ legyen egy TTGD. Jelölje $\mathscr{R}$ a $D=\{\mathrm{VAL}_I(A_1), \ldots, \mathrm{VAL}_I(A_n)\}$-n értelmezett relációknak a halmazát. Mivel $\mathscr{R}$ véges, így SAT $(\mathscr{F})\cap\mathscr{R}$ is véges.

Legyen R, $S \in \mathrm{SAT}(\mathscr{F})\cap\mathscr{R}$ két olyan reláció, amelyek I-t tartalmazzák. Ekkor $R\cap S \in \mathrm{SAT}(\mathscr{F})\cap\mathscr{R}$ is teljesül. (Ez abból a tényből fakad, hogy ha K, $L \in \mathrm{SAT}(\mathscr{F})$, akkor $K\cap L \in \mathrm{SAT}(\mathscr{F})$, vagyis ha $\mathscr{F}$ csak TTGD-kből áll, akkor SAT $(\mathscr{F})$ zárt a metszet műveletre nézve.)

Ekkor a $\mathrm{Chase}_{\mathscr{F}}(I)=\bigcap_{\substack{R\in \mathrm{SAT}(\mathscr{F})\cap\mathscr{R}\\ I\subseteq R}} R$ reláció egyértelműen definiált, és I-ből úgy kapható meg, hogy $T_1, \ldots, T_N$ függőségeket tetszőleges sorrendben alkalmazzuk az I relációból kiindulva addig, amíg olyan relációhoz jutunk, amely semelyik T_i alkalmazásával nem változik. Világos, hogy a végesség miatt a T_i-k alkalmazása véget fog érni, és a kapott reláció tartalmazni fogja a $\mathrm{Chase}_{\mathscr{F}}(I)$-t. Másrészt, ha $\mathrm{Chase}_{\mathscr{F}}(I)$ valódi része lenne a kapott relációnak, akkor legyen T_j az első olyan függőség, amely olyan v sort eredményez, amely nincs benne a $\mathrm{Chase}_{\mathscr{F}}(I)$-ben. Viszont a megfelelő kiértékelés a T_j feltétel relációját a $\mathrm{Chase}_{\mathscr{F}}(I)$-be képezi, így a T_j v eredménysorának is a $\mathrm{Chase}_{\mathscr{F}}(I)$-ben kell lennie.

A 4.1. tétel bizonyítása erre az esetre

1. Először megmutatjuk, hogy ha $\mathscr{F} \models T$, akkor $w \in \mathrm{Chase}_{\mathscr{F}}(I)$. Tegyük fel az ellenkezőjét, azaz, hogy $w \notin \mathrm{Chase}_{\mathscr{F}}(I)$. Ekkor a h identitás leképezés az I-t a $\mathrm{Chase}_{\mathscr{F}}(I)$-be képezi, és $h(w)=w \notin \mathrm{Chase}_{\mathscr{F}}(I)$.
2. Belátjuk azt, hogy ha $w \in \mathrm{Chase}_{\mathscr{F}}(I)$, akkor $\mathscr{F} \models T$. Legyen $R \in \mathrm{SAT}(\mathscr{F})$ és h egy olyan kiértékelés, amely I-t az R-be képezi. Ekkor $h(w) \in h(\mathrm{Chase}_{\mathscr{F}}(I))$, másrészt $\mathrm{Chase}_{\mathscr{F}}(h(I)) \subseteq R$ a $\mathrm{Chase}_{\mathscr{F}}$ definíciója miatt. Így a $h(w) \in R$-hez elég belátni, hogy $\mathrm{Chase}_{\mathscr{F}}(h(I))=h(\mathrm{Chase}_{\mathscr{F}}(I))$.
   A $h(\mathrm{Chase}_{\mathscr{F}}(I)) \in \mathrm{SAT}(\mathscr{F})$ és $h(I) \subseteq h(\mathrm{Chase}_{\mathscr{F}}(I))$ triviális teljesülésből következik, hogy $\mathrm{Chase}_{\mathscr{F}}(h(I)) \subseteq h(\mathrm{Chase}_{\mathscr{F}}(I))$. A másik irányú tartalmazáshoz ve-

gyünk egy $t \in \text{Chase}_{\mathcal{F}}(I)$ sort. Ha $t \in I$, akkor nyilván $h(t) \in \text{Chase}_{\mathcal{F}}(h(I))$. Ha $t \notin I$, akkor vegyük azt a g leképezést, amelyik a t sort behozza a $\text{Chase}_{\mathcal{F}}(I)$ konstruálásakor, és legyen $T_j = \langle I_j, w_j \rangle$ a megfelelő TTGD. Ekkor $t = g(w_j)$ és $h(g(I_j)) \subseteq \text{Chase}_{\mathcal{F}}(h(I))$, ha a $\text{Chase}_{\mathcal{F}}(h(I))$ konstruálásánál a függőségeket ugyanabban a sorrendben alkalmazzuk, mint a $\text{Chase}_{\mathcal{F}}(I)$ konstruálásánál. Így $h(g(w_j)) = h(t) \in \text{Chase}_{\mathcal{F}}(h(I))$ is teljesül. □

A TTGD-k és EGD-k esete

Legyen $\mathcal{F} = \{T_1, \ldots, T_N, E_1, \ldots, E_M\}$ TTGD-knek és EGD-knek egy halmaza, ahol T_i-vel TTGD-t, E_j-vel EGD-t jelölünk. Cseréljük ki az összes E_i-t a neki megfelelő X_{E_i} XGD-vel, és az így kapott rendszert jelöljük $\mathcal{F}^*$-gal, azaz az $\mathcal{F}^* = \{T_1, \ldots, T_N, X_{E_1}, \ldots, X_{E_M}\}$ rendszer csak TTGD-ket tartalmaz. (Az XGD speciális TTGD.)

Legyen F egy TTGD vagy EGD. Legyen F^* F-fel egyenlő, ha F TTGD és X_F XGD-vel egyenlő, ha F egy EGD. Legyen I az F feltétel táblája. Ekkor az 5.2. tétel értelmében az $\mathcal{F} \models F$ implikáció akkor és csak akkor teljesül, ha $\mathcal{F}^* \models F^*$. Az $\mathcal{F}^* \models \models F^*$ implikáció viszont akkor és csak akkor teljesül, ha $\text{Chase}_{\mathcal{F}^*}(I) \in \text{SAT}(F^*)$. Az 5.1. tétel értelmében $\text{Chase}_{\mathcal{F}^*}(I) \in \text{SAT}(F^*)$ akkor és csak akkor teljesül, ha $\text{Kernel}(\text{Chase}_{\mathcal{F}^*}(I)) \in \text{SAT}(F)$, ahol $\text{Kernel}(R) = g(R) = g_1(g_2(\ldots g_M(R)\ldots))$ olyan 5.2. lemmában szereplő g_i leképezésekre, amelyekre az teljesül, hogy ha Z_{ij} egy felcserélhetőségi ekvivalencia osztály, akkor $g_i(x) = \min_{z \in Z_{ij}} z$ minden $x \in Z_{ij}$-re. Definiáljuk a $\text{Chase}_{\mathcal{F}}(I)$ relációt tetszőleges I-re úgy, hogy $\text{Chase}_{\mathcal{F}}(I) := \text{Kernel}(\text{Chase}_{\mathcal{F}^*}(I))$.

A $\text{Chase}_{\mathcal{F}}(\,.\,)$ operáció egyértelműen definiált, és $\mathcal{F} \models F$ akkor és csak akkor teljesül, ha $\text{Chase}_{\mathcal{F}}(I) \in \text{SAT}(F)$. Az így definiált $\text{Chase}_{\mathcal{F}}(I)$ relációt a következő eljárás segítségével kaphatjuk meg. Az $\mathcal{F}$ halmaz függőségeit alkalmazzuk tetszőleges sorrendben I-ből kiindulva, azzal a megszorítással, hogy az EGD-k alkalmazása során mindig a nagyobb értéket tesszük egyenlővé a kisebbel. Ez az eljárás véges lépésben véget ér. Annak a bizonyításához, hogy az így nyert reláció a $\text{Chase}_{\mathcal{F}}(I)$-vel megegyezik, elég észrevenni, hogy egy egyenlővé tétel a fenti g leképezésnek egy részleképezését jelenti, mivel az egyenlővé tett értékek felcserélhető értékek lennének $\text{Chase}_{\mathcal{F}^*}(I)$-ben.

Végezetül kimondunk egy tételt, amely rávilágít arra, hogyan lehet a $\text{Chase}_{\mathcal{F}}(I)$ segítségével egy $\mathcal{F} \models T$ implikációt eldönteni. A bizonyítás a TTGD-k esetéhez hasonlóan történhet.

6.1. TÉTEL. Legyen $\mathcal{F}$ TTGD-k és EGD-k halmaza. Három esetet különböztetünk meg annak megfelelően, hogy egy F függőség a) $\langle I, w \rangle$ alakú TTGD, b) $\langle I, (a, b) \rangle$ alakú EGD vagy c) $\langle I, K \rangle$ alakú TGD. Legyen $\text{Chase}_{\mathcal{F}}(I)$ a fent definiált reláció és g a hozzá tartozó leképezés. Ekkor $\mathcal{F} \models F$ akkor és csak akkor, ha $\text{Chase}_{\mathcal{F}}(I) \in \text{SAT}(F)$ akkor és csak akkor, ha a) $g(w) \in \text{Chase}_{\mathcal{F}}(I)$, b) $g(a) = g(b)$, c) létezik egy olyan h leképezés, amely $g(I)$-n az identitás és $h(K) \subseteq \text{Chase}_{\mathcal{F}}(I)$.

7. Az általánosított függőségek esete

A 4—6 részben az általánosított függőségek egy speciális fajtáját, az attributum korlátozott függőségeket vizsgáltuk, és az ilyen függőségekre vonatkozó implikációs problémát elemeztük. Ebben a részben az attributum korlátozást elhagyjuk, és röviden áttekintjük, hogyan módosulnak a definíciók, lemmák, tételek. A bizonyításokat nem részletezzük, mert azok a megfelelő AGD-kre vonatkozó állítások bizonyításához hasonlóan kaphatók, csak a kiértékelések helyett szimbólum leképezéseket kell venni. Az EGD, TGD, TTGD, XGD jelöléseket és elnevezéseket meghagyjuk, de a definíciójukat a következőképpen módosítjuk.

A TGD egy $\langle I, J\rangle$ relációpár úgy, hogy I bármely sorához létezik egy sor J-ben, hogy a két sornak legalább egy közös egyedi értéke van. Egy R reláció kielégíti az $\langle I, J\rangle$ TGD-t, ha minden olyan h szimbólum leképezést, melyre $h(I)\subseteq R$ teljesül, ki lehet terjeszteni a J-re úgy, hogy $h(J)\subseteq R$ is teljesüljön.

Egy TGD-t TTGD-nek hívunk, ha a J egy olyan w sorral egyenlő, amelyre $\mathrm{VAL}(\{w\})\subseteq\mathrm{VAL}(I)$ teljesül, ahol $\mathrm{VAL}(I)$ az I reláció szimbólumainak halmazát jelöli. Az EGD egy $\langle I, (a, b)\rangle$ pár, ahol $a, b\in\mathrm{VAL}(I)$. Egy R reláció kilégíti az $\langle I, (a, b)\rangle$ EGD-t, ha minden olyan h szimbólum leképezésre, amelyre $h(I)\subseteq R$, teljesül a $h(a)=h(b)$ egyenlőség.

Legyen $E\langle I, (a, b)\rangle$ alakú EGD. Ekkor X_E XGD-n az $\{X_i(a, b),\ X_i(b, a)|i=1, 2, \ldots, n\}$ TTGD-knek a rendszerét értjük, ahol

$$X_i(a, b) = \langle I\cup(x_1, x_2, \ldots, x_{i-1}, a, x_{i+1}, \ldots, x_n),\ (x_1, \ldots, x_{i-1}, b, x_{i+1}, \ldots, x_n)\rangle$$

és

$$X_i(b, a) = \langle I\cup(x_1, x_2, \ldots, x_{i-1}, b, x_{i+1}, \ldots, x_n),\ (x_1, \ldots, x_{i-1}, a, x_{i+1}, \ldots, x_n)\rangle.$$

Az EGD, TGD, TTGD függőségek alkalmazását egy R relációra ugyanúgy definiáljuk, mint ahogy a 4. részben tettük, csak kiértékelések helyett szimbólum leképezéseket kell vennünk.

Az üldözéses *(„Chase")* eljárás változatlan marad, csak most a $\mathrm{VAL}(I)$-re kell feltennünk, hogy teljesen rendezett, és hogy egy $\langle I, (x, y)\rangle$ alakú EGD-ben az x a $\mathrm{VAL}(I)$ legkisebb eleme. A felcserélhetőség fogalma a következőképpen módosul.

7.1. Lemma. Legyen $R\in\mathrm{SAT}(X_E)$, ahol E egy $\langle I, (a, b)\rangle$ alakú EGD és $x, y\in\mathrm{VAL}(R)$. Azt mondjuk, hogy x és y felcserélhetők, ha van olyan h szimbólum leképezés, amelyre $h(I)\subseteq R$ és $h(a)=x, h(b)=y$. Jelöljük ezt $x\sim y$-nal. Azt mondjuk, hogy $x\in\mathrm{VAL}(R)$ felcserélhető érték, ha van olyan $y\in\mathrm{VAL}(R)$, hogy $x\sim y$. Ekkor a felcserélhetőség ekvivalencia reláció a felcserélhető értékek halmazán.

A 7.1. Lemma bizonyítása

1. *Reflexivitás.* Legyen x egy felcserélhető érték. Ekkor van olyan $y\in\mathrm{VAL}(R)$ és egy h szimbólum leképezés, amelyre $h(a)=x$, $h(b)=y$ és $h(I)\subseteq R$. Legyen g a következő szimbólum leképezés; $g(c)=h(c)$, ha $c\neq b$ és $g(b)=h(a)=x$. Ekkor $R\in\mathrm{SAT}(X_E)$ miatt $g(I)\subseteq R$ és $g(a)=g(b)=x$, azaz $x\sim x$.

2. *Szimmetria:* Legyen $x \sim y$, és $h(a)=x$, $h(b)=y$, $h(I) \subseteq R$ egy h szimbólum leképezésre. Definiáljuk a g szimbólum leképezést úgy, hogy az a, b értékek kivételével a h-val egyezzen meg, és $g(a)=h(b)=y$, $g(b)=h(a)=x$ legyen. Mivel $R \in \mathrm{SAT}(X_E)$, ezért $g(I) \subseteq R$, így $y \sim x$ teljesül.
3. *Tranzitivitás:* Legyen $x \sim y$ és $y \sim z$, a megfelelő szimbólum leképezések pedig f és g, azaz $f(a)=x$, $f(b)=g(a)=y$, $g(b)=z$ és $f(I) \subseteq R$, $g(I) \subseteq R$. Definiáljuk a h szimbólum leképezést az alábbi módon: $h(c)=f(c)$, ha $c \neq b$ és $h(b)=g(b)=z$. Mivel $R \in \mathrm{SAT}(X_E)$, ezért $h(I) \subseteq R$, tehát $x \sim z$. □

Jelöljük a felcserélhető értékek halmazát $E(R)$-rel. Ekkor a felcserélhetőségi reláció egy ekvivalencia osztályozást ad az $E(R) \subseteq \mathrm{VAL}(R)$ halmazon:

$$E(R) = \bigcup_{k=1}^{M} Z_k.$$

Legyen $z_j \in Z_j$, $j=1, \ldots, M$ és legyen g a következő szimbólum leképezés:

$$g(x) = \begin{cases} z_j, & \text{ha } x \in Z_j \\ x, & \text{ha } x \notin E(R). \end{cases}$$

Ennek a g-nek a következő tulajdonságai vannak:

7.2. LEMMA. Legyen $R \in \mathrm{SAT}(X_E)$, és g a fenti szimbólum leképezés. Ekkor

(i) $g(R) \in \mathrm{SAT}(E)$.

(ii) $g^{-1}(g(R)) = R$.

(ii) Ha f egy-egy értelmű (i) és (ii) tulajdonságú szimbólum leképezés, akkor $g(R) \equiv f(R)$.

A 7.2. lemma bizonyítása az 5.2. lemma bizonyításához hasonló módon történhet.

Egy $R \in \mathrm{SAT}(X_E)$ reláció magját a fenti g szimbólum leképezéssel adhatja meg: $\mathrm{Kernel}(R)=g(R)$. A magra az 5.1. következmény állításai érvényesek maradnak.

Az 5.3.—5.4. lemma továbbra is igaz marad; a bizonyításokban a kiértékelés helyett értelemszerűen szimbólum leképezést kell venni.

7.3. LEMMA. Legyen $\mathrm{SAT}(X_E) = \bigcup_{Q \in \mathrm{SAT}(E)} \mathrm{Class}(Q)$ az 5.3. lemmában megadott felbontás, és legyen $G=\langle K, (a, b)\rangle$ egy EGD. Ekkor bármelyik $\mathrm{Class}(Q)$-ra a következő állítások valamelyike teljesül:

i) $\mathrm{Class}(Q) \subseteq \mathrm{SAT}(X_G)$ és $Q \subseteq \mathrm{SAT}(G)$,

vagy

$\mathrm{Class}(Q) \cap \mathrm{SAT}(X_G) = \emptyset$.

ii) $\mathrm{Class}(Q) \subseteq \mathrm{SAT}(G)$,

vagy

$\mathrm{SAT}(Q) \cap \mathrm{SAT}(G) = \emptyset$.

A 7.3. lemma bizonyítása az 5.5.—5.6. lemmák bizonyításához hasonlóan történhet.

Legyen $\mathcal{F}=\{E_1, \ldots, E_N\}$ EGD-k egy halmaza és

$$\mathcal{F}^* = \bigcup_{i=1}^{N} X_{E_i}.$$

7.1. TÉTEL. A fenti jelölések mellett SAT $(\mathcal{F}^*)$ diszjunkt osztályokra bontható úgy, hogy minden osztály a gyenge ekvivalencia értelmében egy relációt tartalmaz a SAT $(\mathcal{F})$-ből, és erre a felbontásra a következő állítások teljesülnek.

(i) SAT $(\mathcal{F}^*)=\bigcup_{R\in \mathrm{SAT}(\mathcal{F})} \mathrm{Class}\,(R)$, ahol Class $(R_i)\cap$Class $(R_j)=\emptyset$, ha $R_i \not\equiv R_j$.

(ii) Ha $S\in$Class (R), akkor létezik olyan h szimbólum leképezés, amelyre $h(S)=R$ és $h^{-1}(R)=S$.

(iii) Ha $S\in$Class (R) és $S\in$SAT (G), ahol G egy TGD, akkor Class $(R)\subseteq$ $\subseteq$SAT (G).

(iv) Ha E egy EGD, akkor minden $R\in$SAT $(\mathcal{F})$-re $R\in$SAT (E) akkor és csak akkor, ha Class $(R)\subseteq$SAT (X_E). Ha $S\in$SAT (E) és $S\neq R$, továbbá $S\in$Class (R), akkor Class $(R)\subseteq$SAT (E).

A 7.1. tétel bizonyítása az 5.1. tétel bizonyítása szerint történhet, csak az 5.1., 5.2., 5.5.—5.6. lemmák helyett a 7.1.—7.3. lemmákat kell alkalmazni.

7.2. TÉTEL. Legyen $\mathcal{F}$ TGD-knek és EGD-knek egy halmaza, és legyen $\mathcal{F}^*$ az a halmaz, amelyet úgy kapunk az $\mathcal{F}$-ből, hogy az $E\in\mathcal{F}$ EGD-ket kicseréljük a megfelelő X_E rendszerrel. Legyen F egy EGD vagy TGD, és legyen F^* az F-fel egyenlő, ha F egy TGD és X_E-vel egyenlő, ha F egy E EGD. Ekkor az $\mathcal{F}\models F$ implikációs probléma ekvivalens az $\mathcal{F}^*\models F^*$ implikációs problémával.

A 7.2. tétel bizonyítása az 5.2. tétel bizonyítása szerint történhet.

IRODALOM

[1] BEERI, C., VARDI, M. Y., "Formal system for tuple and equivality generating dependencies", *SIAM J. Computing* **13** (1984) 76—98.

[2] BEERI, C., VARDI, M. Y., "A proof procedure for data dependencies", *JACM* **31** (1984) 718—741.

[3] COSMADAKIS, S. S., KANELLAKIS, P. C., "Two applications of equational theories to database theory", Rewriting Techniques and Applications, Dijon, May,1985, in *Lecture Notes in Computer Science* No **202.** Springer Verlag, 107—123.

[4] DEMETROVICS, J., "Candidate keys and antichains", *SIAM J. Alg. Disc. Math.* **161** (1980) 92.

[5] DEMETROVICS, J., GYEPESI GY., "On the functional dependencies and some generalizations of it", *Acta Cybernetica* **5** (1981) 295—305.

[6] DEMETROVICS, J. GYEPESI, GY., "Some generalized type functional dependencies formalized as equality set on matrices", *Discrete Applied Mathematics* **6** (1983).

[7] DEMETROVICS. J., GYEPESI, GY., "Logical dependencies in relational database", *MTA SZTAKI Tanulmányok* **113** (1980) 59—78.

[8] MITCHELL, J. C., "The implication problem for functional and inclusion dependencies", *Information and Control* **56** (1983) 154—173.

[9] ULLMAN, J. D., *Principles of Database Systems* (Computer Science Press Inc., 1983).

(Beérkezett: 1987. szeptember 1.)

BENCZÚR ANDRÁS, KISS ATTILA ÉS MÁRKUS TIBOR
ELTE SZÁMÍTÓKÖZPONT
1117 BUDAPEST, BOGDÁNFY ÚT 10/B.

ON A GENERAL CLASS OF DATA DEPENDENCIES IN THE RELATIONAL MODELL AND ITS IMPLICATION PROBLEMS

A. BENCZUR, A. KISS, T. MÁRKUS

The present paper deals with an important problem of the relational database theory, with analysis of dependencies. In the first part it is shown, that a sufficiently large class of dependencies ("hypothesis-conclusion" type of dependencies) can be expressed as first-order sentences (HCD-formulas). In the second part of the paper we deal with implication problems, namely with the generalization of implication problems for TTGD, TGD, EGD types of dependencies, introduced by BEERI—VARDI ([2]). For these implication problems we give a generalized version of the so called chase proof procedure. We introduce exchangeability — generating dependency (XGD) and analyze the structure of relation classes satisfying XGDs, from the point of view of invariant dependencies.

Alkalmazott Matematikai Lapok **13** (1987—88) 313—317

EGY ÁLTALÁNOS FELTÉTEL NÉLKÜLI GEOMETRIAI PROGRAMOZÁSI FELADATPÁRRÓL

ILLÉS TIBOR

Budapest

Cikkünk a geometriai programozási probléma egy lehetséges általánosítását mutatja be. Az itt közölt gondolatokat PETERSON [3] és ROCKAFELLAR [5] cikkei inspirálták, de azoktól jelentősen eltér mind formai, mind pedig tartalmi szempontból.

A 2. pontban tetszőleges függvény konjugált függvényével kapcsolatos ismertetőt adunk. A 3. pontban PETERSONtól eltérően egy szimmetrikusabb primál-duál feladatpárt fogalmazunk meg. Ebben a részben tárgyaljuk a feladatpár alaplemmáját is.

1. Bevezetés

Cikkünkben KLAFSZKY [2] dolgozatában megtalálható geometriai programozási feladatpár egy lehetséges általánosítását adjuk, annélkül, hogy megmutatnánk az eredeti feladatról, hogy a jelen dolgozatban tárgyalt feladatpár speciális esete. Erre, trivialitása mellett, azért sem térünk ki, mert a jelen dolgozat egy olyan cikksorozat első eleme, amelyben az általános feltétel nélküli geometriai programozási feladatpár részletes tárgyalását adjuk és így több más speciális eset mellett a geometriai programozási problémáról is, a későbbiekben még szó lesz.

Az általánosításban — mint a nemlineáris programozási feladatpároknál mindig — egy egyenlőtlenség, esetünkben a *Fenchel egyenlőtlenség* játszik központi szerepet. A 2. pontban tetszőleges függvény konjugált függvényének a definícióját adjuk meg és ennek két elemi tulajdonságát igazoljuk. Ebben a pontban kerül kinondásra a *Fenchel egyenlőtlenség* is. A 3. pontban kerül sor a feladatpár definíciójára és az alaplemma igazolására.

Latin kisbetűkkel az R^n tér pontjait, latin nagybetűkkel pedig a halmazait jelöltük. A $g, h, \bar{h}$ függvényeket jelölnek.

2. Konjugált függvények

Konvex függvényekkel kapcsolatban vezette be a konjugált függvény fogalmát FENCHEL [1]. ROCKAFELLAR [5, 6] dolgozta ki részletekbe menő alapossággal a konvex függvények konjugált függvényeinek elméletét és ezt az elméletet sikeresen alkalmazta konvex programozási feladatokra. Néhány évvel később PETERSON [3] nem konvex függvényekre is kiterjesztette a konjugált függvény fogalmát. Ezáltal nem konvex programozási feladatra is alkalmazhatóvá vált ROCKAFELLAR módszere.

Tetszőleges függvény konjugált függvényének itt következő definíciója néhány apró részletben eltér a szakirodalomban eddig használatos definícióktól. Az eltérések

azonban nem elvi jelentőségűek, de a későbbiek során biztosítani fogják azt, hogy a primál és duál feladatpár, PETERSON [3] cikkében levő megfogalmazástól eltérően ne csak formai szimmetriát mutasson feltételeiben, hanem tartalmit is.

2.1. Definíció. Adott a $g: D_g \to \boldsymbol{R}$ tetszőleges függvény, ahol $D_g \subseteq \boldsymbol{R}^n$. Legyen

$$h(y) := \sup_{x \in D_g} \{g(x) - \langle x, y \rangle\},$$

ahol a $h: \boldsymbol{R}^n \to \boldsymbol{R}$ függvény a g függvény *konjugált függvénye* a

$$D_h := \left\{y \in \boldsymbol{R}^n: \sup_{x \in D_g} \{g(x) - \langle x, y \rangle\} < +\infty\right\}$$

halmazon.

Most pedig néhány apró megjegyzést teszünk a konjugált függvénnyel kapcsolatban.

2.1. Megjegyzés. Legyen $A \subseteq D_g$ és

$$\bar{h}(y) := \sup_{x \in A} \{g(x) - \langle x, y \rangle\},$$

ahol a h függvény a g függvény konjugált függvénye a

$$D_{\bar{h}} := \left\{y \in \boldsymbol{R}^n: \sup_{x \in A} \{g(x) - \langle x, y \rangle\} < +\infty\right\},$$

halmazon. Ekkor a $D_h \subseteq D_{\bar{h}}$ és $h(y) \geqq \bar{h}(y)$ $\forall y \in D_h$-n.

A 2.1. megjegyzés nyilvánvaló a 2.1. definícióból és mutatja azt is, hogy hogyan viselkedne a g függvény konjugált függvénye, ha a g függvény értelmezési tartományának valamely részhalmazára definiálnánk. PETERSON [3] említett cikkében a tetszőleges g függvény konjugált függvényét a g függvény értelmezési tartományának egy kúppal elmetszett részére definiálta. A 2.1. megjegyzésből nyilvánvaló, hogy különböző kúpok esetén más és más konjugált függvény adódik.

2.2. Megjegyzés (Fenchel-egyenlőtlenség).
A 2.1. definícióból adódó egyenlőtlenség

$$g(x) - h(y) \leqq \langle x, y \rangle \quad \forall x \in D_g \quad \text{és} \quad \forall y \in D_h.$$

A definícióból még az is látható, hogy konjugált függvény nem mindig létezik, azaz előfordulhat, hogy $D_h = \emptyset$.

2.2. Definíció. Tetszőleges $h: \boldsymbol{R}^n \to \boldsymbol{R}$ konjugált függvény esetén a függvény epigráfja a következő halmaz

$$\operatorname{epi} h := \{(y, \alpha): \alpha \geqq g(x) - \langle x, y \rangle, \quad \text{ahol} \quad y \in D_h \quad \text{és} \quad \alpha \in \boldsymbol{R}\}.$$

Az epigráf ennél általánosabb definíciója megtalálható ROCKAFELLAR [6] könyvében.

2.1. LEMMA. Ha a h függvény létezik akkor

1° a D_h konvex halmaz és a h konvex függvény,
2° a h függvény epigráfja zárt halmaz.

Bizonyítás. 1° Legyen $y_1, y_2 \in D_h$ és $0 \leqq \lambda \leqq 1$. Ekkor a

$$\sup_{x \in A} \{f_1(x)+f_2(x)\} \leqq \sup_{x \in A} f_1(x)+\sup_{x \in A} f_2(x),$$

ahol $A \subseteq \boldsymbol{R}^n$, egyenlőtlenséget figyelembe véve az alábbi összefüggést nyerjük:

$$h(\lambda y_1+(1-\lambda)y_2) = \sup_{x \in D_g} \{g(x)-\langle x, \lambda y_1+(1-\lambda)y_2\rangle\} =$$

$$= \sup_{x \in D_g} \{\lambda[g(x)-\langle x, y_1\rangle]+(1-\lambda)[g(x)-\langle x, y_2\rangle]\} \leqq$$

$$\leqq \sup_{x \in D_g} \lambda\{g(x)-\langle x, y_1\rangle\}+\sup_{x \in D_g} (1-\lambda)\{g(x)-\langle x, y_2\rangle\} =$$

$$= \lambda \sup_{x \in D_g} \{g(x)-\langle x, y_1\rangle\}+(1-\lambda) \sup_{x \in D_g} \{g(x)-\langle x, y_2\rangle\} = \lambda h(y_1)+(1-\lambda)h(y_2).$$

Tehát a h konvex függvény és a D_h konvex halmaz.

2° Az epi h halmaz zártsága a definíciójából nyilvánvaló, hiszen az epih zárt félterek metszete. ∎

ROCKAFELLAR [5] cikke szerint ezt az állítást először FENCHEL igazolta.

3. Az általános feltétel nélküli geometriai programozási feladatpár és alapvető lemmája

Ebben a fejezetben az általános feltétel nélküli geometriai programozási probléma primál és duál feladatát definiáljuk. Definíciónk eltér a PETERSON [3] által adottól. Amíg PETERSONnál a duál célfüggvény — mint ahogy azt a 2.1. megjegyzés mutatja — függ a primál feladat megengedett megoldáshalmazának előállításában szereplő kúptól, addig esetünkben, hiszen a teljes D_g-n definiáltuk a g függvény konjugáltját, ez nem áll fenn. Tehát esetünkben a függvény csak a g függvénytől függ. Az előző gondolatmenet mutatja, hogy PETERSONnál a duál feladat célfüggvénye függött a primál feladat megengedett megoldáshalmazától (és a primál célfüggvény független volt a duál feladat megengedett megoldáshalmazától), azaz a feladatpár egymás megengedett megoldáshalmazától való függése nem volt szimmetrikus. Dolgozatunkban kísérletet teszünk egy — az előző értelemben — szimmetrikusabb feladatpár előállítására. Ebben a szimmetrikus feladatpárban a duál célfüggvénye független lesz a primál feladat megengedett megoldáshalmazától. Kitűzött célunk eléréséhez alkalmas segédeszközt nyújt a tetszőleges függvény konjugált függvényére általunk alkalmazott definíció.

Primál feladat: Legyen $g: D_g \to \boldsymbol{R}$ tetszőleges függvény, ahol $D_g \subseteq \boldsymbol{R}^n$ és $K \subseteq \boldsymbol{R}^n$ tetszőleges kúp.

$$P := D_g \cap K,$$

$$\sup \{g(x) | x \in P\}.$$

Figyeljük meg, hogy a primál feladat definíciójában egyetlen megkötést tettünk, mégpedig azt, hogy a P halmaz a g függvény értelmezési tartományának és egy $\boldsymbol{R}^n$-beli — nem feltétlenül konvex kúpnak a metszete legyen. Nyilván előfordulhat az az eset is, amikor $P=\emptyset$. Ekkor legyen $\sup \{g(x) | x \in P\} = -\infty$.

Duál feladat: Legyen h a primál feladat célfüggvényének a konjugált függvénye. Ekkor

$$D := D_h \cap K^*,$$

$$\inf \{h(x) | x \in D\},$$

ahol K^* a K kúp polárisa.

Ismert, a *Minkowski tételből* (PRÉKOPA [4]), hogy a K^* kúp. Könnyen látható, hogy a K kúp nem feltétlenül konvex kúp, de a polárisa már konvex kúp és a *Farkas-tétel* (PRÉKOPA [4]) értelmében a $K^{**} = \overline{\text{conv}K}$. A fentieket is figyelembe véve igaz a

3.1. Megjegyzés. A 2.1. lemma miatt a duál feladat célfüggvénye h konvex függvény és a D konvex halmaz, tehát a duál feladat konvex programozási feladat.

3.2. Megjegyzés. Mivel a h függvény epigráfja zárt, ezért korlátos D halmaz és $P \neq \emptyset$ esetén létezik $\bar{y} \in D$ úgy, hogy

$$\inf_{y \in D} h(y) = h(\bar{y}),$$

azaz létezik optimális megoldása a duál feladatnak.

A fenti állítás nyilvánvaló, hiszen korlátos halmazon zárt függvény bármely nívóhalmaza is zárt. A nívóhalmazokon h felveszi a minimumát. A nívóhalmazok az epi h részhalmazai. A nívóhalmazokon felvett minimumértékek a zárt epi h halmazon, h konvexitása miatt, alulról korlátos sorozatot alkotnak, ezért a h függvény felveszi a minimumát a D halmazon.

Vezessük be a következő jelöléseket:

$P^* := \{x | x \in P$ és x primál optimális megoldás$\}$,
$D^* := \{x | x \in D$ és x duál optimális megoldás$\}$,

$$\varphi = \sup_{x \in P} g(x) \quad \text{és} \quad \chi = \inf_{x \in D} h(x).$$

3.1. LEMMA *(alaplemma).* Ha $x \in P$ és $y \in D$, akkor

$$g(x) \leqq h(y),$$

és egyenlőség pontosan akkor van, ha $h(y) = \bar{h}(y)$ és

$$g(x) = \sup_{\bar{x} \in P} \{g(\bar{x}) - \langle \bar{x}, y \rangle\}.$$

Bizonyítás. Figyelembe véve a primál és duál feladat definícióját, valamint a 2.1. definíciót, a $P \subseteq D_g$ feltétel miatt adódik a következő összefüggés:

$$(3.1) \qquad \begin{cases} h(y) = \sup\limits_{x \in D_g} \{g(x) - \langle x, y \rangle\} \geqq \sup\limits_{\bar{x} \in P} \{g(\bar{x}) - \langle \bar{x}, y \rangle\} = \bar{h}(y) \geqq \\ \qquad\qquad \geqq g(x) - \langle x, y \rangle \geqq g(x), \end{cases}$$

ahol az utolsó egyenlőtlenség, azért igaz, mert $x \in K$ és $y \in K^*$ feltételekből $\langle x, y \rangle \leqq 0$, azaz $-\langle x, y \rangle \geqq 0$ adódik, tehát

$$h(y) \geqq g(x).$$

Ha $g(x) = h(y)$, akkor természetesen $h(y) = \bar{h}(y)$ és (3.1) utolsó két egyenlőtlenségében egyenlőség áll, azaz

$$g(x) = \sup_{\bar{x} \in P} \{g(\bar{x}) - \langle \bar{x}, y \rangle\} = \bar{h}(y).$$

Megfordítva, (3.1)-ben a $h(y)=\bar{h}(y)$ feltétel miatt az első, a $g(x)=\sup_{\bar{x}\in P}\{g(\bar{x})-\langle\bar{x}, y\rangle\}$ feltétel miatt pedig a második és harmadik egyenlőtlenség is egyenlőséggel teljesül, így $g(x)=h(y)$.

Ezzel az alaplemmát beláttuk. ∎

Az alaplemmából elemi úton nyerhetők az alábbi következmények:

3.1. KÖVETKEZMÉNY. 1° Ha $P\neq\emptyset$ akkor a duál feladat célfüggvénye alulról korlátos. ∎

2° Ha $D\neq\emptyset$ akkor a primál feladat célfüggvénye felülről korlátos.

3.2. KÖVETKEZMÉNY *(gyenge equilibrium)*. Ha $x\in P$ és $y\in D$ esetén $g(x)=h(y)$, akkor $x\in P^*$ és $y\in D^*$, továbbá $\langle x, y\rangle=0$. ∎

Az előző következmény mutatja, hogy ha az optimális megoldások esetén felvett függvényértékek egyenlők, akkor az optimális megoldások kielégítik az ún. *komplementaritási feltételt*, azaz $\langle x, y\rangle=0$.

4. Köszönetnyilvánítás

Köszönetet szeretnék mondani KLAFSZKY EMILnek és TERLAKY TAMÁSnak a munkám során nyújtott hasznos tanácsaikért.

IRODALOM

[1] FENCHEL, W., "On conjugate convex functions", *Canad. J. Math.* (1949) 73—77.
[2] KLAFSZKY, E., "Geometriai programozás és néhány alkalmazása", MTA SZTAKI Tanulmányok **8** (1973).
[3] PETERSON, E. L., "Geometric programing", *SIAM Review*, **18** (1976) 1—51.
[4] PRÉKOPA, A., *Lineáris grogramozás I* (Bolyai János Matematikai Társulat kiadványa, 1968).
[5] ROCKAFELLAR, R. T., "Conjugate convex functions in nonlinear programming".
[6] ROCKAFELLAR, R. T., *Convex Analysis* (Princeton Univ. Press, Princeton, 1970).

(Beérkezett: 1987. június 2.)
(Átdolgozva beérkezett: 1987. szeptember 28.)

ILLÉS TIBOR
MAGYAR TUDOMÁNYOS AKADÉMIA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓINTÉZETE
KENDE U. 13—17
1111 BUDAPEST

ON A GENERAL UNCONSTRAINED GEOMETRIC PROGRAMMING PROBLEM

T. ILLÉS

Our paper presents a possible generalization of geometric programming problems. Such a generalization was proposed by PETERSON [3], based on ROCKAFELLAR's [5] conjugate theory. Using their results we define a slightly different, more symmetric dual of general unconstrained geometric programming problem.

In the second paragraph the conjugate function is defined and some of its properties are demonstrated. In the third paragraph the general unconstrained geometric programming problem and its dual pair is introduced. Finally some fundamental properties of general unconstrained geometric programming problem are proved.

Alkalmazott Matematikai Lapok **13** (1987—88) 319—334

GLOBÁLIS HEURISZTIKUS ELJÁRÁSOK A DISZKRÉT PROGRAMOZÁSBAN: EGY SZTOCHASZTIKUS MEGKÖZELÍTÉS

VIZVÁRI BÉLA

Budapest

A dolgozat munkahipotézis gyanánt egy sztochasztikus modellt állít fel a következő problémára: hogyan kell egy heurisztikus eljárás ismételt alkalmazásait egy diszkrét programozási feladat esetében úgy megszervezni, hogy a számítási költségek egy adott határt ne lépjenek túl és az optimális megoldást a lehető legnagyobb valószínűséggel megkapjuk.

1. Bevezetés

A heurisztikus eljárásokat a diszkrét programozásban abból a szempontból szokták vizsgálni, hogy az általuk szolgáltatott megengedett megoldások mennyire közelítik meg az optimumot (vagy az optimumérték valamilyen becslését), a végrehajtásukhoz szükséges idő mekkora és hogyan aránylik a teljes megoldás idejéhez. Ez az arány egyes esetekben igen nagy *(Hillier típusú módszerek)*, máskor azonban igen kicsi. Dolgozatunkban ezzel az utóbbi esettel foglalkozunk. Ide tartoznak az egyszerűbb struktúrájú feladatok — hátizsák, általánosított hátizsák, halmazfedési probléma — heurisztikus módszerei.

A probléma, amit vizsgálunk, a következő. Egy felhasználó jelentkezik egy olyan nagyméretű problémával, aminek megoldása egzakt módszerekkel reménytelennek tűnik. (A méret relatív nagysága függ a feladat szerkezetétől. A korábban már említett speciális feladatok esetében az általános feladatokhoz viszonyítva többszörös méretű feladatok jelentenek ugyanakkora számítástechnikai nehézséget.) Mi a teendő ekkor? Nyilván heurisztikus módszereket fogunk alkalmazni. A felhasználó a nagyméretű probléma „megoldásá"-ért hajlandó egy bizonyos határig fizetni. Ez a határ jóval fölötte van egy heurisztikus módszer egyszeri végrehajtási költségének. Éppen ezért eljárásunkat ismételten alkalmazni fogjuk. A dolgozatban arra kívánunk módszert adni — a kizárólagosság igénye nélkül —, hogy hogyan lehet a heurisztikus algoritmusok ismételt végrehajtásából egy rendszert szervezni, hogy a felhasználónak minél többet garantálni tudjunk. Ez a több jelen esetben az optimális megoldás megtalálásának nagy valószínűsége lesz adott ráfordítások mellett.

A költségeket mind gépidőben, mind pézben számolhatjuk. Mivel ezek közvetlenül átválthatók egymásba, ezért a továbbiakban mindig gépidőkorlátot feltételezünk.

FARKAS MIKLÓS írja a szimultán tanulásról szóló [3] cikkének befejezésében; „Ahhoz azonban, hogy értelmes kísérletet folytassunk, értékelhető mérési adatokra tegyünk szert, bizonyos elméleti kiindulási alapra van szükségünk. Semmi sem gyakorlatiasabb, mint egy jó elmélet. Itt csupán kiindulási alapot kíséreltünk megadni a mérésekhez és a további elméleti kutatáshoz." A jelen dolgozat is pontosan ebben az értelemben vett munkahipotézis.

2. Heurisztikus módszerek általános tulajdonságai

Eljárásunkat a 0–1-es esetre fogjuk kidolgozni, ezért a heurisztikus módszerek általános tulajdonságait is ennek megfelelően fogalmazzuk meg.

Egy diszkrét programozási feladatot két tényező határoz meg alapvetően: a megengedett pontok halmaza és a célfüggvény.

Egyes heurisztikus eljárások csak speciális feladatokra vonatkoznak, ami mind a megengedett vektorok halmazának, mind a célfüggvénynek jelentheti bizonyos rögzített tulajdonságait. Itt azt feltételezzük, hogy a majdan elvégzendő átalakítások nem vezetnek ki ebből a körből.

Tehát egy feladatot egy

(2.1) $$(S, f)$$

pár határoz meg, ahol

$$S \subset \{0, 1\}^n$$

és

$$f\colon \boldsymbol{R}^n \to \boldsymbol{R}.$$

A megoldandó feladat pedig

(2.2) $$\max f(\mathbf{x})$$

$$\mathbf{x} \in S.$$

A heurisztikus eljárástól csak annyit követelünk meg, hogy egy bináris pontot adjon. Ez nem jelenti azt, hogy ennek a pontnak megengedettnek kell lennie. Ennek az első pillanatra meglepő ténynek az a magyarázata, hogy sok esetben egyetlen megengedett megoldás megadásának nehézsége elméleti szempontból megegyezik a teljes feladat megoldásának nehézségével. Ez a helyzet például az általános diszkrét programozási feladat esetében is.

Tekintsünk most egy (2.2) alakú feladatot. A korábbiak szerint egy heurisztikus eljárás ehhez a feladathoz egy bináris vektort rendel, amit heurisztikus megoldásnak nevezünk. Most a dimenzió rögzítve volt. Ha ebből a feladatból valamely egyszerű transzformációval egy alacsonyabb dimenziós feladatot nyerünk és alkalmazzuk az új problémára ugyanezt a heurisztikus eljárást, akkor nyilván a két heurisztikus megoldás nem lehet független egymástól. Ezeket a tulajdonságokat fogalmazza meg az alábbi definíció.

2.1. Definíció. A χ_n $(n=1, 2, \ldots)$ függvények rendszerét heurisztikus lejárásnak nevezzük, ha

(i) $\chi_n\colon 2^{\{0,1\}^n} \times (\boldsymbol{R}^n \to \boldsymbol{R}) \to \{0, 1\}^n$.

(ii) Legyen k egy 1 és n közé eső egész $(n \geqq 2)$, valamint f egy $n-k$ és g egy n változós függvény, úgy, hogy

$$f(x_1, \ldots, x_{n-k}) \equiv g(x_1, \ldots, x_{n-k}, \delta_{n-k+1}, \ldots, \delta_n),$$

ahol $\delta_j = 0$ v. 1 $(j = n-k+1, \ldots, n)$ rögzített szám. Legyen továbbá S és T az $n-k$, ill. n dimenziós bináris vektorok egy-egy halmaza, úgy, hogy

$$\mathbf{y} \in T \Leftrightarrow \exists \mathbf{x} \in S, \quad \mathbf{y}^T = (x_1, \ldots, x_{n-k}, \delta_{n-k+1}, \ldots, \delta_n)^T.$$

Ekkor

$$\begin{pmatrix} \chi_{n-k}(S, f) \\ \delta_{n-k+1} \\ \vdots \\ \delta_n \end{pmatrix} = \chi_n(T, g). \tag{2.3}$$

A definícióban az (ii) kritérium azt adja, hogy ha a feladatban néhány változó értéke δ_j szinten rögzítve van, akkor a heurisztikus megoldást úgy is megkaphatjuk, hogy a megfelelő alacsonyabb dimenziós heurisztikus megoldást ezekkel a δ_j komponensekkel kiegészítjük. A (2.3) egyenlőség mindkét oldalán tehát egy-egy n-dimenziós vektor áll.

Most megfogalmazunk néhány hipotézist, amelyek a heurisztikus eljárások lényeges tulajdonságait vannak hivatva leírni. Ezek és a később megadandó hipotézisek jogosságára a dolgozat végén visszatérünk.

A bevezetőben már említettük, hogy korlátozni akarjuk a ráfordítások, vagyis a számítások mennyiségét. Ehhez ismernünk kell, hogy ez milyen mennyiségekből tevődik össze. Mivel a heurisztikus eljárásunkat ismételten végre fogjuk hajtani, a számítások egy jelentős része innen fog adódni.

I. Hipotézis. A χ_n függvény egyszeri kiértékelésének, tehát a heurisztikus eljárásnak egy n-változós feladatra való egyszeri végrehajtásának költsége minden természetes n-re adott. Ezt a továbbiakban c_n jelöli.

Mi az optimális megoldást szeretnénk megkapni minél nagyobb valószínűséggel.

II. Hipotézis. A χ_n függvény p_n valószínűséggel ad optimális megoldást egy tetszőleges n-változós feladaton.

2.2. Definíció. Legyen f^* a (2.2) feladat optimális célfüggvényértéke és $\mathbf{x} \in S$. Azt mondjuk, hogy valamely $\varepsilon > 0$-ra $\mathbf{x}$ ε-közelítő megoldás, ha

$$\frac{|f^* - f(\mathbf{x})|}{|f(\mathbf{x})|} < \varepsilon.$$

3. Bináris fák

Ebben a szakaszban tárgyaljuk azokat a kombinatorikus segédeszközöket, amelyekre az eljárásunk támaszkodni fog.

3.1. Definíció. Egy véges irányított fát bináris fának nevezünk, ha minden csúcsnál a befutó élek száma legfeljebb egy és a kifutó élek száma vagy kettő, vagy a csúcs végpont.

3.2. Definíció. Egy irányított gráfban egy csúcsot gyökérpontnak nevezünk, ha az illető csúcsnál a befutó élek száma nulla. Továbbá egy pont végpont, ha a kifutó élek száma nulla.

3.1. TÉTEL. Egy G bináris fában pontosan egy gyökérpont van.

A bizonyítást elhagyjuk, mert az a szokásos gráfelméleti eszközökkel könnyen elvégezhető. Hasonlóan könnyű a következő állítás is.

3.2. LEMMA. Egy G bináris fa gyökérpontjából minden csúcsához vezet irányított út.

3.3. Definíció. Legyen G bináris fa és v a gyökérpontja, w pedig egy tetszőleges további pontja. Azt mondjuk, hogy w a k-adik szinten van, ha v-ből w-be k hosszúságú irányított út vezet. A v pont szintje 0.

A csúcsok szintjei egyértelműen meghatározottak. Ha ugyanis nem ez volna a helyzet, akkor G, illetve a neki megfelelő irányítatlan gráf tartalmazna kört, ami nem lehetséges, lévén G fa.

Mindezek alapján úgy kell elképzelnünk egy bináris fát, hogy egy csúcsból — a gyökérpontból — kiindulva bizonyos pontokból további két kifelé mutató élt húzunk be.

3.4. Definíció. Egy G bináris fát n-szintű teljes bináris fának nevezünk, ha G minden végpontja az n szinten van.

3.3. LEMMA. Az n-szintű teljes bináris fa csúcsainak száma $V_n=2^{n+1}-1$, éleinek száma $E_n=2^{n+1}-2$.

Bizonyítás. Mivel fáról van szó, ezért nyilván elég csak a csúcsokra vonatkozó állítást igazolni. Legyen $0\leqq j<n$. A k-adik és a $(k+1)$-edik szint között a csúcsok száma megduplázódik. Mivel a 0. szinten egy csúcs van ezért a szögpontok száma:

$$\sum_{k=0}^{n} 2^k = 2^{n+1}-1.$$

3.5. Definíció. Legyen G egy olyan bináris fa, hogy bármely végpontjának szintje legfeljebb n ($n\geqq 1$). Legyen $\mathbf{d}$ az a nemnegatív egészekből álló n-dimenziós vektor, amelynek k-adik komponense, d_k, azt mutatja meg, hogy G-nek hány végpontja van a k-adik szinten. Ekkor a $\mathbf{d}$ vektort a G bináris fa stratégiájának nevezzük.

3.4. TÉTEL. Egy $\mathbf{d}$ nemnegatív egészekből álló vektor akkor és csak akkor egy G bináris fa stratégiája, ha

$$\sum_{k=1}^{n} 2^{n-k} d_k = 2^n. \tag{3.1}$$

A bizonyítás BENDE SÁNDOR [2] cikkének 1° és 1°′ pontjaiban megtalálható.

3.5. TÉTEL. Legyen $\mathbf{d}$ egy n-dimenziós vektor, mely a G bináris fa stratégiája. Ekkor G csúcsainak száma

$$2^{n+1}-1-\sum_{k=1}^{n} d_k(2^{n-k+1}-2).$$

Bizonyítás. Vegyünk egy n-szintű teljes bináris fát, ebben egy G-vel izomorf részgráfot. A részgráf egy k szintű végpontja alatt a teljes fának $2^{n-k+1}-2$ csúcsa van. Ha ezt a számot a részgráf összes végpontjára levonjuk a teljes fa csúcsainak számából, a tételbeli egyenletet nyerjük.

A 3.4. tétel kapcsolatban van az információelmélettel. Mivel a tárgyalást nem akarjuk megtörni, ezért erre a Függelékben térünk vissza.

A továbbiakban azt a kérdést vizsgáljuk, hogyan lehet előállítani adott stratégiájú bináris fát. Megadunk egy általános algoritmus keretet, amely lépésenként — azaz egy-egy csúcsból induló két él behúzásával — építi fel a nekünk kívánatos fát. Az algoritmus végrehajtása során a már generált fa megfelel valamely, állandóan változó stratégiának. Ezt a pillanatnyi stratégiát fogjuk jelölni **g**-vel. Az a stratégia pedig, amelyhez tartozó fát akarunk, legyen **d**, ahol $\mathbf{d} \neq \mathbf{0}$.

Bevezetjük minden k szinthez a t_k mennyiségeket. Ezek szemléletes jelentése, hogy potenciálisan t_k végpontot lehet létesíteni, még a gráfban a k szinten, ha minden magasabb j szinten ($j<k$) létrehozunk d_j végpontot. Legyen tehát

$$t_k = \sum_{j=1}^{k} (g_j - d_j) 2^{k-j}.$$

Természetesen t_k is csak a pillanatnyi állapotot tükrözi, az algoritmus minden lépése után változhat.

3.6. TÉTEL. Tekintsünk egy olyan algoritmust, amely bináris fát épít fel és egy rögzített $\mathbf{d} \neq \mathbf{0}$, a (3.1) egyenletet kielégítő nemnegatív egészekből álló n-dimenziós vektor mellett eleget tesz az alábbiaknak:

(i) kezdetben $g_1=2$; $g_j=0$, $j=2, \ldots, n$,
(ii) mindig csak az egy csúcsból kiinduló két élt húzzuk be,
(ii) csak olyan csúcsot bontunk ketté, amelynek k szintjére $t_k>0$,
(iv) az eljárás csak akkor fejeződik be, ha mindazokra a k-szintekre, amelyekre $t_k>0$, a végpontok száma zérus.

Ekkor az algoritmus eredményeképpen kapott bináris fa olyan, hogy nincs n-nél alacsonyabb szintű végpontja és stratégiája éppen **d**.

Bizonyítás. Az (ii) követelményből következik, hogy n-nél alacsonyabb szintű végpontot csak úgy nyerhetünk, ha legalább egy n-szintű pontból kiinduló két élt behúzunk. Ehhez azonban az kell, hogy valamikor $t_n>0$ teljesüljön. Viszont

$$t_n = \sum_{j=1}^{n} (g_j - d_j) 2^{n-j} = \sum_{j=1}^{n} g_j 2^{n-j} - \sum_{j=1}^{n} d_j 2^{n-j}.$$

A (3.1) egyenlőség alapján itt mindkét összeg értéke 2^n, tehát t_n értéke bármely pillanatban nulla. Ebből azonnal következik az algoritmus végessége is, hiszen összesen véges sok pontból induló élek behúzásáról dönthetünk.

Azt kell tehát belátnunk, hogy a végeredményül kapott bináris fa **h** stratégiája megegyezik **d**-vel.

Mindenesetre $h_1=d_1$. Ugyanis az (i) alapján kezdetben $g_1 \geqq d_1$. Másrészt $t_1 = g_1 - d_1$. Ha ez az érték pozitív, akkor van végpont az 1 szinten és a (iv) kritérium szerint az algoritmus még nem fejeződött be. A végességből következik, hogy később valamelyik 1 szintű csúcsból induló éleket is be kell húzni. Azonban amikor $g_1=d_1$ lesz, akkor $t_1=0$, és a továbbiakban 1 szinten levő csúcsból nem indítunk több élt.

Indirekt módon tegyük fel, hogy **h** és **d** nem azonos. Legyen p az a legkisebb index, amelyik komponensben különböznek. Tehát

$$h_k = d_k, \quad k = 1, \ldots, p-1 \quad \text{és} \quad h_p \neq d_p.$$

Ha $h_p > d_p$, akkor az algoritmus befejeztével

$$t_p = h_p - d_p \geqq 1, \tag{3.2}$$

és ebből az is következik, hogy van p szintű végpont. Ekkor viszont (3.2) ellentmond (iv)-nek.

Tehát csak a $h_p < d_p$ eset fordulhat elő. Vizsgáljuk a bináris fát, azután, hogy utoljára p szintű csúcspontból húztunk be éleket. Ezután a lépés után r_p értéke már nem változott. A későbbi változtatások a bináris fém vagy a p szint alatt történtek és ezek a definíció szerint nem játszanak szerepet a t_p érték kialakításában, vagy a p szint fölött. Az utóbbi esetben, ha a változtatás egy $j < p$ szintű csúcsból induló élek behúzása volt, akkor g_{j+1} 2-vel való növelése, illetve g_j 1-gyel való csökkenése miatt t_p változása:

$$2 \cdot 2^{p-j-1} - 2^{p-j} = 0.$$

Vizsgáljuk meg t_p értékét az említett utolsó p szintű változtatás után. Az előbbieknek megfelelően, p kitüntetett szerepét felhasználva:

$$t_p = \sum_{j=1}^{p} 2^{p-j}(h_j - d_j) = h_p - d_p < 0,$$

ami (iii) miatt lehetetlen.

BENDE SÁNDOR is megmutatja az 1° és az 1°′ pontokban, hogy bármely bináris fa felépíthető az (i) és (ii) követelményeknek elget tevő algoritmussal, nála azonban egy csúcs kettévágása az éppen létező fában a két legalacsonyabb szint valamelyikén történik.

4. Egy globális heurisztikus eljárás

A szokásos heurisztikus eljárásokat lokális eljárásoknak gondolhatjuk. Ugyanis a gyorsaság követelménye szükségképpen magában foglalja azt, hogy az eljárás nem tekinti át a feladat egészét, így a szolgáltatott pont is csak valamilyen környezetben lesz jó, nem feltétlenül az egész feladatra vonatkozóan.

Nulla-egy változókon értelmezett optimalizálási feladatok megoldó algoritmusai közül nagyon sok működik úgy, hogy beletartozik az alábbi általános keretbe. A feladatot vagy másképp fogalmazva, a megengedett megoldások halmazát két részre bontjuk azáltal, hogy egy kitüntetett változó az egyikben csak az 1, a másikban a 0 értéket veheti fel. Az így létrejött részhalmazokat más változóra vonatkozó értékadással újabb darabokra vághatjuk ketté. Vegyük észre, hogy így egy bináris fát hozunk létre, ahol a szögpontok a megengedett megoldások bizonyos halmazai, az élek pedig egy halmazból mindig annak valamely részhalmazaiba mutatnak. Ezt a darabolási eljárást mindaddig folytatjuk, míg a kapott kisebb feladatok már valamilyen eszközzel meg nem oldhatók.

Az itt javaslandó globális heurisztikus eljárás alapgondolata az, hogy a részfeladatok megoldását helyettesítjük a heurisztikus algoritmusnak az illető részfeladatra való alkalmazásával. Tehát a keletkező bináris fa minden végpontjában egy lokális heurisztikus eljárást hajtunk végre. Azonban ezek együttesen a teljes feladatra globális információt tartalmaznak azon a szinten, amelyet a bináris fa a feladathoz viszonyítva képvisel.

A továbbiakban olyan **d** stratégia meghatározását vizsgáljuk, hogy ha a megengedett megoldások részhalmazain létrehozunk egy, a **d** stratégiának megfelelő bináris fát, és a fa végpontjai által reprezentált részfeladatokra alkalmazzuk a heurisztikus eljárásunkat, akkor

— minimális számítási költséggel,
— legalább egy adott p_0 valószínűséggel megtaláljuk az optimális megoldást vagy,
— adott korlátot meg nem haladó számítási költséggel,
— maximális valószínűséggel találunk optimális megoldást.

Az az esemény, hogy egy adott ágban a heurisztikus eljárás megtalálja az egész feladat optimális megoldását két esemény egyidejű bekövetkezését követeli meg, nevezetesen: abban az ágban kell lenni az optimális megoldásnak és a heurisztikus eljárásnak meg kell találnia az adott ág által reprezentált (rész)feladat optimális megoldását. Mivel megadhatók feladatok úgy, hogy a különböző feladatokban azonos változókat azonos értéken rögzítve a megengedett megoldások halmaza és a célfüggvény azonos legyen, míg ugyanezen változók más rögzített értékrendszere esetén a megengedett megoldások halmazai lényegesen különböznek egymástól. Ezért a következő hipotézissel élünk:

III. Hipotézis. Az a két esemény, hogy egy adott ágban van optimális (ε-közelítő) megoldás és a heurisztikus eljárás megtalálja azt, független egymástól.

IV. Hipotézis. Annak valószínűsége, hogy az optimális megoldás egy adott ágban van, az ág nagyságával arányos.

Továbbá az alábbi egyszerűsítő feltételezéssel élünk, mely ugyan nem mindenkor, de az esetek döntő többségében igaz:

V. Hipotézis. A feladatnak csak egy optimális megoldása van.

Egy ág annál nagyobb, minél magasabb szinten van a mgfelelő végpont. Ha egy ágban k különböző változót rögzítettünk, vagyis az ág a k szinten van, akkor a keresett valószínűség a IV. hipotézis és a 3.4. tétel alapján:

$$\frac{1}{2^k}.$$

Ez a feltételezés nyilván jogos mindaddig, amíg ezzel ellenkező információt nem nyerünk. Erre azonban csak a feladat megoldása során van lehetőségünk. Tehát a számítások kezdete előtt, amikor a követendő stratégiánkat kívánjuk meghatározni, még szükségképpen ezen feltételezés mellett kell dolgoznunk.

Tehát annak valószínűsége, hogy a bináris fa egy k szintű végpontja által reprezentált feladatban megtaláljuk az optimális megoldást (a 2. szakasz jelöléseit használva) n bináris változó esetén

$$\frac{p_{n-k}}{2^k}.$$

Mivel — az alternatív optimumok esetétől eltekintve — az optimális pont elhelyezkedése a különböző ágakban egymást kizáró eseményeknek felel meg, ezért egy

$\mathbf{d} \neq \mathbf{0}$ stratégiának megfelelő bináris fa esetén az optimális megoldás megtalálásának valószínűsége

$$\sum_{j=1}^{n} \frac{d_j p_{n-j}}{2^j}$$

lesz. Ennek az összegnek kell nagyobbnak lennie az adott p_0 értéknél. A feltételeink tehát a $\mathbf{d}$ stratégiára:

$$\sum_{j=1}^{n} d_j 2^{n-j} = 2^n,$$

$$\sum_{j=1}^{n} \frac{d_j p_{n-j}}{2^j} \geqq p_0, \tag{4.1}$$

$$d_j \geqq 0 \quad \text{egész} \quad k = 1, \ldots, n.$$

Vizsgáljuk meg a felmerülő számítási költségeket. Ezek részben a bináris fa létrehozásából, részben a heurisztikus módszerek végrehajtásából származnak.

Az utóbbiak meghatározása az egyszerűbb, ezért ezzel kezdjük. A $\mathbf{d}$ stratégiát elfogadva az $n-k$ változós feladatok közül d_k-ra fogjuk alkalmazni a módszert, így az összes ilyen költség

$$\sum_{k=1}^{n} c_{n-k} d_k$$

lesz.

A bináris fa létrehozásának ára arányos a fa éleinek számával. A 3.5. tétel alapján tudjuk, hogy a $\mathbf{d}$ stratégia esetén az élek száma

$$2^{n+1} - 2 - \sum_{k=1}^{n} d_k (2^{n-k+1} - 2).$$

Legyen egy új ág létesítéséhez szükséges számítási mennyiség a. Tehát a bináris fa generálásához elvégzett munka

$$a(2^{n+1} - 2) - a \sum_{k=1}^{n} d_k (2^{n-k+1} - 2).$$

Ebből az

$$a(2^{n+1} - 2)$$

tag konstans, így a $\mathbf{d}$ stratégia értékelésében nem játszik szerepet.

Azt kaptuk tehát, hogy a számítási költségek $\mathbf{d}$-től függő része

$$\sum_{k=1}^{n} [a(-2^{n-k+1} + 2) + c_{n-k}] d_k \tag{4.2}$$

lesz. Feladatunk tehát a (4.1) feltételek mellett a (4.2) függvényt minimalizálni.

Ez a megfogalmazás azt jelenti, hogy minimális ráfordítás mellett legalább egy adott valószínűséggel kívánjuk megtalálni az optimális megoldást. Természetesen a probléma megfordítható, hogy adott ráfordítás mellett az optimális megoldás meg-

találásának valószínűségét kívánjuk maximalizálni. Ekkor a probléma matematikai alakja a következő:

$$\max \sum_{j=1}^{n} \frac{d_j p_{n-j}}{2^j},$$

$$\sum_{j=1}^{n} d_j 2^{n-j} = 2^n,$$

(4.3)

$$\sum_{j=1}^{n} [a(2-2^{n-j+1})+c_{n-j}]d_j \leqq K-a(2^{n+1}-2),$$

$$d_j \geqq 0, \quad \text{egész}, \quad j = 1, \ldots, n,$$

ahol K az elvégezhető számítási mennyiség felső korlátja.

Mind a (4.1)—(4.2), mind a (4.3) feladat egy lineáris diszkrét programozási probléma. Megoldása látszólag bonyolultabb, mint az eredeti feladaté, hiszen a változók nem binárisak. A valóságban nem ilyen rossz a helyzet. Először is vegyük észre, hogy a d_k változók közül nagyon sok — a nagy indexűek — bizonyosan a 0 értéket veszi fel, mert egy ésszerű határnál jobban nem fogjuk darabolni a megengedett megoldások halmazát. Nagyon fontos különbség a (4.1)—(4.2), a (4.3) probléma és egy (2.2) alakú feladat között, hogy amíg az utóbbi csak egyszer merül fel és akkor meg kell oldani, addig az előzőek megoldása stratégiát jelent minden konkrét (2.2) probléma kezeléséhez. Végezetül megjegyezzük, hogy amennyiben **g** a (4.1)—(4.2) vagy a (4.3) feladat megoldása, akkor az az $n+1$ dimenziós **d** vektor, amelynek komponenseire:

$$d_1 = 0; \quad d_i = 2g_{i-1}, \quad i = 2, \ldots, n+1$$

jó közelítő megoldása lesz az eggyel nagyobb dimenziós feladatnak.

5. Dinamikus módszerek

Az egzakt eljárások a diszkrét programozásban felhasználnak a heurisztikus módszereken kívül még két olyan elemet, amelyeknek az alkalmazásáról kár volna lemondanunk. Ezek a célfüggvény becslése és a különböző tesztek. Az utóbbiak által szolgáltatható legfontosabb eredmények a következők. A megengedett megoldások egy bizonyos részhalmaza üres, vagy ennek minden vektorában az egyik komponens mindig 1 vagy 0. Az első esetben az adott részhalmazt nem kell tovább vizsgálni, a másodikban pedig a meghatározott változók a megadott értéken leköthetők mindaddig, amíg az adott részhalmazban vizsgálódunk. A bináris fákon megfogalmazva az első esetben a megengedett megoldások bizonyos részhalmazának megfelelő pont egyben végpont, a másik esetben pedig a fa adott pontjából csak egyetlen él indul kifelé. Ezt a lehetőséget korábban kizártuk. Az esetleges zavarokat elkerülendő ilyenkor mindig úgy fogjuk tekinteni a fát, hogy a másik élt is behúztuk, de annak végpontjából további élek nem indulnak ki.

A továbbiakban feltesszük, hogy a (2.2) feladat $f(\mathbf{x})$ célfüggvénye nemnegatív. Az esetek döntő többségében egyszerű átalakításokkal ez elérhető.

A legegyszerűbb esetben, amikor csak felső becslést használunk a heurisztikus eljáráson kívül, módszerünk a következő lesz. Előre eldöntjük, hogy milyen **d** stra-

tégiát válasszunk, és egy ennek megfelelő fát építünk fel fokozatosan. Az eljárást a 3.6. tétel követelményeinek betartásával végezzük. Nagyrészt szabad kezünk van a következő lépésben felbontandó csúcs kiválasztásánál és teljesen tőlünk függ, hogy a felbontás milyen kritérium (mely változó) szerint történik. Könnyen látható, hogy lényegében nem történik más, mint végrehajtjuk egy korlátozás és szétválasztás típusú algoritmus „tetejé"-t. Tehát a következő csúcs kiválasztásánál használhatjuk bármelyik, a korlátozás és szétválasztás típusú eljárásokban szokásos taktikát. Ezek nem használnak fel mást, mint a fa struktúráját és a becsléseket. Részletes leírást ad róluk [5].

Ha közben az egyes végpontok által reprezentált feladatokra alkalmazzuk a heurisztikus eljárásunkat és megengedett megoldáshoz jutunk, akkor ugyancsak a korlátozás és szétválasztás típusú eljárásokban szokásos módon mindazon végpontok további vizsgálata feleslegessé vált, amelyekhez tartozó becslés rosszabb, mint a már ismert megengedett megoldás célfüggvényértéke.

A felbontandó csúcs kiválasztására felmerül egy a szokásostól némileg eltérő lehetőség is. Mindig azt a csúcsot bontjuk fel, ahol a legnagyobb a valószínűsége annak, hogy az adott részhalmazban van az optimális megoldás. Korábban a 4. szakaszban, amíg a megoldandó feladatról semmi információnk nem volt, azt mondottuk, hogy annak a valószínűsége, hogy a keresett pont egy adott ágban van, csak az ág nagyságától függ. Időközben azonban további információkat, nevezetesen becsléseket kaptunk. Nyilvánvaló tehát, hogy a valószínűségnek függnie kell ettől is. Természetes követelménynek tűnik, hogy a keresett valószínűség mind az ág, mind a becslés nagyságában monoton növekedjék. Az algoritmus végrehajtása során az ágak relatív nagysága állandóan változik, ugyanis a bináris fa egyes pontjairól kiderülhet, hogy ott biztosan nem lehet az optimális megoldás, tehát ezekkel a továbbiakban nem kell számolni. Legyen a vizsgált csúcs C, a relatív nagysága q_C. Az összes ág együttes területét 1-nek vesszük, tehát

$$0 < q_C \leqq 1.$$

A C halmazon a felső becslés legyen f_C. Ekkor jelölje p_C annak a valószínűségét, hogy az optimális megoldás C-ben van. Ezt $p_C = p(q_C, f_C)$ alakban kereshetjük a mindkét változójában monoton növő $p(.,.)$ függvényre vonatkozó valamely ésszerű feltételezés mellett. Felbontásra tehát ekkor azt a csúcsot választjuk, ahol a $p(.,.)$ függvény értéke a legnagyobb.

Mielőtt a további részletekbe belemennénk, ki kell térni egy fontos kérdésre. Említettük már, hogy egyes ágak vizsgálata anélkül abbamaradhat, hogy a megfelelő részfeladatra alkalmaznánk a heurisztikus eljárást, illetve a fa szükséges további építését ebből a csúcsból végeznénk. Ha az ilyen végpontok száma megnő, akkor előfordulhat, hogy egy **d** stratégiájú fa létre sem hozható. Legyen **h** az a vektor, amelynek h_k komponensei megadják, hogy a k szinten hány lezárt ág van. Egy ág akkor lezárt, ha nem lehet benne az optimális megoldás (a becslése rossz, vagy megengedett megoldás sem lehet benne), vagy az adott csúcsból kiindulva végrehajtottuk már a heurisztikus eljárást.

A 3.6. tétel jelöléseitől némileg eltérve legyen a pillanatnyi stratégiát leíró vektor $\mathbf{g}+\mathbf{h}$.

Vizsgáljuk meg a k-adik szinten biztosan felbontható csúcsok pillanatnyi számát. Ha

$$h_k \geqq d_k,$$

akkor már megkaptuk a **d** stratégiának megfelelő számú végpontot a k szinten, ezért minden még le nem zárt csúcs, ami ezen a szinten van, felbontható, vagyis összesen g_k. Ha

$$h_k < d_k,$$

akkor a biztosan felbontható csúcsok száma

$$g_k + h_k - d_k.$$

Ha ez az érték negatív, akkor az azt jelenti, hogy ha ezen a szinten levő csúcsokból egyet sem bontunk fel, akkor is a magasabb szinteken levőkből úgy kell további éleket behúzni, hogy a keletkező új végpontok közül legalább $d_k - h_k - g_k$ a k-adik szintre essék. Mindezek alapján a k-adik szinten jelentkező tartalék

$$t_k^* = \sum_{j=1}^{k} \min (g_j, g_j + h_j - d_j) 2^{k-j}. \tag{5.1}$$

Tehát t_k^* az a mennyiség, amit a k-adik szinten levő csúcsok közül fel lehet bontani — figyelembe véve a később esedékes magasabb szinten levő felbontásokat is — anélkül, hogy a **d** stratégiát megsértenénk. A későbbiekben $\mathbf{t}^*$ jelöli az (5.1)-ben definiált mennyiségekből alkotott n-dimenziós vektort. Ez tehát mindig pillanatnyi értéket képvisel.

Most rátérünk a tesztek használatára. Alkalmazásuk csak akkor célszerű, ha a várható számítási megtakarítás nagyobb, mint az ugyanilyen ráfordítás.

Tegyük fel, hogy az a csúcs, ahol a teszteket alkalmazni kívánjuk, az m szinten van és segítségükkel az m szintről a p szintre ugrunk.

Ebben az esetben a tesztek eredménye, hogy az m szinten levő csúcsból induló és majdan generálandó részfát helyettesíthettük egy $p-m$ hosszúságú úttal. A fa egészére nézve ez kétféle megtakarítást jelent. Egyrészt elmarad a heurisztikus módszer alkalmazása bizonyos esetekben, másrészt a fa további kiépítésének egy része ugyancsak szükségtelenné válik. Vegyük sorba a kettőt.

Legyen $\mathbf{t}_0^* = \mathbf{t}^*$ a pillanatnyi tartalékok vektora, $\mathbf{t}_1^*$ pedig a tartalékok vektora a tesztek alkalmazása után. Hasonlóan a lezárt ágakra definiáljuk a $\mathbf{h}_0$ és $\mathbf{h}_1$ vektorokat. Nyilván a komponenseikre igaz a következő:

$$h_{1k} = \begin{cases} h_{0k}, & \text{ha} \quad k \leqq m \\ h_{0k}+1, & \text{ha} \quad m+1 \leqq k \leqq p, \\ h_{0k}, & \text{ha} \quad k > p \end{cases} \qquad g_{1k} = \begin{cases} g_{0k}, & \text{ha} \quad k \neq m, p \\ g_{0m}-1, & \text{ha} \quad k = m \\ g_{0p}+1, & \text{ha} \quad k = p \end{cases}.$$

Ennek alapján számolható a $\mathbf{t}_1^*$ vektor is. A heurisztikus módszerek vizsgálata azért marad el, mert nem lehet létrehozni a **d** stratégiának megfelelő fát, illetve egyes végpontok alkalmatlanok a további vizsgálatra. Ezért a megtakarítás egyrészt azokon a szinteken jelentkezik, ahol a tartalék negatívvá vált, vagy a negatív tartalék tovább csökkent, másrészt azokon a k szinteken, ahol a lezárt ágak száma nőtt, de még nem hajtottunk végre d_k heurisztikus eljárást, vagyis $h_{0k} < d_k$. E két mennyiség nyilván független egymástól, mert a tartalék csak akkor válhat egy szinten negatívvá, ha egy nála magasabb szinten a lezárt ágak száma túl sok. Tehát a heurisztikus módszerek végrehajtásának elmaradásából származó megtakarítás a következő:

$$\sum_{\substack{k: \\ t_{1k}^* < 0}} [\min \{t_{0k}^*, 0\} - t_{1k}^*] c_{n-k} + \sum_{\substack{k: \\ m \leqq k \leqq p \\ h_{0k} < d_k}} c_{n-k}.$$

A fa kiépítésének elmaradásából származó megtakarítást csak becsülni lehet, mivel nem ismert a fa pontos szerkezete és benne a létre nem jövő ágak szerepe. Mivel mi egy, az m szinten levő csúcsban vagyunk, a szerkezeti módosulás csak ezen szint alatt jöhet létre, pontosabban szólva az adott csúcsból induló részfában. Tényleges megtakarítás csak akkor valósul meg, ha elmarad valamilyen ágak létrehozása. Ellenkező esetben ugyanis a később létrehozandó fa szerkezete változik csak meg. Eszerint a k szinten ha $t^*_{1k}<0$

$$w_k = \min\{t^*_{0k}, 0\} - t^*_{1k}$$

végpontot takarítottunk meg; ennyihez nem fog út vezetni az m szintről.

Általában az a helyzet, hogy a 0 szintről az 1 szintre egy él 1 csúcshoz vezet. Ugyanez az él a 2 szintre — közvetve — 2 csúcshoz, a j szintre 2^{j-1} csúcshoz vezet. Tehát a j szinten levő csúcs esetén a hozzá vezető útból az $i-1$ és az i szintek között vezető élnek az adott csúcsra eső költsége

$$\frac{1}{2^{j-i}}.$$

Ezek alapján a mértani haladvány összeg képletét felhasználva a fa kiépítésének elmaradásából származó költségmegtakarítás

$$a \sum_{k=m+1}^{n} w_k \left(2 - \frac{1}{2^{k-m-1}}\right),$$

ahol a egy él létesítésének ára.

Ha ismerjük a tesztek alkalmazása eredményének eloszlását (tehát tudjuk, hogy milyen valószínűséggel ugrunk az m szintről a p szintre), akkor megkaphatjuk a megtakarítás várható értékét. Az egész heurisztikus eljárás kezdetén viszonylag nagy ugrás esetén is aránylag kicsi lehet a megtakarítás, hiszen a **d** stratégia elérésére a fa más ágain még megvan a lehetőségünk. Mégis az itt alkalmazott tesztek alapozzák meg a későbbi nyereséget. Ekkor tehát eleve egy módosított költség/haszon aránnyal célszerű számolni.

A feladat megkezdésekor csak arról lehet képünk, hogy egy „átlagos" feladat esetén az ugrás nagyságának valószínűségei hogyan alakulnak, de ezt a megoldás során a tapasztaltaknak megfelelően állandóan korrigálni kell. Célszerűnek látszik, hogy bizonyos időközönként mindenképpen alkalmazzuk a teszteket.

Az algoritmus végeredményének nemcsak a megismert legjobb megengedett megoldást tekintjük, hanem az egész generált fát az ágakon a célfüggvény felső becsléseivel és a talált megengedett megoldásokkal. Lehetőségünk nyílik arra, hogy az eljárás után annak eredményét utólag értékeljük.

VI. Hipotézis. Az adott feladatosztályra a becslés relatív hibája egy adott eloszlást követ, azaz legyen egy ágon az optimális célfüggvény érték $z<0$, a célfüggvény felső becslése f, $\alpha>1$ pedig valós szám. Ekkor létezik egy olyan $F(\alpha)$ függvény, hogy

$$P\left(\frac{f}{z} < \alpha\right) = F(\alpha). \tag{5.2}$$

A valószínűségi változó szerepét (5.2)-ben z fogja játszani. Legyen ugyanis az ismert legjobb megengedett megoldás célfüggvényértéke z_0. Mi arra vagyunk kíváncsiak, hogy mi a valószínűsége annak, hogy z_0 egyben az optimális célfüggvényérték is. Ehhez az kell, hogy az eljárás során generált ágak egyikében sem legyen jobb megengedett megoldás. Az általunk vizsgált ágban

$$\alpha = \frac{f}{z_0}$$

választás esetén

$$1 - F(\alpha)$$

annak a valószínűsége, hogy az ágban található megengedett megoldások célfüggvény értéke nem haladja meg z_0-t. Tegyük fel, hogy az eljárás folyamán M ágat hoztunk létre. Nekünk csak azon i ágakat kell figyelembe venni, ahol a célfüggvény f_i becslésére

$$f_i > z_0.$$

Ha egy i ág üres, akkor $f_i = -\infty$. Így annak valószínűsége, hogy megtaláltuk az optimális megoldást

$$\prod_{\substack{i=1 \\ f_i > z_0}}^{M} \left(1 - F\left(\frac{f_i}{z_0}\right)\right). \tag{5.3}$$

Ez összevethető azzal a valószínűséggel, amit a **d** stratégiához rendeltünk.

Az (5.3) képlethez hasonlóan bármely $\varepsilon > 0$ esetén

$$\prod_{\substack{i=1 \\ f_i > (1+\varepsilon)z_0}}^{M} \left(1 - F\left(\frac{f_i}{(1+\varepsilon)z_0}\right)\right) \tag{5.4}$$

annak valószínűsége, hogy a talált legjobb megengedett pont ε-közelítő megoldás.

6. A hipotézisek jogosságáról

A hipotézisek sorában az első helyen azt a fentiekben ki nem mondott állítást kell említeni, hogy a 2.1. definíciónak eleget tevő heurisztikus módszer létezik. Az olvasó könnyen meggyőződhet arról, hogy a *Lagrange-szorzókon* alapuló eljárás ilyen (l. pl. [8]), de több más módszerről is be lehet bizonyítani ugyanezt.

Az I. hipotézis minden megszorítás nélkül igaz az egy felhasználót kiszolgáló gépeken javító lépést nem tartalmazó eljárásokra. (Az utóbbi kitétel a *Lagrange szorzós eljárás* esetében pl. azt jelenti, hogy csak egyetlen szorzóvektort választunk.) Jól közelíti a hipotézis a valóságot, ha a javító lépések száma erősen korlátozott. Ez nem mondható el a szomszédság fogalmon alapuló heurisztikára (l. [9]), itt c_n csak várható értéket jelenthet. Ha egy gépen több felhasználó is dolgozik egyszerre, akkor igen jelentős eltérések lehetnek az esetek egy előre nehezen jelezhető kisebb részében. Többnyire azonban az időbeli eltérés nem jelentős.

Azon feladatok közül, amelyekre a jelen dolgozat vonatkozik, a legfontosabb a lineáris 0–1 programozási feladat. Tegyük fel, hogy a megengedett megoldások konvex burka egy nem elfajuló, valódi n-dimenziós poliéder. Ekkor azon célfüggvény

vektorok halmaza, ahol az optimális megoldás nem egyértelmű, több alacsonyabb dimenziós konvex poliéder uniója. Így a szokásos matematikai értelemben nullmértékű halmazt alkot. Ez támasztja alá az V. hipotézis jogosságát.

Az egyik legtöbbet vizsgált NP-teljes probléma az utazó ügynök feladat. Ennek jelentős előnye az általunk vizsgált problémákkal szemben, hogy a heurisztikus eljárások bizonyosan megengedett megoldást adnak. A gyakorlatban ez a diszkrét programozási feladatokról is elmondható [9], bár nem bizonyítható. KRONE és STEIGLITZ mérték a VI. hipotézisben megadott $F(\alpha)$ függvényt az utazó ügynök feladatra [6]. GOLDEN és ALT itt *Weibull eloszlást* tételez fel [4]. Ugyancsak az utazó ügynök feladatra a II. hipotézisben megadott p_n valószínűségekhez hasonlók létezését feltételezi BELLMORE és MALONE [1]. LENSTRA és RINNOOY KAN kritikájukban nem vonják kétségbe az ilyen valószínűségek létezését [7]. Feltételezhető, hogy ezek a gondolatok átvihetők a diszkrét programozási problémákra is. Ez azért nem történt meg eddig, mert ezek a feladatok sokkal nehezebben kiértékelhetők.

IRODALOM

[1] BELLMORE, MANDELL and MALONE, JOHN C., "Pathology of traveling-salesman subtour-elimination algorithms", *Operations Research* **19** (1971) 278—307.

[2] BENDE, S., „A $\sum_{i=1}^{n} \frac{1}{2^{x_i}}=1$ diophantoszi egyenlet és annak gárfelméleti és információelméleti vonatkozása", *Matematikai Lapok* **18** (1967) 323—327.

[3] FARKAS, M., „A szimultán tanulás dinamikai elmélete", *Alkalmazott Matematikai Lapok* **2** (1976) 103—114.

[4] GOLDEN, BRUCE L. and ALT, FRANK B., „Interval estimation of a global optimum for large combinatorial problems", *Naval Research Logistic Quaterly* **26** (1979) 69—76.

[5] KOVÁCS, L. B., *Combinatorial Methods of Discrete Programming* (Akadémiai Kiadó, Budapest 1980).

[6] KRONE, MARTIN J. and STEIGLITZ, KENNETH, "Heuristic-programming solution of a flowshop-scheduling problem", *Operations Research* **22** (1974) 629—638.

[7] LENSTRA, J. K. and RINNOOY KAN, A. H. G., "On the expected performance of branch-and-bound algorithms", *Operations Research* **26** (1978) 347—349.

[8] VIZVÁRI, B., „A Lagrange szorzók használata diszkrét programozási algoritmusokban", *Alkalmazott Matematikai Lapok* **2** (1976) 413—425.

[9] VIZVÁRI, B., "The heuristic methods of discrete programming — I: The method of neighbourhood", *MTA SZTAKI Tanulmányok* **152** (1983) 109—138.

(Beérkezett: 1987. május 4.)

VIZVÁRI BÉLA
MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZET
1111 BUDAPEST, KENDE U. 13—17

GLOBAL HEURISTIC METHODS IN DISCRETE PROGRAMMING: A STOCHASTIC APPROACH

B. VIZVÁRI

We investigate how a heuristic method should be applied several times to solve a large-scale zero-one programming problem. Under some hypotheses, we give integer programming models: (i) to determine the optimal solution with maximal probability when the computation is limited in time; or (ii) to get the optimal solution with a given probabilit y and with minimal quantity of computation. The sequence of the repeated application of the heuristic method is called global heuristic method. In our approach we design such a branch and bound type method where the solution of the subproblems are substituted by the execution of the heuristic method. Finally we discuss the justification of the hypotheses and show similar ideas in the literature.

Függelék: **Az információelmélet és a diszkrét programozás kapcsolatáról**

Az információelmélet vizsgálja azt a problémát, hogy egy viszonylag bő alap *abc* betűit hogyan lehet gazdaságos módon kódolni egy ún. csatorna *abc* kisszámú jelkészletéből alkotott sorozatokká, melyeket kódszavaknak nevezünk. Mi csak olyan kódokkal foglalkozhatunk, ahol a csatorna abc összesen két elemből áll.

F.1. Definíció. Legyen $\varkappa$ kódszavaknak egy halmaza. Ekkor $\varkappa$-t prefix tulajdonságúnak nevezzük, ha egyetlen kódszó sem meghosszabbítása egy másiknak. Tehát nem áll fönn a következő eset:

$$k, m \in \varkappa; \quad k = (k_1, \ldots, k_p), \quad m = (m_1, \ldots, m_q);$$
$$p \leqq q;$$
$$k_i = m_i, \quad i = 1, \ldots, p.$$

Bevezetjük a következő jelölést. Egy tetszőleges $k=(k_1, \ldots, k_p)$ kódszóra $|k|=p$, azaz $|k|$ a k kódszó hossza. Könnyen bebizonyítható az

F.1. TÉTEL. Legyen $\varkappa$ egy prefix kód.
Ekkor

$$\sum_{\substack{k: \\ k \in \varkappa}} \frac{1}{2^{|k|}} \leqq 1. \tag{F.1}$$

Az (F.1) feltétel az ún. *MacMillan egyenlőtlenség*. A tétel megfordítható.

F.2. TÉTEL. Jelölje P természetes számoknak egy véges rendszerét, amelyben egy szám többször is előfordulhat. Ha

$$\sum_{p \in P} \frac{1}{2^p} \leqq 1,$$

akkor létezik olyan $\varkappa$ prefix kód, hogy ha Q jelöli a $\varkappa$ kódszavainak hosszaiból alkotott rendszert, akkor Q megegyezik P-vel, vagyis mindkét rendszerben egy természetes szám pontosan ugyanannyiszor fordul elő.

F.2. Definíció. Egy $\varkappa$ kód telített, ha az (F.1) feltétel egyenlőséggel teljesül.

F.3. Definíció. Egy $\varkappa$ prefix kód maximális, ha egyetlen kódszó sem vehető hozzá a prefix tulajdonság elrontása nélkül.

Igaz a következő:

F.3. TÉTEL. Egy prefix kód akkor és csak akkor telített, ha maximális.

Most térjünk vissza a diszkrét programozáshoz. Ebben az esetben felfoghatjuk úgy a dolgot, hogy van egy alap *abc*-nk: a megengedett megoldások halmaza; egy kódolási eljárásunk: a célfüggvény, mely az előbbi halmazból kiválaszt egyet, nevezetesen az optimális pontot és végül létezik egy dekódolási eljárásunk: a diszkrét programozási feladat megoldási algoritmusa.

Amikor egy leszámlálás típusú eljárást hajtunk végre, akkor mindig kérdéseket teszünk fel, melyekre igyekszünk választ kapni. A kérdés mindig ilyen típusú:

„Igaz-e, hogy az optimális megoldás a vizsgált megengedett megoldások halmazának olyan részhalmazában van, ahol x_k értéke δ".

A válasz elhangzása után már csak a válasznak megfelelő megengedett megoldásokat vizsgáljuk tovább. (Megjegyezzük, hogy a gyakorlatban vannak felesleges kérdések is, mert nem minden kérdésre adható igenlő válasz a kérdés feltevése pillanatában. Így a dekódolási eljárás egy kérdésrendszer lesz, a kódszavak pedig az egyes pontokat definiáló kérdéssorozatokra adandó pozitív és negatív válaszok sorozatai lesznek.

Példaként vizsgáljuk meg a lineáris célfüggvény esetét. Legyen $\mathbf{x}_0$ egy megengedett pont. A $\mathbf{c}$ vektort definiáljuk a következő módon

$$c_j = \begin{cases} 1, & \text{ha} \quad x_{0j} = 1 \\ -1, & \text{ha} \quad x_{0j} = 0 \end{cases}, \quad j = 1, \ldots, n.$$

Nyilván a

$$\max \mathbf{c}'\mathbf{x}$$

célfüggvény esetén $\mathbf{x}_0$ az egyedüli optimális megoldás.

Tehát a kérdésrendszernek olyannak kell lennie, hogy egyértelműen elő kell tudni állítania minden megengedett megoldást. Az egyértelműségből következik a prefix tulajdonság. Amíg felesleges kérdéseket nem teszünk fel, addig a megengedett megoldások megfelelő részhalmaza a bináris fán annyiadik szinten helyezkedik el, amennyi a feltett kérdések száma. Ekkor a kód maximális is lesz. Hiszen ellenkező esetben vagy egy megengedett pontra nem tudnánk eljutni, vagy felesleges kérdést tettünk volna fel.

Tehát a 3.4. tétel a *MacMillan egyenlőséget*, ill. annak megfordítását állítja a megoldási algoritmusunkból származó maximális prefix kódra.

Alkalmazott Matematikai Lapok **13** (1987—88), 335—348

VÉGES ELEMEK MÓDSZERE NEMLINEÁRIS, PARABOLIKUS TÍPUSÚ EGYENLETEK MEGOLDÁSÁRA

FARAGÓ ISTVÁN

Gödöllő

Ebben a dolgozatban a nemlineáris, parabolikus típusú egyenletek véges elemes megoldását tárgyaljuk. Az irodalomban megtalálható klasszikus eredmények többségükben a nemlinearitásra vonatkozó *globális Lipschitz-feltételekre* vonatkoznak. Megmutatjuk, hogy ez a feltétel helyettesíthető bizonyos simasági és pozitivitási feltételekkel. Dolgozatunkban bebizonyítjuk a módszer konvergenciáját és annak nagyságrendjét.

1. Bevezetés

Ebben a dolgozatban a [7] cikkben felvetett problémák közül az instacionárius, nemlineáris folyamatokat modellező parabolikus típusú feladatokat és azok véges elemes megoldási módszereit tárgyaljuk. Ezen feladatosztály specifikuma a nemlinearitás. A klasszikus eredmények [4], [5], [12] a nemlinearitásra vonatkozóan a *globális Lipschitz-feltétel* mellett érvényesek. Megmutatjuk, hogy ez a feltétel kiváltható bizonyos simasági és pozitivitási feltételekkel. Ugyanakkor a konvergencia bizonyítása érvényes marad a klasszikus feltételrendszerre is. Dolgozatunkban a [7] cikk eredményeit terjesztjük ki az ottani terminológiai és jelölésrendszer megtartásával. A második szakaszban a szemidiszkrét közelítő sorozat néhány tulajdonságát bizonyítjuk, a harmadik szakaszban pedig ezen sorozatnak az eredeti feladat megoldásához való konvergenciáját és annak a különböző normákban való nagyságrendjét tárgyaljuk. Megjegyezzük, hogy nem tárgyaljuk a térbeli diszkretizáció után nyert közönséges differenciálegyenlet rendszerre kitűzött *Cauchy-feladatok* numerikus megoldását. Ez meghaladja a jelen dolgozat kereteit. A téma iránt érdeklődőknek figyelmébe ajánljuk a [9] és a [11] dolgozatokat.

2. A feladat kitűzése és a szemidiszkrét közelítő sorozat néhány tulajdonsága

Tekintsük a továbbiakban a következő nemlineáris, másodrendű parabolikus típusú parciális differenciálegyenletet:

$$\frac{\partial u}{\partial t} - L(u) = 0, \quad \forall (x, t) \in Q_T = \Omega \times (0, T], \tag{2.1}$$

$$u(x, t) = 0, \quad \forall (x, t) \in \Gamma \times (0, T]; \tag{2.2}$$

$$u(x, 0) = u_0(x), \quad \forall x \in \bar{\Omega}; \tag{2.3}$$

ahol

$$L(u) = \sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\left(F_{ij}(u, x, t)\frac{\partial u}{\partial x_j}\right) - F_0(u, x, t); \tag{2.4}$$

$\Omega \subset R^N$, F_0, F_{ij}: $R^{N+2} \to R$, u_0: $R^N \to R$ adott leképezések.
Jelölje S2 az alábbi feltételrendszert [8]:

1. a) $F_{ij} \in C^{1,1,1}(Q_T^1)$, ahol $Q_T^1 = R \times Q_T \subset R^{N+2}$;

   b) $F_{ij} = F_{ji}$ $\forall i, j = 1, 2, \ldots, N$.

   c) léteznek olyan F_*, F^* pozitív állandók ($F_* \leqq F^*$), hogy tetszőleges $p \in R^N$ és $(s, x, t) \in Q_T^1$ esetén

$$F_* \|p\|_E^2 \leqq \sum_{i,j=1}^{N} F_{ij}(s, x, t) p_i p_j < F^* \|p\|_E^2,$$

ahol $\|p\|_E^2 = \sum_{i=1}^{N} p_i^2$ az R^N-beli euklideszi norma.

2. a) $F_0 \in C^{1,1,1}(Q_T^1)$;

   b) tetszőleges $s \in R$ és $(x, t) \in Q_T$ esetén $s \cdot F_0(s, x, t) \geqq 0$;

   c) $F_0(0, x, t) = 0$, $\forall (x, t) \in Q_T$;

3. a) $u_0 \in C^3(\bar{\Omega})$;

   b) $\operatorname{supp} u_0 \subset \Omega$.

A továbbiakban S1 jelöli azt a feltételrendszert, amely S2-től pontosan a következőkben tér el: az F_{ij}, F_0 simaságára és pozitivitására vonatkozó 1.a, 2.a és 2.b feltételek helyett a *globális Lipschitz-tulajdonságot* feltételezzük, azaz, hogy létezik olyan $L_0 > 0$ állandó, hogy $g = F_0$ és $g = F_{ij}$ $(i, j = 1, \ldots, N)$ esetén

$$|g(s_1, x, t) - g(s_2, x, t)| \leqq L_0 |s_1 - s_2|, \quad \forall s_1, s_2 \in R \tag{2.5}$$

tetszőleges $(x, t) \in Q_T$-re.

Adjunk példát olyan S2 tulajdonságú F_0 leképezésre, amely nem S1-beli.

Legyen $F_0(s, x, t) = s^3$. Könnyen belátható, hogy az S2 feltételrendszer 2. feltétele teljesül, míg a (2.5) globális liptschitzesség nem. (Megjegyezzük, hogy hasonlóan jó példa az $F_0(s, x, t) = s \cdot |s|^\delta$ típusú függvény, amely $0 \leqq \delta \leqq 2$ esetben a reakciókinetikai matematikai modellekben gyakran szerepel.) Vegyük észre, hogy az S2 feltételrendszerben az F_0 leképezésre kitűzött feltételek szoros összefüggésben állnak a monotonitási feltétellel: ha valamely 2.c tulajdonságú függvény kielégíti a monotonitási feltételt, akkor az egyben 2b tulajdonságú is.

Tekintsük a (2.1)—(2.3) feladatot az S2 feltételrendszer teljesülése mellett. Ekkor a parciális differenciálegyenletek elméletének megfelelően a feladatnak létezik egyetlen, klasszikus értelemben vett megoldása $\bar{Q}_T$-n [13]. Könnyen belátható az alábbi következmény is, ami a (2.1)—(2.3) feladattal modellezett jelenségeknél gyakran lényeges:

2.1. ÁLLÍTÁS. Ha $\min_{x \in \bar{\Omega}} u_0(x) \geqq 0$, akkor a (2.1)—(2.3) feladat $u(x, t)$ megoldása nem negatív $\bar{Q}_T$-n.

Bizonyítás. Nyilvánvaló, hogy az S2 feltételek alapján az L operátor pozitív. Ekkor, ha találunk olyan v és w függvényeket, hogy

$$\frac{\partial v}{\partial t} - L(v) \geqq 0 \quad Q_T\text{-n} \quad \text{és} \quad v(x, 0) \geqq u_0(x) \quad \bar{\Omega}\text{-}n;$$

illetve

$$\frac{\partial w}{\partial t} - L(w) \leqq 0 \quad Q_T\text{-n} \quad \text{és} \quad w(x, 0) \leqq u_0(x) \quad \bar{\Omega}\text{-n},$$

akkor $v \geqq u \geqq w \bar{Q}_T$-n [3]. Jelölje v és w a következő konstans függvényeket: $v = \max_{x \in \bar{\Omega}} u_0(x)$ és $w = 0$.
Ekkor teljesülnek a fenti relációk és így érvényes a

$$0 \leqq u(x, t) \leqq \max_{x \in \bar{\Omega}} u_0(x) \tag{2.6}$$

maximum-elvet is bizonyító állításunk.
Vezessük be az

$$a(t)(z; v, w) = \int_{\Omega} \sum_{i,j=1}^{N} F_{ij}(z, x, t) \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_j} dx \tag{2.7}$$

jelölést. Ekkor a (2.1)—(2.3) feladat $u(x, t) \in \tilde{V}_0$ általánosított megoldása $\forall \varphi \in H_0^1(\Omega)$ esetén kielégíti az

$$\left(\frac{\partial u}{\partial t}, \varphi\right) + a(t)(u, u, \varphi) + (F_0(u, \cdot, t), \varphi) = 0 \tag{2.8}$$

$$(u(\cdot, 0), \varphi) = (u_0, \varphi) \tag{2.9}$$

egyenletet tetszőleges rögzített $t \in (0, T]$ esetén.

Legyen a továbbiakban $(\varphi_i^{(n)}(x)) \subset H_0^1(\Omega)$ $(i = 1, \ldots, n)$ az $\Omega \subset R^N$ tartomány egyenletes felosztásával nyert VEM bázisfüggvény rendszere. Ekkor a

$$H^{(n)} = \operatorname{span}[\varphi_1^{(n)}, \ldots, \varphi_n^{(n)}] \tag{2.10}$$

véges dimenziós altérsorozatban a *diszkrét Galerkin-egyenlet*

$$\left(\frac{\partial u_n}{\partial t}, \varphi_i^{(n)}\right) + a(t)(u_n; u_n, \varphi_i^{(n)}) + (F_0(u_n, \cdot, t), \varphi_i^{(n)}) = 0 \tag{2.11}$$

$$(u_n(\cdot, 0), \varphi_i^{(n)}) = (u_0, \varphi_i^{(n)}) \quad (i = 1, 2, \ldots, n); \tag{2.12}$$

ahol az $u_n(x, t)$ megoldás

$$u_n(x, t) = \sum_{i=1}^{n} g_n \alpha_i^{(n)}(t) \varphi_i^{(n)}(x) \tag{2.13}$$

alakú. A (2.13) alakú közelítésben tetszőleges $n \in N^+$ esetén $g_n > 0$ állandó és csak az Ω tartomány felosztásának paraméterétől függ. Például, ha $\Omega \subset R^1$ és $\varphi_i^{(n)}(x)$ az elsőfokú véges elemek bázisfüggvény rendszere, akkor $g_n = h^{1/2}$, ahol $h = \max_{1 \leqq i \leqq n} h_i$. ($h_i = x_{i+1} - x_i$ a diszkretizációs lépésköz).

Nyilvánvaló, hogy (2.11)—(2.12) az $\alpha^{(n)}(t)=[\alpha_1^{(n)}(t), \ldots, \alpha_n^{(n)}(t)$ együtthatókra nézve egy nemlineáris közönséges differenciálegyenlet-rendszer *Cauchy-feladata.*

2.2. ÁLLÍTÁS. A (2.11)—(2.12) feladatnak tetszőleges $n=1, 2, \ldots$ esetén létezik egyetlen $u_n(x, t)$ megoldása és ezen megoldások sorozata tetszőleges rögzített $t\in[0, T]$ mellett egyenletesen korlátos a maximum és az L_2 normákban.

Bizonyítás. A (2.13) alakú előállítás következtében $u_n(x, t)$ létezésének kérdése ekvivalens az

$$\tilde{M}^{(n)} \frac{d\alpha^{(n)}}{dt} + A_0^{(n)}(\alpha^{(n)}, t)\alpha^{(n)} + \tilde{F}^{(n)}(\alpha^{(n)}, t) = 0; \quad 0 < t \leqq T, \tag{2.14}$$

$$\alpha^{(n)}(0) = \alpha_0^{(n)} \tag{2.15}$$

alakú *Cauchy-feladat* megoldhatóságával, ahol

$$A_0^{(n)} = [a_{ij}^{(n)}]_{ij=1}^n,$$

$$a_{ij}^{(n)} = g_n \int_\Omega \sum_{k,l=1}^N F_{kl}\Big(\sum_{p=1}^n \alpha_p^{(n)} g_n \varphi_p^{(n)}, x, t\Big) \frac{\partial \varphi_i^{(n)}}{\partial x_k} \frac{\partial \varphi_j^{(n)}}{\partial x_l}\, dx,$$

$$M^{(n)} = [m_{ij}^{(n)}]_{ij=1}^n,$$

$$m_{ij}^{(n)} = (\varphi_i^{(n)}, \varphi_j^{(n)}); \quad \tilde{M}^{(n)} = g_n^2 M^{(n)}, \tag{2.16}$$

$$\tilde{F}^{(n)}(\alpha^{(n)}, t) = [\tilde{F}_i^{(n)}(\alpha^{(n)}, t)]_{i=1}^n,$$

$$\tilde{F}_i^{(n)}(\alpha^{(n)}, t) = \int_\Omega F_0\Big(\sum_{p=1}^n g_n \alpha_p^{(n)} \varphi_p^{(n)}, x, t\Big)\varphi_i^{(n)}\, dx,$$

$$\alpha_0^{(n)} = [\alpha_{0,i}^{(n)}]_{i=1}^n; \quad \alpha_{0,i}^{(n)} = u_0(x_i).$$

Vegyük észre, hogy S2 következtében az A_0 tetszőlegesen rögzített $(p, t_0)\in R^n \times R^+$ érték esetén pozitív definit mátrix, továbbá, az $M^{(n)}$ mátrix (amely a $\varphi_i^{(n)}$ bázisfüggvények *Gram-mátrixa*) szintén szimmetrikus, pozitív definit. Mivel F_0 a simaságra tett feltétel következtében lokálisan lipschitzes, ezért megfelelően nagy $n\in N^+$ megválasztása esetén a (2.14)—(2.15) feladatnak létezik egyetlen megoldása $[0, T]$-n [1]. Ugyanakkor megmutatjuk, hogy ez a feltétel elhagyható! Szorozzuk meg a (2.14) egyenletet skalárisan $\alpha^{(n)}(t)$-vel és rögzített $t\in[0, T]$ mellett integráljuk Ω-n! Ekkor

$$\Big(\tilde{M}^{(n)} \frac{d\alpha^{(n)}}{dt}, \alpha^{(n)}\Big) + \big(A_0^{(n)}(\alpha^{(n)}, t)\alpha^{(n)}, \alpha^{(n)}\big) + \big(\tilde{F}^{(n)}(\alpha^{(n)}, t), \alpha^{(n)}\big) = 0, \tag{2.17}$$

minden $t\in(0, T]$-re. Vegyük észre, hogy $A_0^{(n)}$ pozitív definitsége, valamint az S2 tulajdonságok alapján tetszőleges $n\in N^+$ esetén érvényesek az alábbi becslések:

$$\big(A_0^{(n)}(\alpha^{(n)}, t)\alpha^{(n)}, \alpha^{(n)}\big) \geqq 0; \tag{2.18}$$

$$\big(\tilde{F}^{(n)}(\alpha^{(n)}, t), \alpha^{(n)}\big) = \sum_{i=1}^n \Big(\int_\Omega F_0\Big(\sum_{p=1}^n g_n \alpha_p^n \varphi_p^{(n)}, x, t\Big)\varphi_i^{(n)}\, dx\Big)\alpha_i^{(n)} =$$

$$= \int_\Omega F_0\Big(\sum_{p=1}^n g_n \alpha_p^n \varphi_p^{(n)}, x, t\Big)\Big(\sum_{i=1}^n \alpha_i^{(n)} \varphi_i^{(n)}\Big)\, dx = \frac{1}{g_n} \int_\Omega F_0(u_n, x, t) u_n\, dx \geqq 0. \tag{2.19}$$

Mivel

$$\left(\tilde{M}^{(n)}\frac{d\alpha^{(n)}}{dt},\ \alpha^{(n)}\right)=\frac{1}{2}\frac{d}{dt}(\tilde{M}^{(n)}\alpha^{(n)},\ \alpha^{(n)}), \tag{2.20}$$

így a (2.18), (2.19) és (2.20) összefüggések alapján a (2.17) egyenlet az

$$\frac{d}{dt}(\tilde{M}^{(n)}\alpha^{(n)},\ \alpha^{(n)})\leqq 0$$

egyenlőtlenséget eredményezi, amelyből a (2.16) jelölés és a $g_n>0$ feltétel a

$$\frac{d}{dt}(M^{(n)}\alpha^{(n)},\ \alpha^{(n)})\leqq 0$$

becslést adja. Ezt integrálva a $[0,\tau]$ $(0<\tau\leqq T)$ intervallumon az

$$(M^{(n)}\alpha^{(n)}(\tau),\ \alpha^{(n)}(\tau))\leqq(M^{(n)}\alpha^{(n)}(0),\ \alpha^{(n)}(0)) \tag{2.22}$$

egyenlőtlenséget nyerjük.

Mivel $(\varphi_i^{(n)}(x))$ az Ω tartomány egyenletes felosztásával nyert VEM bázisfüggvény-rendszer (az ilyen mindig megadható [15]), ezért ez erősen lineárisan független (majdnem ortogonális) rendszert alkot [6] [10]. Ez azt jelenti, hogy tetszőleges $n=1,2,\ldots$ esetén az $M^{(n)}$ mátrix $\lambda_i^{(n)}$ $(i=1\ldots n)$ sajátértékeire érvényes az

$$0<\lambda_0\leqq\lambda_i^{(n)}\leqq\Lambda_0 \tag{2.23}$$

egyenlőtlenség, ahol a λ_0 és a Λ_0 az n-től független állandók. Így a tetszőleges $c_n\in R^n$ vektor esetén:

$$\lambda_0\|c_n\|_E^2\leqq(M^{(n)}c_n,\ c_n)\leqq\Lambda_0\|c_n\|_E^2. \tag{2.24}$$

Ez a (2.22) egyenlőtlenség a (2.15) összefüggés felhasználásával az

$$\|\alpha^{(n)}(t)\|_E^2\leqq\frac{\Lambda_0}{\lambda_0}\|\alpha_0^{(n)}\|_E^2,\quad\forall t\in[0,T] \tag{2.25}$$

egyenlőtlenséget szolgáltatja. Az $\alpha^{(0)}$ vektor (2.16) alakú előállításának következtében:

$$\|\alpha_0^{(n)}\|_E^2=\sum_{i=1}^{n}\alpha_{0,i}^2=\sum_{i=1}^{n}u_0^2(x_i)\leqq\|u_0\|^2. \tag{2.26}$$

Így (2.25) és (2.26) összevetéséből

$$\|\alpha^{(n)}(t)\|_E^2\leqq C_0,\quad\forall t\in[0,T]; \tag{2.27}$$

ahol

$$C_0=\frac{\Lambda_0}{\lambda_0}\|u_0\|^2.$$

Nyilvánvaló, hogy (2.27) alapján érvényes az

$$|\alpha_i^{(n)}(t)|\leqq C_0;\quad i=1,2,\ldots,n,\quad t\in[0,T] \tag{2.28}$$

becslés az $\alpha^{(n)}$ vektor komponenseire is. Így

$$|u_n(x,t)| = \Big|\sum_{i=1}^{n} \alpha_i^{(n)}(t) g_n \varphi_i^{(n)}(x)\Big| \leqq C_0 \sum_{i=1}^{n} |g_n \varphi_i^{(n)}(x)|. \tag{2.29}$$

A továbbiakban egyrészt felhasználjuk, hogy a $g_n\varphi_i^{(n)}(x)$ bázisfüggvények n-től függetlenül korlátosak, azaz

$$\max_{\substack{x\in\Omega \\ i=1,\ldots,n}} |g_n \varphi_i^{(n)}(x)| \leqq C_1; \tag{2.30}$$

másrészt, hogy tetszőleges $x\in\Omega$ pont csak véges számú supp $\varphi_i^{(n)}$-hez tartozik, (mivel a VEM bázisfüggvények kompakt tartójúak).

Jelölje C_2 ez utóbbi elemek számát. (Ismételten megjegyzzük, hogy C_1 és C_2 csak a közelítés típusától függenek, n-től függetlenek!)

Ekkor a (2.29) alapján

$$|u_n(x,t)| \leqq C_3, \tag{2.31}$$

ahol

$$C_3 = C_0 \cdot C_1 \cdot C_2 \tag{2.32}$$

n-től független, csak a közelítés típusától és az u_0-tól függő állandó. A (2.31) egyenlőtlenségből közvetlenül adódik a

$$\sup_{0\leqq t\leqq T} \|u_n(.,t)\| \leqq C_4 \tag{2.33}$$

becslés, ahol

$$C_4 = C_3(\text{mes}\,\Omega)^{1/2}. \tag{2.34}$$

A (2.33) az $u_n(x,t)$ létezésének n-től való függetlenségét [14], míg (2.31) és (2.33) az egyenletes korlátosságot jelentik.

2.3. ÁLLÍTÁS. Tetszőleges rögzített $t\in[0,T]$ esetén minden $n\in N^+$ mellett a $h_t^{(n)}(x)=F_0(u_n(x,t),x,t)$ függvény $L_2(\Omega)$-beli.

Bizonyítás. Az S2 feltételek alapján F_0 folytonosan differenciálható. így a (2.31) becslés figyelembevételével

$$|h_t^{(n)}(x)| \leqq \max_{\substack{|s|\leqq C_3 \\ (s,t)\in Q_T}} |F_0(s,x,t)| \leqq C_5, \tag{2.35}$$

ami a

$$\|h_t^{(n)}\| \leqq C_6 \tag{2.36}$$

korlátot jelenti, ahol

$$C_6 = C_5(\text{mes}\,\Omega)^{1/2}. \tag{2.37}$$

2.4. ÁLLÍTÁS. Tegyük fel, hogy a (2.1)—(2.3) feladatra teljesül az S2 feltételrendszer. Ekkor létezik olyan L_1 n-től független pozitív állandó, hogy tetszőleges $\varphi\in L_2(\Omega)$ és rögzített $t\in[0,T]$ esetén érvényesek az

$$\big|(F_0(u,.,t)-F_0(u_n,.,t),\ \varphi)\big| \leqq L_1\|u-u_n\|\cdot\|\varphi\|. \tag{2.38}$$

$$\big|(F_{ij}(u,.,t)-F_{ij}(u_n,.,t),\ \varphi)\big| \leqq L_1\|u-u_n\|\cdot\|\varphi\|, \quad i,j=1,\ldots,N \tag{2.39}$$

egyenlőtlenségek, ahol $u(x, t)$ a (2.1)—(2.3) feladat megoldása, $u_n(x, t)$ pedig a (2.11)—(2.12) feladatokkal definiált *Galerkin-féle szemidiszkrét közelítések sorozata.*

Bizonyítás. Jelölje M_c az $L_2(\Omega)$ tér következő korlátos alterét:

$$M_c = \{v | v \in L_2(\Omega);\ \|v\| < C;\ F_0(v, x, t) \in L_2(\Omega),\ C > 0,\ t \in [0, T] \text{ rögzített}\}. \tag{2.40}$$

Ekkor az

$$\Phi_t(v) = F_0(v, x, t) \tag{2.41}$$

$M_c \to L_2(\Omega)$ operátor az S2 feltétel következtében folytonosan differenciálható az M_c korlátos halmazon. Így ezen a *Lipschitz-feltételt* is kielégíti [16], azaz létezik olyan $L_1 > 0$ állandó, hogy

$$\|\Phi_t(v) - \Phi_t(w)\| \leqq L_1 \|v - w\|; \quad \forall v, w \in M_c. \tag{2.42}$$

Így a (2.41) jelölést figyelembe véve tetszőleges rögzített $t \in [0, T]$ mellett érvényes az

$$\|F_0(v, ., t) - F_0(w, ., t)\| \leqq L_1 \|v - w\| \tag{2.43}$$

egyenlőtlenség. A *Cauchy—Schwarz egyenlőtlenség* alkalmazásával tetszőleges $\varphi \in L_2(\Omega)$ esetén ekkor fennáll az

$$\left|(F_0(v, ., t) - F_0(w, ., t),\ \varphi)\right| \leqq L_1 \|v - w\| \cdot \|\varphi\| \tag{2.44}$$

becslés. Jelölje C_6 a következő pozitív, n-től független számot:

$$C_6 = \left\{ \sup_{0 \leqq t \leqq T} \|u(., t)\|;\ C_4 \right\}. \tag{2.45}$$

Jelölje továbbá tetszőleges rögzített $t \in [0, T]$ mellett $v = u(x, t)$ és $w = u_n(x, t)$. Nyilvánvaló, hogy $u(x, t)$ és a 2.3. állítás alapján tetszőleges n-re $u_n(x, t)$ is M_{c_6}-beli, így (2.38) közvetlenül adódik. A (2.39) egyenlőtlenség belátása ezzel teljesen megegyezik.

2.1. Megjegyzés. Ha az S1 feltételrendszer érvényes (azaz a nemlinearitás globálisan lipschitzes), akkor a (2.38) és a (2.39) becslések közvetlenül nyerhetők. Ugyanis, mivel (2.5) érvényes tetszőleges $s_1, s_2 \in R$ esetén, ezért rögzített $(x, t) \in Q_T$ mellett $s_1 = u(x, t)$ és $s_2 = u_n(x, t)$-re is igaz. Így

$$\begin{aligned} \|g(u, ., t) - g(u_n, ., t)\| &= \left(\int_\Omega (g(u, x, t) - g(u_n, x, t))^2 \, dx \right)^{1/2} \leqq \\ &\leqq \left(L_0^2 \int_\Omega (u - u_n)^2 \, dx \right)^{1/2} = L_0 \|u - u_n\|, \end{aligned} \tag{2.46}$$

ami figyelembe véve a $g = F_0$ és a $g = F_{ij}$ jelöléseket éppen a (2.43)-at jelenti. Innen (2.44) felhasználásával a kívánt becslések közvetlenül nyerhetők.

3. A közelítő sorozat konvergenciája

A továbbiakban az $u_n(x, t)$ *Galerkin-sorozat* konvergenciáját és annak nagyságrendjét vizsgáljuk meg. Bevezetve az $e_n(x, t)=u(x, t)-u_n(x, t)$ hibafüggvényt, könnyen ellenőrizhető, hogy tetszőleges $\tilde{u}_n \in E^{(n)}(t)$ esetén érvényes az

$$\frac{1}{2}\frac{d}{dt}\|e_n\|^2+a(t)(u_n;\, e_n, e_n) = [a(t)(u_n;\, u, \tilde{u}_n-u_n)-a(t)(u;\, u, \tilde{u}_n-u_n)]+$$
$$+[a(t)(u_n;\, e_n, u-\tilde{u}_n)]+\left(\frac{\partial e_n}{\partial t},\, u-\tilde{u}_n\right)+(F_0(u, ., t)-F_0(u_n, ., t),\, \tilde{u}_n-u_n), \tag{3.1}$$

azonosság. Adjunk becsléseket a (3.1)-ben szereplő egyes kifejezésekre!

A 2.4. állítás felhasználásával

$$|a(t)(u_n;\, u, \tilde{u}_n-u_n)-a(t)(u;\, u, \tilde{u}_n-u_n)| =$$
$$= \left|\int_\Omega \sum_{i,j=1}^{N} (F_{ij}(u_n, x, t)-F_{ij}(u, x, t))\frac{\partial u}{\partial x_i}\frac{\partial(\tilde{u}_n-u_n)}{\partial x_j}\,dx\right| \leqq$$
$$\leqq \max_{\substack{1\leqq i\leqq N\\(x,t)\in Q_T}}\left|\frac{\partial u}{\partial x_i}\right| L_1 \sum_{i,j=1}^{N}\|u-u_n\|\left\|\frac{\partial(\tilde{u}_n-u_n)}{\partial x_j}\right\| \leqq \tag{3.2}$$
$$\leqq \max_{\substack{1\leqq i\leqq N\\(x,t)\in Q_T}}\left|\frac{\partial u}{\partial x_i}\right| L_1\|e_n\| N \sum_{j=1}^{N}\left\|\frac{\partial(\tilde{u}_n-u_n)}{\partial x_j}\right\| \leqq K_1\|e_n\|\,\|\tilde{u}_n-u_n\|_1,$$

ahol

$$K_1 = \max_{\substack{1\leqq i\leqq N\\(x,t)\in Q_T}}\left|\frac{\partial u}{\partial x_i}\right| N^2 L_1, \tag{3.3}$$

$$\|v\|_1^2 = \sum_{i=1}^{N}\left\|\frac{\partial v}{\partial x_i}\right\|^2, \quad \forall v\in H_0^1(\Omega). \tag{3.4}$$

(Vegyük észre, hogy (3.4) következtében $\|v\|_1 \geqq \left\|\frac{\partial v}{\partial x_i}\right\|$ tetszőleges $i=1, 2, \ldots, N$-re és az ebből következő

$$N\|v\|_1 \geqq \sum_{i=1}^{N}\left\|\frac{\partial v}{\partial x_i}\right\|$$

egyenlőtlenséget alkalmaztuk (3.2)-ben.)

Felhasználva a jól ismert háromszög-egyenlőtlenséget, (3.2) alapján:

$$|a(t)(u_n;\, u, \tilde{u}_n-u_n)-a(t)(u;\, u, \tilde{u}_n-u_n)| \leqq K_1\|e_n\|[\|u-\tilde{u}_n\|_1+\|e_n\|_1]. \tag{3.5}$$

Vegyük észre, ha valamely $A=[a_{ij}]$ $N\times N$ méretű mátrix pozitív definit, akkor tetszőleges $\xi, \eta\in R^N$ vektor esetén

$$\sum_{i,j=1}^{N} a_{ij}\xi_i\eta_j \leqq \Big(\sum_{i,j=1}^{N} a_{ij}\xi_i\xi_j\Big)^{1/2}\Big(\sum_{i,j=1}^{N} a_{ij}\eta_i\eta_j\Big)^{1/2}, \tag{3.6}$$

mivel az $(A\xi, \eta)$ bilineáris alak egy skaláris szorzatot definiál R^N-ben és (3.6) az ezen skaláris szorzatra vonatkozó Cauchy—Schwarz egyenlőtlenség. Ezt figyelembe véve:

$$|a(t)(u_n;\ e_n, u-\tilde{u}_n)| = \left| \int_\Omega \sum_{i,j=1}^N F_{ij}(u_n, x, t) \frac{\partial e_n}{\partial x_i} \frac{\partial (u-\tilde{u}_n)}{\partial x_j}\, dx \right| \leqq$$

$$\leqq \int_\Omega \left(\sum_{i,j=1}^N F_{ij}(u_n, x, t) \frac{\partial e_n}{\partial x_i} \frac{\partial e_n}{\partial x_j} \right)^{1/2} \cdot$$

$$\cdot \left(\sum_{i,j=1}^N F_{ij}(u_n, x, t) \frac{\partial (u-\tilde{u}_n)}{\partial x_i} \frac{\partial (u-\tilde{u}_n)}{\partial x_j} \right)^{1/2} dx \leqq \tag{3.7}$$

$$\leqq F^* \int_\Omega \left[\sum_{i,j=1}^N \left(\frac{\partial e_n}{\partial x_i} \right)^2 \right]^{1/2} \left[\sum_{i,j=1}^N \left(\frac{\partial (u-\tilde{u}_n)}{\partial x_i} \right)^2 \right]^{1/2} dx \leqq$$

$$\leqq F^* \int_\Omega \left[\varepsilon_1 \sum_{i=1}^N \left(\frac{\partial e_n}{\partial x_i} \right)^2 + \frac{1}{4\varepsilon_1} \sum_{i=1}^N \left(\frac{\partial (u-\tilde{u}_n)}{\partial x_i} \right)^2 \right] dx = F^* \varepsilon_1 \|e_n\|_1^2 + \frac{F^*}{4\varepsilon_1} \|u-\tilde{u}_n\|_1^2,$$

ahol felhasználtuk F_{ij} S2 tulajdonságát és az

$$ab \leqq \varepsilon a^2 + \frac{1}{4\varepsilon} b^2 \quad \forall \varepsilon > 0, \quad \forall a, b \in R \tag{3.8}$$

ún. „ε-egyenlőtlenséget".
Alkalmazva (3.8) mellett az

$$(a+b)^2 \leqq 2(a^2+b^2), \quad \forall a, b \in R \tag{3.9}$$

elemi egyenlőtlenséget, a (3.5) és (3.7) összefüggésekből közvetlenül adódik, hogy

$$|a(t)(u_n;\ u, \tilde{u}_n-u_n) - a(t)(u;\ u, \tilde{u}_n-u_n)| + |a(t)(u_n;\ e_n, u-\tilde{u}_n)| \leqq$$
$$\leqq K_2 \|e_n\|_1^2 + K_3 (\|u-\tilde{u}_n\|_1^2 + \|e_n\|^2), \tag{3.10}$$

ahol

$$K_2 = \frac{K_1}{2\varepsilon_2} + F^* \varepsilon_1; \quad K_3 = \max \left\{ \frac{K_1}{2\varepsilon_2} + \frac{F^*}{4\varepsilon_1},\ \varepsilon_2 K_1 \right\} \tag{3.11}$$

és $\varepsilon_1, \varepsilon_2$ tetszőleges pozitív számok.

A (2.38) egyenlőtlenség, (3.8), a háromszög-egyenlőtlenség és a *Friedrichs-egyenlőtlenség* alkalmazásával közvetlen számolással belátható, hogy

$$\left| (F_0(u, .., t) - F_0(u_n, .., t),\ \tilde{u}_n - u_n) \right| \leqq K_4 \|e_n\|^2 + K_5 \|\tilde{u}_n - u\|_1^2, \tag{3.12}$$

ahol

$$K_4 = L_1 \varepsilon_3; \quad K_5 = L_1 K_\Omega / 2\varepsilon_3 \tag{3.13}$$

és K_Ω a *Friedrichs-egyenlőtlenségben* szereplő, az $\Omega \subset R^N$ tartománytól függő állandó.

Az S2 tulajdonságok alapján

(3.14)

$$a(t)(u_n;\ e_n, e_n) = \int_\Omega \sum_{i,j=1}^N F_{ij}(u_n, x, t) \frac{\partial e_n}{\partial x_i} \frac{\partial e_n}{\partial x_j}\, dx \geqq F_* \int_\Omega \sum_{i=1}^N \left(\frac{\partial e_n}{\partial x_i} \right)^2 dx = F_* \|e_n\|_1^2.$$

A

$$K_6 = 2(F_* - K_2); \quad K_7 = 2 \max \{K_3 + K_5,\ K_3 + K_4\} \tag{3.15}$$

jelölésekkel az egyes becslések alapján (3.1)-ből következik az

$$\frac{d}{dt}\|e_n\|^2+K_6\|e_n\|_1^2 \leqq K_7[\|u-\tilde{u}_n\|_1^2+\|e_n\|^2]+2\left|\left(\frac{\partial e_n}{\partial t},\ u-\tilde{u}_n\right)\right| \tag{3.16}$$

egyenlőtlenség.

Válasszuk meg a tetszőleges $\varepsilon_1, \varepsilon_2$ pozitív állandókat úgy, hogy K_6 pozitív legyen. (Ilyen megválasztás nyilvánvalóan mindig lehetséges.) Ekkor a (3.16) egyenlőtlenséget megszorozva $e^{-K_7 t}$-vel, figyelembe véve a

$$\frac{d}{dt}(e^{-K_7 t}\|e_n\|^2) = e^{-K_7 t}\frac{d}{dt}\|e_n\|^2 - K_7 e^{-K_7 t}\|e_n\|^2 \tag{3.17}$$

összefüggést és integrálva egy tetszőleges $(0, \tau)$ $(0<\tau\leqq T)$ intervallumon a következő egyenlőtlenséghez jutunk:

$$\begin{aligned} &e^{-K_7\tau}\|e_n\|^2(\tau)-\|e_n\|^2(0)+K_6\int_0^\tau e^{-K_7 t}\|e_n\|_1^2\,dt \leqq \\ &\leqq K_7\int_0^\tau e^{-K_7 t}\|u-\tilde{u}_n\|_1^2\,dt+2\left|\int_0^\tau e^{-K_7 t}\left(\frac{\partial e_n}{\partial t},\ u-\tilde{u}_n\right)dt\right|. \end{aligned} \tag{3.18}$$

A továbbiakban (3.18) jobb oldalának utolsó tagjára adunk meg becslést. Parciális integrálással közvetlenül adódik, hogy

$$\int_0^\tau e^{-K_7 t}\left(\frac{\partial e_n}{\partial t},\ u-\tilde{u}_n\right)dt = B_1-B_2, \tag{3.19}$$

ahol

$$\begin{aligned} B_1 &= [e^{-K_7 t}(e_n,\ u-\tilde{u}_n)]_{t=0}^{\tau}, \\ B_2 &= \int_0^\tau\left[e^{-K_7 t}\left(\frac{\partial(u-\tilde{u}_n)}{\partial t},\ e_n\right)-K_7 e^{-K_7 t}(u-\tilde{u}_n,\ e_n)\right]dt. \end{aligned} \tag{3.20}$$

Figyelembe véve K_7 pozitivitását, elemi egyenlőtlenségeink felhasználásával közvetlenül megmutatható, hogy

$$B_1 \leqq \left[\varepsilon_4\|e_n\|^2+\frac{1}{4\varepsilon_4}\|u-\tilde{u}_n\|^2\right]_{t=0}^{\tau}. \tag{3.21}$$

Vezessük be a következő jelöléseket:

$$\begin{aligned} &Q_\tau = \Omega\times(0,\tau), \\ &\|v\|_{L_2(Q_\tau)}^2 = \left(\int_0^\tau\|v(.,t)\|^2\,dt\right);\quad v\in L_2(0,T,L_2(\Omega)). \end{aligned} \tag{3.22}$$

Nyilvánvaló, hogy

$$\|v\|_{L_2(Q_T)}^2 \geqq \|v\|_{L_2(Q_\tau)}^2. \tag{3.23}$$

A *Cauchy—Schwarz egyenlőtlenséget* alkalmazva és a (3.23) egyenlőtlenséget figyelembe véve

$$B_2 \leqq \varepsilon_5 \left\| \frac{\partial(u-\tilde{u}_n)}{\partial t} \right\|^2_{L_2(Q_T)} + \frac{1}{4\varepsilon_5} \|e_n\|^2_{L_2(Q_\tau)} + K_7\varepsilon_6 \|u-\tilde{u}_n\|^2_{L_2(Q_\tau)} + \frac{K_7}{4\varepsilon_6} \|e_n\|^2_{L_2(Q_\tau)}. \tag{3.24}$$

Becslésünket összegezve

$$\int_0^\tau e^{-K_7 t} \left(\frac{\partial e_n}{\partial t}, u-\tilde{u}_n \right) dt \leqq K_8 [\|e_n\|^2(\tau) + \|e_n\|^2_{L_2(Q_\tau)}] + \tag{3.25}$$

$$+ K_9 \left[\|u-\tilde{u}_n\|^2_{L_2(Q_T)} + \left\| \frac{\partial(u-\tilde{u}_n)}{\partial t} \right\|^2_{L_2(Q_T)} + \|u-\tilde{u}_n\|^2(\tau) \right],$$

ahol

$$K_8 = \max\{\varepsilon_4;\ 1/4\varepsilon_5 + K_7/4\varepsilon_6\},$$
$$K_9 = \max\{K_7\varepsilon_6;\ \varepsilon_5;\ 1/4\varepsilon_5\}. \tag{3.26}$$

Vezessük be a következő normákat:

$$\|v\|^2_{H_0^1 \times L_2(0,\tau)} = \int_0^\tau \|v\|^2_1 \, dt,$$
$$\|v\|_{L_2 \times C(0,T)} = \sup_{0 \leqq t \leqq T} \|v(.,t)\|. \tag{3.27}$$

Figyelembe véve az $e^{-K_7 T} \leqq e^{-K_7 \tau} < 1$ összefüggést, valamint a *Friedrichs-egyenlőtlenségből* és a (3.27) definícióból következő

$$\|v\|^2_{L_2(Q_\tau)} \leqq K_\Omega^2 \|v\|^2_{H_0^1 \times L_2(0,\tau)}, \quad \tau \in (0, T] \tag{3.28}$$

egyenlőtlenséget, (3.25) alapján (3.18) felírható

$$\|e_n\|^2(\tau)[e^{-K_7 T} - 2K_8] + \|e_n\|^2_{H_0^1 \times L_2(0,\tau)} [K_6 e^{-K_7 T} - 2K_8 K_\Omega] \leqq \tag{3.29}$$

$$\leqq 2K_9 \left\| \frac{\partial(u-\tilde{u}_n)}{\partial t} \right\|^2_{L_2(Q_T)} + 2K_9 \|u-\tilde{u}_n\|^2(\tau) + \|e_n\|^2(0)$$

alakban.

Vegyük észre, hogy (2.9) és (2.12) következtében

$$(u(.,0) - u_n(.,0),\ \varphi_i^{(n)}) = 0; \quad i = 1, 2, \ldots, n. \tag{3.30}$$

Így $e_n(0)$ ortogonális $H^{(n)}$-re, azaz tetszőleges $\tilde{u}_n$ esetén

$$\|e_n\|^2(0) \leqq \|u-\tilde{u}_n\|^2(0). \tag{3.31}$$

Válasszuk meg a tetszőleges ε_4, ε_5 és ε_6 pozitív állandókat úgy, hogy teljesüljenek a

$$K_{10} = e^{-K_7 T} - 2K_8 > 0,$$
$$K_{11} = K_6 e^{-K_7 T} - 2K_8 K_\Omega > 0 \tag{3.32}$$

egyenlőtlenségek. Ekkor (3.29) a

$$K_{10}\|e_n\|^2(\tau)+K_{11}\|e_n\|^2_{H_0^1\times L_2(0,\tau)} \leqq$$

$$\leqq K_{12}\left[\|u-\tilde{u}_n\|^2_{H_0^1\times L_2(0,\tau)}+\left\|\frac{\partial(u-\tilde{u}_n)}{\partial t}\right\|^2_{L_2(Q_T)}+\|u-\tilde{u}_n\|^2_{L_2\times C(0,T)}\right] \tag{3.33}$$

alakot ölti, ahol

$$K_{12}=\max\{K_7+2K_\Omega^2K_9,\ 2K_9+1\}. \tag{3.34}$$

Mivel tetszőleges $\tau\in(0,T]$ mellett

$$\|v\|_{H_0^1\times L_2(0,\tau)}\leqq\|v\|_{H_0^1\times L_2(0,T)}, \tag{3.35}$$

a

$$K_{13}=K_{11}/K_{12};\quad K_{14}=K_{12}/K_{10};\quad K_{15}=2K_{14} \tag{3.36}$$

jelölésekkel (3.33) alapján érvényes az

$$\|e_n\|^2_{L_2\times C(0,T)}+K_{13}\|e_n\|^2_{H_0^1\times L_2(0,T)}\leqq$$

$$\leqq K_{15}\left[\|u-\tilde{u}_n\|^2_{H_0^1\times L_2(0,T)}+\left\|\frac{\partial(u-\tilde{u}_n)}{\partial t}\right\|^2_{L_2(Q_T)}+\|u+\tilde{u}_n\|^2_{L_2\times C(0,T)}\right] \tag{3.37}$$

egyenlőtlenség.

Ezzel bebizonyítottuk a következő tételünket.

3.1. ÁLLÍTÁS. Legyen $u(x,t)$ a (2.1)—(2.3) feladat megoldása, $u_n(x,t)$ a (2.11)—(2.12) feladatok megoldásával nyert szemidiszkrét *Galerkin-féle sorozat*, $\tilde{u}_n(x,t)\in C(0,T;H^{(n)})$ tetszőleges függvény. Ekkor az $e_n(x,t)=u(x,t)-u_n(x,t)$ hibafüggvényre érvényes a (3.37) becslés, ahol K_{13} és K_{15} az n-től független pozitív állandók.

3.1. Megjegyzés. A 2.1. megjegyzés alapján nyilvánvaló, hogy az S1 feltételrendszer esetén (azaz, ha F_0 és F_{ij} az első változójukban globálisan lipschitzesek), a (3.37) hibabecslés ugyanúgy érvényes marad.

3.2. ÁLLÍTÁS. Az S1 és S2 feltételrendszer esetén a *Galerkin-sorozat* konvergens az $L_2\times C(0,T)$ és a $H_0^1\times L_2(0,T)$ normákban egyaránt.

Bizonyítás. Vegyük észre, hogy (3.27) jobb oldala az $u(x,t)$ függvény $C(0,T;H^{(n)})$-beli approximációjától függ. Mivel Ω reguláris, így a spline-függvények $H_0^1(\Omega)$ térben való teljességéből következik az állítás.

A konvergencia nagyságrendjét is az approximáció nagyságrendje határozza meg. Vizsgáljuk meg ezt az $\Omega=(a,b)\times(c,d)\subset R^2$ tartományon. Jelölje Δ az Ω tartomány $(x_i,x_{i+1})\times(y_j,y_{j+1})$ elemi téglalapokra való felbontását, $H^{(m)}(\Delta,\Omega)$ azon Ω-n értelmezett $w(x,y)$ függvényeket, amelyekre $\frac{\partial^{i+j}w}{\partial x^i\partial y^j}\in C^{0,0}(\Omega)$, $(0\leqq i,j\leqq m-1)$ és $w(x,y)$ mindegyik elemi téglalapon valamely $2m-1$-ed fokú polinom. Legyen f olyan függvény, hogy $\frac{\partial^{i+j}f}{\partial x^i\partial y^j}\in C^{0,0}(\Omega)$, ha $i+j<2m$ és $\frac{\partial^{i+j}f}{\partial x^i\partial y^j}\in L_2(\Omega)$, ha $i+j=2m$.

Ekkor létezik olyan $f_h \in H^{(m)}(\Delta, \Omega)$ hogy

$$\left\| \frac{\partial^{i+j}}{\partial x^i \partial y^j} (f - f_h) \right\| \leqq C_f h^{2m-(i+j)}, \tag{3.38}$$

ahol $i+j \leqq 2m-1$, $0 \leqq i, j \leqq m$ és C_f valamely $\| f \|_{2m}$-től függő állandó [2]. Ezt felhasználva (3.37) alapján belátható [4], hogy

$$\|e_n\|^2_{L_2 \times C(0,T)} + \|e_n\|^2_{H_0^1 \times L_2(0,T)} \leqq Ch^{2(2m-1)}. \tag{3.39}$$

Megjegyezzük, hogy a (3.39) hibabecslés elsőfokú közelítés esetén megegyezik a lineáris feladatokra nyert becsléssel [7].
Ha a (3.39) becslésből közvetlenül nyerjük az

$$\sup_{0 \leqq t \leqq T} \|e_n\|(t) \leqq Ch^r \tag{3.40}$$

becslést (ahol r a közelítés fokszáma), akkor ezen normában az elsőfokú közelítés esetén a lineáris feladatokhoz képest $h^{1/2}$ nagyságrenddel romlik a becslés [7].

IRODALOM

[1] BAKER, G. A., DOUGALIS, V. A. and KARAKASIAN, O., "On multistep Galerkin discretizations of semilinear hyperbolic and parabolic equations", *Nonlin. Anal.* **1** (1980) 579—597.
[2] BIRKHOFF, G., SCHULTZ, M. and VARGA, R., "Piecewise Hermite interpolation in one and two variables", *Num. Math.* **11** (1968) 232—256.
[3] DENEL, J. and HESS, P., "Nonlinear parabolic boundary value problems with upper and lower solutions", *Isr. J. of Math.* **29** (1978) 92—104.
[4] DOUGLAS, J. and DUPONT, T., "Galerkin methods for parabolic equations", *SIAM J. Num. Anal.* **7** (1970) 575—626.
[5] DOUGLAS, J., DUPONT, T. and WHEELER, M., "A quasi-projection analysis of Galerkin methods for parabolic and hyperbolic equations", *Math. Comp.* **23** (1978) 345—362.
[6] FARAGÓ, I., „Véges elemek módszere elliptikus típusú feladatok megoldására", *Alk. Mat. Lapok*, **8** (1982) 399—425.
[7] FARAGÓ, I., „Véges elemek módszere lineáris, parabolikus típusú feladatok megoldására", *Alk. Mat. Lapok* **11** (1985) 123—155.
[8] FARAGÓ, I.. "Convergence of semidiscrete Galerkin approximation for parabolic equations with signdeterminated nonlinearity", Num. Math. Coll. Soc. János Bolyai, Nort-Holland, **50** (1987) 371—380.
[9] FARAGÓ, I. and GALÁNTAI, A. "An A-stable three-level method for the Galerkin solution of quasilinear parabolic problems", *Annal. Un. Sc. Budapest de Roland Eötvös, Nom. sec. Computorica*, 1988. (megjelenés alatt)
[10] OMODEI, B. J., "On the numerical stability of the Rayleigh—Ritz method", *SIAM J. Num Anal.* **14** (1977) 1151—1171.
[11] RACHFORD, H. H., "Two level discrete-time Galerkin approximations for second order nonlinear parabolic partial differential equations", *SIAM J. Num. Anal.* **10** (1973) 1010—1026.
[12] WHEELER, M., "A priori L_2-error estimates for Galerkin approximations", *SIAM J. Num. Anal.* **10** (1973) 1723—1759.

(Beérkezett: 1987. december 1).

FARAGÓ ISTVÁN
AGRÁRTUDOMÁNYI EGYETEM MATEMATIKAI ÉS
SZÁMÍTÁSTECHNIKAI INTÉZET
2103 GÖDÖLLŐ

FINITE ELEMENT METHOD FOR SOLVING NONLINEAR PARABOLIC PROBLEMS

I. FARAGÓ

The FEM is considered for nonlinear parabolic equations. The usual global *Lipschitz-condition* of nonlinearity is replaced by a new one related to smoothness and positivity conditions. The convergence of method is proven, the rate of convergence is also estimated.

Alkalmazott Matematikai Lapok 13 (1987—88)

Alkalmazott Matematikai Lapok **13** (1987—88) 349—359

VÉGES ELEMEK MÓDSZERE NEMLINEÁRIS, PARABOLIKUS TÍPUSÚ RENDSZEREK MEGOLDÁSÁRA

FARAGÓ ISTVÁN

Gödöllő

Ebben a dolgozatban a [3] cikk eredményeit terjesztettük ki a nemlineáris, parabolikus típusú differenciálegyenlet-rendszerekre. Bebizonyítjuk a véges elemek módszerének konvergenciáját és a konvergencia nagyságrendjét.

1. Bevezetés

Ebben a dolgozatban a [3] dolgozat eredményeit terjesztjük ki a parabolikus típusú differenciálegyenlet-rendszerekre. A nemlinearitásra vonatkozó *globális Lipschitz-feltétel* vagy a monotonitási feltétel mellett a feladatok numerikus megoldását tárgyalja néhány munka [1], [4]. Dolgozatunkban ezen feltételtől eltérő előjelállandósági és simasági feltételekkel váltjuk ki ezeket. Megmutatjuk, hogy a *Galerkin-féle szemidiszkrét közelítő sorozat* konvergens és a [3] dolgozat eredményeivel megegyező hibabecslést is adunk. A 2. szakaszban felépítjük a *Galerkin-féle szemidiszkrét közelítő sorozatot* és néhány tulajdonságát bizonyítjuk. A 3. szakaszban ezen sorozat konvergenciáját és ennek nagyságrendjét határozzuk meg. Megjegyezzük, hogy a dolgozatban alkalmazott és a [2], [3] cikkekben használt jelölésektől eltérő jelöléseket a dolgozat végén felsoroljuk.

2. A feladat kitűzése és a szemidiszkrét közelítő sorozat néhány tulajdonsága

Tekintsük a továbbiakban a következő nemlineáris, másodrendű parabolikus típusú parciális differenciálegyenlet-rendszereket:

$$\frac{\partial u^k}{\partial t} - \sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\left(F_{ij}^k(u^1, u^2, \ldots, u^L, x, t)\frac{\partial u^k}{\partial x_j}\right) + \tag{2.1}$$

$$+ F_0^k(u^1, \ldots, u^L, x, t) = 0, \quad (x, t) \in Q_T, \quad k = 1, 2, \ldots, L.$$

$$u^k(x, t) = 0; \quad \forall (x, t) \in \Gamma \times (0, T]; \quad k = 1, 2, \ldots, L. \tag{2.2}$$

$$u^k(x, 0) = u_0^k(x); \quad x \in \bar{\Omega}, \; k = 1, 2, \ldots, L. \tag{2.3}$$

A megfelelő jelölések alkalmazásával a (2.1)—(2.3) feladat felírható a tömörebb

$$\frac{\partial \mathbf{u}}{\partial t} - a(\mathbf{u}) + \mathbf{F}_0(\mathbf{u}, x, t) = \mathbf{0}; \quad \forall (x, t) \in Q_T, \tag{2.4}$$

$$\mathbf{u}(x, t) = \mathbf{0}; \quad \forall (x, t) \in \Gamma \times (0, T], \tag{2.5}$$

$$\mathbf{u}(x, 0) = \mathbf{u}_0(x); \quad \forall x \in \bar{\Omega} \tag{2.6}$$

alakban. Tegyük fel, hogy a (2.4)—(2.6) feladatra teljesül az alábbi SS2 feltételrendszer:

1. a) $F_{ij}^k \in C^1(Q_T^L)$; $i, j = 1, \ldots, N$; $k = 1, 2, \ldots, L$.

   b) $\mathbf{F}$ mátrix szimmetrikus;

   c) léteznek olyan F_*, F^* pozitív állandók ($F_* \leqq F^*$), hogy tetszőleges $\mathbf{p} \in R^{LN}$ és $(s, x, t) \in Q_T^L$ esetén

$$F_* |||\mathbf{p}|||_E^2 \leqq [[\mathbf{F}(\mathbf{s}, x, t)\mathbf{p}, \mathbf{p}]] \leqq F^* |||\mathbf{p}|||_E^2;$$

2. a) $F_0 \in C_L^1(Q_T^L)$;

   b) $F_0^k(\mathbf{s}, x, t) s_k \geqq 0$, $\forall (s, x, t) \in Q_T^L$, $k = 1, 2, \ldots, L$.

   c) $\mathbf{F}_0(\mathbf{0}, x, t) = \mathbf{0}$, $\forall (x, t) \in Q_T$;

3. $\mathbf{u}_0 \in \mathbf{C}_L(\bar{\Omega})$ és $\operatorname{supp} \mathbf{u}_0 \subset \Omega$.

A parciális differenciálegyenletek elméletéből ismeretes, hogy az SS2 feltételrendszer mellett a (2.4)—(2.6) feladatnak létezik egyetlen, klasszikus értelemben vett megoldása $\bar{Q}_T$-n [6].

A továbbiakban SS1 jelöli azt a feltételrendszert, amely SS2-től pontosan az F_0^k és F_{ij}^k-re vonatkozó nemlineáritási feltételben tér el: az 1.a, 2.a, és 2.b helyett a globális lipschitzességet tételezzük fel.

A (2.4)—(2.6) feladat általánosított (gyenge) alakja a következő: keressük azt az $\mathbf{u}(x, t) \in C((0, T]; H_{0,L}(\Omega))$ vektort, amelyre tetszőleges $\boldsymbol{\varphi} \in \mathbf{H}_{0,L}^1(\Omega)$ esetén

$$\left\langle \frac{\partial \mathbf{u}}{\partial t}, \boldsymbol{\varphi} \right\rangle + \mathbf{a}_L(t)(\mathbf{u}; \mathbf{u}, \boldsymbol{\varphi}) + \langle \mathbf{F}_0(\mathbf{u}, ., t), \boldsymbol{\varphi} \rangle = 0, \tag{2.7}$$

$$\langle \mathbf{u}(., 0), \boldsymbol{\varphi} \rangle = \langle \mathbf{u}_0, \boldsymbol{\varphi} \rangle. \tag{2.8}$$

A közelítéseket rögzített $t \in (0, T]$ mellett $\mathbf{H}_L^n$-ben keressük, ahol $\varphi^{(k,i)}$ ($i = 1, 2, \ldots, n$, $k = 1, 2, \ldots, L$) adott bázisfüggvények. (Ez azt jelenti, hogy az $\mathbf{u}(x, t)$ vektor komponensei különböző véges dimenziós alterekben is approximálhatók.) Ekkor az $\mathbf{u}_n \in \mathbf{H}_L$ approximáció kielégíti a

$$\left\langle \frac{\partial \mathbf{u}_n}{\partial t}, \boldsymbol{\varphi}_n \right\rangle + \mathbf{a}_L(t)(\mathbf{u}_n, \mathbf{u}_n, \boldsymbol{\varphi}_n) + \langle \mathbf{F}_0(\mathbf{u}_n, ., t), \boldsymbol{\varphi}_n \rangle = 0 \tag{2.9}$$

$$\langle \mathbf{u}_n(., 0), \boldsymbol{\varphi}_n \rangle = \langle \mathbf{u}_0, \boldsymbol{\varphi}_n \rangle \tag{2.10}$$

egyenleteket tetszőleges $\boldsymbol{\varphi}_n \in \mathbf{H}_L^n$ és rögzített $t \in [0, T]$ esetén.

Legyen a továbbiakban $(\varphi_i^{(k,n)})_{i=1,n} \subset H_0^1(\Omega)$ $(k=1, 2, \ldots, L)$ az Ω tartomány egyenletes felosztásával nyert valamely VEM bázisfüggvény-rendszer. Ekkor, mint ismeretes [7], az $\mathbf{u}_n(x, t)$ vektor $u_n^k(x, t) \in H_k^n$ komponensei

$$u_n^k(x, t) = \sum_{i=1}^{n} g_n \alpha_i^{(k,n)}(t) \varphi_i^{(k,n)}(x) \tag{2.11}$$

alakúak, ahol g_n az n-től függő pozitív állandó és feladatunk az

$$\boldsymbol{\alpha}^{(L,n)}(t) = \left(\alpha_i^{(k,n)}(t)\right)_{\substack{i=1,n \\ k=1,L}}$$

$L \cdot n$ dimenziós vektor meghatározása.
Behelyettesítve $u_n^k(x, t)$ alakját a (2.9)—2.10) *Galerkin egyenletbe*, $\boldsymbol{\alpha}^{(L,n)}(t)$-re a következő *Cauchy-feladatot* nyerjük:

$$\tilde{\mathbf{M}}_L^n \frac{d\boldsymbol{\alpha}^{(L,n)}}{dt} + \mathbf{A}_0^L(\boldsymbol{\alpha}^{(L,n)}, t)\boldsymbol{\alpha}^{(L,n)} + \mathbf{F}^L(\boldsymbol{\alpha}^{(L,n)}, t) = 0; \quad 0 < t \leqq T, \tag{2.12}$$

$$\boldsymbol{\alpha}^{(L,n)}(0) = \boldsymbol{\alpha}_0^{(L,n)}, \tag{2.13}$$

ahol

$$\mathbf{A}_0^L(\boldsymbol{\alpha}^{(L,n)}, t) = A_{l,j;k,i}(\boldsymbol{\alpha}^{(L,n)}, t);$$

$$A_{l,j;k,i}(\boldsymbol{\alpha}^{(L,n)}, t) = g_n \int_{\Omega} \sum_{m=1}^{L} \sum_{p,r=1}^{N} F_{pr}^m(u_n, x, t) \frac{\partial \varphi_j^{(l,n)}}{\partial x_p} \frac{\partial \varphi_i^{(k,n)}}{\partial x_r} dx;$$

$$\mathbf{M}_L^n = [M_{l,j;k,i}]_{\substack{l,k=1,L \\ j,i=1,n}};$$

$$M_{l,j;k,i} = \langle\langle \varphi_j^{(l,n)}, \varphi_i^{(k,n)} \rangle\rangle;$$

$$\tilde{M}_L^n = g_n^2 \mathbf{M}_L^n; \tag{2.14}$$

$$F^L(\boldsymbol{\alpha}^{(L,n)}, t) = \left(F_i^l(\boldsymbol{\alpha}^{(L,n)}, t)\right)_{\substack{i=1,n \\ l=1,L}};$$

$$F_i^l(\boldsymbol{\alpha}^{(L,n)}, t) = \int_{\Omega} F_0^l(\mathbf{u}_n, x, t) \varphi_i^{(l,n)}(x)\, dx;$$

$$\boldsymbol{\alpha}_0^{(L,n)} = (\alpha_{0,l,i}^{(L,n)})_{\substack{l=1,L \\ i=1,n}};$$

$$\alpha_{0,l,i}^{(L,n)} = u_0^l(x_i).$$

A (2.12)—(2.13) feladat megoldhatóságát és a megoldások korlátosságát bizonyítja a következő:

2.1. ÁLLÍTÁS. A (2.12)—(2.13) feladatnak tetszőleges $n=1, 2, \ldots$ esetén létezik egyetlen $\mathbf{u}_n(x, t)$ megoldása és ezen megoldások sorozata tetszőleges rögzített $t \in [0, T]$ mellett egyenletesen korlátos a maximum és az $\mathbf{L}_{2,L}(\Omega)$ normákban.

Bizonyítás. Vegyük észre, hogy az SS2 tulajdonságok következtében A_0^L tetszőleges rögzített $(\mathbf{s}, t) \in R^{Ln} \times R^+$ pontban pozitív definit mátrix; továbbá, hogy az $\mathbf{M}^n$

mátrix is pozitív definit. Skalárisan megszorozva a (2.12) egyenletet $\alpha^{(L,n)}(t)$-vel az

(2.15)
$$\left[\left[\tilde{\mathbf{M}}_L^n \frac{d\alpha^{(L,n)}}{dt},\ \alpha^{(L,n)}\right]\right] + \left[\left[\mathbf{A}_0^L(\alpha^{(L,n)}, t)\alpha^{(L,n)},\ \alpha^{(L,n)}\right]\right] + \left[\left[\mathbf{F}^L(\alpha^{(L,n)}, t),\ \alpha^{(L,n)}\right]\right] = 0$$

egyenlőséget nyerjük. Az SS2 tulajdonság következtében

(2.16)
$$\begin{aligned}\left[\left[\mathbf{F}^L(\alpha^{(L,n)}, t),\ \alpha^{(L,n)}\right]\right] &= \sum_{l=1}^{L}\sum_{i=1}^{n} F_i^l(\alpha^{(L,n)}, t)\alpha^{(l,i)}(t) = \\ &= \sum_{l=1}^{L}\sum_{i=1}^{n}\int_\Omega F_0^l(\mathbf{u}_n, x, t)\varphi_i^{(l,n)}(x)\,dx\,\alpha^{(l,i)}(t) = \\ &= \frac{1}{g_n}\sum_{l=1}^{L}\int_\Omega F_0^l(\mathbf{u}_n, x, t)\Big(\sum_{i=1}^{n} g_n\alpha^{(l,i)}(t)\varphi_i^{(l,n)}(x)\Big)\,dx = \\ &= \frac{1}{g_n}\sum_{l=1}^{L}\int_\Omega F_0^l(u_n, x, t)u_n^l(x, t)\,dx \geqq 0.\end{aligned}$$

Így (2.16) és az $\mathbf{A}_0(\alpha^{(L,n)}, t)$ mátrix pozitív definitsége következtében (2.15)-ből következik az

(2.17)
$$\left[\left[\tilde{\mathbf{M}}_L^n \frac{d\alpha^{(L,n)}}{dt},\ \alpha^{(L,n)}\right]\right] \leqq 0$$

egyenlőtlenség, ami ekvivalens tetszőleges $t\in[0, T]$ esetén a

(2.18)
$$\left[\left[\tilde{\mathbf{M}}_L\alpha^{(L,n)},\ \alpha^{(L,n)}\right]\right](t) \leqq 0$$

egyenlőtlenséggel. Ebből az $\tilde{\mathbf{M}}_L$ mátrix alakját és a $g_n>0$ relációt felhasználva kapjuk az

(2.19)
$$\frac{1}{2}\frac{d}{dt}\left[\left[\mathbf{M}_L\alpha^{(L,n)},\ \alpha^{(L,n)}\right]\right](t) \leqq 0, \quad 0 \leqq t \leqq T$$

differenciál-egyenlőtlenséget. Mivel $\mathbf{M}_L^n$ a bázisfüggvények *Gram-mátrixa* és esetünkben ezek erősen lineárisan függetlenek, ezért tetszőleges $\mathbf{c}_n\in R^{Ln}$ vektor esetén

(2.20)
$$\lambda_0^L\left|\left|\left|\mathbf{c}_n\right|\right|\right|_E^2 \leqq \left[\left[\mathbf{M}_L\mathbf{c}_n, \mathbf{c}_n\right]\right] < \Lambda_0^L\left|\left|\left|\mathbf{c}_n\right|\right|\right|_E^2,$$

ahol $\lambda_0^L \leqq \Lambda_0^L$ az n-től független pozitív állandók. Integráljuk a (2.91) egyenlőtlenséget egy tetszőleges $[0, t_1]$ $(0<t_1\leqq T)$ intervallumon! Felhasználva a (2.20) becslést

(2.21)
$$\left|\left|\left|\alpha^{(L,n)}(t_1)\right|\right|\right|_E^2 \leqq \frac{\Lambda_0^L}{\lambda_0^L}\left|\left|\left|\alpha^{(L,n)}(0)\right|\right|\right|_E^2.$$

A jobb oldalra a (2.13) és a (2.14) összefüggések alapján a

(2.22)
$$\left|\left|\left|\alpha^{(L,n)}(0)\right|\right|\right|_E^2 \leqq \left|\left|\left|\mathbf{u}_0\right|\right|\right|_{L_2,L(\Omega)}^2$$

becslést nyerjük. Így

$$\left\|\left\|\alpha^{(L,n)}\right\|\right\|_E^2(t) \leqq C_0^L, \quad 0 \leqq t \leqq T, \tag{2.23}$$

ahol

$$C_0^L = \frac{\Lambda_0^L}{\lambda_0} \left\|\left\|\mathbf{u}_0\right\|\right\|_{L_{2,L}(\Omega)}^2. \tag{2.24}$$

Nyilvánvaló, hogy (2.23) alapján komponensenként is igaz a

$$|\alpha_i^{(l,n)}(t)| \leqq c_0^L; \quad i = 1, \ldots, n; \quad l = 1, \ldots, L \tag{2.25}$$

becslés. A továbbiakban felhasználjuk, hogy a $g_n \varphi_i^{(l,n)}(x)$ függvények n-től függetlenül korlátosak, azaz

$$\max_{x \in \Omega} |g_n \varphi_i^{(l,n)}(x)| \leqq C_1^L; \quad i = 1, \ldots, n; \quad l = 1, \ldots, L \tag{2.26}$$

és azt, hogy tetszőleges $x \in \Omega$ pont csak véges számú supp $\varphi^{(l,n)}(x)$-hez tartozik ($l = 1, 2, \ldots, L$). Jelölje ezek számát $C_{2,l}$ és

$$C_2^L = \max_{l \leqq 1 \leqq L} C_{2,l}. \tag{2.27}$$

(Ismételten megjegyezzük, hogy C_1^L és C_2^L n-től függetlenek.)

Ekkor (2.11) alapján

$$|u_n^l(x, t)| \leqq C_0^L \cdot C_1^L \cdot C_2^L, \tag{2.28}$$

és így

$$\left\|\left\|\mathbf{u}_n\right\|\right\|_{C_L(Q_T)} \leqq C_3^L, \tag{2.29}$$

ahol

$$C_3^L = C_0^L \cdot C_1^L \cdot C_2^L \tag{2.30}$$

n-től független pozitív állandó, amely csak az egyes komponensek közelítésének típusától és az $\mathbf{u}_0$ adott függvénytől függ. A (2.29) egyenlőtlenségből közvetlenül adódik a

$$\sup_{0<t<T} \left\|\left\|\mathbf{u}_n(., t)\right\|\right\| \leqq C_4^L \tag{2.31}$$

becslés, ahol

$$C_4^L = C_3^L (\text{mes } \Omega)^{1/2}. \tag{2.32}$$

A (2.31) becslés $\mathbf{u}_n(x, t)$ n-től való független létezését, míg (2.29) és (2.31) a korlátosságot jelentik.

2.2. ÁLLÍTÁS. Tetszőleges rögzített $t \in [0, T]$ esetén minden $n \in N^+$ mellett a $\mathbf{h}_i^{L,n}(x) = \mathbf{F}_0(\mathbf{u}_n(x, t), x, t)$ vektor $\mathbf{L}_{2,L}(\Omega)$-beli.

Bizonyítás. Az állítást elegendő a $\mathbf{h}_i^{L,n}(x)$ vektor komponenseire belátni, azaz, hogy mindegyik komponens $L_2(\Omega)$-beli. Ez viszont a [3] 2.3. állításából következik.

Vezessük be az

$$\mathbf{F}_{ij}(\mathbf{u}, x, t) = [F_{ij}^1(\mathbf{u}, x, t), \ldots, F_{ij}^L(\mathbf{u}, x, t)] \tag{2.33}$$

jelölést.

2.3. ÁLLÍTÁS. Tegyük fel, hogy az SS2 feltételrendszer teljesül: Ekkor létezik olyan L_1^L, n-től független pozitív állandó, hogy tetszőleges $\varphi \in \mathbf{L}_{2,L}(\Omega)$ és rögzített $t \in [0, T]$ esetén érvényesek az

$$|((\mathbf{F}_0(\mathbf{u}, ., t) - \mathbf{F}_0(\mathbf{u}_n, ., t), \varphi))| \leqq L_1^L \|\|\mathbf{u} - \mathbf{u}_n\|\|_{L_{2,L}(\Omega)} \|\|\varphi\|\|_{L_{2,L}(\Omega)} \tag{2.34}$$

$$|((\mathbf{F}_{ij}(\mathbf{u}, ., t) - \mathbf{F}_{ij}(\mathbf{u}, ., t), \varphi))| \leqq L_1^L \|\|\mathbf{u} - \mathbf{u}_n\|\|_{L_{2,L}(\Omega)} \|\|\varphi\|\|_{L_{2,L}(\Omega)} \tag{2.35}$$

egyenlőtlenségek, ahol $\mathbf{u}(x, t)$ a (2.4)—(2.6) feladat megoldása, $\mathbf{u}_n(x, t)$ pedig a (2.9)—(2.10) feladatokkal definiált *Galerkin-féle szemidiszkrét közelítések sorozata.*

Bizonyítás. Jelölje $\mathbf{M}_{c_6^L}$ az $L_{2,L}(\Omega)$ tér következő korlátos alterét:

$$\mathbf{M}_{c_6^L} = \{\mathbf{v} | \mathbf{v} \in L_{2,L}(\Omega);\ \|\|\mathbf{v}\|\|_{L_{2,L}(\Omega)} < C_6^L;\ \mathbf{F}_0(\mathbf{v}, x, t) \in L_{2,L}(\Omega), \quad C_6^L > 0,\ t \in [0, T] \text{ rögzített}\}. \tag{2.36}$$

Ekkor az

$$\Phi_t(v) = \mathbf{F}_0(\mathbf{v}, x, t) \tag{2.37}$$

$\mathbf{M}_{c_6^L} \to \mathbf{L}_{2,L}(\Omega)$ operátor deriváltja a leképezés *Jacobi mátrixa* [5], [8], azaz

$$\Phi_t'(v) = \left[\frac{\partial F_0^l(\xi, x, t)}{\partial \xi_i}\right]_{\substack{i,l=1,L \\ \xi=\mathbf{v}}}. \tag{2.38}$$

Lássuk be, hogy $\Phi_t'(v)$ korlátos $\mathbf{M}_{c_6^L}$-on, azaz létezik olyan $C_7^L > 0$ állandó, hogy tetszőleges $\mathbf{v} \in \mathbf{M}_{c_6^L}$ és $\mathbf{w} \in \mathbf{L}_{2,L}(\Omega)$ esetén

$$\|\|\Phi_t'(\mathbf{v})(\mathbf{w})\|\|_{L_{2,L}(\Omega)} \leqq C_7^L \|\|\mathbf{w}\|\|_{L_{2,L}(\Omega)}. \tag{2.39}$$

Jelölje $((\Phi_t^l)'(\mathbf{v})(\mathbf{w}))$ a (2.39) bal oldalán levő L tagú összeg l-edik tagját, azaz (2.38) figyelembevételével

$$(\Phi_t^l)'(\mathbf{v})(\mathbf{w}) = \sum_{k=1}^{L} \left.\frac{\partial F_0^l(\xi, x, t)}{\partial \xi_k}\right|_{\xi=\mathbf{v}} w_k. \tag{2.40}$$

Ekkor tetszőleges $l = 1, 2, \ldots, L$ és rögzített $t \in [0, T]$ esetén

$$\|(\Phi_t^l)'(\mathbf{v})(\mathbf{w})\| \leqq \sum_{k=1}^{L} \left\| \left.\frac{\partial F_0^l(\xi, x, t)}{\partial \xi_k}\right|_{\xi=\mathbf{v}} \right\| w_k. \tag{2.41}$$

Tekintsük az

$$\Phi_t^{l,k}(\mathbf{v}) = \left.\frac{\partial F_0^l(\xi, x, t)}{\partial \xi_k}\right|_{\xi=\mathbf{v}} \tag{2.42}$$

$\mathbf{M}_{c_6^L} \to L_2(\Omega)$ leképezést! Az SS2 feltételek alapján $\Phi_t^{l,k}$ folytonos, azaz valamely kompakton korlátos. Így létezik olyan $L_{l,k}$ állandó, hogy

$$\|\Phi_t^{l,k}(\mathbf{v})\| \leqq L_{l,k}, \quad \forall \mathbf{v} \in \mathbf{M}_{c_6^L}. \tag{2.43}$$

Összegezve a (2.41) és (2.42) összefüggéseket:

$$\|(\Phi_t^l)'(\mathbf{v})(\mathbf{w})\| \leqq \sum_{k=1}^{L} L_{l,k}\|w_k\|, \tag{2.44}$$

azaz

$$\left\|\|\Phi_t'(\mathbf{v})(\mathbf{w})\|\right\|_{L_2,L(\Omega)} \leqq \sum_{l=1}^{L}\sum_{k=1}^{L} L_{l,k}\|w_k\|, \tag{2.45}$$

amiből (2.39) közvetlenül következik. Így $\mathbf{M}_{c_6^L}$-en a Φ_t' operátor a *Lipschitz-feltételt* is kielégíti, azaz létezik olyan $L_1^L>0$ állandó, hogy tetszőleges $\mathbf{v}, \mathbf{w}\in\mathbf{M}_{c_6^L}$ esetén

$$\left\|\|\Phi_t'(\mathbf{v})-\Phi_t'(\mathbf{w})\|\right\|_{L_2,L(\Omega)} \leqq L_1^L\left\|\|\mathbf{v}-\mathbf{w}\|\right\|_{L_2,L(\Omega)}. \tag{2.46}$$

A 2.2. állítás alapján tetszőleges $n\in N^+$ esetén $\mathbf{h}_t^n(x)$ véges. Jelölje

$$C_6^L = \max\left\{\sup_{0<t\leqq T}\left\|\|\mathbf{u}(.,t)\|\right\|_{L_2,L(\Omega)};\ \sup_{\substack{0<t\leqq T\\ n\in\mathbf{N}^+}}\left\|\|\mathbf{u}_n(.,t)\|\right\|_{L_2,L(\Omega)}\right\}. \tag{2.47}$$

Ekkor (2.46) és a *Cauchy—Schwarz egyenlőtlenség* alkalmazásával a 2.3 állítás közvetlen belátható.

A (2.35) egyenlőtlenség ezzel megegyező módon bizonyítható. Megjegyezzük, hogy az SS1 feltételrendszerek esetén a (2.34) és (2.35) egyenlőtlenségek közvetlenül beláthatók [1].

3. A Galerkin-sorozat konvergenciája rendszerekre

Ebben a szakaszban a (2.9)—(2.10) feladatokkal definiált $\mathbf{u}_n(x,t)$ *Galerkin-féle szemidiszkrét közelítések* sorozatának a (2.4)—(2.6) feladat $\mathbf{u}(x,t)$ megoldásához való konvergenciáját bizonyítjuk be. (Megjegyezzük, hogy a bizonyítás menete nagyrészt megegyezik [3] 3.1. állításának bizonyításával, ezért az ismétlődő bizonyítástechnikai elemeket mellőzni fogjuk.)

Legyen $\tilde{\mathbf{u}}_n(x,t)\in C(0,T;\ \mathbf{H}_L^n)$ tetszőleges vektor. Ekkor bevezetve a

$$\mathbf{e}_n(x,t) = \mathbf{u}(x,t)-\mathbf{u}_n(x,t) \tag{3.1}$$

hibavektor jelölést, érvényes a következő azonosság:

$$\begin{aligned}\left(\left(\frac{\partial \mathbf{e}_n}{\partial t},\ \mathbf{e}_n\right)\right)+\sum_{l=1}^{L} a_l(t)\,(\tilde{\mathbf{u}}_n;\ \mathbf{e}_n,\mathbf{e}_n) = \\ = \sum_{l=1}^{L}\left(a_l(t)(\mathbf{u}_n;\ \mathbf{u},\tilde{\mathbf{u}}_n-\mathbf{u}_n)-a_l(t)(\mathbf{u};\ \mathbf{u},\tilde{\mathbf{u}}_n-\mathbf{u}_n)\right)+\left(\left(\frac{\partial \mathbf{e}_n}{\partial t},\ \mathbf{u}-\mathbf{u}_n\right)\right)+ \\ +\sum_{l=1}^{L} a_l(t)(\mathbf{u}_n;\ \mathbf{e}_n,\mathbf{u}-\tilde{\mathbf{u}}_n)+\left((\mathbf{F}_0(\mathbf{u},.,t)-\mathbf{F}_0(\mathbf{u}_n,.,t),\ \tilde{\mathbf{u}}_n-\mathbf{u}_n\right)).\end{aligned} \tag{3.2}$$

Adjunk becsléseket a (3.2) azonosság egyes tagjaira!

$$\left(\left(\frac{\partial \mathbf{e}_n}{\partial t},\ \mathbf{e}_n\right)\right) = \frac{1}{2}\frac{d}{dt}\,|||\mathbf{e}_n|||^2_{L_2,L(\Omega)}; \tag{3.3}$$

(3.4)

$$\sum_{l=1}^{L} a_l(t)(\mathbf{u}_n;\ \mathbf{e}_n, \mathbf{e}_n) = \sum_{l=1}^{L} \int_{\Omega} F_{ij}^l(\mathbf{u}_n, x, t)\frac{\partial \mathbf{e}_n^l}{\partial x_i}\frac{\partial \mathbf{e}_n^l}{\partial x_j}\,dx \leqq F_* \sum_{l=1}^{L}\left\|\frac{\partial \mathbf{e}_n^l}{\partial x}\right\|^2 = F_*\,|||\mathbf{e}_n|||^2_{1,L}.$$

A (2.35) egyenlőtlenséget alkalmazva

$$\left|\sum_{l=1}^{L}\{a_l(t)(\mathbf{u}_n, \mathbf{u}, \tilde{\mathbf{u}}_n-\mathbf{u}_n)-a_l(t)(\mathbf{u};\ \mathbf{u}, \tilde{\mathbf{u}}_n-\mathbf{u}_n)\}\right| \leqq$$

$$\leqq \sum_{l=1}^{L}\left\{\sum_{i,j=1}^{N}\int_{\Omega}|F_{ij}^l(\mathbf{u}_n, x, t)-F_{ij}(\mathbf{u}, x, t)|\left|\frac{\partial u^l}{\partial x_i}\right|\left|\frac{\partial(\tilde{u}_n^l-u_n^l)}{\partial x_j}\right|dx\right\} \leqq$$

$$\leqq \sup_{0<t\leqq T}|||\mathbf{u}|||_{1,L}\sum_{i,j=1}^{N}\left(\left(F_{ij}(\mathbf{u}_n, .\,, t)-F_{ij}(\mathbf{u}, .\,, t),\ \frac{\partial(\tilde{\mathbf{u}}_n-\mathbf{u}_n)}{\partial x_j}\right)\right) \leqq \tag{3.5}$$

$$\leqq |||\mathbf{u}|||_{C_L(\Omega)}L_1^L\sum_{i,j=1}^{N}|||\mathbf{u}-\mathbf{u}_n|||_{L_2,L(\Omega)}\left\|\frac{\partial(\tilde{\mathbf{u}}_n-\mathbf{u}_n)}{\partial x_j}\right\|_{L_2,L(\Omega)} \leqq$$

$$\leqq |||\mathbf{u}|||_{C_L(\Omega)}L_1^L N^2\,|||\mathbf{e}_n|||_{L_2,L(\Omega)}|||\tilde{\mathbf{u}}_n-\mathbf{u}_n|||_{1,L}.$$

Bevezetve a

$$K_1^L = |||\mathbf{u}|||_{C_L(\Omega)}L_1^L N^2 \tag{3.6}$$

jelölést ($K_1^L>0$, n-től független állandó) és alkalmazva a háromszög-egyenlőtlenséget, (3.5) alapján:

$$\left|\sum_{l=1}^{L}\{a_l(t)(\mathbf{u}_n;\ \mathbf{u}, \tilde{\mathbf{u}}_n-\mathbf{u}_n)-a_l(t)(\mathbf{u};\ \mathbf{u}, \tilde{\mathbf{u}}_n-\mathbf{u}_n)\}\right| \leqq$$

$$\leqq K_1^L|||\mathbf{e}_n|||_{L_2,L(\Omega)}(|||\mathbf{u}-\tilde{\mathbf{u}}_n|||_{1,L}+|||\mathbf{e}_n|||_{1,L}). \tag{3.7}$$

Az „ε-egyenlőtlenség" alkalmazásával hasonlóan nyerhető az

$$\left|\sum_{l=1}^{L} a_l(t)(\mathbf{u}_n;\ \mathbf{e}_n, \mathbf{u}-\tilde{\mathbf{u}}_n)\right| \leqq F^*\varepsilon_1|||\mathbf{e}_n|||^2_{1,L}+\frac{F^*}{4\varepsilon_1}|||\mathbf{u}-\tilde{\mathbf{u}}_n|||^2_{1,L} \tag{3.8}$$

egyenlőtlenség, ahol $\varepsilon_1>0$ tetszőleges állandó. A (3.7) és (3.8) egyenlőtlenség alapján

$$\left|\sum_{l=1}^{L}\{a_l(t)(\mathbf{u}_n;\ \mathbf{u}, \tilde{\mathbf{u}}_n-\mathbf{u}_n)-a_l(t)(\mathbf{u};\ \mathbf{u}, \tilde{\mathbf{u}}_n-\mathbf{u}_n)\}\right|+\left|\sum_{l=1}^{L} a_l(t)(\mathbf{u}_n;\ \mathbf{e}_n, \mathbf{u}_n-\tilde{\mathbf{u}}_n)\right| \leqq$$

$$\leqq K_2^L|||\mathbf{e}_n|||^2_{1,L}+K_3^L(|||\mathbf{u}-\tilde{\mathbf{u}}_n|||^2_{L_2,L(\Omega)}+|||\mathbf{e}_n|||^2_{L_2,L(\Omega)}), \tag{3.9}$$

ahol

$$K_2^L = \frac{K_1^L}{2\varepsilon_2}+F^*\varepsilon_1,$$

$$K_3 = \max\left\{\frac{K_1^L}{2\varepsilon_2}+\frac{F^*}{4\varepsilon_1};\ \varepsilon_2 K_1^L\right\}. \tag{3.10}$$

A (2.34) egyenlőtlenség, az „ε-egyenlőtlenség" és a *Friedrichs-egyenlőtlenség* alapján

$$(3.11)\qquad \left|\left(\left(\mathbf{F}_0(\mathbf{u}, ., t)-\mathbf{F}_0(\mathbf{u}_n, ., t),\ \tilde{\mathbf{u}}_n-\mathbf{u}_n\right)\right)\right| \leqq K_4^L \left|\left|\left|\mathbf{e}_n\right|\right|\right|^2_{L_{2,L}(\Omega)} + K_5^L \left|\left|\left|\tilde{\mathbf{u}}_n-\mathbf{u}\right|\right|\right|^2_{1,L},$$

ahol

$$(3.12)\qquad K_4^L = L_1\varepsilon_3;\quad K_5^L = \frac{L_1^L K_\Omega}{2\varepsilon_3}$$

(K_Ω a *Friedrichs-egyenlőtlenségben* szereplő, Ω-tól függő állandó.)

Összegezve; a (3.3), (3.4), (3.9) és (3.11) összefüggések alapján a (3.2) azonosságból következik a

$$(3.13)\qquad \begin{aligned} &\frac{d}{dt}\left|\left|\left|\mathbf{e}_n\right|\right|\right|^2_{L_{2,L}(\Omega)} + K_6^L \left|\left|\left|\mathbf{e}_n\right|\right|\right|^2_{1,L} \leqq \\ &\leqq K_7^L\left\{\left|\left|\left|\mathbf{u}-\tilde{\mathbf{u}}_n\right|\right|\right|^2_{1,L} + \left|\left|\left|\mathbf{e}_n\right|\right|\right|^2_{L_{2,L}(\Omega)}\right\} + 2\left|\left(\left(\frac{\partial \mathbf{e}_n}{\partial t},\ \mathbf{u}-\tilde{\mathbf{u}}_n\right)\right)\right| \end{aligned}$$

egyenlőtlenség, ahol

$$(3.14)\qquad \begin{aligned} K_6^L &= 2(F_*-K_2^L), \\ K_7^L &= 2\max\{K_3^L+K_5^L;\ K_3^L+K_4^L\}. \end{aligned}$$

Az ε_1 és ε_2 pozitív állandókat válasszuk meg úgy, hogy a K_6^L állandó is pozitív legyen! A (3.13) egyenlőtlenségre megismételve a [3]-ban leírt, főként technikai elemeket tartalmazó bizonyítást, a következőket kapjuk.

ÁLLÍTÁS 3.1. Legyen $\mathbf{u}(x, t)$ a (2.4)—(2.6) feladat megoldása, $\mathbf{u}_n(x, t)$ a (2.9)—(2.10) feladatokban meghatározott *Galerkin-féle szemidiszkrét közelítések sorozata;* $\tilde{\mathbf{u}}_n(x, t)\in C(0, T;\ H_L^n)$ tetszőleges vektor. Ekkor az $\mathbf{e}_n=\mathbf{u}-\mathbf{u}_n$ hibafüggvény vektora érvényes a

$$(3.15)\qquad \begin{aligned} &\left|\left|\left|\mathbf{e}_n\right|\right|\right|^2_{H_{2,L}\times C(0,T)} + K_{13}^L \left|\left|\left|\mathbf{e}_n\right|\right|\right|^2_{H^1_{0,L}\times L_2(0,T)} \leqq \\ &\leqq K_{15}^L\left\{\left|\left|\left|\mathbf{u}-\tilde{\mathbf{u}}_n\right|\right|\right|^2_{H^1_{0,L}\times L_2(0,T)} + \left|\left|\left|\frac{\partial(\mathbf{u}-\tilde{\mathbf{u}}_n)}{\partial t}\right|\right|\right|^2_{L_{2,L}(Q_T)} + \left|\left|\left|\mathbf{u}-\tilde{\mathbf{u}}_n\right|\right|\right|^2_{L_{2,L}\times C(0,T)}\right\} \end{aligned}$$

becslés, ahol K_{13}^L és K_{15}^L az n-től független pozitív állandók.

3.1. Megjegyzés. A (3.15) becslésből a [3] 3.2. állításával megegyező módon következik az $(\mathbf{u}_n)$ sorozat konvergenciája és annak nagyságrendje.

3.2. Megjegyzés. A nyert eredmények a bizonyítások lényeges megváltoztatása nélkül kiterjeszthetők arra az esetre, amikor a (2.1) egyenlet helyett a

$$(2.1')\qquad \frac{\partial u^l}{\partial t} - \sum_{p=1}^{L}\sum_{i,j=1}^{N}\frac{\partial}{\partial x_i}\left(F_{ij}^{(l,p)}(\mathbf{u}, x, t)\frac{\partial u^p}{\partial x_j}\right) + F_0(\mathbf{u}, x, t) = 0$$

egyenletet vizsgáljuk a (2.2) és (2.3) mellékfeltételekkel. Ekkor SS2-ben még kiegészítőleg feltesszük, hogy az

$$(3.16)\qquad F_M = [F^{(l,p)}]_{l,p=1,L};\quad F^{(l,p)} = [F_{ij}^{(l,p)}]_{i,j=1,N}$$

mátrixok egyenletesen pozitív definitek Q_T^L-en.

A cikkben alkalmazott főbb jelölések jegyzéke:

1. L az ismeretlen függvények száma.
2. $R^{LN} = R^L \times R^N$.
3. $Q_T^L = R^L \times Q_T = R^L \times \Omega \times (0, T]$.
4. $u^1, u^2, \ldots, u^L$ — az ismeretlen függvények.
5. $\mathbf{u} = [u^1, u^2, \ldots, u^L]$.
6. $\mathbf{u}_0 = [u_0^1, u_0^2, \ldots, u_0^L]$.
7. $\mathbf{F}_0(\mathbf{u}, x, t) = [F_0^1(\mathbf{u}, x, t), \ldots, F_0^L(\mathbf{u}, x, t)]$.
8. $a^l(\mathbf{u}) = \sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\left(F_{ij}(\mathbf{u}, x, t)\frac{\partial u^l}{\partial x_j}\right)$.
9. $\mathbf{a}(\mathbf{u}) = [a^1(\mathbf{u}), \ldots, \mathbf{a}^L(\mathbf{u})]$.
10. $\mathbf{F}^l = [F_{ij}^l]_{i,j=1}^N$ $N \times N$ méretű mátrix, $l = 1, \ldots, L$.
11. $\mathbf{F} = \operatorname{diag}[F^1, \ldots, F^L]$ $L \cdot N \times L \cdot N$ méretű mátrix.
12. $a_l(\mathbf{w}; \mathbf{u}, \mathbf{v}) = \sum_{i,j=1}^{N} \int_\Omega F_{ij}^l(\mathbf{w}, x, t)\frac{\partial u^l}{\partial x_i}\frac{\partial v^l}{\partial x_j}\,dx, \quad l = 1, \ldots, L.$
13. $\mathbf{a}_L(t)(\mathbf{w}; \mathbf{u}, \mathbf{v}) = [a_1(\mathbf{w}; \mathbf{u}, \mathbf{v}), \ldots, a_L(\mathbf{w}; \mathbf{u}, \mathbf{v})]$.
14. $\mathbf{H}_L = H \times \ldots \times H$ (H valamilyen függvénytér).
15. $\mathbf{v} \in \mathbf{H}_L$ azt jelenti, hogy a $\mathbf{v} = (v^1, \ldots, v^L)$-re $v^l \in H$.
16. $H_l^n = \operatorname{span}[\varphi_1^{(l,n)}, \ldots, \varphi_n^{(l,n)}]$.
17. $H_L^n = H_1^n \times \ldots \times H_L^n$.
18. $[\mathbf{v}, \mathbf{w}] = \left(\sum_{j=1}^{N} v_j^1 \cdot w_j^1, \ldots, \sum_{j=1}^{N} v_j^L \cdot w_j^L\right); \quad \forall \mathbf{v}, \mathbf{w} \in R^{LN}.$
19. $[[\mathbf{v}, \mathbf{w}]] = \sum_{l=1}^{L}\sum_{j=1}^{N} v_j^l \cdot w_j^l; \quad \|\|\mathbf{v}\|\|_E^2 = [[\mathbf{v}, \mathbf{v}]], \quad \forall \mathbf{v}, \mathbf{w} \in R^{LN}.$
20. $\langle \mathbf{v}, \mathbf{w} \rangle = ((v^1, w^1), \ldots, (v^L, w^L))$.
21. $\langle\langle \mathbf{v}, \mathbf{w} \rangle\rangle = \int_\Omega [[\mathbf{v}, \mathbf{w}]]\,dx, \quad \forall \mathbf{v}, \mathbf{w} \in R^{LN}.$
22. $((\mathbf{v}, \mathbf{w}))_H = \sum_{l=1}^{L} (v^l, w^l)_H$.

23. $|||\mathbf{v}|||^2_{H_L} = ((\mathbf{v}, \mathbf{w}))$.

24. $|||\mathbf{v}|||^2_{L_{2,L}} = \sum_{i=1}^{N} \left|\left|\left| \frac{\partial \mathbf{v}}{\partial x_i} \right|\right|\right|_{L_{2,L}(\Omega)} = |||\mathbf{v}|||^2_{H^1_{0,L}(\Omega)}$.

25. $|||\mathbf{v}|||_{L_{2,L}\times C(0,T)} = \sup_{0<t<T} |||\mathbf{v}(.,t)|||_{L_{2,L}(\Omega)}$.

26. $|||\mathbf{v}|||^2_{H^1_{0,L}\times L_2(0,T)} = \int_0^T |||\mathbf{v}|||^2_{1,L}\, dt$.

IRODALOM

[1] DOUGLAS, J. and DUPONT, T., "Galerkin methods for parabolic equations", *SIAM J. Num. Anal.* **7** (1970) 575—626.
[2] FARAGÓ, I., "Véges elemek módszere lineáris, parabolikus feladatok megoldására", *Alk. Mat. Lapok* **11** (1985) 123—155.
[3] FARAGÓ, I., „Véges elemek módszere nemlineáris, parabolikus feladatok megoldására", *Alk. Mat. Lapok* (ezen szám)
[4] STOYAN, G., "On monotone difference schemes for weakly coupled systems of partial differential equations", *Comp. Math.* **13** (1984) 33—43.

(Beérkezett: 1987. december 1.)

FARAGÓ ISTVÁN
AGRÁRTUDOMÁNYI EGYETEM MATEMATIKAI ÉS
SZÁMÍTÁSTECHNIKAI INTÉZET
2103 GÖDÖLLŐ

FINITE ELEMENT METHOD FOR SOLVING NONLINEAR PARABOLIC SYSTEMS

I. FARAGÓ

In this paper we generalize the results of [3] for nonlinear parabolic systems. The convergence of method is proven, the rate of convergence is also estiamted.

Alkalmazott Matematikai Lapok **13** (1987—88) 361—372

ÁLTALÁNOS HESTENES—STIEFEL TÍPUSÚ KONJUGÁLT IRÁNY REKURZIÓK. II. A FORDÍTOTT REKURZIÓ

HEGEDŰS CSABA J.

Budapest

Egy általános rekurziót ismertetünk, amely egy $m \times n$ típusú komplex elemű A mátrixra nézve generál konjugált vektorpárokat. A következő vektorpár elkészítéséhez négy vektor ismerete szükséges. A módszer jellegzetességei: Nem kell az $\mathbf{A}^H\mathbf{A}$ mátrix elkészítésével a kondíciószámot négyzetelni. A módszer akkor fejeződik be, ha az utolsó pár elemei az $\mathbf{A}^H$ és A mátrixok nullterébe esnek. A korábban ismertetett direkt rekurzióval fordított viszonyban áll, annak vektorait visszafelé képes generálni. Kimutatjuk, hogy a következő vektorok ortogonális vetítésekkel készülnek, szemben a klasszikus módszereket tartalmazó direkt rekurzióval, amely az A-konjugált vektorokból készített ferde vetítésekkel generálja a következő vektorokat. Az itt ismertetett rekurzióból eddig nem ismert új módszerek származtathatók lineáris egyenletrendszerek és egyéb lineáris algebrai alapfeladatok megoldására.

3. A fordított rekurzió[1]

3.1. A fordított rekurzió alapegyenletei

A *fordított rekurzió* is *Hestenes—Stiefel típusú* abban az értelemben, hogy segítségével a mátrix oly módon transzformálható egyszerűbb alakra, hogy az új mátrix szinguláris értékei megegyeznek az eredetiével. Az első cikkben ismertett *direkt rekurzióval* fordított viszonyban áll: Ha a direkt rekurzióról átkapcsolunk a fordítottra, akkor a direkt rekurzió vektorait fordított sorrendben kapjuk vissza. Most a $\mathbf{v}_1$ és $\mathbf{u}_1$ első konjugált vektorokkal indítunk, s az i-edik lépésben először a $\mathbf{v}_{i+1}$, $\mathbf{u}_{i+1}$ vektorokat készítjük el, s ezután generáljuk az $\mathbf{r}_{i+1}$, $\mathbf{q}_{i+1}$ vektorokat. A fordítottság ellenére mégis igaz, hogy ugyanazon $\mathbf{v}_1$ és $\mathbf{u}_1$ vektorokkal indulva a két rekurzió ugyanazt a konjugált pár sorozatot készíti, feltéve, hogy a direkt rekurziónál megismert és most is létező oldallépést egyik módszernél sem alkalmaztuk.

A fordított rekurzióhoz a következő rekurziós egyenletrendszer tartozik:

$$\mathbf{L}_i\mathbf{V}^H - \delta_i \mathbf{e}_i \mathbf{v}_{i+1}^H = \mathbf{D}_r^{-1}\mathbf{R}^H\mathbf{C}, \quad \mathbf{A}\mathbf{U}\mathbf{D}_a^{-1} = \mathbf{R}\mathbf{L}_i \tag{3.1}$$

$$\mathbf{D}_a^{-1}\mathbf{V}^H\mathbf{A} = \mathbf{F}_i\mathbf{Q}^H, \qquad \mathbf{U}\mathbf{F}_i - \varepsilon_i \mathbf{u}_{i+1}\mathbf{e}_i^T = \mathbf{K}\mathbf{Q}\mathbf{D}_q^{-1} \tag{3.2}$$

$$\mathbf{D}_a = \mathbf{V}^H\mathbf{A}\mathbf{U}, \quad \mathbf{D}_r = \mathbf{R}^H\mathbf{C}\mathbf{R}, \quad \mathbf{D}_q = \mathbf{Q}^H\mathbf{K}\mathbf{Q}. \tag{3.3}$$

Itt megtartottuk az előző [1] cikk jelöléseit és konvencióit: A $\mathbf{v}_j$, $\mathbf{u}_j$ vektorok A-konjugált párokat jelölnek, $\mathbf{A} \in C^{m,n}$, az $\mathbf{r}_j$ vektorok C-ortogonális, a $\mathbf{q}_j$ vektorok

[1] A dolgozat a szerző korábbi dolgozatának a folytatása, ezért a szakaszok számozása folytatólagos.

K-ortogonális rendszert képeznek, és a kisbetűs vektorokból az első i vektort egy ugyanolyan nagybetűvel jelölt mátrixba fogtuk össze. Így a $\mathbf{Ve}_j = \mathbf{v}_j$, $j = 1, 2, \ldots, i$ vektorok képezik a $\mathbf{V}$ mátrixot, s hasonlóképp készült az $\mathbf{U}$, $\mathbf{R}$ és $\mathbf{Q}$ mátrix.

$\mathbf{L}_i$ i-edrendű alsó *Hessenberg mátrix*, amely egyszerű esetben diagonálmátrix, vagy felső bidiagonális mátrix alakú. $\mathbf{F}_i$ i-edrendű felső *Hessenberg mátrix*, mely diagonálmátrix vagy alsó bidiagonális mátrix alakra egyszerűsödhet. Bővítésük az alábbi séma szerint történik:

$$\mathbf{L}_{i+1} = \left[\begin{array}{c|c} \mathbf{L}_i & -\delta_i \mathbf{e}_i \\ \hline \mathbf{l}_{i+1}^T & \end{array}\right], \quad \mathbf{F}_{i+1} = \left[\begin{array}{c|c} \mathbf{F}_i & \\ \hline -\varepsilon_i \mathbf{e}_i^T & \mathbf{f}_{i+1} \end{array}\right], \tag{3.4}$$

ahol $\mathbf{l}_{i+1}$ és $\mathbf{f}_{i+1}$ $i+1$-dimenziós vektorok. Értelmezés szerint $\mathbf{l}_j^T$ egy j-elemű sorvektor az $\mathbf{L}_i$, $j \leqq i$ mátrixban, amely a j-edik sor első elemével kezdődik, és a diagonálelemmel végződik. Hasonlóan $\mathbf{f}_j$ az $\mathbf{F}_i$, $j \leqq i$ mátrix j-edik oszlopának az első elemtől a diagonálelemig tartó része. A diagonálelemek az $\mathbf{r}_j$ és $\mathbf{q}_j$ vektorok hosszát skálázzák, a kodiagonális elemek pedig a $\mathbf{v}_j$, $\mathbf{u}_j$ vektorokét. Most is megmaradunk csak az $\mathbf{r}_j$, $\mathbf{q}_j$ vektorok skálázása mellett, így a kodiagonális elemeket adó ε_i és δ_i paraméterek értékét 1-nek választjuk, ha az nullától különböző.

Az $i+1$-edik konjugált vektorok (3.1) első és (3.2) második egyenletéből adódnak, ha az utolsó sort, ill. oszlopot vesszük. Amennyiben $\mathbf{D}_r$ és $\mathbf{D}_q$ diagonálmátrixok, az eredmény:

$$\delta_i \mathbf{v}_{i+1}^H = \mathbf{l}_i^T \mathbf{V}^H - \mathbf{r}_i^H \mathbf{C}/\|\mathbf{r}_i\|_C^2, \quad \varepsilon_i \mathbf{u}_{i+1} = \mathbf{U}\mathbf{f}_i - \mathbf{K}\mathbf{q}_i/\|\mathbf{q}_i\|_K^2. \tag{3.5}$$

A δ_i és ε_i paraméter értékeit a következőképpen választjuk:

$$\delta_i = \begin{cases} 1, & \text{ha } \mathbf{l}_i^T \mathbf{V}^H - \mathbf{r}_i^H C/\|\mathbf{r}_i\|_C^2 \neq \mathbf{0}, \\ 0, & \text{egyébként}, \end{cases} \tag{3.6}$$

$$\varepsilon_i = \begin{cases} 1, & \text{ha } \mathbf{U}\mathbf{f}_i - \mathbf{K}\mathbf{q}_i/\|\mathbf{q}_i\|_K^2 \neq \mathbf{0}, \\ 0, & \text{egyébként}. \end{cases} \tag{3.7}$$

Ha δ_i és ε_i egyaránt nulla, akkor a rekurzió befejeződik.

Az $\mathbf{l}_i^T$ és $\mathbf{f}_i$ vektor alakja általában

$$\mathbf{l}_i^T = \sum_{j=h(i)}^{i} \lambda_j \mathbf{e}_j^T, \quad \mathbf{f}_i = \sum_{j=k(i)}^{i} \varphi_j \mathbf{e}_j, \tag{3.8}$$

ahol a $h(i)$ és $k(i)$ egész értékekre vonatkozóan az alábbi két eset lehetséges:

$$h(i) \leqq i, \quad k(i) = i, \quad \varepsilon_i \in \{0, 1\}, \quad \delta_{i-1} = 1, \tag{3.9}$$

$$h(i) = i, \quad k(i) \leqq i, \quad \delta_i \in \{0, 1\}, \quad \varepsilon_{i-1} = 1. \tag{3.10}$$

A direkt rekurziónál megismert három építési forma most is létezik. Ezek a diagonális építés, a diagonál-folytatás és az oldallépés. A továbbiakban ezeket ismertetjük.

3.2. *Bidiagonális építés módszere*

Ekkor az **L** mátrixot felső, az **F** mátrixot pedig alsó bidiagonális mátrixként építjük, azaz $h(i+1)=k(i+1)=i+1$ és $\delta_i=\varepsilon_i=1$.

Először az **A**-konjugált vektorokat állítjuk elő (3.5) segítségével, amelyben ekkor az $\mathbf{l}_i$ és $\mathbf{f}_i$ vektoroknak csak az i-edik eleme nemzérus:

$$\mathbf{v}_{i+1}^H = \lambda_i \mathbf{v}_i^H - \mathbf{r}_i^H \mathbf{C}/\|\mathbf{r}_i\|_C^2, \quad \mathbf{u}_{i+1} = \varphi_i \mathbf{u}_i - \mathbf{K}\mathbf{q}_i/\|\mathbf{q}_i\|_K^2. \tag{3.11}$$

Elkészítve az $\mathbf{L}_{i+1}$ és $\mathbf{F}_{i+1}$ mátrixokat (3.4) alapján, majd helyettesítve (3.2) első és (3.1) második egyenletébe, az utolsó sor, ill. oszlop szolgáltatja az $i+1$-edik ortogonális vektorokat. Amennyiben a $\mathbf{D}_a$ mátrix diagonálmátrix, a következőket kapjuk:

$$\begin{aligned} \lambda_{i+1}\mathbf{r}_{i+1} &= \mathbf{r}_i + \mathbf{A}\mathbf{u}_{i+1}/\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_{i+1}, \\ \varphi_{i+1}\mathbf{q}_{i+1}^H &= \mathbf{q}_i^H + \mathbf{v}_{i+1}^H\mathbf{A}/\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_{i+1}. \end{aligned} \tag{3.12}$$

A bidiagonális építés sikeres, ha a kapott új vektorok egyike sem zérus, továbbá

$$\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_{i+1} \neq 0.$$

3.3. *Diagonál-folytatás módszere*

Ekkor az **L** és **F** mátrix közül az egyiket diagonális, a másikat pedig bidiagonális mátrixként építjük.

Bal oldali diagonál-folytatás esetén az alkalmazhatóság feltétele: $\delta_i=0$ és $\varepsilon_i=1$. Innen látható, hogy (3.5)-ből nem tudjuk meghatározni a $\mathbf{v}_{i+1}$ vektort, csak az $\mathbf{u}_{i+1}$ vektort. Folytassuk ekkor az **L** mátixot diagonálmátrixként, amikoris $\mathbf{l}_{i+1}^T=\mathbf{e}_{i+1}^T$ és $\delta_{i+1}=0$. Helyettesítsünk az i index helyére $i+1$-et (3.1) egyenleteiben, így az utolsó sor, ill. oszlopból a következő $i+1$-edik vektorokat kapjuk:

$$\mathbf{v}_{i+1}^H = \mathbf{r}_{i+1}^H\mathbf{C}/\|\mathbf{r}_{i+1}\|_C^2, \quad \mathbf{r}_{i+1} = \mathbf{A}\mathbf{u}_{i+1}/\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_{i+1}. \tag{3.13}$$

Ezek megegyeznek az előző cikkünk (2.11) egyenleteivel, amelyek a (2.12) megoldásokra vezetnek:

$$\begin{aligned} &\text{a)}\ \mathbf{r}_{i+1} = \mathbf{A}\mathbf{u}_{i+1}/\|\mathbf{A}\mathbf{u}_{i+1}\|_C, \quad \mathbf{v}_{i+1}^H = \mathbf{r}_{i+1}^H\mathbf{C}, \\ &\text{b)}\ \mathbf{r}_{i+1} = \mathbf{A}\mathbf{u}_{i+1}, \qquad\qquad\quad \mathbf{v}_{i+1}^H = \mathbf{r}_{i+1}^H\mathbf{C}/\|\mathbf{r}_{i+1}\|_C^2. \end{aligned} \tag{3.14}$$

Az a) megoldásnál $\|\mathbf{r}_{i+1}\|_C=1$, a b) megoldásnál $\mathbf{v}_{i+1}^H\mathbf{A}\mathbf{u}_{i+1}=1$.

Hasonló módon a jobb oldali diagonál-folytatás formulái:

$$\mathbf{f}_{i+1} = \mathbf{e}_{i+1}, \quad \varepsilon_i = \varepsilon_{i+1} = 0, \tag{3.15}$$

$$\begin{aligned} &\text{a)}\ \mathbf{q}_{i+1}^H = \mathbf{v}_{i+1}^H\mathbf{A}/\|\mathbf{v}_{i+1}^H\mathbf{A}\|_K, \quad \mathbf{u}_{i+1} = \mathbf{K}\mathbf{q}_{i+1}, \\ &\text{b)}\ \mathbf{q}_{i+1}^H = \mathbf{v}_{i+1}^H\mathbf{A}, \qquad\qquad\quad \mathbf{u}_{i+1} = \mathbf{K}\mathbf{q}_{i+1}/\|\mathbf{q}_{i+1}\|_K^2. \end{aligned} \tag{3.16}$$

A rekurzió befejeződik, ha $\mathbf{A}\mathbf{u}_{i+1}=0$, ill. $\mathbf{v}_{i+1}^H\mathbf{A}=0$.

3.4. Oldallépés módszere: bidiagonális építés és csere az ortogonális vektoroknál az egyik oldalon

Az oldallépés most is a nemkívánt zérus osztó elkerülésének általános módszere. A direkt rekurzióhoz képest annyi a különbség, hogy az i- és $i+1$-edik ortogonális vektorok helyett a bidiagonális építéssel nyert $i-1$- és i-edik ortogonális vektorokat cseréljük fel.

Rátérünk a bal oldali oldallépés átrendezéseire. A (3.1) egyenletek a következő particionált alakba írhatók:

$$\text{(3.17)}\qquad \left[\begin{array}{c|c|c} \mathbf{L}_{i-2} & -\mathbf{e}_{i-2} & \\ \hline & & -\mathbf{e}_{i-1} \\ \tilde{\mathbf{l}}_{i-1}^T & & \\ \hline & \tilde{\mathbf{l}}_i^T & -\mathbf{e}_i \end{array}\right] \left[\begin{array}{c} \mathbf{V}_{i-1}^H \\ \tilde{\mathbf{v}}_i^H \\ \tilde{\mathbf{v}}_{i+1}^H \end{array}\right] = \left[\begin{array}{c} \mathbf{D}_{r,i-2}^{-1}\mathbf{R}_{i-2}^H\mathbf{C} \\ \tilde{\mathbf{r}}_{i-1}^H\mathbf{C}/\|\tilde{\mathbf{r}}_{i-1}\|_C^2 \\ \tilde{\mathbf{r}}_i^H\mathbf{C}/\|\tilde{\mathbf{r}}_i\|_C^2 \end{array}\right],$$

$$\text{(3.18)}\qquad \mathbf{A}\mathbf{U}_i\tilde{\mathbf{D}}_a^{-1} = [\mathbf{R}_{i-2}\ \tilde{\mathbf{r}}_{i-1}\ \tilde{\mathbf{r}}_i] \left[\begin{array}{c|c|c} \mathbf{L}_{i-2} & -\mathbf{e}_{i-2} & \\ \hline & & -\mathbf{e}_{i-1} \\ \tilde{\mathbf{l}}_{i-1}^T & & \\ \hline & \tilde{\mathbf{l}}_i^T & \end{array}\right],$$

ahol a megváltozó mennyiségeket felül hullámmal jelöltük, az $\mathbf{L}_i$ mátrixot a (3.4) séma alapján az $\mathbf{L}_{i-2}$ mátrixból építettük fel, és az alsó index a $\mathbf{V}$, $\mathbf{R}$ mátrixoknál arra utal, hogy azok hány vektorból készültek. A bidiagonális építéssel kezdünk, amihez $\tilde{\mathbf{l}}_i^T = \tilde{\lambda}_i\mathbf{e}_i^T$ tartozik.

Vezessük be a következő i- és $i+1$-edrendű segédmátrixokat:

$$\mathbf{P}_i = \begin{bmatrix} \mathbf{I}_{i-2} & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},\quad \mathbf{Z}_i = \begin{bmatrix} \mathbf{I}_{i-1} & 0 \\ 0 & -\tilde{\lambda}_i^{-1} \end{bmatrix},\quad \mathbf{G}_{i+1} = \begin{bmatrix} \mathbf{I}_{i-1} & 0 & 0 \\ 0 & -\tilde{\lambda}_i^{-1} & 1 \\ 0 & 0 & \tilde{\lambda}_i \end{bmatrix}.$$

Szorozzuk (3.17)-et balról a $\mathbf{P}_i$ mátrixszal, az $\mathbf{L}$ és $\mathbf{V}$ mátrix közé írjuk be a $\mathbf{G}_{i+1}\mathbf{G}_{i+1}^{-1}$ szorzatot és szorozzunk az egyikkel balra, a másikkal jobbra. Az eredmény:

$$\left[\begin{array}{c|c|c|c} \mathbf{L}_{i-2} & -\mathbf{e}_{i-2} & 0 & 0 \\ \hline 0 & & -1 & 0 \\ \hline \tilde{\mathbf{l}}_{i-1}^T & & \tilde{\lambda}_i^{-1} & -1 \end{array}\right] \left[\begin{array}{c} \mathbf{V}_{i-1}^H \\ -\tilde{\lambda}_i\tilde{\mathbf{v}}_i^H + \tilde{\mathbf{v}}_{i+1}^H \\ \tilde{\lambda}_i^{-1}\tilde{\mathbf{v}}_{i+1}^H \end{array}\right] = \left[\begin{array}{c} \mathbf{D}_{r,i-2}^{-1}\mathbf{R}_{i-2}^H\mathbf{C} \\ \tilde{\mathbf{r}}_i^H\mathbf{C}/\|\tilde{\mathbf{r}}_i\|_C^2 \\ \tilde{\mathbf{r}}_{i-1}^H\mathbf{C}/\|\tilde{\mathbf{r}}_{i-1}\|_C^2 \end{array}\right].$$

(3.18)-ban az $\mathbf{R}$, $\mathbf{L}$ mátrixok közé az $\mathbf{I}=\mathbf{P}_i\mathbf{P}_i$ szorzatot visszük be, és jobbról mindkét oldalt szorozzuk a $\mathbf{Z}_i$ mátrixszal:

$$\mathbf{A}\mathbf{U}_i\tilde{\mathbf{D}}_a^{-1}\mathbf{Z}_i = \mathbf{A}\mathbf{U}_i\mathbf{D}_a^{-1} = [\mathbf{R}_{i-2}\ \tilde{\mathbf{r}}_i\ \tilde{\mathbf{r}}_{i-1}] \left[\begin{array}{c|c|c} \mathbf{L}_{i-2} & -\mathbf{e}_{i-2} & 0 \\ \hline 0 & & -1 \\ \hline \tilde{\mathbf{l}}_{i-1}^T & & \tilde{\lambda}_i^{-1} \end{array}\right].$$

A kapott alakokról leolvashatók, hogy beleillenek a (3.1)—(3.4), vagy az ezeknek megfelelő (3.17)—(3.18) általános sémába. Így a következő új hullám nélküli mennyiségeket vezethetjük be:

$$\mathbf{l}_{i-1}^T = \mathbf{0}, \quad \lambda_i = \tilde{\lambda}_i^{-1}, \quad \mathbf{l}_i^T = [\tilde{\mathbf{l}}_{i-1}^T \ \lambda_i], \tag{3.19}$$

$$\mathbf{r}_{i-1} = \tilde{\mathbf{r}}_i, \quad \mathbf{r}_i = \tilde{\mathbf{r}}_{i-1}, \quad \mathbf{v}_{i+1}^H = \tilde{\lambda}_i^{-1}\tilde{\mathbf{v}}_{i+1}^H, \tag{3.20}$$

$$\mathbf{v}_i^H = -\tilde{\mathbf{r}}_i^H \mathbf{C}/\|\tilde{\mathbf{r}}_i\|_C^2 = -\tilde{\lambda}_i \tilde{\mathbf{v}}_i^H + \tilde{\mathbf{v}}_{i+1}^H, \tag{3.21}$$

$$\mathbf{v}_i^H \mathbf{A}\mathbf{u}_i = -\tilde{\lambda}_i \tilde{\mathbf{v}}_i^H \mathbf{A}\mathbf{u}_i, \tag{3.22}$$

ahol a (3.22) összefüggést a $\mathbf{D}_a^{-1} = \tilde{\mathbf{D}}_a^{-1}\mathbf{Z}_i$ egyenlőségből nyertük.

A jobb oldali oldallépés átrendezései hasonlóan adódnak:

$$\mathbf{f}_{i-1} = \mathbf{0}, \quad \varphi_i = \tilde{\varphi}_i^{-1}, \quad \mathbf{f}_i^T = [\tilde{\mathbf{f}}_{i-1}^T \ \varphi_i], \tag{3.23}$$

$$\mathbf{q}_{i-1} = \tilde{\mathbf{q}}_i, \quad \mathbf{q}_i = \tilde{\mathbf{q}}_{i-1}, \quad \mathbf{u}_{i+1} = \tilde{\varphi}_i^{-1}\tilde{\mathbf{u}}_{i+1}, \tag{3.24}$$

$$\mathbf{u}_i = -\mathbf{K}\tilde{\mathbf{q}}_i/\|\tilde{\mathbf{q}}_i\|_K^2 = -\tilde{\varphi}_i \tilde{\mathbf{u}}_i + \tilde{\mathbf{u}}_{i+1}, \tag{3.25}$$

$$\mathbf{v}_i^H \mathbf{A}\mathbf{u}_i = -\tilde{\varphi}_i \mathbf{v}_i^H \mathbf{A}\tilde{\mathbf{u}}_i. \tag{3.26}$$

Az oldallépés átrendezéseit mindig először az egyik oldalon kell végrehajtanunk, s ezután készülnek az *i-edik ortogonális vektorral kezdve* a másik oldali vektorok bidiagonális építéssel.

A fordított rekurziónál a direkt rekurzióval ellentétben nem a kodiagonális elemek tolódnak el a rekurziós mátrixokban, hanem a diagonális elemek. Ha például a h-adik szinttől minden lépésben oldallépéssel generáljuk a bal oldali vektorokat az i-edik szintig, akkor az $\mathbf{L}_i$ mátrix a következő alakú:

$$\mathbf{L}_i = \begin{bmatrix} \ddots & & & & \\ & \lambda_{h-1} & -1 & & \\ & & 0 & -1 & \\ & & \vdots & \ddots & \ddots \\ & & 0 & 0 \ \dots & -1 \\ & & \lambda_h & \lambda_{h+1} \dots \lambda_{i-1} & \lambda_i \end{bmatrix} \begin{matrix} \\ \\ \\ \\ (h \\ \\ (i \end{matrix}, \tag{3.27}$$

és a konstrukcióból nyilvánvaló, hogy $\lambda_h \neq 0$.

Oldallépés diagonál-folytatáskor is tehető azon az oldalon, ahol a bidiagonális építés történik. Ilyenkor a diagonálmátrix a másik oldalon bidiagonális mátrixba fog kinyílni.

3.5. A fordított rekurzió működésének alaptétele

Vezessük be a direkt rekurziónál már megismert konvenciót: A bidiagonális építéssel nyert vektorokat felül hullámmal jelöljük, és őket próbavektoroknak tekintjük. Hullám nélkül az elfogadott vektorokat fogjuk jelölni. Elfogadható lehet bármely építéssel készített vektor, amely sikerrel bővít, tehát amelyre a kapott $i+1$-edik vektorok nemzérusok és $\mathbf{v}_{i+1}^H \mathbf{A}\mathbf{u}_{i+1} \neq \mathbf{0}$. Az alaptétel analóg a 2.1. tétellel:

3.1. TÉTEL. Amennyiben a $\mathbf{v}_{i+1}^H\mathbf{A}$ vagy $\mathbf{Au}_{i+1}$ vektoroknak legalább az egyike nemzérus, a (3.1)—(3.4) fordított rekurzió építési módszereinek valamelyikével az $\{\mathbf{r}_j\}_{j=1}^i$, $\{\mathbf{q}_j\}_{j=1}^i$ és $\{\mathbf{v}_j, \mathbf{u}_j\}_{j=1}^i$ vektorrendszerek a kívánt ortogonalitási tulajdonságok megtartásával sikerrel bővíthetők. Mégpedig:

1. A bidiagonális építés sikeres, ha $\tilde{\mathbf{v}}_{i+1}^H\mathbf{A}\tilde{\mathbf{u}}_{i+1}\neq\mathbf{0}$;
2. A diagonál-folytatás sikeres a bal oldalon, ha $\tilde{\mathbf{v}}_{i+1}=\mathbf{0}$ és $\mathbf{A}\tilde{\mathbf{u}}_{i+1}\neq\mathbf{0}$, jobb oldalon, ha $\tilde{\mathbf{u}}_{i+1}=\mathbf{0}$ és $\tilde{\mathbf{v}}_{i+1}^H\mathbf{A}\neq\mathbf{0}$;
3. Ha $\tilde{\mathbf{v}}_{i+1}^H\mathbf{A}\neq\mathbf{0}$, akkor a bidiagonális építés vagy a bal oldallépés legalább egyike sikeres. Ha $\mathbf{A}\tilde{\mathbf{u}}_{i+1}\neq\mathbf{0}$, akkor a bidiagonális építés vagy a jobb oldallépés legalább egyike sikeres.

3.6. Az alaptétel igazolása

Ismét több tétel igazolása vezet majd el az alaptétel teljes bizonyításához. A bizonyítás indukciós: Feltesszük, hogy az i-edik szintig minden állítás igaz, s az $i+1$-edik szintre bizonyítunk. Az első konjugált vektorok meghatározzák az első ortogonális vektorokat, és az 1-elemű vektor-rendszerekre az állítások triviálisan teljesülnek. Az algebrai hasonlóság miatt a bizonyításokat csak az egyik (bal) oldali vektorokra fogjuk elvégezni.

3.2. TÉTEL. A $\mathbf{v}_{i+1}$ vektor ortogonális az $\mathbf{Au}_j$, $j=1, 2, \ldots, i$ és az $\mathbf{u}_{i+1}$ vektor ortogonális az $\mathbf{A}^H\mathbf{v}_j$, $j=1, 2, \ldots, i$ vektorokra. Ennek következménye, hogy $\mathbf{D}_a$ diagonálmátrix.

Bizonyítás. Szorozzuk össze (3.1), ill. (3.2) két-két egyenletét, bal oldalt a ballal, jobb oldalt a jobbal:

$$\delta_i\mathbf{e}_i\mathbf{v}_{i+1}^H\mathbf{AUD}_a^{-1}=\mathbf{0},\quad \varepsilon_i\mathbf{D}_a^{-1}\mathbf{V}^H\mathbf{Au}_{i+1}\mathbf{e}_i^T=\mathbf{0}. \tag{3.28}$$

Ezekből az ortogonalitási relációk következnek, feltéve, hogy δ_i és $\varepsilon_i\neq 0$. Arra az esetre, amikor δ_i vagy ε_i egyike zérus, a bizonyítást a következő tétel egyik lépésében fogjuk megtenni. ∎

Vezessük be a következő projektorokat:

$$\mathbf{P}_i^R=\mathbf{I}_m-\mathbf{RD}_r^{-1}\mathbf{R}^H\mathbf{C}, \tag{3.29}$$

$$\mathbf{P}_i^Q=\mathbf{I}_n-\mathbf{KQD}_q^{-1}\mathbf{Q}^H, \tag{3.30}$$

ahol az i index azt jelöli, hogy az $\mathbf{R}$, ill. $\mathbf{Q}$ mátrixban i oszlopvektor található.

3.3. TÉTEL. Az $i+1$-edik vektorok a következő projektor-összefüggésekkel állíthatók elő:

$$\mathbf{P}_i^R\mathbf{Au}_{i+1}/\mathbf{v}_{i+1}^H\mathbf{Au}_{i+1}=\delta_i\mathbf{r}_i+\mathbf{Au}_{i+1}/\mathbf{v}_{i+1}^H\mathbf{Au}_{i+1}=\lambda_{i+1}\mathbf{r}_{i+1}, \tag{3.31}$$

$$\mathbf{v}_{i+1}^H\mathbf{AP}_i^Q/\mathbf{v}_{i+1}^H\mathbf{Au}_{i+1}=\varepsilon_i\mathbf{q}_i^H+\mathbf{v}_{i+1}^H\mathbf{A}/\mathbf{v}_{i+1}^H\mathbf{Au}_{i+1}=\varphi_{i+1}\mathbf{q}_{i+1}^H, \tag{3.32}$$

$$\mathbf{v}_{i+1}^H = \begin{cases} \lambda_h \mathbf{v}_h^H \mathbf{P}_i^R, & \text{ha} \quad \delta_i = 1 \quad \text{és} \quad \mathbf{l}_i^T = \sum_{j=h}^{i} \lambda_j \mathbf{e}_j^T, \\ \|\mathbf{A}\mathbf{u}_{i+1}\|_C^{-\alpha} \mathbf{u}_{i+1}^H \mathbf{A}^H \mathbf{C} \mathbf{P}_i^R, & \alpha \in \{1, 2\}, \quad \text{ha} \quad \delta_i = 0, \end{cases} \tag{3.33}$$

$$\mathbf{u}_{i+1} = \begin{cases} \varphi_k \mathbf{P}_i^Q \mathbf{u}_k, & \text{ha} \quad \varepsilon_i = 1 \quad \text{és} \quad \mathbf{f}_i = \sum_{j=k}^{i} \varphi_j \mathbf{e}_j, \\ \|\mathbf{A}^H \mathbf{v}_{i+1}\|_K^{-\alpha} \mathbf{P}_i^Q \mathbf{K} \mathbf{A}^H \mathbf{v}_{i+1}, & \alpha \in \{1, 2\}, \quad \text{ha} \quad \varepsilon_i = 0. \end{cases} \tag{3.34}$$

Bizonyítás. Több lépésben fog történni.

1. A $\mathbf{P}_i^R$ és $\mathbf{P}_i^Q$ (3.29)—(3.30)-ban bevezetett projektorok (3.1) és (3.2) egyenleteivel átírhatók a

$$\mathbf{P}_i^R = \mathbf{I}_m - \mathbf{A}\mathbf{U}\mathbf{D}_a^{-1}\mathbf{V}^H + \delta_i \mathbf{r}_i \mathbf{v}_{i+1}^H \tag{3.35}$$

$$\mathbf{P}_i^Q = \mathbf{I}_n - \mathbf{U}\mathbf{D}_a^{-1}\mathbf{V}^H\mathbf{A} + \varepsilon_i \mathbf{u}_{i+1}\mathbf{q}_i^H \tag{3.36}$$

alakba. Tegyük fel, hogy $\varepsilon_i = \delta_i = 1$. Ekkor a (3.28) összefüggésekkel adódik (3.31) és (3.32) első egyenlősége. A második egyenlőséget a bidiagonális építés (3.12) egyenleteiből kapjuk, amelyeket azzal a feltevéssel nyertünk, hogy $\mathbf{D}_a$ diagonálmátrix. Ha $\delta_i = 0$, akkor a bal diagonálépítés (3.13) formuláit ellenőrizve ismét kiadódik (3.31). Erre az esetre azonban még szükséges a $\mathbf{v}_{i+1}$, $\mathbf{u}_{i+1}$ vektorok **A**-konjugáltságát igazolni. Oldallépéskor a bidiagonális építés vektorait kell átrendezni, s a (3.19)—(3.22) formulákat nyomon követve (3.31) ismét teljesül.

2. A $\delta_i = \varepsilon_i = 1$ esetre az előző cikk 1.1. tételéből következik, hogy az $\mathbf{r}_j$ vektorok **C**-ortogonálisak, a $\mathbf{q}_j$ vektorok **K**-ortogonálisak, minthogy (3.31) és (3.32) alapján azok a (3.29)—(3.30) projektorokkal készülnek, amelyek most **C**- és **K**-ortogonális vektorokból, mint speciális konjugált irányokból vannak felépítve.

3. Befejezzük a 3.2. tétel igazolását a bal oldali diagonálépítés esetére. Legyen $\delta_i = 0$ és $\varepsilon_i = 1$. Ekkor az $\mathbf{u}_{i+1}$ vektor a bidiagonális építés (3.11) formulájából számítható és (3.28) alapján $\mathbf{V}^H\mathbf{A}\mathbf{u}_{i+1} = \mathbf{0}$ igaz. A $\mathbf{P}_i^R$ projektor az $\mathbf{A}\mathbf{u}_{i+1}$ vektort most helybenhagyja, emiatt az $\mathbf{A}\mathbf{u}_{i+1}$ vektor az előző pontban mondottak szerint **C**-ortogonális az $\mathbf{r}_j$, $j = 1, 2, \ldots i$ vektorokra. Így a (3.13)-ban választott $\mathbf{r}_{i+1}$ vektor a kívánalmaknak megfelel. De akkor a **C**-ortogonaitás miatt (3.13) és (3.1)-ből helyettesítéssel adódik a

$$\mathbf{v}_{i+1}^H \mathbf{A}\mathbf{U} = \mathbf{r}_{i+1}^H \mathbf{C}\mathbf{R}\mathbf{L}_i\mathbf{D}_a / \|\mathbf{r}_{i+1}\|_C^2 = 0$$

összefüggés, amiből látható, hogy a 3.2. tétel, s így (3.31)—(3.32) minden esetben igaz lesz.

4. A (3.33) formulát ellenőrzéssel igazoljuk. Ha $\delta_i = 1$, akkor a (3.35) alakot felhasználva kapjuk:

$$\lambda_h \mathbf{v}_h^H \mathbf{P}_i^R = \lambda_h \mathbf{v}_h^H \mathbf{r}_i \mathbf{v}_{i+1}^H.$$

Így már csak azt kell igazolnunk, hogy $\lambda_h \mathbf{v}_h^H \mathbf{r}_i = 1$. Ehhez szorozzuk be (3.1) második összefüggését a $\mathbf{V}^H$ mátrixszal, ezzel a

$$\mathbf{V}^H\mathbf{R}\mathbf{L}_i = \mathbf{I}_i = \mathbf{L}_i\mathbf{V}^H\mathbf{R} \tag{3.37}$$

egyenlőségekre jutunk, ahonnan

$$1 = \mathbf{e}_i^T \mathbf{L}_i \mathbf{V}^H \mathbf{R} \mathbf{e}_i = \mathbf{l}_i^T \mathbf{V}^H \mathbf{r}_i = \sum_{j=h}^{i} \lambda_j \mathbf{v}_j^H \mathbf{r}_i.$$

Az $\mathbf{L}_i$ mátrix (3.27) alakjából (3.1) segítségével kiolvasható, hogy

$$\mathbf{v}_j^H = -\mathbf{r}_{j-1}^H \mathbf{C}/\|\mathbf{r}_{j-1}\|_C^2, \quad j = h+1, \ldots, i, \tag{3.38}$$

így a C-ortogonalitás folytán csak a $\mathbf{v}_h$ vektor ad járulékot a szummába, ami 1-gyel egyenlő.

Ha $\delta_i = 0$, akkor (3.33) második formulájának ellenőrzése közvetlenül megtehető, ha (3.13), (3.14) alapján tekintetbe vesszük, hogy $\mathbf{r}_{i+1}$ egyirányú az $\mathbf{A}\mathbf{u}_{i+1}$ vektorral, s a (3.29) projektor a $\mathbf{v}_{i+1}^H$ vektorral egyirányú $\mathbf{r}_{i+1}^H\mathbf{C}$ vektort helybenhagyja.

Ezzel a 3.3. tétel igazolását befejeztük. ∎

A következő tétel az építési módszerek sikerességének eldöntésekor lesz a segítségünkre.

3.4. TÉTEL. A bal oldali oldallépéskor fennáll a

$$\varphi_i \|\mathbf{q}_i\|_K^2 \mathbf{v}_{i+1}^H \mathbf{A}\mathbf{u}_{i+1} = \tilde{\lambda}_i^{-2} (\mathbf{u}_i^H \mathbf{A}^H \tilde{\mathbf{v}}_i)^{-1} \|\mathbf{A}^H \tilde{\mathbf{v}}_{i+1}\|_K^2 + \tilde{\lambda}_i^{-1} \|\tilde{\mathbf{q}}_i\|_K^2 \tilde{\varphi}_i \tilde{\mathbf{v}}_{i+1}^H \mathbf{A}\tilde{\mathbf{u}}_{i+1}, \tag{3.39}$$

a jobb oldali oldallépéskor pedig

$$\lambda_i \|\mathbf{r}_i\|_C^2 \mathbf{v}_{i+1}^H \mathbf{A}\mathbf{u}_{i+1} = \tilde{\varphi}_i^{-2} (\tilde{\mathbf{u}}_i^H \mathbf{A}^H \mathbf{v}_i)^{-1} \|\mathbf{A}\tilde{\mathbf{u}}_{i+1}\|_C^2 + \tilde{\varphi}_i^{-1} \|\tilde{\mathbf{r}}_i\|_C^2 \tilde{\lambda}_i \tilde{\mathbf{v}}_{i+1}^H \mathbf{A}\tilde{\mathbf{u}}_{i+1} \tag{3.40}$$

összefüggés.

Bizonyítás. Csak a (3.39) formulát igazoljuk, amikor az oldallépést a bal oldalon tettük. Az A-konjugált tulajdonság és (3.11) második egyenlete miatt

$$\mathbf{v}_{i+1}^H \mathbf{A}\mathbf{u}_{i+1} = -\mathbf{v}_{i+1}^H \mathbf{A}\mathbf{K}\mathbf{q}_i/\|\mathbf{q}_i\|_K^2. \tag{3.41}$$

A $\mathbf{q}_i$ vektor (3.12) alapján $\mathbf{q}_{i-1}$ és $\mathbf{A}^H\mathbf{v}_i$ lineáris kombinációjaként fejezhető ki. De $\mathbf{q}_{i-1}$ a $\mathbf{K}$ mátrixszal szorzódik, és (3.11)-ből $\mathbf{K}\mathbf{q}_{i-1}$ az $\mathbf{u}_{i-1}$ és $\mathbf{u}_i$ vektor lineáris kombinációja. Emiatt $\mathbf{q}_{i-1}$ is kiesik:

$$\mathbf{v}_{i+1}^H \mathbf{A}\mathbf{u}_{i+1} = -\mathbf{v}_{i+1}^H \mathbf{A}\mathbf{K}\mathbf{A}^H \mathbf{v}_i/(\varphi_i \mathbf{u}_i^H \mathbf{A}^H \mathbf{v}_i \|\mathbf{q}_i\|_K^2).$$

Helyettesítsük ide $\mathbf{v}_i$ kifejezését (3.21)-ből és $\mathbf{v}_{i+1}$-et (3.20)-ból:

$$\mathbf{v}_{i+1}^H \mathbf{A}\mathbf{u}_{i+1} = -\frac{\|\mathbf{A}^H \tilde{\mathbf{v}}_{i+1}\|_K^2}{\tilde{\lambda}_i \varphi_i \|\mathbf{q}_i\|_K^2 \mathbf{u}_i^H \mathbf{A}^H \mathbf{v}_i} + \frac{\tilde{\mathbf{v}}_{i+1}^H \mathbf{A}\mathbf{K}\mathbf{A}^H \tilde{\mathbf{v}}_i}{\varphi_i \|\mathbf{q}_i\|_K^2 \mathbf{u}_i^H \mathbf{A}^H \mathbf{v}_i}.$$

A második tagban $\mathbf{A}^H\tilde{\mathbf{v}}_i$ (3.12)-ből $\tilde{\varphi}_i\tilde{\mathbf{q}}_i$ és $\mathbf{q}_{i-1}$ különbségével arányos. Helyettesítéskor $\mathbf{q}_{i-1}$ éppúgy kiesik, mint korábban. A $\mathbf{K}\tilde{\mathbf{q}}_i$ vektor (3.11) alapján az $\mathbf{u}_i$ és $\tilde{\mathbf{u}}_{i+1}$ vektorok lineáris kombinációja, s az A-konjugáltság miatt a fenti két helyettesítés után a $\mathbf{K}\mathbf{A}^H\tilde{\mathbf{v}}_i$ vektor helyett az $\tilde{\mathbf{u}}_{i+1}$ vektor áll, s átrendezés után kapjuk a (3.39) formulát. ∎

A (3.39)—(3.40) formulákból leolvasható, hogy a $\mathbf{v}_{i+1}^H \mathbf{A}\mathbf{u}_{i+1}$ és $\tilde{\mathbf{v}}_{i+1}^H \mathbf{A}\tilde{\mathbf{u}}_{i+1}$ belső szorzatoknak cs..k az egyike lehet zérus, amennyiben $\|\mathbf{A}^H\tilde{\mathbf{v}}_{i+1}\|_K \neq 0$, vagy $\|\mathbf{A}\tilde{\mathbf{u}}_{i+1}\|_C \neq 0$. Ezekből következik a 3.1. tétel 3) állítása. A 3.1. tétel 1) és 2) állításait a most igazolt tételek segítségével egyszerű ellenőrizni, s ezzel a 3.1. tétel biznyításának is a végéhez érkeztünk. ∎

3.7. A rekurziós mátrixok inverze

A következőkben megvizsgáljuk, hogyan állítható elő általában a fordított rekurzióhoz tartozó **L** rekurziós mátrix inverze. Ennek ismeretére a rekurziók vizsgálatakor szükség lehet. Például egy hatodrendű mátrix

$$\mathbf{L}_6 = \left[\begin{array}{cc:ccc:c} \lambda_1 & -1 & & & & \\ & \lambda_2 & -1 & & & \\ \hdashline & & 0 & -1 & & \\ & & 0 & 0 & -1 & \\ & & \lambda_3 & \lambda_4 & \lambda_5 & -1 \\ \hdashline & & & & & \lambda_6 \end{array}\right]. \tag{3.42}$$

Innen meggyőződhetünk arról, hogy **L** egy blokk-felső háromszögmátrix, ahol a diagonális blokkok felső bidiagonális mátrixok, vagy elemi *Frobenius mátrixok*. Az egyszerűsége miatt nem foglalkozunk azzal az esettel, amikor a diagonálblokk egységmátrix. Jelöljük a k-adrendű bidiagonális blokkot $\mathbf{B}_k$-val, a *Frobenius blokkot* pedig $\mathbf{\Phi}_k$-val. A k-adrendű elemi nilpotens mátrix

$$\mathbf{N}_k = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & 0 & \ddots & \\ & & & \ddots & 1 \\ & & & & 0 \end{bmatrix} \tag{3.43}$$

segítségével a k-adrendű bidiagonális és *Frobenius blokk* a

$$\mathbf{B}_k = \mathbf{D}_k - \mathbf{N}_k, \quad \text{ill.} \quad \mathbf{\Phi}_k = \mathbf{e}_k \mathbf{a}^T - \mathbf{N}_k$$

alakban írható fel, ahol $\mathbf{D}_k$ diagonálmátrix, $\mathbf{e}_k \mathbf{a}^T$ pedig egy k-adrendű diád.

Az inverz felépítését most is megkönnyíti a felső háromszögmátrixokra érvényes partícionált inverz:

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{0} & \mathbf{A}_{22}^{-1} \end{bmatrix}. \tag{3.44}$$

Ha $\mathbf{D}_k = \langle \lambda_1, \lambda_2, \ldots, \lambda_k \rangle$, akkor a $\mathbf{B}_k$ bidiagonális mátrix inverzére (3.44) felhasználásával egyszerűen adódik

$$(\mathbf{B}_k^{-1})_{ij} = \begin{cases} \prod_{l=i}^{j} \lambda_l^{-1}, & i \leqq j, \\ 0, & i > j. \end{cases} \tag{3.45}$$

A $\mathbf{\Phi}_k$ *Frobenius mátrix* inverze ugyancsak egyszerűen adható meg. Legyen $\mathbf{a}^T = [\lambda_1 \lambda_2 \ldots \lambda_k]$, ekkor az $\mathbf{e}_k^T \mathbf{N}_k = 0$ és $\mathbf{N}_k^T \mathbf{N}_k = \mathbf{I}_k - \mathbf{e}_1 \mathbf{e}_1^T$ összefüggések segítségével ellenőrizhető, hogy

$$\mathbf{\Phi}_k^{-1} = \mathbf{e}_1 \mathbf{b}^T - \mathbf{N}_k^T, \quad \text{ahol} \quad \mathbf{b}^T = (\mathbf{a}^T \mathbf{N}_k^T + \mathbf{e}_k^T)/\mathbf{a}^T \mathbf{e}_1 = \lambda_1^{-1}[\lambda_2 \ \lambda_3 \ldots \lambda_k, \lambda_1]. \tag{3.46}$$

A (3.27) alaknál már láttuk, hogy a *Frobenius blokk* az oldallépésből úgy áll elő, hogy a vezető helyen levő φ_1 elem nemzérus. Ha a diagonál-blokkok inverze ismert, akkor (3.44) felhasználásával a teljes $\mathbf{L}$ mátrix inverze is elkészíthető, amelynek általános elemét (3.45)-höz hasonlóan felírhatjuk, ha bevezetjük a vezető és belső elem fogalmát. Az ortogonális rekurzióhoz tartozó $\mathbf{L}$ mátrixban a λ_i elem vezető elem, ha az az $\mathbf{L}$ mátrix főátlójában helyezkedik el. Egyébként λ_i belső elem. Vezessük be az alábbi jelöléseket:

$$\nu_i = \begin{cases} \lambda_i, & \text{ha } \lambda_i \text{ vezető elem,} \\ 1, & \text{ha } \lambda_i \text{ belső elem,} \end{cases} \tag{3.47}$$

$$\alpha_i = \begin{cases} 1, & \text{ha } \lambda_i \text{ vezető elem,} \\ 0, & \text{ha } \lambda_i \text{ belső elem,} \end{cases} \tag{3.48}$$

$$\beta_i = \begin{cases} \lambda_{i+1}, & \text{ha } \lambda_{i+1} \text{ belső elem,} \\ 1, & \text{egyébként.} \end{cases} \tag{3.49}$$

Ekkor $\mathbf{L}$ inverze a következőképpen adható meg:

$$(\mathbf{L}^{-1})_{ij} = \begin{cases} \alpha_i \beta_j \prod_{l=i}^{j} \nu_l^{-1}, & i \leqq j, \\ \alpha_i - 1, & i = j+1, \\ 0, & \text{egyébként.} \end{cases} \tag{3.50}$$

A formula (3.44) alapján látható be. Csak azt kell észrevennünk, hogy a $-\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}$ minormátrix egy olyan diád, amely $\mathbf{A}_{11}^{-1}$ utolsó oszlopvektorából és $\mathbf{A}_{22}^{-1}$ első sorvektorából épül fel. A formula egyaránt érvényes egy bidiagonális blokk és egy *Frobenius mátrix* inverzére.

Az $\mathbf{F}$ mátrix inverze az $\mathbf{L}$ mátrixra kapott eredményekből transzponálással nyerhető.

3.8. Példa

Megismételjük a 2.7. példát a (3.11)—(3.12) rekurziós összefüggésekkel. A $\mathbf{v}_1$ és $\mathbf{u}_1$ vektort most is ugyanannak választjuk, mint a 2.7. példában. Az

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & -1 \\ 2 & 0 & 1 \\ 3 & 1 & 1 \end{bmatrix}$$

mátrixszal és a $\mathbf{v}_1^T = [0\ 0\ 1]$, $\mathbf{u}_1^T = [1\ 0\ 0]$ induló vektorokkal a következőket kapjuk:

$$\mathbf{r}_1 = \frac{1}{3}\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \mathbf{v}_2 = \frac{1}{14}\begin{bmatrix} -3 \\ -6 \\ 5 \end{bmatrix}, \quad \mathbf{r}_2 = \frac{1}{2}\begin{bmatrix} -4 \\ -1 \\ 2 \end{bmatrix}, \quad \mathbf{v}_3 = \frac{1}{6}\begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix},$$

$$\mathbf{q}_1 = \frac{1}{3}\begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{u}_2 = \frac{1}{11}\begin{bmatrix} 2 \\ -3 \\ -3 \end{bmatrix}, \quad \mathbf{q}_2 = \frac{1}{2}\begin{bmatrix} 2 \\ -3 \\ -3 \end{bmatrix}, \quad \mathbf{u}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Az $\mathbf{u}_3$ vektor most is ugyanúgy zérusnak adódott, mint a 2.7. példában. Szembetűnő még az is, hogy a $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ vektorok ugyanolyan irányúak, mint először a 2.7. példában, az $\mathbf{r}_j, \mathbf{q}_j$ vektorok azonban különbözőek.

A zérus $\mathbf{u}_3$ vektor megszüntetésére most is két lehetőségünk van. A diagonál-folytatásnak megfelelő (3.16b) szerint

$$\mathbf{q}_3 = \mathbf{A}^T\mathbf{v}_3 = \frac{1}{3}\begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}, \quad \mathbf{u}_3 = \mathbf{q}_3\|\mathbf{q}_3\|^{-2} = \frac{3}{2}\begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \quad \text{és} \quad \mathbf{r}_3 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix},$$

ahol a még hiányzó $\mathbf{r}_3$ vektort (3.12)-ből határoztuk meg. Ennek a folytatásnak megfelelő $\mathbf{F}$ és $\mathbf{L}$ mátrixok:

$$\mathbf{F} = \begin{bmatrix} 1 & & \\ -1 & 1 & \\ & & 1 \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} 1 & -1 & \\ & 1 & -1 \\ & & 1 \end{bmatrix}. \tag{3.51}$$

A zérus vektor megszüntetésének másik módja a bal oldali oldallépés az $i=2$ szinten. Az új $\mathbf{v}_2$ vektort (3.21)-ből határozzuk meg, $\mathbf{r}_1$ és $\mathbf{r}_2$ felcserélődik, s a $\mathbf{v}_3$ vektor változatlan marad. Végeredményben a következő vektor-együttesre jutunk:

$$\mathbf{r}_1 = \frac{1}{2}\begin{bmatrix} -4 \\ -1 \\ 2 \end{bmatrix}, \quad \mathbf{v}_2 = \frac{1}{21}\begin{bmatrix} 8 \\ 2 \\ -4 \end{bmatrix}, \quad \mathbf{r}_2 = \frac{1}{3}\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \mathbf{v}_3 = \frac{1}{6}\begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}, \quad \mathbf{r}_3 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix},$$

$$\mathbf{q}_1 = \frac{1}{3}\begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{u}_2 = \frac{1}{11}\begin{bmatrix} 2 \\ -3 \\ -3 \end{bmatrix}, \quad \mathbf{q}_2 = \frac{1}{9}\begin{bmatrix} 9 \\ 25 \\ -52 \end{bmatrix}, \quad \mathbf{u}_3 = \frac{7}{310}\begin{bmatrix} 7 \\ -15 \\ -6 \end{bmatrix}, \quad \mathbf{q}_3 = \frac{1}{7}\begin{bmatrix} 7 \\ -15 \\ -6 \end{bmatrix}.$$

A bal oldali oldallépéshez az

$$\mathbf{F} = \begin{bmatrix} 1 & & \\ -1 & 1 & \\ & -1 & 1 \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} 0 & -1 & \\ 1 & 1 & -1 \\ & & 1 \end{bmatrix}$$

mátrixok tartoznak.

Mint látható, a diagonál-folytatás mellett is ugyanazokat a $\mathbf{v}_j$, $\mathbf{u}_j$ vektorokat kapjuk, mint a 2.7. példánál. Azonban az oldallépés itt más vektorokat eredményezett.

E cikk folytatásában látni fogjuk, hogy a direkt és fordított rekurziók akkor állítanak elő egyirányú $\mathbf{A}$-konjugált párokat, ha nem alkalmaztunk egyetlen oldallépést sem.

IRODALOM

[1] HEGEDŰS, Cs. J., „Általános Hestenes—Stiefel típusú konjugált irány rekurziók. I. A direkt rekurzió", *Alkalmazott Matematikai Lapok*, **13** (1987—88) 83—96.

(Beérkezett: 1987. szeptember 14.)

HEGEDŰS CSABA J.
MTA KÖZPONTI FIZIKAI KUTATÓ INTÉZET
1525 BUDAPEST, PF. 49.

GENERAL RECURSIONS OF HESTENES—STIEFEL TYPE FOR CONJUGATE DIRECTIONS. II. THE REVERSE RECURSION

Cs. J. HEGEDŰS

A general recursion is given for generating A-conjugate vectors, where A is an $m \times n$ complex matrix. The knowledge of four vectors is necessary when producing a next pair. Features are: Forming $\mathbf{A}^H\mathbf{A}$, thus squaring the condition number is not needed. The generation ends if the vectors of the last pair are in the zero spaces of $\mathbf{A}^H$ and A. The method is reversely related to the previously given direct recursion, whose vectors can be generated here in reverse order. The next vectors are shown to be produced by ortogonal projections, in contrast to the direct recursion, where skew projections, composed from A-conjugate vectors do the same job. From this sheme new methods can be derived for solving linear systems and other standard problems of linear algebra.

Alkalmazott Matematikai Lapok **13** (1987—88) 373—389

MÓDSZEREK ÉS ALGORITMUSOK RITKA SZIMMETRIKUS MÁTRIXOKRA

GERGELY JÓZSEF

Budapest

Az [1] dolgozatban ritka mátrixokra vonatkozó hat feladat megoldását ismertettem és közöltem azok FORTRAN programjait. A dolgozat folytatásaképpen most szimmetrikus ritka mátrixokra vonatkozó feladatokat vizsgálok. Először a normálmátrixszal, mint a szimmetrikus mátrixok egy fontos speciális típusával foglalkozom. Azután a ritka szimmetrikus lineáris egyenletrendszerek numerikus megoldásának legfontosabb módszereit tárgyalom. Végül néhány megjegyzést és összefoglaló értékelést adok.

A dolgozatban a vizsgált módszereknek csak rövid leírását adom, részletesebben megtalálhatók azok a numerikus szakirodalomban ([5], [6]). A módszerek számítógépeken való megoldási algoritmusait ALGOL terminológiájú PASCAL nyelven fogalmazom meg.

1. Ritka szimmetrikus mátrixok

A dolgozatban olyan négyzetes

$$\mathbf{S} = \{s_{ij}\}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n$$

mátrixokat vizsgálok, amelyek egyrészt szimmetrikusak, vagyis $s_{ij}=s_{ji}$, másrészt sok bennük a nulla (zéró). A nem nulla (nem zéró) elemek jelölésére az „*NZ*'' rövidítést használom. Ezek mátrixbeli elrendezéséről előre semmiféle szabályosságot nem tételezek fel.

Ritka szimmetrikus mátrixok számítógépen való tárolásához és kezeléséhez a következő módot célszerű használni: csak a főátló és a felette levő *NZ* elemeket tároljuk sorfolytonosan (ami ugyanazt jelenti, mint a főátló és az alatta levő elemek tárolása oszlopfolytonosan). A tároláshoz három egydimenziós tömböt célszerű használni:

*S*1 (.), a *NZ* elemek tömbje;

JS (.), a *NZ* elemek oszlopindexeinek tömbje;

IS (*I*), $I=1, \ldots, n+1$, pointertömb. *IS*(*I*) mutatja az *I*-edik sor előző *NZ* elemének helyét az *S*1, illetve annak oszlop indexét a *JS* tömbben. $IS(n+1)=IS(n)+1$. A dolgozatban több módszert vizsgálok majd az

$$\mathbf{Sx} = \mathbf{b} \tag{1.1}$$

lineáris egyenletrendszer megoldására. Az egyértelmű megoldhatóság kedvéért mindig felteszem, hogy a mátrix nem szinguláris. Ez eleve teljesül, ha azt teszem fel, vagy bizonyítom, hogy a szimmetrikus mátrix pozitív definit. Ez azt jelenti, hogy tetszőleges $\mathbf{x}\neq\mathbf{0}$ vektorral képzett $(\mathbf{x}, \mathbf{Sx})>0$. (Ha csak az $(\mathbf{x}, \mathbf{Sx})\geqq 0$ áll fenn, akkor **S** pozitív szemidefinit.)

Az algoritmusokat PASCAL nyelven fogalmazom meg. Azonban csak az algoritmusok számítási részével foglalkozom. Az algoritmusokból akkor lesznek használható programok, ha kiegészítjük azokat a megfelelő deklaráló utasításokkal és adatkezelő input/output utasításokkal. Felhívom azonban a figyelmet arra, hogy a vizsgált algoritmusokat megvalósító FORTRAN programok a szerzőnél többnyire megtalálhatók.

2. Normálmátrixok

A szimmetrikus pozitív definit mátrixok egy gyakori speciális esete a normálmátrix, amik az

$$\mathbf{S} = \mathbf{A}^T\mathbf{P}\mathbf{A} \tag{2.1}$$

mátrix szorzattal képződnek, ahol

A $m \times n$-es (esetünkben ritka) téglalap mátrix,
P szimmetrikus pozitív definit súlymátrix.

A dolgozatban az $m > n$ esettel foglalkozok.
Normálmátrixokhoz többek közt mérési eredmények legkisebb négyzetek módszerével való kiértékelésénél jutunk [5]. Ebben az esetben az **A** mátrix az

$$\mathbf{A}\mathbf{x} = \mathbf{l} \tag{2.2}$$

mérési egyenlet „alakmátrixa" vagy átviteli mátrixa,
x a mért objektumot leíró paraméterek,
l a mért mennyiségek vektora. A legkisebb négyzetek módszerénél a (2.2) egyenletből képezett

$$\mathbf{v} = \mathbf{A}\mathbf{x} - \mathbf{l}$$

javítások vektoráról követeljük meg, hogy az azokból képzett

$$F(\mathbf{x}) = \mathbf{v}^T\mathbf{P}\mathbf{v}$$

kvadratikus funkcionál legyen minimális. A minimum szükséges feltételéből vezethető le az

$$\mathbf{S}\mathbf{x} = \mathbf{b}$$

normálegyenlet-rendszer, aminek **S** mátrixát a (2.1) képlettel számíthatjuk. A számítás egyszerűsödik, ha a mérések függetlenek. Ebben az esetben a **P** súlymátrix diagonális pozitív diagonális elemekkel, $\mathbf{b} = \mathbf{A}^T\mathbf{P}\mathbf{l}$.

A normálmátrix és a normálegyenlet-rendszer vizsgálatához először bebizonyítom, hogy a normálmátrix pozitív definit.

2.1 TÉTEL: Ha az **A** mátrix rangja n (az oszlopok száma), akkor az $\mathbf{S} = \mathbf{A}^T\mathbf{P}\mathbf{A}$ normálmátrix pozitív definit.

Bizonyítás. A pozitív definit **P** mátrix felbontható

$$\mathbf{P} = \mathbf{R}^T\mathbf{R}$$

szorzat alakban, ahol **R** nem szinguláris jobb felső háromszögmátrix. Ezután

$$\mathbf{S} = \mathbf{A}^T\mathbf{P}\mathbf{A} = (\mathbf{A}^T\mathbf{R}^T)(\mathbf{R}\mathbf{A})$$

írható, ami $\mathbf{Q}=\mathbf{R}\mathbf{A}$ bevezetésével a normálmátrix.

$$\mathbf{S} = \mathbf{Q}^T\mathbf{Q}$$

faktorizálásába megy át. Minthogy **R** nem szinguláris, ezért a **Q** mátrix rangja is n lesz, vagyis a **Q** mátrix oszlopvektorai $\mathbf{q}_i$, $i=1, \ldots, n$ is lineárisan független vektorrendszert alkotnak.

Tekintsük a $\mathbf{q}_i$, $i=1, \ldots, n$ lineárisan független vektorok $\mathbf{q}_i$, $i=1, \ldots, k$ alrendszerét $k=1, \ldots, n$-re. Képezzük a $\mathbf{q}_i$ vektorokból, mint oszlopvektorokból $i=1, \ldots, k$-ra a $\mathbf{Q}_{(k)}$ és az $\mathbf{S}_{(k)}=\mathbf{Q}_{(k)}^T\mathbf{Q}_k$ mátrixokat és azok determinánsait $\det(\mathbf{S}_{(k)})$.

A $\det(\mathbf{S}_{(k)})$ determinánsok a $\mathbf{q}_i$, $i=1, \ldots, k$ lineárisan független vektorokból képzett *Gram-féle determinánsok* pozitívek, ami azt jelenti, hogy az **S** normálmátrix bal felső minormátrixainak determinánsai pozitívak. Ez viszont azt jelenti, hogy az **S** normálmátrix pozitív definit.

A tételből következik, hogy az

$$\mathbf{S}\mathbf{x} = \mathbf{b}, \quad \mathbf{S} = \mathbf{A}^T\mathbf{P}\mathbf{A}, \quad \mathbf{b} = \mathbf{A}^T\mathbf{P}\mathbf{l} \tag{2.3}$$

normálegyenlet-rendszer egyértelműen megoldható.

Legyen a (2.3) egyenletrendszer megoldása

$$\mathbf{x}' = \mathbf{S}^{-1}\mathbf{b},$$

amit a (2.2) egyenletbe helyettesítve kapjuk a mérések „kiegyenlített" eredményeit, ami a legkisebb négyzetek elmélete szerint a mérések várható értékeinek vektora, vagyis

$$E(\mathbf{l}) = \mathbf{A}\mathbf{x}'.$$

A továbbiakban algoritmusokat ismertetek a normál mátrixok számítására. A (2.2) egyenletrendszer **A** együttható mátrixa általában ritka mátrix. Számítógépes tárolására használjuk a következő három egydimenziós tömböt:

*A*1 (.) a *NZ* elemek tömbje;

JA (.) a *NZ* elemek oszlopindexeinek tömbje;

IA (*I*) $I=1, \ldots, m+1$, pointertömb. *IA* (*I*) mutatja az *I*-edik sor első *NZ* elemének helyét az *A*1, illetve a *JA* tömbben. (Az $\mathbf{A}^T$ transzponált mátrixot természetesen nem tároljuk külön). $IA(m+1)=IA(m)+1$.

Az **S** mátrixot a (2.1) képlet szerint számítjuk. Az **S** tároláshoz az 1. pontban ismertetett *S*1, *JS* és *IS* tömböket használjuk. A **P** súlymátrixot diagonálisnak tételezzük fel és a *P*1 tömbben tároljuk. Mint azt már az előzőekben leírtam **A** $m\times n$-es, az **S** $n\times n$-es ritka mátrixok (**S** szimmetrikus).

A normálmátrix számításához legegyszerűbbnek tűnő algoritmus a következőképpen fogalmazható meg.

(PASCAL terminológiát használva):

Algoritmus ($N1$)

```
begin s:=1;
c:=0;
IS(1):=1;
for i:=1 to n do
begin for h:=i to n do
  begin for j:=1 to m do
    begin j1:=IA(j); j2:=IA(j+1)-1;
      for k:=j1 to j2 do
       if JA(k)=i then
        for q:=j1 to j2 do
        if JA(q)=h then
        c:=c+A1(k)*A1(q)*P1(j)
    end;
        S1(s):=c; JS(s)=h; s:=s+1
    end;
    IS(i+1)=s
end.
```

Az $N1$ algoritmus a szimmetrikus ritka mátrix tárolásának megfelelően alakítja ki a normál mátrixot. A normál mátrixban az NZ elemek száma és elhelyezkedése előre nem ismeretes, ezért az algoritmus úgy működik, hogy soronként haladva számítja ki és helyezi el az NZ elemeket. Azonban ez a természetesnek tűnő megoldás a számítási időben nagyon gazdaságtalan, hiszen többször végig kell vizsgálni az **A** mátrixot.

A számítási idő csökkentése érdekében megadunk egy másik, $N2$ algoritmust. Az $N2$ algoritmus bonyolultabbaz $N1$ algoritmusnál. Az $N2$ algoritmus lényeges eleme, hogy előre kiszámítja a normál mátrix NZ elemeinek indexeit egy IT tömbbe, majd ezeket átszervezi az S struktúrájának megfelelően a JS és az IS tömbökbe. Ezután a már feltöltött JS és IS tömbök segítségével számítja a normál mátrixot.

Algoritmus ($N2$)

```
begin for i:=1 to n do s(i):=1;
for i:=1 to m do
```

```
  begin i1:=IA(i); i2:=IA(i+1)-1;
    for j:=i1 to i2 do
      begin j1:=JA(j);
        for h:=i1 to i2 do
          begin h1:=JA(h); if j1<=h1 then
            for l:=1 to s(j1) do
              if IT(j1,l)<>h1 then
                begin
                  s(j1):=s(j1)+1;
                  IT(j1,s(j1)):=h1
                end
          end
      end
  end;
        k:=1;
        for i:=1 to n do
          begin IS(i):=k; for j:=1 to s(i) do begin
            JS(k):=IT(i,j); k:=k+1 end
          end; IS(n+1):=k;
          for i:=1 to k-1 do S1(i):=0;
          for i:=1 to m do
            begin i1:=IA(i); i2:=IA(i+1)-1;
              for j:=i1 to i2 do
                begin j1:=JA(j)
                  for k:=i1 to i2 do
                  begin k1:=JA(k);
                    if j1<=k1 then
                    begin l1:=IS(j1); l2:=IS(j1+1)-1;
                      for l:=l1 to l2 do
                      S1(l)=S1(l)+A1(j1)*A1(k1)*P1(i)
                    end
                  end
                end
            end
end.
```

A (2.2) mérési egyenletben szereplő **A** mátrixról feltételeztük, hogy ritka mátrix, de az *NZ* elemek számát, se azok elhelyezkedését nem kellett előre vizsgálnunk. A mátrix struktúráját a mérés jellege vagy konkrét végrehajtása szabja meg. Még kevesebbet tudunk az **S** normálmátrix *NZ* elemeinek elhelyezkedéséről. Viszont az *N*2 algoritmusban használt *IT* indextömbnek legalább annyi oszlopnak kell lenni, mint a normál mátrix egy sorában maximálisan előforduló *NZ* elemek száma, amit viszont előre nem ismerünk. Ezért az *IT* tömböt sor túlcsordulási lehetőséggel kell felépíteni és így kevesebb oszlopot kell használni az *IT* tömbben. Ennek a megoldásával az algoritmus nem foglalkozik.

3. Ritka szimmetrikus egyenletrendszerek megoldása

Nagyméretű lineáris egyenletrendszerek megoldásánál még nagyobb számítógépeknél is problémát jelent, hogy a számítógép operatív memóriája kicsinek bizonyulhat az egyenletrendszer egyidejű tárolására. A probléma kisebb, ha a mátrix szimmetrikus, hiszen elég tárolni csak a félmátrixot, de sokkal többet segíthet az, hogyha a mátrix ritka is. Ebben az esetben csak az *NZ* mátrixelemeket kell tárolni és csak azokkal kell műveleteket végezni. Azonban a programok sokkal bonyolultabbak lesznek, ezért külön megfontolandó kérdés, hogy milyen méret felett és milyen fokú ritkaság esetén célszerű a használatuk.

A következőkben a ritka, szimmetrikus mátrixú

$$\mathbf{Sx} = \mathbf{b} \tag{3.1}$$

lineáris egyenletrendszer megoldódására használatos módszereket tekintem át. Először a direkt módszereket vizsgálom, amelyek faktorizálják az **S** mátrixot, majd visszahelyettesítéssel oldják meg az egyenletrendszert. Azután a sávredukcióval foglalkozom, amivel az **S** mátrixból minél kisebb szélességű szalagmátrixot hozunk létre és a szalagmátrixos módszerekkel oldjuk meg az egyenletrendszert. Ezután a konjugált gradiens és az iterációs módszerek használatát vizsgálom, ritka szimmetrikus mátrixok esetén.

4. Direkt eljárások

A direkt eljárások többsége *Gauss elimináción* alapszik. Ez azt jelenti, hogy ritka mátrixos technikával először faktorizálják az **S** mátrixot a következő alakban:

$$\mathbf{S} = \mathbf{L}\mathbf{D}^{-1}\mathbf{L}^T (= \mathbf{L}\mathbf{U},\ \mathbf{U} = \mathbf{D}^{-1}\mathbf{L}^T), \tag{4.1}$$

ahol **L** az alsó háromszög mátrix, $\mathbf{L}^T$ a transzponáltja, **D** a diagonális elemek mátrixa.

Ezután az $\mathbf{Sx}=\mathbf{b}$ egyenletrendszerből $\mathbf{LUx}=\mathbf{b}$ egyenletrendszer lesz, ami az

$$\mathbf{Ly} = \mathbf{b},\quad \mathbf{Ux} = \mathbf{y} \tag{4.2}$$

egyenletrendszerekre esik szét. (4.2) egyenletrendszerekből visszahelyettesítésekkel kapjuk az

$$\mathbf{y} = \mathbf{L}^{-1}\mathbf{b},\quad \text{majd az}\quad \mathbf{x} = \mathbf{U}^{-1}\mathbf{y}$$

megoldást. (Tárolni faktorizálás után csak az **L** és a **D** mátrixot kell). Ezen az elven működnek az IBM Matematikai Szubrutin Könyvtár ritka szimmetrikus egyenletrendszert megoldó programjai, [2]. A könyvtár tíz programot tartalmaz ritka szimmetrikus mátrixokra. Ezek közül kettő (MSNS és MSNO) elvégzi a (4.1) faktorizálást, kettő (MBSS és MBSO visszahelyettesítéssel megoldja a (4.2) egyenleteket. Két programmal (MPRS és MPRO) a megoldás pontosságát javíthatjuk iteratív módon.

A (4.1) faktorizálás is két lépésre bontható. Első lépésként csak szimbolikus faktorizálást végzünk, ami azt jelenti, hogy kialakítjuk azokat az indexeket, amelyekre szükség lesz az eliminációközbeni telítődés *(fill-in)* miatt. Erre a célra két program szolgál (MSSS és MSSO). A szimbolikus faktorizálás után kerülhet sor a tényleges faktorizálásra, amit ugyancsak két program lát el (MSSN és MNSO)

Az IBM könyvtári programok használatához az **S** mátrix előzőekben leírt tárolási módja szükséges. A szubrutinok paraméter-listáin az **S** tárolásához az *IA*, a *JA* és az *A* paraméterek szolgálnak. A szubrutinok úgy működnek, hogy az *IA*, *JA* és *A* tömbökben adott **S** mátrixot átrendezik a dinamikusan bővülő *IU*, *JU* és *U* tömbökbe. A paraméterlistán szereplő *N* a mátrix rendszáma, *IBOT* a *JU* és *U* tömbök maximális mérete, amit előre meg kell becsülni. Ha ezt a telítettség miatt túllépnénk, hibajelzéssel a számítás leáll. *DI* a diagonális elemek tömbje, *IK* a faktorizálás közben szükséges szorzások száma, *IER* a hibajelzés kódját tartalmazza.

A következő hibajelzések lehetségesek: a mátrix szinguláris, a főelem = 0, túlcsordul az *IU* és az *U* tömb, a mátrix megadása rossz, a főelem nagyon kicsi, a hiba pontosan nem határozható meg stb. Ezeken kívül szerepel még néhány paraméter, amikre a faktorizálás közben van szükség. A (4.2) egyenleteket megoldó szubrutinok természetesen használják a jobb oldal **B** és a megoldás **X** vektorát.

Ha az **S** pozitív definit, a (4.1) egyenletrendszer a fenti módszerrel mindig megoldható. Ha a pozitív definitség nincs biztosítva, akkor főelem választásra van szükség. Az IBM szubrutinok főelemválasztással működnek.

A másik leggyakrabban használt direkt eljárás a *Cholesky (vagy négyzetgyökös) módszer*. A módszer lényege, hogy a szimmetrikus **S** mátrixot úgy faktorizálja

$$\mathbf{S} = \mathbf{R}^T\mathbf{R} \tag{4.3}$$

alakra, hogy az **S** mátrix s_{ij} elemei és az **R** mátrix r_{ij} elemei közt teljesüljenek az

$$s_{ij} = \sum_{k=1}^{i} r_{ki}r_{kj}, \quad i < j \quad \text{és az} \quad s_{ii} = \sum_{k=1}^{i} r_{ki}^2 \tag{4.4}$$

összefüggések, amik a (4.3) faktorizálásból adódnak. **R** jobb felső háromszög mátrix. A (4.4) összefüggésből a következő képletekkel számíthatók ki az r_{ij} mátrixelemek:

$$r_{11} = \sqrt{s_{11}}, \quad r_{1j} = \frac{s_{1j}}{r_{11}}, \quad j = 1, \ldots, n,$$

$$r_{ii} = \sqrt{s_{ii} - \sum_{k=1}^{i-1} r_{ki}^2}, \quad i > 1, \quad r_{ij} = \frac{s_{ij} - \sum_{k=1}^{i-1} r_{ki}r_{kj}}{r_{ii}}, \tag{4.5}$$

$$r_{ij} = 0, \quad \text{ha} \quad i > j.$$

Ha a (4.5) képletekkel kiszámítottuk az $\mathbf{R}$ mátrix elemeit, akkor az $\mathbf{Sx}=\mathbf{b}$ egyenletrendszert a (4.2)-höz hasonló

$$\mathbf{R}^T\mathbf{y} = \mathbf{b}, \quad \mathbf{Rx} = \mathbf{y} \tag{4.6}$$

egyenletrendszerekből visszahelyettesítésekkel oldjuk meg.

A *Cholesky-módszer* stabilitási tulajdonságai jobbak, mint a *Gauss elimináción* alapuló megoldásoké. A mátrix ritkasága is jól kihasználható a megoldás közben. A *Gauss eliminációhoz* hasonló telítődésre itt is számolni kell. Pozitív definit szimmetrikus S mátrix esetén a (3.1) egyenletrendszer mindig megoldható. Ha S csak pozitív szemidefinit, akkor már szingularitást is kell vizsgálni. Ha az S negatív definit (nem szinguláris) a (3.1) egyenletrendszer ugyancsak megoldható, de akkor már módosítani kell az algoritmust.

A *Cholesky-módszer* programját is célszerű úgy elkészíteni, hogy számolás közben a mátrixot három tömbben tároljuk: *IU* egy pointer tömb, az *U*1 tömbben tároljuk az *NZ* elemeket, *JU*-ban pedig az *U*1 elemekhez tartozó oszlopindexeket. *IU*(*I*) tömbelem mutatja az S mátrix *I*-edik sorának főátlóban álló elemének helyét az *U*1 tömbben (tegyük fel, hogy a főátlóban *NZ* elemek vannak, ha a mátrix pozitív definit, akkor ez teljesül).

Az *U*1 és a *JU* tömbök dinamikusan bővülő tömbök. Minden elemhez tartozik egy pointer, ami megmutatja a vizsgált elemet ugyanabban a mátrixsorban követő elem helyét az *U*1 (illetve a *JU*) tömbben. A pointerek is egy tömbben helyezhetők el, aminek annyi eleme lesz, mint a *JU*, (ill. *U*1) tömböknek. Így a mátrix egy sorának *NZ* elemeit a pointer tömbbel vezérelten egy láncolattal érhetjük el. Ha egy sorban egy új *NZ* elem lép fel, akkor a pointer tömbelem kijelöli a helyét az *U*1 (illetve a *JU*) tömbökben és az új pointer átvált az *U*1 (illetve *JU*) tömb első még le nem foglalt helyére.

Rátérek a *Cholesky-módszer* algoritmusának ismertetésére. Azonban az *U*1 és *JU* tömbök kezelésének bonyolult algoritmusát nem részletezem. Az algoritmusban a következő megjegyzésekkel hivatkozom rájuk:

$$(* \; r_{ij} \neq 0 \quad \text{elem keresés} \; *)$$

$$(* \; r_{ij} \neq 0 \quad \text{elem elhelyezés} \; *).$$

Az első azt jelenti, hogy a program megvizsgálja, hogy a mátrix *i*-edik sorának *j*-edik oszlopában álló elem zéró, vagy *NZ*. A második pedig azt, hogy a mátrix *i*-edik sorának *j*-edik oszlopába kell elhelyezni az $r_{ij} \neq 0$ elemet. Az *S*1, *IS* és *JS* tömbökben tárolt $n \times n$-es *S* mátrix *NZ* elemeinek száma kezdetben legyen *n*1. A *Cholesky faktorizáció* egy algoritmusa:

Algoritmus (*C*):

```
begin
for i:=1 to n do IU(i):=IS(i);
for i:=1 to n1 do begin JU(i):=JS(i); U1(i):=S1(i) end;
U1(1):=SQRT(U1(1));
```

```
for i:=IU(1) to IU(2)-1 do U1(i):=U1(i)/U1(1);
for i:=1 to n do
begin f:=0;
  for k:=1 to i-1 do
  begin (*r_ki≠0 elem keresés, ha r_ki≠0, U1(K1)=r_ki*);
    f:=f+U1(K1)*U1(K1)
  end
  U1(IU(i)):=SQRT (U1(IU(i))-f);
  for j:=i to n do
  begin f:=0; for k:=1 to i-1 do
    begin (*r_kj≠0 elem keresés, ha r_kj≠0, U1(KJ)=r_kj*);
          (*r_ki≠0 elem keresés, ha r_ki≠0, U1(K1)=r_ki*);
      f:=f+U1(KJ)*U1(KI)
    end;
          RIJ:=(RIJ-f)/U1(IU(I));
          (*RIJ a mátrix i-edik sorának j-edik eleme*);
          (* RIJ elhelyezése az U1 tömbbe *);
          (* RIJ indexének elhelyezése a JU tömbbe *)
  end
end
end.
```

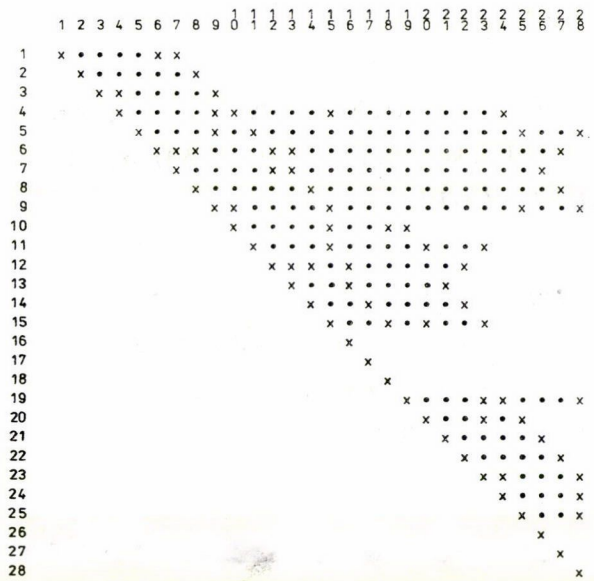
5. Sávosítás

Nagyméretű ritka, szimmetrikus mátrixú lineáris egyenletrendszerek megoldásának egyik leghatékonyabb módszere azon alapszik, hogy először az egyenletrendszer mátrixát a lehető legszűkebb szélességű szalagmátrixszá alakítjuk át és az egyenletrendszert mint szalagmátrixos egyenletrendszert oldjuk meg. Az ilyen egyenletrendszerek megoldására nagyon jó módszerek és jól működő programok találhatók. [1], [2]. A szimmetrikus mátrixok sávosítására, illetve a sávszélesség csökkentésére is jól kidolgozott elméletek, módszerek és programok találhatók. A módszerek még csak vázlatos ismertetésére sem térek ki, hiszen ez egy külön dolgozatnak is elég nagy téma. A [3] dolgozatban a sávosítás elméleti kérdéseinek tisztázása és a módszerek nagy választéka található. Jelen dolgozatban az a célom, hogy minél kevesebb elmé-

leti meggondolással a gyakorlatban könnyen használható, egyszerű módszert adjak a sávszélesség csökkentésére.

A módszer nem biztosítja a minimális sávszélességet, de könnyen végrehajtható és sávszélesség-csökkentést eredményez.

A módszert egy számpéldán szemléltetve mutatom be. A számpéldát a [3] dolgozat 21. mellékletéből vettem. A példa egy rúdszerkezet statikai számításához használt merevségi mátrix sávszélesség redukciójának vizsgálatával kapcsolatos. Az idézett melléklet 3. ábrájából indulok ki. Az ott szemléltetett 28×28-as szimmetrikus ritka mátrix *NZ* elemeit *X* jellel, a zérus elemeket ponttal jelölöm. Csak a főátló és felette levő félmátrixot tüntetem fel az 1. ábrán. Könnyen kiszámítható, hogy a mátrix szalagszélessége = 23.



1. ábra

A [3] dolgozat jelöléseit és elnevezéseit használva az ábra alapján egy szintstruktúrát képezek az 1-ből kiindulva a következőképpen. Az 1-es után veszem az első sor *NZ* elemeinek oszlopindexeit

1, 6, 7

Ezután a 6., majd a 7. oszlopban lefelé haladva a főátlóig, majd onnan jobbra a sor végére leírom a haladás mentén talált *X*-ek sor, illetve oszlop-indexeit, ha azok még nem kerültek a listába. Ily módon a következő számokat kapom:

8, 12, 13, 27, 26

Ezután a 8, 12, 13, 27, majd 26-ik oszlopban fentről lefelé haladva a főátlóig, majd onnan jobbra a sorok végéig írom ki a talált X-ek indexeit (ha azok még nem szerepeltek). Ezek a következők lesznek:

2, 14, 16, 22, 21

Most ezek oszlopai, illetve sorai mentén megismétlem az előző lépéseket. Ezt addig ismétlem, míg az összes sorszám listára nem kerül. Ha valahol megakad a fenti eljárás még mielőtt az összes sorszám listára nem került (azért, mert elfogynak az induló oszlopok számai), akkor újra kezdem a legkisebb, még listára nem került sorszámmal. (Ebben az esetben a mátrix diagonális blokk mátrixokká bomlik, ami esetünkben a 14-edik sor után be is következik.) Az eljárás befejezésével a következő új sorrendet kaptam:

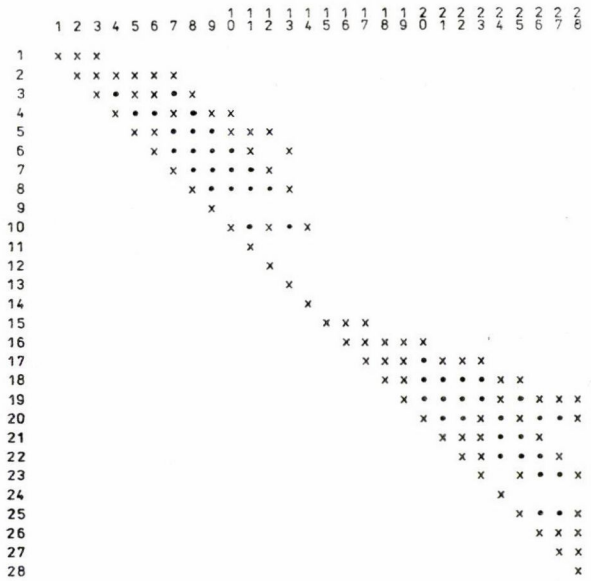
1, 6, 7, 8, 12, 13, 27, 26, 2, 14, 16, 22, 21, 17,

3, 4, 9, 10, 15, 24, 5, 25, 28, 18, 19, 11, 20, 23.

Átrendezem a mátrixot az új sorrendnek megfelelően. Vagyis az eddigi 6. sor, illetve oszlop kerüljön a 2. helyre, a 7. kerüljön a 3. helyre stb. Az átrendezett mátrixot a 2. ábra szemlélteti.

A 2. ábrából látható, hogy a mátrix szalagmátrixszá alakult át, aminek szalagszélessége = 9.

Az ismertetett módszer valójában az E. H. CUTHILL és J. MCKEE módszerének egyszerűsített változata [4]. Ugyanis csak a mátrix fogalmait használtam és a mátrixhoz rendelhető gráf és annak fogalmait elhagytam a megfogalmazásban. A *Cuthill—McKee-módszer* egyik legismertebb módszer a sávszélesség csökkentésére. A módszer



2. ábra

nagyon könnyen programozható, hatékony és gyors. A módszer részletes elemzésével nem kívánok foglalkozni, de a közölt példán ; jellel elválasztva megmutatom a generált sorrend szintstruktúra szintjeit:

1, 6, 7; 8, 12, 13, 27, 26; 2, 14, 16, 22, 21; 17

3; 4, 9; 10, 15, 24, 5, 25, 28; 18, 19, 11, 20, 23.

A példán szemléletesen végrehajtott módszer általánosan is könnyen megfogalmazható:

a) Választunk egy sorszámot. Ez lesz a struktúra „gyökere", ami az első szint (a példa esetében ez 1 volt).

b) A választott sorszám oszlopában haladva a főátlóig, majd onnan a sor mentén a sor végéig megkeressük a mátrix NZ elemeit és listába írjuk azok sor, illetve oszlop indexeit. Ezek alkotják a második szintet és így tovább;

c) A k-adik szinthez tartozó sorszámok oszlopai mentén haladva, a főátlóig, majd sor szerint haladva a sorok végéig talált NZ elemek sor, illetve oszlop indexei alkotják a $k+1$-edik szintet.

Az így kialakult szintstruktúra szerint átrendezzük a mátrixot. Az átrendezéssel a mátrix sávszélessége nem nő, de általában csökken [3, 4].

6. Konjugált gradiens és iterációs módszerek

Ezeknek a módszereknek közös sajátossága, hogy a számítás döntő része mátrix-vektor szorzásokból, esetünkben ritka, szimmetrikus mátrixnak vektorral való szorzásából áll ($\mathbf{u}=\mathbf{S}\mathbf{v}$). Ezért először ennek a részfeladatnak az algoritmusát fogalmazom meg. Legyen az $n\times n$-es $\mathbf{S}$ szimmetrikus ritka mátrix az 1. pontban leírtaknak megfelelően adva az IS, JS és $S1$ tömbökben, a $\mathbf{v}$ és $\mathbf{u}$ vektorok pedig a V és U tömbökben $(V(i), U(i), i=1, \ldots, n)$. Az

$$\mathbf{u}=\mathbf{S}\mathbf{v}$$

mátrix-vektor szorzást végző algoritmust a következő eljárásban adom meg:

Algoritmus (V):

```
procedure P1(n, IS, JS, S1, U, V);
(* deklaráló rész *);
begin
for i:=1 to n do
begin
  U(i):=0;
  for j:=1 to i-1 do
```

```
    begin
      j1:=IS(j);
      j2:=IS(j+1)-1;
      for h:=j1 to j2 do
        if JS(h)=i then U(i):=U(i)+S1(h)*V(j)
    end;
      j1:=IS(i);
      j2:=IS(i+1)-1;
      for j:=j1 to j2 do
      begin
        k:=JS(j);
        U(i):=U(i)+S1(j)*V(k)
      end
end
end.
```

A mátrix-vektor szorzásra közölt eljárás látszólag bonyolult, azonban nagyon ritka mátrixok esetén számítási időben megtérül a bonyolultság.

A konjugált gradiens módszer elvi alapjaival nem kívánok foglalkozni, ez megtalálható legtöbb numerikus analízissel foglalkozó könyvben [5], [6]. Jelen dolgozatomban csak a módszer ritka mátrixok esetében való végrehajtását vizsgálom. Ennek részletes elemzése megtalálható a (7) könyvben.

Feltételezem, hogy S szimmetrikus pozitív definit mátrix. Ebben az esetben stabilitási probléma kevésbé jelentkezik és a konjugált gradiens módszert, mint véges iterációs módszert tekintem, vagyis n ismeretlenes egyenletrendszer megoldására n lépést hajtok végre.

Tetszőlegesen választott $\mathbf{x}_1$ vektorból kiindulva

$$\mathbf{v}_1 = \mathbf{r}_1 = \mathbf{b} - \mathbf{S}\mathbf{x}_1 \tag{6.1}$$

és ezután $i=1, \ldots, n$-re számítjuk a következő képleteket, [5]:

$$\begin{aligned} \alpha_i &= \mathbf{v}_i^T \mathbf{r}_i / \mathbf{v}_i^T \mathbf{S}\mathbf{v}_i \\ \mathbf{x}_{i+1} &= \mathbf{x}_i + \alpha_i \mathbf{v}_i \\ \mathbf{r}_{i+1} &= \mathbf{r}_i - \alpha_i \mathbf{S}\mathbf{v}_i \\ \beta_i &= -\mathbf{v}_i^T \mathbf{S}\mathbf{r}_{i+1} / \mathbf{v}_i^T \mathbf{S}\mathbf{v}_i \\ \mathbf{v}_{i+1} &= \mathbf{r}_{i+1} + \beta_i \mathbf{v}_i . \end{aligned} \tag{6.2}$$

A (6.1) és (6.2) képletek számítását megvalósító algoritmus programja a *P*1 eljárást felhasználva a következő:

```
Algoritmus(K):
begin P1 (n, IS, JS, S, U, X1);
  for i:=1 to n do
  begin
        V(i):=B(i)-U(i);
        R(i):=V(i)
  end;
  for i:=1 to n do
begin P1 (n, IS, JS, S, U, V);
    a2:=0;
  for i:=1 to n do
    begin
    a2:=a2+V(i)*U(i)
    end;
    a1:=0;
    for j:=1 to n do a1:=a1+V(j)*R(j);
    a:=a1/a2;
    for j:=1 to n do
    begin X(j):=X(j)+a*V(j);
          R(j):=R(j)-a*U(j)
    end; P1 (n, IS, JS, S, U, R);
    a3:=0;
    for j:=1 to n do
    begin
        a3:=a3+V(j)*U(j);
    end; b:=-a3/a2
    for j:=1 to n do V(j):=R(j)+b*V(j)
  end
end.
```

Az $\mathbf{Sx}=\mathbf{b}$ egyenletrendszer iterációs megoldásai közül egyedül a *Gauss—Seidel iterációval* foglalkozom. Ismeretes a következő tétel [5], [6]:

6.1. TÉTEL. Ha az S szimmetrikus mátrix pozitív definit, akkor a *Gauss—Seidel iteráció* a kezdeti vektortól függetlenül konvergens.

Induljunk ki tetszőleges $\mathbf{x}^{(1)}=(x_1^{(1)}, \ldots, x_n^{(1)})$ vektorból. Az algoritmust a *Gauss—Seidel iterációt* megvalósító

$$x_i^{(k+1)} = -\frac{1}{s_{ii}}\left[\sum_{j=1}^{i-1} s_{ij}x_j^{(k+1)} + \sum_{j=i+1}^{n} s_{ij}x_j^{(k)} - b_i\right] \quad (i=1, \ldots, n) \tag{6.3}$$

képlet segítségével írom fel. Az S szimmetrikus együttható mátrixot az 1. pontban ismertetett *S1*, *IS* és *JS* tömbökben, az $\mathbf{x}$ és $\mathbf{b}$ vektorokat az *X* és *B* tömbökben tároljuk. Az iteráció az $\mathbf{x}=\mathbf{b}$ közelítésből indul ki. A módszert megvalósító algoritmus a következő:

Algoritmus (GS):

```
begin
  c: for i:=1 to n do
  begin X(i):=-B(i);
    for j:=1 to i-1 do
    begin j1:=IS(j);
        j2:=IS(j+1)-1;
        for k:=j1 to j2 do
        if JS(k)=i then X(i):=X(i)+S1(k)*X(j)
    end;
    j1:=IS(i);
    j2:=IS(i+1)-1;
    for j:=j1 to j2 do X(i):=X(i)+S1(j)*X(JS(j));
    X(i):=-X(i)/S1(IS(i));
  end;
  if (* convergencia feltétel teljesül *) stop
  else go to c
end.
```

Az algoritmusban nem részleteztem az iteráció konvergenciájának teljesülését. Konkrét esetben valamilyen konvergencia kritérium teljesülését kell ellenőrizni.

Ebben a pontban az S ritka szimmetrikus mátrixú lineáris egyenletrendszer megoldását vizsgáltam. Abban az esetben, ha az S mátrix a 2. pontban tárgyalt normál

mátrix, az egyenletrendszer megoldásához nincs szükség az S tényleges előállítására. A (2.1) képlet felhasználásával az $\mathbf{Sx}=\mathbf{b}$ lineáris egyenletrendszer ugyanis felírható

$$\mathbf{A}^T\mathbf{PAx} = \mathbf{b} \tag{6.4}$$

alakban. A (6.4) egyenletrendszert megoldó konjugált gradiens módszer (6.1) és (6.2) képletei, valamint a *Gauss—Seidl-módszer* (6.3) képlete a következő alakba mennek át:

$$\mathbf{v}_1 = \mathbf{r}_1 = \mathbf{b}-\mathbf{A}^T\mathbf{PAx}_1 \tag{6.5}$$

$$\alpha_i = \mathbf{v}_i^T\mathbf{r}_i/\mathbf{v}_i^T\mathbf{A}^T\mathbf{PAv}_i$$

$$\mathbf{x}_{i+1} = \mathbf{x}_i+\alpha_i\mathbf{v}_i$$

$$\mathbf{r}_{i+1} = \mathbf{r}_i-\alpha_i\mathbf{A}^T\mathbf{PAv}_i \tag{6.6}$$

$$\beta_i = -\mathbf{v}_i^T\mathbf{A}^T\mathbf{PAr}_{i+1}/\mathbf{v}_i^T\mathbf{A}^T\mathbf{PAv}_i$$

$$\mathbf{v}_{i+1} = r_{i+1}+\beta_i\mathbf{v}_i$$

$$x_i^{(k+1)} = -\frac{1}{s_{ii}}\Big[\sum_{j=1}^{i-1}\mathbf{a}_i^T\mathbf{Pa}_j x_j^{(k+1)}+\sum_{j=i+1}^{n}\mathbf{a}_i^T\mathbf{Pa}_j x_j^{(k)}-\mathbf{b}_i\Big], \quad s_{ii} = \mathbf{a}_i^T\mathbf{Pa}_i \tag{6.7}$$

ahol $\mathbf{a}_i$ az A mátrix i-edik oszlopa. A (6.5), a (6.6) és a (6.7) képletekben kijelölt számítások csak az A mátrixot használjuk. Ha a konjugált gradiens módszert véges iterációs módszerként használom, vagy a *Gauss—Seidel módszerben* nincs n-nél több iterációs lépésre szükségünk, akkor a (6.5), (6.6) és (6.7) képletekkel számolva az eredményt a normál mátrix előállításával azonos nagyságrendű lépésekben megkaphatjuk, így használatuk előnyös lehet.

7. Összefoglalás

Összefoglalásul néhány megjegyzést szeretnék tenni. Először a ritka mátrix inverzének számítására térek ki. Mint ismeretes, ritka mátrix inverze lehet teljesen telt mátrix is. Ez természetesen igaz szimmetrikus esetben is. Ezért nem célszerű arra törekednünk, hogy nagyméretű ritka mátrix inverzét számítsuk ki. Ehelyett az inverz faktorizált alakját célszerű előállítani [8]. Ez a következőket jelenti. Ritkamátrixos technikával végezzük el a (4.1) faktorizálást *Gauss-eliminációval*, vagy a (4.3) faktorizálást *Cholesky módszerrel*. (4.1)-ben az **L**, (4.3)-ban az **R** háromszög mátrixok megőrzik a ritka mátrixos struktúrát (bár a telítettség miatt több NZ elemet tartalmaznak, mint az S mátrix). Ha a (4.1) faktorizálás után a (4.2) egyenletrendszerek helyett az

$$\mathbf{Ly}_i = \mathbf{e}_i, \quad \mathbf{Ux}_i = \mathbf{y}_i, \tag{7.1}$$

a (4.6) egyenletrendszerek helyett pedig az

$$\mathbf{R}^T\mathbf{y}_i = \mathbf{e}_i, \quad \mathbf{Rx}_i = \mathbf{y}_i \tag{7.2}$$

egyenletrendszereket oldjuk meg visszahelyettesítéssel, ahol $\mathbf{e}_i$ az i-edik egységvektor, akkor megoldásul az $\mathbf{S}^{-1}$ inverzmátrix i-edik oszlopát kapjuk. Vagyis az inverzet oszloponként állítjuk elő. Ily módon az inverz mátrix tetszőleges elemét kiszámíthatjuk a (7.1), vagy a (7.2) egyenletrendszerekből visszahelyettesítéssel.

A ritka szimmetrikus mátrixok sajátértékeinek számításával az [1] dolgozatban foglalkoztam. A *szimmetrikus Láncos módszerrel* az S mátrixot szimmetrikus háromátlós mátrixá transzformáltam, majd annak sajátértékeit számoltam implicit *QR módszerrel.*

A nagyméretű ritka szimmetrikus egyenletrendszerek megoldásához alkalmazandó módszer kiválasztásánál alapvető szempontok: a mátrix mérete, ritkasága, pozitív definitsége, a rendelkezésre álló számítógép operatív (belső) memóriájának nagysága és hogy a gépen milyen programok állnak rendelkezésre. Ha például a mátrix pozitív definitsége nincs biztosítva, olyan programot kell választani, amelyek főelem választással dolgoznak. Ilyenek az IBM szubrutinkönyvtár programjai. Ha a számítógépünk memóriája szűkös a feladat méreteihez képest, akkor célszerű az iterációs módszereket használni. Ugyanis azok alkalmazásával a mátrix változatlan marad a számítás teljes menetében (telítődés nem lép fel) és így a memória elég lehet a tárolásra. Sőt ebben az esetben olyan megoldás is alkalmazható, hogy a mátrixot külső tárban helyezzük el, hiszen annak csak soronkénti használatára van szükség.

A sávszélesség redukálásán alapuló módszerek feltétlen előnyösek a többi módszerekkel szemben, hiszen egyrészt itt a mátrix csak átrendeződik, telítettségi probléma nem lép fel és a keletkezett szalagmátrixos egyenletrendszer megoldására nagyon jó, viszonylag egyszerű, memória-takarékos módszerek és programok készíthetők, vagy találhatók [1], [2].

IRODALOM

[1] GERGELY, J., „Módszerek és programok ritka mátrixokra", *AML* **6** (1980) 407—442.
[2] *IBM Subroutine Library Mathematics* (System /360, /370, 1130, 1800)
[3] ARANY, I., „Nagyméretű, ritka, szimmetrikus mátrixok hatékony számítógépes kezelése", *AML* **11** (1985) 1—90.
[4] CUTHILL, E. H. and MCKEE, J., "Reducing the bandwidth of sparse symmetric matrices", *Proc. 24th National Conference. ACM* (1969) 157—172.
[5] A. RALSTON, *Bevezetés a numerikus analízisbe* (Műszaki Kiadó, Budapest 1969).
[6] N. SZ. BAHVALOV, *A gépi matematikai numerikus módszerei* (Műszaki Kiadó, Budapest, 1977).
[7] REID, J. S., "Large sparse sets of linear equation", *Procedings of the Oxford conference,* Academic Press, London, 1971.
[8] TEWARSON, R. P., *Sparse Matrices* (Academic Press, New York and London, 1973.)

(Beérkezett: 1987. március 3.)
(Átdolgozva beérkezett: 1987. december 3.)

DR. GERGELY JÓZSEF
FÖLDMÉRÉSI INTÉZET
1051 BUDAPEST, GUSZEV U. 19.

METHODS AND ALGORITHMS FOR SPARSE SYMMETRICAL MATRICES

J. GERGELY

Methods and programs were discussed in paper [1] for sparse matrices. As a continuation of that, methods and algorithms are discussed for sparse symmetrical matrices in this paper. First I deal with the normal matrixes which are an important, special case of it. Then the more important numerical methods for the solution of the system of the sparse symetrical linear equations are studied. Brief formulations of the methods and algorithms in PASCAL language are written donon.

Alkalmazott Matematikai Lapok **13** (1987—88) 391—396

KÖNYVISMERTETÉS

Melvyn W. Jeter: *Mathematical programming* — An introduction to optimization. A könyv egy általános, nem túl részletes bevezető a matematikai programozás néhány alaptémájába. Az alább részletezendő tíz fejezetre oszlik a könyv, melyekben a lineáris programozás, a hálózati folyamok és a nemlineáris programozás néhány alapvető problémáját, módszerét tárgyalja.

Az első, bevezető fejezet, mint a legtöbb hasonló könyv, néhány példát ad matematikai programozási feladatokra. A szerző érzékelteti a paraméterezés és a post optimális analízis fontosságát. Végül történeti megjegyzések következnek.

A második fejezet az elemi lineáris algebra egy rövid összefoglalója. Ismertetésre kerülnek a lineáris algebra mindazon eredményei, amelyek szükségesek a lineáris programozás elméletének és algoritmusainak kidolgozásához.

A harmadik fejezet a primál szimplex és a kétfázisú szimplex módszert tartalmazza. A szimplex módszert tábla formában tárgyalja, de kitér a geometriai interpretációra is.

A negyedik fejezet a lineáris programozás dualitás elméletét és további technikáit tartalmazza. Előbb a dualitáselmélet, a post optimális analízis, majd a duál szimpla módszer és a lineáris komplementaritási feladat és annak megoldására alkalmas Lemke-féle komplementáris pivotálási algoritmus következik.

Az ötödik fejezet a szimplex módszer numerikus megvalósításával kapcsolatos problémákat tárgyalja. Így tartalmazza a módosított szimplex módszert, a bázisinverz szorzat formában való előállítását, a bázisinverz eliminációs formáját. Ismerteti a primál-duál algoritmust, a paraméterezés problémáját, a degeneráció és ciklizálás lehetőségét. Végül érinti a dekompozíció lehetőségét is.

A hatodik fejezet nagyon rövid betekintést ad a hálózati folyamok elméletébe. Néhány alapfeladaton (maximális folyam, transzshipment) keresztül bemutat néhány alapvető technikát (cimkézés, szimplex adaptáció, primál-duál algoritmus adaptációja).

Az utolsó négy fejezet a nemlineáris programozás néhány alapvető technikáját tárgyalja.

A hetedik fejezet egy rövid áttekintést ad a konvex analízis elméletéből. Itt kerülnek ismertetésre a későbbi fejezetekben alapvető szerepet játszó tételek (konvex függvények tulajdonságai, differenciálhatóság stb.).

A nyolcadik fejezet a konvex programozás talán legfontosabb eredményeit, az optimalitási feltételeket tartalmazza. A szerző tárgyalja a feltétel nélküli, az egyenlőséges és egyenlőtlenséges feltételek, valamint nemnegatív változók esetén az optimalitási feltételeket. A fejezet a *Kuhn—Tucker-tétel* ismertetésével zárul.

A kilencedik fejezet a feltétel nélküli optimalizálás technikáival foglalkozik. Egy dimenzióban ismerteti a legfontosabb keresési technikákat (pl. *Fibonacci,* aranymetszés, görbe illesztés). Röviden a többdimenziós optimalizálást is érinti.

Az utolsó fejezetben a feltételes optimalizálás büntetőfüggvényes módszereivel foglalkozik. Itt a *Barrier függvényeket* és a kvadratikus büntetőfüggvényeket tárgyalja.

A könyv szerzője a könyv céljaként azt jelöli meg, hogy egy tankönyvet szeretne adni viszonylag csekély (egy félév lineáris algebra és kevés analízis) matematikai háttérrel rendelkező hallgatók kezébe. Ezt a célt és a benne megszabott feltételeket betartja. A könyv tárgyalása korrekt. A nehezen és hosszadalmasnak vélt bizonyításokat gyakran elhagyja a szerző, csak a tételek ismertetésére szorítkozik. Az algoritmusok megértését jól segíti a sok végigszámolt mintafeladat és a könyv bőséges feladat anyaga. Az egyes fejezetek végén kielégítő irodalomjegyzék található.

Véleményem szerint a kitűzött célnak a könyv megfelel, akár hazánkban hasznos segédeszköz lehet rövid, áttekintő kurzusok esetén (pl. műszaki egyetemen, főiskolán), de mindenképpen kevés például matematikusok oktatásában. Amiben még eredeti céljának sem felel meg a mű, az a könyv

anyaga, és az egyes témákra fordított terjedelem. Még egy rövid áttekintő könyvben is legalább érinteni kellett volna például a lineáris programozás újabb eredményeit. Gondolok itt például a *Bland szabályra*, mely a ciklizálás elkerülését teszi lehetővé. Említeni kellett volna az elméleti hatékonyság (exponenciális, polinomiális algoritmusok) problémáját is. A könyv anyagából sajnos teljesen kimaradtak a diszkrét programozás alapfeladatai, melyeket minden alapkurzusnak tartalmaznia kell.

Összefoglalva, a M. W. JETER könyve hasznos segédkönyv, bőséges példaanyaggal, de fontos, lényeges témakörök maradnak említés nélkül, így a könyv olvasása, illetve használása esetén erre feltétlenül figyelmet kell fordítani.

Terlaky Tamás

Többváltozós statisztikai analízis. Szerkesztette Móri F. Tamás és Székely J. Gábor. Műszaki Könyvkiadó, Budapest, 1986. 393 oldal. A többváltozós statisztikai analízis a matematikai statisztika egyik legfontosabb fejezetévé nőtte ki magát, és az utóbbi évtizedekben — elsősorban a számítógépek alkalmazása nyomán — egyre nagyobb gyakorlati jelentőségre tett szert. Ezért üdvözlendő, hogy a témakörről átfogó jellegű munka jelent meg magyar nyelven.

A jelen könyv 14 matematikus közös munkája, és elsősorban matematikus olvasóközönség számára készült. A *Bolyai János Társulatban* megtartott továbbképző tanfolyam anyagát adják közre a szerzők. Kár, hogy a szerkesztők nem említik az *Eötvös Loránd Tudományegyetem Valószínűségszámítási Tanszékének* a szerepét, hiszen itt folyt először rendszeres oktatás a témakörben hazánkban.

A könyv matematikai szempontból szigorú tárgyalásmódban adja elő az anyagot, pontos definíciókat és tételeket mond ki, és — a terjedelem adta korlátok között — bizonyításokat is közöl. A szerzők és a szerkesztők igen nagy munkával a többváltozós analízis szinte minden fontos fejezetét feldolgozták, így művük igen gazdag tényanyagot tár az olvasó elé.

A könyv lényegében a témakörről szóló standard szakkönyvekhez (l. [1], [2], [3]) hasonló felosztásban tárgyalja a többváltozós statisztikát. A szerzők érdeklődésének megfelelően bizonyos részek (pl. a változók közötti kapcsolat vizsgálata, a kontingenciatáblázatok elemzése) bővebben kerültek kifejtésre, míg más részek (pl. az osztályozások) a szokásosnál rövidebben szerepelnek. A többdimenziós normális eloszláson alapuló fő irányvonal természetesen egybeesik az általánosan elterjedt tárgyalásmóddal. A szerzők nemcsak a klasszikus elméletre, hanem az újabb eredményekre is utalnak. A témában tovább elmélyülni kívánók számára pedig bőséges irodalomjegyzéket állítottak össze.

A jelen mű értékes segédkönyv a többdimenziós statisztikával megismerkedni és azt alkalmazni kívánók számára, a felhasználhatóságának azonban megvannak a korlátai. Mivel nem nyújtja a tárgyalt témakör egységes és szisztematikus felépítését és gyakran kevés magyarázatot ad, tankönyvként nem ajánlható. Az alkalmazásokat illusztráló példákat, esettanulmányokat alig tartalmaz a könyv, ezért a gyakorlati felhasználók számára is csak részben szolgálhat iránymutatóul.

A mű egyes fejezetei eltérő stílusúak és mélységűek, így azokat külön-külön ismertetjük. Az egyes szerzők egyéni felfogását tiszteletben tartjuk, néhány hiányosság mellett azonban nem mehetünk el szó nélkül. Ezek elsősorban bizonyos részek nem megfelelően kiérlelt volta, pontatlanságok és helyenként zavaróan sok sajtóhiba.

I. *A többdimenziós normális eloszlás* (TUSNÁDY GÁBOR)

A szerző először az n-dimenziós standard normális eloszlást ismerteti, majd ennek segítségével vezeti be a többdimenziós normális eloszlást. Ezután a feltételes eloszlások és a kvadratikus alakok tárgyalása következik. Kiszámolja a normális eloszlás sűrűségfüggvényét és momentumgeneráló függvényét is. A többdimenziós normalitásvizsgálatról annyit jegyez meg, hogy szemben a valós esettel, itt nincs jól használható módszer.

Az anyag tárgyalásmódja logikus és pedagógiai szempontból is jó. A fejezet hiányossága, hogy a széles körben használt karakterisztikus függvényeket meg sem említi. A függetlenség és a korrelálatlanság kapcsolatát sem emeli ki. A tárgyalás tömör, a bizonyítások vázlatosak, sőt néhány esetben elnagyoltak.

II. *A Wishart-mátrix* (TUSNÁDY GÁBOR)

A szerző megadja a p-dimenziós, n szabadsági fokú, Σ kovarianciájú *Wishart-mátrix* definícióját, és meghatározza a momentumgeneráló függvényét. A sűrűségfüggvényét a technikai nehézségek miatt először a $p=2$ esetben számolja ki, ezután tér rá az általános eset tárgyalására. A bétamátrixok ([3] szerint első típusú bétaeloszlások) definíciója és sűrűségfüggvényének meghatározása, valamint a *Wilks-eloszlás* definíciója zárja a fejezetet.

A fejezet első részében a *Wishart-mátrix* bevezetése világos, jól áttekinthető. A későbbi részekben fellelhető néhány következetlenség és több sajtóhiba.

III. *Véletlen mátrixok saját értékeinek eloszlása és aszimptotikus viselkedése* (MICHALETZKY GYÖRGY)

A fejezet első részében a *Wishart-* és a *bétamátrixok* saját értékeinek sűrűségfüggvényét határozza meg a szerző, a kimondott tételeket részletesen bizonyítja is. A második részben a *Wishart-mátrix* saját értékeinek aszimptotikus viselkedését írja le, többnyire már bizonyítás nélkül. Utal néhány újabb eredményre is.

IV. *A többdimenziós normális eloszlás várható érték vektorára és szórásmátrixára vonatkozó becslés és hipotézisvizsgálat* (MICHALETZKY GYÖRGY)

A szerző sok értékes információt gyűjt egybe a témakörről, ahol lehet bizonyításokat is közöl. A fejezet elején áttekinthetően leírja a várható érték és a szórás maximum-likelihood becslését. Az 1.3 megjegyzést sokkal pontosabban kellett volna kimondani, hisz gyakran kell rá hivatkozni. A megengedhető becslésekkel kapcsolatban részletesen tárgyalja a *Stein-problémát*. Hipotézisvizsgálatra a likelihood-hányados próbát alkalmazza. Nem említi azonban a témakörben használatos aszimptotikus sorfejtéseket ([3], 7. fejezet).

V. *A változók közötti kapcsolat vizsgálata* (RUDAS TAMÁS)

A szerző az alábbi négy csoportba sorolva mutatja be a kapcsolat mérésére szolgáló módszereket:

1. Normális eloszlású változók közötti kapcsolat mérésére a korrelációs együtthatót, valamint a parciális és a többszörös korrelációt használja. Ezen mennyiségek maximum-likelihood becslésének az eloszlását [1] alapján írja le.

A további esetekben nemparaméteres módszereket tárgyal.

2. Két folytonos eloszlású változó kapcsolatának mérésére a korrelációs együtthatót, a *Kendall-féle* τ és a *Spearman-féle rangkorrelációs együtthatót* használja. A rangkorrelációs módszereket egy bővebb elméletbe is beilleszti.

3. Olyan csoportosított mérési eredmények esetén, amikor a csoportok között rendezés van, a *Goodman—Kruskal-féle* γ-t alkalmazza.

4. Abban az esetben, amikor a csoportok között nincs rendezés, a χ^2-próbát, a *Yule-féle Q-* és *Y-*, valamint a *Csuprov-féle T-*, a *Cramér-féle C-* és a *Goodman—Kruskal-féle* (más néven *Guttmann-féle*) λ-statisztikát ajánlja.

A fejezet fő érdeme, hogy megpróbál bizonyítást adni a statisztikai kézikönyvek által bizonyítás nélkül közölt, általánosan ismert eredményekre. Azonban a rendelkezésre álló néhány oldalon ilyen sokrétű témakör tényleges kifejtésére nincs lehetőség. Kifejezetten alkalmazott témakörről lévén szó, nagy hátrányt jelent, hogy a szerző a bevezetett fogalmakhoz kevés magyarázatot fűz, ezek használatát nem ismerteti, mintapéldákat nem közöl. Mivel a fejezet megfogalmazása igen tömör, a benne szereplő számos elírás és sajtóhiba különösen zavaró.

VI. *A változók számának csökkentése: főkomponens- és faktoranalízis* (REJTŐ LÍDIA)

A főkomponens-analízis témaköréből a standard tankönyvekben (pl. [2]) megtalálható bevezető anyagot közli a szerző. Egyszerűen és érthetően írja le a főkomponensek alapvető tulajdonságait. Normális eloszlásból vett minta esetén tárgyalja a főkomponensekkel kapcsolatos problémákat: bebizonyítja az empirikus kovarianciamátrix saját értékeinek és saját vektorainak aszimptotikus viselkedéséről szóló tételt, ismerteti a saját értékek egyezését vizsgáló próbát.

A faktormodell definíciója szerepel a könyvben, majd ezzel kapcsolatban néhány egyszerű állítást is közöl (ezek között a 4.4. lemma bizonyítása nem teljes, l. [2] 276. old.). A faktormodell identifikációjáról felsorolt tételeket bizonyítás nélkül közli. Normális eloszlású alapsokaság esetén foglalkozik a faktormodellben szereplő mennyiségek maximum-likelihood becslésével, itt az átviteli mátrix becslését részletesen leírja. A faktorok magyarázatánál fontos forgatásokat csak megemlíti. A főkomponens- és faktoranalízis gyakorlati jelentőségét, alkalmazási módját nem részletezi.

VII. *Kanonikus korrelációanalízis* (LENGYEL TAMÁS)

A fejezet első részében példákkal indokolja a szerző a két változó közötti kapcsolat mérésének a jelentőségét. A többszörös korreláció segítségével világítja meg a kanonikus korreláció fogalmát. Sajnálatos, hogy a 2.1. definíció megfogalmazása pontatlan. Ezután a szerző rávilágít a kanonikus faktorok (más néven kanonikus korreláció vektorok [2]) és a korrelációs mátrix szinguláris felbontásának a kapcsolatára. Foglalkozik a kanonikus korrelációk optimális tulajdonságaival is.

A fejezet hasznos rövid összefoglalást nyújt a kanonikus korrelációanalízisről, gyakorlati alkalmazására azonban kevés útmutatást ad.

VIII. *Regressziós modellek* (SZÉKELY J. GÁBOR)

A regresszió fogalmának bevezetése után a nemparaméteres regresszióról közöl tételeket a szerző. A szokásos többváltozós regressziós modell esetén adja meg a legkisebb négyzetes becslést. Foglalkozik a változók számának csökkentésére alkalmas *stepwise-módszerrel* is. Ezután a *ridge-becslést* és általánosításait tárgyalja. A robusztus becslések rövid általános ismertetését követően a robusztus regressziót vizsgálja.

A szerző a témakör sok érdekes és fontos sajátosságára hívja fel a figyelmet rövid megjegyzések és bőséges irodalmi utalás formájában. Másrészt egyetlen kérdéskört sem fejt ki elég részletesen. így a regressziószámítással megismerkedni vágyók és a felhasználók ismét kevés útmutatást kapnak.

IX. *Becslés és hipotézisvizsgálat lineáris modellben. Szórásanalízis* (MICHALETZKY GYÖRGY)

A szerző definiálja a lineáris modellt, majd rátér a benne szereplő paraméterek maximum-likelihood becslésére. Vizsgálja a becslés explicit megadhatóságát és speciális esetben az eloszlását. Közli a *Gauss—Markov-tételt.* A regressziós paraméterre vonatkozóan a *Wilks-* és a *Roy-próbát* ismerteti. A lineáris modell általános vizsgálatából származó eredményeket a szórásanalízisben alkalmazza. Egyetlen faktor esetére a *Wilks-próbát* adja meg. Bonyolultabb szórásanalízisbeli feladatokat (pl. két faktor esetét) is beágyaz a lineáris modellbe.

X. *Osztályozási módszerek* (GULYÁS OTTÓ)

A fejezet első részében az osztályozás problémáját világosan veti fel a szerző. A döntésfüggvények általános elméletéből indul ki, megadja az optimális *Bayes-döntést,* majd a szemléletes kétcsoportos, ismert sűrűségfüggvényű esetet mutatja be.

Speciális eljárásokat ismertet: a lineáris és szakaszonként lineáris szeparálást, a legközelebbi társ módszert. A fejezet végén leírja a clusteranalízis feladatát, szól a dinamikus és a hierarchikus clusterszerkesztésről.

A fejezet egészét tekintve elmondható, hogy a problémák felvetése kristálytiszta, sok fontos részletre rávilágít a szerző, a konkrét eljárások is használhatók. A rövid tárgyalás eredménye, hogy

az osztályozás összetett problémakörét nem ismerteti, így klasszikus fogalmak (pl. *Fisher-féle diszkriminancia-függvény, Mahalanobis-távolság*) maradtak ki, ill. egyes részek nem a valóságos súlyuknak megfelelő arányban szerepelnek. Talán szerencsésebb lett volna az ismert művek ([2]) mintájára külön fejezetet szentelni a diszkriminancia-analízisnek is és a clusteranalízisnek is.

XI. *Kontingenciatáblázatok loglineáris elemzése* (RUDAS TAMÁS)

A polinomiális eloszlásról szóló rövid bevezetés után megadja a loglineáris modell definícióját. A két- és háromdimenziós táblázatok esetén leírja a függetlenség jellemzését speciális loglineáris modellek segítségével. A minimális diszkrimináló információ elvét és a maximum-likelihood módszert, valamint ezek kapcsolatát tárgyalja. Egy iterációs eljárást is ismertet az eloszlás maximum-likelihood becslésének kiszámítására. Bizonyos esetekre hipergráfok segítségével zárt képletet ad meg a maximum-likelihood becslésre. Egy modell helyességének vizsgálatára a χ^2- és a loglikelihood statisztikát használja.

A fejezet érdeme, hogy betekintést nyújt a statisztika olyan részterületébe, melyet a standard kézikönyvek többsége nem ismertet. A tárgyalásmód azonban nem alkalmas a témakörrel való első megismerkedésre. A sok fogalom, jelölés, részeredmény, utalás sűrűjéből nehezen bontakozik ki az olvasó számára a tulajdonképpeni cél. A loglineáris modellek alkalmazási módja nem nyer kellő megvilágítást. A szerző utal ugyan az V. és XII. fejezettel való kapcsolatra, az összedolgozottság azonban hiányzik.

XII. *A minimális diszkrimináló információ módszere; kontingenciatáblázatok elemzése* (CSISZÁR IMRE)

Először a d-dimenziós kontingenciatáblázatokat és a távolságminimalizáló módszerekkel megoldható statisztikai feladatokat ismerteti a szerző. Az információs divergenciát mint az eloszlások egy speciális f-eltérését vezeti be. Ezen fogalom segítségével tárgyalható a minimális diszkrimináló információ módszere (MDI) és a maximum-likelihood módszer is. Részletesen vizsgálja az információs divergencia tulajdonságait. A külső és a belső feltételekkel meghatározott feladatok MDI-módszerrel történő megoldásával foglalkozik.

A szerző elegánsan, logikus felépítésben tárgyalja az MDI-módszert. A fejezet önálló egységként olvasható, témája jelentősen eltér a könyv fő irányvonalától és vitatható alkamazott statisztikai jelentősége is.

XIII. *Többdimenziós skálázás* (TELEGDI LÁSZLÓ)

A többdimenziós skálázás feladatának körülhatárolása után, annak a klasszikus megoldását ismerteti. Bevezeti a távolság- és a hasonlóságmátrixot, majd leírja a klasszikus megoldás algoritmusát és ezen megoldás optimális tulajdonságait. A nem-metrikus algoritmusok közül a Shepard—Kruskal-féle eljárást ismerteti. Foglalkozik a többszörös többdimenziós skálázással is.

A tárgykörrel kapcsolatos alapvető tudnivalókat érthetően foglalja össze; lényegében a [2] műben megtalálható anyagot közli. Sajnálatos, hogy a [2]-ben található, a témakör megértését nagyban elősegítő, szemléletes példákhoz hasonlóakat nem mutat be.

XIV. *A szimultán döntés paraméteres módszerei* (KUN ANDREA)

Normális alapeloszlás esetén több várható érték egyidejű összehasonlítását vizsgálja a szerző. A bevezetésben érthetően mutat rá a szimultán hipotézisvizsgálati módszerek szükségességére. Először az egyváltozós eset történeti áttekintését adaja. Többek között *Newman, Tukey, Dunnett, Scheffé* és *Krishnaiah próbáit* említi. A többváltozós esetben *Roy és Krishnaiah módszerét* írja le. A regressziós modellekkel is foglalkozik. Végül röviden szól a szórások összehasonlításáról.

Sok eljárást ír le kevés magyarázattal és szemléletes példák nélkül.

XV. *A szimultán döntés nemparaméteres módszerei* (HAJTMAN BÉLA)

A szimultán és a nemparaméteres módszerekről szóló általános bevezető után a szerző rátér a szimultán összehasonlítás tárgyalására az egyváltozós esetben. A paraméteres módszerekről szóló XIV. fejezetben már ismertetett, TURKEY, SCHEFFÉ és DUNNETT nevéhez fűződő eljárások nemparaméteres változatát mutatja be. Végül ezen módszerek többváltozós esetre történő kiterjesztésével foglalkozik.

A szerző jól érthető módon vázolja fel a nemparaméteres módszerek általános elveit, az ismertetett eljárások egy részét konkrét példákkal is illusztrálja. A módszerek matematikai hátterének részletezésében nem mélyed el.

XVI. *A maximum-likelihood-módszer* (MICHALETZKY GYÖRGY ÉS SZÉKELY J. GÁBOR)

Független, azonos eloszlású minta esetén vizsgálja a fejezet a maximum-likelihood becslés tulajdonságait. Bizonyítást nyer a likelihood-egyenlet gyökének konzisztenciájáról és aszimptotikus normalitásáról szóló tétel. Több példát is közöl a maximum-likelihood becslés viselkedésének szemléltetésére. Végül a likelihood-hányados statisztika aszimptotikus tulajdonságairól bizonyít tételeket a szerzőpáros.

XVII. *A többváltozós statisztikai analízis számítógépes eljárásai* (GAUDI ISTVÁN)

A BMDP statisztikai programcsomagot ismerteti a szerző. Leírja a programcsomag részeit és haszhálatának módját. A többváltozós statisztikai eljárások közül clusteranalízisre, diszkriminancia-analízisre, illetve regresszióanalízisre szolgáló programot mutat be részletesen. Ezen programok alkalmazását egy-egy generált adathalmaz feldolgozásával és a futási eredmények közlésével szemlélteti.

Függelék: Lineáris algebrai segédeszközök (BOLLA MARIANN)

A többdimenziós analízisről szóló más művek mintájára ez a könyv is függelékben sorolja fel a nélkülözhetetlen mátrixelméleti ismereteket. Tömören és világosan foglalja össze a mátrixok felbontásáról, a bilineáris formákról, az általánosított inverzről szóló tudnivalókat.

Itt jegyezzük meg, hogy a könyv listát közöl a legfontosabb jelölésekről, kislexikonba gyűjti az alapvető statisztikai fogalmakat, név- és tárgymutatót azonban nem tartalmaz.

Összefoglalva megállapíthatjuk, hogy a matematikai statisztikával foglalkozók széles rétege, egyetemi hallgatók, oktatók és kutatók is haszonnal forgathatják a jelen művet, sőt az a felhasználók számára is értékes információkkal szolgálhat a módszerek matematikai hátterének megismerésében.

A bírálatban felhasznált irodalom:

[1] ANDERSON, T. W., *An Introduction to Multivariate Statistical Analysis* (Wiley, New York, 1958).
[2] MARDIA, K. V.; KENT, J. T. and BIBBY, J. M., *Multivariate Analysis* (Academic Press, New York, 1979).
[3] SRIVASTAVA, M. S. and KHATRI, C. G., *An Introduction to Multivariate Statistics* (Nort Holland, New York, 1979).

A debreceni Kossuth Lajos Tudományegyetem Alkalmazott Matematikai és Valószínűségszámítási Tanszék oktatói a fenti könyvet hasznosan alkalmazzák az oktatásban. Véleményüket összefoglalta

Fazekas István

A kiadásért felelős az Akadémiai Kiadó és Nyomda Vállalat főigazgatója
Műszaki szerkesztő: Sándor István
A kézirat a nyomdába érkezett: 1988. május 18. — Terjedelem: 18,2 (A/5 ív)
88-1975 — Szegedi Nyomda — Felelős vezető: Surányi Tibor igazgató

ÚTMUTATÁS A SZERZŐKNEK

Az Alkalmazott Matematikai Lapok csak magyar nyelvű dolgozatokat közöl. A kéziratok gépelését olyan formában kérjük, hogy minden gépelt oldal 25, egyenként átlag 50 betűhelyes sort tartalmazzon. A közlésre szánt dolgozatokat három példányban kell beküldeni.

A kéziratok szerkezeti felépítésének a következő követelményeket kell kielégíteni. A fejlécnek tartalmaznia kell a dolgozat címét, a szerző teljes nevét, valamint annak a városnak a nevét, ahol a szerző dolgozik. A fejléc után egy, képletet nem tartalmazó, legfeljebb 200 szóból álló kivonatot kell minden esetben megadni. A dolgozatot címmel ellátott szakaszokra kell bontani, és az egyes szakaszokat arab sorszámmal kell ellátni. Az esetleges bevezetésnek mindig az első szakaszt kell alkotnia. Az irodalomjegyzék mindig az utolsó szakasz kell hogy legyen, és azt nem kell sorszámmal ellátni. Az irodalomjegyzék után, a kézirat befejezéseképpen fel kell tüntetni a szerző teljes nevét és a munkahelye (illetve lakása) pontos postai címét. A dolgozatban előforduló képleteket szakaszonként újrakezdődően, a képlet előtt két zárójel közé írt kettős számozással kell azonosítani. Természetesen nem szükséges minden képletet számozással ellátni. Az esetleges definíciókat és tételeket (segédtételeket és lemmákat) ugyancsak szakaszonként újrakezdődő, kettős számozással kell ellátni. Kérjük a szerzőket, hogy ezeket, valamint a tételek bizonyítását a szövegben kellő módon emeljék ki. Minden dolgozathoz csatolni kell egy angol, német, francia vagy orosz nyelvű, külön oldalra gépelt összefoglalót. Amennyiben lehetséges, kérjük a nyomtatás számára különösen nehézkes matematikai jelölések használatának az elkerülését.

A dolgozat ábráit és az esetleges lábjegyzeteket a dolgozat végén, különálló lapokon kérjük beküldeni. Mind az ábrákat, mind a lábjegyzeteket a dolgozat szakaszokra bontásától független, folytatólagos arab sorszámozással kell ellátni. Az ábrák elhelyezését a dolgozat megfelelő helyén, széljegyzetként feltüntetett, ábraazonosító sorszámokkal kell megadni. A lábjegyzetekre a dolgozaton belül az azonosító sorszám felső indexkénti használatával lehet hivatkozni.

Az irodalmi hivatkozások formája a következő. Minden hivatkozást fel kell sorolni a dolgozat végén található irodalomjegyzékben, a szerzők, illetve társszerzők esetén az első szerző neve szerint alfabetikus sorrendben úgy, hogy külön, de folytatólagos sorszámozású listát alkossanak a latin és a cirill betűs nevű szerzők műveire vonatkozó hivatkozások, és mindkét részben a megfelelő alfabetikus sorrend legyen kialakítva. A folyóiratban megjelent cikkekre [1], a könyvekre [5], a kötetben megjelent dolgozatokra [4], a disszertációkra [3] és a gépi program leírásokra [2] a következő minta szerint kell hivatkozni:

[1] Farkas, J., Über die Theorie der einfachen Ungleichungen, *Journal für die reine und angewandte Mathematik* **124** (1902) 1—27.

[2] Kéri, G., „DUALSIMP", rutin a CDC 3300-as gépekre (Magyar Tudományos Akadémia Számítástechnikai és Automatizálási Kutató Intézete, CDC 3300 felhasználói ismertetők 2. 1973. május) 19—20.

[3] Prékopa, A., „Sztohasztikus rendszerek optimalizálási problémáiról", doktori értekezés. Magyar Tudományos Akadémia, Budapest, 1970.

[4] Prabhu, N. U., "Recent research on the ruin problem of collective risk theory", in: *Inventory Control and Water Storage* Ed. A. Prékopa (János Bolyai Mathematical Society and North-Holland Publishing Company, Amsterdam—London, 1973) 221—228.

[5] Zoutendijk, G., *Methods of Feasible Directions* (Elsevier Publishing Company, Amsterdam and New York, 1960).

A dolgozatok szövegében az irodalmi hivatkozás számait szögletes zárójelben kell megadni, mint például [5] vagy [4, 76—78]. A szerzők a dolgozatukról 100 darab különlenyomatot kapnak ezek költsége — nyomtatott oldalanként 25 forint — a szerzői díjat terheli.

Ára: 64 Ft

ISSN 0133—3399

TARTALOMJEGYZÉK

INDEX

ALKALMAZOTT MATEMATIKAI LAPOK

A MAGYAR TUDOMÁNYOS AKADÉMIA MATEMATIKAI ÉS FIZIKAI TUDOMÁNYOK OSZTÁLYÁNAK KÖZLEMÉNYEI

XIII. KÖTET

AKADÉMIAI KIADÓ, BUDAPEST

1988

TARTALOMJEGYZÉK

INDEX

88-1975—Szegedi Nyomda — Felelős vezető: Surányi Tibor igazgató