8

Táblázatok átrendezése

(Áttekinthetőség – első látásra)

Dendrogramok, kladogramok, háromszög-diagramok, gyökér-nélküli fa-gráfok, komponensek és egyéb tengelyek, és így tovább... Mind megannyi mesterséges, a megértést közvetve segítő matematikai eszköz, amelyek értelmezése – akárhogy is vesszük – igényel némi ismeretet az előállításuk módjáról, sokszor pedig még a szakembert is nehéz helyzetbe hozhatja. Megkönnyítené a dolgunkat, ha a végeredmény első pillantásra mindenki számára felfogható módon érzékeltetné az adatszerkezetet. Mi lenne akkor, ha nem vezetnénk be semmiféle új matematikai objektumot, vagyis a végeredmény ugyanolyan típusú lenne, mint a kiindulás? Ilyen szempontból az adatmátrixok jöhetnek elsősorban számításba, amelyek sorai és/vagy oszlopai valamilyen feltételrendszer szerint úgy rendezhetők át, hogy ezután már – puszta ránézésre – olyasmit is észrevehetünk az adatok alapján, ami azelőtt rejtve volt előttünk. Hasonlóan érdekes lehet a távolság- vagy hasonlóság-mátrixok átrendezése is, bár ezt már valamilyen más módszerrel együttesen szoktuk alkalmazni. Ebben a fejezetben olyan eljárásokat tekintünk át, amelyek – más program előzetes futtatásával vagy anélkül – alkalmasak az ilyen intuitíve értelmes átrendező műveletre. Nem ígérjük persze, hogy maga az átrendezés algoritmusa is mentes lesz a matematikától, de a végeredmény biztosan. Azért nem került ez a fejezet előbbre, mert elég sok ponton jól jönnek majd az osztályozásról és az ordinációról szerzett eddigi ismereteink. Először csak a változók sorrendjének átrendezéséről lesz szó, majd rátérünk olyan módszerek tárgyalására is, amelyek osztályozós, ill. ordinációs szempontokat érvényesítenek az elemzésben.

8.1 Változók rangsorolása fontosságuk alapján

Az adatmátrixokban a változókat rendszerint teljesen véletlenszerű vagy önkényes sorrendben adjuk meg, pl. neveik szerinti abc felsorolásban. Mindezt nyugodtan megtehetjük, hiszen a többváltozós elemzés eredményének – ha minden egyéb azonos – teljesen függetlennek kell lennie ettől a sorrendtől (ha ez nem áll fenn, akkor nagy baj van, mert a módszer rosszul de-

finiált, vagy a számítógépes programot írták meg hibásan). Felmerülhet az igény azonban, hogy a változók sorrendje ne akármilyen, hanem az adatszerkezetbeli fontosságuknak megfelelő legyen. Legelöl szerepeljenek a meghatározó, döntő fontosságú változók, majd lefelé haladva a táblázatban sorakozzanak az egyre kisebb jelentőségűek, vagy az elhanyagolhatóak. A kulcskérdés persze az, hogy mi is valójában a *fontosság*, mert ennek bizony – mint meglátjuk – többféle meghatározása lehetséges. A fontosság először is *mérhető*, objektív formában kifejezhető, s az, hogy milyen függvénnyel mérünk, megfelel majd a vele kapcsolatos elképzeléseinknek. Továbbá, a változók rangsorolása attól is függ, hogy ez minden egyéb elemzés nélkül, ill. azt megelőzve – mintegy előzetes tájékozódásként – történik-e (*a priori* rangsorolás) vagy pedig valamely többváltozós adatelemzést követően, utólagosan (*a posteriori* rangsorolás), tükrözendő a változóknak az illető vizsgálatban betöltött "szerepét" és súlyát. Ez utóbbi szorosan kapcsolódik az eredmények értékelésének témaköréhez. Dale et al. (1986) egyébként a rangsorolások három fő funkcióját emeli ki:

- A legfontosabb változók kiválasztása, mert a számítógépes program nem tudja kezelni az összes változót. Ez a probléma ma már egyre kevésbé súlyos, tekintve a számítógépek egyre növekvő kapacitását.
- Bonyolult, sokváltozós esetek leegyszerűsítése egyváltozósra (pl. a diszkriminancia függvények komplex sokváltozós elkülönítést tesznek lehetővé, míg a dichotomikus határozókulcsokban egy-egy változó a lényeges minden lépésben).
- Az irreleváns, a mintázat lényegi részeihez hozzá nem járuló változók kiszűrése. Ezek rendszerint "háttér-zajt" produkálnak csupán, így elhanyagolásuk révén az adatszerkezet lényeges jellemzői világosabban kimutathatók.

Itt nem szerepel ugyan a táblázatok átrendezése, de ezt is a rangsorolás egyik fontos – bár nem minden esetben hangsúlyos – céljának tekinthetjük.

8.1.1 Előzetes (a priori) rangsorolás

A rangsor felállításához azt kell kimutatni, hogy melyik változónak a legnagyobb a részesedése az adatstruktúra meghatározásában. Ennek mérése attól függ elsősorban, hogy milyen skálán vettük fel az adatokat. Intervallum- és arányskálán mért változóknál meghatározható a kovariancia vagy korrelációs, esetleg a keresztszorzat mátrix (3.68-70 formulák). Prezencia/abszencia, vagyis bináris adatok esetében emellett információelméleti mérőszámok és a χ^2 statisztika jönnek szóba elsősorban, és a nominális skála esetén is ezek jelentik a megoldást. Ezen kívül még egy választás elé kerülünk: vagy az *eliminációs* vagy pedig az *egyszerű* rangsorolásos technikát választjuk.

Eliminációs módszer. A sorrend felállítása itt több lépésben történik, de legfeljebb annyiban, ahány változónk van. Először kiválasztjuk a legfontosabbat, majd ennek részesedését kivonjuk, eliminálju az adatokból (Orlóci 1973, 1978). Így az adatstruktúrának a most kiválasztott változótól – valamilyen kritérium szerint – független összetevői maradnak csak meg. Az elimináció után megkeressük a második legfontosabb változót, és így tovább. Mindezt addig folytatjuk, amíg a maradvány (reziduális) 0-ra nem csökken. Ezt biztosan elérjük az utolsó változónál, bár az is lehetséges, hogy a már rangsorolt változók jóval előbb elérik a 100 %-os

részesedés szintjét, s a megmaradt változókra már semmi sem jut: közöttük további sorrendet nem is lehet felállítani.

Az eliminációs technikát először az intervallum skálán mért adatokra, az $\mathbf{S}_{nxn} = \{s_{jk}\}$ keresztszorzat, kovariancia vagy korrelációs mátrix elemzésével mutatjuk be. Mint látjuk, itt valójában a nyers, vagy a centrált, vagy pedig a standardizált adatok négyzetösszegéből való részesedés a rangsorolás alapja. A lépések a következők:

1. A kezdő sorszám r=1. Kiszámítandó a tr{**S**} mennyiség, amely a teljes négyzetösszeg (keresztszorzat esetén) vagy a variancia (a centrált és standardizált esetben).

2. Minden j oszlopra előállítjuk az elemek négyzetösszegét, s ezt osztjuk az s_{jj} értékével. Az r sorszámot a legnagyobb eredményt adó változó kapja. Formálisan: megkeresendő az a változó, amelyre a

$$g_j = \sum_{k=1}^{n} s_{jk}^2 / s_{jj} \tag{8.1}$$

mennyiség maximális. Jelöljük ezt a változót h-val. Ennek relatív fontossága százalékban $100 \times g_h/\text{tr}\{\mathbf{S}\}$.

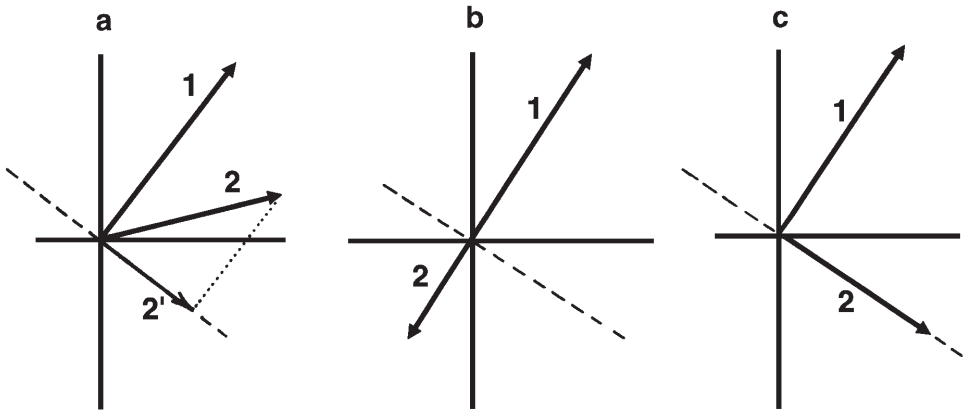
3. A h változó hatását most kivonjuk **S**-ből. A mátrix egy eleme – beleértve az átló elemeit is – ekkor a következőképpen számítandó át:

$$s_{jk} = s_{jk} - \frac{s_{jh} s_{kh}}{s_{hh}} \tag{8.2}$$

Ezek után a mátrix h-adik sorában és oszlopában minden érték 0 lesz, a többi pedig olyan arányban csökken, amennyire az illető változó "kovariál" h-val.

4. r értékét 1-gyel megnöveljük. Ha még van nem zérus érték a mátrixban, akkor visszatérünk a 2. lépéshez. Ellenkező esetben a rangsorolás befejeződött.

Ezzel kiszűrtük a teljes négyzetösszeg (vagy variancia) megmagyarázásához szükséges minimális számú *eredeti* változót. A procedúra megértését geometriai illusztráció segítheti elő, amint a PCA esetében. A változókat egy m-dimenziós tér pontjaiként kell elképzelnünk, az s_{jj} értéke ekkor a j pontra mutató vektor négyzete (önmagával vett skaláris szorzata, C függelék), tr{**S**} pedig az összes vektor hosszának a négyzetösszege. Minden egyes változót sorra veszünk úgy, hogy a vektort tengelyként fogjuk fel. Mindegyikre létezik egy rá merőleges hipersík, amelyre a többi változó vektorai rávetíthetők. A levetített hosszak és az eredeti hosszak különbségei a 8.1 képlet számlálójában szereplő tagok. A legfontosabb változó tehát az lesz, amelyik saját magával együtt (hiszen a levetített saját-rész 0 hosszúságú) a legnagyobb négyzetösszeg csökkenést eredményezi. A 8.1a ábra ezt az m=n=2 esetre mutatja be. A 8.1b ábra érzékelteti, hogy ha a változók teljesen korrelálnak (a vektorok egy egyenesen vannak), akkor egyikük "felesleges". Amikor a változók eleve ortogonálisak, akkor – az előző szituáció ellentéteként – a két változó nem magyaráz semmit egymásból (8.1c ábra). A legfontosabb változó kiválasztása után a dimenzionalitás eggyel csökken, és a kapott *altér*ben új változót keresünk.



8.1 ábra. Az eliminációs rangsorolás egy lépésének geometriai szemléltetése. Mindhárom esetben az 1. vektort tüntetjük ki. Ekkor **a**: a 2. változónak az elsőtől független részesedése a szaggatott vonalra vetítés után maradó vektor (2') hosszával arányos, **b:** a 2. változónak nincs az elsőtől független részesedése, és **c:** a két vektor ortogonális egymásra.

Mindebből látszik, hogy a módszer alapja a négyzetösszeg ortogonális összetevők szerinti felbontása. A felbontás azonban nem mesterséges változók (komponensek) szerint történik, így a sorba rendezett változók kumulatív %-os részesedése mindig alatta marad az ugyanolyan számú sorba rendezett főkomponens %-os részesedésének. (Éppen elérné azt, ha az eredeti változók egybeesnének a komponensekkel, amire gyakorlatilag nincs esély valós adatmátrixok esetében.) Az eredeti változók hallatlan előnye a komponensekkel szemben azonban az, hogy "ismerjük" őket, közvetlenül értelmezhetők.

Példaként vizsgáljuk meg az A1 táblázat változóit az eliminációs módszerrel, mindhárom kritérium alapján. Az eredményeket a 8.1 táblázat összesíti. Keresztszorzat esetében 8, míg a másik kettőnél – a centrálás miatt – eggyel kevesebb változó éri el a 100 %-ot. Ez nem véletlen, hiszen a kiinduló szimmetrikus mátrix rangja (C függelék) a rangsorolható változók számát is befolyásolja. A keresztszorzat esetében (8.1A) kissé meglepő az eredmény, hiszen a legfontosabbnak olyan faj bizonyult, amely igen kis négyzetösszegű (18,0) más fajokhoz képest (pl. a BRO ERE 3020,0 és a SES SAD pedig 4916,0 négyzetösszeget ad). Ebből is látszik, hogy nem az abszolút értékek számítanak (legkevésbé persze a korreláció esetén), hanem az *irányultság*, amit a fajvektor a sokdimenziós térben képvisel. Márpedig a CAR HUM-mal egybeeső tengellyel a teljes négyzetösszeg 41,9 %-a megmagyarázható, s ez nagyobb, mint bármelyik más faj esetében. A kovarianciára egészen más sorrendet kaptunk (8.1B), mutatva a centrálás hatását. Itt már egyértelműen a nagy varianciájú fajok dominálnak, míg a kis varianciájúak fel sem bukkannak a rangsorban. Standardizálás hatására (8.1C) – mint várható – megint más sorrend alakul ki, amelyre az jellemző, hogy a kumulatív százalékok (utolsó oszlop a táblázatban) lassabban növekednek, mint az előző két rangsorban.

Mikor érdemes az eliminációs rangsorolást alkalmazni? Nos, minden olyan esetben, amelyben túl sok változónk van és az alkalmazandó számítógépes módszer

- csak jóval kevesebb változóval tud dolgozni, mint amennyi az adatokban szerepel és
- olyan alapelven próbálja meg a dimenzionáltság csökkentését, amelyet maga a rangsorolásos technika is alkalmaz (kompatibilitás).

8.1 táblázat. Az A1 cönológiai tabella fajainak rangsorolása az eliminációs módszerrel három mérőszám szerint. A 100 % elérése után megmaradó fajok nem szerepelnek a táblázatban. Kisebb eltérések a kerekítési hibákból adódhatnak.

	Rangszám	Változó	Specifikus rész	Relatív fontosság	Kumulatív %
A	1	CAR HUM	5297.278	41.935	41.935
Keresztszorzat	2	SES LEU	3629.493	28.733	70.668
	3	BRO ERE	2656.635	21.031	91.699
	4	CHR GRY	549.148	4.347	96.046
	5	FUM PRO	284.417	2.252	98.298
	6	SCA CAN	123.509	0.978	99.275
	7	CAM SIB	50.065	0.396	99.672
	8	SES SAD	41.487	0.328	100.000
		Total:	12632.000	100.000	
B	1	SES SAD	651.905	53.642	53.642
Kovariancia	2	BRO ERE	318.132	26.178	79.820
	3	SES LEU	161.852	13.318	93.138
	4	CHR GRY	59.445	4.891	98.029
	5	FES PAL	18.822	1.549	99.578
	6	SCA CAN	4.483	0.369	99.947
	7	KOE CRI	0.647	0.053	100.000
		Total:	1215.286	100.000	
C	1	CAR LIP	4.061	33.840	33.840
Korreláció	2	FUM PRO	2.372	19.763	53.603
	3	CHR GRY	1.961	16.345	69.949
	4	SES SAD	1.576	13.131	83.080
	5	SES LEU	0.951	7.925	91.004
	6	BRO ERE	0.882	7.346	98.350
	7	FES PAL	0.198	1.650	100.000
		Total:	12.000	100.000	

A rangsorolást követően az adatmátrix mérete erőteljesen redukálható anélkül, hogy a végeredmény jelentősen megváltozna. A centrált PCA például a rangsorban első három faj alapján (93 %, 8.1B táblázat) gyakorlatilag ugyanolyan eredményt ad az első két komponensre, mint amikor az összes faj benne van az elemzésben (ki lehet próbálni!). Nincs értelme azonban adott rangsort alapul venni a "felesleges" változók kiszűrésében, ha az eliminációs technika logikailag nem kompatibilis a módszerrel, mint az osztályozások esetében. Ekkor az egyszerű rangsorolásos technikák közül válasszunk. Az elimináció, bár táblázatok átrendezésére elvileg alkalmas lenne, mégsem jön számításba ilyen szempontból (egyetlen kivételként e fejezetben), mert a változók elhagyása miatt a táblázat nem igazán informatív.

Prezencia/abszencia esetben, Orlóci (1976a) javaslatára, alkalmazható még a változók hozzájárulása azok kölcsönös információjához (3.115 formula). A sorrendet itt úgy határozzuk meg, hogy minden lépésben megvizsgáljuk: melyik változó kiesése okozza a függvényérték legnagyobb csökkenését. A legfontosabb változónak ugyanis azt tekinthetjük, amelyik a legtöbb információt tartalmazza az összes többire nézve. Ennek elhagyása után megkereshető a második legfontosabb változó, ami a maradék kölcsönös információ java részéért "felelős", és így tovább. Az utolsó két helyen – ha addig nem érjük el a nullát – szükségképpen "holtverseny" van. E módszer hátránya, hogy nagy adatmátrixokra rendkívül számításigényes. A formula kibővítése lehetőséget nyújt a többállapotú nominális változók rangsorolására is.

Az információstatisztikák mellett a 2^n kontingenciatáblák χ^2 elemzése is segíthet (vö. Fienberg 1970).

Egyszerű rangsorolás. A numerikus osztályozásban, akár hierarchikus, akár nem-hierarchikus, az azonos módon "viselkedő" változók erősítik egymást, és ha sok változó ugyanazt "mondja", akkor az osztályozás is általánosabb érvényű lesz. Nem volna értelme tehát az első változó kiszűrése után a vele erősen korreláló ill. asszociálódó változókat idő előtt kiiktatni. Más típusú rangsorra van itt szükség, ami a változó abszolút részesedését mutatja. Miután nem ortogonális felbontást végzünk, az összes változó rangsorolására is lehetőség nyílik, és a rangsor szerint átrendezett adattáblázat is informatív lesz. Először a változó varianciájára gondolhatunk, mondván, hogy a kis varianciájú változók valószínűleg sokkal kevésbé értékesek az osztályok elkülönítésében, mint a nagy varianciát felmutató változók (más kérdés, hogy utólag mégis interpretatívak lehetnek, de erről már szóltunk, vö. 5.5.3 rész). Ezt a típusú rangsorolást – tudatosan vagy kevésbé tudatosan – igen sokan használják szerte a világon, amikor pl. feldolgozhatatlanul terjedelmes cönológiai táblázataikból a ritka fajokat egyszerűen elhagyják. Szóba jöhet a keresztszorzat, a kovariancia és a korreláció is – de elimináció nélkül. Ez azt jelenti, hogy az eliminációs algoritmus 2. lépésében kapott értékek alapján végzünk egyszerű rangsorolást (Podani 1994). Ez – ha visszagondolunk a 8.1 ábra értelmezésére – végül is a változók fontosságát attól teszi függővé, hogy *saját irányultságukban* mennyire képviselik a többieket. Azaz mennyire "reprezentatív" az egész adatmátrixot tekintve a változó vektora az *m*-dimenziós térben. Az egyéni módon viselkedő vagy csak sztochasztikus zajt okozó változók ebben a rangsorban bizonyosan hátulra kerülnek. (Az egyes változókra kapott fontossági értékek formailag összegezhetők ugyan, s így százalékos "hozzájárulás" is meghatározható, de ez csak arra alkalmas, hogy a változók egymáshoz viszonyított relatív fontosságát megmutassuk.)

Vizsgáljuk most meg az A1 táblázat fajait az egyszerű rangsorolás segítségével (8.2 táblázat). Az első két oszlop az összvarianciából való részesedést mutatja, s inkább csak tájékozódásra való. Ez a variancia ugyanis – mint a CAR HUM példája is mutatja – önmagában semmit nem mond a változók közötti kapcsolatokról. A relatíve kis mennyiségben jelenlevő faj is fontos lehet tehát. A keresztszorzat és a kovariancia alapon most a két sorrend hasonlóbb egymáshoz, mint a 8.1 táblázatban. Ezt a rangsorolást ajánlhatjuk minden olyan esetben, amikor a változók abszolút mennyisége döntő (pl. euklidészi távolság v. eltérésnégyzet-összeg alapján osztályozunk). A korrelációs adatstruktúra szerinti rangsor pedig inkább olyankor jöhet számításba, amikor standardizált adatokkal kívánunk dolgozni. Átrendezett tabellát is érdemes készíteni, mint például a kovariancia szerint:

SES	SAD	0	0	0	0	0	0	4	70
CAR	HUM	1	0	0	0	0	0	1	4
FES	PAL	20	11	5	15	25	4	6	2
SES	LEU	25	15	0	8	25	1	1	0
BRO	ERE	5	7	18	0	1	0	50	11
CAR	LIP	2	0	1	1	3	1	0	0
CAM	SIB	0	1	0	0	0	0	2	1
CHR	GRY	30	8	5	0	4	0	0	0
FUM	PRO	3	11	7	5	7	12	3	2
SCA	CAN	1	10	0	0	0	0	2	8
CEN	SAD	1	1	1	4	1	2	3	3
KOE	CRI	5	1	2	1	1	0	2	1

amelynek négy első sorában jól láthatjuk az adatokban rejlő változás fő "felelőseit". A rangsor végén szereplő fajok ilyen szempontból inkább "zaj"-változóknak számítanak. Óvatosságra

8.2 táblázat. Az A1 cönológiai tabella fajainak egyszerű rangsorolása négy szempont szerint.

	Variancia			Keresztszorzat			Kovariancia			Korreláció		
	Species		%	Species		%	Species		%	Species		%
1	SES SAD	604.50	49.74	CAR HUM	5297.2	12.1	SES SAD	651.9	16.9	CAR LIP	4.0	10.8
2	BRO ERE	280.28	23.06	SES SAD	5212.1	11.9	CAR HUM	594.9	15.4	SES LEU	3.8	10.3
3	SES LEU	119.69	9.84	FES PAL	3840.8	8.8	FES PAL	374.2	9.7	FES PAL	3.8	10.2
4	CHR GRY	104.12	8.56	SES LEU	3723.0	8.5	SES LEU	363.7	9.4	CAM SIB	3.3	8.8
5	FES PAL	69.14	5.68	KOE CRI	3673.3	8.4	BRO ERE	322.8	8.3	CAR HUM	3.2	8.7
6	SCA CAN	16.26	1.33	CAM SIB	3673.1	8.4	CAR LIP	297.5	7.7	SES SAD	3.2	8.7
7	FUM PRO	13.92	1.14	BRO ERE	3626.1	8.3	CAM SIB	289.2	7.5	CHR GRY	3.1	8.5
8	KOE CRI	2.26	0.18	CEN SAD	3407.1	7.8	CHR GRY	227.6	5.9	KOE CRI	2.6	7.2
9	CAR HUM	1.92	0.15	SCA CAN	3024.2	6.9	FUM PRO	207.7	5.4	BRO ERE	2.5	6.8
10	CEN SAD	1.42	0.11	CAR LIP	2964.5	6.8	SCA CAN	200.6	5.2	CEN SAD	2.5	6.8
11	CAR LIP	1.14	0.09	CHR GRY	2796.7	6.4	CEN SAD	167.7	4.3	FUM PRO	2.4	6.4
12	CAM SIB	0.57	0.04	FUM PRO	2311.3	5.3	KOE CRI	148.3	3.8	SCA CAN	2.3	6.2

kell azonban intenünk a tekintetben, hogy ezeket osztályozásra teljesen alkalmatlannak tekintsük. Ha nincs is beleszólásuk a fő csoportok kialakulásába, a kevéssé fontos változóknak is lehet szerepük az osztályozás finomabb részleteiben.

Egyszerű rangsorolásra még számos más módszer is alkalmazható. Kiszámítható például minden egyes változónak a többivel adott *többszörös korrelációja* (ami a kanonikus korreláció – 7.2 rész – speciális esete, az egyik csoportban n–1, a másikban pedig egy változóval). A többszörös korrelációk értékei adják a rangsorolás alapját. Rohlf (1977) és Orlóci (1978) tárgyalja részletesen ezt az eljárást, megemlítve, hogy lényegesen számításigényesebb, mint a többi módszer.

Dale & Williams (1978) az egész adattáblázatot egy kontingencia-táblázatnak tekinti (ami a COA alapja is egyben), majd a sor- és oszlopösszegek alapján kiszámítja minden érték eltérését az arra a helyre várható adattól (a 3.36 formula számlálójában lévő mennyiség). Ezen eltérések abszolút értékeinek összege ("eident value") adja minden változóra a rangsorolás alapját. A stratégia eliminációs változata is elképzelhető, amikor is minden lépésben csak a legfontosabb változót keressük meg, ezt kihagyva a mátrixból újraszámoljuk az eltéréseket, és így tovább.

Bináris változók egyszerű rangsorolása a numerikus osztályozás hőskorában is a χ^2 függvény felhasználásával történt, mintegy a divizív osztályozó folyamat részeként (l. az 5.3.2 részt). Minden klasszifikációs lépésben kiszámolták a változók közötti asszociációs koeffiensek mátrixát, s ennek oszlopösszegei adták az alapot a rangsoroláshoz (5.7 formula). A legnagyobb összegű változó tekinthető ui. a többit legjobban magyarázó változónak. A kis cella-gyakoriságokra kevésbé érzékeny 5.8 formula talán még inkább megfelelő a bináris változók egyszerű, *a priori* rangsorolására.

8.1.2 Utólagos (a posteriori) rangsorolás

A változók fontosságának meghatározása egy eredmény kialakulásában szinte minden többváltozós elemzés szerves része kellene, hogy legyen, melyet szinte természetes módon követhet az adatmátrix átrendezése. Erről már szóltunk egyszer-kétszer az előzőekben is, pl. a hierarchikus osztályozás értékelésével kapcsolatban (5.3.3 rész). Most röviden felvetünk néhány rangsorolási lehetőséget, a többváltozós módszerek főbb csoportjainak megfelelően.

Minden esetben lényeges, hogy a rangsorolás mérőszáma logikailag kompatibilis legyen a többváltozós elemzés során alkalmazott távolság- s egyéb függvényekkel.

Változók szerepe a partíciókban. A *k*-közép módszer "jósági" kritériumában (*J*, 4.1 függvény) a változók összhatása additív (*i* szerinti összegzés!). A *J* felbontása változók szerinti összetevőkre ennek alapján nem okozhat nehézséget, majd az összetevők nagyság szerint emelkedő sorrendje megadja a változók hozzájárulásának "erősorrendjét". Az ideális, a partíciót tökéletesen megmagyarázó változó 0-val járul a *J* értékéhez (ami azt jelenti, hogy a változó minden egyes osztályon belül konstans értéket vesz fel), míg az osztályozást nem támogató változók hozzájárulása a legnagyobb. Az index-független particionáló módszer esetében már jóval rejtettebb a változók szerepe. Először ugyanis különbözőségeket számolunk, majd ezeknek képezzük az átlagait, s emiatt viszonylag nehéz követni a változók hatását. Az 5.3.3 rész végén leírt általános értékelő módszert azonban éppen az ilyen esetekre dolgoztuk ki. A Ψ_{ik} mérőszám azt fejezi ki, hogy *k* csoport esetén milyen mértékben járul az *i* változó az osztályokon belüli távolságokhoz (vagy különbözőségekhez) az osztályközötti hozzájárulásokhoz képest. (Ezek kiszámítását egyes távolság- és különbözőség-indexekre l. Podani 1997). A Ψ függvény felhasználásával a változók – a particionálásnál választott különbözőségi indexszel összhangban – sorba rendezhetők. Lágy osztályozásoknál a változók hozzájárulásai a "fuzzy" eltérésnégyzet-összeghez a 4.6 és a 4.7 összefüggések segítségével számolhatók ki, majd ezután rangsorolhatók – a *k*-közép módszerhez hasonlóan – emelkedő sorrendben.

A három *Iris* faj lágy osztályozásában az *f* lágysági paraméter 1,25-ös értéke mellett (4.9 ábra) a négy változó hozzájárulásai a következő sorrendet adják: BLSZ 10,7%, KLSZ 19,7 %, KLH 33,2 % és BLH 36,4 %. Ez nem lehet különösebben meglepő, mert nyers adatokkal dolgoztunk, s a méretek is ilyen sorrendben növekednek. Az *f*=2,5 esetben sem sokat változik a helyzet, bár a két hossz-méret helyet cserél, és a külső lepelhossz lesz az osztályozásnak leginkább ellentmondó tulajdonság.

További lehetőségek a változók fontosságának értékelésére a következők. Egy változó csoportok közötti és csoportokon belüli varianciájának (ha nem 0) a hányadosát, formálisan az *F*-statisztika alkalmazását javasolta Jancey (1979). A csoport-közötti és a teljes variancia hányadosát pedig Lance & Williams (1977) alkalmazta. Ez utóbbi szerzők bináris és nominális adatok esetében minden egyes változóra felírtak egy kontingenciatáblát (sorok az osztályok, oszlopok a tulajdonság egyes állapotai) és a Cramér indexszel (3.37) mérték a változó diszkriminatív erejét.

Hierarchikus osztályozások. A hierarchikus osztályozások partíciók sorozataként foghatók fel, így a változók szerepe minden egyes hierarchikus szintre külön-külön értékelhető a már említett módszerek valamelyikével (tipikus példa erre Lance & Williams (1977) módszere). Egy változó, amely kiemelkedő az objektumok – mondjuk – két osztályra történő felosztásában, már erőteljesen ellentmondhat a három vagy több osztályba csoportosításnak, amelyet persze más változók viszont támogathatnak. Emiatt nincs különösebb értelme olyan módszert keresni, amely a változók globális, a teljes hierarchiát meghatározó szerepét rangsorolná.

Kladogramok. A változók fontossága egy kladisztikai hipotézisben a konzisztencia index (6.9) és az összetartási index (6.11) felhasználásával értékelhető. Az adott kladogramot egyértelműen támogató karakterek az 1-es értéket veszik fel, s természetszerűleg ezek kerülhetnek az átrendezett adattáblázat első soraiba, majd ezeket követik az egyre csökkenő értéket adó tulajdonságok. Az egyezések miatt a sorbarendezés sok esetben csak részleges lehet.

A változók súlya az ordinációban. A rangsorolás alapja ekkor sokféleképpen megválasztható, és természetesen attól függ elsősorban, hogy milyen ordinációs módszert alkalmaztunk. Mivel az ordinációt rendszerint két dimenzióban ábrázoljuk, számunkra többnyire az az érdekes, hogy az 1. és 2. tengelyen kapott elrendezést mely változók értelmezik a legjobban. A *főkomponens* elemzésben a rangsorolás alapja az lehet, hogy a változók saját varianciájából hány százalékot fed le a két kiválasztott komponens, tehát a 7.12 formulát kell alkalmaznunk.

> A 7.1 táblázat alsó részében, az első két oszlopban lévő százalékok összeadásával megkapjuk, hogy a 7.2 ábrán látható ordináció leginkább a SES SAD (99 %), BRO ERE (87,7 %), CAR HUM (86,8 %), SES LEU (63,8 %), FES PAL (62,5 %), CAM SIB (61,6 %), és a CAR LIP (54,5 %) "véleményét" tükrözi, s legkevésbé a KOE CRI (6,2 %) egyezik vele. Ez a sorrend elég jól megegyezik az *a priori* egyszerű rangsorral a kovariancia alapján (8.2 táblázat). A KOE CRI elhagyása tehát igen kis mértékben változtatta volna meg az eredményt.

A standardizált PCA esetében a változók rangsorolása hasonlóan történik. Egy változó és a két kiválasztott komponens közötti korrelációk négyzetösszege pontosan megadja a megmagyarázás mértékét (emlékeztetőül: egy változónak az összes komponenssel vett korrelációi 1-es négyzetösszeget adnak). A *kanonikus korreláció elemzés*ben a 7.26-27 függvények alkalmasak a két változócsoport tagjainak sorbarendezésére, külön-külön természetesen. A *korreszpondencia-elemzés*ben a változók pozícióinak az origótól vett távolsága ad információt fontosságukról. Minél nagyobb ez a távolság, annál lényegesebb az illető változó szerepe az objektumok elrendeződésében. Csakúgy mint a standardizált PCA-nál, az éppen vizsgált két komponensen lényegtelen változók az origó közelébe kerülnek. A *többdimenziós skálázás*ban szóba se jön a változók értékelése, hiszen ezekre nincs is közvetlenül szükség. A *diszkriminancia-elemzés*ben pedig a változók kommunalitása (7.79 formula) lehet a sorbarendezés alapja, amint ezt a 7.2 táblázat már példázta is.

Átrendezett táblázatok. Mindeddig visszafelé lapoztunk a könyvben, most pedig egy kicsit előrefelé tekintünk. Az adattáblázatok blokkos (8.2.3) vagy átlós (8.3) szerkezetének optimalizálását követően megállapítható az egyes változók (és az objektumok!) relatív hozzájárulása az eredményhez[1]. Blokk-osztályozásoknál a módszer a jackknife eljárás alapelvét követi: a blokkok "élességét" mérő függvényt meghatározzuk úgy is, hogy az adott változót kihagyjuk, s az ilymódon redukált mátrixra valamint a teljes mátrixra kapott két érték különbségét kiszámítjuk. A χ^2 esetében ez a különbség negatív és pozitív is lehet: negatív irányú eltérés (a χ^2 csökkenése a változó kihagyására) azt jelenti, hogy az illető változó jelenléte elősegíti a blokkosodást, míg a pozitív változás annak a jele, hogy a változó zavarja a blokk-szerkezetet, és eltávolítása az eredetinél erősebben strukturált adatmátrixot eredményezne. A rangsor tehát a legnegatívabb eltérést okozó változóval kezdődik s a legnagyobb különbséget adókkal záródik. Ha a blokk-szerkezet mérőszáma az entrópia vagy az eltérésnégyzet-összeg, akkor a változás legfeljebb csak csökkenés lehet. Itt azok a változók a legjobbak, amelyek kihagyása kis csökkenést eredményezne, míg a viszonylag nagy csökkenést adó változók a blokk struktúrának leginkább ellentmondóak. Az átlós szerkezet optimalizálásában a változók hozzájárulása additív, s a 8.10 függvény szerint könnyen megkapható. Minél

1 Természetesen – az eddigiekkel ellentétben – ebben az esetben nem egy újabb táblázat szerkesztése az *a posteriori* rangsorolás célja, hanem a blokkok értelmezésének a megkönnyítése.

nagyobb a hozzájárulás mértéke, annál kevésbé egyértelmű a változó helyzete az átrendezett mátrixban. Mindezekre példákat is láthatunk majd az alábbiakban.

8.2 Blokk-osztályozás

A változók rangsorolása természetesen csak részlegesen alkalmas – ha egyáltalán alkalmas – táblázataink átrendezésére, hiszen nem feledkezhetünk meg az objektumokról sem. Amennyiben mind a változók, mind pedig az objektumok értelmes módon csoportokba oszthatók, azaz osztályozhatók, akkor a táblázatos átrendezésnek célszerűen tükröznie kell e csoportokat. Ennek az a hallatlan nagy interpretatív előnye, hogy a változók osztályai jól értelmezhetik az objektumok osztályait és viszont. A sorok és az oszlopok szerinti klasszifikáció ugyanis a táblázatot téglalap alakú részmátrixokra, ún. blokkokra darabolja – minden egyes blokk mutatva az adott változócsoport és objektum-osztály kölcsönös viszonyát. Bináris adatok esetében például ez a kapcsolat akkor a legegyértelműbb, ha bizonyos blokkok csupa 1-esből, a többiek pedig 0-ból állanak. A blokkok szerinti strukturáltság azonban nemigen látszik egy szabadon felírt adatmátrixban; az ilyen típusú adatszerkezet feltárása a blokk-osztályozás feladata. Az alapproblémát a 8.2 ábra egyszerű mátrixa illusztrálja.

Adatmátrixok blokkos háttérszerkezetének keresése a tudomány legkülönfélébb területein merülhet fel. A biológiában például nagyméretű növénycönológiai tabellák megfelelő átrendezése a kezdetektől számítva egyik fő célja a Zürich-Montpellier-i iskola követőinek (vö. Braun-Blanquet 1965, Mueller-Dombois & Ellenberg 1974). Ez, számítógép és megfelelő módszerek hiányában, manuálisan igen fáradtságos munka volt, bizonytalan értékű végeredménnyel. Kézenfekvő megoldásként kínálkozik az, hogy végezzük el a változók osztályozását és az objektumok osztályozását ugyanabból az adatmátrixból, ugyanazzal a módszerrel, majd az átrendezést a kapott csoportok szerint végezzük el. Az első ilyen vizsgálat Williams & Lambert (1961a,b) nevéhez fűződik. Az attribútum dualitás elvének megfelelően az asszociáltság analízis módszerét (5.3.2 rész) alkalmazták a cönológiai kvadrátokra (normál elemzés) a fajok χ^2 -összegzése szerint, majd a fajokra (inverz elemzés) a kvadrátok χ^2 összegeit figyelembe véve. A dendrogramokat megfelelő helyeken elmetszve kapott csoportok szerint rendezték át az adattáblázatot. Módszerük "*nodal analysis*" néven vált ismertté, utalva arra, hogy az átren-

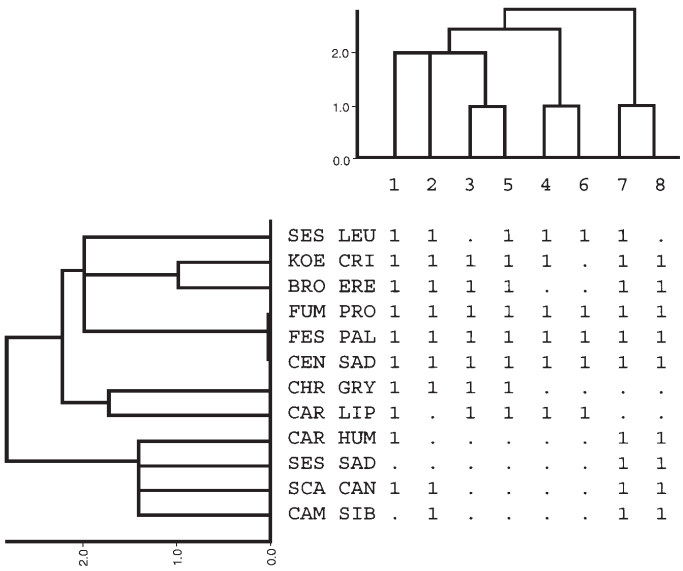
a

```
..1..11..
.1..1...1
1..1...1.
.1..1...1
..1..11..
1..1...1.
..1..11..
.1..1...1
1..1...1.
```

b

```
111......
111......
111......
...111...
...111...
...111...
......111
......111
......111
```

8.2 ábra. A teljesen rendezetlen mátrix (**a**) elfedi előlünk a sorok és az oszlopok közötti erős interakciót (**b**), melynek felderítése a blokk-osztályozásra vár. (A szemléletesség kedvéért a 0-k helyett pontok szerepelnek.)



8.3 ábra. Az A1 táblázat binarizált változatának blokkos átrendezése a sorok és oszlopok *euklidészi távolság + teljes lánc* módszerrel való osztályozását követően. Az ábra egyúttal illusztrálja az ilyen típusú mátrixoknál gyakori egyezéseket (egyszerű lánc feloldással, vö. 5.2 rész). A blokkok kijelölése a dendrogramok alapján *részben* önkényes, csakúgy mint az objektumok sorendje.

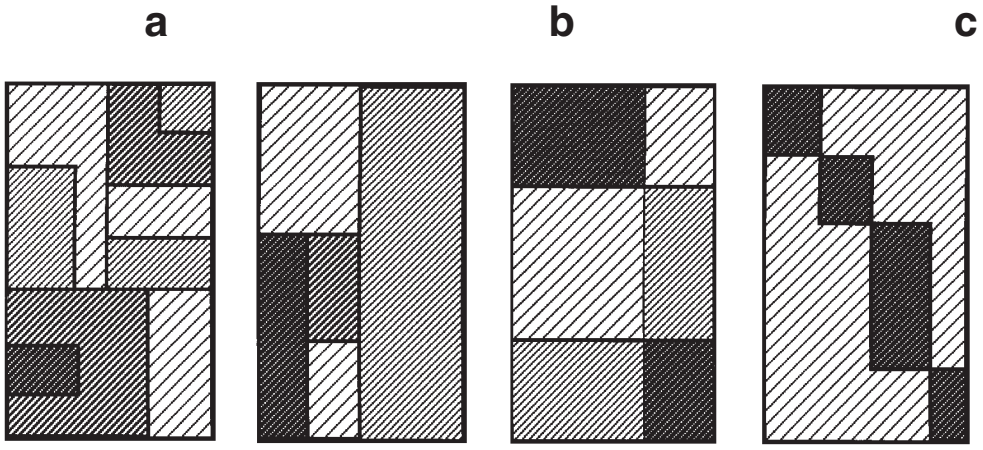
dezés után a blokkok jól mutatják a faj-kvadrát egybeeséseket, csomósodásokat[2]. A divizív módszer helyett természetesen bármilyen más hierarchikus és nem-hierarchikus osztályozás is elképzelhető, amely mindkét irányban alkalmas a táblázat elemeinek osztályozására. A két osztályozás "egymásra vetítéséből" pedig kialakítható az átrendezett mátrix, abban a reményben, hogy a változók és objektumok csoportjainak kölcsönössége maximálisan kirajzolódik (8.3 ábra).

Ez azonban nem mindig van így. A két osztályozás – eltekintve persze attól, hogy ugyanabból az adatmátrixból történik – némileg "független" egymástól. Az oszlopok osztályozása során nem vesszük figyelembe, hogy a változók is csoportosulhatnak, és fordítva: a változók osztályozásából is hiányzik az objektumcsoportok osztályainak ismerete. A változók és az objektumok közötti "interakció" akkor mutatkozik meg igazán a táblázatban, ha a blokkokat közvetlenül állítjuk elő valamilyen kereső vagy optimalizációs technikával (Gordon 1981). Erre a célra új, speciális módszerekre van szükség, így ezt a részt akár a klasszifikációs fejezetek kései folytatásának is tekinthetjük.

A blokk-osztályozás módszereit négy fő csoportra oszthatjuk aszerint, hogy milyen megszorításokat alkalmazunk a sorok ill. az oszlopok klasszifikációjában:

- A legegyszerűbb esetben voltaképpen nincs semmiféle sor- vagy oszlop szerinti osztályozás, az átrendezés feladata a táblázaton belüli maximálisan homogén blokkok, adatcsoportosulások kikeresése (8.4 a ábra).
- A részleges blokk-osztályozásban a sorok *p* az oszlopok pedig *q* osztályba tartoznak, de egy sor szerinti blokk egyidejűleg kettő vagy több oszlop szerinti osztályt is jellemezhet és fordítva (8.4b ábra).

2 Greig-Smith (1983) tekinti át a cönológiában kifejlesztett hasonló, számítógép-orientált módszereket.



8.4 ábra. A blokk-osztályozás alaptípusai. **a:** Megszorítás nélküli blokkok, **b**: részleges blokkosítás. **c**: kereszt-partíció, általános eset ($p \neq q$), **d**: blokk-szeriálás ($p=q$), Árnyékolás utal a blokkok belső homogenitására.

- Ha az átrendezett mátrixban bármely érték sor szerinti besorolása egyértelműen megadja az oszlop szerinti osztályba tartozást és viszont, akkor teljes blokkosításról, vagy kereszt-partícióról beszélhetünk (8.4c ábra). A $p \neq q$ itt megengedett,
- Ha viszont kikötjük a $p=q$ feltételt, és a sorok ill. oszlopok osztályai között egy-egyértelmű megfeleltetést keresünk, akkor a 8.3 rész felé átmenetet mutató problémáról, a blokk-szeriálásról (8.4d ábra) van szó. Ekkor figyelmünket az átlós blokkokra összpontosítjuk, az átlón kívülre esőket "egy kalap alá" véve.

8.2.1 Blokkok keresése megszorítások nélkül

Ilyen típusú módszereket elsősorban Hartigan (1975) könyvében találhatunk. Egyik módszere, a "*two-way joining*" v. *kétutas összevonó* algoritmus bináris adatokra való. Az egyezési koefficiens (3.6) komplementjét alkalmazza távolságfüggvényként, s az elemzés minden lépésében az egymáshoz legközelebbi két sort vagy oszlopot vonja össze, azaz helyezi el egymás mellé a mátrixban. A maximálisan homogén blokkok száma az elemzés közben alakul ki. Egy, a blokkon belüli homogenitást kontrolláló küszöbérték bevezetésével a módszer intervallumskálán mért változókra is alkalmassá tehető.

Az A1 táblázat binarizált változatának kétutas összevonó elemzése látható a 8.5 ábrán. A módszer annyi blokkot alakít ki, amennyi minimálisan szükséges ahhoz, hogy mindegyiken belül csak azonos értékek szerepeljenek. Eredményül elég sok blokkot kaptunk, s ezek elhatárolása is tartalmaz önkényes elemeket. A 8.3 ábrán látható eredménytől is feltűnően nagy az eltérés. Elképzelhetjük, hogy nagyobb adatmátrixok esetén e módszerrel könnyen kaphatunk áttekinthetetlen és emiatt nehezen interpretálható eredményt.

Hartigan (1981) másik módszere kategorizált (intervallumokra osztható) adatokra alkalmas, és a vezető (*leader)* algoritmus (vö. 4.1.4 rész) segítségével választja ki a blokkok kezdő elemeit, amelyek egy előre megadott küszöbértéknél távolabb vannak a többitől (a kezdő elem az 1. sorhoz és 1. oszlophoz tartozó érték). Az egyes lépésekben felváltva tekinti a sorokat, ill. az oszlopokat. Ha túl sok blokkot kapunk (pl. mindössze egy-egy értékkel), akkor túl alacsonyra vettük a küszöbértéket, s érdemes egy magasabbal próbálkozni.

	1	7	8	2	3	5	4	6
BRO ERE	1	1	1	1	1	1	.	.
CEN SAD	1	1	1	1	1	1	1	1
FES PAL	1	1	1	1	1	1	1	1
FUM PRO	1	1	1	1	1	1	1	1
KOE CRI	1	1	1	1	1	1	1	.
SES LEU	1	1	.	1	.	1	1	1
CHR GRY	1	.	.	1	1	1	.	.
SCA CAN	1	1	1	1	.	.	.	.
CAM SIB	.	1	1	1	.	.	.	.
CAR HUM	1	1	1	.	.	.	.	.
SES SAD	.	1	1	.	.	.	.	.
CAR LIP	1	.	.	.	1	1	1	1

8.5 ábra. Az A1 táblázat binarizált változatának értékelése a kétutas összevonás módszerével. A jobb áttekinthetőség kedvéért a zérus értékeket pontok helyettesítik.

A prezencia/abszencia esetre Bruelheide & Flintrop (1994) is egy küszöbérték alkalmazását javasolja: a blokkot azok a változók alkotják, amelyek a blokk objektumainak legalább ε százalékában megvannak és fordítva. A módszer sorok és oszlopok fokozatos elhanyagolásával alakítja ki a blokkokat. A kapott eredmény azonban sok esetben nem igazi blokkosítás, mert egyes blokkok elemei szétszórtan helyezkedhetnek el a mátrixban (l. a szerzők 8. táblázatát). Eckes (1995) az eltérésnégyzet-összeg minimalizálását egy agglomeratív stratégiával próbálja meg elérni ("*centroid effect method*"). A viszonylag bonyolult algoritmus a táblázatban lévő értékeket aszerint vonja össze blokkokba, hogy az eltérésnégyzet-összeg növekedése minimális legyen – a módszer tehát az 5.5 kritérium szerint működő hierarchikus módszer adaptálása blokkokra. A blokkokat a fúziók leállításával kapjuk a szerző szerint akkor, amikor a kritérium "erőteljesen növekedik", s ez persze némi önkényességet visz az elemzésbe.

8.2.2 Az adatmátrix részleges blokkosítása

Gordon (1981) több, részlegesen particionáló módszert is említ, külön kiemelve a Hartigan (1972) -féle divizív eljárást. Ez intervallum/arány-skálán felvett adatokra alkalmas, ugyanis a blokkokon belüli eltérésnégyzet-összeget minimalizálja. A kapott blokkokon belül az értékek tehát a lehető leghasonlóbbak egymáshoz. Kezdetben nincs semmiféle kikötés a blokkok számára vonatkozóan. Jelölje $\bar{z}_{ij}$ annak a blokknak az átlagértékét, amelybe x_{ij} tartozik, s ekkor a feladat a következő mennyiség minimalizálása:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \bar{z}_{ij}) \qquad (8.3)$$

A minimumot Hartigan (1972) egy hierarchikus stratégiával próbálja meg elérni. Az adatmátrixot, később pedig az egyes blokkokat szukcesszíve osztja két részre az oszlopok v. a sorok szerint, attól függően, hogy melyik adja a maximális csökkenést J értékében. A sorok és oszlopok sorrendisége azonban nagymértékben befolyásolja az eredményt, s nem világos, hogy a módszer mennyire képes a permutációkat is figyelembe venni. Ez a divizív stratégia tehát csak akkor használható, ha a sorrendet valamilyen más módszerrel egyértelműen meghatároztuk ill. rögzítettük. Dale & Anderson (1973) monotetikus divíziókkal éri el az adattáblázat hasonló jellegű felosztását.

8.2.3 Kereszt-partíciók

Ebben az esetben a feladat a változók p csoportba, az objektumok pedig q számú csoportba történő felosztása olymódon, hogy a kapott *kereszt-partíció*, azaz a mátrix $p \times q$ darab blokkja kielégítsen valamilyen optimalitási feltételt. Podani & Feoli (1991) három ilyen "blokk-élesség" kritériumot emel ki a lehetőségek tárházából:

- a blokkok eltérésnégyzet-összege intervallum és arányskála esetén (8.3 függvény, jele legyen most $J_{(p,q)}$);
- a blokkokon belüli súlyozott entrópiaösszeg nominális karakterekre:

$$H_{(p,q)} = \sum_{i=1}^{p} \sum_{j=1}^{q} \left(k_i k_j \log k_i k_j - \sum_{h=1}^{s} f_{hij} \log f_{hij} \right) \qquad (8.4)$$

  amelyben k_i az i-edik változócsoport elemeinek száma, k_j a j objektumcsoport elemszáma, s a karakterállapotok száma (s=2, 3, ...) és f_{hij} a h-adik karakterállapot gyakorisága az ij blokkban;

- a blokkokat egy $p \times q$ méretű kontingenciatáblázat celláinak tekintjük, az ij blokkon belüli értékek összegét pedig az illető cellához tartozó gyakoriságnak (f_{ij}). Ekkor alkalmazható a 3.36 függvény, amit most jelöljünk $\chi^2_{(p,q)}$-val. A formula nyilvánvalóan megfelel bináris adatok feldolgozására, de formálisan akkor is alkalmazható, ha az adatmátrix elemei gyakoriságértékek (pl. egyedszámok).

Feladatunk az első két kritérium minimalizálása vagy a harmadik maximalizálása, mert így kapunk maximálisan homogén blokkokat. Sejthető, hogy egy nagyon nehéz problémával állunk szemben, hiszen adott n, m, p és q mellett a lehetséges táblázat-átrendezések száma a Stirling formulával (4.17) számítva $S_{(n,p)}S_{(m,q)}$, ami rendszerint csillagászati szám (a blokkokon belüli sorrend itt érdektelen). A biztosan az abszolút optimumot adó algoritmus hiányában tehát kénytelenek vagyunk valamilyen más kereső technikát alkalmazni. Podani & Feoli (1991) heurisztikus eljárása egy iteratív procedúra, amely az adatmátrixban azt a sort vagy oszlopot helyezi át minden lépésben egy másik csoportba, amelyik a legnagyobb javulást eredményezi bármelyik kritériumot alkalmazzuk is. Az iteráció akkor ér véget, ha már nincs olyan sor és oszlop, amelynek áthelyezése tovább javíthatná az eredményt. A $J_{(p,q)}$ kritérium esetében ez a módszer voltaképpen egy kétutas k-közép osztályozás. Iteratív módszerről lévén szó, a végeredményt nagymértékben befolyásolja a kiindulás, és – az adatszerkezettől függően – az iterációk könnyen konvergálhatnak valamilyen szuboptimális konfigurációba. Nincs tehát garancia arra, hogy akár több száz, különböző random kiindulásból végrehajtott elemzés meg fogja találni a legjobb megoldást. Bizonyos azonban, hogy ezek legjobbika közel lehet az abszolút optimumhoz. A módszer relatív számításigényessége a mai számítógép-korban már nem jelenthet komoly problémát még nagyméretű mátrixok esetében sem.

A bináris adatoknak az az előnye az illusztrálás szempontjából, hogy mindhárom kritériummal kompatibilisek. A p=q=2 paraméterek mellett az A1 adatmátrix binárissá konvertált adataiból 100-100 elemzést végezve, és mindegyik sorozatból az optimális eredményt kiválasztva kaptuk a 8.6 ábrán látható átrendezett táblázatokat. Az eredmények részletezése a következő:

A χ^2 statisztika alapján az iterációk egészen rossz eredménynél (χ^2=1,75, egyetlen egyszer) is megakadtak, míg a maximális χ^2-et (10,1) adó átrendeződés 42-szer jött létre (8.6a ábra) s köztes jóságú eredmények is bőven előfordultak. A *J* mérőszámnál ugyanolyan volt a legjobb átrendezés (J_{max}=12,75 mellett), de ez mind a 100 esetben kijött! Az entrópiafüggvény esetében a helyekre nézve ugyanez, a fajokra azonban egészen más az optimális eredmény (8.6b ábra, H_{max}=223,57, 62-szer). Itt viszont ehhez nagyon közeli szuboptimális értékek (224,76 ill. 225,54) is nagy számban jelentkeztek (ezek csupán az oszlopok osztályozásában tértek el a legjobbtól). Látható tehát, hogy a három kritérium nem feltétlenül vezet ugyanarra az eredményre, s ha igen, akkor sem ugyanolyan hatékonysággal, így a "legjobb" átrendezésre könnyen adódnak alternatívák. Az **a** ábra tiszta 0-ból álló blokkja, vagy a **b** ábra két, majdnem tisztán 1-esből álló blokkjai egyszerre nem jelentkezhettek az eredményben, de a többirányú vizsgálat kimutatta őket.

Érdemes megvizsgálni, hogy mely változók magyarázzák legjobban az eredményt ill. mondanak leginkább ellene annak (a változó kihagyása után számított új érték százalékában kifejezve). Az értékelés módozatait már említettük a 8.1.2 részben. A rangsor élén az átrendezést leginkább támogató, a végén pedig az ellentmondó fajok szerepelnek:

	Δchi^2 %		ΔJ%		ΔH%
1	SCA CAN -27,08	1	FUM PRO -2,52	1	SES SAD -5,83
2	CAM SIB -20,31	2	FES PAL -2,52	2	CAR HUM -7,24
3	CAR HUM -20,31	3	CEN SAD -2,52	3	CAM SIB -7,24
4	SES SAD -13,54	4	SCA CAN -2,61	4	SCA CAN -8,69
5	CAR LIP -11,38	5	CAR HUM -5,88	5	CHR GRY -9,31
6	CEN SAD -2,88	6	CAM SIB -5,88	6	CAR LIP -11,54
7	FES PAL -2,88	7	KOE CRI -7,28	7	BRO ERE -12,17
8	FUM PRO -2,88	8	SES SAD -10,46	8	SES LEU -12,49
9	SES LEU -2,16	9	SES LEU -12,04	9	KOE CRI -13,01
10	CHR GRY -1,44	10	BRO ERE -12,61	10	FUM PRO -13,01
11	KOE CRI 1,55	11	CAR LIP -18,49	11	FES PAL -13,01
12	BRO ERE 6,42	12	CHR GRY -22,69	12	CEN SAD -13,01

Annak ellenére tehát, hogy maga az átrendezett táblázat ugyanaz az első két kritériumra nézve, a fajok fontossági sorrendjében vannak ingadozások! A CAR LIP pl. erősen diszkriminál a két csoport között, tehát kihagyása viszonylag nagy χ^2 csökkenésre vezetne. Ugyanez a faj azonban a bal alsó blokk eltérésnégyzet-összegét jelentősen növeli. Az Olvasóra bízzuk a fenti táblázat további részleteinek megvizsgálását. Hasonlóképpen érdemes megadni az objektumok sorrendjét is:

a

	1	2	7	8	3	4	5	6
CAM SIB	.	1	1	1	.	.	.	.
CAR HUM	1	.	1	1	.	.	.	.
SCA CAN	1	1	1	1	.	.	.	.
SES SAD	.	.	1	1	.	.	.	.
BRO ERE	1	1	1	1	1	.	1	.
CAR LIP	1	.	.	.	1	1	1	1
CEN SAD	1	1	1	1	1	1	1	1
CHR GRY	1	1	.	.	1	.	1	.
FES PAL	1	1	1	1	1	1	1	1
FUM PRO	1	1	1	1	1	1	1	1
KOE CRI	1	1	1	1	1	1	1	.
SES LEU	1	1	1	.	1	1	1	.

b

	1	2	7	8	3	4	5	6
CEN SAD	1	1	1	1	1	1	1	1
FES PAL	1	1	1	1	1	1	1	1
FUM PRO	1	1	1	1	1	1	1	1
KOE CRI	1	1	1	1	1	1	1	.
BRO ERE	1	1	1	1	1	.	1	.
CAM SIB	.	1	1	1	.	.	.	.
CAR HUM	1	.	1	1	.	.	.	.
CAR LIP	1	.	.	.	1	1	1	1
CHR GRY	1	1	.	.	1	.	1	.
SCA CAN	1	1	1	1	.	.	.	.
SES LEU	1	1	1	.	.	1	1	1
SES SAD	.	.	1	1	.	.	.	.

8.6 ábra. Az A1 táblázat prezencia/abszencia skálára transzformált változatának blokk osztályozása a χ^2 és *J* statisztika (**a**) ill. a *H* (**b**) optimalizációjával. p=q=2.

		Δchi^2 %			ΔJ %			ΔH %
1	8	-25,92	1	5	-2,94	1	4	-11,57
2	7	-23,08	2	3	-7,19	2	6	-11,57
3	5	-18,27	3	4	-12,09	3	3	-13,24
4	3	-15,90	4	1	-13,40	4	5	-14,94
5	4	-13,70	5	7	-14,71	5	2	-17,79
6	6	-11,42	6	2	-17,65	6	8	-17,79
7	2	6,00	7	6	-17,65	7	1	-19,14
8	1	11,72	8	8	-20,26	8	7	-19,14

Itt már nem lephet meg bennünket az, hogy az első két esetben más az objektumok sorrendje, hiszen az okok ugyanazok, mint a változóknál.

Kötött blokk-osztályozás. Az osztályozásról és ordinációról szóló fejezetekben már tárgyaltunk néhány eljárást, amelyek az elemzés menetét bizonyos korlátok között tartják. A blokk osztályozásban ilyen korlátozás lehet az, ha az oszlop vagy a sorok szerinti partíciót nem engedjük megváltoztatni. Például adott a mintavételi helyek egy klasszifikációja (mondjuk sok egyéb osztályozás konszenzusaként, 9.4 alfejezet), és ehhez keressük a legoptimálisabb blokk-szerkezetet. Ekkor az elemzés során csak a sorok besorolása változhat. Fordított szituáció is elképzelhető, amikor a változók partícióját rögzítjük, és ehhez keressük a legjobb objektum-klasszifikációt (mondjuk egy határozókulcs készítésével kapcsolatosan).

Koncentráció-elemzés. Prezencia/abszencia adatok blokk-osztályozását követően lehetőségünk van a sorok és az oszlopok osztályai közötti kölcsönös megfeleltetés ordinációs elemzésére is ("*analysis of concentration*", Feoli & Orlóci 1979). Ez voltaképpen az osztályok szimmetrikusan súlyozott korreszpondencia elemzése (7.3 alfejezet) a blokkokon belüli f_{ij} összegek alábbi átalakítása után:

$$F_{ij} = \frac{\dfrac{f_{..} f_{ij}}{n_{ij}}}{\displaystyle\sum_{g=1}^{p}\sum_{h=1}^{q} \frac{f_{gh}}{n_{gh}}} \tag{8.5}$$

amelyben F_{ij} az új érték, n_{ij} pedig az ij blokk mérete. Ilymódon a blokkok méretében mutatkozó különbségeket eltüntetjük, azaz minden blokk egyformán fontos lesz (Orlóci & Kenkel 1985). A lehetséges ordinációs tengelyek száma t=min {p–1, q–1}. Az átalakított blokkok alapján számolt χ^2 (ami nem egyezik meg az iterációk során maximalizált értékkel) a következőképpen alkalmassá tehető az átrendezés relatív jóságának a mérésére:

$$RD = \frac{\chi^2}{t f_{..}} \tag{8.6}$$

("relative divergence"). *RD* értéke 0-tól 1-ig terjed, jelezve a blokkok élességét a minimális ill. maximális határ között. Ennek segítségével p és q különböző értékeire végrehajtott mátrixátrendezések közül kiválaszthatjuk a legélesebb blokk-szerkezetet mutató eredményt.

8.2.4 Blokk-szeriálás

Az előző rész módszerei csak a blokkok belső homogenitását veszik tekintetbe, a sorok és oszlopok osztályainak táblázatbeli sorrendjét szabadon választjuk meg. A blokk-osztályozás

talán legspeciálisabb módszerei viszont arra törekszenek, hogy az átló mentén elhelyezkedő blokkok és a többi közötti kontrasztot maximalizálják, s ezáltal minél egyértelműbb megfeleltetést keressenek a változók és objektumok csoportjai között (8.2 és 8.4d ábra). Ekkor tehát, mint említettük, $p=q$. Míg a kereszt-particionálásnál minden blokkot egyformán fontosnak tekintünk, az átlós szerkezetre összpontosító blokk-szeriálás[3] módszere (Marcotorchino 1991) az átlón kívüleső blokkokat gyakorlatilag egyetlen egységként kezeli. Blokk-szeriálásra leginkább prezencia/abszencia adatok esetében merül fel az igény, hogy az objektumok csoportjait minél egyértelműbben definiálhassuk a változók egy-egy csoportjával. Az **X** prezencia/abszencia adatmátrix blokk-szeriálása p csoport szerint (sorok egy csoportja A_k, az oszlopoké B_k) a Garcia - Proth (1985) féle kritérium maximalizálását jelenti:

$$GP_p = \sum_{i=1}^{p} \sum_{i\in A_k, j\in B_k} x_{ij} + \sum_{k=1}^{p} \sum_{i\notin A_k, j\notin B_k} (1 - x_{ij}) \qquad (8.7)$$

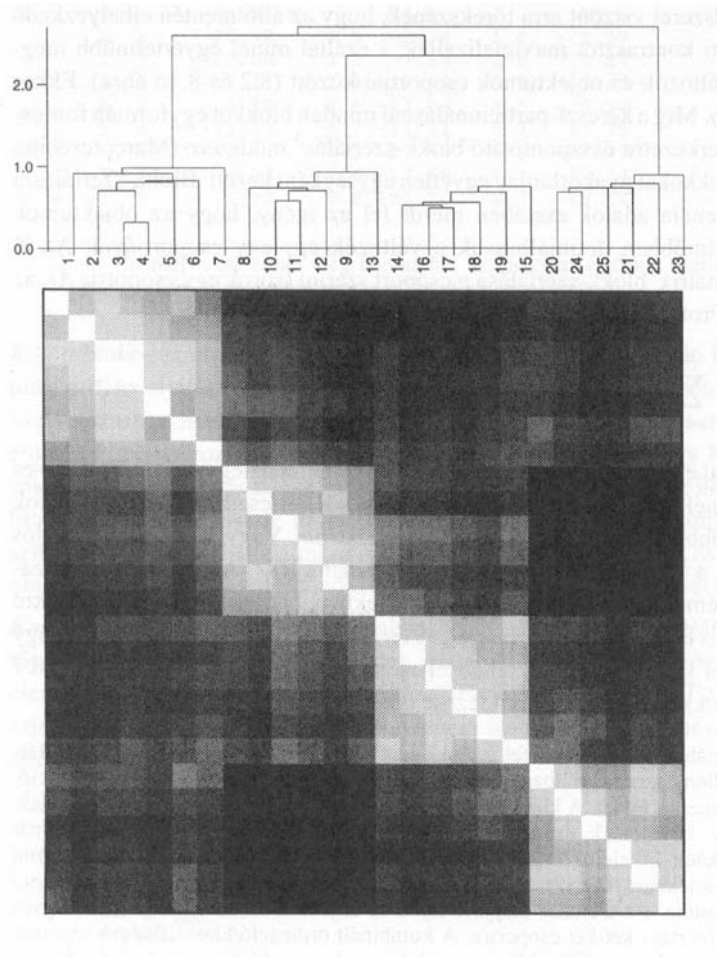
ami szavakban talán sokkal egyszerűbben kifejezhető, mint képletben: legyen minél több 1-es az átlós blokkokban, s minél kevesebb az átlón kívüliekben. Ideális esetben az átlós blokkok csupa 1-esből állanak, a többiek pedig 0-ból, ekkor GP_p értéke nm. Vagyis, GP_p/nm az átlós *blokk-élesség* egy relatív, a [0,1] intervallumba eső mérőszáma lesz. Az index optimalizálásánál éppen olyan problémákba ütközünk, mint az előzőekben, bár viszonylag kis n értékre (n<30) pontos algoritmus is ismert (Marcotorchino 1991). Elképzelhető egyébként az is, hogy az előző részek módszerei közelítőleg vagy akár teljesen optimális eredményt adnak a 8.7 függvény szerint is, s a blokkok manuális átrendezése is megoldható kicsi p esetében.

A növényökológiában akkor merül fel a blokk-szeriálás szükségessége, ha a vizsgált társulások egy fő grádiens szerint változnak, és a grádiens mentén jól elkülönülő vegetációtípusok, nodumok ismerhetők fel. A kiinduló adatok azonban nem feltétlenül bináris típusúak, hanem gyakoriságok is lehetnek, ezért a Garcia-Proth kritérium nem alkalmazható . A már említett kétutas indikátor-faj elemzés (Hill 1979a; **TWINSPAN** program, 5.3.1 rész) viszont szóba jöhet ilyen esetekben, bár nincs egy explicit módon kifejezett, a 8.7-hez hasonló optimalizációs kritériuma. Az elemzés alapja a fajok és objektumok ordinációja és az egyes tengelyek szerinti felosztása két-két csoportra. A kombinált ordináció/klasszifikáció végeredménye éppen egy átlós elrendezésű táblázat, amelyben – ha az objektumok és a fajok csoportjai között valóban egyértelmű megfeleltetés van – a blokkok jól felismerhetők, és p értékét nem is kell előre megadni. Ha a blokkos szerkezet hiányzik, akkor csupán a reciprok átlagolás szerinti szeriálást végeztük el (tartalom nélküli divíziókkal), amit a következő alfejezet tárgyal. Wildi (1989) bonyolultabb módon kombinálja az ordinációs és klasszifikációs módszereket, azzal a céllal, hogy a zaj-elemek (akár fajok, akár objektumok) minél kevésbé zavarják az átlós blokkok felismerését.

Távolságmátrixok blokk-szeriálása. Mindeddig csak adatmátrixok átrendezéséről beszéltünk, de a blokk-szeriálással kapcsolatban meg kell említenünk a távolság- és hasonlóság-mátrixokat is. Legyen **D** az m objektumra számított távolságmátrix. A feladat most az, hogy az átló mentén lévő blokkokba kis távolság- (vagy nagy hasonlóság-) értékek kerüljenek, éles kontrasztban a többi blokkal, ilymódon definiálva objektumok osztályait. Miután most az *i*-

3 A következő, 8.3 alfejezetben térünk rá a szó jelentésének tisztázására.

4 A 8.7 mérőszám persze módosítható olymódon, hogy bármilyen nem-ordinális adattípusra alkalmas legyen, a lényeg az, hogy az átlós blokkokon belüli értékek és az átlón kívüli értékek közötti eltérés maximális legyen.



8.7 ábra. A 4.3 ábra pontjai között számolt távolságmátrix átrendezése és árnyékolása az objektumok egyszerű lánc-módszerrel történt hierarchikus osztályozásával (5.6b ábra). Általában nem enynyire egyértelmű az eredmény. A rajz a **SYN-TAX** Mac programmal készült, s a 0 (fehér) és a maximális 7,87-es (fekete) távolságérték között az árnyalatok átmenete folytonos.

edik sor megegyezik az *i*-edik oszloppal, a feladat kissé más természetű, mint eddig volt: egy sor áthelyezése a megfelelő oszlop áthelyezésével automatikusan együttjár. Ez a probléma önmagában ritkán merül fel: a mátrix átrendezése rendszerint az objektumok egy korábbi osztályozására épül, s csupán *a posteriori* illusztrációul szolgál. Távolságmátrixok blokkszeriálása szoros kapcsolatban van a *mátrix árnyékolás* klasszikus problémakörével: ha a növekvő távolságértékeket egyre sötétedő színekkel helyettesítjük, akkor az átló mentén lévő blokkoknak kell a legvilágosabbaknak, a távolesőknek pedig a legsötétebbeknek lenniük[5]. A legegyszerűbb megoldás – és az általános gyakorlat – az, hogy az objektumok hierarchikus osztályozását végrehajtva rendezzük át a távolságmátrixot (8.7 ábra). Miután azonban egy dendrogram 2^{m-1}-féleképpen írható fel anélkül, hogy maga a hierarchia változna (vö. 5.1 rész, 5.2 ábra), a blokkok kialakítása részben saját megítélésünkre van bízva. Gale et al. (1984)

5 A 8.4 ábra árnyékolt blokkjai már arra az intuícióra építettek, hogy voltaképpen bármilyen blokk-osztályozás érzékeltethető árnyékolt mátrixokkal. A 8.5-6 ábrákon – bináris adatokról lévén szó – ennek még nem volt jelentősége.

megmutatta azonban, hogy a megfelelő dendrogram megtalálása a **D** mátrix (nem blokk-) szeriálásával függ össze, tehát itt az ideje, hogy a következő alfejezetre térjünk.

8.3 Szeriálás

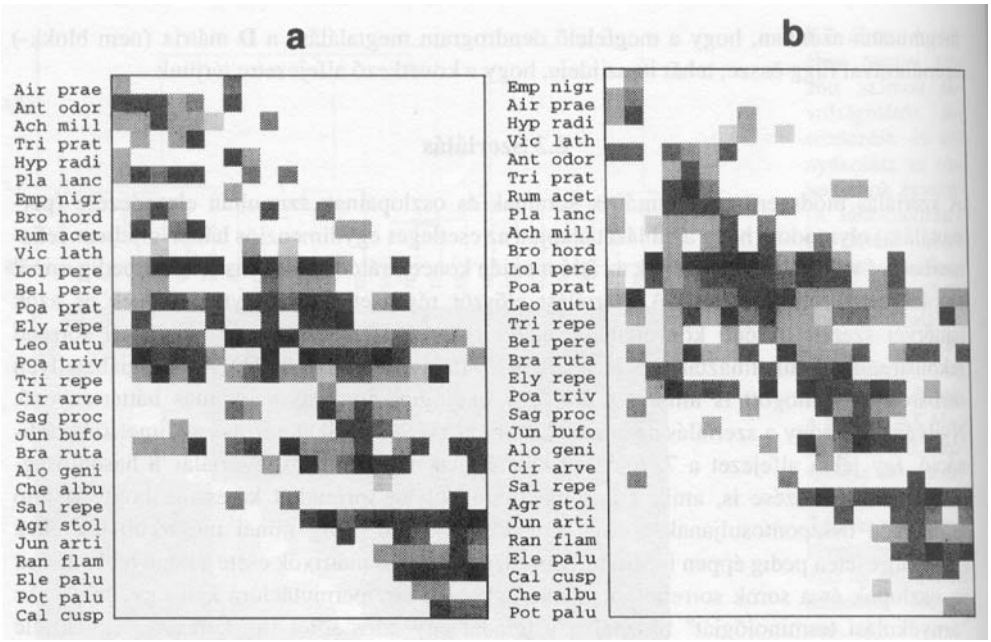
A szeriálás módszere az adatmátrix sorainak és oszlopainak szimultán elrendezése (permutálása) olymódon, hogy a táblázat alapján az esetleges egydimenziós háttér-grádiens felismerhetővé váljon: a nagy értékek az átló mentén koncentrálódjanak a kis értékek pedig minél távolabbra kerüljenek attól. A szeriálást először régészeti és őslénytani leletek és azok ismérvei szerint történő közvetett és relatív réteg-datálásra, vagyis egy időbeli sorrend rekonstruálására alkalmazták (vö. Kendall 1970, 1971, Goldmann 1971). A biológiában ilyen sorbarendezés mögött is állhat az idő, egy ökológiai grádiens vagy más háttértényező. Nyilvánvaló, hogy a szeriálás az adattáblázatra közvetlenül alkalmazott egydimenziós ordináció, így jelen alfejezet a 7. fejezet folytatásának is tekinthető. Szeriálás a hasonlóság-mátrixok átrendezése is, amikor az objektumok olyan sorrendjét keressük, hogy az átló környékén összpontosuljanak a nagy értékek, a kicsik pedig minél messzebb legyenek (távolság esetén pedig éppen fordítva). Eme szimmetrikus mátrixok esete a könnyebb, hiszen az oszlopok és a sorok sorrendje ugyanaz, így csak egy permutációra kell ügyelnünk. Az "árnyékolási terminológiát" használva a feladat egy átlós sötét sáv keresése, ami kifelé mindkét irányban fokozatosan világosodik (távolságoknál fordítva).

8.3.1 Adatmátrixok szeriálása

A sorok és oszlopok átrendezett sorrendje valamilyen ordinációs módszer alkalmazásával közvetett módon előállítható. Erre – gyakorisági adatok esetében – megfelelőnek látszik a reciprok átlagolás (vagyis a korreszpondencia analízis, 7.3.1-2 rész) módszere, hiszen ennek lényege éppen a szimultán ordináció. Ez a módszer jelenti az alapját a **TWINSPAN** tabellaátrendező programnak is. A főkomponens elemzés is számításba jöhet, hiszen az első tengelyre kapott objektum-koordináták ill. a változók korrelációinak nagyság szerinti sorrendje alkalmas lehet a sorbarendezésre, és az átrendezett táblázat előállítására. Különösen hatékony az ordináció-alapú átrendezés, ha egy erős háttér-dimenzió dominál az adatokban, azaz az első sajátérték relatíve igen nagy.

> A dünevegetáció adatok (A4 táblázat) korreszpondencia elemzéséből származó legnagyobb sajátérték 0,53 (25 %) egy "közepesen erős" háttérgrádiens jelenlétére utal. A koordináták sorrendje szerinti átrendezéssel a 8.8a ábra árnyékolt táblázatát kapjuk. A nem-zérus értékek egy meglehetősen széles sávban, de jól elkülönülten húzódnak a főátló mentén. A mindenütt előforduló fajok a táblázat közepére kerültek. A grádiens azonosításához további információra van szükség, s ehhez érdemes figyelembe venni az ugyanezen adatokból kapott CCOA értékelés eredményét is (7.17 ábra).

Az ordinációs módszerek egy adott tengely szerint, az összvariancia egy kitüntetett "irányában" rendezik át az adatokat. A táblázatban rejlő többi, lineárisan független összetevő így elvész, s egyáltalán nem mutatkozik meg az eredményben. Ez, relatíve kicsiny első sajátérték esetén nem vezet általánosabb érvényű átrendezéshez. Az ordinációval egyébként nem egy, az átrendezés "jóságát" vagy hatékonyságát mérő függvényt optimalizálunk, így a kapott eredmény csak önmagában érdekes, más átrendezésekkel nem vethető össze. A közvetlen módsze-



8.8 ábra. A hollandiai dünevegetáció-adatok A4 táblázatának átrendezése a korreszpondencia elemzés 1. tengelye szerint (**a**) és a 8.10 kritérium minimalizálásával (**b**).

rek viszont az átlós elrendezést valamilyen kritériumváltozó optimalizálásával igyekeznek megvalósítani. A feladat persze elég nagy, hiszen a megvizsgálandó sorbarendezések száma éppen $n!m!/4$ (a sorok permutációinak a száma szorozva az oszlopok permutációinak a számával majd osztva 4-gyel, hiszen a fordított sorrendek számunkra azonosak). Ez gyakorlatilag azt jelenti, hogy az összes lehetőség kipróbálására nincs mód, s ugyanakkor nem ismerünk még olyan algoritmusokat sem, amelyek bármekkora méretű adattáblázatra véges időn belül megtalálnák az optimumot.

McCormick et al. (1972) az alábbi "szomszédsági feltétel" maximalizálását javasolták az n sor és az m oszlop permutációira:

$$MC_{P(n),P(m)} = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}(x_{ij+1} + x_{ij-1} + x_{i+1j} + x_{i-1j}) \qquad (8.8)$$

amelyben $x_{0j} = x_{n+1j} = x_{i0} = x_{im+1} = 0$. Ez a függvény nyilvánvalóan csak a lokális viszonyokra koncentrál, és inkább csak az ideálishoz közeli esetekben ad elfogadható eredményt. Az optimális elrendezés kereséséhez először belátjuk, hogy a 8.8 függvény felbontható az alábbi összegre

$$MC_{P(n),P(m)} = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}(x_{ij+1} + x_{ij-1}) + \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}(x_{i+1j} + x_{i-1j}) = MC_{P(n)} + MC_{P(m)} \quad (8.9)$$

vagyis az oszlopok ill. a sorok szerinti összetevőkre. Az oszlopokkal és a sorokkal tehát különkülön foglalkozhatunk, s az átrendezett tabellát a két maximumot adó permutációk alapján

állítjuk elő. Az optimális sorrend közelítésére a szerzők az "utazó ügynök" (*traveling salesman*) problémájának egy heurisztikus megoldását javasolják.

> Elég csak a sorokról beszélnünk, hiszen az oszlopokra ugyanez az algoritmus alkalmazandó. Az eljárás lényege, hogy vesszük az 1. sort, mint kezdőelemet, s ezután egyenként megvizsgáljuk a többi sort, hogy az első elé vagy alá helyezendő-e ahhoz, hogy $MC_{P(n)}$ értékét maximálisan növelje. A maximumot adó sor megfelelő elhelyezése után a következő lépésben a megmaradt n–2 sort vizsgáljuk meg, amelyek mindegyike már 3 helyen próbálható ki. Ezután n–3 sor valamelyikének optimális helyét keressük a 4 lehetséges közül és így tovább mindaddig, amíg minden sort el nem helyeztünk a táblázatban. Mivel az eredmény csak a kezdő sor kiválasztásától függ, az egész procedúrát megismételjük n–1-szer, hogy minden sor egyszer kezdőelem lehessen. Ezután a legjobb eredményt fogadjuk el, bár távolról sem biztos, hogy ez az abszolút (globális) optimum a sorok elrendezésére. A keresést az oszlopokra is megismételjük, s a két maximumot adó permutációk segítségével felírható a végeredmény ami – még egyszer hangsúlyozzuk – nem biztos, hogy a globális optimum az illető táblázatra nézve.

Más függvények nemcsak az értékek szomszédságára ügyelnek, hanem az átrendezett táblázat úgynevezett Robinson-tulajdonságát (Robinson 1951) "figyelik". Egy mátrix akkor rendelkezik ezzel a tulajdonsággal, ha az átlótól kifelé haladva mindkét irányban monoton csökkennek az értékek a sorokban is és az oszlopokban is. Adatmátrixok esetében ez a kívánalom "benne van" az alábbi mérőszámban

$$\Psi_{(n,m)} = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}\left[\left|\frac{m\times i}{n} - j\right| + \left|\frac{n\times j}{m} - i\right|\right] \qquad (8.10)$$

(Podani 1994) amely voltaképpen minden x_{ij} értéket az átlótól vett oszlopbeli és sorbeli eltérésekkel súlyoz. Az eltérés azoknak a soroknak a száma (a j-edik oszlopon belül), plusz azoknak az oszlopoknak a száma (az i-edik soron belül) amennyit x_{ij}-nek mozdulnia kellene, hogy az átlóba kerüljön. Ez rendszerint nem egész szám. Ha tehát egy nagy érték távolesik az átlótól, akkor nagy lesz a hozzájárulása a 8.10 mennyiséghez, vagyis Ψ értékét minimalizálnunk kell. Ekkor sértjük meg legkevésbé a Robinson-feltételt is. Mivel a 8.10 összeg nem bontható fel oszlop ill. sorok szerinti komponensekre – ellentétben a 8.8-cal – nincs lehetőség a fenti algoritmus alkalmazására. A blokk osztályozáshoz (8.2.3) hasonlóan azonban eljárhatunk úgy, hogy egy kezdeti átrendezést javítunk minden iterációs lépésben a maximális javulást eredményező sor vagy oszlop áthelyezésével. Ez természetesen nagy mátrixok esetében rendkívül időigényes lesz, s a lokális optimumok nagy száma miatt sok párhuzamos futtatást kell végeznünk véletlen elrendezésű mátrixokból kiindulva (s nincs garancia a globális optimum megtalálására sem). Nem árt valamilyen ordinációs (pl. RA) alapon nyugvó kezdeti sor- és oszlopelrendezésből is kiindulni, ahogy az alábbi illusztrációban láthatjuk.

> Az RA átrendezéssel való összehasonlítás kedvéért a módszert példaképpen az A4 táblázat növényzeti adataira alkalmazzuk. 50 random kiindulást követően a legjobb eredmény Ψ = 5078 lett, az átrendezett színezett mátrixot a 8.8b ábra mutatja. Ez – legalábbis a 8.10 kritérium szerint – jobb átrendezés, mint az RA-alapú eredmény, amelyre egyébként Ψ = 5698. Ez utóbbit a módszer jelentősen tovább javította, Ψ = 5093-ig, erősen megközelítve tehát a random kiindulásokból kapott legjobb átrendezést. (Az 50 futás legtöbbje egyébként jobb eredményt adott, mint az RA módszer.) A két színezett mátrix vizuális összehasonlítása egyértelműsíti, hogy a Ψ mérőszám segítségével jobban koncentráljuk a nagy értékeket az átló mentén, mint a reciprok átlagolással, ugyanakkor – mintegy kompromisszumként – néhány kisebb érték távolabbra is kerülhet az átlótól, mint az RA elrendezésben. Nagyon lényeges

eltérés a két mátrix között nincsen, de ez nyilván adatfüggő, így mindig érdemes mindkét módszert kipróbálni.

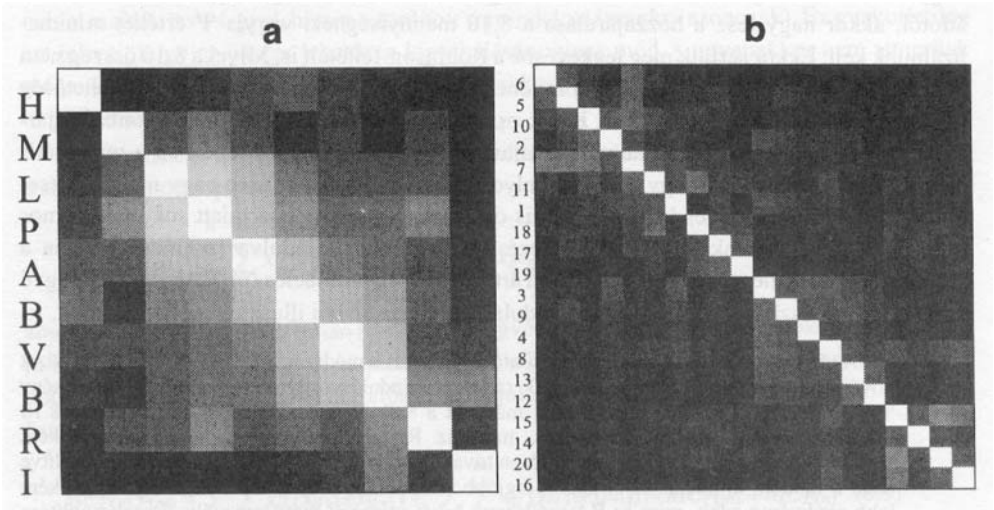
8.3.2 Távolság és hasonlósági mátrixok szeriálása

Ez annyiban könnyebb feladat az adatmátrixoknál, hogy csupán egyetlen sorrendet kell tekintenünk. Ebben az esetben is elvégezhetjük az átrendezést egy ordinációs objektumsorrend alapján. A 8.10 mérőszám is alkalmazható, annál is inkább, mert képlete szimmetrikus mátrixok esetében a következő alakra egyszerűsödik:

$$\Psi_{(m)} = \sum_{i=1}^{m-1} \sum_{j=1}^{m} s_{ij} |i - j|. \qquad (8.11)$$

Hasonlósági mátrixokra ezt ugyanazzal az algoritmussal minimalizáljuk, mint a 8.10 függvényt nyers adatok esetében, távolságmátrixokra viszont fordított a helyzet: a Ψ értékét maximalizálni kell.

Az európai nagyvárosok távolságmátrixából (A4 táblázat) éppen száz véletlenszerűen generált kiinduló elrendezést próbáltunk ki. Tekintettel arra, hogy a mátrix mérete viszonylag kicsiny, a legjobb eredmény (Ψ=354400) 64 esetben jött létre az iteratív relokáció módszerével (8.9a ábra). Az ábra jól demonstrálja azt az esetet, amikor nem lehet minden nagy értéket eltávolítani az átló közeléből, mert Helsinki, Madrid és Istanbul egy nagy háromszög csúcsaiként szerepelnek, s távolságviszonyaik egy dimenzióban nem mutathatók meg maradéktalanul. A mátrix sarkaiban lévő magas értékek az ordinácós patkó-jelenséggel analóg helyzetre utalnak. Valamivel "sikeresebb" a szeriálás az A4 táblázat objektumaira (oszlopok) számolt euklidészi távolságmátrixból (Ψ=18410; 8.9b ábra). Ötven futásból 14-szer kaptuk meg ezt az eredményt. Az objektumok sorrendisége azonban nem egyezik meg a 8.8 ábrán láthatóval, s két távoli objektum (3 és 19) közel került egymáshoz a diagram közepén, ugyancsak a patkó-jelenség analógiájaként. Az átló menti elrendeződés egyidejűleg a blokkok vizuális azonosítására is alkalmas, és az objektumok két fő csoportra történő osztályozását, s a



8.9 ábra. Európai nagyvárosok (A7 táblázat) távolságmátrixából (**a**) és az A4 táblázat oszlopaira (mintavételi helyek) számított euklidészi távolságmátrixból (**b**) végzett szeriálás a 8.11 mennyiség maximalizálásával.

baloldalinak további kettéosztását sugallja. Említettük is Gale et al. (1984) megjegyzését, miszerint a szeriálás közvetve alkalmas lehet egy osztályozás felismerésére is.

A Robinson-feltétel ebben az esetben közvetlen módon alkalmazható, amint azt Hubert et al. (1982) megmutatták. Javaslatuk szerint etalonként elő kell állítani a "Robinson mátrixot", amelynek egy x_{ij} eleme $m-|i-j|$. Az átlóban tehát minden érték m, s ettől kifelé haladva teljesen szabályos a csökkenés mindkét irányban. A legjobban átrendezett hasonlósági mátrixnak ekkor azt tekinthetjük, amelyik maximális pozitív korrelációt mutat a Robinson mátrixszal (távolságok esetében pedig a legnegatívabb korrelációt adó mátrix átrendezést keressük).

8.4 Irodalmi áttekintés

A rangsoroló és táblázat-átrendező módszerek irodalma az előző fejezetekéhez képest jóval szűkebb, s java részére már hivatkoztunk is. A változók rangsorolását tekintve Orlóci (1978) műve ajánlható elsősorban. A blokk-osztályozás legtöbb – belső blokkokat kereső – algoritmusa Hartigan (1975) könyvéből ismerhető meg. Miután a biológiában a táblázat-átrendezés igénye leginkább a cönológia/ökológia területén merül fel, a legtöbb ilyen tárgyú monográfiában bőven találunk utalásokat félig-meddig kézi és teljesen gépi ("automatizált") módszerekre is. Marcotorchino (1991) cikke és a benne található bibliográfia jó kiindulópont a blokk-szeriálás matematikai háttere iránt komolyabban érdeklődők számára. A szeriálás mindmáig legfontosabb forrásműve a Hodson et al. (1971) szerkesztette kötet, amelyben a már említett Kendall-féle cikken kívül még vagy négy további közlemény foglalkozik a problémával, igaz régészeti aspektusból. A mátrixok árnyékolásával – elsősorban növényökológiai és cönológiai alkalmazásokat bemutatva – részletesen McIntosh (1978) foglalkozik.

8.4.1 Számítógépes programok

A fejezet témájához csatlakozó számítógépes programokat is röviden "elintézhetjük". A 8.3 táblázat foglalja össze az eddig is említett programcsomagok opcióit. A **SYN-TAX** (Podani 1989c) rutinjai a változók rangsorának kiszámítása mellett az átrendezett táblázatot a változók új rangsora szerint is ki tudják menteni. A program Macintosh változata végtelen számú színárnyalatban mutatja be az átrendezett mátrixokat, akár blokk-osztályozás, akár szeriálás volt a módszer (ezzel készültek az ábrák is). A **Statistica** több színben, az értékeket kategóriákra osztva adja meg az átrendezett táblázatot a kétutas összevonás végrehajtása után.

BASIC programokat közöl változók rangsorolására és a koncentráció-elemzésre Orlóci (1978) és Orlóci & Kenkel (1985). Cönológus körökben kétségkívül a legismertebb program a **TWINSPAN**, amely – publikus forráskódja alapján (Hill 1979a) – "beépült" más programcsomagokba is (pl. **PC-ORD**, McCune 1986). Megemlíthető még a **TABORD** program (Maarel et al. 1978), melynek ma is működő változatáról nincs tudomásom.

8.5 Kérdezz - Válaszolok!

K: *Ha a rangsorolásos módszerek egy része nem is való táblázatos átrendezésre, akkor miért itt tárgyalod őket?*

V: Valóban említettem, hogy bizonyos esetekben a rangsorolás erre egyáltalán nem alkalmas. Úgy gondoltam azonban, hogy a rangsorolásos módszerekről egy helyen, együttesen érdemes egy tömör összefoglalót adni, s erre a legalkalmasabbnak mégiscsak ez a fejezet tűnt. Meg-

8.3 táblázat. Változók rangsorolása és táblázatok elrendezése különböző programcsomagokban.

	Statistica	BMDP	SYN-TAX
Eliminációs rangsorolás			+
Egyszerű rangsorolás			+
Kétutas összevonás	+		
Blokk keresés ("leader")		+	
Iteratív relokációs blokk osztályozás			+
Koncentráció-elemzés			+
Iteratív szeriálás (távolság- és adatmátrixokból)			+

győződésem ugyanis, hogy a publikált adattáblázatokban a változók sorrendjét mindig célszerű valamilyen objektív módszerrel meghatározni.

K: *Feltűnik nekem, hogy a rangsorolásoknál vagy csak a priori vagy csak a posteriori esetekről beszélsz. Nincs valamiféle köztes állapot is?*

V: Sejtem mire gondolsz: mi volna, ha az elemzés minden lépésében megnéznénk a változók fontosságát, majd ennek figyelembevételével – nagyobb súlyt adva a fontosaknak – tovább finomítanánk az eredményt, és az iterációkat addig folytatnánk, amíg a végeredmény stabilizálódik. Jó példa eme elképzelés gyakorlati megvalósítására a Jancey & Wells (1987) proponálta iteratív klasszifikációs módszer. A divizív klasszifikáció minden egyes lépésébe külön rangsorolás van beépítve, biztosítva azt, hogy minden változó azon a hierarchikus szinten kerüljön előtérbe, ahol a rangsor elején van, míg a rangsor végén levők az adott szinten zajelemként kimaradnak. Fowlkes et al. (1988) is áttekintik az ilyen célú módszereket, bár az ő javaslatuk nem tartalmaz rangsorolást, hanem a változók legjobb részhalmazának egyszeri kiválasztását minden lépésben. Ezt a beépített kiszűréses technikát "*forward selection*"-nek nevezi a szakirodalom.

K: *Tetszik nekem, hogy a blokk-struktúra kialakulásáért felelős, ill. annak leginkább ellentmondó változókat a kihagyogatásos technikával választod ki. Nem lehet ugyanezt alkalmazni – mondjuk – a változók ordináció-beli fontosságának újtípusú értékelésére is?*

V: Biztosan van rá lehetőség. Például úgy, hogy az összes változó alapján készült ordinációt (a referencia alapot) összehasonlítod az egy-egy változó kihagyásával készült ordinációkkal. Ha a referenciához a kihagyással is hasonló eredményre jutunk, akkor az illető változó kevéssé fontos, hiszen nélküle is visszakapnánk a keresett eredményt. Azok a változók azonban, amelyek elhanyagolása nagymértékben megváltoztatja az eredményt, nyilván fontosak voltak a konfiguráció kialakításában. A referenciához való összehasonlítás módjáról azonban eddig még nem szóltunk, erre majd a következő, 9. fejezetben kerül sor.

K: *Ezek szerint akkor nem is igaz, hogy a változók globális jelentősége dendrogramok esetében nem mérhető! Azt írod ugyanis, hogy csak szintenként lehet (és érdemes) az értékelést végrehajtani.*

V: Éles eszed van, erre nem is gondoltam; de már csak nem írom át újra azt a részt még egyszer! Valóban, hierarchikus osztályozásoknál is megtehetjük, hogy a vizsgálandó dendrogram mellé legyártunk *n* darab dendrogramot, egy-egy változó kihagyásával, természetesen ugyan-

azzal a módszerrel. Mindegyiket összevetjük a referenciával, s a nagy eltérést okozó változókat deklaráljuk fontosnak. Dendrogramok összehasonlítására pedig elég sokféle módszerünk van, lásd ugyancsak a következő fejezetben. Mind ordinációk, mind dendrogramok esetében elég számításigényes persze ez a munka, talán ezért nem próbálta ki még – tudtommal – senki sem eddig. Jó téma – mondjuk – egy szakdolgozathoz!

K: *Miért nem lehet koncentráció-elemzést végrehajtani akkor, ha nem bináris-típusú az adatmátrixunk?*

V: Az elemzés elvileg azokra az esetekre alkalmazható, amelyekre a korreszpondencia-analízis is, vagyis amikor a χ^2 függvény. A lényeg az, hogy adataid bináris típusúak vagy gyakoriságok legyenek. Más jellegű adatoknál egyszerűen nincs értelme az alkalmazott függvényeknek. Magát a blokk-osztályozó módszert persze lefuttathatod más adattípussal is, és kaphatsz értelmes eredményt, de a blokkok korreszpondencia-elemzésének – még egyszer mondom – nem lenne értelme.

K: *Van egy gyanum, miszerint az "utazó ügynök" módszer nem is sokkal hatékonyabb a teljes leszámlálásnál...*

V: OK, számoljunk utána. Egy sor kiválasztása után két helyre jut n–1 sor, majd három helyre kell kipróbálni n–2 sort és így tovább, míg végül n helyre illesztjük be az utolsóként megmaradt sort. Ezek összegződnek, de mivel minden sor lehet kiválasztott, az egészet n-nel meg kell szorozni. Ugyanez végrehajtandó az oszlopokra is, és a két eredményt összeadva megkapod, hogy hány lépésre volt szükséged:

$$n\sum_{i=2}^{n} i(n-i+1) + m\sum_{j=2}^{m} j(m-j+1) \qquad (8.12)$$

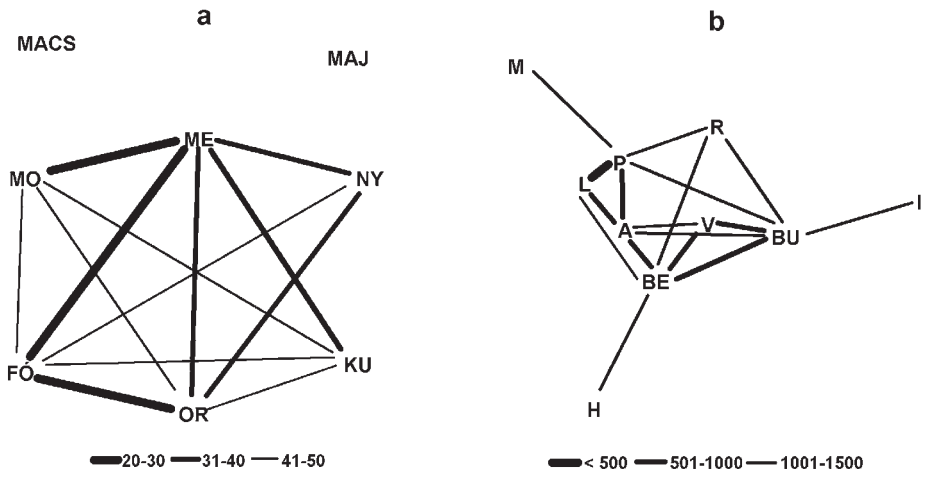
Ez pedig bizonyosan kevesebb az összes lehetséges sorbarendezések számánál. n=8 és m=7 mellett például $n!m!/4$= 50 803 200, de a 8.12 mennyiség jóval kisebb, csak 1275. Vagyis ilymódon távolról sem nézünk meg minden lehetőséget, s egyáltalán nem kizárt, hogy a globális optimumot adó sorbarendezést is elmulasztjuk. El sem képzelheted, mekkora a különbség mondjuk n=100 és m=80 esetében a teljes felmérés és a heurisztikus keresés között!

K: *A szeriálásnál nem említed, hogy az objektumok fontossági sorrendbe helyezhetők lennének aszerint, hogy mennyire támogatják az adott átrendezést.*

V: Ja, igen... Ezt elfelejtettem. A 8.10 és a 8.11 képleteket megnézve azonban beláthatod, hogy bármely sor vagy oszlop hozzájárulása Ψ-hez gond nélkül megkapható, s ennek alapján felállítható a rangsor.

K: *A mátrixok elemeinek egy színskálával történő megjelenítése akár művészi tevékenységnek is felfogható a szememben. Némely diagramot akár Mondrian is magáénak vallhatná. Hallottam azonban színezett gráfokról is. Van ezeknek valami köze a mi témánkhoz?*

V: Köze természetesen van, amennyiben az adatainkban rejlő adatstruktúra illusztrációjának egy másik lehetőségét adják, s – a tabelláris átrendezéshez hasonlóan – nem igényelnek különösebb matematikai eszközöket. A színezett gráfokat a mi esetünkben *plexus* gráfoknak nevezik, amelyeknek szögpontjai a vizsgálat objektumai, vagy a változói (pl. fajok). A közöttük futó éleket attól függően "színezzük" vagy rajzoljuk különböző vastagságúra, hogy milyen



8.10 ábra. a: Emlősök immunológiai távolságmátrixából (A5) ill. **b:** a nagyvárosok távolságmátrixából (A7) készített plexus gráfok. A távolságértékek kategorizálása mindkét esetben önkényes, de sokszor a szignifikancia-szint határozza meg a kategóriákat (pl. χ^2 és korreláció esetében, persze formálisan).

nagy a távolságuk, hasonlóságuk vagy korrelációjuk (konvenció szerint a növekvő hasonlóságot egyre vastagabb élekkel érzékeltetjük). A plexus gráf tehát az árnyékolt $n \times n$-es vagy $m \times m$-es mátrix alternatívája, s nem véletlen, hogy McIntosh (1978) ugyanabban a cikkben tárgyalta mind a két lehetőséget. A diagramot viszonylag könnyű felrajzolni kevés számú objektumra (8.10a ábra), de nagy mátrix esetében már nehéz a szögpontokat úgy elhelyezni a papír síkjában, hogy a gráf éleit könnyedén megrajzolhassuk. Segítségül kétdimenziós ordinációt alkalmazhatunk (pl. Matthews 1978, Matus & Tóthmérész 1990). A 8.10b ábrán láthatod az európai nagyvárosok 7.18 ordinációjának felhasználásával készített plexus gráfot. Jól látható, hogy a gráfban szándékosan nem húzunk meg minden élt, a nagy távolságokat ugyanis éppen az élek hiányával érzékeltetjük. A módszer elősegítheti bármely ordinációs eredmény értelmezését (pl. Whittaker 1987, Moskát 1991). Ez azonban már végérvényesen átvezet bennünket az eredmények értékeléséhez és összehasonlításához, vagyis a következő – és egyben utolsó – fejezethez (l. a különböző típusú eredmények összevetéséről szóló részt).