7

Ordináció

(A hatékony dimenzió-redukálás "művészete")

A 2.1 részben már említettük, hogy az adatmátrixnak kétféle geometriai reprezentációja képzelhető el: a változók mint dimenziók alkotta térben az objektumokat pontok képviselik, vagy fordítva: az objektumokat feleltetjük meg a tengelyeknek és ekkor a változók lesznek pontok. A 2.2 részben már meg is ismerkedtünk néhány módszerrel, amelyekkel – egyszerű módon – bepillanthatunk a sokdimenziós adatstruktúrákba. A megelőző három fejezet módszerei az adatszerkezetet speciális szempontok szerint elemzik az osztályok ill. az evolúciós mintázatok feltárásával, ezért náluk a dimenzionalitás redukciója legfeljebb közvetetten vagy rejtetten jelentkezik. Erre a fejezetre maradt minden olyan eljárás ismertetése, melyeknek már elsődleges feladata a sok dimenzió behelyettesítése kevés számú, de az eredeti adatstruktúrát többé-kevésbé jól tükröző dimenzióval. Az ezt célzó elemzéseket Goodall (1954) nyomán *ordináció* néven foglalhatjuk össze, bár a módszerek távolról sem alkotnak matematikailag egységes csoportot (pl. a többdimenziós skálázás eljárásaira "*scaling*" néven is hivatkozhatunk). Az elemzett objektumok többnyire egy halmazba tartoznak, de itt mutatjuk be a diszkriminancia-elemzést is, amelyben az új tengelyek keresésével objektumok *a priori* meglevő csoportjai között mutatkozó eltéréseket tárjuk fel. De nemcsak az objektumok, hanem a változók is besorolhatók egymástól logikailag elkülönülő csoportokba, s a dimenzió redukciót ennek figyelembevételével is végrehajthatjuk, mint például a kanonikus korreláció és a kanonikus korreszpondencia-elemzés esetében. Az ordináció fogalmát voltaképpen tehát minden más eddigi értelmezésnél tágabban fogjuk majd fel ebben a könyvben, *ordinációs módszer alatt értve minden olyan eljárást, amelyben a dimenzionalitás csökkentése mesterséges változók bevezetésével történik.* Ezeket a különféle módszerek esetében, a hagyományoknak megfelelően, más és más elnevezés illeti, mint például komponens, faktor, kanonikus tengely és így tovább.

Míg a kladisztikában a legnagyobb "szellemi megterhelést" a sok-sok, esetleg ismeretlen vagy nem eléggé tisztázott fogalom jelenti, az ordinációs metodológia elsajátításának lényeges

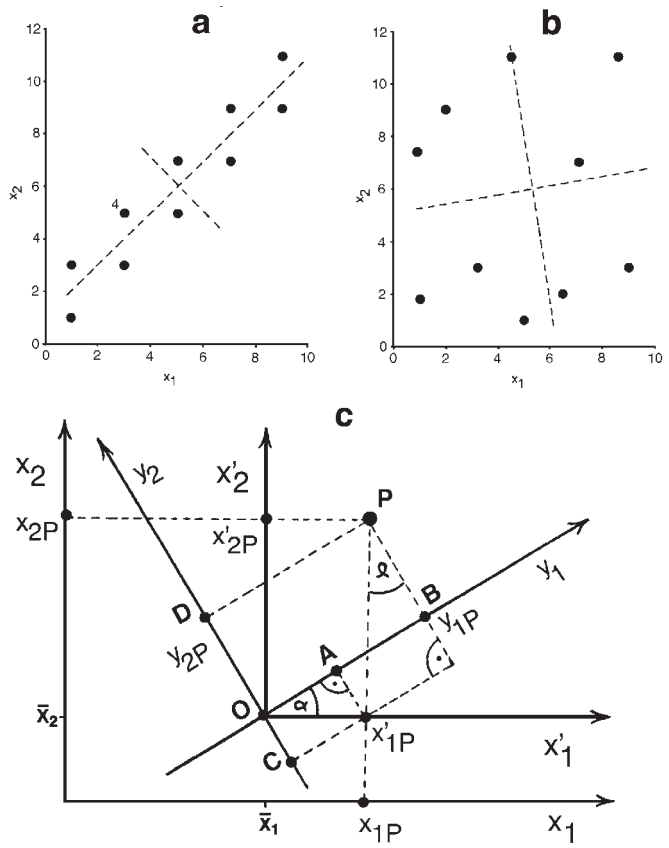
feltétele a mátrixszámítás alapjainak az ismerete. Enélkül nehezen vagy egyáltalán nem írhatók le és nem érthetők meg a legegyszerűbb módszerek alapvető lépései sem. Mindenesetre megpróbálkozunk majd a lehetetlennel, hogy a matematikai részletezés előtt minden módszerről egy intuitíve érthető bevezető jellemzést is adjunk (némi grafikus segédlettel, mivel a biológus Olvasó – mint a bevezetőben említettük – elsősorban vizuális típus). Mindenképpen javasoljuk azonban az alaposabb elmélyedést, hogy az adott problémához legalkalmasabb ordinációs módszert mindig ki tudjuk választani. Szemben a klasszifikáció eljárásaival, az ordinációk alkalmazásához bizonyos kezdeti feltételeknek is teljesülniük kell[1], ellenkező esetben az eredmények értelmezésében könnyen tévútra juthatunk.

7.1 A legfontosabb ordinációs módszer: a főkomponens analízis

A főkomponens elemzés (*principal components analysis*, általános rövidítéssel[2]: PCA) központi szerepet tölt be a többváltozós adatstruktúra-feltárásban, csakúgy mint a variancia analízis a hagyományos biometriában, így a többi módszernél részletesebben tárgyaljuk. Kifejlesztése Pearson (1901) és – elsősorban – Hotelling (1933, 1936) munkásságának köszönhető. Széles körű elterjedése és valós, nagyméretű problémákra való alkalmazása azonban csak a megfelelően gyors számítógépek kifejlesztésével vált lehetségessé. A módszer lényege többféleképpen is elmondható, számunkra a grafikus illusztráció a legjobb kiindulópont. A 7.1a ábra egy nagyon egyszerű esetet mutat be, hiszen a pontfelhő dimenzionalitása eleve kettő és ezt kell "leegyszerűsítenünk". (A könnyű illusztrálhatóság kedvéért választjuk ilyennek a példát, realisztikus esetekben persze a leegyszerűsítendő dimenzionalitás sokkal nagyobb.) Megfigyelhető, hogy a két változóra (az x_1 és x_2 tengelyre) a tíz pont összvarianciájából (3.108 egyenlet) kb. azonos rész jut (vetítsük le gondolatban a pontokat az egyes tengelyekre). Ha azonban egy teljesen új tengelyt fektetünk a pontokra olymódon, hogy az egybeessen a pontfelhő fő irányával (hosszú, szaggatott vonal az ábrán), akkor ez már az összvarianciának a jelentős hányadát megmagyarázza, míg az erre merőleges második új tengelyre csak az összvariancia töredéke jut. Ezeket az új tengelyeket nevezzük *komponens*eknek. Összefoglalva az eddigieket: a pontok helyzetét változatlanul hagyva az eredeti koordinátarendszert egy új koordinátarendszerrel helyettesítettük úgy, hogy az első új tengely (komponens) maximális varianciát sűrítsen magába, s a lehető legkevesebbet hagyja a második komponensre. A főkomponens analízisben sokkal több kiinduló változó esetén is hasonlóan járunk el: először a legnagyobb variancia-hányadot lefedő komponenst keressük ki, ezt követően a megmaradó varianciát legjobban magyarázó másodikat, és így tovább. A komponensek száma tehát nem feltétlenül kevesebb, mint az eredeti változóké volt: a variancia-hányadok "átrendezése" nem jelenti automatikusan a dimenziók számának csökkentését (a lehetséges komponensek számát l. lentebb). Az új dimenziók egy része azonban – a rájuk eső jelentéktelen variancia-hányad miatt – számunkra teljesen érdektelen lesz. Az átrendezhetőség hátterében a változók közötti pozitív (vagy negatív) lineáris korrelációk (3.70 képlet) állanak, de ez a komponensekre már nem igaz: közöttük a lineáris korreláció értéke 0. Következésképpen, ha az eredeti változók

1 Ez alól a nem-metrikus többdimenziós skálázás (7.4.2 rész) egyértelműen felmenthető.

2 Eltekintve Digby & Kempton (1987)-tól, mert ők a PCP rövidítést részesítik előnyben. A lényeg persze az, hogy a rövidítés egy könyvön belül következetes legyen.



7.1 ábra. A főkomponens analízis grafikus illusztrációja. **a:** hatékony variancia-sűrítés korrelált változók esetén; **b:** lineárisan korrelálatlan változók esetén a komponensek sem segíthetnek, **c:** vázlat a P pont koordinátáinak kiszámításához a komponensek alkotta térben. (Megjegyzés: az y_{1P} és B, valamint az y_{2P} és D egy pontra vonatkozik!)

eleve korrelálatlanok, akkor a főkomponens analízis nem eredményez lényeges változást, legfeljebb a koordináta-rendszert csúsztatja el a súlypontba (7.1b ábra). Ekkor ugyanis nincsenek "kitüntetett irányok", amelyre hatékonyabb komponenseket illeszthetnénk. A PCA "sikerességének" az tehát a feltétele, hogy a változók lineárisan korreláljanak egymással, ami biológiai objektumok esetében gyakorlatilag mindig teljesül. A PCA alkalmazásának részeredménye éppen az egymással korreláló változócsoportok azonosítása, mint a későbbiekben látni fogjuk.

Az alábbiakban először megmutatjuk, hogy a komponensek és a tengelyek közötti szögek és a pontokhoz tartozó eredeti értékek felhasználásával – némi geometriai ismeret birtokában – megkaphatók az új koordináták. A 7.1c ábrán az eredeti két változó jele x_1 és x_2, míg a komponenseket y_1 és y_2 jelöli. Az áttekinthetőség kedvéért mindössze egyetlen egy, P-vel jelölt pontot tüntettünk fel. α az x_1 változó és az y_1 komponens közötti szög. Az első lépésben az adatokat *centrál*juk (2.2 képlet), azaz minden értékből kivonjuk az adott változó átlagértékét. Ennek révén az új koordináta-rendszer origója, O, a ponthalmaz súlypontjába kerül (az x'_1 és x'_2 tengelyek metszéspontja). A P pont centrálás után kapott koordinátáit jelölje x'_{1P} és x'_{2P}, vagyis

$$x'_{1P} = x_{1P} - \bar{x}_1, \text{ ill., } x'_{2P} = x_{2P} - \bar{x}_2 \qquad (7.1)$$

A P pont koordinátáit az új tengelyeken az $\overline{OA}$ és az $\overline{AB}$, illetve az $\overline{OC}$ és $\overline{CD}$ szakaszok felhasználásával számíthatjuk ki elemi trigonometriai megfontolások révén:

$$y_{1P} = \overline{OA} + \overline{AB} = \cos\alpha\, x'_{1P} + \sin\alpha\, x'_{2P} \qquad (7.2a)$$

$$y_{2P} = \overline{OC} + \overline{CD} = -\sin\alpha\, x'_{1P} + \cos\alpha\, x'_{2P} \qquad (7.2b)$$

Mivel $\sin\alpha = \cos(90^o - \alpha)$, és a $(90^o - \alpha)$ a második változónak az első komponenssel bezárt szöge, a 7.2 összefüggések egyedül a cosinus függvény fehasználásával is felírhatók:

$$y_{1P} = \cos\alpha\, x'_{1P} + \cos(90^o - \alpha)\, x'_{2P} \qquad (7.3a)$$

$$y_{2P} = -\cos(90^o - \alpha)\, x'_{1P} + \cos\alpha\, x'_{2P} \qquad (7.3b)$$

Szavakban kifejezve a fentieket: a P pont koordinátája az y_1 komponensen egyenlő az x_1 és az y_1 közötti szög cosinusa szorozva az x'_1-en levő értékkel, plusz az x_2 és az y_1 közötti szög cosinusa szorozva az x'_2-n lévő értékkel. Az új koordinátákat *főkomponens értékek*nek ("*component scores*") nevezzük, ezek tehát a centrált adatok és a szögek ismeretében határozhatók meg. Mátrixalgebrai formában a fentieket a következőképpen írhatjuk fel:

$$\mathbf{y} = \mathbf{V}'(\mathbf{x} - \bar{\mathbf{x}}), \quad \text{vagyis} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos\alpha & \cos(90-\alpha) \\ -\cos(90-\alpha) & \cos\alpha \end{bmatrix} \begin{bmatrix} x_{1P} - \bar{x}_1 \\ x_{2P} - \bar{x}_2 \end{bmatrix} \qquad (7.4)$$

A **V** mátrix tehát egy forgató mátrix, amely a P pontot egy új koordináta-rendszerbe helyezi át. A mátrix *h*-adik *oszlopa* az eredeti változók és a *h* komponens közötti szögek cosinusait tartalmazza ("iránycosinusok"), a 7.3 szorzásokkal azonos felírásmódot tehát **V** transzponáltjával érjük el. A **V** mátrixra érvényes, hogy **V'V = I**, ami az oszlopok *ortonormalitásának* a feltétele (azaz a komponensek merőlegesek egymásra, l. C függelék).

Már "csak" annyi maradt hátra, hogy az iránycosinusokat tartalmazó **V** mátrixot kiszámítsuk. A **V** mátrix révén valójában az eredeti változók közötti kovarianciákat (és a változók varianciáit) visszük át 0 kovarianciákba (amelyek 0 korrelációknak felelnek meg) és az "átrendezett" varianciákba. Egyenlet formájában mindezt a következőképpen írhatjuk fel:

$$\mathbf{V'C\,V = L}. \qquad (7.5)$$

amelyben **C** a változók variancia-kovariancia mátrixa, **L** pedig a komponensek variancia-kovariancia mátrixa. Ez utóbbiban, a 0 kovarianciák miatt, csak az átlóban vannak pozitív értékek (diagonális mátrix), ezeket λ_h jelöli a továbbiakban. A **V** mátrix egy $\mathbf{v}_h$ oszlopvektora és a hozzá tartozó λ_h variancia kielégíti az alábbi mátrixegyenletet:

$$(\mathbf{C} - \lambda_h\mathbf{I})\,\mathbf{v}_h = \mathbf{0}. \qquad (7.6)$$

Ennek megoldásához abból indulunk ki, hogy

$$|\,\mathbf{C} - \lambda_h\mathbf{I}\,| = 0. \qquad (7.7)$$

A determináns kifejtésével λ_h-ra több megoldást is kapunk (lásd lentebb), mindegyikhez tartozik egy $\mathbf{v}_h$ vektor, amelyre a $\mathbf{v}_h'\mathbf{v}_h = \Sigma_i\, v_{ih}^2$ feltételt kell kikötnünk (vagyis a vektor hossza 1, mert csak így kapunk iránycosinusokat). E feltétel mellett – a λ_h-k ismeretében – a 7.6 egyenlet alapján megkapjuk a **V** mátrixba írt $\mathbf{v}_h$ oszlopvektorokat. A λ értékeket a **C** mátrix *sajátértékeinek*, a **v** vektorokat pedig a mátrix s*ajátvektorainak* nevezzük (C függelék).

Az elmondottakat egy teljes számítási példával illusztráljuk. A 7.1a ábrán bemutatott kétváltozós esetet vesszük alapul, amelyben a tíz pont koordinátái az alábbiak:

1. változó: 1 1 3 3 5 5 7 7 9 9
2. változó: 1 3 3 5 5 7 7 9 9 11

A két változó variancia-kovariancia mátrixa a következő:

$$= \begin{bmatrix} 8{,}89 & 8{,}89 \\ 8{,}89 & 10{,}0 \end{bmatrix}$$

(az első változó varianciája teljesen véletlenül megegyezik a két változó kovarianciájával). Először meg kell oldanunk a 7.7 egyenletet. Felírhatjuk, hogy:

$$\mathbf{C}-\lambda\mathbf{I}=\begin{bmatrix}8{,}89 & 8{,}89\\ 8{,}89 & 10{,}0\end{bmatrix}-\begin{bmatrix}\lambda & 0\\ 0 & \lambda\end{bmatrix}=\begin{bmatrix}8{,}89-\lambda & 8{,}89\\ 8{,}89 & 10{,}0-\lambda\end{bmatrix}$$

Mivel egy 2×2-es mátrix determinánsa

$$\begin{vmatrix}a & c\\ b & d\end{vmatrix}=ad-bc$$

megkapjuk, hogy

$$(8{,}89-\lambda)\,(10{,}0-\lambda)-8{,}89^2=\lambda^2-18{,}89\lambda+9{,}87=0,$$

amelynek két megoldása: $\lambda_1 = 18{,}35$ és $\lambda_2 = 0{,}54$.

Ezen a ponton megállhatunk egy pillanatra: míg a két változó varianciája 8,89 és 10,0 volt, azaz *közel azonos* részesedést jelzett az összvarianciából (47% ill. 53%), az új varianciák már jelentékenyen *eltérnek* egymástól. Az első komponens 18,35-ös varianciája az összvariancia 97%-a, így csak csekélyke 3% marad a második komponensre. A variancia-sűrítés tehát rendkívül hatékony. A sajátértékek összege, mint látjuk, egyenlő az összvarianciával, vagyis a **C** mátrix átlójába írt értékek összegével (tr {**C**}). Egy λ_h sajátérték százalékos relatív fontossága tehát $100 \times \lambda_h$ / tr{**C**}. Azt is észrevehetjük, hogy a sajátértékek szorzata (9,9) megegyezik a variancia-kovariancia mátrix determinánsával ($10 \times 8{,}89 - 8{,}89^2$), amit egyébként *általánosított varianciának* nevezünk.

A sajátértékek ismeretében már meghatározhatjuk a sajátvektorokat. Az előbb kapott két lambdára, az 1-es vektorhossz kikötésével, az alábbi mátrixot kapjuk (a számolás részleteit mellőzzük):

$$\mathbf{V}=\begin{bmatrix}0{,}685 & -0{,}729\\ 0{,}729 & 0{,}685\end{bmatrix}\ \text{azaz}\ \mathbf{V}'=\begin{bmatrix}0{,}685 & 0{,}729\\ -0{,}729 & 0{,}685\end{bmatrix}$$

Például, a 4. pont új koordinátái a 7.3ab képletek alapján, a két változó 5-ös ill. 6-os átlagaival centrálva:

$$y_{14} = 0{,}685\,(3-5) + 0{,}729\,(5-6) = -2{,}099$$
$$y_{24} = -0{,}729\,(3-5) + 0{,}685\,(5-6) = 0{,}773$$

melynek helyességéről a 7.1a ábra alapján meg is győződhetünk. Ha valaki mindezt valamely közhasználatú programcsomaggal is leellenőrzi, és azt tapasztalja, hogy a főkomponens-értékek előjeleiben eltérés van, az ne vonja mindjárt kétségbe a program használhatóságát: az előjelváltás (vagyis a tükrözés) voltaképpen teljesen önkényes. Vegyük észre, hogy a **V** mátrix vektorai mindkét irányban normáltak (azaz elemeik négyzetösszege 1). Nemcsak a sajátvektor hossza egységnyi, hanem egy adott változóhoz tartozó összes iránycosinus négyzetösszege is 1, azaz $\Sigma_h\, v_{ih}^2 = 1$. Az ortonormalitás a sorokra is fennáll: **VV'= I**.

A kétváltozós kiindulás csupán egyszerű illusztrációja volt a főkomponens-analízis számításmenetének. A leírtak természetesen érvényesek kettőnél több változóra is (a számolást persze jobb, ha a számítógépre hagyjuk). A PCA leírásakor úgy is fogalmazhatunk, hogy a komponensek az eredeti változók lineáris kombinációi (amint a 7.3 egyenletek ezt mutatták is) a fentiekben ismertetett feltételek mellett (pl. Manly 1986). A főkomponens-analízist a legkisebb négyzetek elvén működő regresszió-analízis általánosításának is tekinthetjük (Jongman et al. 1987). A 7.1a ábrán látható első komponenst voltaképpen úgy fektettük le, hogy a pontoknak az egyenestől vett távolság-négyzetösszegét minimalizáltuk. A komponens és az eredeti tengelyek közötti szög pedig – a sajátérték számítást megkerülve – egy iterációs eljárással is meghatározható.

Meg kell említenünk a komponensek lehetséges számát is. A 7.7 egyenlet megoldása t pozitív sajátértéket eredményez, az alábbi korlátozással:

$$t \leq \min \{n, m-1\} \quad (7.8)$$

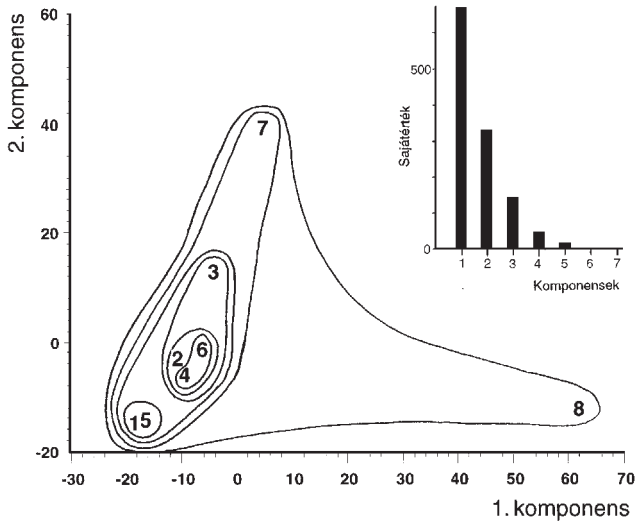
ahol – az eddigieknek megfelelően – n a változók száma, m pedig a pontok (objektumok) száma. Abban esetben tehát, amikor a pontok száma meghaladja a változókét, legfeljebb n komponens vonható ki az adatokból. Ennél természetesen kevesebb is lehet, mégpedig akkor, amikor a változók közötti kapcsolatrendszerben egyértelmű függvénykapcsolatok jelentkeznek (az eredeti változók valamelyike előállítható a többi változó lineáris kombinációjaként). Ha pedig $m - 1 < n$, akkor már ez dönt a maximális komponensek számát illetően. Ennek pedig az az oka, hogy a sok változó mintegy "túldefiniálja" a kis számú pontot, hiszen m pont távolságviszonyainak a feltüntetéséhez maximum m–1 dimenzióra van csak szükség. (Két pont távolsága az egyenes mentén – azaz egy dimenzióban – hűen ábrázolható; három pont távolságait két dimenzióban tökéletesen feltüntethetjük, míg négy ponthoz három dimenzió elegendő, és így tovább.) A pozitív sajátértékek számát, vagyis t-t, a **C** mátrix *rangjának* nevezzük, s ez valójában az adatokban rejlő háttér dimenziók száma, az adatrendszer *valós dimenzionalitása* (C függelék).

Felmerül a kérdés, hogy a valós dimenzionalitáson belül hány, valóban értelmes – vagy legalábbis annak tűnő – dimenziónk van, s mennyi az elhanyagolható, csupán a sztochasztikus ingadozást magyarázó dimenziók száma. Ha a sajátértékeket nagyság szerint sorba rendezzük, akkor a dimenziók relatív fontossága is érzékelhetővé válik, amelynek grafikus illusztrációja pl. egy lejtős oszlopdiagram ("*scree diagram*", Cattell 1966, lásd a 7.2 ábrát) lehet. A javaslat szerint keressük meg azt a pontot a diagramon, ahonnan a sajátértékek már csak nagyon lassan csökkennek tovább. Az azt megelőző dimenziók tehát "fontosak" lennének, a többiek pedig nem. Ez a szubjektív módszer – bár egyes esetekben még működhet is – semmiképpen sem ajánlható általános megoldásként. A Kaiser-féle kritérium (Mardia et al. 1979) szerint viszont csak azokat a komponenseket kell figyelembe vennünk, amelyekhez az átlagosnál nagyobb sajátérték tartozik. Eszerint általában jóval kevesebb az "értelmes" komponensek száma, mint a Cattell-módszer alapján. Statisztikai próbára is van módunk, de ennek feltétele a kiinduló adatok normális eloszlása: a Bartlett-féle izotrópia-teszt révén a kisebb sajátértékek eltéréseit vizsgáljuk, s így keressük meg azt a dimenziót, amelyet követően a sajátértékek közötti különbség már nem szignifikáns (Mardia et al. 1979).

A főkomponens elemzést először az A1 táblázat cönológiai adataival illusztráljuk, amelyben nyolc kvadrátot 12 faj borításértékei jellemeznek. A PCA variancia-sűrítő hatása szembeszökő: az első két komponens (7.2 ábra) az összvariancia 82 %-át magyarázza (55 ill. 27 %), míg a többiekre jutó rész már szinte elhanyagolható. A nyíltabb gyepek felvételei viszonylag közel állanak egymáshoz, és a zártabb gyepek (a 7-es, és különösen a 8-as felvétel) távoli pozíciója jól tükrözi a fajok borításában jelentkező mennyiségi eltéréseket. Az eredmény – az ábra tanúsága szerint – jól egybevág a nyers borításértékekből számolt euklidészi-távolságok hierarchikus osztályozásával.

7.1.1 Komponens-kovariancia és -korreláció: a változók ordinációja

A főkomponens analízis rendkívüli előnye, hogy a pontok közötti távolságviszonyok elemzésén, a variancia-sűrítésen túlmenően a változók kapcsolatrendszerének alaposabb értékelésére is alkalmas. A komponens értékekből és az eredeti adatokból a 3.69 formula alapján pl. kiszámíthatjuk a komponensek és a változók közötti kovarianciát, ami lehetőséget ad a komponensek – mint matematikai konstrukciók – konkrét, a valós problémának jobban meg-



7.2 ábra. Az A1 táblázat cönológiai felvételeinek ordinációja a főkomponens-elemzés segítségével. A jobb áttekinthetőség kedvéért a tengelyek nem az origóban metszik egymást (s a további ábrákon is így lesz). A kontúrok a hierarchikus osztályozás egyes lépéseit jelzik. Az oszlopdiagram a hét sajátérték relatív nagyságát tükrözi.

felelő interpretációjára is. A kovariancia képletére azonban itt nincs is szükségünk, mert a 7.6 összefüggésből következően:

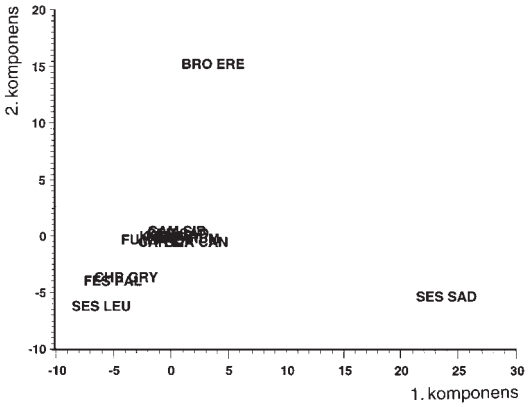
$$\mathbf{Cv} = \lambda\mathbf{v}. \tag{7.9}$$

amelyben $\lambda_h v_{ih}$ felel meg az *i* változó és a *h* komponens közötti *kovarianciának*. A kovariancia standardizálása a változó szórásával *(s_i)* és a komponens szórásával (a komponens varianciája a sajátérték, a szórás tehát nyilván $\sqrt{\lambda_h}$) adja a keresett, ún. *komponens korrelációt*:

$$r_{ih} = \lambda_h v_{ih}/s_i\sqrt{\lambda_h} = v_{ih}\sqrt{\lambda_h}/s_i \tag{7.10}$$

(más néven: főkomponens-súly, "*loading"*). A komponens-kovarianciák és -korrelációk alkalmasak a *változók ordinációs diagramjának* a felrajzolására. A tengelyek ekkor, ugyanúgy, mint az objektumok esetén, az egyes komponenseknek, a pontok viszont az eredeti változóknak felelnek meg. Mivel a változók egymás közötti összefüggéseit – eddig legalábbis – a kovariancia alapján fejeztük ki, logikus, hogy a változók és a komponensek kapcsolatát is hasonlóképpen mérjük. Az *i* változó koordinátája a *h* tengelyen tehát $\lambda_h v_{ih}$ lesz, és elvileg bármekkora értéket felvehet. A korreláció alkalmazása sem teljesen érdektelen, de ez inkább a standardizált PCA – 7.1.4 rész – esetén interpretálható jobban. Mindenesetre megjegyezzük, hogy korreláció esetén egy pont koordinátái nyilván a [–1,1] intervallumba esnek. Miután a komponensek egymásra merőlegesek, egy változó nem korrelálhat akárhogyan velük: a pontok szükségképpen az origóra felrajzolt egységsugarú körön belülre kerülnek (az összes komponensre: az egységsugarú hipergömb felületére; vagyis $\sum_{h=1}^{i} r_{ih}^2 = 1$. Megjegyzendő még, hogy r_{ih} éppen a *h* komponens és az *i*-edik pontra mutató vektor közötti szög cosinusa.

A komponens-kovarianciák és -korrelációk alapján nyert faj-ordinációk összehasonlítását elvégezhetjük az A1 táblázat elemzésével. A 7.3 ábra diagramja a változókat a komponensekkel adott kovarianciájuk szerint rendezi el. Látható, hogy a komponenseket elsősorban a legalább egy helyen nagy borításértékű (azaz nagy varianciájú) fajok határozzák meg, az elsőt a *Sesleria* a másodikat pedig a *Bromus*. Mindkettővel többé-kevésbé ellentétes tendenciát képvisel a *Seseli*, a *Festuca* és a *Chrysopogon*, míg a még kisebb borítású fajok az origóban

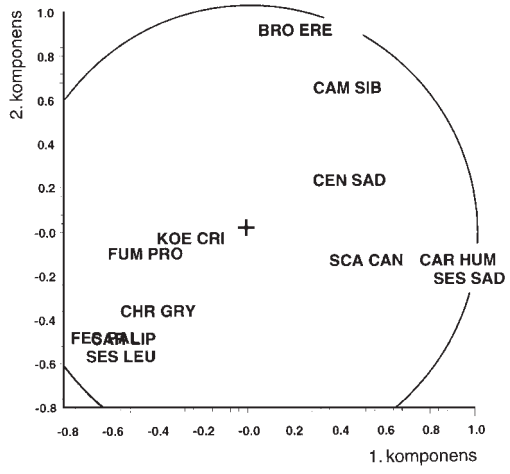


7.3 ábra. Az A1 táblázatban szereplő fajok ordinációja a komponens-kovarianciák alapján. A pontok a fajnevek közepén helyezkednek el. Több név teljesen átfedi egymást az origó körül, ezek a kis varianciájú fajokat jelzik.

tömörülnek. A korrelációk alkalmazásával a borításbeli különbségek "eltűnnek" és a pontok jobban szétterülnek (7.4 ábra). Az 1. komponenssel erős pozitív korrelációt a zárt gyep fajai, negatívat pedig a nyílt gyep fajai adnak. Ez a legerősebb kontraszt, a legfőbb trend, amit az egyszerű adattáblázatból leszűrhetünk. A *Bromus erectus* azonban láthatóan függetlenségre "törekszik" a többi fajjal szemben, hiszen mindenütt megjelenhet kisebb vagy egészen nagy borítással is. Ezt a 2. komponenssel adott magas korrelációja tükrözi. Az origóhoz közel eső fajok *(Koeleria, Scabiosa, Centaurea)* pedig a legkevésbé függenek a többitől, ezek szerepe elhanyagolható a társulások elkülönítésében is. Általában elmondható, hogy *minél közelebb van a pont a kör kerületéhez, annál teljesebb mértékben megmagyarázza az illusztrált két komponens az illető változó viselkedését korrelációs értelemben.*

7.1.2 Százalékos hozzájárulások

Adott komponens szerepe egy objektum vagy egy változó megmagyarázásában a grafikus ábrázoláson túlmenően kvantitatíve is kifejezhető. A *k* objektum "varianciájának" (pontosabban az origótól vett eltérésnégyzet-összegének) a *h* komponens az alábbi százalékban felel meg:



7.4 ábra. Az A1 táblázatban szereplő fajok ordinációja a komponens-korrelációk alapján. A pontok a fajnevek közepén helyezkednek el. Hasonlítsuk össze a 7.3 ábra elrendezésével!

7.1 táblázat. Az A1 táblázat adataiból kapott elsõ öt komponens százalékos részesedései a felvételekre és a változókra. Minden sor legmagasabb értékét vastag szedés emeli ki.

Felvételek	Komponensek				
1	32,788	20,828	**39,445**	6,354	0,575
2	**49,861**	6,096	0,196	0,723	42,499
3	7,354	**63,758**	12,651	13,412	2,024
4	33,613	12,784	**38,315**	5,207	6,086
5	**41,815**	30,638	1,654	25,245	0,035
6	9,825	0,455	**82,130**	6,476	0,006
7	1,258	**92,687**	5,169	0,872	0,000
8	**96,230**	3,636	0,134	0,000	0,000
Fajok					
BRO ERE	4,488	**83,167**	11,954	0,351	0,028
CAM SIB	19,279	**42,352**	10,104	3,185	18,178
CAR HUM	**85,061**	1,790	11,184	0,361	0,460
CAR LIP	**30,611**	23,965	3,248	8,452	17,107
CEN SAD	**19,884**	5,256	7,073	6,169	13,545
CHR GRY	15,597	12,942	**48,084**	23,269	0,108
FES PAL	**39,173**	23,352	16,754	19,037	1,473
FUM PRO	20,446	0,968	**42,720**	1,465	26,912
KOE CRI	6,194	0,099	**67,650**	17,206	7,508
SCA CAN	26,722	1,818	2,256	0,417	**63,037**
SES LEU	31,618	**32,224**	28,702	5,829	1,436
SES SAD	**94,276**	4,752	0,925	0,038	0,003

$$z_{hk} = 100 \times y_{hk}^2 / \sum_{j=1}^{t} y_{jk}^2 \qquad (7.11)$$

(y_{hk} a k objektum értéke a h komponensen). Ennek alapján megállapítható, hogy az egyes pontok helyzetéért hány komponens felelős, azonosíthatók a kevés objektumot befolyásoló komponensek, és a centroidhoz közeli, "átlagos" objektumok is kikereshetők. Az i változó varianciájának az a része, amelyet a h komponens magyaráz, a következő:

$$w_{hi} = 100 \times v_{ih}^2 \lambda_h / \sum_{j=1}^{t} v_{ji}^2 \lambda_j \qquad (7.12)$$

Ha az i változó és a h komponens közötti korreláció magas, akkor a fenti módon kiszámított százalékérték is magas lesz, és a többi komponensre már csak kis rész juthat. Ha egy változóra közel egyenletes százalékértékek jutnak, akkor ezt a változót az adatstruktúra jellemzéséből akár ki is hagyhatjuk.

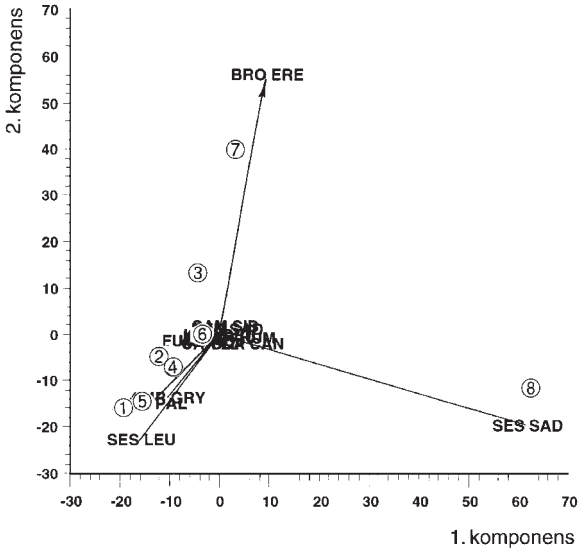
Az Olvasóra bízzuk a példaadatokra vonatkozó százalékértékek (7.1 táblázat) és a 7.2-3 ábrák összevetését. Mint látható, egyes kvadrátok ill. fajok értelmezésében az eddig mellőzött komponensek is számításba jöhetnek. A *Scabiosa canescens* varianciájának jelentős része pl. az 5. komponensen fejeződik ki, mutatva ennek a fajnak az egyedi viselkedését (kicsi vagy nagy borítása is lehet a nyílt és a zárt gyepben egyaránt). Ez azonban *összhatásában* szinte elhanyagolható, hiszen az 5. komponensre csak 1,5 % variancia jut.

7.1.3 Objektumok és változók együttes ordinációja: a "biplot"

Felmerül az a lehetőség, hogy az objektumok és a változók külön-külön felrajzolt diagramjait valamilyen formában egyesíteni kellene. Gabriel (1971, 1981) javasolta először, hogy a megfelelő módon meghatározott koordináták alapján készüljön egy olyan diagram, amely mind az objektumokat, mind pedig a változókat feltünteti, egy diagramba sűrítve minden lényeges információt. Az egyesített diagramot *biplot*nak (=kettős szórásdiagram) nevezzük (a *bi-* előtag nem a bemutatott dimenziók számára, hanem az egyesített ordinációk számára utal!). Miután rendszerint a változók koordinátáit más skálán vesszük fel, mint az objektumokét, a változók koordinátáit egyszerűen beszorozzuk egy alkalmasan megválasztott számmal, hogy a két ordináció felrajzolható legyen. A változókat képviselő pontokhoz pedig az origótól kiinduló nyilakat húzunk, amelyek majd megkönnyítik a diagram értelmezését.

A biplot szerkesztésére többféle módszer ismeretes. A Gabriel-féle eredeti javaslat értelmében a biplot nemcsak egyszerűen két ordináció egymásra vetítése, hanem a változók és az objektumok olyan ábrázolása, amely lehetővé teszi az eredeti adatok – a sajátértékek százalékos "sikerétől" függő – rekonstruálását. A Gabriel-féle biplotban az objektumok koordinátáit az ismert módon kapjuk meg, míg a változók koordinátáit a sajátvektorok megfelelő értékei (az iránycosinusok) adják, beszorozva egy önkényes skála faktorral. (Erre az esetre ter Braak (1983) az *euklidészi biplot* elnevezést alkalmazza, mert az objektumok között az euklidészi távolságokat közelítjük.) A változóra mutató nyíl és egy komponens hajlásszöge annál kisebb, minél inkább egybevág a komponens a változóval a sokdimenziós térben.

Az A1 táblázat elemzéséből származó euklidészi biplotot a 7.5 ábrán láthatjuk. A kisborítású, ill. kis varianciájú fajok az origó körül tömörülnek, úgyhogy a fajnevek nem is ismerhetők fel a rajzon. A fajok relatív elhelyezkedése rendkívül hasonló a kovariancia-szerinti ordinációhoz (7.3 ábra). Ennek megfelelően a *Bromus*, a *Sesleria* és a *Seseli* hatása erőteljesebben jelentkezik az eredményen, mint a többi fajé. A *Sesleria* a nyúlfarkfüves zárt gyep, a *Bromus* a rozsnokos zárt gyep, a *Seseli* pedig a nyílt gyep "irányába" mutat. Jelentős még – és az ábrán is jól látszik – a *Chrysopogon gryllus* és a *Festuca pallens* ugyanilyen irányban kifejtett hatása.



7.5 ábra. Euklideszi biplot: a 7.2 ordinációra a fajokat a sajátvektorok szerint vetítettük (a sajátvektorokat beszorzó konstans: 67,05, a fajok koordinátáit tehát megkapjuk, ha a tengelyeken megadott értéket osztjuk e konstanssal).

Az adatrekonstruálás, a fenti példában a fajok abundanciáinak illesztése a következőképpen értendő. A *j* kvadrátból az *i* fajt képviselő nyílra irányított merőleges egyenessel az ún. *illesztett komponens értéket* ("*fitted score*") kapjuk meg (az önkényes szorzóval való osztás után, természetesen), amely az *i* faj *j* kvadrátbeli borításának egyfajta becslése. Ha az illesztett érték pozitív, akkor az adott faj az átlagnál várhatóan magasabb értékkel rendelkezik az illető kvadrátban, ha negatív, akkor pedig az átlagnál kisebbel (ui. az adatokat az elemzés során *centráltuk*). Ez az illesztés annál jobb, minél nagyobb variancia hányadot összesít a két bemutatott komponens. (Az első két komponensre kapott euklidészi biplot alapján végzett illesztés a legjobb közelítést adja az adatokhoz. Lásd C függelék: szinguláris érték-felbontás, [C50 összefüggés], amelyben **US** felel meg az objektumok koordinátáinak, **V'** pedig a sajátvektoroknak).

> A 7.5 ábrán például hosszabbítsuk meg képzeletben a *Bromus*-ra mutató egyenest az origón túl, erre nézve negatív irányba. Az 1., 2., 4., 5. és 6. felvételek adnak az átlagos borításnál (11,5) kisebbet, úgyhogy az illesztett érték jól egybevág az eredetivel. A 8. felvételre (11-es értékkel) ez kevéssé igaz, mert már a pozitív régióba került. A 3. és a 7. pontok pedig az átlag felett vannak, erre a fajra nézve ugyancsak egybehangzóan az eredeti adatokkal.

Egy másik lehetőség az ún. Mahalanobis biplot, amelyhez az objektumok koordinátáit a következőképpen számítjuk át: $u_{hj} = y_{hi} / \sqrt{(\lambda_h \times (m-1))}$. Az átalakítás eredményeképpen minden egyes komponensen egyforma lesz a variancia, mert a koordináták vektora egységnyi hosszúságú. (Amit most kaptunk, az éppen az adatmátrix szinguláris érték-felbontásából származó baloldali mátrix egy eleme, vö. a C50 képlettel a C függelékben). A pontok közötti távolságok most a Mahalanobis-féle általánosított távolságokat tükrözik (az összes komponensre) ill. annak legjobb közelítései az első két komponensen. A biplotban a változók koordinátáit a komponens-kovarianciák (7.9) adják, a vektorok hossza a változó szórásával arányos, a közöttük levő szögek pedig a változók közötti korrelációkkal arányosak. Az adatok a szinguláris érték-felbontás értelmében itt is rekonstruálhatók (**U** felel meg az objektumok koordinátáinak, **SV'** pedig a változókénak a C50-ben).

Bizonnyal felmerül ezen a ponton az Olvasóban a kérdés, hogy miért nem használhatjuk a biplot elkészítésében az objektumok euklidészi távolságait közelítő ordinációt (az euklidészi biplotból) és a változók közötti kovarianciákat (a Mahalanobis biplotból), vagyis miért ne vetíthetnénk egymásra a 7.3 és a 7.2 ábrát? Ezt természetesen megtehetjük, s eredményül egy meglehetősen szemléletes diagramot fogunk kapni (illusztrációt a korrelációra mutat be majd a 7.6c ábra). Számos szerző azonban ezt nem tekinti valódi biplotnak, mert az eredeti adatok előállítására (helyesebben becslésére) nem alkalmas. Ennek ellenére Rohlf (in: Marcus 1993) megvizsgálásra érdemesnek tekinti ezt a lehetőséget. Magam is ezen a véleményen vagyok, s a továbbiakban Rohlf-biplot néven fogok hivatkozni erre. A biplot fontossága ugyanis nem feltétlenül abban van, hogy az optimális adatrekonstrukciót adja, hanem abban, hogy az ordinációk egymásra vetítése mennyire könnyíti meg az egyik értelmezését a másikkal és viszont. Sok esetben egyébként nincs lényeges különbség az euklidészi- és a Rohlf-biplot között, mert a változók sajátvektorok, ill. a kovarianciák szerinti ordinációja nem tér el lényegesen (mint a jelen példában sem).

7.1.4 Standardizált PCA

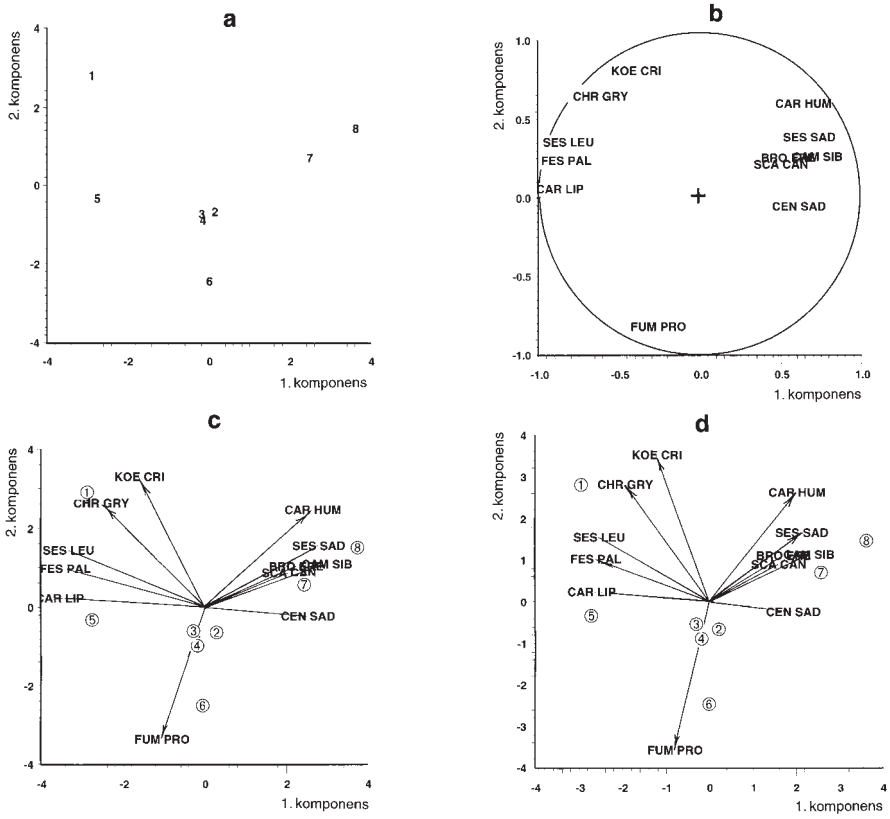
Az eddigiekben a főkomponens-analízis "alapváltozatát" mutattuk be, melyben az egyedüli adat-átalakítási művelet a centrálás volt (innen a név: *centrált* PCA). Ennek hatását láttuk is az eredményben: a főkomponensek helyzetét a nagy varianciájú változók (a példában: fajok) határozták meg elsősorban, a kis varianciájú változók rovására. Gyakran előadódhat azonban olyan eset, amikor ezt el szeretnénk kerülni, és minden változót egyforma súllyal akarunk szerepeltetni az elemzésben (pl. minden fajt egyformán fontosnak tekintünk, függetlenül az abszolút borításértékektől). Amennyiben az egyes változókat eltérő jellegű skálán mérjük, akkor pedig egyenesen "kötelező" lesz az egyforma súlyozás: ellenkező esetben az analízis eredménye csupán a mérési skálák önkényes megválasztásának a tükrözője lesz. Ekkor a főkomponens-elemzés a 2.4 standardizálást alkalmazza, vagyis a centráláson túlmenően minden értéket a szórással is elosztunk. Ennek hatására természetesen megváltozik a pontfelhő alakja a sokdimenziós térben, de ilyenkor is lesz értelme a fő irányok megkeresésének.

Az ún. *standardizált* PCA a 7.1.1-3 részben leírtaktól a következőkben különbözik:

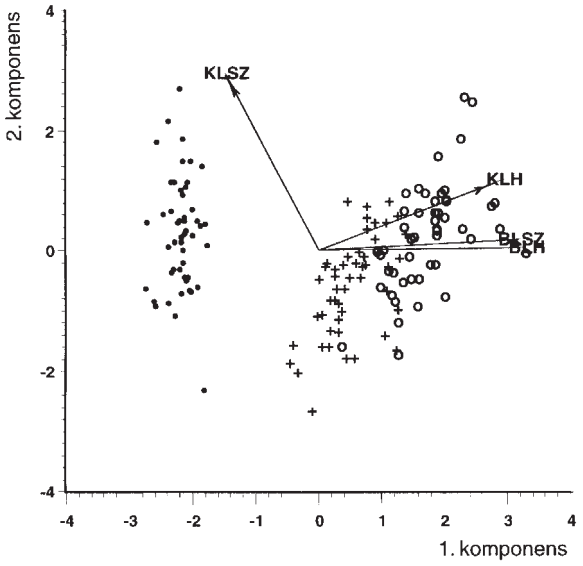
- A kiindulás a variancia-kovariancia mátrix helyett a változók **R** korrelációs mátrixa (ez a standardizálást eleve tartalmazza). A nyers adatok 2.4 szerinti standardizálását tehát felesleges előre elvégezni.
- Mivel a standardizálás révén minden változó varianciája 1 lesz, az összvariancia értéke n, csakúgy, mint a sajátértékek összege. A h komponens %-os részesedése az összvarianciából ennek megfelelően $100 \times \lambda_h/n$.
- Az 1-nél kisebb sajátértékekhez tartozó komponenseket nyugodtan kihagyhatjuk az eredmények értékelése során, hiszen ez a variancia kisebb, mint bármely (standardizált) változóé.
- A h komponens és az i változó korrelációja az $r_{ih} = v_{ih}\sqrt{\lambda_h}$ alakra egyszerűsödik (s voltaképpen megegyezik a kovarianciával, így majd felhasználható lesz a Rohlf-biplot elkészítésére).

A változók kiegyenlített hatását az eredményre az A1 táblázat standardizált főkomponens elemzése mutatja (7.6 ábra). Az első két komponens ekkor már "csak" 63 %-nak felel meg, szemben a kovarianciákból kapott 82 %-kal. A felvételek helyzete is megváltozik, a 7-es és a 8-as közelebb kerültek egymáshoz mert a kettejük különbségét elsősorban okozó *Bromus* és *Sesleria* hatását csökkentettük. Ezzel szemben az 1. és 5. felvételek eltávolodtak egymástól, mert a közöttük mutatkozó kisebb eltéréseket a standardizálás felnagyította (7.6a ábra). Az 1. komponens egyértelműen a nyílt-zárt gyep ellentétet tükrözi. A standardizálás révén a fajok közötti csoportosulások is megváltoztak: a nyílt-zárt gyep szembeállásnak megfelelően két fő csoportot figyelhetünk meg (7.6b ábra). A *Fumana* "különállása" azt jelenti, hogy ez a faj egyik csoporthoz sem "vonzódik", nem korrelál semmivel sem. A korrelációs Rohlf- (7.6c ábra) és az euklidészi (7.6d ábra) biplot közötti kismértékű eltéréseket a $\sqrt{\lambda_h}$ szorzó megléte ill. hiánya okozza (lásd a fenti felsorolás utolsó pontját).

Egy cönológiai elemzésben magunk döntjük el, hogy alkalmazzuk-e a standardizálást, vagy sem, de a taxonómiai vizsgálatokban értelmes eredményre általában csak a standardizált PCA vezethet. Így van ez az *Iris* adatok esetében is (A2 táblázat), hiszen a hosszúság-értékek többszörösei a szélességeknek. Az elemzés kimutatja, hogy míg a belső lepel szélessége és hosszúsága szinte tökéletesen korrelál, a külső lepelre vonatkozó méretek korrelálatlanok (a rájuk mutató két nyíl között csaknem derékszög van, 7.7 ábra, vö. a 2.3 ábrával). Az első komponens gyakorlatilag a belső lepel méretével függ össze, a második pedig leginkább a külső



7.6 ábra. Az A1 cönológiai adattáblázat értékelése standardizált főkomponens analízissel. **a:** felvételek ordinációja, **b:** fajok ordinációja a komponens-korrelációk alapján, **c:** korrelációs Rohlf biplot, **d**: euklidészi biplot.



7.7 ábra. Az A2 táblázat egyedeinek standardizált főkomponens elemzéséből kapott korrelciós Rohlf biplot. Jelek: •: *I. setosa*, +: *I. versicolor*, o: *I. virginica*. Rövidítések: KLH: külső lepel hossza, KLSZ: külső lepel szélessége, BLH: belső lepel hossza, BLSZ: belső lepel szélessége.

lepel szélességével (ez így is van: a táblázat alapján láthatjuk, hogy az *I. setosa* belső lepellevelei sokkal kisebbek a másik két fajnál). A két komponens 73 ill. 23 %-ot magyaráz az összvarianciából, azaz minden lényegeset. Míg azonban az első komponens többé-kevésbé alkalmasnak tűnik a fajok elválasztására, a második már egyáltalán nem (az *Iris* fajok elkülönülését majd még egyszer megvizsgáljuk a diszkriminancia-elemzés segítségével).

7.1.5 Nem-centrált PCA

A kovarianciák, ill. a korrelációk mátrixából végrehajtott PCA elemzések megegyeznek az adatok centrálásában, melynek eredményeképpen a komponensek a pontfelhő súlypontjában, az "átlag"-ban metszik egymást. A centrálás azonban ki is hagyható, ha a változók **K** keresztszorzat mátrixából indulunk ki (3.68 összefüggés). Ezt a típusú elemzést *nem-centrált főkomponens analízis*nek nevezzük[3]

- A centrálás elmaradása azt jelenti, hogy a komponensek szükségképpen az eredeti koordináta-rendszer origóján mennek át. E megszorítás miatt a komponensek általában *nem ortogonálisak*, a közöttük lévő korreláció 0-tól jelentékenyen eltérhet.
- A variancia helyett a pontok origótól vett eltérésének *négyzetösszegét* mérjük, ezért a sajátértékekre a következő összefüggés lesz érvényes:

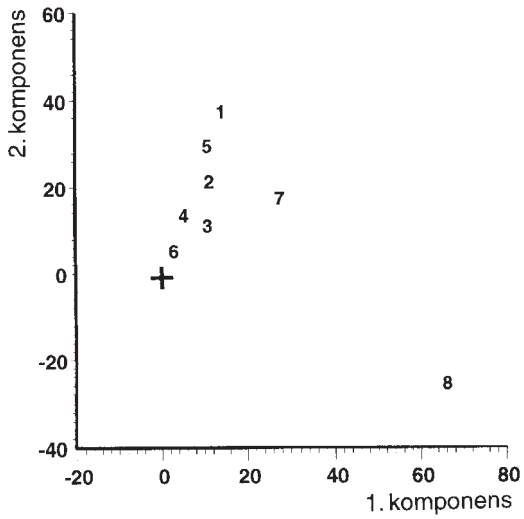
$$\sum_{h=1}^{t} \lambda_h = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}^2 = \mathrm{tr}\{\mathbf{K}\} \qquad (7.13)$$

A $100 \times \lambda_h$ / tr{**K**} mennyiség tehát az eltérésnégyzet-összegnek, vagyis **K** átlós értékei összegének a *h* komponensen lefedett része százalékban.

- Mivel a sajátértékeknek itt négyzetösszeg-értelmezése van, a komponens korrelációkat csak a 3.70 formulával kaphatjuk meg az eredeti változók és a komponens-értékek összevetésével. A biplot ugyan itt is megrajzolható, de értelmezése nehézkesebb, főleg az ortogonalitás hiánya miatt.

Felmerülhet akkor a kérdés, hogy mire is alkalmas egyáltalán a nem-centrált PCA? A lényeg itt a pontok és az origó eltérésnégyzet-összegének maximalizálása minden tengelyen. Ezt a tényt ökológiai mintavételi egységek (pl. kvadrátok) ordinációjában használhatjuk ki a következőképpen: A négyzetösszeg annál nagyobb, minél nagyobb egy-két faj dominanciája a többivel szemben, azaz minél kisebb a kvadrát *diverzitása* (a távolságok négyzetösszege voltaképpen *fordítottan arányos* a Simpson diverzitással). Más szóval, az ordinációban a kis diverzitású kvadrátok az origótól távolra, a nagy diverzitásúak pedig az origó közelébe kerülnek, közvetlen diverzitás-interpretációt biztosítva a kutató számára (Carleton 1980, ter Braak 1983, Digby & Kempton 1987). Ehhez százalékos borításértékekből, vagy még inkább relatív dominancia-adatokból (azaz standardizálás az objektum összege szerint, vö. 2.20 formula) kell kiindulnunk: nyers egyedszámok viszont közvetlenül nem használhatók (mert kvadrátonként nagyon eltérő lehet az össz-egyedszám). A nem-centrált PCA másik előnye akkor jelentkezik, ha az objektumok erőteljes csoportosulást mutatnak, mert ekkor minden egyes komponens egy adott csoporttal ad csak magas koordinátákat, a többivel pedig 0-hoz közeli komponens-értékeket kapunk (Pielou 1984). Az ilymódon nyert eredményeket érdemes távolság-alapon végrehajtott osztályozásokkal is egybevetni.

3 A centrálás kihagyása mellett a változó szórásával még oszthatunk is, így a PCA egy negyedik változatához jutunk. Ennek azonban már nincs különösebb értelme, így ezt a lehetőséget nem tárgyaljuk.



7.8 ábra. Nem-centrált PCA az A1 táblázat adataira. Az origó valódi helyzetét, a félreértések elkerülése végett + jelöli.

Vizsgáljuk meg, hogy mire jutunk az A1 táblázat nem-centrált elemzéséből. Az objektumok ordinációjában (7.8 ábra) az origóhoz a legközelebb a 6. felvétel található: itt a leginkább egyenletes ugyanis a fajok borítása (csupán egy kiugró értékkel: 12% a *Fumana* esetében). Az origótól távolodva a diverzitás csökken (egyre nagyobb a fajok közötti borításbeli eltérés, ami különösen a 8. felvételben szembetűnő).

7.1.6 A patkó-jelenség és ennek szerepe a háttérgrádiensek azonosításában

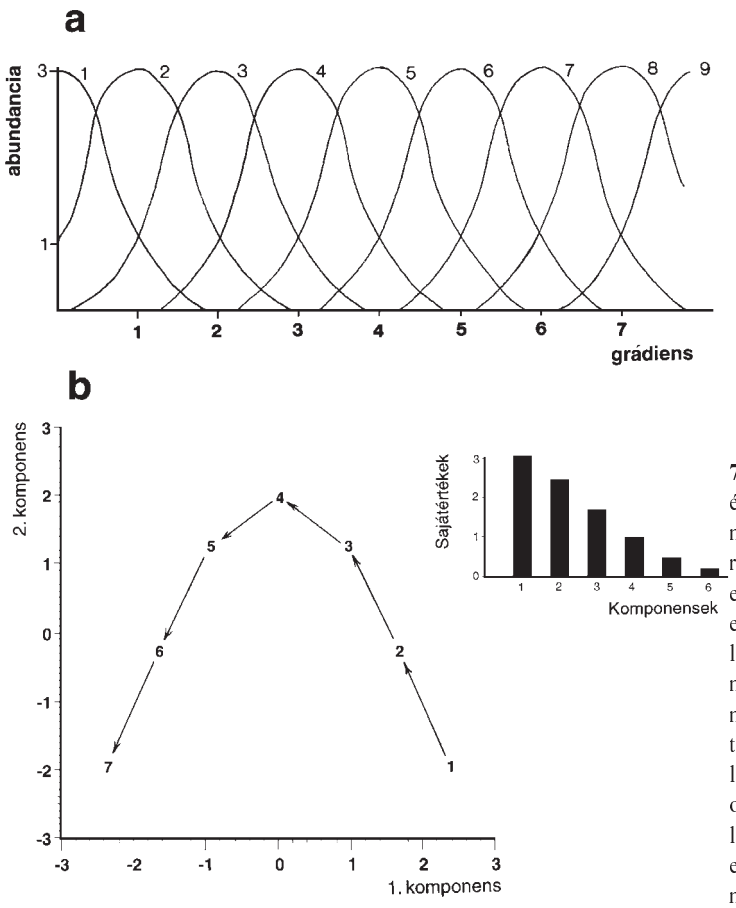
A többváltozós módszerekkel kapcsolatban sokszor felmerül a kérdés, hogy milyen körülmények között, milyen feltételek teljesülése esetén alkalmazhatók, s mikor nem? A főkomponens-elemzésről sokszor leírják, hogy az adatoknak nem árt a többváltozós normális eloszlást követniük. A PCA eredményességét azonban távolról sem csökkenti, ha a változók eloszlása más, pl. egyenletes (más kérdés persze, hogy a komponensek számára vonatkozó statisztikai próba ekkor már nem végezhető el, mert itt igenis feltétel a normális eloszlás). A PCA ordinációk interpretációját inkább egy másik probléma nehezíti meg, mégpedig a változók közötti esetlegesen nem-lineáris kapcsolatrendszer. Mindezt az alábbi adatmátrix segítségével fogjuk szemléltetni (7 objektum = sor, 9 változó = oszlop):

$$\begin{matrix} 1\,3\,1\,0\,0\,0\,0\,0\,0 \\ 0\,1\,3\,1\,0\,0\,0\,0\,0 \\ 0\,0\,1\,3\,1\,0\,0\,0\,0 \\ 0\,0\,0\,1\,3\,1\,0\,0\,0 \\ 0\,0\,0\,0\,1\,3\,1\,0\,0 \\ 0\,0\,0\,0\,0\,1\,3\,1\,0 \\ 0\,0\,0\,0\,0\,0\,1\,3\,1 \end{matrix} \qquad (7.14)$$

Mint látható, az egyes objektumok "fokozatosan" mennek át egymásba, lefelé haladva minden lépésben egy változó eltűnik, egy új megjelenik, kettő pedig – fontosságát tekintve – felcserélődik egymással. Azt mondhatnánk, hogy itt valójában egy mozgató grádiens van a háttérben, amely az objektumokra és a változókra hat. Az ökológiában például valamilyen környezeti tényező (mondjuk a tengerszint feletti magasság) módosul ilyeténképpen, míg erre a fajok különböző optimumgörbékkel leírható módon reagálnak. A 7.9a ábra egy ilyen hipotetikus – és idealizált – esetet mutat be; a fenti adatmátrix ennek egy további leegyszerű-

sítéseként fogható fel. Ha valaki azt várja, hogy a grádiens mentén felvett mintavételi egységek PCA ordinációja szépen sorba rendezi a pontokat, rekonsturálva a háttér-grádienst, az súlyosan csalatkozik. A fenti táblázat standardizált komponens-elemzése ugyanis a hét pontot egy patkó (vagy ív) mentén helyezi el az első két komponens alkotta térben (7.9b ábra). A sajátértékek egymáshoz viszonyított fontossága pedig éppen nem hirtelen (ami *egy* kitüntetett háttérváltozó fontosságára utalna), hanem csak fokozatosan csökken. Azt gondolhatnánk tehát, hogy maga a módszer – műtermék jelleggel – torzítja el az egyébként számunkra teljesen nyilvánvaló, és ezért el is várt elrendeződést valamely egyenes mentén.

A "jelenség" okát értelmezve kiderül, hogy nem "műtermékről", sem "torzításról", hanem nagyon is értelmes módon magyarázható dologról van szó, amely régen ismeretes az ordinációs módszerek irodalmában ("horseshoe effect", Kendall 1971, vagy "arch effect"). A kilenc faj kapcsolata az egész grádienst alapul véve távolról sem lineáris. Például az 1. faj csökkenése mellett a 2. faj abundanciája először nő, majd csökken, végül mindegyikük 0 értéket vesz fel, de más fajpárokra is hasonló helyzet adódik. Vannak olyan fajpárok is, amelyek szinte össze sem hasonlíthatók, hiszen amikor az egyik értékei változnak a másik konstans módon 0 marad, és fordítva. Ennek következtében a grádiens mentén az 1. objektumtól a 4.-ig haladva már teljesen kicserélődnek a fajok, az 1. és 4. objektum közötti távolság eléri a maximumot. Az 1. és az 5. ill. az ezt követő objektumok között a távolság már nem nőhet tovább, hiszen nincs lehetőség további kicserélődésre. Mindezt az ordináció voltaképpen



7.9 ábra. A patkó-jelenség és magyarázata. **a:** A 7.14 mátrix kilenc változójának reakciója egy háttér-grádiensre nem lineáris, ezáltal egymással sem korrelálnak lineárisan. **b:** Ennek eredményeképpen a grádiens mellett észlelt hét objektum egy ív mentén helyezkedik el a PCA ordinációban. A kis oszlopdiagram a sajátértékek egyenletes csökkenését mutatja.

hitelesen ábrázolja, mert "megpróbálja" az 1. objektum távolságát a 4., az 5., a 6. és 7. objektumokkal azonosra venni. A patkó-jelenséget valójában némi gyakorlattal magunk is könnyen felismerhetjük, s ennek előfordulásából minden esetben az adatok nem-lineáris összefüggéseire kell következtetnünk.

A patkó-jelenség – ellentétben egyes pesszimista vélekedésekkel – nem zavarja a PCA ordináció értelmezhetőségét. Sokan vannak azonban, akik mindenképpen egy "egyenes mentén" akarják elrendezni az objektumokat, így mindenféle "patkó-egyengető korrekciókat" javasolnak. Phillips (1978) írja le pl. a *polinomiális* ordinációt, amely egy parabolát illeszt a pontokra az 1-2. dimenzióban, s ebből kap új koordinátákat. (Az 1. és 3. tengely között egyébként rendszerint egy harmadfokú, az 1. és 4. komponens között pedig negyedfokú polinom fejezi ki az összefüggést, innen a *polinomiális* elnevezés). A regresszió és a "kinyújtás" eredményeképpen az objektumok már jobban illeszkednek egy egyenesre. A gond az ilyen automatikus kiigazításokkal azonban az, hogy *eleve feltételezik* az egydimenziós háttérgrádiens jelenlétét, amelyre a változók optimum-görbe szerint reagálnak. Nem tudjuk tehát, hogy mi lett volna az eredmény egyébként. A logika inkább azt diktálja, hogy egy teljesen szabványos analízist végezzünk el először, s legfeljebb ezt követően nyúljunk a korrekciós módszerekhez, s a kapott diagramokat hasonlítsuk is össze.

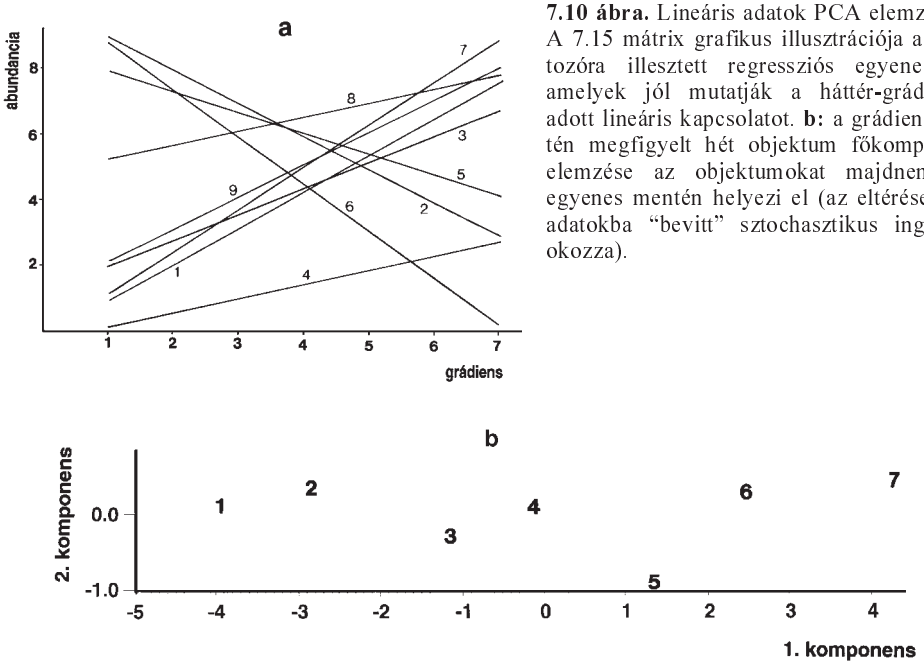
A patkó jelenség ökológiai ordinációkban elsősorban akkor fordul elő, ha nagy a háttérgrádiens mentén bekövetkezett változások "sebessége" ("*species turnover*", β-*diverzitás,* Whittaker 1967). Ilyen esetekben szóba jöhet egy – a patkó-egyengetésnél szellemesebb – megoldás, a legrövidebb út kiigazításának módszere, amelyre a többdimenziós skálázás ismertetésekor visszatérünk (7.4.1 rész); ez a PCA esetében ugyanis nem alkalmazható.

Az ökológiai grádiens egy rövid szakaszán felvett objektumsorban a változók kapcsolata már megközelítően lineáris lehet, s ekkor az objektumok az ordinációs diagramon többé-kevésbé egy egyenes mentén helyezkednek majd el. Ezt mutatjuk be az alábbi, ugyancsak 7×9-es adatmátrix felhasználásával:

$$\begin{array}{ccccccccc} 1 & 9 & 2 & 0 & 8 & 9 & 1 & 5 & 2 \\ 2 & 7 & 3 & 0 & 8 & 7 & 2 & 5 & 3 \\ 3 & 6 & 3 & 1 & 7 & 5 & 4 & 6 & 4 \\ 4 & 6 & 4 & 1 & 6 & 3 & 5 & 6 & 5 \\ 5 & 5 & 4 & 2 & 6 & 1 & 7 & 7 & 6 \\ 6 & 4 & 6 & 2 & 5 & 1 & 8 & 7 & 7 \\ 8 & 3 & 7 & 3 & 4 & 0 & 9 & 8 & 8 \end{array} \qquad (7.15)$$

A kilenc változó (az oszlopokban!) a háttér-grádiens mentén lineárisan változik, mindegyikükre egyenes is illeszthető (7.10a ábra). Ennek következtében a változók között erős – pozitív és negatív – lineáris korreláció mérhető, bár ezt némi "zajjal" szándékosan csökkentettük. A standardizált PCA ebben az esetben "ideálisan" viselkedik. A háttér-grádiens hatása az első komponensben jelentkezik, amely az összvariancia 96 %-át magyarázza meg. A "zaj" miatt persze az objektumok nem illeszkednek tökéletesen a regressziós egyenesre, a linearitástól mutatkozó eltérés azonban az összvariancia rendkívül kis része – pontosan 4 %-a – csupán (7.10b ábra).

A PCA eredmények értékelésében (különösen a patkótól "mentes" esetben) gyakori, hogy a komponenseket *utólag* valamilyen, az elemzésben nem szerepeltetett változóval, pl. környezeti tényezővel igyekszünk azonosítani. Kiszámítható például a tengelyek és a külső tényezők lineáris korrelációja, ami a komponensek értelmezésében segíthet. Ökológiai grádiensek ilymódon történő kimutatását ter Braak & Prentice (1988) *indirekt grádiens elemzés*nek nevezi, hiszen a grádiens feltárása nem közvetlenül a háttérváltozók alapján, hanem a fajok



7.10 ábra. Lineáris adatok PCA elemzése. **a:** A 7.15 mátrix grafikus illusztrációja a 9 változóra illesztett regressziós egyenesekkel, amelyek jól mutatják a háttér-grádienssel adott lineáris kapcsolatot. **b:** a grádiens mentén megfigyelt hét objektum főkomponens-elemzése az objektumokat majdnem egy egyenes mentén helyezi el (az eltéréseket az adatokba "bevitt" sztochasztikus ingadozás okozza).

felhasználásával megy végbe. Ez szembeállítható a *direkt* grádiens elemzéssel, amelyben az objektumok ordinációja a környezeti változókat is figyelembe veszi (lásd a 7.2.5 és 7.3.5 részeket).

7.1.7 Faktoranalízis

Röviden foglalkoznunk kell a főkomponens-elemzéssel technikai rokonságban álló, attól azonban az alapelveket tekintve mégis lényegesen különböző módszerrel, a *faktoranalízis*sel (rövidítve: FA) is. A rövid ismertetés annál is inkább lényeges, mert gyakran összekeverik a kettőt – például olymódon, hogy a PCA eredmények értékelésekor komponensek helyett faktorokról beszélnek. Így azután sokszor nem is igazán világos, hogy voltaképpen milyen elemzést is hajtottak végre az illető vizsgálatban.

A döntő elvi különbség a PCA és a FA között az, hogy míg a komponensek az összvariancia lehető legnagyobb hányadának a megmagyarázását szolgálják, a *faktor*ok az egyes változók közötti kovarianciát (ill. a melegen ajánlható standardizált esetben: a *korrelációt*) fedik le maximálisan. A faktorok olyan hipotetikus háttértényezők, amelyek a változók közötti kapcsolatokat értelmezik – ezáltal az összvarianciának egy kisebb, bár rendszerint jelentékeny részéért felelősek csupán. A modell szerint ugyanis van egy olyan variancia-hányad is, amely kizárólag az egyes változókra jellemző, következésképpen ezt a közös faktorok nem magyarázhatják. Ez a változók saját, *specifikus* varianciája, amelyet az ún. specifikus v. egyedi faktorok okoznak (éppen *n* ilyen faktor van). A faktoranalízis tehát olyan ordinációs módszer, amely elsősorban a változók kapcsolatrendszerének a feltárását célozza, míg az objektumok elrendezése másodlagos fontosságú, ill. el is marad. Így a FA szerepe a biológiai adatfeltárás-

ban eleve kisebb[4]. Anélkül, hogy a faktoranalízis módszereit részletesebben bemutatnánk, megemlítjük még a következőket:

- A faktoranalízis során általunk előre megadott p számú faktorba igyekszünk belesűríteni a közös varianciát. Miután p megadása önkényes, azaz a modell szabadon változtatható, sok szerző a FA-t inkább "művészi" tevékenységnek, mintsem objektív statisztikai módszernek tekinti, s nem igazán ajánlja az adatfeltárásban (pl. Kendall 1975, Chatfield & Collins 1980). Mások, pl. Jolliffe (1986) szerint nincs értelme a kérdésnek, hogy melyik módszer a jobb, mert a FA is értékes információval szolgálhat az adatelemző számára.
- A faktorok, akárcsak a komponensek, páronként korrelálatlanok. Az i faktor és a j változó korrelációját, a_{ij}-t, a változó *faktorsúly*ának nevezzük. A faktorsúlyok alapján történik a változók ordinációs diagramjának a felrajzolása (hasonlóan a PCA komponens korrelációinak illusztrálásához). Az ortogonalitás azonban nem kötelező: a faktorok úgy is meghatározhatók, hogy korrelációjuk nem 0 (Cattell 1978 említ példákat, amikor ez lényeges).
- A változók korrelációs mátrixának átlójában lévő egyesek, az "önkorrelációk", a modell szerint részben a közös faktorok hatásának részben pedig a specifikus (egyedi) faktorok hatásának az összegzései. Más szóval,

$$r_{jj} = 1 = \sum_{i=1}^{p} a_{ij}^2 + e_j \qquad (7.16)$$

amelyben az összeget a j változó kommunalitásának, az e_j mennyiséget pedig egyedi varianciának nevezzük. Minél nagyobb p, annál nagyobb a kommunalitás részesedése az önkorrelációkból, azaz határesetben, amikor p megegyezik a korrelációs mátrix rangjával, a specificitás 0 lesz, s voltaképpen a FA megegyezik a PCA-val. Egyébként a PCA és a FA eredménye gyakran rendkívül hasonló.

- A faktoranalízis legismertebb eljárása az ún. *főfaktor*-módszer, amely valójában a PCA iteratív alkalmazása a kommunalitások becslésére. A módszerrel megismerkedni kívánóknak Jahn & Vahle (1974) könyvét ajánlhatjuk elsősorban.

7.2 Két változócsoport értékelése kanonikus korreláció-elemzéssel

A főkomponens-elemzés során az összes változót "egy kalap" alá vesszük, mondván, hogy ezek logikailag azonos jellegűek: növényfajok borítás értékei vagy morfológiai karakterek, stb. Előadódhat azonban olyan helyzet is, hogy a vizsgálatban szereplő változókat nincs értelme együtt kezelni, mert azok valójában két csoportra oszthatók. Ökológiai vizsgálatokban például a mintavételi egységeket gyakran jellemezzük mind a bennük talált fajok alapján mind pedig környezeti változók figyelembevételével. Bár – elvileg – ekkor is alkalmazható a PCA, de ezzel nem kapunk információt egy fontos dologról, mégpedig a csoportok mint egységek

4 A faktoranalízis elsősorban a társadalom-tudományokban, ill. a pszichológiában népszerű. Korábban egyike volt a legelőször alkalmazott többváltozós módszereknek a biológiában is.

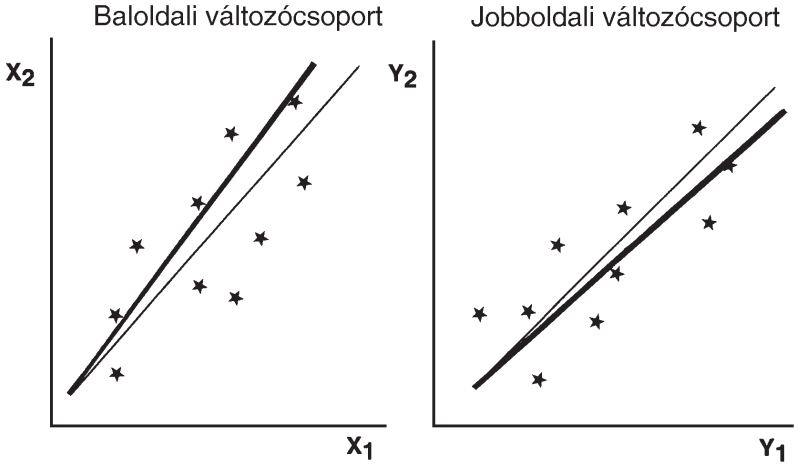
közötti összefüggésekről. A fent említett esetben például nagyon is érdekelheti az ökológust az, hogy milyen a kapcsolat a környezeti változók csoportja és a biológiai változók csoportja között (azaz: megjósolható-e a fajösszetétel a környezeti változók ismeretében és viszont). Erre a vizsgálódásra ad lehetőséget a PCA-val szoros kapcsolatban álló *kanonikus korreláció elemzés* (CCA; más könyvekben COR, pl. Jongman et al. 1987).

A módszer Hotelling (1936) munkásságával vált ismertté. Alapgondolata röviden az, hogy mindkét változócsoportban külön-külön megkeressük a változók lineáris kombinációit olymódon, hogy ezek maximálisan korreláljanak egymással. Másképpen fogalmazva: a CCA felfogható egy dupla főkomponens elemzésnek, amelyet az első, ill. a második változócsoportra kapott tengelyek lehető legjobb illesztése (forgatása) követ. A CCA-ban kapott két új tengelyt *kanonikus változók*nak, a változócsoportok kapcsolatát, azaz a két kanonikus változó közötti korrelációt pedig *kanonikus korrelációnak* nevezzük. Az első tengelypár meghatározása után az elemzés természetesen tovább folytatható az előzőekre merőleges újabb tengelyek keresésével, és ezek korrelációinak meghatározásával. Míg tehát a főkomponens elemzés célja a teljes variancia megmagyarázása egy változócsoportra, a CCA már a két változócsoport közötti kovarianciára összpontosít (Cooley & Lohnes 1971, Gittins 1979).

A CCA grafikus illusztrációja talán jobban megvilágítja a kérdést (7.11 ábra). Tételezzük fel, hogy mindkét csoportban két változónk van, amelyeket x_1 és x_2, valamint y_1 és y_2 jelöl. Ha külön-külön hajtunk végre főkomponens-elemzést, akkor a vékony vonallal jelzett főkomponensek adódnak (a második komponensekről most megfeledkezhetünk). Az objektumok koordinátái a baloldali ábrán azonban nem korrelálnak maximálisan a megfelelő koordinátákkal a jobboldali rajzon. Mindkét tengelyt el kell forgatnunk (vastag vonalak a 7.11 ábrán), hogy a koordináták közötti korrelációt maximalizálhassuk.

A CCA számításmenete a változók korrelációs mátrixából indul ki, amelyet célszerűen a két csoport szerint osztunk fel négy részmátrixra az alábbi módon:

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix}. \qquad (7.17)$$



7.11 ábra. A kanonikus korreláció elemzés szemléltetése. A két változócsoportra kapott komponensek (vékony vonalak) szinte sosem esnek egybe a kanonikus változókkal (vastag vonalak).

$\mathbf{R}_{11}$ tartalmazza tehát az első csoportban lévő n_1 számú változó (baloldali változók) korrelációit, $\mathbf{R}_{22}$ a második csoport n_2 számú változójának (jobboldali változók) korrelációit, míg $\mathbf{R}_{21}$ ill. ennek transzponáltja $\mathbf{R}_{12}$ a két csoport kereszt-korrelációit összesíti. Az alapproblémát, miszerint a két csoport közötti korrelációt kell maximalizálnunk a belső korrelációkhoz képest, vagyis a kettő "hányadosát", a mátrix algebra nyelvén a következőképpen írhatjuk fel: $\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$. A feladat ezek után e mátrix sajátérték-elemzése, amely a PCA-nál már ismertetett módon indul el:

$$(\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12} - \lambda_j\mathbf{I})\,\mathbf{v}_j = \mathbf{0}. \tag{7.18}$$

azzal a különbséggel, hogy az $\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$ mátrix nem szimmetrikus. A $\mathbf{v}_j$ sajátvektorokat itt is úgy számítjuk ki, hogy egységnyi hosszúak legyenek. Ezt követően normálással meghatározzuk a második változócsoport kanonikus súlyait (a j kanonikus változó értékeit):

$$\mathbf{c}_j = \mathbf{v}_j / \sqrt{(\mathbf{v}_j'\mathbf{R}_{22}\mathbf{v}_j)} \tag{7.19}$$

így a változó varianciája 1 lesz. Az első változócsoportra pedig a következő módon kapjuk meg a súlyokat, ugyancsak a j kanonikus változóra:

$$\mathbf{b}_j = (\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{c}_j) / \sqrt{\lambda_j} \tag{7.20}$$

A normált sajátvektorokat megszorozva az eredeti adatokkal kapjuk meg az objektumok koordinátáit az első változócsoportra:

$$\mathbf{T} = \mathbf{X}'\mathbf{B}, \tag{7.21}$$

illetve a második változócsoportra:

$$\mathbf{U} = \mathbf{Y}'\mathbf{C}, \tag{7.22}$$

ahol $\mathbf{X}'$ az adatmátrix első része (mérete: $m \times n_1$), az $\mathbf{Y}'$ ($m \times n_2$ -es mérettel) pedig az adatmátrix második fele (mint látható, itt a sorokban vannak az objektumok, a változók pedig az oszlopokban). Az összes változót előzőleg centráljuk és egységnyi szórásúra standardizáljuk. A $\mathbf{B}$ és $\mathbf{C}$ mátrixok a 7.19 és 7.20 formulákkal meghatározott kanonikus változókat tartalmazzák, méretük $n_1 \times q$, illetve $n_2 \times q$ (q értékét lásd lentebb).

7.2.1 A kanonikus korrelációk és szignifikancia-próbájuk

A pozitív sajátértékek száma a CCA elemzésben $q = \min\{n_1, n_2\}$ feltéve, hogy $m > n_1, n_2$. A sajátértékek négyzetgyökei mérik a jobb és baloldali változócsoportok kanonikus korrelációját:

$$|R_j| = \sqrt{\lambda_j} \tag{7.23}$$

amelyből éppen q darab van, ezek célszerűen csökkenő sorrendbe rendezendők. Az abszolútérték jel arra utal, hogy a kanonikus korrelációk is éppúgy –1 és 1 közé esnek, mint két változó hagyományos korrelációja, de az előjelet nem ismerjük meg az elemzésből. Az tehát megegyezés kérdése, hogy a $\mathbf{b}_j$ és $\mathbf{c}_j$ közötti lineáris korrelációt az abszolút értékkel mérjük.

Amennyiben a megfigyelések függetlenek egymástól, és a többváltozós normalitás feltétele is teljesül, akkor statisztikai próbával ellenőrizhető, hogy a kanonikus korreláció különbözik-e 0-tól. A Bartlett féle Λ (lásd Cooley & Lohnes 1971) a kanonikus korrelációk együttes szignifikanciáját teszteli, amellett, hogy az első 0, 1, 2, ..., k, ..., q kanonikus változót fokozatosan elhagyjuk:

$$\Lambda = \prod_{j=k+1}^{q}(1-\lambda_j) \qquad (7.24)$$

Bartlett szerint Λ a χ^2 eloszlást követi a következő átalakítás után:

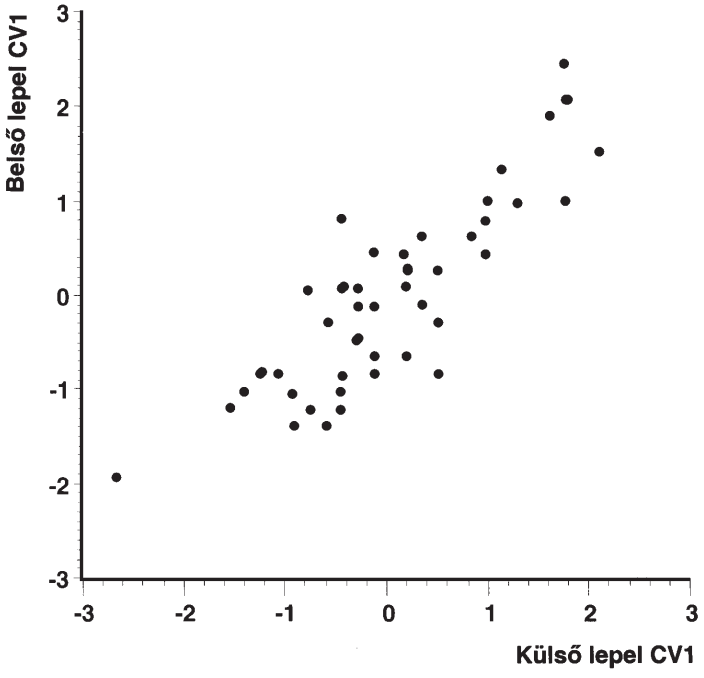
$$X^2 = -[m - 1 - 0{,}5\,(n_1 + n_2 + 1)\ln\Lambda \qquad (7.25)$$

$(n_1 - k)\,(n_2 - k)$ szabadsági fok mellett. Tehát, ha k=0-ra a próba szignifikáns eredményt ad, de k=1-re már nem, akkor csak az első kanonikus korrelációnak van "értelme".

A szignifikancia-teszttel kapcsolatos problémákat Gittins (1979, 1985) részletesen taglalja. Figyelmeztet, hogy az alapfeltételek teljesülése mellett is elővigyázatosnak kell lennünk. A sajátértékek nagyságának közvetlen tesztje segíthet, de ehhez az ún. *legnagyobb sajátérték* eloszlást kell ismernünk.

A fenti számításokat követően az objektumok ordinációja többféleképpen ábrázolható. Általános gyakorlat az objektumokat egy olyan kétdimenziós diagramon feltüntetni, amelyben a vízszintes tengely a baloldali, a függőleges tengely pedig a jobboldali csoportból származó első – így a legfontosabb – kanonikus változó. Ha a két változócsoport közötti összefüggés erős (amelyet a 7.23 formula fejez ki), akkor az objektumok közel esnek egy "átlós" egyeneshez. Gyengébb kapcsolatra, vagy annak teljes hiányára utal az objektumok szórt elrendeződése. Érdemes az objektumok ordinációját külön-külön is felrajzolni mindkét változócsoportra és az első két kanonikus változóra, különösen akkor, ha a két első kanonikus korreláció egyaránt magas értékű, ill. szignifikáns. Erre a diagramra változók és a kanonikus tengelyek közötti korrelációk is rávetíthetők.

Hajtsuk végre a harmadik *Iris* faj (*I. virginica*) CCA értékelését az A2 táblázat figyelembevételével! A szemléltetés kedvéért választott példa relatíve egyszerű: a baloldali változócsoport a külső lepelre, a jobboldali pedig a belső lepelre vonatkozik, s a feladat nyilván e két változócsoport közötti összefüggésrendszer feltárása (bár logikailag természetesen ezek egy



7.12 ábra. Az *Iris virginica* 50 egyedének kanonikus korreláció elemzése. A bemutatott ordináció az egyedeket a külső lepelkörre (vízszintes tengely) ill. a belső lepelkörre (függőleges tengely) kapott 1. kanonikus változó alapján rendezi el.

változócsoportba is besorolhatók). Az ordinációs diagramok közül először azt vizsgáljuk meg, amelyben a vízszintes tengely a külső, a függőleges tengely a belső lepelre kapott 1. kanonikus változó (7.12 ábra). Az egyedek egy "átló" mentén helyezkednek el, kevéssé szóródnak, mutatva a két változócsoport viszonylag szoros kapcsolatát (R_1=0,86, χ^2=75,9, p<<0,001). Megjegyzendő, hogy a második kanonikus változók is jelentősek (R_2=0,47, χ^2=12,0, p<0,001). Az *Iris versicolor* esetében is hasonló eredményt kapunk, bár R_1 "csak" 0,76. Az *Iris setosa* CCA elemzése azonban már igencsak eltérő eredménnyel járt: mindkét kanonikus korreláció alacsony (pl. R_1=0,32), és egyik kanonikus változó sem tekinthető szignifikánsnak! E fajnál tehát a külső lepel nem prediktív a belsőre vonatkozóan és viszont. Ez igazolja, hogy a fajokra közösen alkalmazott CCA elmosta volna a fajok közötti eltéréseket. A három faj együttes analízise *egy* erős kanonikus korrelációt ad egyébként (vö. Podani 1994), ami azt jelenti, hogy *génusz* szinten már igenis erős a kapcsolatrendszer a két változócsoport között).

7.2.2 Korreláció az eredeti változókkal

A kanonikus változók és az eredeti változók között kétféle korreláció kiszámítására nyílik lehetőség a CCA során.

Csoporton belüli korrelációk. Az egyes változók hozzájárulása a saját változócsoportból kapott kanonikus korrelációhoz rendkívül hasznos lehet az eredmények interpretációja szempontjából. A baloldali *i* változó (amelyet az $\mathbf{x}_i$ vektor képvisel) korrelációja az erre a csoportra kapott *j* kanonikus változóval a következő:

$$\rho(\mathbf{x}_i, \mathbf{b}_j) = \sum_{k=1}^{n_1} r_{ik} b_{kj} \qquad (7.26)$$

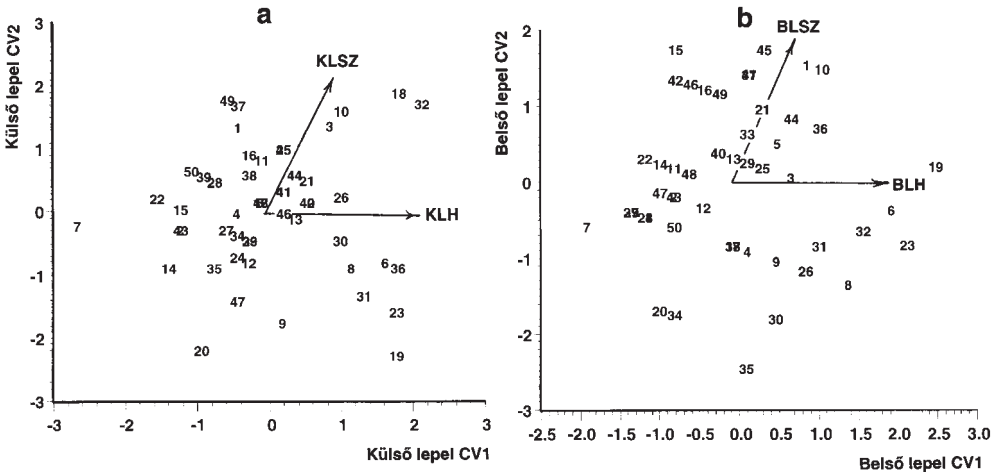
("*structure correlation*"). Hasonlóképpen kapható meg a jobboldali változócsoportban az *i* változó ($\mathbf{y}_i$ vektor) és a megfelelő *j*-edik kanonikus változó korrelációja:

$$\rho(\mathbf{y}_j, \mathbf{c}_j) = \sum_{k=1}^{n_2} r_{ik} c_{kj} \qquad (7.27)$$

Az r_{ik} értékek mindkét egyenletben az eredeti változók közötti lineáris korrelációkat jelölik. A 7.26-27 egyenletek segítségével azonosíthatjuk azokat a változókat, amelyek legmarkánsabb módon "képviselik" saját csoportjukat a másikkal való összehasonlításban.

Az *Iris virginica* elemzéséből kiderül, hogy az 1. kanonikus változóval mindkét csoportban a lepel hossza korrelál legerősebben, sőt gyakorlatilag azonosak, hiszen a korrelációk közel vannak 1-hez. A 2. kanonikus változót viszont a szélesség adatok értelmezik (0,88 ill. 0,94-es korrelációval). Mindezt két biplot-szerű diagramon ábrázolhatjuk is, külön-külön a két változócsoport 1-2. kanonikus változóira (A CCA biplot elkészítésének lehetőségeit ter Braak (1990) vizsgálta meg részletesen). A 7.19-20 egyenletek révén nyert értékek lesznek az objektumok koordinátái, a változók koordinátáit pedig a 7.26-27 egyenletekből kapjuk, önkényes átskálázással rávetítve az objektumok ordinációjára (7.13a-b ábra). A két ordináció között szemmel látható egyezések vannak, de ezt egyelőre nem tudjuk pontosabban megítélni a pontok nagy száma és az átfedések miatt.

Csoportok közötti korrelációk. Ezek révén az egyik csoport eredeti változói és a másik csoport kanonikus változói közötti kapcsolat értékelhető (Gittins 1979). A baloldali csoport *i* változójának és a jobboldali csoport *j* kanonikus változójának korrelációja a következő:



7.13 ábra. Az *Iris virginica* 50 egyedének CCA vizsgálata **a:** a külső lepelre vonatkozó változócsoport alapján, **b:** a belső lepelre vonatkozó változócsoport alapján. A nyilak a tulajdonságok és a kanonikus változók közötti korrelációkat érzékeltetik (korrelációs Rohlf-biplot).

$$\rho(\mathbf{x}_i, \mathbf{c}_j) = \sum_{k=1}^{n_2} r_{ik} c_{kj} \tag{7.28}$$

Hasonlóképpen számítható ki a jobboldali csoport *i* változójának és a baloldali csoport *j* kanonikus változójának a korrelációja:

$$\rho(\mathbf{y}_i, \mathbf{b}_j) = \sum_{k=1}^{n_1} r_{ik} b_{kj} \tag{7.29}$$

amelyekben az r_{ik} értékek az $\mathbf{R}_{12}$ mátrixba írt korrelációk. Egy csoport-közötti korreláció négyzete az eredeti változó varianciájának azon hányada, amelyet a másik csoport kanonikus változója is megmagyaráz. A csoportok közötti korrelációk nem alkalmasak grafikus ábrázolásra.

Az *Iris virginica* elemzése azt mutatja, hogy a lepel-hossz megjósolhatósága a legnagyobb a másik változócsoportból. Pl. a külső lepel hosszának varianciáját a belső lepel 1. kanonikus változója 75 %-ban értelmezi. A szélességeknél ez már legfeljebb a 18 %-ot éri el.

7.2.3 Variancia és redundancia

Kiszámíthatjuk azt is, hogy adott változócsoport teljes varianciájának hányad részét értelmezi a saját kanonikus változója. Ez valójában a négyzetre emelt csoporton *belüli* korrelációk átlaga, vagyis a baloldali változókra a *j*-edik kanonikus változó által magyarázott variancia a következő:

$$100 \times \sum_{i=1}^{n_1} \rho(\mathbf{x}_i, \mathbf{b}_j)^2 / n_1 \tag{7.30}$$

Hasonlóképpen adódik a százalékérték a jobboldali változókra:

$$100 \times \sum_{i=1}^{n_2} \rho(\mathbf{y}_i, \mathbf{c}_j)^2 / n_2 \qquad (7.31)$$

Az *Iris virginica* elemzésében a baloldali csoportban (külső lepel méretei) a variancia 61 ill. 39 %, a jobboldaliban (belső lepel méretei) pedig 55 ill. 45 %. A varianciák segítségével jól érzékeltethető, hogy a kanonikus változók egészen mások, mint a komponensek. A külső lepel két méretére alkalmazott PCA például az első főkomponensben 74 %-ot összesít, s így csak 26% marad a másodikra. Ez pedig lényegesen jobb varianciasűrítés, mint a 61 és 39 %. De itt nem is ez volt a célunk.

A *redundancia* adott változócsoport teljes varianciájának az a része, amelyet a másik változócsoport egy kanonikus változója meg tud magyarázni. Ez tehát a csoportok-*közötti* analógiája a fenti varianciáknak (Gittins 1979), és a négyzetre emelt csoport-közötti korrelációk átlagának felel meg. A baloldali csoportot a jobboldali csoport *j*-edik kanonikus változója a következőképpen értelmezi:

$$100 \times \sum_{i=1}^{n_1} \rho(\mathbf{x}_i, \mathbf{c}_j)^2 / n_1 \qquad (7.32)$$

míg a jobboldali változócsoport redundanciája a baloldali csoport *j*-edik kanonikus változója alapján a következő:

$$100 \times \sum_{i=1}^{n_2} \rho(\mathbf{y}_i, \mathbf{b}_j)^2 / n_2 \qquad (7.33)$$

A redundancia-értékek összege *j* szerint az adott változócsoport teljes redudanciája a másik változócsoport kanonikus változóinak értelmében. Ez tehát a varianciának azon része, amelyet a másik változócsoporttal értelmezhetünk. A teljes redundancia egy aszimmetrikus mérték, vagyis a jobboldali csoport redundanciája általában nem ugyanaz, mint a baloldalié.

Ez így van az *Iris* példában is. A külső lepel két változóját a belső lepel kanonikus változói 55 %-ban értelmezik, míg a fordított irányban csak 51,5 %-os a megmagyarázás mértéke.

7.2.4 Megjegyzések a CCA használhatóságával kapcsolatban

Számos szerző egyetért abban, hogy a CCA eredmények interpretációja rendszerint problematikusabb, mint más többváltozós módszer esetében. Bock (1975) rámutat, hogy a kanonikus változók bizonyos "kompromisszum" eredményei: az ortogonalitás feltétele mellett a csoportok közötti kovarianciát igyekszünk velük maximalizálni úgy, hogy közben a csoporton belüli variancia minimális legyen (vö. a varianciákkal kapcsolatos fenti összehasonlítással a 7.31 egyenlet alatt). Ennek következtében a kanonikus változók és a főkomponensek csak kivételes esetben esnek egybe. Amint Rohlf (In: Legendre & Legendre 1983, p. 330) rámutat, ez egészen odáig mehet, hogy valamelyik változócsoportra kapott kanonikus változó annak teljes varianciájából csak igen kis részt magyaráz meg, így nincs is interpretatív értéke. A kanonikus korreláció ekkor egy jelentés nélküli kapcsolatot maximalizál (vö. Pimentel 1979). Ezt a problémát voltaképpen elkerülhetjük, ha valamelyik változócsoportra PCA-t alkalmazunk, és a CCA elemzésbe az eredeti adatok helyett a főkomponens értékeket vonjuk be (pl. Digby & Kempton 1987, p. 82, vagy Ludwig & Reynolds 1988, p. 299) vagy pedig egyszerűen mindkét változócsoportot PCA segítségével vizsgáljuk csak meg (Williams & Lance 1968, Shaukat &

Uddin 1989). Ennek az lehet a további előnye, hogy az $\mathbf{R}_{11}$ vagy $\mathbf{R}_{22}$ esetleges szingularitási problémái is megoldódnak (ui. ha például az objektumok száma kisebb, mint a változóké, akkor ezek a mátrixok nem invertálhatók). Azt is megtehetjük, hogy a standard CCA elemzést mindig végrehajtjuk az eredeti változók és a komponens értékek alapján is, és utána összehasonlítjuk az eredményeket. Ezt azonban itt hely hiányában már nem szemléltethetjük.

7.2.5 Redundancia-analízis

Az *Iris* CCA értékelésében a két változócsoport kapcsolatát szimmetrikusnak fogtuk fel, hiszen összehasonlításukban egyik irány sem tüntethető ki. Ökológiai vizsgálatokban, ha például az egyik csoportban fajok borításai, a másikban pedig környezeti változók szerepelnek, ez a szimmetria nem igazán érvényesül: a fajok reagálnak a környezeti változókra, de ez nem áll fenn a fordított irányban! A megjósolhatóság problémája nem szimmetrikus, ellentétben azzal, amire a CCA bevezetésében céloztunk. Voltaképpen itt nem logikus a kanonikus változókat úgy meghatározni, hogy mindkét változócsoportban lineáris kombinációkat keresünk és ezek korrelációját maximalizáljuk (ahogy a CCA teszi). Ez sokszor nehezen értelmezhető, amint a 7.2.4 részben már említettük. Fajok és környezeti változók esetén inkább azt kellene elérni, hogy az ordinációs tengelyek csak a környezeti változókra utaljanak, hiszen ezek hatnak a fajok borítás-értékeire (direkt grádiens analízis). Legyenek tehát a tengelyek a környezeti változók olyan lineáris kombinációi, amelyek ugyanakkor a lehető legnagyobb varianciát magyarázzák meg a mintavételi helyeknek (objektumoknak) fajok szerinti ordinációjában. Más szóval: az objektumokat a PCA-hoz hasonló módon ordináljuk, azzal a megszorítással, hogy a komponensek a lehető legjobban értelmezzék a környezeti változókat is. Erre alkalmas a Rao (1964) által kifejlesztett *redundancia-analízis* (RDA) módszere, amely elsősorban ter Braak & Prentice (1988) munkássága révén vált ismertté az ökológia irodalmában. Mivel a tengelyek nem "szabadon" illeszkednek az objektumokra, hanem a környezeti változócsoport által megszabottan, ter Braak & Prentice a *kötött* ("*constrained*") ordináció elnevezést javasolta. Mindeddig még ritkán alkalmazták, elsősorban azért, mert az ugyancsak kötött stratégiájú CCOA (7.3.5 rész) vált inkább népszerűvé (vö. Birks et al. 1996).

> A RDA módszere lényegében véve egy olyan kanonikus korreláció elemzés, amelyben nem törődünk azzal, hogy mi történik a fajok között. Korrelációjuk teljesen közömbös számunkra, csak a környezeti változók csoportján belüli korrelációk és a fajok-környezeti változók korrelációi kellenek. Ha a faj-adatok az $\mathbf{X}$ részmátrixban vannak, akkor a 7.18 egyenletben $\mathbf{R}_{11}$ helyére az $\mathbf{I}$ egységmátrixot írjuk (ennek átlójában 1-esek vannak, a többi érték 0).

7.3 Korreszpondencia elemzés

A PCA vagy a CCA biplot alkalmazásával az objektumok és a változók közötti kapcsolatokat szemléltethetjük az ordinációban. Mindkét módszer esetében azonban voltaképpen különkülön állítjuk elő a változók és az objektumok ordinációit, s azokat csak ezután, különféle "trükkök" alkalmazásával vetítjük egymásra. Felmerülhet a kérdés, vajon lehetséges-e olyan biplot-ot előállítani, amelyben a változók és az objektumok optimális egymásra illesztése egyidejűleg és közvetlen módon alakítható ki? A válasz: van erre alkalmas módszer. A legjobb illeszkedés, vagy kölcsönös megfeleltetés (más szóval: korreszpondencia) feltárása a *korreszpondencia-elemzés* célja. Erre a legkülönbözőbb tudományágakban, s azokon belül is

eltérő területeken már régóta felmerült az igény, így nem csodálkozhatunk azon, hogy a szakirodalom a rokon jellegű módszereket más és más néven illeti. Így például megmutatható, hogy a reciprok átlagolás, a "*dual scaling*", kontingencia-tábla elemzés és más eljárások (vö. Legendre & Legendre 1983) valójában a *l'analyse des correspondances* néven, a francia "iskola" (Benzécri et al. 1973) által kifejlesztett, és széles körben népszerűsített módszernek a változatai, melyekre ma a *"correspondence-analysis"* (COA[5]) gyűjtőnéven hivatkozunk.

7.3.1 Egy intuitív példa és a reciprok átlagolás

A módszer lényegét leginkább az alábbi egy-dimenziós ökológiai példa felhasználásával érthetjük meg. Tételezzük fel, hogy egy, a PCA-nál már említett ökológiai – mondjuk nedvességbeli – háttér-grádiens kimutatását szeretnénk elvégezni mintavételi egységek (pl. cönológiai felvételek) alapján. A felvételekben talált fajokat – korábbi ökológiai vizsgálatok összegzéseképpen – nedvességigényük szerint "pontozhatjuk", mondjuk 0-tól 10-ig. Minden egyes felvételre, minden egyes benne talált faj egyedszámát megszorozzuk a faj nedvességpontjával (vagy "súllyal") s ezeket összeadjuk. A kapott értéket ezután a felvétel össz-egyedszámával elosztjuk, hogy a felvételek között mutatkozó egyedszámbeli eltéréseket kiegyenlítsük. Adott felvétel értéke ezek után annál kisebb, minél inkább dominálnak benne a szárazságtűrő fajok, s annál magasabb, minél nedvességigényesebb fajok szerepelnek benne nagy arányban. Ezek alapján a felvételek sorba rendezhetők, amely már jó közelítése lehet egy valós, nedvesség-grádiensnek. Ez a lényege a Whittaker-féle (1967) *súlyozott átlagolás* módszerének. Nem kell azonban megállnunk ezen a ponton, mert most a felvételek pozíciói szerint tovább finomítható a fajok 0-10-es skálája: egy faj új súlyértékét megkapjuk, ha a felvételek koordinátáit megszorozzuk a faj egyedszámával, ezeket összeadjuk, majd a faj össz-egyedszámával elosztjuk. Ezután célszerű egy átskálázási (standardizálási) műveletet végrehajtani, hogy a súlyértékek egységnyi varianciát adjanak, 0 átlaggal. Az új súlyokból a felvételek egy tovább javított ordinációja állítható elő, abból azután a fajok súlyait számoljuk át megint, és így tovább. Ezt a műveletsort voltaképpen addig folytathatjuk, amíg két, egymás utáni lépésben elért változások már nem haladnak meg egy előre megadott küszöbértéket. A kapott eredmény a felvételek és a fajok egydimenziós ordinációja, amelyben a fajok helyzete jól értelmezi a felvételek pozícióit, és viszont. Ez az iteratív módszer *reciprok átlagolás* (RA, "reciprocal averaging", Hill 1973) néven ismert az ökológiai irodalomban.

Az RA számításmenete az alábbiakban foglalható össze. Legyenek most az $\mathbf{X}_{n,m}$ adatmátrix soraiban a fajok (változók), és oszlopaiban a felvételek (objektumok). A felvételek koordinátáit az első dimenzió mentén a következőképpen kapjuk meg

$$b_j = \frac{1}{\lambda^{1-\alpha}} \sum_{j=1}^{n} a_i \frac{x_{ij}}{u_j}, \quad u_j = \sum_{i=1}^{n} x_{ij} = x_{.j} \tag{7.34}$$

ahol a u_j felvételre vonatkozó összeg, $1/\lambda^{1-\alpha}$ pedig – a fentiekben már említett – skálázási paraméter α=0,5 mellett (de erről lesz még szó a későbbiekben). a_i az i faj súlyértéke ugy-

5 Más szerzők a CA elnevezést részesítik előnyben, amely viszont sokszor a *cluster analysis* rövidítéseként szerepel az irodalomban. E betűszavak tehát semmiképpen sem univerzális jelentésűek, s használatuk inkább csak egy adott cikken vagy könyvön belül következetes.

anezen a tengelyen. Az objektum koordinátáit tehát a változók súlyozott relatív hozzájárulásai adják meg. Az RA az objektumösszeggel való standardizálást eleve tartalmazza. A változók koordinátáira hasonló egyenletet írhatunk fel

$$a_i = \frac{1}{\lambda^{\alpha}} \sum_{j=1}^{m} b_j \frac{x_{ij}}{t_i}, \quad t_i = \sum_{j=1}^{m} x_{ij} = x_{i.} \tag{7.35}$$

Vagyis, a változó pozíciója az objektumok súlyozott relatív hozzájárulásaiból adódik. Ebben a változóra vonatkozó összeggel standardizáltunk.

A 7.34-35 ún. *átmeneti egyenletek* közötti kölcsönös viszony nyilvánvaló: az egyiket a másik segítségével fejezzük ki, a megoldás tehát mindenképpen iteratív jellegű. Az iterációk egy stabilis végeredménybe konvergálnak, függetlenül a kezdeti kiinduló súlyoktól. A kiindulási állapot legfeljebb az iteráció sebességére hat. A módszer tehát akkor is alkalmazható, ha a kezdetben semmiféle információnk nincs a fajok környezeti optimumáról (ellentétben a súlyozott átlagolás módszerével), és rendszerint az objektumokra adjuk meg a kiindulási sorrendet. A 7.34-35 egyenleteknek több megoldása is van, ezek mindegyike megfelel egy tengelynek. Ha az első tengelyre már stabilis eredményt kaptunk, akkor ennek hatását kivonva meghatározható egy erre merőleges második tengely is, és így tovább. Az RA részletes számításmenetét Hill (1973) és Pimentel (1979) közli.

7.3.2 A korreszpondencia-elemzés számításmenete

Míg az RA jól szemlélteti azt, hogy mire is törekszünk az elemzés során, az összes lehetséges tengelyt legegyszerűbben és leghatékonyabban a sajátértékelemzés segítségével határozhatjuk meg. Ezt a részt akár át is ugorhatja a mátrixok iránt kevésbé fogékony Olvasó. Az átmeneti egyenleteket mátrixalgebrai formában a következőképpen írhatjuk fel:

$$\mathbf{B} = \mathbf{U}^{-1}\mathbf{X'}\,\mathbf{A}\,\mathbf{R}^{-1} \tag{7.36}$$

$$\mathbf{A} = \mathbf{T}^{-1}\,\mathbf{X}\,\mathbf{B}\,\mathbf{R}^{-1} \tag{7.37}$$

ahol $\mathbf{U}^{-1}$, $\mathbf{T}^{-1}$ és $\mathbf{R}^{-1}$ diagonális mátrixok $1/u_j$, $1/t_i$ illetve $1/\sqrt{\lambda}$ elemekkel. A 7.37-et behelyettesítve 7.36-ba kapjuk:

$$\mathbf{B} = \mathbf{U}^{-1}\,\mathbf{X'}\,\mathbf{T}^{-1}\,\mathbf{X}\,\mathbf{B}\,\mathbf{R}^{-2} \tag{7.38}$$

Mivel $\mathbf{U}^{-1} = \mathbf{U}^{-1/2}\,\mathbf{U}^{-1/2}$ és $\mathbf{T}^{-1} = \mathbf{T}^{-1/2}\,\mathbf{T}^{-1/2}$, a fenti formula a következőképpen is felírható

$$\mathbf{B} = \mathbf{U}^{-1/2}\,\mathbf{U}^{-1/2}\,\mathbf{X'}\,\mathbf{T}^{-1/2}\,\mathbf{T}^{-1/2}\,\mathbf{X}\,\mathbf{B}\,\mathbf{R}^{-2} \tag{7.39}$$

amely némi átalakítás után az alábbi alakot ölti:

$$\mathbf{R}^2\,\mathbf{U}^{-1/2}\,\mathbf{B} = (\mathbf{T}^{-1/2}\,\mathbf{X}\,\mathbf{U}^{-1/2})'\,(\mathbf{T}^{-1/2}\,\mathbf{X}\,\mathbf{U}^{-1/2})(\mathbf{U}^{1/2}\,\mathbf{B}) \tag{7.40}$$

amelyben $\mathbf{U}^{-1/2}$ és $\mathbf{T}^{-1/2}$ diagonális mátrixok $1/\sqrt{u_j}$ illetve $1/\sqrt{t_i}$, elemekkel; $\mathbf{U}^{1/2}$ tartalmazza a $\sqrt{u_j}$ értékeket az átlóban. Ha bevezetjük a következő jelöléseket: $\mathbf{Z} = \mathbf{T}^{-1/2}\,\mathbf{X}\,\mathbf{U}^{-1/2}$, $\mathbf{V} = \mathbf{U}^{1/2}\,\mathbf{B}$ és $\Lambda = \mathbf{R}^2$, akkor a 7.40 egyenlet a PCA leírásából már ismert alakba egyszerűsödik (vö. 7.6 egyenlet):

$$(\mathbf{Z'Z} - \lambda\mathbf{I})\,\mathbf{v} = \mathbf{0} \tag{7.41}$$

vagyis a Λ mátrixban levő λ-k valamint a **V** mátrixban összesített **v** értékek a **Z'Z** kereszt-szorzat-mátrix sajátértékei illetve sajátvektorai. Ezek meghatározását követően, a $\mathbf{v} = \mathbf{U}^{1/2}\mathbf{b}$ összefüggés alapján megkapjuk az objektumok koordinátáit:

$$\mathbf{B} = \mathbf{U}^{-1/2}\,\mathbf{V} \tag{7.42}$$

Ezután **A**, a változók koordinátáinak mátrixa, a 7.35 átmeneti formula segítségével könnyen előállítható.

7.3.3 Megjegyzések a COA végrehajtásával és az eredmények értékelésével kapcsolatban

A korreszpondencia-elemzés számítógépes megvalósításakor és az eredmények értékelésében vegyük tekintetbe az alábbiakat:

1) Az **X** adatmátrixot célszerű először az összes érték összegével, a *főösszeggel* ($x_{..}$) standardizálni (minden értéket ezzel elosztani):

$$x'_{ij} = x_{ij} / x_{..} \tag{7.43}$$

melynek révén az új értékek összege 1 lesz. Természetesen egy ilyen standardizálás csak olyankor valósítható meg, ha a változók teljesen azonos jellegűek (pl. fajok prezencia/abszencia vagy egyedszám-értékei), és nincsenek negatív adatok. Ebből is kiderül, hogy teljesen eltérő típusú és különféle skálákon mért változók értékelésére a COA nem alkalmas. A COA voltaképpen az adatmátrixot egy nagy kontingencia-táblaként kezeli, amelyben a sorok, ill. az oszlopok csak logikailag összetartozó, egymással összemérhető dolgok lehetnek, maguk a táblázat értékei pedig gyakoriságok vagy így értelmezhető adatok (mint pl. borítási %).

2) Az átmeneti formulákra van egy *triviális* megoldás (a_i=1, b_j=1), amely éppen egy λ=1 sajátértéknek felel meg. Ez az adatmátrix sorainak és oszlopainak a súlypontját magyarázza meg és nincs semmi jelentősége. Ettől a "felesleges" dimenziótól centrálás révén még az elemzés előtt megszabadulhatunk az alábbi módon:

$$y_{ij} = x_{ij} - x_{i.}x_{.j} / x_{..} \tag{7.44}$$

Más szóval, minden értékből kivonjuk a sor- ill. az oszlopösszeg alapján "várt" értéket. Ezt a "trükköt" már ismerjük a klasszikus biometriából, a kétutas kontingencia-táblákra számított χ^2 statisztikából. Mindez még inkább megerősíti azt a fenti megjegyzést, hogy a COA voltaképpen kontingencia-tábla elemző módszer.

A fentiekben említett két módosítás egyidejűleg bekerül az elemzésbe, ha a kiinduló **Z** mátrixot a következő formulával számítjuk ki:

$$z_{ij} = \frac{x_{ij}x_{..} - x_{i.}x_{.j}}{x_{..}[x_{i.}x_{.j}]^{1/2}} \tag{7.45}$$

Mivel a sajátértékelemzést a **Z'Z** mátrixon végezzük, a számítások rövidebb ideig tartanak, ha az **X** mátrixot úgy olvassuk be, hogy az oszlopok száma ne legyen nagyobb a sorok számánál (ui. ekkor lesz **Z'Z** kisebb). Ezt megtehetjük, hiszen a COA az objektumokat és a változókat (α=0,5 esetén) teljesen szimmetrikusan kezeli, azok felcserélhetők egymással, vagyis a COA tökéletesen megfelel az attribútum dualitás (lásd 2.1 rész) elvének. Mindezt nemigen tehetnénk meg pl. a PCA esetében. Az eredmény értékelésekor persze vigyázzunk arra, hogy mi is volt valójában a sorokban és az oszlopokban.

3) A kapott sajátértékek rendszerint kisebbek 1-nél. Négyzetgyökük a *kanonikus korreláció*

$$R_i = \sqrt{\lambda_i} \qquad (7.46)$$

amely az objektum- és változó-koordináták kölcsönösségét fejezi ki, azaz – a példában – mennyire megbízhatók a fajok a felvételek elrendezésére és fordítva: milyen informatívak a felvételek a fajok elrendezésére nézve az *i* tengelyen. Minél nagyobb R_i, annál inkább megfeleltethető a két sorrend egymásnak. Az átmeneti formulákban (mivel α=0,5 volt), voltaképpen a kanonikus korreláció reciprokát használtuk.

A sajátértékek összege a kontingencia-táblaként felfogott adatmátrixra adódó χ^2. Mivel korábban a főösszeggel elosztottunk minden értéket, az eredeti adatmátrix esetében:

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} \sum_{i=1}^{t} \lambda_i = x_{..} \sum_{i=1}^{t} \lambda_i \qquad (7.47)$$

A pozitív sajátértékek számára, *t*-re, ugyanazok a megállapítások érvényesek, mint a PCA esetében. Ez az adatmátrix értékei és a várt értékek közötti eltérések megmagyarázásához szükséges ortogonális dimenziók száma. Ha tehát az adatmátrixban minden érték az oszlop- és a sorösszegek alapján várható értékkel megegyező (azaz χ^2=0), akkor minden sajátérték 0. Más szóval, a sajátértékek nagyságából következtethetünk az adatmátrix strukturáltságának, a várttól vett eltérésének a mértékére.

4) A sorokra és az oszlopokra kapott koordinátákat külön ordinációs diagramokon ábrázolhatjuk (értelmezésük a szokásos módon történik), és minden további nélkül felhasználhatjuk a COA biplot készítésére is. Ez utóbbi interpretációja azonban más, mint a PCA esetében (s gyakran nem biplotnak, hanem *"joint plot"*-nak nevezik, vö. Oksanen 1987). Amíg ott a változókra mutató nyilak iránya és relatív hossza volt interpretatív értékű, a COA biplotban a felvételeket és a változókat reprezentáló pontok közelségének *is* lehet jelentősége, s a nyilak elmaradhatnak. A COA biplot értelmezése és értelmezhetősége azonban nagymértékben függ az α paramétertől és a sajátértékek nagyságától. A fentiekben leírt COA eredményében a változókat (a mátrix sorait) és az objektumokat (a mátrix oszlopait) szimmetrikusan tekintettük, amit az α=0,5 beállítással értünk el. Ennek a paraméternek azonban szabadon változtatható az értéke a [0,1] intervallumon belül (vö. ter Braak 1985), s ezért végtelen számú lehetséges eredményt kaphatunk. Ezek közül még kettőnek van kitüntetett szerepe, elsősorban ökológiai ordinációkban.

Ha α=1, akkor a felvételek (most: oszlopok) koordinátáit a fajok koordinátáinak súlyozott átlaga adja meg. Ekkor a biplotban az *i* faj azokhoz a felvételekhez kerül legközelebb, amelyekben a legnagyobb arányban fordul elő, vagyis a pozíció az *i* faj optimumhelyét "becsli". Lehetnek olyan fajok, amelyek optimuma kívül esik a vizsgált felvételeken, ezért értékeik rendszerint nagyobb tartományt ölelnek fel, mint a felvételeké. Ha van olyan felvétel, amelyben csak egy faj szerepel, akkor ez a felvétel egybeesik az illető faj pozíciójával a biplotban. Sok COA program eleve ezt az opciót tartalmazza csak (pl. **DECORANA**, Hill 1979b). Az α=0 esetben fordított a helyzet: a fajok (most: sorok) koordinátáit a felvételek koordinátáinak a súlyozott átlaga adja meg (átskálázás nélkül, ui. $1/\lambda^{\alpha}$=1). Ekkor rendszerint a fajokat reprezentáló pontok közelebb kerülnek az origóhoz, és a felvételeknek jut szélesebb értéktartomány. Ha van olyan faj, amely éppen egy helyen fordul csak elő, akkor ez pontosan egybeesik azzal a hellyel a biplotban. Az α=0,5 beállítás e két konfigurációnak az átlagát adja. Ha a

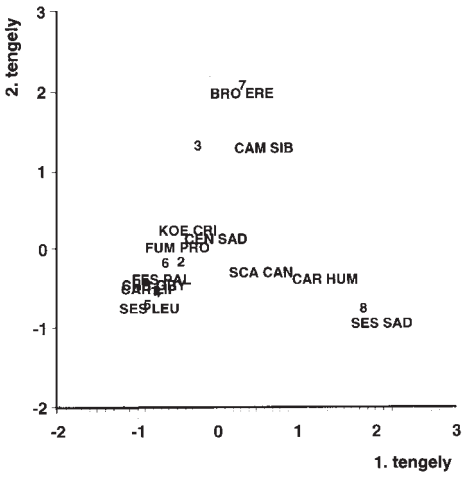
sajátértékek közel állanak 1-hez, vagyis az adattáblázat erősen strukturált (a χ^2 értéke magas), akkor az α skálázási paraméter változtatása csak kis különbségeket eredményez a biplotban. Ilyenkor a sorok és az oszlopok közelsége határozott értelmű. Alacsony sajátértékek esetén – vagyis kevéssé strukturált adatmátrixokra – viszont a pozicionális közelség nem egyértelmű, s csak a szögek és irányok adnak útmutatást. Mindezt példák illusztrálják majd a következő részben (7.16a-c ábrák).

5) A 7.43-44 standardizálásokat, illetve a COA 7.41 mátrix-egyenletét tekintve talán nem lep meg bennünket a következő állítás: a korreszpondencia-elemzés a PCA egy speciális formája (Greenacre & Vrba 1984 pl. ilyen értelemben mutatja be a módszert). Míg a centrált PCA az objektumok közötti euklidészi távolságokat őrzi meg az ordinációban, a COA az úgynevezett χ^2-távolságokat konzerválja (3.67 formula). Pontosabban: α=1 mellett az oszlopok ordinációjában – az összes tengelyre nézve – a *j* és *k* pontok távolságnégyzete arányos a

$$CHISQD_{jk} = \sum_{i=1}^{n} \frac{(x_{ij} / x_{ij} - x_{ik} / x_{.k})^2}{x_{i.}} \qquad (7.48)$$

mennyiséggel. (Ez voltaképpen a sor- és oszlopösszeggel standardizált adatokból számolt euklidészi távolság négyzete.) $CHISQD_{jk} = 0$, ha az eredeti adatmátrixban a két objektum pontosan egyenlő arányban tartalmazza a változókat (az egyik objektum a másiknak pontosan *q*-szorosa, *q* egy tetszőleges nemnegatív szám). Amennyiben α=0, a sorok ordinációja tartja meg a távolságviszonyokat (a sorokra a 7.48-hoz hasonló egyenlet írható fel). $CHISQD_{hi} = 0$, ha a *h* és *i* változó mindig azonos arányban jelentkezik az objektumokban (pl. az egyik faj mindig *q*-szor nagyobb egyedszámban jelentkezik, mint a másik).

Most már itt az ideje, hogy a módszert egy konkrét példán is bemutassuk. Vegyük az A1 táblázat adatait, amelyeket egy 12×8-as kontingenciatáblázat értékeinek fogunk fel. A jelenlegi elrendeződés éppen kedvező az analízis gyorsasága szempontjából, mert az oszlopok száma a kisebb. (Ha azonban – mondjuk – 30 faj jellemez több száz felvételt, akkor inkább a fajokat tegyük az oszlopokba.) A szimmetrikus COA eredménye, biplot formájában, a 7.14 ábrán látható. A PCA eredményekkel összehasonlítva túlságosan nagy eltérések nem mutatkoznak, ezeket a sor- ill. oszlopösszeggel való standardizálás okozza. Az aszimmetrikus COA eredményeket nem mutatjuk be, mivel a viszonylag magas első sajátérték miatt (λ_1=0,7) nem



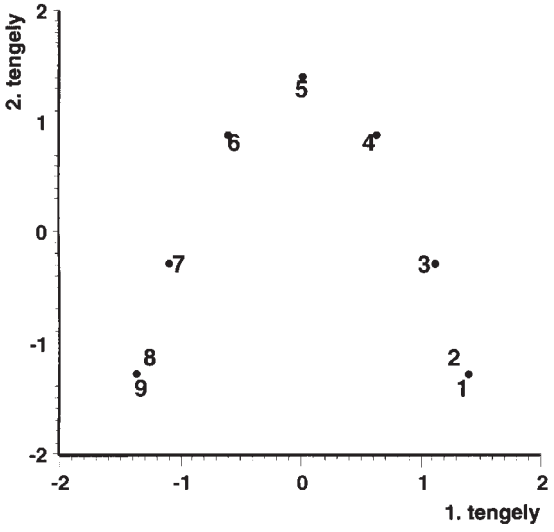
7.14 ábra. Az A1 táblázat adatainak korreszpondencia-elemzése (α=0,5). Hasonlítsuk össze az eredményt a PCA diagramokkal (7.2-6 és 7.8 ábrák).

különböznek jelentékenyen a 7.14 ábra diagramjától. Most már eléggé gyakorlottak lehetünk a patkó-jelenség felismerésében, amely mintha most is jelentkezne az eredményben.

7.3.4 A patkó-jelenség és az adatok linearitása a korreszpondencia-elemzésben

Mivel a PCA és a COA lényegileg ugyanarra a sajátérték-problémára vezethető vissza, a patkó-jelenség a COA esetében sem ismeretlen. A 7.1.6 részben megadott 7.14 adatmátrix elemzése a 7.15 ábra biplotját eredményezi, amelyen – mint talán várható is – egy kettős ívet figyelhetünk meg.

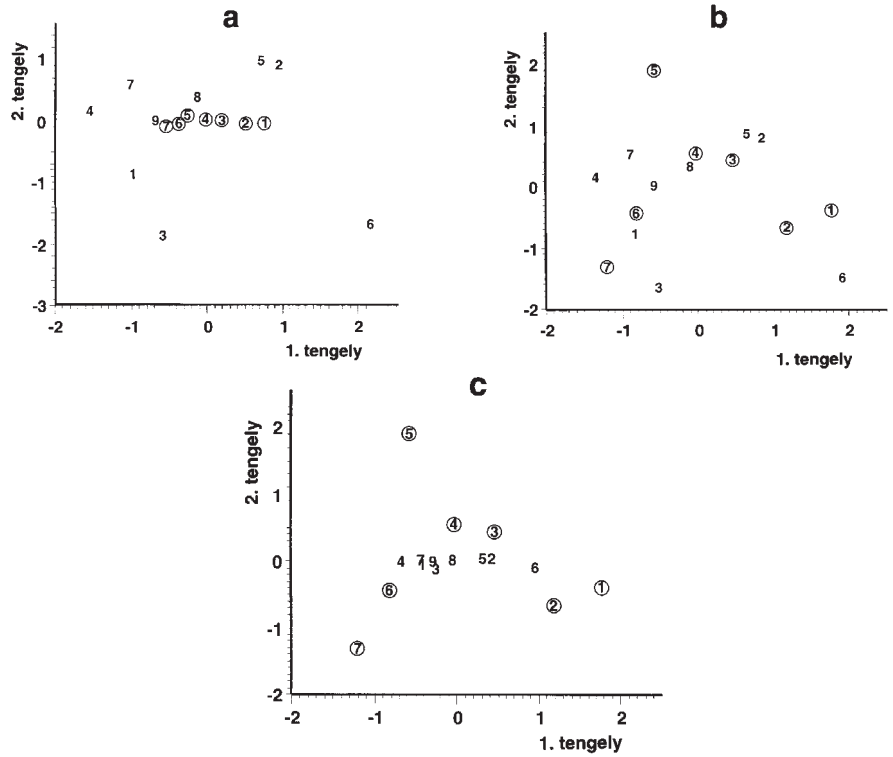
Az ok már ismert: az adatok nem-lineáris jellege, melynek következtében a pontok közötti távolságokat csak úgy tudja az elemzés többé-kevésbé hűen megtartani, ha a pontokat patkó alakba rendezi. Ugyanakkor a változók és az objektumok közötti "korreszpondencia" kiválóan tükröződik az eredményben. Az interpretációt tehát csak részben zavarja – ha zavarja – a patkó-hatás (Greenacre 1984), ennek ellenére "kiküszöbölését" számos szerző fontosnak találja. A Hill (1979b) és Hill & Gauch (1980) által kidolgozott "*detrended correspondence analysis*" (DCA) módszere például meglehetősen önkényes módon szegmensekre osztja az első tengelyt, majd a szegmensek vertikális elcsúsztatásával ér el jobb illeszkedést az egyenesre. Mivel a DCA módszere rendkívül népszerű mind a mai napig az ökológiai adatok ordinációjában (lásd pl. a *Vegetatio*-ban vagy a *Journal of Vegetation Science*-ben megjelent cikkeket), megerősítjük a PCA-nál mondottakat: önmagában semmiképpen sem ajánljuk a DCA használatát, legfeljebb a standard COA-val párhuzamosan (újat a DCA úgysem fog mondani, legfeljebb "esztétikai" javulást ér el). A DCA potenciális veszélye, hogy "fekete doboz" módjára működik, és esetleg ökológiailag értelmes információt veszíthetünk általa (Pielou 1984). A "detrendeljünk vagy ne detrendeljünk" kérdése egyébként elég kemény szakmai viták tárgya (vö. pl. Gauch 1982, Kenkel & Orlóci 1986, Minchin 1987, Wartenberg et al. 1987, Peet et al. 1988, Oksanen 1988, Knox 1989). Reyment (1991) pl. a "pudding próbája az evés" közhellyel nyitva hagyja a kérdést, mondván, hogy sok múlik az adott vizsgálati szituáción is. Jongman et al. (1987) és ter Braak & Prentice (1988) a szegmentálás helyett a polinomiális regresszió módszerét ajánlja (mint Phillips, 1978, a PCA esetében) és ez opcióként szerepel ter Braak (1988) **CANOCO** programjában is.



7.15 ábra. A korreszpondencia-elemzés eredménye (α=0,5) erősen nem-lináris adatszerkezet esetén. A második tengely láthatólag másodfokú függvénye az első tengelynek. Az első két sajátérték magas (0,92 illetve 0,72), így az α paraméter változtatása nem módosítja lényegesen a pontok relatív helyzetét.

Nézzük meg most, hogy miképpen "vizsgázik" a COA módszere a linearitás feltételét teljesítő változók esetén. A kiindulás, akár a PCA esetében, a 7.15 mátrix lesz, amelyben az oszlopok felelnek meg a változóknak (a közöttük fennálló közelítőleg lineáris kapcsolat nyilvánvaló).

Az elemzés jól jelzi, hogy az adatmátrix egyes értékei már nem olyan feltűnően "váratlanok", mint a 7.14 mátrix esetében, hiszen az első sajátérték mindössze 0,19, a második pedig 0,003 (relatíve persze az első sajátérték kiemelt fontosságú, e tekintetben nincs különbség a PCA-tól). Ennek következtében az α paraméter megválasztása már döntően befolyásolja a pontok közelségét a biplotban (7.16a-c ábra). Ha a sorok ("felvételek") koordinátáit a fajok koordinátáinak a súlyozott átlaga adja meg (7.16a ábra), akkor a hét objektum majdnem teljesen illeszkedik az 1. komponensre, tehát a lineáris elrendeződést a COA is kimutatja. A χ^2-távolság-viszonyok ezen a diagramon a felvételek között tükröződnek. A fajok (oszlopok) a "külső körön" helyezkednek el, mutatva, hogy legtöbbjüknek nem esik az optimuma a felvételek közelébe. Amint arra ter Braak & Prentice (1988) rámutat, a COA valójában a fajok *unimodális* reakcióját feltételezi a grádiensre (ellentétben a PCA-val), s ennek hiánya vezet az alacsony sajátértékekre. (Persze az unimodális viselkedés, mint fent láttuk, óhatatlanul előhívja a patkó-jelenséget.) A fordított esetben (7.16c ábra) az oszlopok (fajok) koordinátáit a felvételek koordinátáinak súlyozott átlaga adja meg, s így a fajok kerülnek az origó közelébe, ugyancsak egy sorba rendezve. A közöttük lévő távolságok közelítőleg arányosak a χ^2-távolságokkal. A szimmetrikus COA (7.16b ábra) a két előző konfiguráció "átlagának" felel meg. A pontok közelségéből nem következtethetünk egyértelműen arra, hogy az egyes fajok mely helyeket jellemzik. Az irányoknak azonban határozott jelentésük van.



7.16 ábra. A 7.15 mátrix (227. oldal) adatainak korreszpondencia-elemzése az α skálázási paraméter három különböző értékére (**a:** α=1, **b:** α=0,5, **c**: α=0). Az **a** esetben a sorok értékeit az oszlopok koordinátáinak súlyozott átlaga adja, a **c** esetben ez fordítva van, a **b** eset pedig kompromisszumot jelent.

7.3.5 Kanonikus korreszpondencia-elemzés

A COA módszerének is van kötött formája, a *kanonikus korreszpondencia analízis* (CCOA, ter Braak 1986, 1987), amely ma már rendkívül széles körben használatos az ökológiai adatelemzésben. “Komoly” folyóiratok szinte el sem fogadják a fajok és a környezeti változók közötti összefüggésekről szóló cikkeinket (pl. grádiens elemzés), ha a CCOA alkalmazásáról “megfeledkezünk”. A módszer alapjairól és az eredmények értelmezéséről mindenképpen szólnunk kell tehát.

Az alapgondolat hasonló az RDA-éhoz (7.2.5 rész), csak PCA helyett korreszpondencia-elemzést alkalmazunk. Az objektumokat nem tisztán a fajadatok alapján ordináljuk, mert a tengelyeknek – amellett, hogy a lehető legnagyobb varianciát magyarázzák meg az adatokból – a környezeti változók lineáris kombinációjaként kell adódniuk. Az egyes CCOA tengelyek egymással korrelálatlanok, e tekintetben már nincs eltérés a standard COA-tól. A megszorítás miatt azonban az összvarianciából kisebb hányad esik a CCOA tengelyekre, mint amekkorát a COA esetében elérhetnénk. Ezt az “áldozatot” kell meghoznunk annak érdekében, hogy a tengelyek közvetlenül interpretálhatók legyenek. Miután a tengelyek helyzetét a környezeti változók befolyásolják, a CCOA a direkt grádiens elemzés egyik fő módszerének tekinthető (ter Braak & Prentice 1988).

A CCOA számításmenete legkönnyebben a reciprok átlagolással (7.3.1 rész) több lépésben megegyező iteratív algoritmus alapján érthető meg (Jongman et al. 1987). Az **X** adatmátrix soraiban vannak a fajok, oszlopaiban pedig az objektumok. A **Z** adatmárix soraiban a környezeti változók, oszlopaiban pedig ugyanolyan sorrendben, mint **X**-ben, az objektumok szerepelnek. Legyen a környezeti változók száma q. Ezek után az első tengelyen az objektumok és a fajok koordinátái a következőképpen határozhatók meg:

1. Az objektumok (felvételek, helyek, stb.) önkényes, de különböző koordinátáiból (b_j) indulunk ki.

2. Minden egyes faj koordinátáit meghatározzuk az objektumok koordinátáinak súlyozott átlagaként, vagyis $a_i = \Sigma_j\, b_j\, x_{ij} / t_i$, ahol t_i az i faj értékeinek összege (sorösszeg).

3. A fajok új értékei alapján súlyozott átlagolással kiszámítjuk az objektumok új koordinátáit, vagyis $b_j = \Sigma_i\, a_i\, x_{ij} / u_j$, ahol u_j a j objektumban lévő fajok értékeinek az összege (oszlopösszeg). Ezek az ún. "WA - weighted average - score"-ok.

4. Az objektum-koordinátákat (“függő” változó), egyenként az mennyiség szerint súlyozva, a *többszörös regresszió* módszerével illesztjük a környezeti változókra (“független” változók). A kapott c_h kanonikus koefficiensek segítségével az alábbi összefüggésből

$$b_j = c_0 + \sum_{h=1}^{q} c_h z_{jh} \qquad (7.49)$$

megállapítjuk a j objektumnak a regressziós egyenesre illesztett értékét, ami majd az új koordináta lesz. (z_{jh} a h környezeti változó értéke a j objektumban.) Ezeket nevezik a szakirodalomban "LC - linear combination - score"-oknak.

5. Az új koordinátákat átskálázzuk (standardizáljuk): a súlyozott értékekből kivonjuk az átlagot, s elosztjuk a szórással.

6. Ha a most kapott objektumkoordináták és az előző iterációs ciklusban számolt koordináták eltérése nem nagyobb egy előre megadott ε küszöbértéknél, akkor az iteráció véget ér. Ellenkező esetben visszatérünk a 2. lépésre.

A reciprok átlagolástól vett fő különbség tehát a 4. lépésben alkalmazott regresszióelemzés. Az iterációs algoritmus bármilyen kiindulásból ugyanabba a végeredménybe konvergál, legfeljebb a lépések száma változó. Az iteráció végeztével az objektumok tengelyén a pontok szórásnégyzete éppen az első sajátértéknek felel meg. Az 1. tengely hatását kivonva ezután meghatározható egy második CCOA tengely is, amely lineárisan korrelálatlan az elsővel, majd megállapíthatunk egy harmadik tengelyt, amely ortogonális az első kettőre, és így tovább (ter Braak 1987). Annyi tengelyünk lehet, ahány környezeti változónk van.

A CCOA eredmények értékelésében a következőket vehetjük figyelembe:

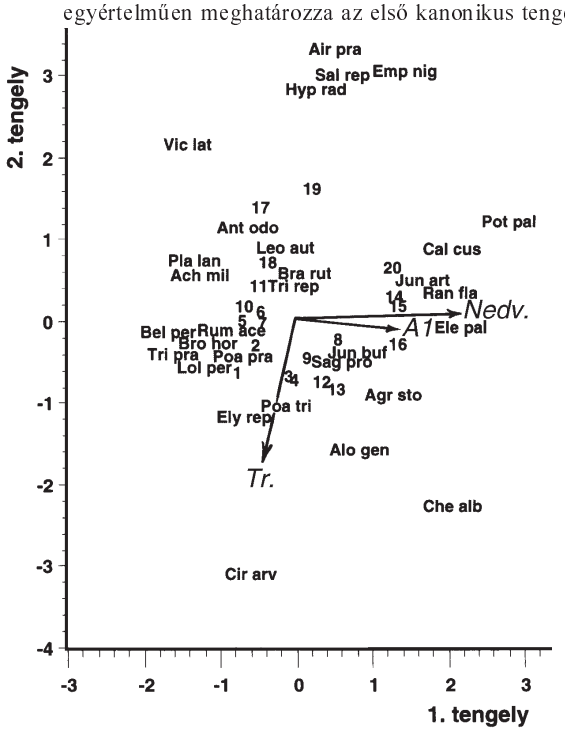
- A WA és LC eredmények rendszerint nem nagyon különböznek, bár az LC értékeket jelentősen befolyásolhatják a környezeti változók "hibái". Így inkább a WA értékeket javasolhatjuk.
- A környezeti változók és a fajok kapcsolatának erősségét adott tengelyre az ún. *faj/környezet korreláció* fejezi ki. Ez az objektumoknak az utolsó regresszió előtti koordinátái és a változók lineáris kombinációjára (az új tengelyre) illesztett koordinátái közötti lineáris korreláció. Ennek nagysága azonban önmagában nem elegendő, mert a sajátértéket is célszerű figyelembe vennünk. Ugyanis alacsony sajátértékű tengely is adhat magas faj/környezet korrelációt, de ez a tengely a varianciának csak kis hányadát magyarázza meg, s ezért mégsem tulajdoníthatunk neki nagy jelentőséget.
- Az ordinációs tengelyek és a környezeti változók közötti korrelációk igen fontosak a tengelyek interpretációjában. Ezek a korrelációk azonban, mivel a tengelyek a vizsgálatba bevont *összes* környezeti változó lineáris kombinációi, már változhatnak, ha új változókat veszünk bele az elemzésbe, vagy ha egyes változókat kihagyunk (ez ugyanis megváltoztatja a többszörös regressziót).
- A fajok és az objektumok koordinátáit biplotban ábrázolhatjuk, a COA-nál már ismertetett módon. Azon objektumok, amelyben egy faj relatíve magas értékkel szerepel, közel lesznek a faj pozíciójához (α-tól függően, mint láttuk). Ebben a diagramban feltüntetjük a környezeti változókat is, hogy ezek hatása is érzékelhető legyen. Arány- vagy intervallum-skálán mért környezeti változókat nyilak segítségével mutatjuk be (hasonlóan a PCA-hoz). A nyilak irányának és a fajok helyzetének együttes interpretációja a következő: a fajokat képzeletben levetítjük a nyílra, mint a környezeti változónak megfelelő “tengelyre”. Ekkor egy fajsorrendet kapunk, amely nagyjából (nem tökéletesen) tükrözi a fajoknak a környezeti változó szerinti sorbarendezését. Így azonosíthatók azok a fajok, amelyek a változóval pozitívan vagy negatívan kapcsolódnak (lásd a lenti példát). A hosszú nyilak erősebben korrelálnak a tengelyekkel, mint a rövidek, így elsősorban ezek jöhetnek számításba a fajösszetétel

elemzésében. Nominális változókat is bevonhatunk az elemzésbe; ezeket annyi bináris változóval helyettesítjük, amennyi az illető nominális változó állapotainak a száma (vö. 1.4.1 rész). Ezek a változók rendszerint nem nyilakkal szerepelnek a diagramban (Jongman et al. 1987, p. 142), hanem annyi pontként, ahány állapotuk van. A változó egy adott állapotára vonatkozó koordinátáját azon objektumok súlyozott koordinátáinak átlaga adja meg, amelyekben ezt az állapotot megfigyeltük.

- Mivel az objektumok távolságai közvetlenül nem szerepelnek az elemzésben, a patkójelenség a CCOA esetében ritkább, mint a COA-nál, és ha előfordul, akkor sem kifejezett. Ter Braak & Prentice (1988) az esetleges patkó-jelenséget a túl sok, az elemzésbe "feleslegesen" bevont környezeti változónak tulajdonítja, s ezek kihagyásával a "probléma" megszüntethető. (A kihagyás persze mindig egy önkényes művelet marad.) Ha a környezeti változók száma viszonylag kevés, és jól kifejeződnek a CCOA tengelyeken, a patkó valószínűleg teljesen el is marad.

A kanonikus-korreszpondencia elemzést ter Braak (1988) példaadataival szemléltetjük (A4 táblázat). Az adatmátrixban 30 faj és 20 cönológiai felvétel (mintavételi hely) szerepel, ez utóbbiakat három környezeti változó is jellemzi: az A1 talajszint mélysége, a talaj nedvességtartalma és a trágyázás mennyisége. Az elemzés révén megmutatható, hogy a dünevegetáció fajösszetételében milyen mértékű és irányú változásokat okoznak eme környezeti jellemzők.

Az elemzést úgy hajtottuk végre, hogy a mintavételi helyek pozícióját a fajok koordinátáinak a súlyozott átlaga adja (7.17 ábra). Ennek következtében számos faj került a diagram szélére. A három környezeti változó közül az A1 szint magassága és a nedvességtartalom egyértelműen meghatározza az első kanonikus tengelyt (0,56 és 0,90-es korrelációval), míg a



7.17. ábra. A dünevegetáció (A4 táblázat) kanonikus korreszpondencia-elemzése. A diagramon háromféle típusú pontok szerepelnek: mintavételi helyek (1-20), fajok és a környezeti változók (A1, Nedv. és Tr.). Emiatt akár *triplot*nak is nevezhetjük (Šmilauer 1992).

második tengelyt inkább a trágyázás mennyisége befolyásolja (r = –0,79). Mindezt a nyilak iránya és relatív hossza is érzékelteti a diagramon. Az első két tengelyhez tartozó sajátértékek 0,42 ill. 0,23, amelyek a fajok mintavételi helyek mátrixnak "sima" COA elemzéséből kapott sajátértékeinél (0,53 ill. 0,40) kisebbek. Ez várható is, hiszen a CCOA tengelyeknek a környezeti változók lineáris kombinációinak kell lenniük, és ezáltal a legritkábban esnek egybe a COA tengelyekkel, amelyek tehát – önmagukban – hatékonyabb variancia-sűrítők. Mindazonáltal a különbség nem jelentős, amit a magas faj/környezet korrelációk mutatnak (0,925 az első tengelyen, 0,816 a másodikon). A fajok pozíciójának és a nyilaknak a kölcsönös értelmezéséhez vegyük a nedvességtartalmat. Ha a nyilat képzeletben mindkét irányban meghosszabbítjuk és erre levetítjük a fajokat képviselő pontokat, akkor többé-kevésbé megkapjuk a fajok nedvességigény szerinti sorrendjét: a jobb szélen azok a fajok szerepelnek, amelyek a nedvesebb helyeken szerepelnek, a baloldalon pedig a szárazabb körülmények között élő fajok sorakoznak. Ez a sorba rendezés annál hűségesebben tükrözi a valós viszonyokat, minél magasabb a két tengelynek megfelelő sajátérték relatív fontossága (ebben az esetben 25 % és 15 %).

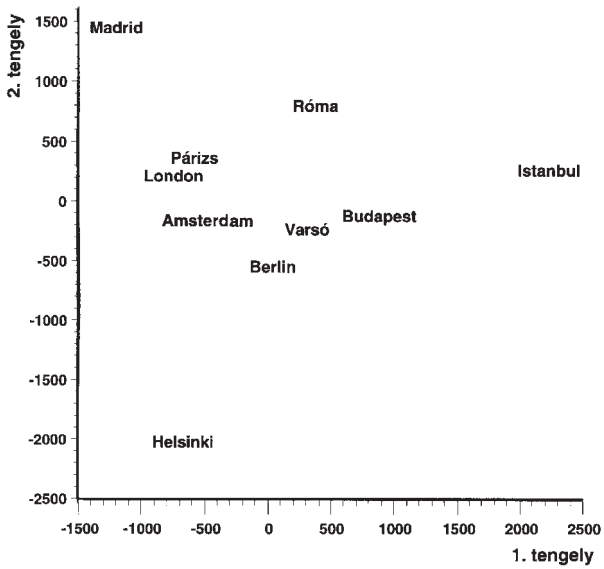
7.4 Többdimenziós skálázás

Az előző részekben megismert ordinációs eljárások közös sajátossága, hogy a számítások során a nyers adatokra mindvégig szükség van. Sok esetben azonban a mintavételezés vagy a mérés közvetlenül távolság- vagy különbözőség-mátrixot eredményez, mint erre már a kladisztikáról szóló részben (6.2 alfejezet) utaltunk (ismert példa a Sarich-féle immunológiai távolságok A5 mátrixa). Felmerül a kérdés, ha távolságok alapján lehetőségünk van az evolúciós utak rekonstrukciójára, akkor van-e mód hatékony dimenzió-redukcióra is? E kérdést persze nem tettük volna fel, ha nem az *igen* lenne rá a válasz: az ún. *többdimenziós skálázás* (rövidítve: t. d. s.) módszerei képesek az objektumok távolságmátrixából ordinációt előállítani.

A többdimenziós skálázás tematikája önmagában is rendkívül szerteágazó, figyelmünket azonban két területre összpontosítjuk. A *metrikus* t.d.s. módszere az elgebrai megoldást tekintve közvetlen rokonságban áll a főkomponens-elemzéssel. A *nem-metrikus* t. d. s. eljárásai sokkal kevesebb feltételt szabnak a kiinduló mátrixszal szemben, mint a metrikus módszerek, és sok esetben az egyetlen megoldást adják az adott problémára. Algoritmikus elveik alapján teljes mértékben különböznek az eddigiektől.

7.4.1 Metrikus többdimenziós skálázás avagy a főkoordináta módszer

A módszer eredetileg Torgerson (1952) nevéhez fűződik, de igazán Gower (1966) munkássága révén *főkoordináta módszer* néven vált népszerűvé (alkalmas rövidítése: PCoA, míg Digby & Kempton [1987] a PCO betűszót használja). A PCoA azért metrikus, mert az ordinációban megőrzi az objektumok közötti távolságviszonyokat (akár csak a PCA). Annyi ordinációs tengelyt állítunk elő, amennyi a kiinduló mátrixban lévő metrikus információ tökéletes megtartásához szükséges. Alkalmazásának feltétele tehát, hogy a távolságok teljesítsék a metrikus axiómákat (3.1.1 rész), bár – mint ezt majd a későbbiekben részletezzük – ezek kismértékű megsértése sem teszi lehetetlenné a PCoA eredmények interpretálhatóságát. A főkoordináta-módszer illusztrálásának tipikus példája nem biológiai ugyan, de – szemléletessége miatt – sok könyv ezt említi először (pl. Manly 1986) és mi is ezt tesszük. Nagyvárosok közötti úttávolságok félmátrixa gyakran szerepel a térképek hátoldalán. Ennek alapján a PCoA képes a városok relatív pozícióját, azaz egy térképet rekonstruálni, bár ennek



7.18 ábra. Tíz európai nagyváros relatív elhelyezkedésének rekonstrukciója az A6 úttávolság-mátrixból a főkoordináta módszer alkalmazásával.

sikere az utak kanyargósságának a függvénye. Ha az utak nem nyílegyenesek (s rendszerint nem olyanok), akkor a PCoA eredménye az első két dimenzióban csupán megközelíti a valós helyzetet, és további dimenziók kellenek a "kanyargósság" megmagyarázására. Teljesen egyenes utak esetében viszont a távolságmátrix belső dimenzionalitása (rangja, C függelék) kettő, így a térkép a papír síkjában torzítás nélkül előállítható PCoA-val.

Vegyük példaként tíz európai nagyváros úttávolságainak mátrixát (A7 táblázat). A mátrix főkoordináta-elemzése (7.18 ábra) a városok egymáshoz viszonyított földrajzi helyzetét meglehetősen jól reprodukálja, a városok közötti távolságok a tengelyeken feltüntetett lépték szerint elég pontosan leolvashatók a diagramról. Az irányokkal viszont mintha bajban lennénk: el kell forgatni és tükrözni a teljes konfigurációt, hogy az égtájak is érzékelhetőek legyenek. A két bemutatott tengely hatékonyságának értékelésére később térünk ki.

A főkoordináta-elemzés két fő lépésben hajtható végre. Az első lépés az igazi "trükk": a távolságok felhasználásával egy szimmetrikus mátrixot állítunk elő, amely éppen a későbbiekben meghatározandó koordinátákból kiszámítható keresztszorzat mátrixnak fogható fel (csakúgy, mint a kovariancia vagy a korrelációs mátrix a PCA-ban, vagy a **Z'Z** mátrix a COA-ban). A következő lépés ezen mátrix sajátérték elemzése, amely a már ismert módon a sajátértékeket és sajátvektorokat, ezekből pedig magukat a koordinátákat eredményezi.

Az $m \times m$-es méretű **A** keresztszorzat-mátrixot az $\mathbf{X}_{n,m}$ koordinátákból, ha azok már ismertek lennének, az alábbi formula segítségével kapnánk meg:

$$a_{jk} = \sum_{i=1}^{n} x_{ij} x_{ik} \qquad (7.50)$$

vagyis, mátrixalgebrai megfogalmazásban:

$$\mathbf{A} = \mathbf{X'X} . \qquad (7.51)$$

A pontok relatív helyzete nem változik meg, ha a meghatározandó koordinátákat centráljuk, s ezáltal majd egy triviális sajátértéket eleve kiküszöbölünk (hasonlóan a COA-hoz), vagyis

$$\sum_{j=1}^{m} x_{ij} = 0, \text{ minden } i \text{ változóra} \tag{7.52}$$

Ebből következően az **A** mátrixban a sorösszegek és az oszlopösszegek értéke is 0:

$$\sum_{j=1}^{m} a_{jk} = \sum_{k=1}^{m} a_{jk} = 0 \tag{7.53}$$

Most feltételezzük, hogy a kezdeti, *négyzetre emelt* távolságok, d^2_{jk} a keresett koordináták alapján a következők:

$$d^2_{jk} = \sum_{i=1}^{n} (x_{ij} - x_{ik})^2 \tag{7.54}$$

ami ekvivalens a következő felírással:

$$d^2_{jk} = \sum_{i=1}^{n} [x^2_{ij} + x^2_{ik} - 2x_{ij}x_{ik}] = \sum_{i} x^2_{ij} + \sum_{i} x^2_{ik} - 2\sum_{i} x_{ij}x_{ik} \tag{7.55}$$

A 7.50 összefüggés alapján ez az alábbiak szerint írható át:

$$d^2jk = a_{jj} + a_{kk} - 2a_{jk} \tag{7.56}$$

Ezután 7.56-ból a_{jk}-t kifejezzük:

$$a_{jk} = 1/2\ [-d^2_{jk} + a_{jj} + a_{kk}] \tag{7.57}$$

amelyet, itt nem részletezett behelyettesítések után (l. például Pielou 1984, p. 184) teljes egészében átírhatunk a távolságnégyzetek felhasználásával:

$$a_{jk} = 1/2\ [d^2_{jk} - d^2_{j.} - d^2_{.k} + d^2_{..}] \tag{7.58}$$

ahol

$$d^2_{j.} = \frac{1}{m} \sum_{k=1}^{m} d^2_{jk} \tag{7.59}$$

a j objektum és a többi objektum közötti távolságnégyzetek átlaga, és

$$d^2_{..} = \frac{1}{m^2} \sum_{j} \sum_{k} d^2_{jk} \tag{7.60}$$

pedig az összes távolságnégyzet (ide értve az átlóban lévő 0-kat) átlaga.

A fenti levezetésből látszik, hogy a PCoA a 7.58 egyenlet szerint kapott **A** mátrixból indul ki. Miután a már ismert módon meghatároztuk ennek sajátértékeit és sajátvektorait:

$$(\mathbf{A} - \lambda\mathbf{I})\ \mathbf{v} = \mathbf{0} \tag{7.61}$$

úgy, hogy a sajátvektorok egységnyi hosszúságúak, a sajátértékeket pedig nagyság szerint csökkenő sorrendbe tesszük, akkor a mátrixok spektrálfelbontásának tételét (C függelék) alkalmazhatjuk. Eszerint az **A** szimmetrikus mátrix a következőképpen is felírható:

$$\mathbf{A} = \mathbf{V}\,\Lambda\,\mathbf{V}' = (\mathbf{V}\,\Lambda^{1/2})(\Lambda^{1/2}\,\mathbf{V}') \qquad (7.62)$$

amelyben Λ egy diagonális mátrix a sajátértékekkel. A 7.51 és 7.62 egyenletek alapján megkapjuk a keresett koordinátákat:

$$\mathbf{X} = \Lambda^{1/2}\,\mathbf{V}' = [\sqrt{\lambda_1}\,\mathbf{v}_1, \sqrt{\lambda_2}\,\mathbf{v}_2, \ldots, \sqrt{\lambda_m}\,\mathbf{v}_m] \qquad (7.63)$$

A főkoordináta módszer megértéséhez valamint eredményének értékeléséhez az alábbiakat vehetjük figyelembe:

- Egy $n \times m$-es adatmátrixból, az n változó között számított kovarianciákból kiinduló PCA (centrált PCA) és az m objektum közötti távolságnégyzetek mátrixából végrehajtott PCoA teljesen azonos objektum-ordinációt eredményez, legfeljebb a koordináták előjelében lehet eltérés. Mindez nem lephet meg bennünket: a megoldás alapja, a sajátértékelemzés, ugyanis közös a két módszerben. A COA (az objektumok koordinátáit a fajok koordinátáinak súlyozott átlagaként véve) és a χ^2-távolságokon alapuló PCoA az összes dimenzióra nézve ugyanazt a távolságstruktúrát tárja fel, bár itt az első két dimenzióra már eltérések adódhatnak (vö. Digby & Kempton 1987).
- Ha a kiinduló távolságmátrix hiánytalanul megfeleltethető euklidészi távolságok segítségével, akkor legfeljebb m–1 pozitív sajátértéket kapunk, s az m-edik értéke 0. Ebben az esetben az **A** mátrix átlójában a pontoknak a centroidtól vett távolságnégyzete szerepel. Ezek összege, vagyis tr { **A** }, az összes pontra vonatkozó négyzetösszeg, amelyet a pontok közötti páronkénti távolságok segítségével is kifejezhetünk (vö. 3.106 egyenlet). Ez a mennyiség éppen a sajátértékek összegével egyenlő:

$$\mathrm{tr}\{\mathbf{A}\} = \sum_{j=1}^{m} a_{jj} = \sum_{j}\sum_{k} d_{jk}^2 / 2m = \sum_{k=1}^{m-1} \lambda_k \quad (7.64)$$

Következésképpen, az első t dimenzió a teljes távolságstruktúrát

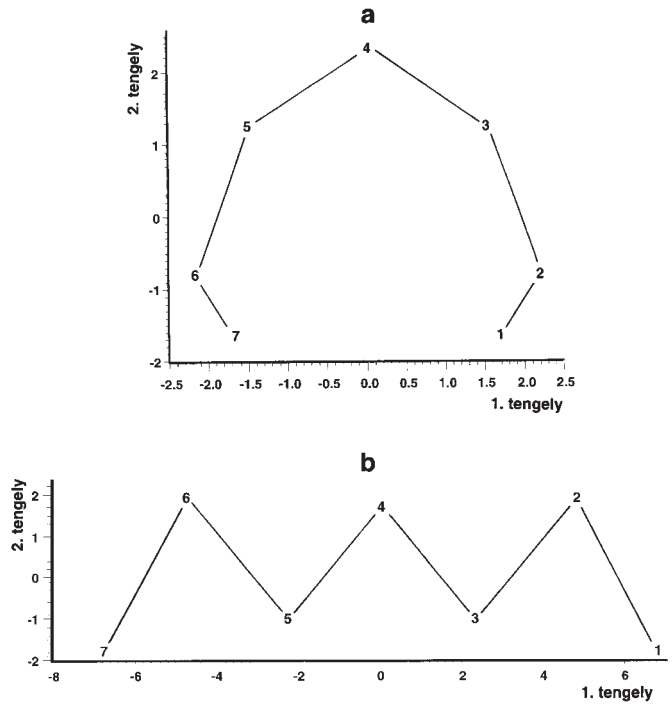
$$100 \times \sum_{k=1}^{t} \lambda_k \,/\, \sum_{k=1}^{m-1} \lambda_k \quad (7.65)$$

százalékban magyarázza meg. Egy két-dimenziós PCoA diagram, amely a teljes négyzetösszeg 20-30 %-át értelmezi csupán, sok esetben félrevezethető lehet bizonyos pontpárok közelségét illetően. Amelyek az első két dimenzióban közel állanak, még nem biztos, hogy az összes dimenzióra nézve is közeliek. Mindezt a minimális feszítőfa (5.4.3 rész) segítségével ellenőrizhetjük legegyszerűbben, amelyre majd a 9. fejezetben is kitérünk.

- Ha egyes sajátértékek negatívak, akkor ez annak a jele, hogy a kiinduló mátrix nem feleltethető meg tökéletesen az euklidészi térben. Néhány, relatíve kicsiny negatív sajátérték voltaképpen még figyelmen kívül hagyható, és a nagy pozitív sajátértékekhez tartozó tengelyek továbbra is interpretálhatók maradnak. Nagy negatív sajátértékek már gondot okoznak, mert ekkor a kiinduló különbözőségi struktúra már csak nagy torzításokkal ábrázolható az euklidészi térben, és ekkor a PCoA eredményét nem szabad elfogadnunk kritika nélkül. Ebben az esetben a nem-metrikus skálázás (következő rész) módszerei jelenthetnek megoldást.

A 7.18 ábra két dimenziója 46,4 ill. 38,3 %-ot magyaráz meg a teljes négyzetösszegből. A 10×10-es távolságmátrix átskálázása a síkba tehát 84,7 %-os "sikerrel" járt, a variancia többi részét az úttávolságok és a légvonalbeli távolságok eltérései okozzák. Mindehhez három pozitív sajátérték tartozik (λ_3=8,8, λ_4=5,8 és λ_5=0,7), így a mátrix rangja 5. Még nagyobb a variancia-sűrítés hatékonysága az immunológiai távolságok A5 mátrixának elemzésében, ugyanis itt λ_1=63,1 ill. λ_2=22,7 (s van még további négy kicsiny pozitív sajátérték). A diagramot nem mutatjuk be, elegendő megjegyezni, hogy az első tengely a "majom" és a többi taxon nagy eltérését magyarázza, míg a második tengelyen a macska különül el a többiektől. A maradék hat faj egy csoportba tömörül az origó körül az 1-2. tengelyeken, különbségeik inkább a hátralévő tengelyeken fejeződnek ki.

A patkó-jelenség és a legrövidebb út szerinti kiigazítás flexibilis módszere. A 7.14 mátrix soraiban lévő objektumokra kiszámított euklidészi távolságmátrixból kapott PCoA eredmény – mint már sejtjük – úgyanúgy mutatja a patkó-jelenséget (7.19a ábra), mint a centrált PCA. Itt válik igazán nyilvánvalóvá az, amit a 7.1.6 részben már említettünk: a patkó alakú elrendezés azért adódik, hogy a pontok közötti távolságok a lehető leghűségesebben tükröződjenek az eredményben. A grádiens mentén haladva, ha az első objektumtól vett távolság eléri a maximumot, akkor tovább lépve ez már nem nőhet tovább. Ez adta az ötletet Williamson (1978) és Clymo (1980) számára, hogy az egyetlen közös fajt sem tartalmazó objektumok közötti távolságokat számoljuk át; pontosabban: növeljük meg, hogy ezáltal a grádiens mentén a távolságok növekedése az eddigiekkel arányosan folytatódjék. Javaslatuk szerint a kérdéses két objektum között egy olyan sorozatot kell keresni a többi objektumból, amelyben a szomszédos objektumok legalább egy faj jelenlétében (de nem mennyiségében) megegyeznek, és a sorozat mentén a páronkénti távolságok összege minimális ("*shortest path*"). Ez a távolságösszeg adja azután a két objektum "kiigazított" v. módosított távolságát.



7.19 ábra. a: A PCoA sem mentes a patkó jelenségtől, ha a kiinduló adatok nem lineárisak (7.14 mátrix). **b:** A távolságok grádiens melletti megnövelése Williamson és Clymo módszerével viszont már közelítőleg felfedi a háttérgrádiens hatását is.

Például, a 7.14 mátrixban az 1. és a 4. objektum (sor) távolsága eléri a maximumot (EU_{14}= 4,69), s ez az eredeti távolság az 1-5, 1-6 és 1-7 párosításokban is. Az 1-4 párra a legrövidebb út az 1-2-4, a 3,16 és 4,47 távolságokkal, melyek összege 7,63, s ez lesz a megnövelt új érték. Hasonló módon kapjuk az *EU*'15=8,94, *EU*'16=12,1 és *EU*'17=13,4 módosított távolságokat. E manipulációk eredményessége a 7.14 mátrixból számolt eredeti és módosított távolság-mátrixok PCoA analízisével értékelhető igazán (7.19 ábra). A módosított mátrix elemzésekor azonban negatív sajátértékek is adódtak, s az összehasonlítás kedvéért megadjuk mindegyiket: 148,37; 18,76; 1,53; 0,13; 0,00; –2,26 és –8,52. Az első sajátérték tehát nagyságrendileg meghaladja a többit, s az objektumok koordinátái az első tengelyen lényegileg jól tükrözik a grádiens melletti elhelyezkedést. Abszolút értékben azonban λ_2 csupán kétszerese λ_7-nek, ezért a második tengely melletti elrendeződés nem interpretálható.

A távolságok megnövelése miatt adódó negatív sajátértékek jelenléte vagy elmaradása további vizsgálódások témája lehetne. További "hátrány" az, hogy a módosításokhoz az eredeti adatokra is szükség van, s ez a PCoA esetében eredetileg nem volt követelmény.[6]

7.4.2 Nem-metrikus többdimenziós skálázás

Ebben a fejezetben eddig olyan módszereket tárgyaltunk, amelyek – közvetlenül vagy közvetve – az adatokban lévő metrikus információ megőrzésével adnak ordinációt. A PCA, COA és PCoA lineáris adatszerkezetet feltételeznek, s a linearitás nem teljesülése többé-kevésbé zavaró lehet a végeredményben. Kisebb elmozdulást a nem-lineáris irányba még minden módszer elvisel (robusztusság), de a jelentős eltérés már a patkó-jelenséggel "terhelt" eredménnyel jár. A nem-metrikus t.d.s. ("*non-metric multidimensional scaling*", NMDS) módszere, első közelítésben legalábbis, minden ilyen kiinduló feltételtől mentes, és bármilyen módon származtatott távolság- vagy különbözőségi mátrixot vizsgálhatunk vele.

A módszer lényege, hogy a távolságértékek közötti *különbségeket* (amelyek a metrikus információ hordozói) teljesen figyelmen kívül hagyjuk, és az értékek *nagyság szerinti sorrendjére* vagyunk csak tekintettel[7]. Célunk az, hogy előre kikötött számú dimenzióban (ez rendszerint és érthető módon 2) úgy rendezzük el az objektumokat reprezentáló pontokat, hogy a közöttük levő távolságok *sorrendisége* a lehető legjobban megközelítse a távolságok eredeti nagyság szerinti sorrendjét (Shepard 1962, Kruskal 1964). A lényeg tehát az, hogy az objektumok közötti távolságok v. különbözőségek (d_{jk}) és a pontok közötti ordinációs távolságok (δ_{jk}) közötti kapcsolat *monoton l*egyen. Miután az eredményt végülis metrikus koordináták formájában kapjuk, talán helyesebb volna a módszert ordinális skálázásnak nevezni (Gordon 1981), de az "ordinális ordináció" elnevezés már igencsak furcsa volna (itt derül ki persze, hogy a biológusok által jobban kedvelt ordináció elnevezés matematikailag nem igazán helytálló, hiszen azon rendszerint többet értünk puszta sorbarendezésnél).

A nem-metrikus t. d. s. általánosan ismert, Kruskal-féle algoritmusa iteratív jellegű. A pontok egy kezdő konfigurációját (amely lehet általunk megadott vagy random) finomítjuk

6 Bradfield & Kenkel (1987) flexibilis eljárása már nemcsak azokat a távolságokat számítja át, amelyekre nem volt közös faj a két összehasonlított objektumban, hanem azokat is, amelyek legfeljebb k fajban egyeznek meg (k értéke változtatható, k=1, 2, vagy 3, stb.). Ez a módszer az extrém magas β-diverzitású (gyors fajcseréjű) grádiensek értékelésében lehet hatékony elsősorban.

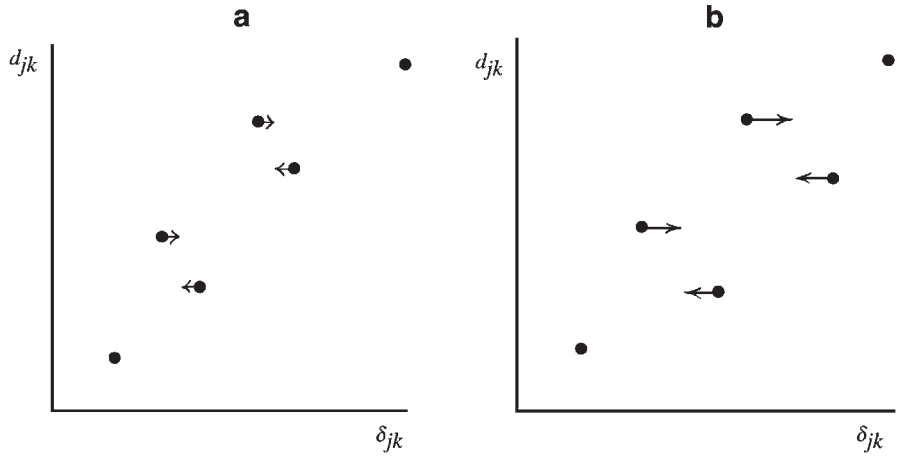
7 Az itt jelentkező információveszteség ahhoz hasonlítható, amikor egy intervallumskálán mért változót átalakítunk ordinális skálájúvá.

számos lépésen keresztül mindaddig, amíg további lényeges javulást már nem érhetünk el. Minden egyes iterációs lépés két részből áll:

1) Az eredeti távolságok (v. különbözőségek) sorrendjét és az ordináció-beli távolságok sorrendjét a legkisebb négyzetek elvén alapuló *monoton regresszió* módszerével vetjük össze (Kruskal 1964). Ez nem jelenti a két sorrend közvetlen összehasonlítását, mint azt először gondolnánk, mert csupán azt nézzük meg, hogy mennyire kell megváltoztatnunk az ordinációban lévő távolságokat ahhoz, hogy az eredeti távolságokkal való monotonitást elérjük. Ez a 7.20 ábra diagramjaiból érthető meg a leginkább. Vegyünk négy objektumot, a közöttük levő távolságok száma tehát 6. A függőleges tengelyen mérjük fel az eredeti távolságértékeket, a vízszintesen pedig az ordinációbeli távolságokat. Az egyes pontok egy-egy objektumpárnak felelnek meg. A két diagram erősen eltérő helyzeteket jelenít meg. A 7.20a ábrán igen kicsiny változtatásokra van szükség ahhoz, hogy a monotonitást elérjük, a b ábrán viszont már jelentékenyebb a differencia. Nyilvánvaló, hogy az első megoldás jobb, mint a második. Mindez persze kvantitatíven is kifejezhető például a Kruskal-féle *stressz*-függvénnyel, amely az ábrán kis nyilakkal jelölt eltérések négyzetével számol:

$$ST = \left[\frac{\sum_{j<k} (\delta_{jk} - \hat{\delta}_{jk})^2}{\sum_{j<k} \delta_{jk}^2} \right]^{1/2} \qquad (7.66)$$

Ez egyszerű euklidészi távolság amelyet a [0,1] intervallumba normálunk (egyébként az eltérésnégyzetnek nem lenne felső határa, így a regressziós eredmények összehasonlítása nehézkesebb lenne). A $\hat{\delta}_{jk}$ jelenti azt az értéket, amelybe a δ_{jk} távolságot módosítani kellene a monotonitás eléréséhez, vagyis $\hat{\delta}_{jk}$ nem egy konkrét távolságnak, hanem kettő vagy több távolságérték átlagának felel meg (a 7.20a és b ábrán is két ilyen átlagérték szerepel). Az ST=0 esetben az ordinációs távolságsorrend tökéletesen illeszkedik az ere-



7.20 ábra. Az eredeti (d_{jk}) és az ordinációs (δ_{jk}) távolságok összehasonlítása Shepard diagrammal, két, lényegesen eltérő szituációban (lásd a szöveget). A nyilak a monotonitás eléréséhez szükséges változások mértékét illusztrálják.

deti távolságok sorrendjére, ezeket tehát nem kell megváltoztatni. (Természetesen az értékek radikálisan különbözőek lehetnek, hiszen a távolságok megváltoztatására elég nagy szabadságunk van anélkül, hogy a sorrendiséget befolyásolnánk). A stressz-függvény tájékoztat bennünket arról, hogy az ordináció mennyire hatékonyan ábrázolja a távolságok sorrendi viszonyait.

2. Ha két iterációs lépés között *ST* csökkenése kisebb mint egy ε küszöbérték (pl. 0,001), akkor leállhatunk az elemzéssel, és az utoljára kapott konfigurációt véglegesnek tekintjük. Egyéb esetben folytatjuk az iterációt a pontok elcsúsztatásával olymódon, hogy *ST* értéke tovább csökkenthető legyen. Ezt a legmeredekebb lejtő ("*steepest descent")* módszerrel érjük el (Kruskal 1964, ill. Brambilla & Salzano 1981 írja le részletesen az algoritmust). Ez lényegében véve a stressz minden egyes koordináta szerinti parciális deriváltjának kiszámításán alapszik, és azt az "irányt" jelöli ki, amelyben a koordináták megváltoztatása *ST* maximális csökkenésére vezet.

Az új konfiguráció meghatározását követően megkapjuk *ST* új értékét, és így tovább. Ha a változás már jelentéktelen, a végső konfigurációt úgy határozzuk meg, hogy súlypontja 0 legyen és a súlyponttól vett távolságok négyzetösszege pedig 1. Ez csupán jótanács, ennek révén ugyanis könnyebb az összehasonlítás más eredményekkel. A konfiguráció egyébként elforgatható vagy bármilyen önkényes szorzóval átskálázható, a NMDS eredményben erre vonatkozó kitüntetettség nincs, ellentétben az előző módszerekkel.

A végső ordinációs diagramon túlmenően az NMDS eredmények bemutatásához szervesen hozzátartozik az eredeti és az ordinációs távolságok grafikus összehasonlítása, az ún. Shepard-diagram, amire a 7.20 ábra már például szolgált. Ezen a diagramon annál jobban szórnak a pontok, minél rosszabb az ordinációs sorrend az eredetihez képest, egyébként viszont átlós irányban tömörülnek (lásd még a 7.21b ábrát).

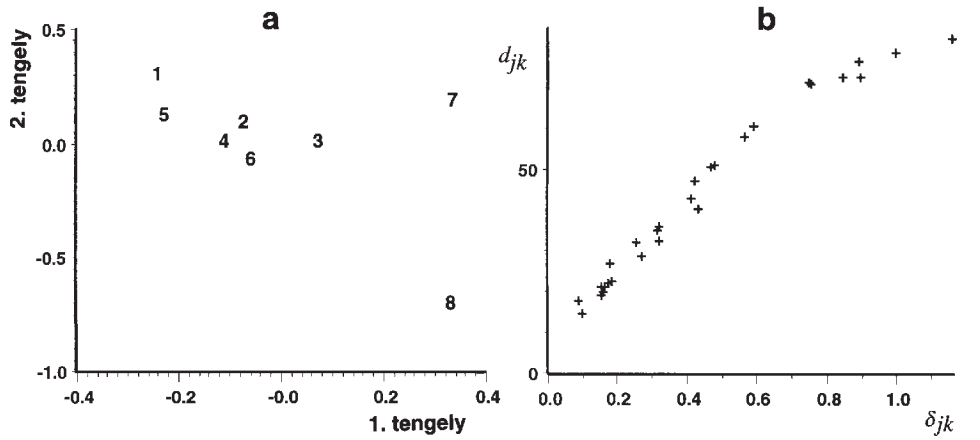
Az NMDS végrehajtásában és az eredmények értékelésében hasznos szempontok a következők:

- A kiinduló mátrixban bármilyen különbözőségek lehetnek, a metrikus axiómákat akár drasztikusan is megsérthetik. A módszer úgy is programozható, hogy hiányzó értékeket is megengedünk. Amikor a PCoA nagy negatív sajátértékeket produkál, s így az eredményt kétségessé teszi, a NMDS marad az egyetlen használható ordinációs módszer. Egy nagyszabású összehasonlító vizsgálatban Kenkel & Orlóci (1986) azt találták, hogy az NMDS + hurtávolság relatíve hatékony volt a két-dimenziós háttérgrádiensek (="*coenoplane*") feltárásában. Minchin (1987) hasonlóképpen ezt a módszert találta a leghatékonyabbnak. Az NMDS alkalmazhatóságát illetően mások a korlátokat hangsúlyozzák (Gauch et al. 1981, Digby & Kempton 1987), a következő pontokban felsoroltakkal érvelve.
- A dimenziók számát a felhasználó előre adja meg (e tekintetben tehát a faktoranalízisre emlékeztet, hiszen ott a faktorok számát magunknak kell megadnunk). Kiindulásul mindjárt kérhetünk 2 dimenziót, hiszen ez a dimenzionalitás az, amit a papír síkjában is ábrázolhatunk. Ha pl. 4-et választunk kezdetként, akkor a négy-dimenziós megoldásból egy három-dimenziósat, majd abból egy két-dimenziósat lehet

előállítani. Ellentétben azonban a metrikus ordináció módszereivel, a kétdimenziós megoldás nem egyszerűen a harmadik dimenzió elhagyásából származik! Az *ST* és a dimenziók száma közötti összefüggés grafikusan is illusztrálható. *ST* nyilván csökken a dimenziószám növelésével, s ahol nem tapasztalunk hirtelen csökkenést, ott akár meg is állhatunk. Általános szabály azonban nincs az optimális dimenziószám maghatározására.

- A stressz értékek nagyságával kapcsolatban sem adható általános szabály. Az *ST*=0,05-ös értéket már rendszerint nagyon jónak tarthatjuk, bár megítélésünk a pontok számától és a dimenziók számától nagymértékben függhet (abszolút érvényű kritériumaink nincsenek) és a 0,1 – 0,2 közötti értékek is általában még elfogadhatók.
- Csakúgy, mint sok más iterációs módszernél (lásd még a a 8. fejezetet; viszont kivétel a reciprok átlagolás) az elemzés végeredménye függ a kiinduló konfigurációtól. Az analízis tehát nem feltétlenül konvergál ugyanabba a végeredménybe, akár eltalálhatja az abszolút legjobbat (*globális* optimum) de relatíve egészen rossz *lokális* optimumokban is "megrekedhet". Ezt a problémát úgy kerülhetjük meg, hogy az elemzést random konfigurációkból többször is végrehajtjuk, s a legkisebb stresszt adó ordinációt fogadjuk el végeredményül (Shepard 1980). A metrikus módszerekkel kapott kétdimenziós ordinációk is kiváló kezdőpontnak bizonyulhatnak a globális optimum megkeresésére vagy legalább a megközelítésére.
- Másik gyakorlati kérdés a kiinduló távolságértékek nagyságára vonatkozik. Kruskal (1977) beszámolt egy vizsgálatról, amelyben a távolságokat három csoportba (nagy, közepes és kicsiny) osztották, s megvizsgálták, hogy az egyes csoportok elhanyagolása milyen mértékben befolyásolja a végeredményt. Kiderült, hogy a kis távolságok kevéssé hatnak a végeredményre, a legnagyobb távolságok elhagyása viszont nagymértékben megváltoztatta az ordinációt. A kis távolságok megítélésével kapcsolatos esetleges bizonytalanságaink tehát – úgy tűnik – nem járnak komolyabb következményekkel.
- Bár a linearitást, mint feltételt, nem mondtuk ki, nagy β-diverzitású ökológiai grádiensek esetén a távolságsorrendek megőrzése is csak a patkó-jelenség eltűrésével lehetséges. Az előző pontból következően tehát a nagy távolságok átalakítása a legközelebbi út módszerével erőteljes "javulást" eredményezhet (pl. Podani 1994).
- A tengelyek értelmezésekor ne feledjük, hogy – ellentétben a PCA és PCoA tengelyekkel – közöttük a lineáris korreláció nem feltétlenül 0! Az NMDS ordinációkban az irányok önkényesek, a teljes konfiguráció elforgatható (néhány szerző ezért a PCA alkalmazását javasolja az NMDS koordinátákra). Emiatt a tengelyek azonosítása biológiai vagy egyéb külső változók segítségével problematikusabb, mint az indirekt lineáris módszerek esetében volt.

Az A1 táblázat cönológiai felvételeinek NMDS elemzését több, különböző random kiindulásból is elvégeztük, két dimenzióra, a felvételek közötti euklidészi távolságok mátrixából. (Az tehát "nem akadály", hogy a nyers adatok is megvannak, bár a többdimenziós skálázást tipikusan olyan esetekre alkalmazzuk, amikor csak a különbözőségek ismeretesek.) Kiválasztottuk a legjobb eredményt, azaz a legkisebb stresszt adó konfigurációt (7.21a ábra). *ST* értéke 0,006, ami arra utal, hogy a kétdimenziós ábrázolás csaknem teljes mértékben megőrzi



7.21 ábra. Cönológiai felvételek (A1 táblázat) NMDS ordinációja (**a**) és a hozzátartozó Shepard-diagram (**b**).

a távolságok sorrendjét, s mindezt a Shepard-diagram pontjainak megközelítőleg átlós elhelyezkedése is igazolja (7.21b). A pontok elrendeződésében – az ezzel az eredménnyel logikusan összevethető – metrikus ordinációhoz[8] (centrált PCA, 7.2 ábra) képest az a legfeltűnőbb különbség, hogy a távolságok kiegyenlítődnek: a nagy távolságok kissé csökkennek, a kicsik pedig megnövekszenek. A pontok tehát jóval egyenletesebben szóródnak, mint a metrikus ordinációkban, s ez általános jellemzője a NMDS eredményeknek. Ennek nyilvánvaló oka az, hogy a NMDS-ben csak a sorrendiség számít, az abszolút eltérések nem. Egyéb tekintetben az NMDS eredmény nem mond újat a PCA-hoz képest, viszont egy erősen eltérő kritérium figyelembevételével *megerősíti* a korábban tapasztaltakat.

A nem-metrikus többdimenziós skálázás más módszerei. Az NMDS Kruskal-féle algoritmusa voltaképpen csak egy lehetőség az objektumok nem-metrikus módon történő elrendezésére. Egy jól ismert módosítás (Sibson 1972) az ún. *lokális* NMDS (rövidítve: LNMDS), amelyben nem törekszünk arra, hogy a teljes távolságsorrend megmaradjon. Eme szigorú feltétel helyett megelégszünk azzal, hogy minden egyes objektumnak az összes többivel adott eredeti távolságainak (különbözőségeinek) sorrendjét próbáljuk meg maximálisan megőrizni az ordinációban. A d_{jk}-ra nézve például nem lényeges, hogy hogyan viszonyul a d_{lm}-hez, de a d_{jm} és d_{jl} távolságokhoz képest felvett helye a sorrendben már fontos. Az LNMDS minden egyes pont saját környezetét nézi (innen a "lokális" elnevezés) s emiatt – Prentice (1977) véleménye szerint – alkalmasabb lehet ökológiai háttérgrádiensek értékelésére, mint az eredeti módszer (hiszen a környezet maga is változhat a grádiens mentén: egy adott távolságkülönbség a grádiens elején nem biztos, hogy olyan fontos marad a végén is). Az LNMDS esetében a Shepard-diagram persze elveszti értelmét.

A távolságok sorrendjének reprodukálása csak az egyik lehetőség nem-metrikus ordinációra. A kontinuitás-elemzés ("*continuity analysis*" vagy "*parametric mapping*", Shepard & Carroll 1966, Noy-Meir 1974) minimális számú új dimenziót keres, amellyel a változók (pl.

8 A standardizált PCA eredménye nyilván nem komparábilis a nyers adatokon alapuló NMDS eredménnyel, hiszen nemcsak a módszer, hanem az alapadatok is eltérnek. Ha pedig két vagy több dolgot is egyidejűleg megváltoztatunk, akkor az összehasonlításból nem derülhet ki, hogy mi okozza az esetleges különbségeket az eredményben. Minderre a 9. fejezetben részletesen visszatérünk majd ("komplex összehasonlítások").

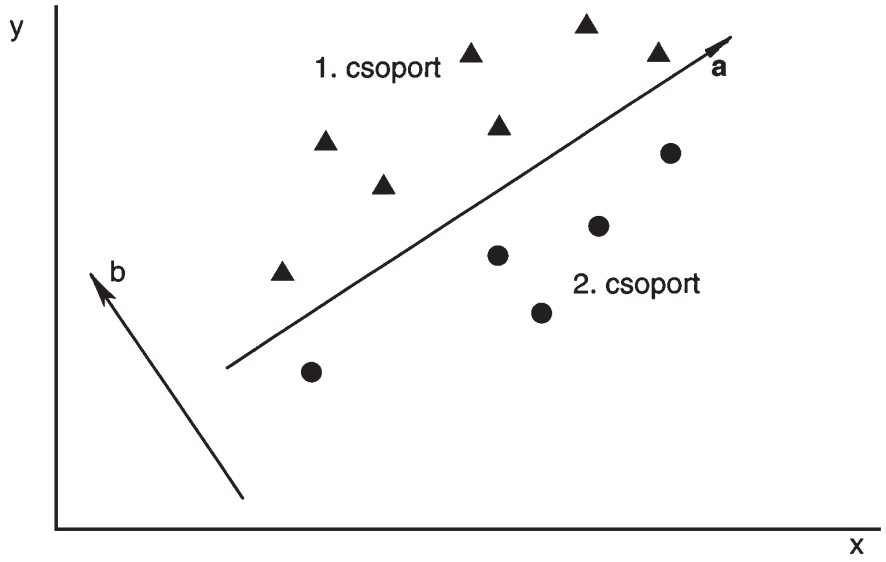
fajok) olyan függvénykapcsolatot mutatnak, hogy a pontok illeszkedése a lehető legjobb legyen (a függvények alaptulajdonságaira nincs semmiféle kikötés). A stressz helyett ebben az esetben az függvényértékektől való eltéréseket minimalizáljuk. A kontinuitás itt valójában a függvényekkel számolt sokdimenziós felületre történő illeszkedés jóságát ("*smoothness*") jelenti.

7.5 Csoportok elkülönítõ ordinációja: a diszkriminancia elemzés

Az eddigi ordinációs módszereknél az objektumokat egységes csoportként kezeltük, s legfeljebb – a CCA és a CCOA esetében – a változóknál tettünk két csoport között különbséget. Az is lehetséges azonban, hogy kiinduláskor az *objektumok*nak létezik egy – valamilyen külső szempontrendszer szerint létrehozott, *a priori* – osztályozása $k \geq 2$ csoportba. Ekkor felmerülhet a következő ordinációs probléma: találjunk lineárisan korrelálatlan új tengelyeket olymódon, hogy ezek a lehető legjobban megmagyarázzák a csoportok közötti különbségeket és ne törődjenek a csoporton belüli tendenciákkal. Vagyis az ordinációnak maximális hatékonysággal kell a *csoportok elválását* feltárnia. A főkomponens-elemzéssel szemben tehát nem a teljes varianciát, hanem a csoportok közötti varianciát maximalizáljuk, s időközben a csoportokon belüli varianciát pedig minimalizálni igyekszünk (Mardia et al. 1979). Ha ugyanazt a ponthalmazt mindkét módon is elemezzük, akkor a csoportosítástól függően egészen eltérő lehet az új tengelyek helyzete, amint azt a 7.22 ábra két szélsőséges esete illusztrálja. A csoportok elválását maximalizáló ordinációs módszer a *diszkriminancia-elemzés* vagy "*canonical variate analysis*" (CVA). Az angolszász irodalomban a CVA rövidítés egyértelműen az ordinációs célú, azaz a dimenzió-redukciót előtérbe helyező alkalmazásokra korlátozódik. A "linear discriminant function analysis" (LDFA) vagy "multigroup discriminant analysis" (MDA) elnevezések pedig – bár lényegileg ugyanarról a módszerről van szó – inkább arra utalnak, amikor az elemzés célja a legjobban diszkrimináló változók megkeresése, valamint új objektumok besorolása már létező osztályok valamelyikébe. Ez utóbbi esetben azonban nem is cél az ordinációs diagram elkészítése. Jelen könyvben a diszkriminancia-analízis elnevezést és a CVA rövidítést használjuk, megjegyezve, hogy elsősorban az ordinációs funkciót emeljük ki. A CVA egyébként szoros kapcsolatban áll a többváltozós variancia-elemzéssel (MANOVA) is, amelyet e könyvben külön nem tárgyalunk.

A CVA a változók közötti keresztszorzat mátrixokon alapszik (a 3.68 függvény centrált adatok alapján, vagy a 3.69 függvény az m–1-el való osztás nélkül), ezeket *diszperziós* mátrixoknak nevezzük. Mindegyiknek $n \times n$ a mérete. A **T** mátrixot a teljes objektumhalmazra számítjuk ki, függetlenül attól, hogy az egyes objektumok melyik csoportba tartoznak (*teljes diszperziós mátrix*). A $\mathbf{W}_1$, $\mathbf{W}_2$,...,$\mathbf{W}_k$ mátrixokat külön-külön határozzuk meg a k csoport mindegyikére (centrálás a csoportra vonatkozó átlag szerint!). Ezek az ún. *csoporton belüli diszperziós mátrixok*, amelyek összeadása, $\mathbf{W} = \sum_{i=1}^{k} \mathbf{W}_i$ eredményezi az *egyesített (közös) csoporton-belüli diszperziós mátrixot*. A keresztszorzatoknak az a része, amely a csoportok közötti eltéréseket értelmezi egyszerűen megkapható e két mátrix különbségeként:

$$\mathbf{A} = \mathbf{T} - \mathbf{W} \qquad (7.67)$$



7.22 ábra. A főkomponens-elemzés és a diszkriminancia analízis összehasonlítása egy mesterséges ponthalmaz felhasználásával. Az 1. főkomponens (a) a teljes adathalmaz varianciájának fő irányával esik egybe, míg a kanonikus változó (csak egy van, b) a két csoport közötti legjobb elválasztás irányát adja meg.

Ekkor – a biometriából már ismert variancia-analízisre visszagondolva – azt ajánlhatnánk, hogy a csoportok közötti és a csoportokon belüli diszperziónak valamiféle hányadosát kellene maximalizálnunk. Mivel mátrixokat osztani nem tudunk egymással (a mátrixalgebrában nem létezik az **A**/**W** művelet), az **A** mátrixot megszorozzuk a **W** mátrix inverzével, hogy az arányosságot matematikailag is kifejezhessük. Ezt a mátrixot kell sajátértékelemzésnek alávetni az alábbi egyenlet figyelembevételével:

$$(\mathbf{W}^{-1}\mathbf{A} - \lambda\mathbf{I})\,\mathbf{v} = \mathbf{0} \qquad (7.68)$$

A főkomponens-elemzéssel fennálló analógia most válik csak nyilvánvalóvá, bár – mint említettük – az új tengelyek a maximális szeparálódás és nem a maximális variancia irányába mutatnak. További eltérés a PCA-tól az, hogy az **A** és **W** mátrixok szimmetrikusak ugyan, de a $\mathbf{W}^{-1}\mathbf{A}$ szorzatmátrix nem az, s ezért az eredményül kapott sajátvektorok *nem lesznek ortogonálisak* egymásra, holott lineárisan korrelálatlanok. Ez azt jelenti, hogy az eredményeket nem célszerű egy derékszögű koordináta-rendszerben feltüntetni. Az esetleges torzítás hatása azonban megfelelő transzformációval (7.70) csökkenthető.

A $\mathbf{v}_j$ sajátvektorokat a megszokott módon határozzuk meg: mindegyiket egységnyi hosszúságúra normáljuk. Az objektumok koordinátáit az új koordináta-rendszerben (a tengelyeket kanonikus tengelyeknek nevezzük) azonban nem kapjuk meg közvetlenül a sajátvektorokból. Először a $\mathbf{c}_j$ *diszkriminánsúlyokat* vagy *kanonikus változókat* kell meghatároznunk. A CVA első népszerűsítői (Cooley & Lohnes 1971) a következő átalakítást javasolták:

$$\mathbf{c}_j = \frac{\mathbf{v}_j}{\left(\mathbf{v'}_j \frac{\mathbf{T}}{m-1} \mathbf{v}_j\right)^{1/2}} \tag{7.69}$$

amelyben m az objektumok száma (mint eddig), és $\mathbf{T}/(m–1)$ a teljes variancia/kovariancia mátrix (a képletben az osztó egy skalármennyiség!). Ennek révén minden kanonikus tengelyen egységnyi lesz az összvariancia, tehát a csoporton belüli szóródás egyenlőtlenül fejeződik ki a tengelyeken (7.23a ábra). Következésképpen, a csoportok szórása aránytalanul elnyújtottá válik a kanonikus térben, s a kevésbé fontos tengelyek túlhangsúlyozódnak. Ha azonban a sajátvektorokat az egyesített csoporton-belüli variancia/kovariancia mátrix [$\mathbf{W}/(m–k)$] felhasználásával normáljuk:

$$\mathbf{c}_j = \frac{\mathbf{v}_j}{\left(\mathbf{v'}_j \frac{\mathbf{W}}{m-k} \mathbf{v}_j\right)^{1/2}} \tag{7.70}$$

(Mardia et al. 1979), akkor a varianciák már nem egyenlően oszlanak meg a kanonikus tengelyeken, és a csoportok közötti különbségek sokkal inkább kifejeződnek. Továbbá, minden egyes tengely hozzájárulása a csoporton-*belüli* varianciához azonos lesz, s a csoportok szórásképe – két dimenzióban – nagyjából kör alakúvá válik (7.23b ábra, sok dimenzióban pedig hipergömbről beszélhetünk). Ez a – "*spherizing*"-nek nevezett – átalakítás mindenképpen jó választásnak tűnik, hiszen a kanonikus változóktól nem a csoporton belüli, hanem a csoportok közötti diszperzió magyarázását várjuk. Az ortogonalitás hiánya kör alakú szórásképen esetén jóval kisebb torzítást jelent, mint az elnyújtott pontfelhők esetében.

Miután a kanonikus változókat meghatároztuk, az s-edik objektumnak a j tengelyen felvett koordinátáját a centrált adatok felhasználásával kapjuk meg az alábbiak szerint:

$$e_{js} = \sum_{i=1}^{n} c_{ij}(x_{is} - \bar{x}_i), \tag{7.71}$$

amelyben $\bar{x}_i$ az i változóra vonatkozó átlag a *teljes* adatmátrixban. A centrálás eredményeképpen az ordinációs pozíciók súlypontja a kanonikus tengelyek metszéspontjába, azaz az origóba kerül.

A diszkriminancia-elemzés értékelése során a következőket érdemes szem előtt tartani:

A kapott sajátértékek eleget tesznek a következő összefüggésnek:

$\lambda_j = (\mathbf{v}'_j \mathbf{A} \mathbf{v}_j) / (\mathbf{v}'_j \mathbf{W} \mathbf{v}_j)$ (7.72)

A sajátértékek nagyságát nem befolyásolja a kanonikus változók 7.69-70 szerinti átskálázása. Az alábbi arány

$$\frac{\lambda_j}{\mathrm{tr}(\mathbf{W}^{-1}\mathbf{A})} = \frac{\lambda_j}{\sum_{i=1}^{q} \lambda_i} = \frac{\lambda_j}{T^2} \tag{7.73}$$

kifejezi, hogy a csoportok közötti variancia hányad része esik a j kanonikus változóra (Mardia et al. 1979). q a kanonikus változók száma (lásd a következő bekezdést). A

nevező, vagyis a sajátértékek összege (s egyben a szorzatmátrix nyoma), a Hotelling-féle T^2 néven ismert a szakirodalomban. Ez a mennyiség a csoportok közötti különbség statisztikai tesztelésére használható (mi azonban a Bartlett-féle próbával ismerkedünk meg, 7.74 formula). A Hotelling-féle T^2 jelentése az általánosított (Mahalanobis-) távolság (3.96 egyenlet) felidézésével tovább finomítható. A T^2 ugyanis a csoportok súlypontjai és a teljes objektumhalmaz főátlaga közötti általánosított távolságok súlyozott középértéke (a nagyobb csoport távolsága ezáltal erőteljesebben jut érvényre, mint a kisebb létszámú csoportoké).

- A lineárisan korrelálatlan (de nem feltétlenül ortogonális) kanonikus tengelyek száma $q = \min\{k–1, n\}$. Vagyis, amikor a csoportok száma kisebb, mint a változóké, k–1 kanonikus tengely elegendő a k csoport közötti kapcsolatok maradéktalan ábrázolásához. Egy kétdimenziós "szokványos" ordinációhoz tehát minimum három objektumcsoport kell. Nyilvánvaló az is, hogy q nem lehet nagyobb az eredeti változók számánál. Megjegyzendő azonban, hogy a változók száma nem haladhatja meg az objektumokét, mert akkor a **W** mátrix nem invertálható (C függelék). A módszer annál hatékonyabb, minél nagyobb az objektumok száma a változók számához képest.
- Ha két szigorú feltétel, a változók csoporton-belüli többváltozós normális eloszlása és a csoporton-belüli varianciák/kovarianciák homogenitása (vagyis azonossága, a sztochasztikus ingadozást persze megengedve) teljesül, akkor az ordináción túlmenően a kanonikus változók *szignifikancia*-próbája is lehetővé válik. A Bartlett-teszt (Cooley & Lohnes 1971) alkalmas annak eldöntésére, hogy a p legfontosabb kanonikus változó eltávolítása után maradó $q–p$ számú változó szignifikánsan hozzájárul-e a csoportok elválásához. A statisztikát a következőképpen számítjuk ki:

$$X^2 = -\left(m - 1 - \frac{n+k}{2}\right)\ln\left(\prod_{j=p+1}^{q}\frac{1}{1+\lambda_j}\right) \qquad (7.74)$$

mely a χ^2 -eloszlást követi $(n – p)$ $(k – p – 1)$ szabadsági fok mellett. Ha tehát arra vagyunk kíváncsiak, hogy az 1. változó szignifikáns-e, akkor a fenti statisztikát p=0 mellett számítjuk ki, s a kapott értéket a táblázatos χ^2 küszöbértékkel összehasonlítjuk. Ha a statisztika meghaladja azt, akkor a változó szignifikáns eltérést jelez a csoportok között. Ha nem haladja meg a küszöbértéket, akkor természetesen a többi tengely sem lesz szignifikáns és nincs értelme tovább keresgélni.

- A fenti formulában "benne van" a Wilks-féle Λ vagy más néven, a *determináns-hányados:*

$$\Lambda = \prod_{j=1}^{q}\frac{1}{1+\lambda_j} = \frac{|\mathbf{W}|}{|\mathbf{T}|} \qquad (7.75)$$

amelyet "mellékesen" amúgy is kiszámítunk a Bartlett próba során. Λ értéke 0 (=a csoport centroidok maximális elválása) és 1 (=a csoportok centroidjai statisztikailag megkülönböztethetetlenek) közé esik.

Felmerülhet az Olvasóban a kérdés, hogy kis Λ, ill. szignifikáns Bartlett-teszt esetén miképpen mutatható ki: voltaképpen mely csoportok térnek el szignifikánsan egymástól s melyek nem? A helyzet hasonló az egyváltozós variancia-analízis utáni vizsgálódáshoz, amikor minden csoportot párosával értékelünk, hogy kikeressük a nagy különbségeket (szimultán összehasonlítások). Ebben a kötetben nem célunk, hogy a statisztikai hipotézis-vizsgálatok témakörét részletesebben megvizsgáljuk, ezért csak megemlítjük: a szignifikánsan eltérő csoport-párok kiválasztása sok elméleti és gyakorlati nehézséggel jár. Ugyanazt a módszert, amit a k=2 esetre nyugodtan alkalmazhatunk (mint pl. a T^2-n alapuló F-próbát, Sváb 1979: 126), már nem használhatjuk minden párosításban a k>2 esetben, mert ez az "elsőfajú" hiba halmozódásához, s ezáltal téves következtetések levonásához vezet. A próbát tehát szigorítani kell, amelyre sokféle lehetőség kínálkozik. A szimultán összehasonlítások problémakörét Kun (1986) összefoglalójából ismerhetjük meg igazi "mélységeiben", de szó lesz még róla a 9. fejezetben is, egészen más kontextusban.

- Vezessünk be k számú bináris csoportbatartozási indikátorváltozót, amelyre g_{hi}=1, ha a h objektum az i csoportba tartozik, ill. g_{hi}=0, ha máshova. Ennek alapján belátható, hogy a CVA *a kanonikus korreláció-elemzés* (CCA) *speciális esete:* az egyik (mondjuk a baloldali) változócsoportot az eredetiek, a másikat (a jobboldalit) pedig eme új indikátorváltozók alkotják (Bartlett 1938, lásd még Cooley & Lohnes 1971: 249, ter Braak & Prentice 1988). A kanonikus változók tehát az eredeti változók olyan lineáris kombinációi, amelyek maximálisan korrelálnak az indikátorváltozók lineáris kombinációival. A CVA-ból adódó j-edik kanonikus korreláció a következő:
  $$R_j = (\lambda_j / (1 + \lambda_j))^{1/2} \qquad (7.76)$$
  Ez abszolút értékben megegyezik a megfelelő módon elvégzett CCA-ból származó kanonikus korrelációval (7.23 függvény). Minél jobban elválasztja egymástól a kanonikus tengely a csoportokat, annál magasabb a 7.76 koefficiens értéke.
- Ha **R** az n változó korrelációs mátrixa, akkor a j kanonikus változó és az eredeti változók közötti korrelációk ("*structure coefficients*" vagy "*loadings*") a következőképpen kaphatók meg:
  $$\mathbf{s}_j = \mathbf{R}\mathbf{c}_j \qquad (7.77)$$
  amelyben $\mathbf{c}_j$ a 7.69 formula szerint veendő figyelembe. Ezek a korrelációk egyébként függetlenek a sajátvektorok 7.69 vagy 7.70 normalizálásától. (A fenti egyenlet azonban érvényét veszti, ha $\mathbf{c}_j$-t a 7.70 formulával számoljuk). Az alternatív CCA elemzésben ugyanezek a korrelációk az eredeti (baloldali) változócsoport tagjainak a saját kanonikus változójukkal vett csoporton belüli korrelációiként (7.26 egyenlet) adódnak. Ebből szinte azonnal következik, hogy a CVA ordinációban a pontok relatív elhelyezkedése megegyezik a baloldali változócsoportra kapott kanonikus változók szerinti CCA ordinációval.
  A 7.77 korrelációkat felhasználhatjuk a csoportok között legjobban diszkrimináló eredeti változók kiválogatására, amelynek – talán nem kell mondanunk – rendkívül nagy interpretatív értéke lehet. A CVA ordinációs koordinátákból és a korrelációkból biplot is szerkeszthető, s ez grafikus módon szemlélteti a változók és a tengelyek kapcsolatát. A biplot koordináták lehetnek önkényesek (a két ordináció egymásra vetítéséből), s ekkor csak az irányoknak és a relatív hosszúságoknak van jelentőségük. Dillon & Goldstein (1984) javaslata szerint egyébként a korrelációkat a megfelelő

egyváltozós *F*-értékekkel kell megszorozni, ilymódon a változók közötti különbségek még jobban kifejeződnek.

A CVA-t végrehajtó számítógépes programok az eredmények listáját gyakran az egyváltozós *F*-hányadosok felsorolásával kezdik. Az F_i érték a csoportok közötti és a csoporton belüli variancia hányadosa az *i* változóra, s ennek nagysága már az elemzés legelején tájékoztat bennünket a csoportok közötti különbséget leginkább magyarázó változók kilétéről. A nagy F_i-t adó változók lesznek elsősorban azok, amelyek magas korrelációt adnak a kanonikus változókkal.

- Megvizsgálhatjuk azt is, hogy a változók korrelációs mátrixában lévő variancia (tr {**R**} = *n*) hányad részét magyarázza meg a *j* kanonikus változó:

$$100 \times \frac{\sum_{i=1}^{n} s_{ij}^2}{n} \tag{7.78}$$

Ha $(k-1) \times n$, ami ritkán fordul elő, akkor tr {**R**}-t teljes mértékben megmagyarázzák a kanonikus változók, és a kumulatív százalékok összege 100 lesz. Egyéb esetben, azaz általában, a kumulatív százalékok összege 100 alatt marad.

- Az *i* változó *kommunalitását* a következő egyenlet fejezi ki:

$$h_i = \sum_{j=1}^{q} s_{ij}^2 \tag{7.79}$$

Ennek csak akkor van interpretatív értéke, ha $(k-1) < n$. Az alacsony kommunalitású változók varianciáját a kanonikus változók összessége sem magyarázza meg, így ezeknek a változóknak nincs jelentőségük a csoportok elválasztásában. Az 1-hez közeli kommunalitású változók viszont igen fontosak ebből a szempontból. A $(k-1) \times n$ esetben minden kommunalitás 1, így nincs interpretatív értékük.

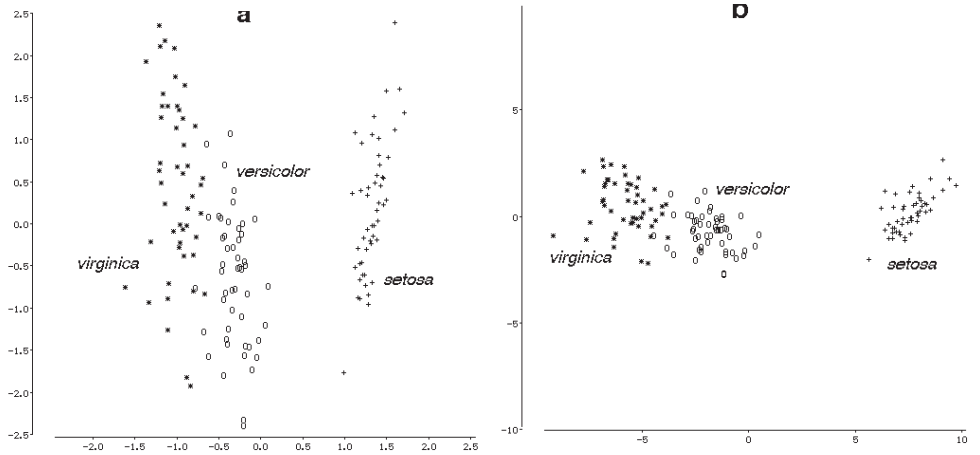
- A CVA ordinációs térben célszerű a csoportok *súlypontjait* (centroidjait) is feltüntetni. Ha a koordinátákat a 7.69 egyenlet szerinti normálással kaptuk meg, vagyis a csoportok szórása hipergömbszerű, akkor minden súlypont körül felrajzolhatjuk a csoportok *izodenzitási körét* (térben: gömböt), melynek sugara $r = \sqrt{(4\chi^2_{2,\alpha})/2}$ (Giri 1977, lásd még Dillon & Goldstein 1984). Az α=0,05 esetben, ami a biológiában általánosan alkalmazott 95 %-os valószínűségi szintnek felel meg, a sugár éppen 2,45 egység. Ez a kör *várhatóan* a csoport – mint statisztikai populáció – objektumainak 95 %-át tartalmazza. A sugár nem függ a csoportok elemszámától, így minden csoportra azonos kört kapunk. Van azonban még egy kör, ami a centroidok köré berajzolható az ordinációs diagramon, s ez már változó sugarú lesz. Ez a *konfidencia-kör*, amely a csoport – mint statisztikai populáció – *várható értékét* $100(1-\alpha)$ %-os valószínűséggel tartalmazza. Ennek sugara $r = \sqrt{(4\chi^2_{2,\alpha}/m_i)}$, ahol m_i az *i*-edik csoport elemszáma (Mardia et al. 1979). Itt is a 95 %-os valószínűségi szintet alkalmazzuk a leggyakrabban. Szinte mondanunk sem kell, hogy mindkét körnek csak akkor van értelme, ha a

7.2 táblázat. Az *Iris* adatok (A2 táblázat) CVA elemzésének összefoglalása. A kanonikus változók a 7.70 szerinti normálásra vonatkoznak, a táblázat többi értékét a normálás módja nem befolyásolja.

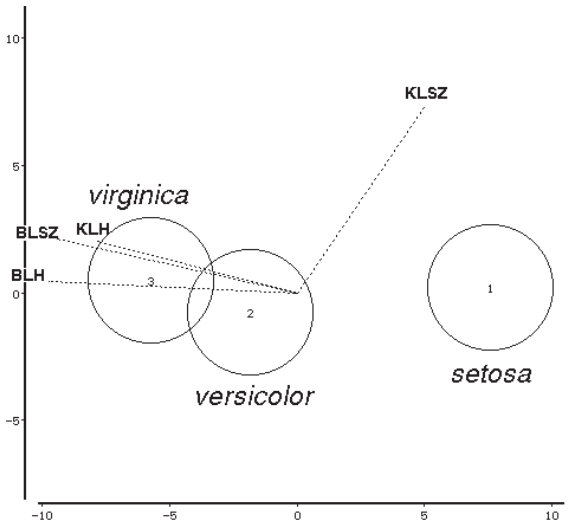
Változó	*F* -hányados	Korreláció az 1. kanonikus változóval	Korreláció a 2. kanonikus változóval	1. kanonikus változó	2. kanonikus változó	Kommunalitás
KLH	118	–0,791	0,206	0,723	–0,107	0,668
KLSZ	48	0,521	0,765	0,157	0,224	0,857
BLH	1180	–0,985	0,046	–0,212	–0,834	0,973
BLSZ	960	–0,973	0,221	–0,285	–0,274	0,996
Kanonikus korreláció				0,985	0,475	
Sajátérték				31,83	0,29	
Külső variancia %				99,09	0,91	
Részesedés a korrelációkból				70,36	16,97	

többváltozós normalitás, a varianciák és kovarianciák homogenitása és a random mintavétel feltételei teljesülnek, egyébként csak félrevezetnek bennünket.

A CVA legcélszerűbben – és leghagyományosabb módon – az *Iris* adatok segítségével illusztrálható, hiszen ez az adathalmaz volt az, amit maga Fischer (1936) használt eredetileg a diszkriminancia elemzés bemutatására. A mátrixban eleve adott három csoport, az *Iris* fajok, és a CVA segítségével megnézhetjük, hogy milyen mértékű elválás mutatkozik közöttük. A 7.69 és 7.70 normálások közötti különbség szembetűnő a 7.23a és b ábrákon. (Persze csak akkor igaz ez, ha a tengelyeken azonos a skála, mert egyébként a pontok relatív helyzete nem változik.) A skálázástól függetlenül jól látszik az *Iris setosa* elkülönülése a másik két fajtól, s e tekintetben megerősítettük a PCA (7.7 ábra) és a fuzzy osztályozás (4.9 ábra) eredményeit. A másik két faj azonban jobban elkülönül egymástól az első tengelyen, mint a PCA esetében bármely tengelyen. Az első kanonikus korreláció ennek megfelelően igen magas, és a csoportok közötti variancia csaknem teljesen megmagyarázódik ezen a tengelyen (7.2 táblázat). A máso-



7.23 ábra. A három *Iris faj* (A2 táblázat) diszkriminancia-elemzéséből kapott ordinációs diagramok kétféle normálás szerint. **a:** normálás a teljes diszperziós mátrix szerint (7.69 egyenlet); **b**: normálás az egyesített csoporton-belüli diszperziós (variancia/kovariancia) mátrixszal (7.70 egyenlet). Jelek: + *Iris setosa,* O: *Iris versicolor*, *: *Iris virginica.*



7.24 ábra. Az *Iris* adatok CVA biplotja. A három faj centroidjai: 1: *I. setosa*, 2: *I. versicolor*, 3: *I. virginica*. A centroidok körül a két kanonikus változóra kiszámítható izodenzitási körök láthatók. A változók diszkriminatív erőssége felmérhető, ha a változókra mutató nyilakat képzeletben meghosszabbítjuk, s az így kapott egyenesre rávetítjük a köröket. Vessük össze az eredményt a PCA biplottal (7.7 ábra) is!

dik kanonikus korreláció értéke is nagynak tűnhet, de ez önmagában nem sokat jelent, hiszen a hozzátartozó külső-variancia-hányad kevesebb 1 %-nál! Kiindulva abból, hogy a szignifikancia próba feltételei teljesülnek, a Bartlett próba eredményeit is célszerű megvizsgálnunk. Mindkét tengelyt megtartva az X^2 értéke 545,2 (a kritikus χ^2 érték d.f.=8 és α=0,05 mellett 15,5), ami azt jelenti, hogy a három csoport azonos statisztikai populációból csak igen kis valószínűséggel származhat, vagyis a fajok között van különbség. Az 1. tengely elhagyása után is marad egyébként X^2=37,2 (d.f.=3, α=0,05 mellett a kritikus érték χ^2=7,85), ami arra utal, hogy még a 2. tengelynek is van szerepe a fajok elválasztásában, bár ez már jóval gyengébb. Mindezt az izodenzitás-körök is alátámasztják (7.24 ábra): az *I. setosa* teljesen elkülönül, a másik két faj kismértékben átfed az 1. tengelyen, míg a 2. tengelyen az *I. versicolor* mutat enyhe elválást. A konfidencia-köröket nem mutatjuk be az ábrán, mert azok rendkívül kis átmérőjűek, a csoportok centroidjai tehát egyértelműen elkülönülnek egymástól, összhangban a Bartlett-teszt eredményével.

Térjünk most rá az eredeti bélyegek értékelésére. A három faj elkülönülését legerőteljesebben a belső lepel méretei teszik lehetővé (7.2 táblázat és 7.24 ábra). Az 1. tengellyel igen magas korrelációt adnak, s nagyon magas az *F*-hányadosuk is. A külső lepel hossza már jóval kevésbé diszkriminatív, míg a szélesség esetében a leggyengébb a fajok elkülönülése. Ez a tengelyekkel adott korrelációkból is jól látható. Mindezzel összhangban van a kommunalitások nagysága is.

7.6 Morfometriai ordináció

Az *Iris* adatok különféle elemzéseivel voltaképpen már eddig is érintettük a biológiai adatelemzés egy speciális területét, a *morfometriát*. Ennek elsődleges célja az alakbeli és méretbeli változatosság vizsgálata és elemzése, különös tekintettel e két tényező elválasztására. Az eddig ismertetett dimenzió-redukáló módszerek több-kevesebb sikerrel alkalmazhatók a morfometriában (míg azelőtt szinte kizárólagosak voltak e területen, vö. Blackith & Reyment 1971 klasszikus monográfiájával). Ma már azonban számos olyan speciálisan morfometriai célú eljárás áll rendelkezésünkre, amely sokkal alaposabb vizsgálódást tesz lehetővé, s az alakbeli változás értelmezését is megkönnyíti (Rohlf & Marcus 1993). E módszerek – a statisztikai értékelésen és a biológiai interpretáció elősegítésén túlmenően –

adatfeltárásra is alkalmasak a taxonómiai és evolúció-biológiai vizsgálatokban, így mindenképpen szólnunk kell róluk. A téma azonban – szinte már közhelyként mondjuk, ha valami "új" következik – olyan szerteágazóvá vált röpke tíz esztendő alatt, hogy e kötetben csak egy rövid összefoglalást adhatunk – különös tekintettel az ordinációs szempontokra –, megmutatva a tovább-informálódás lehetőségeit mindazoknak, akik úgy érzik, hogy problémáik csak ilymódon oldhatók meg.

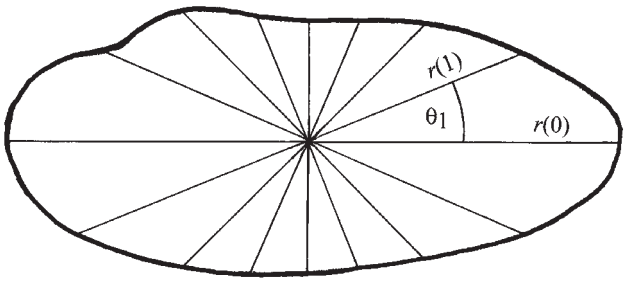
Az *Iris* példákban voltaképpen *távolságértékekkel* dolgoztunk: a lepellevelek bizonyos kitüntetett pontjai (csúcsa, töve, szélső pontok mindkét oldalon) közötti távolságok szerepeltek változóként. Ez sok más esetben is így van, az egyedeken felvett *tájékozódási pontok* (mérőpontok, kulcspontok, "landmark"-ok) közötti távolságok adják a morfológiai bélyegeket.[9] E távolságok azonban nem alkalmasak arra, hogy az eredeti alakot pontosan reprodukáljuk belőlük, vagyis a méretek alkalmazásával nem használunk fel minden alakbeli információt. Amennyiben a vizsgált objektumok teljes alakját szeretnénk elemzés tárgyává tenni, sokkal kifinomultabb technikákat kell igénybe vennünk. A "kifinomultság" nem azt jelenti, hogy az alkalmazandó adatfeltáró módszerek gyökeresen eltérnének az eddig megismertektől, hanem arra utal, hogy az adatrögzítés módszerei lényegesen mások. Az esetek jelentős részében ugyanis a speciális módon nyert adatokat később éppen a már jól ismert és bevált módszerek értékelik. Megjegyezzük továbbá, hogy eme új adattípusok – minden látszólagos és valós előnyeik ellenére – nem teszik feleslegessé a korábbi, "tradicionális" morfometria távolságokra alapozott eljárásait, amint azt pl. Reyment (1990) és Marcus (1990, 1993) is hangsúlyosan kiemeli.

7.6.1 Kontúr-elemzés

A szervezetek alakjának teljesebb figyelembevételére az első lehetőség az objektum *kontúr*jának, külső körvonalának ("*outline*") elemzése. Rohlf (1990a) tekinti át részletesen azokat a módszereket, melyek révén a teljes kontúrvonalra (zárt kontúr), vagy két kitüntetett kulcspont közötti szakaszra (nyitott kontúr) függvényeket illeszthetünk. A kapott függvények paramétereit – mint input adatokat – szokványos többváltozós elemzésnek vethetjük alá. Ez a megközelítés persze teljes mértékben "elfeledkezik" a kontúrvonalon belülre eső jellegekről, s ezért csak akkor célszerű alkalmazni, ha az objektumok kifejezetten szegények belső bélyegekben (pl. Ostracoda és Mollusca héjak esetében).

Figyelmünket a továbbiakban a zárt kontúrral leírható alakokra összpontosítjuk, mert ezek lényegesen fontosabbak – és gyakoribbak – a morfometriai vizsgálatokban, mint a nyitott görbék. Az elemezni kívánt objektumokon találnunk kell egy kulcspontot amely biológiailag "azonos jelentésű" (azaz *homológ*) minden esetben. Ettől a ponttól kezdjük a görbe leírását és ide térünk vissza. Célszerű egy másik homológ kulcspont kijelölése is, mert kettőjük segítségével minden objektum egyértelműen elhelyezhető egy derékszögű koordináta-rendszerben. A standardizált elhelyezésmód kötelező, mert máskülönben az objektumok összehasonlítása értelmét veszti. Az objektum alakját vagy a centroidtól (vagy más középponttól) húzott su-

9 A *landmark* voltaképpen olyan speciális mérőpontnak felel meg, ahol valamilyen struktúrák kicsúcsosodnak, kereszteződnek, stb. Az objektum szélső pontjait inkább *pseudo-landmark*-nak nevezik (Rohlf & Marcus 1993), bár ez a megkülönböztetés a mi céljaink szempontjából most nem lényeges. A landmark-ok finomabb osztályozását lásd Bookstein (1991) könyvében.



7.25 ábra. Egyenlő szögben felvett rádiuszok alkalmazása kontúrvonalak leírására, az *Unio pictorum* példáján.

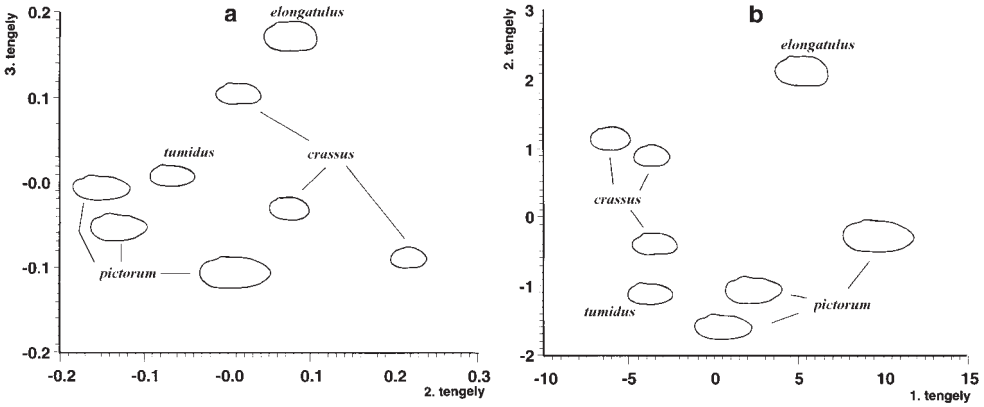
garak hosszúságai, vagy pedig a kontúr mentén megfelelő sűrűségben felvett pontok *x,y* koordinátái képviselik majd az elemzésben. A zárt görbe matematikai leírására a legismertebb módszerek a következők:

- Nem túl bonyolult[10] kontúrvonalak egyszerű leírására alkalmas a *rádiusz-függvény* (Scott 1980, Lohmann & Schweitzer 1990). Az objektum súlypontjából az első kulcspontig húzott sugárt tekintjük majd hivatkozási alapnak. Ettől számítva egyenlő közökben (helyesebben *szögek*ben) sugárirányú egyeneseket húzunk a kontúrvonalig; legyen a sugarak száma *p* (7.25 ábra). A rádiusz-függvény valójában az elfordulási szög és a hozzátartozó sugár hossza közötti összefüggést adja meg [r , θ] értékpárok formájában. Az objektumok alakja elég jó közelítéssel leírható a *p* számú hosszértékkel, különösen akkor, ha *p* elég nagy. (A sugarak alkalmazásával voltaképpen a kontúrvonal szisztematikus "mintavételezését" hajtjuk végre, amely annál hatékonyabb, minél több pontot veszünk fel.) A hosszértékek egy $p \times m$-es mátrixba összesíthetők, amelyet azután standardizált főkomponens elemzésnek vethetünk alá. Ez a PCA egy speciális esete, hiszen a korrelációt az objektumok (és nem a változók, vagyis a kontúr adott pontjaira mutató rádiuszok) között számítjuk ki. Lohmann & Schweitzer (1990) *alakkomponens-elemzés* ("*eigenshape analysis*") néven tárgyalja a PCA ilyen speciális alkalmazásait (lásd még lentebb). A PCA diagramok közül az objektumok és a komponensek közötti korrelációk diagramja lesz igazán érdekes, amit *Unio* kagylók kontúrvonalainak elemzésével szemléltetünk.

  A vizsgálatban 4 faj szerepel, az *U. pictorum* és *U. crassus* három, más és más lelőhelyről származó egyeddel, az *U. tumidus* és *U. elongatulus* pedig egy-egy egyeddel (részleteket lásd az A8 táblázatban). A rádiusz-értékekből végrehajtott alakkomponens-elemzés egy igen magas sajátértéket adott (97 %), ami nem szokatlan, ha a kontúrok erősen hasonlítanak egymásra (a legkisebb korreláció az *U. pictorum* és az *U. crassus* között volt (*COR*=0,926), a legmagasabb pedig az *U. pictorum* és *U. tumidus* között (*COR*=0,99). Ez a nagy sajátérték lényegében véve egy általános méretbeli komponenst fed le, az 1. komponensen mind a nyolc kontúrvonal nagyon magas értékekkel szerepel (0,971 és 0,994 között), s így nincs értelme ábrázolni. Emiatt – bár kicsiny variancia jut rájuk – a második és a harmadik komponens jelentősége megnövekszik, s az egyedeket a 2-3. dimenzióban ábrázoljuk (7.26a ábra).

  Felmerül persze egy "szokványos" standardizált PCA végrehajtásának a lehetősége is, amelyben a változók a sugarak hosszértékei, az objektumok pedig maguk a kagylópéldányok.

10 A "nem túl bonyolultság" tartalma majd később, a harmadik módszer tárgyalásában válik nyilvánvalóvá.



7.26 ábra: *Unio* kagylóteknők kontúrvonalának komponens-elemzése **a:** az *Unio* egyedek között számolt korrelációból a 2. és 3. alakkomponensre és **b:** a rádiuszok közötti korreláció alapján az 1-2. komponensekre. A **b** ábrán a vízszintes és a függőleges tengely skálája erősen eltér!

Ebben az esetben is igen magas 1. sajátértéket kaptunk (91,3 %), de jelentős még a 2. sajátérték is (5,2 %). Az így kapott ordinációs diagramot tüntetjük fel a 7.26b ábrán.

Az adatelemző dilemmája nehezen kerülhető meg: melyik eredményt vegyük elsősorban tekintetbe az *Unio* egyedek alak-szerinti ordinációjában? Az alakkomponens-értékelés kétségtelen "hátránya", hogy a bemutatott ordináció mindössze kb. 2 %-nyi összvarianciát fed le, a többi a méret általános komponense. Ezzel szemben a normál PCA első két tengelye 96,5 %-ot értelmez, és sokkal jobban "szétdobja" a nyolc egyedet. Egyik ábrán sem tűnik azonban erősnek a populáción belüli "összetartás", az egy fajhoz tartozó egyedek ugyanis nem kerülnek közel egymáshoz, vagyis a kontúrvonal önmagában *nem elegendő* a fajok egyértelmű elválasztásához.

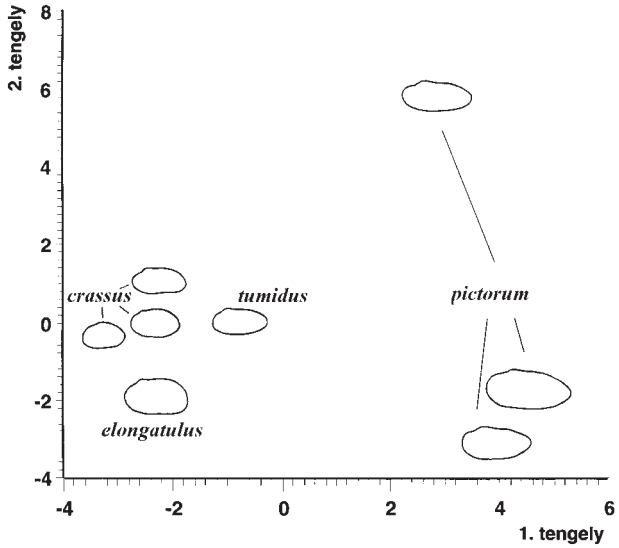
- A fenti módon megmért rádiuszok az ún. *Fourier-analízis* (harmonikus analízis) segítségével egy függvénysor összegeként is előállíthatók (Rohlf 1990a). Az elemzés azt a matematikai törvényszerűséget használja fel, hogy – Fourier francia matematikus tétele szerint – minden "görbe" előállítható egyszerű "hullámok" (harmonikusok) összegeként. Az első (referencia) sugárral θ szöget adó sugár, vagyis r(θ), hossza a következő sor segítségével közelíthető:

$$r(\theta) = a_0 + \sum_{i=1}^{k} a_i \cos i\theta + b_i \sin i\theta \qquad (7.80)$$

ahol k a kiszámított harmonikusok száma ($k < p/2$), és

$$a_0 = \sqrt{\frac{2}{p}} \sum_{j=1}^{p} r_j, \quad a_i = \sqrt{\frac{2}{p}} \sum_{j=1}^{p} r_j \cos i\theta_j, \quad b_i = \sqrt{\frac{2}{p}} \sum_{j=1}^{p} r_j \sin i\theta_j, \qquad (7.81a\text{-}c)$$

A Fourier-elemzés a k számú harmonikusra becsli az a_i és b_i paramétereket, amelyek az objektum alakjának absztrakt, közvetett leírására használhatók fel. A $h_i = a_i^2 + b_i^2$



7.27 ábra. Az *Unio* teknők standardizált PCA elemzése a rádiuszokra (A8 táblázat) illesztett Fourier-együtthatók alapján.

mennyiség a harmonikus *amplitúdó*, a sor *i*-edik tagjának relatív "hozzájárulása" a kontúrhoz. Több objektumról származó Fourier-koefficiensek (ha az 1. sugár minden esetben homológ kulcspontra mutat) egy nyers adatmátrixot szolgáltatnak a többváltozós elemzés számára. Ha azonban nincs homológ kulcspont, csak a harmonikus amplitúdók jöhetnek számításba, de ez már kétségtelen információvesztéssel jár. Maguknak a harmonikusoknak nemigen tulajdoníthatunk biológiai jelentőséget, de leíró – következésképpen ordinációs – célra viszont alkalmazhatók (Rohlf 1993a).

Az *Unio* teknők értékelésében a Fourier-együtthatók szabványos PCA értékelése lényegesen eltérő eredményt ad, legalábbis a sajátértékek relatív nagyságát illetően (λ_1=32 %, λ_2=23 % és λ_3=16 %). Az egyedek ordinációs elrendeződése (7.27 ábra) már nem annyira különbözik az előzőektől, de azoknál talán valamivel jobban interpretálható. Az *U. pictorum* az 1. tengelyen jól elválik a többi fajtól, az *U. crassus* pedig egy viszonylag kompakt csoportot alkot a tengely másik végén. Az *U. elongatulus* elkülönülése viszont kevéssé kifejezett, mint a 7.26b diagramon. Az *U. tumidus* minden ábrán "átmeneti" pozícióban van a *pictorum* és a *crassus* egyedek között.

- A fenti két módszer "hibájaként" leginkább azt róhatjuk fel, hogy a súlypont kijelölése valójában egy teljesen önkényes lépés, és egy másik – biológiailag esetleg még logikusabb – referenciapont alapján könnyen eltérő eredményt kaphatunk. Bonyolultabb körvonalakra pedig, amikor ugyanaz a sugár esetleg két v. több helyen is metszi a kontúrt, már egyáltalán nem alkalmazhatók. E problémák legismertebb módon a Zahn & Roskies (1972) javasolta "alakfüggvény" révén küszöbölhetők ki,

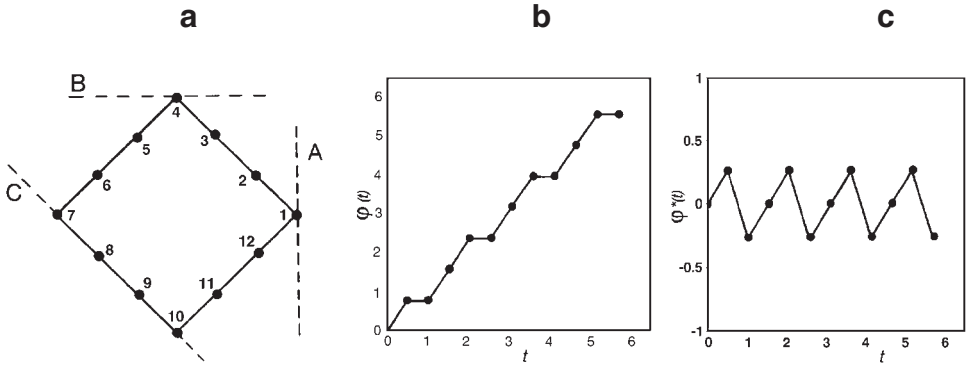
$$\varphi^*(t) = \varphi(t) - t \qquad (7.82)$$

amely viszont szintén megkívánja egy kezdő tájékozódási pont kijelölését. A t a kezdőponttól vett távolság a kontúrvonalon olymódon normálva, hogy a teljes kontúr hossza éppen 2π radián legyen. $\varphi(t)$ a szögeltérés a 0 pontban a kontúrhoz húzott érintő és a kezdőponttól t távolságban lévő pontban húzott érintő között, radiánban. A függvény értéke egy szabályos kör esetében minden pontban 0, egyéb formák

esetében tehát a körtől, mint referencia-alaptól, való alakbeli eltérést mérünk vele. E módszerben nem a súlypontból húzott sugarak elrendeződése, hanem a kontúr mentén egyenlő közökben felvett t távolságok biztosítják a körvonal szabályos "mintavételezését" (nem árt legalább 100 egyenlő részre felosztani minden egyes kontúrt, vö. Reyment 1991). Ezután a $\varphi^*(t)$ függvény értékeit mátrixba egyesítve *alakkomponens-elemzést* hajthatunk végre (Lohmann 1983, Lohmann & Schweitzer 1990). Itt is fennáll az előző pontból ismert lehetőség: a 7.82 függvény értékeit Fourier-analízissel elemezhetjük, s a kapott koefficienseket többváltozós analízisnek vethetjük alá (Rohlf 1993a).

A Zahn-Roskies alakfüggvény kiszámítását a négyzet egyszerű példáján mutatjuk be (7.28a ábra). Az egyik csúcstól indulva 12 pontot jelölünk ki egyenlő távolságra egymástól (mivel a négyzet kerületét 2π-re normáltuk, ez a térköz $2\pi/12=0{,}52$ lesz). Az 1. ponthoz húzott érintő lesz a referencia alap. A $\varphi(t)$ függvény monoton növekvő (7.28b ábra), a $\varphi^*(t)$ függvény pedig a kört képviselő egyenes körül "oszcillál" (7.28c ábra), mutatva a négyzet és a kör közötti különbség szabályos váltakozását. Belátható, hogy a négyzet elforgatásával a $\varphi^*(t)$ értéke nem változik. E módszer sem mentes azonban bizonyos problémáktól. A kezdőpont egyértelmű kijelölésén túlmenően az is lényeges, hogy az órajárás szerint vagy azzal ellentétes irányban értékeljük-e a kontúrvonalat (bár ez a négyzetnél éppen nem így van).

- A kontúrvonal leírására legáltalánosabban alkalmazható az ún. elliptikus Fourier-elemzés, amely koordinátákból indul ki. A "független változó" itt is a körvonal mentén felmért távolság a $[0,2\pi]$ intervallumban, a keresett függvény pedig a felvett pontokra vonatkozó Δx és Δy koordinátáknak a t távolsággal való együttes megváltozását írja le harmonikusok összegeként, akár a rádiusz-függvény esetében (Kuhl & Giardina 1982). Minden egyes harmonikusra négy Fourier-együttható adódik (kettő az x, kettő pedig az y koordinátákra) és kettő konstanst is figyelembe kell vennünk. A meglehetősen terjedelmes képleteket most mellőzzük, az Olvasó megtalálhatja ezeket Rohlf (1990a, 1993a) munkáiban. A módszer rendkívüli előnye, hogy az iránynak sőt a kezdőpont kijelölésének sincs befolyása a végeredményre (legalábbis a módszer megvalósításában az **EFA** programban, vö. Rohlf & Ferson 1992). A kontúr mentén



7.28 ábra. A Zahn-Roskies alakfüggvény meghatározása a négyzet esetén. **a:** A négyzet kerületén kijelölt 12 pont. A az 1. ponthoz, B a 4. ponthoz, C pedig a 8. és a 9. ponthoz húzott érintő. **b**: a $\varphi(t)$ függvény, **c:** a $\varphi^*(t)$ függvény a 12 pontra.

nem kell egyenlő távolságban felvenni a pontokat és igen bonyolult, akár önmagát keresztező kontúrvonal elemzésére is felhasználható. Biológiai – és ordinációs célú – alkalmazásokat találunk Rohlf & Archie (1984) ill. Ferson et al. (1985) cikkeiben.

7.6.2 Mérõpontok felhasználása ordinációra

A körvonalon alapuló elemzésekkel két alapvető problémát említhetünk. Az egyik, hogy az egyed belsejét teljesen figyelmen kívül hagyják (ezt már említettük is), másrészt pedig azt, hogy az alak változásának biológiai magyarázata ilymódon szinte lehetetlen (Bookstein 1991). A megoldást egyértelműen a mérőpontok kijelölése és alkalmazása jelenti, hisz ezek is kiindulópontot jelenthetnek az adatfeltáró munkához, s ugyanakkor az "új geometriai morfometria" kifinomultabb elemzési lehetőségeit is megengedik. Az összehasonlítani kívánt objektumokat koordinátarendszerbe helyezzük.

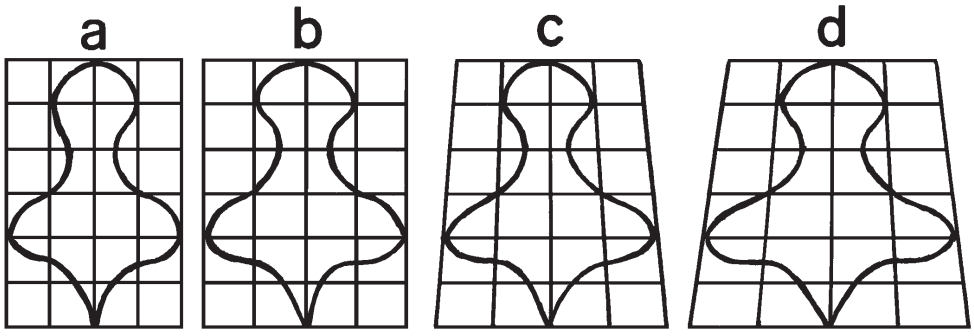
- Az objektumok összehasonlítása a koordináták közvetlen felhasználásával is lehetséges, ha kiválasztunk két – nem túl közeli – kulcspontot, s a közöttük húzható egyenest tekintjük referenciának. Ezután minden objektumot úgy helyezünk el a koordináta-rendszerben, hogy a referencia-egyenes ("*baseline*", azaz alapvonal) ráessen az x tengelyre, éppen a –0,5 és 0,5 értékek között. Ezzel a standardizálással kapjuk az ún. *Bookstein-féle alak-koordinátákat* (Bookstein 1991). Ha összesen p kulcspontunk van, akkor a többváltozós elemzés inputját minden objektumra $2(p–2)$ érték jelenti majd. Loy et al. (1993) munkája példa a Bookstein-koordináták többváltozós elemzésére (vakond-koponyák vizsgálatában, pl. diszkriminancia elemzéssel – és osztályozással is).
- A koordináták másik közvetlen felhasználási lehetőségét jelentik a *szuperpozíciós* módszerek. Ekkor a feladat az egyik objektum forgatása és nagyítása/kicsinyítése olymódon, hogy a homológ kulcspontokat tekintve maximálisan illeszkedjen a vele összehasonlítandó másik objektumra.[11] Az objektumpár távolsága a homológ kulcspontok közötti eltérések négyzetösszegeként definiálható. m objektum között minden párosításban meghatározható a négyzetösszeg, s a kapott $m \times m$-es távolságmátrix a már ismert módokon értékelhető (pl. Chapman 1990, Sanfilippo & Riedel 1990). A szuperpozíciós módszereknek több verziója ismeretes: ezeket, és a velük kapcsolatos problémákat Rohlf & Slice (1990), Chapman (1990) és Rolhf (1990b) tekintik át részletesen. Az utóbbi cikk konklúziója szerint a szuperpozíciós technika leginkább akkor használható, ha a különbségeket viszonylag kevés számú kulcspont okozza, vagy amikor az eltérések a kulcspontokon közelítőleg véletlenszerűen oszlanak meg.
- A koordináták közvetett alkalmazása jellemzi a geometriai morfometria legújabb, "forradalmi"-nak nevezett irányzatát. A gyökerek egészen Thompson (1917) híres könyvéig nyúlnak vissza: ebben a szerző az alak (pl. koponya, falevél stb) változását egy négyzetrács segítségével szemléltette a 7.29 ábrán látható módon. Thompson gondolkodásmódja hosszú ideig csak leíró szinten érvényesült a biológiai alak megvál-

11 Ezt Prokrusztész-módszer néven ismerjük a többváltozós elemzés irodalmában, s általánosan ordinációk összehasonlításában alkalmazzuk (a kulcspont-koordináták is – két- vagy háromdimenziós – speciális ordinációk). A részletes ismertetést lásd a 9. fejezetben.

tozásával kapcsolatos elképzelésekben. Nemrégiben derült ki, hogy a mechanikában már régebben rendelkezésre áll az az eszköztár, amely rendkívül kifinomult módon képes az alakváltozás elemzésére, nagymértékben elősegítve a biológiai interpretációt is. E módszerek jellemzése igen nagy helyet és az eddigieknél is alaposabb matematikai ismereteket igényelne, így csak röviden utalunk rájuk. Az ordinációs alkalmazás egyébként is csak "melléktermékként" merül fel, de a teljesség kedvéért, és e módszerek várható népszerűsége miatt is meg kell említenünk. A részletek iránt érdeklődők elsősorban Bookstein (1991) könyvéből, illetve a Rohlf & Bookstein (1990) és Marcus et al. (1993, 1996) szerkesztette cikkgyűjteményekből tájékozódhatnak.

A geometriai morfometria az alakváltozás két fő komponensét különíti el. Az *affin* (vagy uniform = egyenletes) alakváltozás minden olyan átalakítást tartalmaz, ami a mérettel, forgatással, tükrözéssel kapcsolatos, és ilyennek tekinti az egyirányú és azonos mértékű (homogén) megnyújtást vagy összenyomódást is (7.29b ábra). A nem-affin vagy *deformációs* elváltozások ezzel szemben nem kitüntetett irányúak, inhomogének, az egyes mérőpontokon más és más mértékűek lehetnek, és szabálytalan torzulásokat eredményeznek (7.29c ábra). Ha a vizsgált objektumok halmazában van egy kitüntetett referencia-objektum (ez lehet egy típuspéldány vagy az összes egyed általánosított Prokrusztész elemzésével [9.4.3 rész] kapott "átlagos" objektum), akkor a geometriai morfometria eszközeivel ezen elváltozások elkülöníthetők, komponensekre bonthatók, amelyek végeredményben ordinációs input adatokat is szolgáltatnak.

Kiindulásképpen képzeljünk el egy rendkívül vékony, sima fémlemezt, amelyen számos pontot megjelölünk, akárcsak a biológiai objektumokon. Ezt kissé meghajlítva, meggyűrve a mérőpontok vertikális irányban elmozdulnak. A torzuláshoz szükséges idealizált "energiát" az ún. hajlítási energiamátrix, $\mathbf{L}^{-1}$, segítségével fejezhetjük ki. Ennek mérete $p \times p$, (ahol p a mérőpontok száma), és a megfelelő mérőpontok közötti távolságokból ill. a referencia-objektum koordinátáiból kapott mátrix invertálásával állítjuk elő (l. Rohlf 1993b: 137-138). Az affin átalakításhoz az energia 0, hiszen semmi hajlítás nem történt, csak nyújtás vagy összepréselés. A hajlítási energiamátrixot formálisan kiszámíthatjuk a referencia-objektum és a j-edik példány (célobjektum) között is (arról egyáltalán nincs szó persze, hogy a biológiai objektumok a fémlemezhez hasonló módon viselkednének, s ezért *formális* csupán az alkalmazás). Az $\mathbf{L}^{-1}$ mátrixból és a célobjektum koordinátáiból megkapjuk az ún. vékonylemezes interpolációs függvényt ("*thin plate spline*"), amely homogén és inhomogén összetevők segítségével pontosan leírja a referencia objektum egy adott mérőpontjának leképezését a célobjektumba.



7.29 ábra. Egy hipotetikus levélalak (**a**) megváltozása a Thompson-féle transzformációs négyzetrácsban. **b:** homogén alakváltozás, **c:** torzulás, **d**: a két változás összege.

Az inhomogén transzformációt jellemző energiamátrix spektrálfelbontása (C függelék) a főkomponensekhez hasonló ortogonális vektorokat eredményez (*"principal warps"*), amelyeket *főtorzulás*ok vagy deformációs komponensek névvel illethetünk. Az egyes vektorok a különböző geometriai léptékben bekövetkezett inhomogén változások mértékét jellemzik, és akár rendszertani bélyegként is alkalmazhatók (Rohlf & Marcus 1993). Az utolsó 3 sajátérték szükségképpen 0, így az ehhez tartozó vektorokra a továbbiakban nincs szükség. A két objektum közötti koordináták különbségéből, a vektorokból és a sajátétékekből a PCA-hoz hasonló módon kaphatjuk meg a célobjektum *deformációs értékeit, p*–3 számút mind az x mind pedig az y tengelyre (*"partial warp scores"*, Rohlf 1993b).

Ha van egy *m* objektumból álló mintaösszletünk, akkor természetesen külön-külön meghatározhatjuk mindegyikre a deformációs értékeket. Ezek egy 2(*p*–3) hosszúságú vektorba írhatók[12], s ezután egy *m* × 2(*p*–3) méretű **W** mátrixban egyesíthetők. Itt léphet be a vizsgálatba a már ismert többváltozós módszerek valamelyike, például a főkomponens elemzés. A **W** mátrix főkomponens analízise (amelyet a szakirodalom *relative warp analysis* néven emleget) végül is az objektumok és a referencia közötti eltérések lineáris kombinációját állítja elő. A kapott koordináták segítségével a populáció egyedei szórásdiagramban ábrázolhatók. Nem feltétlenül kell persze PCA-t alkalmaznunk, hiszen a diszkriminancia elemzés sőt numerikus osztályozás is szóba jöhet. Újabban az a lehetőség is felmerült, hogy megfelelő átalakítás után a koordinátákat kladisztikai elemzésben használjuk fel (Zelditch et al. 1995). Naylor (1996) ennek ellenőrzésére halak evolúcióját szimulálta, s a kapott optimális fák egyike pontosan megegyezett a "valódi" evolúciós fával – jelezve, hogy nagyon is érdemes ebben az irányban tovább vizsgálódni.

7.7 Irodalmi áttekintés

Az ordinációs módszerek irodalma, hasonlóan a klasszifikációéhoz, rendkívül terjedelmes és nehezen áttekinthető, még akkor is, ha csak a biológiai alkalmazásokra szorítkozunk. Mindenesetre megjegyzendő, hogy amennyiben egy kézikönyv címe utalást tartalmaz a "többváltozós módszerek"-re, akkor csaknem biztosak lehetünk abban, hogy az adatfeltáró ordinációs módszerek nagy hangsúllyal szerepelnek benne. A "többváltozós statisztika" említése a címben viszont inkább a formális statisztikával (pl. szignifikancia-próbákkal, többváltozós normalitás, stb) való kapcsolatot jelzi, s céljainkkal kevésbé egybeeső tartalomra utal. A biológiai ordinációt általánosságban – bár egyes módszereket részletezve is – taglaló művek száma igen nagy, így csak kiragadott példaként említünk néhányat. Az ökológus/cönológus számára a Whittaker (1973, 1978) szerkesztette kötetek sok hasznos információval szolgálhatnak, elsősorban a kezdetekről. Greig-Smith (1983) és Kershaw & Looney (1985) is haszonnal forgatható. Gauch (1982) a formalizmus szinte teljes mellőzésével próbál bevezetni bennünket az ordináció témájába; inkább kevesebb, mint több sikerrel. Haladók számára ajánlható – ebben a sorrendben – Ludwig & Reynolds (1988), Pielou (1984), Orlóci (1978), Legendre & Legendre (1983[13], 1987) és Digby & Kempton (1987). Az egyre inkább előtérbe kerülő kötött ordinációs stratégiákról külön kötet még nincs, csak review (ter Braak & Prentice 1988) illetve cikkgyűjtemény (ter Braak 1996). A taxonómus számára Sneath & Sokal (1973) jelenti e tekintetben is a kiindulópontot, Dunn & Everitt (1982) csak bevezető jellegű. A modernebb taxonómiai könyvek – a kladisztika előretörésének megfelelően – az ordinációt már sokkal kisebb súllyal vagy egyáltalán nem tárgyalják (kivétel Stuessy 1990). Az általános művek közül az ordinációs módszerek adatstruktúra feltáró funkcióját emeli ki Gordon (1981), míg Cooley &

12 Természetesen létezik háromdimenziós eset is, de ezt most az egyszerűség kedvéért mellőzzük. Ekkor x, y és z koordináták szerepelnek a kissé bonyolultabb elemzésben.

13 A könyv második kiadása 1997 első felében várható.

Lohnes (1971), Mardia et al. (1979), Chatfield & Collins (1980), Dillon & Goldstein (1984) általánosabb tárgyalásmódot követ.

A főkomponens elemzés részletes áttekintése számos – általában nem biológiai – példával megtalálható Jolliffe (1986) könyvében. A szerző részletesen elemzi a PCA és más többváltozós módszerek kapcsolatát, együttes alkalmazásának lehetőségeit. Természetesen a PCA leírása minden, többváltozós módszerekkel foglalkozó könyv szerves része, így ezeket talán már nem érdemes felsorolni. Megemlítendő, hogy könyvünkkel ellentétben Jongman et al. (1978) a PCA-t mint a legkisebb négyzetek elvének egyik alkalmazását tárgyalja, s egy iteratív algoritmus leírásával a PCA megértésének – és megértetésének – egy alternatív lehetőségét mutatja be. Rao (1973) (még Bookstein 1991:39) is rámutat arra, hogy a főkomponenssúlyok (melynek négyzetösszegét most a sajátértékkel s nem 1-gyel tesszük egyenlővé) vektorának mátrix-szorzata egy olyan 1-es rangú mátrixot eredményez, melynek elemei minimális eltérésnégyzetet adnak a kiinduló kovariancia mátrix elemeivel. A PCA legújabb összefoglalása megtalálható Jackson (1991) könyvében, amelyből a biplot-technikákkal kapcsolatosan is részletesen tájékozódhatunk.

Bár a faktor analízis témáját csak egy rövid bekezdés erejéig érintettük, ez nem jelenti azt, hogy a biológiában hasonlóan egyértelműen mellőzött lenne. Cattell (1978) például csak a biológiai alkalmazásokkal foglalkozik. Sajnos Reyment & Jöreskog (1993) könyvének a címe ("*Applied Factor Analysis in the Natural Sciences*") némiképpen félrevezető, hiszen a kötetben valójában az ordinációs módszerekről esik szó, és a *sensu stricto* faktoranalízis ("True factor analysis" címmel) egy alfejezetnyi rész csupán. Holott Wright (1954) elég korán rámutatott a faktor-analízis és a PCA közötti különbségekre, úgy tűnik a terminológiai zűrzavar a "faktorok" körül nem igazán tisztul, s kár, hogy ez az új könyv csak belezavar ebbe.

A kanonikus korrelációelemzés mindmáig legjobb összefoglalója Gittins (1985). Ebben a könyvben mindenki megtalálhatja a további tájékozódáshoz alkalmas irodalom listáját. A korreszpondencia-analízis "bibliája" Greenacre (1984) könyve (lásd még van Rijkevorsel et al. 1988), bár a módszerről más címen – és nyelven – is olvashatunk (Benzécri et al. 1973).

A kanonikus korreszpondencia-elemzés népszerűsége rendkívül gyorsan növekedett az utóbbi években. Mindez jól felmérhető, ha megvizsgáljuk a módszer 1986-1993 közötti alkalmazásainak teljes bibliográfiáját (Birks et al. 1996). Reyment (1991) szerint a módszer nem feltétlenül korlátozódik recens ökológiai alkalmazásokra, s bemutat egy "paleós" példát is. A diszkriminancia-analízis is szerepel szinte minden többváltozós szakkönyvben, különös súlyt fektet erre a módszerre Mardia et al. (1979).

A morfometriai módszerek klasszikus ordinációs tematikájú irodalmát 1970-ig Blackith & Reyment (1971) könyvéből ismerhetjük meg leginkább. Az azóta eltelt időszak fejleményeit maga Reyment (1990, 1991) tekinti át legrészletesebben, amiből kiderül, hogy az ordinációs módszerek vesztettek ugyan hangsúlyos szerepükből, de továbbra is szerves részei a morfometria eszköztárának, különösképpen a mérőpont-adatok értelmezésében (l. még Marcus 1990, 1993). Ezzel szögesen ellentétes véleményen van Bookstein, aki rendszeresen "alulbecsli" a deszkriptív, ordinációs jellegű statisztikai feldolgozások fontosságát (pl. Bookstein 1990, 1991, 1993), mert szerinte az ordináció a legjobb esetben is információvesztéssel jár s nem alkalmas az alak biológiai interpretációjára. Kifejezetten a forradalmian új geometriai morfometria mellett száll síkra, melynek legáltalánosabb bevezetése (Bookstein 1991) minden, a téma legújabb fejleményei iránt érdeklődő olvasónak ajánlható. Ordinációk – speciális tengelyek mentén – persze még ebben a kötetben is fontosak maradnak. A morfometria legeslegújabb eredményeit illetően pedig a Marcus et al. (1996) szerkesztette kötetet érdemes forgatni, amelyben örvendetes módon hazai kutatók cikkei is szerepelnek (pl. Demeter et al.

1996 számítógépes/kamerás adatrögzítő eljárásáról, amely nagymértékben megkönnyíti a koordináták felvételét). A morfometria fő szakirodalma mindenesetre ma még néhány kötetnyire korlátozódik, amit az is mutat, hogy "kék könyv", "narancsszínű könyv", "fekete könyv" stb. megjelöléssel illetik azokat, de a színek közeljövőbeli elfogyása bizton megjósolható.

7.7.1 Számítógépes programok

Egy átlagos felhasználó – a könyvben ismertetett módszereket illetően – leginkább az ordinációs szoftverekkel van elkényeztetve. A "nagy", kereskedelmi forgalomban kapható statisztikai programcsomagok is rendszerint tartalmaznak néhány ordinációs eljárást, bár nem biztos, hogy éppen az ordinációs cél domborodik ki a dokumentációban és a felhasználói környezetben. A jelen céljainkra leginkább megfelelő ill. könnyen elérhető szoftvereket foglaltuk össze a 7.3 táblázatban, teljességre nem törekedve (lásd még a B függeléket).

Egy ordinációs program értékelésében sok tényezőt kell figyelembe vennünk. Fontos például, hogy milyen grafikus lehetőséggel egészül ki a numerikus eredmények listája. A főkomponens-elemzésben és a korreszpondencia-analízisben is igen fontos, hogy a biplot azonnal megjelenjen, amelyre sok programban nincs lehetőség (kivétel pl. a **SYN-TAX**, amelynek új változata sokféle biplot automatikus megjelenítését teszi majd lehetővé). A **CANOCO** program, amely igen sok ordinációs és ordináció-értékelő opciót tartalmaz (hiszen speciálisan erre készült), nem rendelkezik saját grafikus rutinokkal, hanem a P. Šmilauer által kidolgozott **CANODRAW-LITE** vagy **CANODRAW 3.0** (Šmilauer 1992) programokra "hagyja" a grafikát, amely kimondottan jó minőségű, közleményekben azonnal felhasználható rajzokat (biplotot, és triplotot is) szolgáltat. A **CANOCO**-t egyébként elsősorban a kanonikus korreszpondencia- és a redundancia-elemzésre ajánlhatjuk, egyéb módszerekre kevésbé, mert azok más programcsomagokban könnyebben elérhetőek. A **Statistica** grafikus outputjának nagy előnye, hogy javítható, ízlés szerint átalakítható. További szempont a felhasználói környezet: az opciók könnyű megadhatósága, a választási lehetőségek menüs/ablakos tálalása, a "súgó" (help) jelenléte és így tovább. E tekintetben a **Statistica** tűnik a legjobb választásnak, amelyben a faktoranalízis sok válfaját is megtaláljuk (jól megkülönböztetve a főkomponens elemzéstől), bár más fontos módszerek, sajnos, hiányoznak (7.3 táblázat). A nem-metrikus többdimenziós skálázás módszerét – némi leegyszerűsítéssel –, csak többdimenziós skálázásnak nevezi a

7.3 táblázat. Ordinációs módszerek különféle számítógépes programcsomagokban.

Módszer	**Statistica**	**SYN-TAX**	**NT-SYS**	**CANOCO**	**NuCoSA**	**BMDP**
Főkomponens-elemzés	+	+	+	+	+	+
Faktoranalízis	+					+
Kanonikus korreláció	+	+	+	+		+
Redundancia analízis				+		
Korreszpondencia elemzés		+	+	+	+	+
Kanonikus korreszpondencia-elemzés				+		
Főkoordináta módszer		+	+	+	+	
Nem-metr. többdim. skálázás	+	+	+		+	
Diszkriminancia-elemzés	+	+	+	+		+
Vékonylemezes interpolációs függvény			+			

felhasználói kézikönyv, megfeledkezve a metrikus módszerről. A **BMDP** parancs-nyelvezete pedig talán a legkörülményesebb. A programok hardver-igénye sem mellékes: a 7.3 táblázat programjai DOS/WINDOWS környezetben futnak, de a **SYN-TAX**-nak van Macintosh verziója is (ráadásul grafikájában flexibilisebb, mint a DOS változat).

Nem szerepel a táblázatban, de megemlítendő még jó néhány más program is, mert kifejezetten biológusok számára készültek. Orlóci (1978), Orlóci & Kenkel (1985) és Ludwig & Reynolds (1978) sok ordinációs módszer BASIC nyelvű forráskódját adja meg. FORTRAN nyelvű lista van Gauch (1977) **ORDIFLEX** kézikönyvében. A Wildi-féle **MULVA-5** programcsomag (Wildi & Orlóci 1996) sok ordinációs módszere leginkább ökológusok és cönológusok nagy táblázatfeldolgozó igényeit elégítheti ki. Az egyengetett ("detrended") korreszpondencia-elemzés – hogy a kritika ellenére erről is szóljunk azért – mindmáig legnépszerűbb programja a **DECORANA** (Hill 1979b), de ennél ma már jobb a **CANOCO**, pl. a "kiegyengetési" opciókat tekintve.

Külön téma a morfometriai adatelemzés, amely számunka annyiban érdekes, hogy milyen programokat használjunk a többváltozós módszerek input adatainak előállítására. Az alakkomponens elemzés Macintosh programját MacLeod (1993) fejlesztette ki. PC-re alkalmas WINDOWS szoftverek a **tpsRelw** és **tpsSplin**, amelyek a modern morfometria eszköztárát teszik elérhetővé, bár ezek egy része a **NT-SYS** programcsomagban is megvan (további részletekről a B függelékben megadott Internet információ segíthet).

7.8 Kérdezz – Válaszolok!

K: *Nyilván nem állítod – nem is állíthatod –, hogy minden ordinációs módszerre jutott hely e könyvben, de – az ökológiai irodalmat böngészve – egy dolog feltűnő: nem szólsz egy szót sem a polár-ordinációról. Sokfelé láttam ezt említeni, s ezért kíváncsi vagyok: mi az ördög ez voltaképpen és miért nem szerepel a könyvben?*

V: A polár-ordinációnak elsősorban történeti jelentősége van; ma már nemigen használják. A módszert ökológusok (Bray & Curtis 1957) "spekulálták ki" még akkor, amikor a számítógépek nem tették lehetővé nagy adatmátrixok gyors elemzését, mondjuk a PCA segítségével. Lényege az, hogy a távolságmátrix alapján kiválasztjuk a két egymástól legtávolabbi objektumot, s ezt tekintjük az első ordinációs tengely két pólusának. Feltételezzük ugyanis, hogy a vizsgált közösségekre nézve ezek jelentik valamely ökológiai háttérgrádiens végpontjait. Az összes többi objektum közbülső helyét a két végpont-objektumhoz való relatív hasonlóság határozza meg. Ezután egy második ordinációs tengely is megkapható a második legtávolabbi objektumpár kiválasztásával. Részletesebb leírást Gauch (1982) könyvében találhatsz, de nem nagyon biztatlak a keresgélésre, mert már magam is rendkívül elavultnak tekintem a módszert. A **NuCoSA** programcsomagban egyébként benne van, ha ki akarod próbálni.

K: *Van még más is, amiről nem ejtettél szót?*

V: Hogyne, bőven. Gauss-ordináció, maximum likelihood-ordináció és még sorolhatnám, de ezekről már nemigen szólhatok részletesebben, mert akkor sosem érnénk a könyv végére.

K: *Ha már volt fuzzy osztályozás, akkor van-e fuzzy ordináció?*

V: Nem tudom, hogy jutott eszedbe ez a kérdés, nyilván az előző fejezetekből "extrapolálsz", mint eddig sokszor. Gondolom úgy véled, hogy egy fuzzy ordinációban a pont pozíciója lesz bizonytalan (amennyire az objektum osztályba tartozása a bizonytalan a fuzzy osztályozások-

ban). Ilyen értelemben azonban nincs olyan módszer, amely közvetlenül fuzzy ordinációt adna, de a 9. fejezetben majd említendő konszenzus ordináció (9.18 ábra) akár fuzzy ordinációként is értelmezhető. Van azonban lehetőség ordinációt szerkeszteni fuzzy alapokon. Roberts (1986) javasolta először, hogy fuzzy halmazokból kiindulva állítsunk elő ordinációt, de ez az ordináció bizonyos értelemben "direkt", mert rendelkeznünk kell a fajok és a környezet kapcsolatáról szóló adatokkal v. legalábbis feltételezésekkel.

Továbbmenve: ordináció lehetséges osztályozásból is! Olvasd el Feoli & Zuccarello (1986) nálunk is könnyen hozzáférhető cikkét ebben a témában!

K: *Rövid kérdés: ordináció vagy klasszifikáció?*

V: Igen, volt idő, amikor ez valóban kérdés volt, például a növényökológusok körében. Gondoljunk az elhíresült kontinuum vitára a 60-as évek végéről, amelyben a klasszifikáció és az ordináció hívei "veszekedtek", hogy melyik az előbb való. Ma már nyugodtan mondhatjuk, hogy az osztályozás és az ordináció együttes alkalmazása többet mond az adatstruktúráról, mint bármelyikük külön-külön. Ha mindenáron meg akarod állapítani, hogy mégis melyik legyen az elsődleges, akkor azt mondanám, hogy sose osztályozzunk ordináció nélkül, míg az ordináció jól megvan klasszifikációs ellenőrzés nélkül is.

K: *Mondd, nem akarsz egy rövid döntési kulcsot is mellékelni az ordinációs módszerek kiválasztására, hasonlóan a 3. fejezetbeli kulcshoz? Ezzel megkönnyítenéd a kezdő felhasználó dolgát.*

V: Megpróbálhatjuk, bár a legfontosabb lépéseket már a 0.1 ábra is bemutatta. Nos, íme egy bővített kulcs, amely persze csak egy a lehetségesek közül:

1a Az objektumokat vagy a változókat eleve csoportokba osztjuk (kanonikus módszerek)..... **2**
1b Semmiféle *a priori* csoportosítás nincs **5**
2a Az objektumok 2 vagy több csoportba vannak beosztva. A változók egységes halmazt képviselnek ..... *Diszkriminancia elemzés*
2b A változók két csoportot alkotnak, az objektumok egyet **3**
3a A változók csoportjai közötti viszony szimmetrikus, egyikük sem kitüntetett ..... *Kanonikus korreláció elemzés*
3b A változók 1. csoportja megszabja a 2. csoport szerinti ordinációt (kötött ordináció) **4**
4a A 2. csoport változói között lineáris a kapcsolat ..... *Redundancia elemzés*
4b A 2. csoport változói unimodális reakciót adnak a háttérgrádiensre ..... *Kanonikus korreszpondencia elemzés*
5a Az elemzett objektumok távolság- (különbözőség-) mátrixa áll csupán rendelkezésünkre, ill. ha az eredeti adatok is megvannak, a változók ordinációja most mellékes **6**
5b Az eredeti nyers adatok is megvannak, és az objektumok és változók ordinációja egyaránt lényeges számunkra **7**
6a Az ordinációban megtartjuk a metrikus információt ..... *Főkoordináta módszer*
6b A metrikus információ elvész, csak a távolságértékek sorrendisége lényeges ..... *Nem-metrikus többdimenziós skálázás*
7a Az összvariancia közös részét magyarázzuk csupán ..... *Faktoranalízis*
7b A teljes variancia megmagyarázására törekszünk **8**
8a Az adatstruktúra – közelítőleg – lineáris ..... *Főkomponens elemzés*

8b Az adatstruktúra unimodális jellegű, gyakorisági adataink vannak . *Korreszpondencia elemzés*

Természetesen a döntéshez olyasmi is kell, amire csak az elemzés közben derül fény, ezért a fenti kulcs semmiképpen sem helyettesítheti az értelmes, többirányú vizsgálódást.

K: *Úgy tűnik számomra, hogy a patkó-jelenség kizárólagosan csak ökológiai ordinációkban, hosszú, gyors fajcserékkel jellemzhető háttér-grádiensek esetében "fenyeget". Egy taxonómusnak vagy morfológusnak tényleg nem kell tartania ettől?*

V: A patkó-jelenség természetesen nemcsak az ökológiai adatok ordinációjának lehetséges kísérő jelensége. Reyment (1991: 51) be is mutat egy példát, amelyben *Leptograpsus* rákok egyedeinek főkoordináta-ordinációja produkál egy csaknem tökéletes parabola-menti elrendeződést. Ennek a Reyment-féle magyarázata ("a majdnem egyenlő változók közötti nagyon magas korrelációk") nem világosít fel bennünket az okokról. Az elemzést megismételtem többféle módszerrel is, és a patkó-jelenség csak akkor adódott, ha Manhattan-metrikával hasonlítottam össze az egyedeket, más esetben nem (a Reyment által alkalmazott Gower index "kvantitatív" esetre valójában Manhattan metrika, vö. 3.6 rész). Euklidészi távolságból pl. egyáltalán nem ilyen, hanem a várt eredmény jött ki, s bevallom, a jelenség magyarázatával még tartozom. A dolog annál is inkább "zavaró", mert ugyanazon Manhattan-távolságmátrixból a nem-metrikus többdimenziós skálázás távolról sem patkó-szerű elrendeződést, hanem méretbeli sorrendet adott. Tapasztalataim szerint egyébként akkor is kaphatunk patkó-szerű elrendeződést, amikor az adatmátrix sorainak vagy oszlopainak az összege konstans (ez a statisztikában "záródás" vagy *closure* néven ismert). Ha például a változók összege minden egyes objektumra 100-at ad (vagyis objektum-szerinti százalékokról van szó, mondjuk talajminták százalékos anyagtartalma), akkor a változók ordinációjában figyelhetjük meg gyakran – nem mindig – az ívet. Ez fordítva is igaz: amikor az objektumok összege ad 100-at minden egyes változóra, az objektumok kerülnek a patkóra. A Reyment-féle illusztrációban pedig a változók értéktartomány-szerinti standardizálása szerepel, s ez közelítően konstans összegre vezethetett minden változóra. Azt hiszem ezt a problémát érdemes lenne jobban megvizsgálni.

K: *És térsor vajon elképzelhető-e az ordinációs térben?*

V: A kérdést már vártam, s bizonnyal nem lepődsz meg nagyon, ha válaszom igenlő. Nemcsak olyan térsorra gondolok persze, amelyet az ordinációs térben voltaképpen egy adattranszformációs függvény, vagy mondjuk a mintavételi feltételek fokozatos megváltoztatása generál, hanem olyanra is, amelyet magának az ordinációs módszernek a szukcesszív változtatása hoz létre. Vagyis a primer sor maga is ordinációs. Láthattad a korreszpondencia elemzésről szóló részben, hogy α értéke bizonyos korlátok között szabadon módosítható. Nos, α függvényében egy ordináció-sorozatot készíthetünk, amely megint csak jobban tükrözi az adatok szerkezetét, mint α bármely, önkényesen kiragadott értéke. Hasonlóan változtatható paraméter szerepel pl. Noy-Meir (1974) "*catenation*" módszerében. A biplot szerkesztése is elképzelhető egy térsor mentén. Gondolj vissza az euklidészi és Mahalanobis biplotra, amelyek Jackson (1991) javaslata szerint csupán két szélső esetei egy "biplot grádiens"- nek, melynek végtelen számú esetei egy hatványkitevő változtatásával egyszerűen előállíthatók.

K: *A morfometriai ordinációról szóló rész mintha egy kicsit kilógna ebből a fejezetből, ugyanis itt szinte több szó esik az újszerű adattípusokról (kontúr, mérőpontok, stb.), mint*

magáról az ordinációról. Az itt alkalmazott ordinációs módszerek voltaképpen ugyanazok, amelyekről az előző fejezetrészek szólnak!

V: Ebben van némi igazad, de úgy éreztem, hogy a könyv olvasása közben – feltéve, ha akad valaki, aki e kötetet szabályosan, oldalról oldalra tanulmányozza végig –, az Olvasó ekkor "érik meg" igazán a téma befogadására. Az osztályozás és különösképpen az ordináció módszereit ismerve viszont már jöhet a "nehezebb falat". Az alak leírása önmagában a 2. fejezetben még eléggé érdektelen lett volna.

K: *Mi van akkor, ha nekem nem szokványos fajok × helyek mátrixom van, hanem egy harmadik szempontom, – vagy dimenzióm? – is van. Például évek munkájával összeállítottam egy fajok × helyek × időpontok táblázatot, s ezt szeretném részletesen kiértékelni. Úgy tűnik, erről mintha megfeledkeztél volna, holott a biológiában egyáltalán nem lehet ritka az ilyen típusú adathalmaz.*

V: Nos, megfogtál, mert erre eddig valóban nem gondoltam. A háromutas (*three-way* vagy *three-mode*, de persze nem "három-dimenziós") "mátrixok" vagy inkább "tömbök" elemzésére többféle lehetőség adódik. Először is természetes, hogy a tömböt valamelyik szempont szerint kétutas "szeletek"-re bonthatod, s ezek a már ismert módon elemezhetők. Ha pl. az idő szerint osztod fel a kiinduló adatokat *p* számú mátrixra, akkor az egyes időpontokra kapott "sima" ordinációs elemzéseket összehasonlíthatod egymással (lásd majd a 9. fejezetet), s ebből az összehasonlításból hámozhatod ki az időbeli trendeket. Azt is teheted – bár ez eléggé "quick-and-dirty" ("gyors és nem igazán matematikai") eljárás –, hogy minden egyes kétutas szeletet egy vektorrá "nyújtasz ki", s ezeket a vektorokat egy új "adatmátrixba" egyesítve hajtod végre az ordinációs értékelést (erre példa volt a "relative warp analysis"). Sokkal jobb persze a kifejezetten ilyen céllal kidolgozott nem-metrikus INDSCAL módszer (Carroll & Chang 1970), a faktor-analízis háromutas kiterjesztése (PARAFAC; Harshman 1970, Tucker 1972), a korreszpondencia-elemzés háromutas verziója (Carlier & Kroonenberg 1996). Természetesen a PCA-nak is megvan a háromutas megfelelője (Kroonenberg 1983). Ha ezeket a cikkeket megkeresed, persze gondban lehetsz, mert a terminológia a legkevésbé sem biológiai. Ennek ellenére ajánlhatom ezeket a komolyabb utánaolvasás céljára. Nekem azonban már nincs helyem és időm a probléma – és a lehetséges megoldások – további részletezésére.