æ

# TÖBBVÁLTOZÓS STATISZTIKAI MÓDSZEREK

A többdimenziós normális eloszlás kulcsszerepet játszik itt, bár a legtöbb módszer akkor is alkalmazható, ha a háttéreloszlás többdimenziós folytonos és csak a változók páronkénti kovarianciájára szorítkozunk.

*Definíció.* Az $\mathbf{X} \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$ véletlen vektor $p$-dimenziós normális eloszlású $\mathbf{m}$ várható érték vektorral és $\mathbf{C}$ kovarianciamátrixszal, ha sűrűségfüggvénye

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{C}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})}, \quad \mathbf{x} \in \mathbb{R}^p.$$

## Főkomponensanalízis

Legyen $\mathbf{X} \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$, és tegyük fel, hogy a $\mathbf{C}$ kovarianciamátrix pozitív definit. A modell a következő:

$$\mathbf{X} = \mathbf{V}\mathbf{Y} + \mathbf{m},$$

ahol $\mathbf{m} = \mathbb{E}\mathbf{X}$, $\mathbf{V}$ $p \times p$-s ortogonális mátrix (azaz $\mathbf{V}^{-1} = \mathbf{V}^T$), $\mathbf{Y}$ pedig független komponensű, $p$-dimenziós normális eloszlású véletlen vektor.

Megadjuk a fenti előállítást. Mivel $\mathbf{V}$ invertálható, ezért

$$\mathbf{Y} = \mathbf{V}^{-1}(\mathbf{X} - \mathbf{m}) = \mathbf{V}^T(\mathbf{X} - \mathbf{m}).$$

Jelölje $\mathbf{C} = \mathbf{U}\Lambda\mathbf{U}^T$ az $\mathbf{X}$ véletlen vektor kovarianciamátrixának spektrálfelbontását. Ezzel $\mathbf{Y}$ kovarianciamátrixának diagonálisnak kell lennie. A spektrálfelbontás egyértelműsége értelmében

$$\begin{aligned}\mathbb{E}\mathbf{Y}\mathbf{Y}^T &= \mathbb{E}\left[\mathbf{V}^{-1}(\mathbf{X}-\mathbf{m})(\mathbf{X}-\mathbf{m})^T\mathbf{V}\right] = \mathbf{V}^{-1}\mathbb{E}\left[(\mathbf{X}-\mathbf{m})(\mathbf{X}-\mathbf{m})^T\right]\mathbf{V} = \\ &= \mathbf{V}^{-1}\mathbf{C}\mathbf{V} = \mathbf{V}^{-1}\mathbf{U}\Lambda\mathbf{U}^T\mathbf{V} = (\mathbf{V}^{-1}\mathbf{U})\Lambda(\mathbf{V}^{-1}\mathbf{U})^T\end{aligned}$$

diagonális mátrix fődiagonálisában csökkenő elemekkel akkor és csak akkor, ha $\mathbf{V}^{-1}\mathbf{U} = \mathbf{I}_p$, azaz $\mathbf{V} = \mathbf{U}$. (Itt kihasználtuk, hogy $\mathbf{V}$, $\mathbf{U}$, következésképpen $\mathbf{V}^{-1}\mathbf{U}$ is ortogonális mátrix.) Megjegyezzük, hogy többszörös multiplicitású sajátértékek esetén az $\mathbf{U}$ mátrix megfelelő oszlopai sem egyértelműek. Így

$$\mathbf{X} = \mathbf{U}\mathbf{Z} + \mathbf{m}$$

lesz a kívánt felbontás, ahol $\mathbf{Z}$ jelöli a $\mathbf{V} = \mathbf{U}$ választás melletti $\mathbf{Y}$-t, azaz

$$\mathbf{Z} = \mathbf{U}^{-1}(\mathbf{X} - \mathbf{m}) = \mathbf{U}^T(\mathbf{X} - \mathbf{m}).$$

Ezt a $\mathbf{Z}$-t az $\mathbf{X}$ véletlen vektor *főkomponensvektor*ának, komponenseit pedig *főkomponensek*nek nevezzük.

Vegyük észre, hogy a $k$-adik főkomponens az $\mathbf{X}-\mathbf{m}$ változó komponenseinek az $\mathbf{u}_k$ vektor koordinátáival vett lineáris kombinációja:

$$Z_k = \mathbf{u}_k^T(\mathbf{X}-\mathbf{m}) \qquad (k=1,\ldots,p),$$

ahol $\mathbf{u}_k$ a $\mathbf{C}$ mátrix $\lambda_k$ sajátértékéhez tartozó normált sajátvektora ($\mathbf{U}$ $k$-adik oszlopa), és $Z_k$ varianciája $\lambda_k$; $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.

$\mathbf{X}$ fenti felbontása eleget tesz az alább ismertetendő optimalitási kritériumnak (a főkomponenseket ezzel is be lehetne vezetni): Az első főkomponens, $Z_1$ szórása maximális az $\mathbf{X}-\mathbf{m}$ véletlen vektor komponenseinek összes lehetséges normált (egységvektorral képzett) lineáris kombinációéi között; $Z_2$ szórása maximális az összes lehetséges, $Z_1$-től független normált lineáris kombinációéi közt; s.í.t. a $k$-adik főkomponens, $Z_k$ szórása maximális az összes lehetséges, $Z_1,\ldots,Z_{k-1}$-től független normált lineáris kombináció szórása közt ($k=3,\ldots,p$).

A főkomponens modellhez nem kell szükségképpen feltenni a többdimenziós normalitást. Tetszőleges $p$-dimenziós folytonos eloszlású véletlen vektor $\mathbf{C}$ pozitív szemidefinit kovarianciamátrixára elvégezhető a $\mathbf{C}=\mathbf{U}\Lambda\mathbf{U}^T$ spektrálfelbontás, a $\mathbf{Z}=\mathbf{U}^T(\mathbf{X}-\mathbf{m})$ véletlen vektorról pedig csak annyi mondható el, hogy koordinátái korrelálatlanok (normális eloszlás esetén ez a függetlenséget is jelentette), varianciáik pedig $\mathbf{C}$ sajátértékeivel egyeznek meg, csökkenő sorrendben. Így a "függetlenség" szó helyébe tehát mindenütt a "korrelálatlanság" lép. A valóságban persze a felbontást az $\mathbf{X}$ véletlen vektor empirikus kovarianciamátrixán végezzük el, nem törődve azzal, hogy milyen eloszlásból származik a mintánk. Megjegyezzük azonban, hogy diszkrét eloszlásokra a módszer nem javasolt, hanem ott hasonló céllal más eljárásokat vezethetünk be.

A fentiek alapján a főkomponens transzformáció egyben azt is jelenti, hogy ha az $\mathbf{u}_1,\ldots,\mathbf{u}_p$ sajátvektorok alkotta bázisra térünk át, akkor ezekben az irányokban a transzformált változó varianciája maximális.

Amennyiben adott egy $n$-elemű minta, azaz $n$ pont a $p$-dimenziós térben ($n>p$), a főkomponensanalízis szemléletesen a következőképpen képzelhető el: $\mathbf{u}_1$ az az irány, amely mentén a mintaelemek szóródása a legnagyobb. Ha az $\mathbf{u}_1,\ldots,\mathbf{u}_{k-1}$ irányokat már megkonstruáltuk, akkor az $\mathbf{u}_k$ irány olyan lesz, amely merőleges az előzőekre, továbbá emellett az ortogonalitási feltétel mellett ebben az irányban a mintaelemek szóródása a legnagyobb. Grafikusan lehetséges a mintapontok kirajzolása az első néhény főkomponenspár síkjában 2-dimenziós ábrákon.

## Faktoranalízis

A főkomponensanalízisnél láttuk, hogy a módszer alkalmas a változók számának csökkentésére. A faktoranalízis célja eleve ez: nagyszámú korrelált változó magyarázata kevesebb korrelálatlannal (többdimenziós normális eloszlás esetén a korrelálatlan helyett független mondható). Ezek a *közös faktorok* azonban nem magyaráznak meg mindent a változókból, csak azoknak az ún. "közös részét". Ezen kívül van a változóknak egy "egyedi része" is, amelynek leválasztása szintén a modell feladata. A közös faktorokra itt nem úgy kell gondolni, mintha közvetlenül megfigyelhető változók lennének.

A $k$-faktor modell tehát a következő. Adott a $p$-dimenziós $\mathbf{X}$ véletlen vektor $\mathbf{m}$ várható érték vektorral és $\mathbf{C}$ kovarianciamátrixszal, többdimenziós normalitás

esetén $\mathbf{X} \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$. Adott $k$ $(1 \leq k < p)$ egészre keressük az

$$\mathbf{X} = \mathbf{A}\mathbf{f} + \mathbf{e} + \mathbf{m}$$

felbontást, ahol $\mathbf{A}$ $p \times k$-as mátrix, az $\mathbf{f}$ *közös faktor* $\mathbf{0}$ várható érték vektorú, korrelálatlan komponensű, $k$-dimenziós véletlen vektor, komponensei 1 szórásúak, az $\mathbf{e}$ *egyedi faktor* $p$-dimenziós korrelálatlan komponensű véletlen vektor, ráadásul komponensei még $\mathbf{f}$ komponenseivel is korrelálatlanok. A modell feltevései formálisan:

$$\begin{aligned} \mathbb{E}\mathbf{f} = \mathbf{0}, \quad & \mathbb{E}\mathbf{f}\mathbf{f}^T = \mathbf{I}_k, \\ \mathbb{E}\mathbf{e} = \mathbf{0}, \quad & \mathbb{E}\mathbf{e}\mathbf{e}^T = \mathbf{D} \quad \text{diagonális mátrix}, \\ \mathbb{E}\mathbf{f}\mathbf{e}^T = \mathbf{0} \quad & \text{a} \quad k \times p\text{-es azonosan 0 mátrix}. \end{aligned}$$

Koordinátákra lebontva ez a következőt jelenti:

$$X_i = \sum_{j=1}^{k} a_{ij} f_j + e_i + \mu_i, \quad i = 1, \ldots, p.$$

Mivel $e_i$ és $f_j$ korrelálatlanok, $X_i$ varianciája

$$c_{ii} = \sum_{j=1}^{k} a_{ij}^2 + d_{ii},$$

ahol $d_{ii}$ a $\mathbf{D}$ diagonális mátrix $i$-edik diagonális eleme nem más, mint az $e_i$ változó ($i$-edik egyedi faktor) varianciája. Tehát $X_i$ varianciájából a $\sum_{j=1}^{k} a_{ij}^2$ részt magyarázzák a közös faktorok – ezt nevezzük az $X_i$ változó *kommunalitás*ának –, $d_{ii}$ pedig az *egyedi variancia*.

A modell paraméterei az $\mathbf{A}$ és $\mathbf{D}$ mátrixok. Az $\mathbf{A}$ mátrixot *faktorsúly-mátrix*nak (más terminológiával átviteli mátrixnak) nevezzük. Ezekkel a modell mátrixalakja a következő:

$$\mathbf{C} = \mathbf{A}\mathbf{A}^T + \mathbf{D}.$$

Látható, hogy $\mathbf{X}$ tetszőleges átskálázás után is leírható a $k$-faktor modellel, ugyanis

$$\mathbf{S}\mathbf{X} = (\mathbf{S}\mathbf{A})\mathbf{f} + \mathbf{S}\mathbf{e} + \mathbf{S}\mathbf{m}$$

teljesíti a modell feltételeit. Az is látható, hogy az $\mathbf{A}$ faktorsúly-mátrix sorainak tetszőleges elforgatása után (azaz az $\mathbf{A}\mathbf{O}$ transzformáció után is, ahol $\mathbf{O}$ $k \times k$-as ortogonális mátrix) faktorsúly-mátrix marad a fenti modellben.

Még adott $k$ esetén is nehéz megtalálni a fenti felbontást. Az egyértelműség kedvéért szokás ezen kívül még további kényszerfeltételeket tenni az $\mathbf{A}$ mátrixra. Például többdimenziós normális eloszlású $\mathbf{X}$, $\mathbf{f}$, $\mathbf{e}$ esetén a $k$-faktor modell paramétereinek maximum likelihood becslését keresve fel szokták tenni, hogy a $\mathbf{C}$ kovarianciamátrix nem-szinguláris, az

$$\mathbf{A}^T\mathbf{D}^{-1}\mathbf{A}$$

mátrix pedig diagonális, diagonális elemei különbözőek, és nem-csökkenő sorrendbe vannak rendezve. Ez a feltétel bizonyos egyértelműséget biztosít a faktorok maximum likelihood becsléséhez, és a számolásokat is egyszerűbbé teszi.

A faktorok számát, $k$-t "kicsire" célszerű választani. Kérdés azonban, hogy milyen $k < p$ természetes számokra írható le az $n$-dimenziós $\mathbf{X}$ véletlen vektor a $k$-faktor modellel. Ehhez számoljuk össze a modell paramétereit: $\mathbf{A}$-ban és $\mathbf{D}$-ben összesen $pk+p$ ismeretlen paraméter van, a kényszerfeltétel azonban a diagonálison kívüli elemek 0 voltára vonatkozón $(1/2)(k^2-k) = (1/2)k(k-1)$ egyenletet jelent (ez megegyezik a $k \times k$-as forgatások szabad paramétereinek számával). Alapvetően pedig van $(1/2)p(p+1)$ egyenletünk (a $\mathbf{C}$ kovarianciamátrix különböző elemei a szimmetria miatt). A felírható egyenletek és a szabad paraméterek számának különbsége:

$$s = (1/2)p(p+1) + (1/2)k(k-1) - (pk+p) = (1/2)(p-k)^2 - (p+k).$$

Általánosságban $s \leq 0$ esetén várható az egyenlet algebrai megoldásának létezése. Ekkor

$$k \geq (2p+1-\sqrt{8p+1})/2.$$

A faktormodell identifikálhatóságán azt értjük, hogy rögzített $k$ esetén egyértelműen meg tudjuk adni $\mathbf{D}$-t és $\mathbf{A}$-t.

Adott $k < p$ természetes szám esetén a $\mathbf{C} = \mathbf{A}\mathbf{A}^T + \mathbf{D}$ egyenlet pontosan akkor oldható meg, ha van olyan $p \times p$-s diagonális $\mathbf{D}$ mátrix (fődiagonálisában nemnegatív elemekkel), hogy a $\mathbf{C} - \mathbf{D}$ mátrix pozitív szemidefinit és rangja nem nagyobb $k$-nál. Ez az állítás azonban nem sok gyakorlati jelentőséggel bír, a gyakorlatban numerikus eljárások használhatók. Az SPSS-ben többféle módszer közül lehet választani, a főkomponensanalízis mint ezek egyike van feltüntetve.

A végén a faktorsúlymátrixot tetszőlegesen elforgatva (legyen ehhez $\mathbf{O}$ $k \times k$-as ortogonális mátrix), az $\mathbf{AO}$ mátrixszal, mint faktorsúlymátrixszal a $k$-faktor modell továbbra is fennáll, a faktorsúlyok pedig esetleg jobban reprezentálhatók a gyakorlatban (jobban mutatják, hogy melyik faktorban melyik változó játszik szerepet). A faktorok *rotáció*jára különböző módszerek állnak rendelkezésünkre, például a VARIMAX módszer, ami sok 0-hoz közeli és viszonylag kevés $\pm 1$-hez közeli faktorsúlyt eredményez. A faktoroknak ilyen módon a felhasználó konkrét jelentést tud tulajdonítani. Például ha változóink gépkocsik különböző műszaki jellemzői, akkor faktora lehet például a teljesítőképességnek, gyorsíthatóságnak, stb.

## Többváltozós regresszióanalízis

A többváltozós regressziós problémában az $Y$ valószínűségi változót (*függő változó*) szeretnénk az $X_1, \ldots, X_p$ valószínűségi változók (*független változók*) függvényével közelíteni legkisebb négyzetes értelemben. Amennyiben ismerjük az $Y, X_1, \ldots, X_p$ véletlen vektor együttes eloszlását (tegyük fel, hogy ez abszolút folytonos, az együttes sűrűségfüggvényt jelölje $f(y, x_1, \ldots, x_p)$), akkor

$$\mathbb{E}(Y - g(X_1, \ldots, X_p))^2$$

minimumát a $p$-változós $g$ függvények körében $Y$-nak az $X_1, \ldots, X_p$ változók adott értéke mellett vett feltételes várható értéke szolgáltatja:

$$(3.1) \quad g_{opt}(x_1, \ldots, x_p) = \mathbb{E}(Y | X_1 = x_1, \ldots, X_p = x_p) = \frac{\int_{-\infty}^{\infty} y f(y, x_1, \ldots, x_p) dy}{\int_{-\infty}^{\infty} f(y, x_1, \ldots, x_p) dy},$$

ezt nevezzük regressziós függvénynek.

Adott $f$ sűrűségfüggvény mellett sem mindig triviális a fenti integrál kiszámolása, általában azonban $f$ nem adott, csak egy statisztikai mintánk van a függő és független változókra az $(Y^{(m)}, X_1^{(m)}, \ldots, X_p^{(m)})$, $(m = 1, \ldots, n)$ független, $(p+1)$-dimenziós megfigyelések formájában. A legegyszerűbb ilyenkor a fenti minimumot a lineáris függvények körében keresni, ezt nevezzük *lineáris regresszió*nak. Erre az esetre vezethető vissza olyan függvényekkel való közelítése $Y$-nak, amely az $X_i$ változók lineáris függvényének monoton (például exponenciális, logaritmikus) transzformációja. Ilyenkor az inverz transzformációt alkalmazva $Y$-ra, az így kapott új függő változón hajtunk végre lineáris regressziót az eredeti független változók alapján.

A másik érv a lineáris regresszió mellett az, hogy amennyiben $Y, X_1, \ldots, X_p$ együttes eloszlása $(p+1)$-dimenziós normális, akkor a feltételes várható érték vevés valóban lineáris függvényt ad megoldásul. A többdimenziós normalitás pedig a centrális határeloszlás tételre hivatkozva elég általánosan feltehető, vagy legalábbis közelíthető vele eloszlásunk.

Harmadik érv: még akkor is, ha eloszlásunk nem közelíthető normálissal, de abszolút folytonos, előfordulhat, hogy pusztán a változók között második momentumokra akarunk hagyatkozni, azaz az együttes kovarianciával akarunk csak dolgozni (amely mintánkból becsülhető). Ilyenkor is alkalmazható az alább ismertetendő módszer, hiszen ennek is – a főkomponens- és faktoranalízishez hasonlóan – az a sajátossága, hogy csak a második momentumig bezárólag használ momentumokat.

Térjünk rá tehát a lineáris regresszióra. A legjobb

$$Y \sim l(\mathbf{X}) = a_1 X_1 + \cdots + a_p X_p + b$$

lineáris közelítést keressük legkisebb négyzetes értelemben, azaz minimalizálni akarjuk az

$$\mathbb{E}(Y - (a_1 X_1 + \cdots + a_p X_p + b))^2$$

kifejezést az $a_1, \ldots, a_p$ és $b$ együtthatókban általában egy $n$-elemű minta alapján $(n \geq p)$.

Az egyváltozós esethez hasonlóan megmutatható, hogy az $a_1, \ldots, a_p$ együtthatók megoldásai a

$$\mathbf{Ca} = \mathbf{d}$$

egyenletnek, ahol $\mathbf{a} = (a_1, \ldots, a_p)^T$, $\mathbf{C}$ jelöli a $p \times p$-s empirikus kovarianciamátrixot, a $\mathbf{d} \in \mathbb{R}^p$ vektor pedig az $Y$ változónak $\mathbf{X}$ komponenseivel vett empirikus (kereszt)kovarianciáit tartalmazza. A fenti lineáris egyenletrendszernek létezik egyértelmű megoldása, ha $|\mathbf{C}| \neq 0$, ami teljesül, ha az $X_1, \ldots, X_p$ változók között nincsen lineáris kapcsolat. A megoldás:

$$\hat{\mathbf{a}} = \mathbf{C}^{-1}\mathbf{d} \quad \text{és} \quad \hat{b} = \bar{Y} - \hat{\mathbf{a}}^T \bar{\mathbf{X}}.$$

Ezután hipotéziseket vizsgálhatunk a regresszió "jóságára". Ha a $H_0 : \mathbf{a} = \mathbf{0}$ hipotézist szignifikánsan el tudjuk utasítani, akkor értelmes a regresszió. Mindezt $F$-próbával tesszük az $Y$ minta teljes szóródásának $(\sum_{i=1}^n (Y_i - \bar{Y}_i)^2)$ felbontásával a regresszió által képviselt és a véletlen hiba $(\sum_{i=1}^n (Y_i - \hat{a}_1 X_{i1} - \cdots - \hat{a}_p X_{ip} - \hat{b})^2)$ okozta részre, varianciaanalízisbeli technikákkal.

Az $l(\mathbf{X}) = a_1X_1+\cdots+a_pX_p+b$ jelöléssel $Y$ és $l(\mathbf{X})$ korrelációját $Y$ és $(X_1,\ldots,X_p)$ *többszörös korrelációjának* nevezzük. Ha az $Y_1$ és $Y_2$ független változókat ugyanazon $\mathbf{X}$ független változókkal lineárisan közelítjük, akkor az $Y_1 - l_1(\mathbf{X})$ és $Y_2 - l_2(\mathbf{X})$ változók korrelációját $Y_1$ és $Y_2$ *parciális korrelációjának* nevezzük, miután $(X_1,\ldots,X_p)$ ún. hatását belőlük kiküszöböltük.

## Varianciaanalízis

Gyakorlati alkalmazásokban olyan mintákat vizsgálunk, melyeket különböző körülmények közt figyeltünk meg, és célunk éppen annak a megállapítása, vajon ezek a körülmények jelentősen befolyásolják-e a mért értékeket. Tehát mintánkat eleve csoportokba osztottan kapjuk, feltesszük azonban, hogy a különböző csoportokban felvett minták egymástól függetlenek, normális eloszlásúak és azonos szórásúak.

Például több gépen, vagy többféle technológiával gyártott alkatrészek valamilyen mérhető jellemzőjét vizsgáljuk, és az érdekel bennünket, vajon a gyártó gép vagy a gyártási technológia befolyásolja-e az alkatrész mért tulajdonságát. Ha egyszerre mindkét hatás, esetleg azok kölcsönhatása is érdekel bennünket, akkor kétszempontos varianciaanalízisről beszélünk, ha külön vizsgáljuk az egyes tényezők hatását, akkor egyszempontos a varianciaanalízis. Természetesen bevezethetnénk további szempontokat is. A szempontok száma szerint ismertetjük a legfontosabb modelleket.

*Egyszempontos varianciaanalízis*

Valamilyen szempont alapján (például különböző kezelések) $k$ csoportban külön végzünk megfigyeléseket. Az egyes csoportokban a mintaelemek száma általában nem egyenlő: jelölje $n_i$ az $i.$ csoportbeli mintaelemek számát, $n = \sum_{i=1}^{k} n_i$ pedig az összminta elemszámát. Az $i.$ csoportban az $X_i \sim \mathcal{N}(b_i, \sigma^2)$ valószínűségi változóra vett mintaelemeket

$$X_{ij} \sim \mathcal{N}(b_i, \sigma^2), \quad (j = 1,\ldots,n_i)$$

jelöli. Ezek egymás közt és különböző $i$-kre is függetlenek, azonos szórásúak. A várható értékekre a $b_i = m + a_i$ felbontást alkalmazzuk, ahol $m$ a várható értékek súlyozott átlaga, $a_i$ pedig az $i.$ csoport hatása:

$$m = \frac{1}{n}\sum_{i=1}^{k} n_i b_i, \qquad a_i = b_i - m \quad (i = 1,\ldots,k).$$

Könnyen látható, hogy

$$\sum_{i=1}^{k} n_i a_i = 0.$$

Ezekkel a jelölésekkel az egyszempontos modell

$$X_{ij} = m + a_i + \varepsilon_{ij} \qquad (j = 1,\ldots,n_i;\ i = 1,\ldots,k)$$

alakban írható, ahol az $\varepsilon_{ij} \sim \mathcal{N}(0,\sigma^2)$ független valószínűségi változók véletlen hibák.

Vezessük be a csoportátlagokra ill. a teljes mintaátlagra az

$$\bar{X}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \quad (i = 1, \ldots, k) \quad \text{ill.} \quad \bar{X}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}$$

jelöléseket! Belátható, hogy a paraméterek legkisebb négyzetes becslései

$$\hat{m} = \bar{X}_{..} \quad \text{és} \quad \hat{a}_i = \bar{X}_{i.} - \bar{X}_{ii} \quad (i = 1, \ldots, k)$$

lesznek.

A gyakorlati alkalmazók terminológiájával élve: a fenti kvadratikus alakok segítségével a mintaelemek teljes mintaátlagtól vett eltéréseinek négyzetösszega ($Q$) felbomlik csoportok közötti (between, $Q_a$) ill. csoportokon belüli (within, $Q_e$) részre a következőképpen:

$$\begin{aligned} Q &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}[(X_{ij} - \bar{X}_{i.}) + (\bar{X}_{i.} - \bar{X}_{..})]^2 = \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_{i.} - \bar{X}_{..})^2 = \\ &= \sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2 = Q_a + Q_e. \end{aligned}$$

A fenti felbontásokat az alábbi ún. **ANOVA** (ANalysis Of VAriances) táblázatban foglaljuk össze.

| A szóródás oka | Négyzetösszeg | Szabadsági fok | Empirikus szórásnégyzet |
|---|---|---|---|
| Csoportok között | $Q_a = \sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2$ | $k-1$ | $s_a^2 = \frac{Q_a}{k-1}$ |
| Csoportokon belül | $Q_e = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2$ | $n-k$ | $s_e^2 = \frac{Q_e}{n-k}$ |
| Teljes | $Q = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{..})^2$ | $n-1$ | - |

A fenti modellben először az $m = 0$ hipotézist teszteljük. Ha ezt elutasítjuk (az összes várható érték nem 0, azaz van ún. főhatás), akkor a

$$H_0 : a_1 = \cdots = a_k = 0, \quad \text{tömören} \quad \mathbf{a} = \mathbf{0}$$

hipotézist vizsgáljuk. Bevezetve az

$$s_a^2 = \frac{Q_a}{k-1} \qquad \text{ill.} \qquad s_e^2 = \frac{Q_e}{n-k}$$

kifejezéseket, ezek azonos ($\sigma^2$) szórásúak, függetlenek, hányadosuk pedig $H_0$ fenállása esetén $F$-eloszlást követ $k-1$ ill. $n-k$ szabadsági fokkal:

$$F = \frac{s_a^2}{s_e^2} = \frac{Q_a}{Q_e} \cdot \frac{n-k}{k-1} \sim \mathcal{F}(k-1, n-k).$$

és ez az $F$ is szigorúan monoton csökkenő függvénye a likelihood hányados statisztikának.

Tehát $H_0$ tesztelésére $F$-próba alkalmazható, mellyel tulajdonképpen arról döntünk, hogy a csoportok közötti eltéréseket mérő $s_a^2$ szignifikánsan nagyobb-e, mint a csoportokon belüli ingadozásokat mutató $s_e^2$ (utóbbi ingadozásokat csak a véletlen eltérések hozzák létre). Ha a fenti $F$-statisztika nagyobb vagy egyenlő, mint az $\mathcal{F}(k-1, n-k)$-eloszlás $1-\varepsilon$ szinthez tartozó kritikus értéke, akkor $1-\varepsilon$ biztonsággal ($\varepsilon$ szignifikanciával) elutasítjuk $H_0$-t, azaz az $a_i$ várható értékek között van olyan, ami nem egyenlő a többivel; különben pedig elfogadjuk $H_0$-t. Az ANOVA programokban általában feltüntetik azt a legkisebb $\varepsilon$ értéket, amely mellett a csoportok közti eltérés még szignifikáns.

Ha $H_0$-t elutasítottuk, hipotéziseket vizsgálhatunk és konfidenciaintervallumokat szerkeszthetünk az egyes csoportok várható értékére és tetszőleges két csoport várható értékének különbségére.

*Kétszempontos varianciaanalízis (interakcióval)*

Itt is két különböző szempont alapján kialakított $k \cdot p$ csoportban végzünk megfigyeléseket, de cellánként több (mondjuk minden cellában $n$) megfigyelést. Az előző rész példájával élve: $k$ féle technológiával $p$ féle gépen gyártanak alkatrészeket és mérik azok szakítószilárdságát. Itt azonban feltételezzük, hogy a kétféle szempont hatása nem független, (nem mindegy, hogy melyik gépen melyik gyártási technológiát alkalmazzuk).

Jelölje $X_{ijl}$ az első szempont alapján $i$-edik, a második szempont alapján pedig $j$-edik csoportban végzett $l$-edik megfigyelést, példánkban az $i$-edik technológiával a $j$-edik gépen gyártott $l$-edik termék szakítószilárdságát ($i = 1, \ldots, k$; $j = 1, \ldots, p$; $l = 1, \ldots, n$).

Tehát összmintánk elemszáma $kpn$. A mintaelemek függetlenek és $X_{ijl} \sim \mathcal{N}(m + a_i + b_j + c_{ij}, \sigma^2)$, azaz lineáris modellünk most a következő:

$$X_{ijl} = m + a_i + b_j + c_{ij} + \varepsilon_{ijl}, \qquad (i = 1, \ldots, k;\ j = 1, \ldots, p)$$

ahol az $\varepsilon_{ijl} \sim \mathcal{N}(0, \sigma^2)$ független valószínűségi változók véletlen hibák. Itt $a_i$-k jelölik az egyik, $b_j$-k a másik tényező hatásait, $c_{ij}$-k pedig az interakciókat. Feltesszük ($m$-be való beolvasztással elérhető), hogy

$$\sum_{i=1}^{k} a_i = 0, \qquad \sum_{j=1}^{p} b_j = 0,$$

$$\sum_{i=1}^{k} c_{ij} = 0 \qquad (j = 1, \ldots, p) \quad \text{és}$$

$$\sum_{j=1}^{p} c_{ij} = 0 \qquad (i = 1, \ldots, k).$$

A megoldáshoz lásd az ANOVA-táblázatokat.

*Több szempont, kísérlettervezés*

Három vagy több szempont is bevezethető, ekkor azonban az általános modellben vizsgálnunk kell az összes kétszeres, háromszoros, esetleg többszörös interakciót, ami hihetetlenül elbonyolítja a táblázatot (a programcsomagok általánosan csak háromszempontos varianciaanalízist tartalmaznak).

Három szempont esetén tegyük fel, hogy mindegyik szempont szerint $k$ csoportunk van. Ez $k^3$ cellát, és ugyanennyi megfigyelést jelent (ha cellánként csak egy megfigyelést végzünk). Amennyiben nincsenek interakciók a modellben, elég $k^2$ számú kísérletet végeznünk a következő, ún. *latin négyzet* elrendezéssel: az első két faktor $(i, j)$ szintjét a harmadik $l = l(i, j)$ szintjével kombináljuk olyan módon, hogy a harmadik faktor mindegyik szintje minden $i$ és minden $j$ index esetén pontosan egyszer forduljon elő. Ilyen elrendezés persze több is van (csak ciklikus $k!$ db.).

Ezt a módszert először mezőgazdasági parcellákon alkalmazták, ahol gabonafajták sikértartalmát hasonlították össze maga a fajta, az öntözés és az alkalmazott műtrágya szempontjából. Például 5 különböző öntözési mód és az alkalmazott 5-féle műtrágya szerint egy $5 \times 5$-ös parcellát alakítottak ki, ahol a 25 cellába 5 fajta gabonát vetettek úgy, hogy minden sorban és minden oszlopban előforduljon az 5 fajta gabona mindegyike ( ciklikus permutációnál a legegyszerűbb az ültetés, ahol az azonos fajta gabonát rézsutosan ültetik).

## Kanonikus korrelációanalízis

Most két véletlen vektor együttes struktúráját szeretnénk leírni. Mindkettőt ugyanazokon az objektumokon figyeljük meg, ezért úgy is elképzelhetjük őket, hogy egy többdimenziós véletlen vektor komponenseit osztjuk két csoportba, és a két változócsoport közti összefüggést vizsgáljuk. Például pszichiátriai betegek felvett adatait két csoportra osztjuk: az adatok egyik része ún. klinikai paramétereket tartalmaz (ezek vérszérumszint, EEG-adatok elemzéséből adódó folytonos valószínűségi változók), másik része pedig pszichiátriai tesztek eredményeiből áll (melyek különböző skálákon szerzett pontszámok, általában %-os értékek, szintén tekinthetők folytonos valószínűségi változóknak), és azt szeretnénk vizsgálni, hogy mely klinikai paraméterek ill. mely pszichiátriai teszteredmények vannak a legszorosabb kapcsolatban egymással.

Legyen tehát $\mathbf{X} = (X_1, \ldots, X_p)^T$ illetve $\mathbf{Y} = (Y_1, \ldots, Y_q)^T$ $p$- ill. $q$-dimenziós véletlen vektor $\mathbf{0}$ várható érték vektorral. A rájuk vonatkozó $n$ független megfigyelést az $n \times p$-es $\mathbf{F}$ ill. az $n \times q$-as $\mathbf{G}$ adatmátrix tartalmazza ($n > \max\{p, q\}$). Ekkor

$$\widehat{\mathbf{C}}_{11} = \frac{1}{n}\mathbf{F}^T\mathbf{F}, \quad \widehat{\mathbf{C}}_{22} = \frac{1}{n}\mathbf{G}^T\mathbf{G}, \quad \widehat{\mathbf{C}}_{12} = \frac{1}{n}\mathbf{F}^T\mathbf{G}$$

az empirikus kovarianciák ill. keresztkovarianciák mátrixa. Többdimenziós normális eloszlás esetén ezek az elméleti $\mathbf{C}_{11}$, $\mathbf{C}_{22}$, $\mathbf{C}_{12}$ kovarianciamátrixok maximum likelihood becslését adják. Feltehető, hogy $\mathbf{C}_{11}$ és $\mathbf{C}_{22}$ nem-szinguláris, ellenkező esetben ugyanis az $\mathbf{X}$ vagy $\mathbf{Y}$ véletlen vektor komponensei közt lineáris összefüggések lennének, ekkor azonban alacsonyabb dimenziós véletlen vektorokkal lennének helyettesíthetők, azaz $p$ vagy $q$ csökkenthető lenne.

Keresünk olyan $\mathbf{a}_1 \in \mathbb{R}^p$ és $\mathbf{b}_1 \in \mathbb{R}^q$ vektorokat, hogy

$$\mathrm{Corr}\,(\mathbf{a}_1^T\mathbf{X}, \mathbf{b}_1^T\mathbf{Y})$$

maximális legyen. (Amennyiben $\mathbf{X}$ és $\mathbf{Y}$ többdimenziós normális eloszlásúak, ez a Rényi-féle maximálkorreláció feladatának speciális esete.) Ezután a $k = 2, 3, \ldots, l = \min\{p, q\}$ indexekre keressük azokat az $\mathbf{a}_k \in \mathbb{R}^p$ és $\mathbf{b}_k \in \mathbb{R}^q$ vektorokat, melyekkel

$$\mathrm{Corr}\,(\mathbf{a}_k^T\mathbf{X}, \mathbf{b}_k^T\mathbf{Y})$$

maximális és az $\mathbf{a}_k^T\mathbf{X}$ ill. $\mathbf{b}_k^T\mathbf{Y}$ valószínűségi változók korrelálatlanok az $\mathbf{a}_i^T\mathbf{X}$ ill. $\mathbf{b}_i^T\mathbf{Y}$ valószínűségi változókkal $(i = 1, \ldots, k-1)$, azaz a maximumot a

$$\mathrm{Corr}\,(\mathbf{a}_k^T\mathbf{X}, \mathbf{a}_i^T\mathbf{X}) = 0, \quad \mathrm{Corr}\,(\mathbf{b}_k^T\mathbf{Y}, \mathbf{b}_i^T\mathbf{Y}) = 0 \qquad (i = 1, \ldots, k-1)$$

kényszerfeltételek mellett keressük. A $k$-adik maximum értéke $\varrho_k$, a $k$-adik *kanonikus korrelációs együttható.*

A feladat megoldása az empirikus kovarianciákból és keresztkovarianciákból számolt mátrix szinguláris felbontásával történik. Az eredményeket az $\mathbf{a}$ és $\mathbf{b}$ vektorok együtthatóival interpretálhatjuk. Megjegyezzük, hogy a kanonikus korrelációk a többszörös korreláció természetes általánosításai, amennyiben az $\mathbf{Y}$ célváltozó is többdimenziós.

## Diszkriminanciaanalízis

Objektumokat szeretnénk a rajtuk végrehajtott többdimenziós megfigyelések alapján előre adott osztályokba besorolni. Például pácienseket klinikai- vagy pszichiátriai teszteredményeik alapján szeretnénk beteg- ill. kontrollcsoportba, vagy többféle betegcsoportba besorolni; vagy egy új egyedet mért értékei alapján valamely ismert fajba akarunk besorolni.

A módszert úgy kell elképzelni, hogy első lépésben egy ún. tanuló-algoritmust hajtunk végre. Az objektumoknak kezdetben létezik egy osztálybasorolása. Ezt úgy adjuk meg, hogy a megfigyelt többdimenziós, folytonos eloszlású valószínűségi változó komponensein kívül bevezetünk egy, az osztálybatartozásra jellemző diszkrét valószínűségi változót, mely annyiféle értéket vesz fel, ahány osztály van; ez utóbbit egy szakértő a mérésektől függetlenül állapítja meg. Az egyes osztályok adatai alapján diszkrimináló algoritmust készítünk, és megnézzük, hogy az algoritmus szerint melyik osztályba kerülnének eredeti objektumaink. Amennyiben a téves osztálybasorolások száma nem túl nagy, úgy tekintjük, hogy az algoritmus által adott diszkrimináló függvény a továbbiakban is használható az adott csoportok elkülönítésére. Az előbbi példákkal élve, ha a jövőben bejön egy páciens, akkor a mért klinikai- vagy pszichiátriai teszteredményei alapján a diszkrimináló függvény segítségével tudjuk őt beteg- ill. kontrollcsoportba, vagy többféle betegcsoport egyikébe besorolni; vagy elég sok egyedet megfigyelve a fajokból, egy új egyedet már mért tulajdonságai alapján be tudunk sorolni. A téves csoportbasorolásnak általában pozitív valószínűsége van (sőt ez minden csoportra más és más), speciális esetekben ezeket az algoritmus végén meg is tudjuk határozni, és természetesen csak ezekkel a valószínűségekkel bocsátkozhatunk "jóslatokba"

egy újonnan bejövő objektum osztálybasorolása tekintetében. Időnként a közben beérkező objektumokat is hozzávéve a mintához, az algoritmus felfrissíthető, ezáltal a diszkrimináló függvény is módosulhat, sőt a mintaelemszám növelésével javulhat, de $n \to \infty$ esetén sem várható el tökéletes osztálybasorolás.

Egyes alakfelismerő programok is alapulhatnak egy tanulómintán előzetesen végrehajtott diszkriminanciaanalízisen, más néven identifikáción. A most bevezetendő séma döntéselméleti fogalmakat használ, azokat is a legegyszerűbb formában. Hasonló alakfelismerést végez pl. a postán az irányítószámokat leolvasó automata, vagy egy computer tomográf, amely daganatokat diagnosztizál.

A tényleges osztályozás figyelembevételével bevezetjük a következőket. Jelölje $k$ az osztályok számát, továbbá

a. jelölje az egyes osztályokhoz tartozó $p$-dimenziós mintaelemek sűrűségfüggvényét $f_1(\mathbf{x}), \ldots, f_k(\mathbf{x})$ (abszolút folytonos eloszlásokat feltételezünk);

b. jelölje $\pi_1, \ldots, \pi_k$ az egyes osztályok a priori valószínűségeit;

Az a.-beli sűrűségeket osztályonként becsüljük a mintatákból, a b.-beli a priori valószínűségek pedig lehetnek az egyes osztályok relatív gyakoriságai. Így visszük bele "tudásunkat" az alábbi algoritmusba.

Ha már adva lenne a $p$-dimenziós mintatér egy $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_k$ partíciója, akkor a $\mathbf{x} \in \mathcal{X}$ mintaelemet akkor soroljuk a $j$-edik osztályba, ha $\mathbf{x} \in \mathcal{X}_j$. A cél az, hogy a legkisebb veszteséggel járó partíciót megkeressük. Ehhez jelölje $r_{ij} \geq 0$ $(i, j = 1, \ldots k)$ azt a veszteséget, ami akkor keletkezik, ha egy $i$-edik osztálybelit a $j$-edik osztályba sorolunk (a veszteségek nem feltétlenül szimmetrikusak, de feltesszük, hogy $r_{ii} = 0$), és legyen $L_i$ az $i$-edik osztálybeliek besorolásának átlagos vesztesége (rizikója):

$$L_i = \int_{\mathcal{X}_1} r_{i1} f_i(\mathbf{x})\, d\mathbf{x} + \cdots + \int_{\mathcal{X}_k} r_{ik} f_i(\mathbf{x})\, d\mathbf{x}, \quad (i = 1, \ldots, k),$$

ahol összegeztük a veszteségeket azokra az esetekre, mikor az $i$-edik osztálybelit az $1., \ldots, k.$ osztályba soroltuk.

Az egyes $L_i$ veszteségek helyett az

$$L = \sum_{i=1}^{k} \pi_i L_i$$

*átlagos Bayes-féle veszteséget (rizikót)* minimalizáljuk.

$$L = \sum_{i=1}^{k} \pi_i \sum_{j=1}^{k} \int_{\mathcal{X}_j} r_{ij} f_i(\mathbf{x})\, d\mathbf{x} = \sum_{j=1}^{k} \int_{\mathcal{X}_j} \sum_{i=1}^{k} \pi_i r_{ij} f_i(\mathbf{x})\, d\mathbf{x} = -\sum_{j=1}^{k} \int_{\mathcal{X}_j} S_j(\mathbf{x})\, d\mathbf{x},$$

ahol az

$$S_j(\mathbf{x}) = -[\pi_1 r_{1j} f_1(\mathbf{x}) + \cdots + \pi_k r_{kj} f_k(\mathbf{x})]$$

függvényt $j$-edik *diszkrimináló informáns*nak nevezzük, és argumentumában az $\mathbf{x}$ mintaelem szerepel $(j = 1, \ldots, k)$. A negatív előjel miatt $S_j$-k növekedése az átlagos veszteség csökkenését eredményezi, azaz a

$$\sum_{j=1}^{k} \int_{\mathcal{X}_j} S_j(\mathbf{x})\, d\mathbf{x}$$

kifejezést szeretnénk maximalizálni a mintatér összes lehetséges mérhető partícióján.

Célszerűnek tűnik tehát egy $\mathbf{x}$ mért értékekkel rendelkező objektumot abba az osztályba sorolni, melyre diszkrimináló informánsa a legnagyobb értéket veszi fel. Ennek az eljárásnak a jogosságát a következő tény biztosítja: legyen az $\mathcal{X}$ mintatér $\mathcal{X}_1^* \cup \cdots \cup \mathcal{X}_k^*$ partíciója olyan, hogy $\mathbf{x} \in \mathcal{X}_j^*$-ból $S_j(\mathbf{x}) \geq S_i(\mathbf{x})$ következik az összes $i \neq j$ indexekre $(j = 1, \ldots, k)$. Akkor az $\mathcal{X}_1^*, \ldots, \mathcal{X}_k^*$ osztályozással az $L$ átlagos veszteség minimális lesz.

Most néhány egyszerűsítő feltevést vezetünk be. Ha az $r_{ij}$ veszteségekre nincsenek adataink, és az összes téves besorolást egyformán akarjuk büntetni, akkor jobb híján az $r_{ij} = 1$ $(i \neq j)$ és $r_{ii} = 0$ választással élünk. Ezzel

$$S_j(\mathbf{x}) = -\sum_{i=1}^{k} \pi_i r_{ij} f_i(\mathbf{x}) = -\sum_{i \neq j} \pi_i f_i(\mathbf{x}) = -\sum_{i=1}^{k} \pi_i f_i(\mathbf{x}) + \pi_j f_j(\mathbf{x}) = \pi_j f_j(\mathbf{x}) + c,$$

ahol a $c$ konstans nem függ $j$-től. Valójában tehát az $\mathbf{x}$ mért értékekkel rendelkező objektumot az $l$. osztályba soroljuk, ha

$$\pi_l f_l(\mathbf{x}) = \max_{j \in \{1,\ldots,k\}} \pi_j f_j(\mathbf{x}).$$

Tegyük fel, hogy az egyes osztályoknak különböző paraméterű, $p$-dimenziós normális eloszlások felelnek meg. Azaz, ha $\mathbf{X} \in \mathcal{N}_p(\mathbf{m}_j, \mathbf{C}_j)$, akkor

$$f_j(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{C}_j|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_j)^T \mathbf{C}_j^{-1}(\mathbf{x}-\mathbf{m}_j)}.$$

Tekintsük az osztálybasorolás alapját képező $\pi_j f_j(\mathbf{x})$ mennyiségek természetes alapú logaritmusát, a logaritmus monoton transzformáció lévén ez ugyanarra a $j$-re lesz maximális, mint az eredeti kifejezés, sőt az összes $j$-re közös $\ln \frac{1}{(2\pi)^{p/2}}$-től is eltekinthetünk. Az így kapott módosított $j$-edik diszkrimináló informánst $S_j'$-vel jelöljük, és alakja miatt *kvadratikus diszkriminancia szkór*nak is szokás nevezni:

$$S_j'(\mathbf{x}) = -\frac{1}{2}\ln|\mathbf{C}_j| - \frac{1}{2}(\mathbf{x}-\mathbf{m}_j)^T \mathbf{C}_j^{-1}(\mathbf{x}-\mathbf{m}_j) + \ln \pi_j.$$

Ha a kovarianciamátrixok azonosak: $\mathbf{C}_1 = \cdots = \mathbf{C}_k = \mathbf{C}$, akkor $S_j'(\mathbf{x})$-ből a $j$-től független $-\frac{1}{2}\ln|\mathbf{C}|$ és a kvadratikus alak kifejtésében fellépő, $j$-től ugyancsak független $-\frac{1}{2}\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}$ rész elhagyható, a maradék pedig $\mathbf{x}$ lineáris függvényeként írható. Ezt nevezzük *lineáris informán*snak:

$$S_j''(\mathbf{x}) = \mathbf{m}_j^T \mathbf{C}^{-1}\mathbf{x} - \frac{1}{2}\mathbf{m}_j^T \mathbf{C}^{-1}\mathbf{m}_j + \ln \pi_j.$$

Eljárásunk tehát a következő: minden osztályra kiszámoljuk az $S_j''(\mathbf{x})$ értékét $(j = 1, \ldots k)$, és objektumunkat abba az osztályba soroljuk, amelyikre az $S_j''(\mathbf{x})$ lineáris informáns értéke a legnagyobb. A 3.1. Tétel garantálja, hogy ekkor átlagos veszteségünk minimális lesz.

Amennyiben csak két osztályunk van, objektumunkat az $\mathbf{x}$ megfigyelés alapján az első osztályba soroljuk, ha $S_1''(\mathbf{x}) \geq S_2''(\mathbf{x})$, különben a másodikba. Azaz az $S_1''(\mathbf{x}) - S_2''(\mathbf{x})$ különbség előjele fogja eldönteni az osztálybatartozást. De

$$S_1''(\mathbf{x}) - S_2''(\mathbf{x}) = L(\mathbf{x}) - c,$$

ahol

$$L(\mathbf{x}) = (\mathbf{m}_1^T - \mathbf{m}_2^T)\mathbf{C}^{-1}\mathbf{x} \quad \text{és}$$
$$c = \frac{1}{2}(\mathbf{m}_1^T\mathbf{C}^{-1}\mathbf{m}_1 - \mathbf{m}_2^T\mathbf{C}^{-1}\mathbf{m}_2) - \ln\pi_1 + \ln\pi_2.$$

A fenti $L(\mathbf{x})$-et *Fisher-féle diszkriminancia függvény*nek is szokták nevezni, és ennek alapján döntjük el az osztálybatartozást: ha $L(\mathbf{x}) \geq c$, akkor objektumunkat az első, ha pedig $L(\mathbf{x}) < c$, akkor a második osztályba soroljuk. Az $L(\mathbf{x})$ lineáris kifejezésben az egyes $x_i$ változók együtthatói egyfajta súlyokként is szolgálnak, azok a változók fejtik ki a legerősebb hatást a két csoport diszkriminálásában, amely a legnagyobb súllyal szerepelnek. Így például megtudhatjuk, mely klinikai tünetek a legjelentősebbek egy diagnózis készítésekor, mikor első látásra nem különböző beteg–kontroll, vagy két betegcsoportot akarunk elkülöníteni; vagy a másik példában megtudjuk, mely mutatók a leglényegesebbek a fajspecifikációban.

Ha az átlagos veszteséget akarjuk minimalizálni, normális eloszlású minták esetén a fenti eljárás keresztülvihető az egyes osztályokban számolt empirikus kovarianciamátrixokkal és az osztályok relatív gyakoriságaival becsült apriori valószínűségek segítségével. Létezhetnek azonban ún. látens osztályok (pl. egy újfajta betegség, újfajta faj), ami ronthat a módszer alkalmazhatóságán. Szükség van ezért különféle hipotézisvizsgálatokra. Pl. két osztály esetén, az első osztályba való besorolhatóság a

$$T_1 = \frac{[(\mathbf{m}_2 - \mathbf{m}_1)^T\mathbf{C}^{-1}(\mathbf{X} - \mathbf{m}_1)]^2}{(\mathbf{m}_2 - \mathbf{m}_1)^T\mathbf{C}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)} \sim \chi^2(1)$$

statisztikával, míg a második osztályba való besorolhatóság a

$$T_2 = \frac{[(\mathbf{m}_2 - \mathbf{m}_1)^T\mathbf{C}^{-1}(\mathbf{X} - \mathbf{m}_2)]^2}{(\mathbf{m}_2 - \mathbf{m}_1)^T\mathbf{C}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)} \sim \chi^2(1)$$

statisztikával tesztelhető. Ha mind $T_1$, mind $T_2$ szignifikánsan nagyobb az 1-paraméterű $\chi^2$-eloszlás adott (pl. 95%-os) kvantilisénél, akkor egy látens harmadik osztály jelenlétére gyanakodhatunk.

A diszkrimináló informánsokban szereplő paramétereket a mintából becsüljük, minél több a paraméter, annál pontatlanabb az egyes paraméterek becslése; azt is mondhatjuk, hogy a paraméterek a konkrét mintához vannak adaptálva. Ezért, ha az eljárás rizikóját a nem megfelelő osztályba sorolt egyedek száma alapján az alább ismertetendő módon becsüljük, a valódi veszteségfüggvénynél kisebb torzított becslést kapunk. E torzítás kivédésére alkalmazzák az ún. *cross-validation (keresztkiértékelés)* módszert: a paramétereket a minta egy része (60% a szokásos hányad) alapján becsüljük, míg az osztályozás minőségét a paraméterbecslésben fel nem használt mintaelemekkel teszteljük (40%).

Az $L$ átlagos bayesi rizikó becslése: $\hat{L} = \sum_{i=1}^{k} \pi_i \hat{L}_i$, ahol

$$\hat{L}_i = \sum_{j=1}^{k} \frac{n_{ij}}{n_{i.}} r_{ij}, \qquad i = 1, \ldots, k.$$

Itt $n_{ij}$ jelenti azon $i$-edik osztálybeli elemek számát, amelyeket az algoritmus a $j$-edik osztályba sorolt, továbbá $n_{i.} = \sum_{j=1}^{k} n_{ij}$.

## Klaszteranalízis

A diszkriminanciaanalízistől eltérően itt nem adott osztályokkal dolgozunk, hanem magukat az osztályokat (*klaszter*eket) keressük, azaz objektumokat szeretnénk osztályozni a rajtuk végrehajtott többdimenziós megfigyelések alapján (ugyanez megtehető a változókkal is az objektumok alapján).

A minimalizálandó veszteségfüggvény, aminek segítségével az osztályozást végrehajtjuk – egyelőre csak vázlatosan – a következő. Az $n$ db. objektum a $p$-dimenziós mintatér pontjainak tekinthető ($p < n$), és euklideszi metrikában dolgozunk. Tekintsük minden egyes osztályra az adott osztálybeli objektumok súlypontját, és vegyük az objektumok négyzetes eltérését (távolság-négyzetét) a súlyponttól. Az így kapott mennyiségeket utána összegezzük az osztályokra és keressük azt az osztályszámot, hozzá pedig az osztályokat, melyekre ez a veszteség minimális.

Arra vonatkozóan, hogy hogyan alakult ki ez a veszteségfüggvény, röviden utalunk a varianciaanalízisre, ahol a

$$T = W + B$$

szórásnégyzet-felbontás alapvető. A minta teljes (Total) varianciáját a csoportokon belüli (Within) és a csoportok közötti (Between) varianciákra bontjuk fel.

Az objektumok minden egyes partíciójához létezik ilyen felbontás, és a *klaszterezés* (osztálybasorolás) annál homogénebb, minél kisebb $W$ a $B$-hez képest, azaz a

$$\frac{W}{B} = \frac{W}{T - W}$$

kifejezést szeretnénk minimalizálni, ami ($T$ fix lévén) $W$ minimalizálásával ekvivalens. Ezt a $W$ mennyiséget fogjuk tehát alább definiálandó veszteségfüggvényünkbe beépíteni. (Persze, varianciáról csak akkor van értelme beszélnünk, ha feltesszük, hogy mintapontjaink valamely háttéreloszlásból származnak. Jobb híján ezt vehetjük egyenletesnek, ilyenkor egy rész-pontrendszer alapján számolt várható értéknek a pontrendszer súlypontja fog megfelelni, és az ettől vett négyzetes eltérések átlaga adja a pontrendszer varianciáját.)

Legyenek $\mathcal{C}_1, \ldots, \mathcal{C}_k$ a klaszterek (ezek a mintateret alkotó objektumok partícióját jelentik diszjunkt, nem-üres részhalmazokra). A $j$. klaszter súlypontja

$$\mathbf{s}_j = \frac{1}{|\mathcal{C}_j|} \sum_{\mathbf{x}_i \in \mathcal{C}_j} \mathbf{x}_i.$$

A $\mathcal{C}_j$-beliek négyzetes eltéréseinek összege $\mathbf{s}_j$-től:

$$W_j = \sum_{\mathbf{x}_i \in \mathcal{C}_j} \|\mathbf{x}_i - \mathbf{s}_j\|^2 = \frac{1}{|\mathcal{C}_j|} \sum_{\substack{\mathbf{x}_i, \mathbf{x}_{i'} \in \mathcal{C}_j \\ i<i'}} \|\mathbf{x}_i - \mathbf{x}_{i'}\|^2.$$

(Az utolsó egyenlőség egyszerű geometriai meggondolásból adódik, így még a súlypont kiszámolása sem szükséges.)

Megjegyezzük, hogy a fenti euklideszi távolságok az eredeti adatok ortogonális transzformációira invariánsak, a célfüggvény csak a pontok kölcsönös helyzetétől függ. Ezekután keresendő a

$$W = \sum_{j=1}^{k} W_j \to min.$$

veszteség-minimum, amelynek fizikai jelentése a $k$ db. súlypontra vonatkozó tehetetlenségi (inercia) nyomatékok összege.

A minimalizálás persze az összes lehetséges $k$-ra ($1 \leq k \leq n$), és emelett az összes lehetséges klaszterbesorolásra vonatkozik, ami nem oldható meg $n$-ben polinomiális időben. Helyett inkább néhány jól bevált algoritmust használunk:

a. *k-közép* (MacQueen) módszer: a minimalizálandó veszteségfüggvény

$$W = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{C}_j} \|\mathbf{x}_i - \mathbf{s}_j\|^2.$$

Itt $k$ adott (geometriai vagy előzetes meggondolásokból adódik), és induljunk ki egy kezdeti $\mathcal{C}_1^{(0)}, \ldots, \mathcal{C}_k^{(0)}$ klaszterbesorolásból (pl. kiszemelünk $k$ távoli objektumot, és mindegyikhez a hozzájuk közelieket soroljuk, egyelőre csak durva megközelítésben). Egy iterációt hajtunk végre, a lépéseket jelölje $m = 1, 2, \ldots$. Tegyük fel, hogy az $(m-1)$-edik lépésben az objektomoknak már létezik egy $k$ klaszterbe sorolása: $\mathcal{C}_1^{(m-1)}, \ldots, \mathcal{C}_k^{(m-1)}$, a klaszterek súlypontját pedig jelölje $\mathbf{s}_1^{(m-1)}, \ldots, \mathbf{s}_k^{(m-1)}$ (a 0. lépésbeli besorolásnak a kezdő klaszterezés felel meg). Az $m$-edik lépésben átsoroljuk az objektumokat a klaszterek között a következőképpen: egy objektumot abba a klaszterbe sorolunk, melynek súlypontjához a legközelebb van. Pl. $\mathbf{x}_i$-t az $l$. klaszterbe rakjuk, ha

$$\|\mathbf{x}_i - \mathbf{s}_l^{(m-1)}\| = \min_{j \in \{1,\ldots,k\}} \|\mathbf{x}_i - \mathbf{s}_j^{(m-1)}\|$$

(ha a minimum több klaszterre is eléretik, akkor a legkisebb indexű ilyenbe soroljuk be), azaz $\mathbf{x}_i \in \mathcal{C}_l^{(m)}$ lesz. Kétféle módon is el lehet végezni az objektumok átsorolását: vagy az összes objektumot átsoroljuk az $(m-1)$-edik lépésben kialakult klaszter-súlypontokkal számolva, majd a régi súlypontok körül kialakult új klasztereknek módosítjuk a súlypontját, vagy pedig az objektumokat $\mathbf{x}_1, \ldots, \mathbf{x}_n$ szerint sorravéve, mihelyt egy objektum átkerül egy új klaszterbe, módosítjuk annak súlypontját. Így a végén nem kell már újra súlypontokat számolnunk, és az iterációszám is csökkenhet, ui. célratörőbb ("mohó") az algoritmus. Miután az összes objektumot átsoroltuk, az új $\mathcal{C}_1^{(m)}, \ldots, \mathcal{C}_k^{(m)}$ klaszterezásből és az új $\mathbf{s}_1^{(m)}, \ldots, \mathbf{s}_k^{(m)}$ súlypontokból kiindulva ismét teszünk egy lépést. Meddig? Választhatunk többféle leállási kritériumot is, pl. azt, hogy az objektumok már stabilizálódnak a klaszterekben, és a klaszterek nem változnak az iteráció során. Mivel a $W$ veszteségfüggvény minden egyes lépésben csökken, és véges sok objektumunk van, egy ilyen állapot véges lépésen belül bekövetkezik, ezt $W$ stacionárius pontjának nevezzük. A "mohó" algoritmus elég gyors, és végállapota általában nem függ a kezdő klaszterbesorolástól, szerencsés kezdeti klaszterezés esetén néhány lépésben végetér. A leállási szabály lehet az is, hogy bizonyos ésszerű korlát alatt marad a klaszterbeli pozíciójukat változtató objektumok száma. Egyéb feltételeket is szoktak a kialakult klaszterekre róni, pl. hogy *jól szeparált*ak legyenek: ez azt jelenti, hogy a legnagyobb, azonos klaszterbeliek közti távolság is kisebb, mint a létező legkisebb, különböző klaszterbeliek közti távolság. Beizonyítható, hogy az objektumok ilyen geometriai struktúráját a fenti legkisebb négyzetes kritériumra épített veszteségfüggvényt minimalizáló iteráció megtalálja.

b. Az ún. *agglomeratív* ill. *divizív* módszerek a klaszterszámot fokozatosan csökkentik ill. növelik. Ezek közül is az ún. *hierarchikus* eljárások terjedtek el, ahol úgy csökkentjük ill. növeljük a klaszterszámot, hogy minden lépésben bizonyos klasztereket összevonunk ill. szétvágunk. Például nézzünk egy agglomeratív, hierarchikus eljárást. A kezdeti klaszterszám $k^{(0)} = n$, tehát kezdetben minden objektum egy külön klasztert alkot. Az iteráció a következő: tegyük fel, hogy az $m$. lépésben már csak $k^{(m)}$ db. klaszterünk van. Számítsuk ki a klaszter-középpontokat (súlypontokat). Ezek euklideszi távolságai egy $k^{(m)} \times k^{(m)}$-es, szimmetrikus ún. távolság-mátrixot alkotnak (fődiagonálisa 0). Azokat a klasztereket, melyek távolsága egy adott korlátnál kisebb, egy klaszterbe vonjuk össze, ilyen módon egy lépésben persze kettőnél több klaszter is összevonódhat. Végül, legfeljebb $n$ lépésben már minden összeolvad, és csak egy klaszterünk lesz. Nem szükséges persze végigcsinálni az összes lépést. Agglomeratív eljárások esetén a $W$ veszteségfüggvény általában monoton nő, azt kell megfigyelni, hol ugrik meg drasztikusan. Ha végigcsináljuk az összes lépést, az ún. dendrogramot szemlélve próbálunk meg egy ésszerű klaszterszámot találni. Ilyen agglomeratív, hierarchikus eljárás a *legközelebbi szomszéd* módszer is, amely akkor is összevon két klasztert, ha létezik közöttük egy lánc, amelyben az egymás utáni elemek már közelebb vannak egymáshoz egy adott korlátnál.

c. Ha a célfüggvényt még klaszterenként egy súllyal is ellátjuk, akkor súlyozott klaszterezésről beszélünk (pl. lehet a klaszterek elemszámának valamely függvényét tekinteni, vagy más módon súlyozni a homogenitás mértékét).

d. Nemcsak diszjunkt, hanem átfedő klasztereket is létrehozhatunk.

e. Gráfelméleti módszerek is léteznek, ahol az objektumokat egy gráf csúcspontjainak képzeljük, az élek pedig a köztük való távolságok valamely monoton csökkenő függvényével vannak súlyozva (minél közelebb van egymáshoz két objektum, annál nagyobb közöttük a hasonlóság). A klaszterezés ilyenkor a kezdeti súlyozott gráf szintgráfjainak megkeresését jelenti, ahol egy adott korláthoz tartozó szintgráf úgy jön létre, hogy csak a korlát alatti (vagy feletti) súllyal rendelkező éleket tartjuk meg. Ez lényegében a b.-beli dendogramon végzett műveleteknek felel meg, vannak azonban bonyolultabb gráf-struktúrákat kereső algoritmusok is. Pl. a minimális feszítő fa keresése a legközelebbi szomszéd módszerrel mutat rokonságot.

f. Valószínűségszámítási módszerek, melyek feltételezik, hogy az egyes klaszterbeli pontok más-más eloszlásból származnak, a kapott minta pedig ezek keverékeloszlásából adódik. Nekünk a keveréket kell felbontani komponenseire, a felbontás "jóságára" vonatkozóan pedig hipotéziseket vizsgálhatunk.

Előfordulhat, hogy az objektumok között nem akarunk, vagy nem tudunk közvetlenül euklideszi távolságokat számolni . Ennek különösen akkor van jelentősége, mikor méréseink nem folytonos eloszlású, hanem diszkrét valószínűségi változókra vonatkoznak, és struktúrájukat gráffal, hipergráffal, vagy egyszerűen csak valamilyen hasonlósági mérőszámokkal tudjuk jellemezni. Ilyenkor elsődleges feladatunk az, hogy objektumainkat valamilyen euklideszi térbe ágyazzuk be, azután használhatjuk csak a fenti a.b.c. metrikus módszereket. Ilyen célú eljárás az ún. *többdimenziós skálázás*.

## SZTOCHASZTIKUS FOLYAMATOK

A pénzügyi folyamatokat sokszor időben változó valószínűségi változókkal írják le. Az $\{X_t : t \in \mathbb{R}\}$ halmazt sztochasztikus folyamatnak vagy idősornak nevezzük, ahol $t$ az idő. Amennyiben az idő folytonosan változik, folytonos idejű, ha pedig az $X_t$ változók is folytonosan veszik fel értékeiket a valós számegyenes ($\mathbb{R}$) valamely tartományában, akkor folytonos álapotterű sztochasztikus folyamatról beszélünk, általában ilyeneket vizsgálunk. Ha az $X_t$ értékeket az idő ($t$) függvényében ábrázoljuk, a sztochasztikus folyamat egy lehetséges lefutását (trajektóriáját) kapjuk.

Definiáljuk a leggyakrabban vizsgált típusú idősorokat.

*Definíció..* Az $\{X_t : t \in \mathbb{R}\}$ idősort *független növekményű*nek nevezzük, ha az $X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, \ldots, X_{t_n} - X_{t_{n-1}}$ valószínűségi változók függetlenek tetszőleges $t_1 < t_2 < \cdots < t_n$ időpontok esetén. A folyamatot *stacionárius növekményű*nek nevezzük, ha az $X_{t+h} - X_t$ növekmény eloszlása csak $h$-tól függ tetszőleges $t, h$ esetén.

Példa erre az alább definiálandó Wiener-folyamat.

*Definíció..* A $\{W_t : t \geq 0\}$ idősort (standard) *Wiener-folyamat*nak, másképpen *Brown-mozgás*nak nevezzük, ha

a. $W_0 = 0$ és $W_t$ trajektóriái 1 valószínűséggel folytonosak.

b. Tetszőleges $0 \leq t_1 < t_2 < \cdots < t_n$ időpontok esetén a $(W_{t_1}, W_{t_2}, \ldots, W_{t_n})$ véletlen vektor eloszlása $n$-dimenziós normális.

c. Független, stacionárius növekményű, ahol $W_{t+h} - W_t$ eloszlása normális (Gauss), 0 várható értékkel és $h$ varianciával (szórásnégyzettel) tetszőleges $t, h > 0$ esetén.

*Definíció..* A folytonos idejű $\{X_t : t \in \mathbb{R}\}$ sztochasztikus folyamatot *Markov-folyamat*nak nevezzük, ha tetszőleges $t_1 < t_2 < \cdots < t_n < t$ időpontok esetén

$$\mathbb{P}(a < X_t \leq b \,|\, X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_n} = x_n) = \mathbb{P}(a < X_t \leq b \,|\, X_{t_n} = x_n)$$

minden $a < b$ és $x_1, x_2, \ldots x_n \in \mathbb{R}$ értékrendszerre.

Az olyan Markov-folyamatokat, amelyeknek majdnem minden trajektóriája folytonos függvény, *diffúziós folyamat*oknak nevezünk. Ilyen például a Wiener-folyamat.

Definíciója miatt egy Markov-folyamatot egyértelműen meghatározunk, ha tetszőleges $s < t$ és $x \in \mathbb{R}$ értékekre megadjuk $X_t$ feltételes eloszlását az $X_s = x$ feltétel mellett. Ezt az $f(X_t = y | X_s = x)$ feltételes sűrűségfüggvény definiálja, $y$ helyen felvett értékét $p(y, t|x, s)$ jelöli, melyet a diszkrét eset analógiájára *átmenet-valószínűség*nek nevezünk.

*Definíció..* Az $\{X_t : t \in \mathbb{R}\}$ Markov-folyamatot *stacionárius átmenet-valószínűségű*nek (*homogén*nek) mondjuk, ha a $p(y, t|x, s)$ átmenet-valószínűség csak a $t - s = h$ különbségtől függ.

A továbbiakban $p_h(x, y)$ jelöli egy stacionárius átmenet-valószínűségű Markov-folyamat átmenet-valószínűségeit. Ilyen folyamatra példa a Wiener-folyamat.

Ezután térjünk át olyan sztochasztikus folyamatok tárgyalására, melyek időben homogének.

*Definíció..*

a. Az $\{X_t : t \in \mathbb{R}\}$ sztochasztikus folyamatot *erősen stacionárius*nak nevezzük, ha tetszőleges $n \in \mathbb{N}$ és $h \in \mathbb{R}$ esetén ($\mathbb{N}$ a természetes számok halmaza) $X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_n+h}$ együttes eloszlása magegezik $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ együttes eloszlásával, bármik legyenek is a $t_1, t_2, \ldots, t_n$ időpontok.

b. Az $\{X_t : t \in \mathbb{R}\}$ sztochasztikus folyamatot *gyengén stacionárius*nak nevezzük, ha tetszőleges $h \in \mathbb{R}$ esetén a $\mathrm{Cov}\,(X_t, X_{t+h})$ kovariancia csak $h$-tól függ (és független $t$-től).

Amennyiben egy folyamatot stacionáriusnak mondunk, a gyenge stacionaritást értjük alatta. Megjegyezzük, hogy egyetlen nem konstans, stacionárius növekményű folyamat sem stacionárius, így például a Wiener-folyamat sem az.

Stacionárius folyamatok vizsgálatánál fontos szerepet játszik az ún. kovarianciafüggvény:

$$R(h) = \mathrm{Cov}\,(X_h, X_0) = \mathrm{Cov}\,(X_{t+h}, X_t), \qquad h > 0.$$

Mivel az első és második momentum a normális eloszlást egyértelműen meghatározza, minden gyengén stacionárius folyamathoz van erősen stacionárius Gauss-folyamat, amelynek ugyanaz a kovarianciafüggvénye.

Gyakran idősorunk diszkrét: csak megszámlálható (a gyakorlatban véges) sok (általában ekvidisztans) időpontban figyeljük meg (a folytonos idejű folyamatokat is sokszor diszkretizáljuk). Diszkrét idejű idősorunk felírható $X_0, X_1, \ldots, X_n, \ldots$ alakban (néha mindkét irányban végtelen sorozat), ahol $X_0$ a kezdeti állapotot jelenti. Ilyenkor az autokovarianciafüggvény csak egész helyeken értelmezett:

$$R(i) = \mathrm{Cov}\,(X_i, X_0) = \mathrm{Cov}\,(X_{n+i}, X_n), \qquad i = 1, 2, \ldots.$$

Ha ezt még leosztjuk $X_i$ és $X_0$ szórásával, az ún. autokorrelációkhoz jutunk. Ezek sorozatának fontos szerepe van a stacionárius folyamatok karakterizálásában (spektrálanalízisében).

Fontos szerepet játszanak majd a továbbiakban az alábbi típusú, diszkrét idejű stacionárius folyamatok.

*Definíció..* Az $X_0, X_1, \ldots, X_n, \ldots$ folyamatot *$p$-edrendű autoregressziós folyamat*nak nevezzük és $AR(p)$-vel jelöljük, ha megfelelő $a_1, a_2, \ldots, a_p \in \mathbb{R}$ együtthatókkal tetszőleges $n \in N$ esetén

$$X_n = a_1 X_{n-1} + a_2 X_{n-2} + \cdots + a_p X_{n-p} + \xi_n, \tag{1.1}$$

ahol $\{\xi_n\}$ 0 várható értékű, korrelálatlan valószínűségi változók sorozata, melyeknek közös a szórásnégyzetük és $\xi_n$ korrelálatlan az $X_{n-1}, X_{n-2}, \ldots$ változóktól is.

Adott minta alapján az $a_i$ együtthatók becslése egy többváltozós lineáris regressziós feladat. Amennyiben az

$$x^p - a_1 x^{p-1} - \cdots - a_p$$

polinom (esetleg komplex) gyökei 1-nél kisebb abszolút értékűek, akkor az (1.1) egyenletnek van stacionárius megoldása.

Ennél általánosabb a következő folyamat.

*Definíció..* Az $X_0, X_1, \ldots, X_n, \ldots$ stacionárius folyamatot *$p$-edrendű autoregressziós és $q$-adrendű mozgóátlag folyamat*nak nevezzük és $ARMA(p, q)$-val jelöljük, ha megfelelő $a_1, a_2, \ldots, a_p$ és $b_1, \ldots, b_q$ valós együtthatókkal tetszőleges $n \in N$ esetén

$$X_n = a_1 X_{n-1} + a_2 X_{n-2} + \cdots + a_p X_{n-p} + b_1 \varepsilon_n + b_2 \varepsilon_{n-1} + \cdots + b_q \varepsilon_{n-q+1},$$

ahol $\{\varepsilon_n\}$ 0 várható értékű, 1 szórású, korrelálatlan valószínűségi változók sorozata (amennyiben normális eloszlásúak is, azaz $\mathcal{N}(0, 1)$-változók, az ilyen sorozatot *fehérzaj*nak nevezzük).