

Többváltozós statisztika

Nagy-György Judit

Szegedi Tudományegyetem, Bolyai Intézet

Ezek a módszerek több változó együttes vizsgálatára vonatkoznak.

Alapvető típusaik:

- ▶ többdimenziós eloszlásokra vonatkozó hipotézisvizsgálatok (általában többdimenziós normális eloszlás paramétereire vonatkoznak),
- ▶ varianciaanalízisek,
- ▶ klasszifikációs módszerek (diszkrimináló illetve klaszterező eljárások),
- ▶ dimenziócsökkentés,
- ▶ ...

Egyszempontos varianciaanalízis (ANOVA)

Tekintsük a $\xi_{i,j} \sim N(\mu_i, \sigma^2)$, $i = 1, \dots, r$, $j = 1, \dots, n_i$ független mintákat.

A várható értékek felírhatók a következő formában: $\mu_i = \mu + a_i$, ahol μ a várható értékek n_i értékekkel súlyozott átlaga, a_i pedig az i -edik csoportthatás.

$$H_0 : \mu_1 = \dots = \mu_r,$$

azaz minden csoportthatás 0.

Legyen $n = n_1 + \dots + n_r$. Képezzük a következő statisztikákat:

$$\bar{\xi}_i = \frac{\sum_{j=1}^{n_i} \xi_{i,j}}{n_i}, \quad \bar{\xi} = \frac{\sum_{i=1}^r \sum_{j=1}^{n_i} \xi_{i,j}}{n}.$$

$$SST = \sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{i,j} - \bar{\xi})^2,$$

$$SSB = \sum_{i=1}^r n_i (\bar{\xi}_i - \bar{\xi})^2, \quad SSW = \sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{i,j} - \bar{\xi}_i)^2.$$

Állítás

$$SST = SSB + SSW.$$

Bizonyítás

$$\begin{aligned} SST &= \sum_{i=1}^r \sum_{j=1}^{n_i} ((\xi_{i,j} - \bar{\xi}_i) + (\bar{\xi}_i - \bar{\xi}))^2 \\ &= SSW + SSB + 2 \sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{i,j} - \bar{\xi}_i)(\bar{\xi}_i - \bar{\xi}) \\ &= SSW + SSB + 2 \sum_{i=1}^r (\bar{\xi}_i - \bar{\xi}) \underbrace{\sum_{j=1}^{n_i} (\xi_{i,j} - \bar{\xi}_i)}_0. \end{aligned}$$

Visszatérve H_0 vizsgálatára:

- ▶ a próbastatisztika

$$F = \frac{SSB}{SSW} \cdot \frac{n-r}{r-1},$$

ami H_0 mellett $F(r-1, n-r)$ eloszlású.

- ▶ Tehát a kritikus érték α -hoz $F_\alpha = F_{r-1, n-r}^{-1}(1-\alpha)$.
- ▶ Döntés: ha $F < F_\alpha$, akkor H_0 -t elfogadjuk, különben elvetjük.

Megjegyzések

- ▶ ANOVA a kétmintás t-próba általánosítása: könnyen látható, hogy ha $r = 2$ és $\Delta_0 = 0$, akkor $F = t^2$, és $F_\alpha = t_\alpha^2$.
- ▶ Itt is feltétel a szórások egyezése, ami az általánosított F -próbával tesztelhető. Ha a szórások nem egyenlőek, akkor ANOVA helyett a többmintás Welch-próbát alkalmazzák.
- ▶ Az ANOVA-t diszkrét és normális eloszlású változó függetlenségének tesztelésére is használják.
- ▶ Van többszemponyos varianciaanalízis is.

Lineáris regresszió

Tekintsük a (ξ, η) véletlen vektort. Keressük η -t legjobban közelítő lineáris függvényt, azaz a -t és b -t, amelyre $E((\eta - (a\xi + b))^2)$ **négyzetes eltérés minimális**. Ennek megoldása

$$a = \frac{\text{Cov}(\xi, \eta)}{D^2(\xi)} = r(\xi, \eta) \frac{D(\eta)}{D(\xi)},$$

$$b = E(\eta) - \frac{\text{Cov}(\xi, \eta)}{D^2(\xi)} E(\xi).$$

Definíció

A fenti paraméterekkel felírt $y = ax + b$ egyenes a **regressziós egyenes**, a, b paraméterek a **lineáris regressziós együtthatók**.

Megjegyzés

Ha ξ, η függetlenek, akkor $a = r(\xi, \eta) \frac{D(\eta)}{D(\xi)} = 0$.

Regressziós egyenes paramétereinek becslése

Cél $Q(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2$ minimalizálása, ez a (legkisebb négyzetek módszere). $Q(a, b)$ minimumát az alábbi helyen veszi fel:

$$\hat{a} = r_n(\xi, \eta) \frac{D_n(\eta)}{D_n(\xi)}, \quad \hat{b} = \frac{\sum_{i=1}^n y_i}{n} - \hat{a} \frac{\sum_{i=1}^n x_i}{n}.$$

Definíció

Az $\hat{a}, \hat{b}$ értékeket *becsült lineáris regressziós együtthatóknak*, az $y = \hat{a}x + \hat{b}$ egyenest *becsült lineáris regressziós egyenesnek* nevezzük.

Megjegyzések

- ▶ $Q(a, b)$ az $(x_1, y_1), \dots, (x_n, y_n)$ pontok együttes négyzetes távolsága az $y = ax + b$ egyenestől.
- ▶ A becsült lineáris regressziós egyenest η_{n+1} érték becslésére használjuk, ha x_{n+1} adott csak.

Varianciaanalízis lineáris modellben

Tekintsük a $\eta = ax + b + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ regressziós modellt és az $\eta_1, \dots, \eta_n$ mintát.

$$H_0 : a = 0.$$

Vegyük a regressziós egyenes paramétereinek $\hat{a}$, $\hat{b}$ becsléseit és legyen $\hat{\eta}_i = \hat{a}x_i + \hat{b}$. Képezzük a következő négyzetösszegeket:

$$SST = \sum_{i=1}^n (\eta_i - \bar{\eta})^2, \quad SSR = \sum_{i=1}^n (\hat{\eta}_i - \bar{\eta})^2, \quad SSE = \sum_{i=1}^n (\eta_i - \hat{\eta}_i)^2.$$

► a próbastatisztika $F = \frac{SSR}{SSE}(n-2),$

ami H_0 mellett $F(1, n-2)$ eloszlású.

► Tehát a kritikus érték α -hoz $F_\alpha = F_{1, n-2}^{-1}(1 - \alpha).$

► Döntés: ha $F < F_\alpha$, akkor H_0 -t elfogadjuk, különben elvetjük.

Állítás

$$SST = SSR + SSE.$$

Bizonyítás

$$SST = \sum_{i=1}^n ((\eta_i - \hat{\eta}_i) + (\hat{\eta}_i - \bar{\eta}))^2 = SSR + SSE + 2 \underbrace{\sum_{i=1}^n (\eta_i - \hat{\eta}_i)(\hat{\eta}_i - \bar{\eta})}_{=0},$$

$$2 \sum_{i=1}^n (\eta_i - \hat{a}x_i - \hat{b})(\hat{a}x_i + \hat{b} - \bar{\eta}) = -\hat{a} \frac{\partial Q}{\partial a} - (\hat{b} - \bar{\eta}) \frac{\partial Q}{\partial b} = 0.$$

Állítás

$$SSR = SST \cdot r_n^2(x, \eta).$$

Bizonyítás

$$\sum_{i=1}^n (\eta_i - \hat{a}x_i - \hat{b})^2 = \hat{a}^2 \sum_{i=1}^n (x_i - \bar{x})^2 = r_n^2(x, \eta) \frac{V_n(\eta)}{V_n(x)} n V_n(x).$$

Következmény

$$SSE = SST(1 - r_n^2(x, \eta)) \quad \text{és} \quad \frac{SSR}{SSE} = \frac{r_n^2(x, \eta)}{1 - r_n^2(x, \eta)}.$$

Többváltozós regresszió

A modell $\eta = a_1x_1 + \dots + a_rx_r + b + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$.

$$H_0 : a_1 = \dots = a_r = 0.$$

- ▶ Vesszük $a_1, \dots, a_r, b$ legkisebb négyzetes becsléseit, $\hat{\eta}_i = \hat{a}_1x_{i,1} + \dots + \hat{a}_rx_{i,r} + \hat{b}_i$ és képezzük SST , SSR és SSE négyzetösszegeket.
- ▶ A próbastatisztika most $F = \frac{SSR}{SSE} \cdot \frac{n-r-1}{r}$, ami H_0 mellett $F(r, n-r-1)$ eloszlású.

Megjegyzések

- ▶ Az x_i -k a teljes variancia $r_n^2(x, \eta)$ -részét magyarázzák.
- ▶ Ha (ξ, η) kétdimenziós normális eloszlású és ξ, η függetlenek, akkor $\sqrt{n-2} \frac{r_n(\xi, \eta)}{\sqrt{1-r_n^2(\xi, \eta)}}$ aszimptotikusan $t(n-2)$ eloszlású.
- ▶ Az általános regressziós modell $\eta = f(\xi) + \varepsilon$. Általában a nemlineáris regressziót a lineárisra vezetik vissza.

Főkomponensanalízis

Legyen $\mathbf{X}$ egy d -dimenziós véletlen vektor $\mathbf{m}$ várható érték vektorral és pozitív definit $\mathbf{C}$ kovarianciamátrixszal. Cél a változók számának csökkentése minél kevesebb információ vesztesével.

$\mathbf{C}$ szimmetrikus, spektrálfelbontása

$$\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T,$$

- ▶ ahol $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_d)$ a $\mathbf{C}$ mátrix $\lambda_1 \geq \dots \geq \lambda_d > 0$ sajátértékeinek diagonális mátrixa,
- ▶ $\mathbf{V}$ oszlopvektorai pedig a hozzá tartozó ortonormált sajátvektorrendszer a megfelelő sorrendben.

Definíció

A $\mathbf{Y} = \mathbf{V}^{-1}(\mathbf{X} - \mathbf{m}) = \mathbf{V}^T(\mathbf{X} - \mathbf{m})$ az $\mathbf{X}$ főkomponensvektora, $\mathbf{Y}$ k -adik komponense (Y_k) a k -adik főkomponens.

Állítás

$\mathbf{Y}$ kovarianciamátrixa Λ .

Bizonyítás

$$\mathbf{E}(\mathbf{Y}) = \mathbf{E}(\mathbf{V}^T(\mathbf{X} - \mathbf{m})) = \mathbf{V}^T(\mathbf{E}(\mathbf{X}) - \mathbf{m}) = \mathbf{0},$$

$$\begin{aligned}\mathbf{E}(\mathbf{Y}\mathbf{Y}^T) &= \mathbf{E}(\mathbf{V}^T(\mathbf{X} - \mathbf{m})(\mathbf{X} - \mathbf{m})^T\mathbf{V}) \\ &= \mathbf{V}^T\mathbf{E}((\mathbf{X} - \mathbf{m})(\mathbf{X} - \mathbf{m})^T)\mathbf{V} = \mathbf{V}^T\mathbf{C}\mathbf{V} = \Lambda.\end{aligned}$$

Tétel

$\mathbf{x}_1, \dots, \mathbf{x}_d$ ortonormált vektorrendszer. Ekkor

$$\sum_{i=1}^k \mathbf{x}_i^T \mathbf{C} \mathbf{x}_i \leq \sum_{i=1}^k \lambda_i, \quad k = 1, \dots, d.$$

Ha $\mathbf{x}_i = \mathbf{v}_i, i = 1, \dots, d$, akkor egyenlőség teljesül.

Következmény

Y_k szórása maximális az $\mathbf{z}^T(\mathbf{X} - \mathbf{m})$ alakú vv-k között, amelyre $\|\mathbf{x}\| = 1$ és $\mathbf{z}^T(\mathbf{X} - \mathbf{m})$ korrelálatlan Y_i -vel minden $i < k$ -ra.

Tétel

A $k(< d)$ -dimenziós $\mathbf{P}$ projekciók közül $E(\|\mathbf{X} - \mathbf{P}\mathbf{X}\|)$ kifejezést a $\mathbf{v}_1, \dots, \mathbf{v}_k$ által feszített altérre vetítő projekció minimalizálja.

(Tehát $\mathbf{X}$ négyzetes középben vett legjobb k -dimenziós becslésének első k koordinátája a sajátvektorok bázisában az első k főkomponens, a többi pedig 0.)

k megválasztása

$$\frac{\lambda_1 + \dots + \lambda_k}{\lambda_1 + \dots + \lambda_d}$$

hányados azt mutatja, hogy az első k főkomponens a teljes variancia hányadrészét magyarázza.

Megjegyzés

A gyakorlatban egy minta empirikus kovarianciamátrixából indulunk, ennek sajátértékei és sajátvektorai lesznek λ_i és $\mathbf{v}_i$ ($i = 1, \dots, d$) becslései.

Definíció

Legyen $\mathbf{X}$ egy d -dimenziós véletlen vektor $\mathbf{m}$ várható érték vektorral és $\mathbf{C}$ kovarianciamátrixszal. A *k -faktor modell* a következő:

$$\mathbf{X} = \mathbf{A}\mathbf{Y} + \mathbf{W} + \mathbf{m},$$

- ▶ ahol $\mathbf{A}$ *átviteli mátrix vagy faktorsúlymátrix* $d \times k$ -as,
- ▶ az $\mathbf{Y}$ *közös faktor* k -dimenziós véletlen vektor, amelyre $E(\mathbf{Y}) = \mathbf{0}$ és $E(\mathbf{Y}\mathbf{Y}^\top) = \mathbf{I}_k$, valamint
- ▶ $\mathbf{W}$ *egyedi faktor* egy d -dimenziós véletlen vektor, amelyre $E(\mathbf{W}) = \mathbf{0}$, $E(\mathbf{W}\mathbf{W}^\top) = \mathbf{D}$ diagonális mátrix, valamint $E(\mathbf{Y}\mathbf{W}^\top) = \mathbf{0}$.

A modell paraméterei $\mathbf{A}$ és $\mathbf{D}$, feltevése, hogy a faktorkomponensek korrelálatlanok, továbbá a közös faktorok szórásnégyzete 1.

Állítás

$\mathbf{C} = \mathbf{A}\mathbf{A}^\top + \mathbf{D}$. (Ezt nevezzük a modell mátrixalakjának.)

Bizonyítás

$X_i = \sum_{j=1}^k a_{i,j} Y_j + W_i + m_i \quad (1 \leq i \leq d)$ szórásnégyzete
 $c_{ii} = \sum_{j=1}^k a_{i,j}^2 + d_{i,i}$, mivel a faktorkomponensek korrelálatlanok.

($\sum_{j=1}^k a_{i,j}^2$ az X_i változó **kommunalitása**, $d_{i,i}$ az **egyedi variancia**.)

Tétel

Adott $k < d$ esetén a k -faktor modellnek pontosan akkor van megoldása, ha van olyan $d \times d$ -s $\mathbf{D}$ nemnegatív elemű diagonális mátrix, hogy $\mathbf{C} - \mathbf{D}$ egy legfeljebb k -adrangú, pozitív szemidefinit.

Bizonyítás

Ha van megoldás, akkor $\mathbf{C} - \mathbf{D} = \mathbf{A}\mathbf{A}^\top$ rangja legfeljebb k , mivel $\mathbf{A}$ egy $d \times k$ -as mátrix. Fordítva, ha $\mathbf{C} - \mathbf{D}$ szimmetrikus, rangja $r(\leq k)$, akkor felírható $\mathbf{A}\mathbf{A}^\top$ alakban, ahol $\mathbf{A}$ $d \times r$ -es.

Megjegyzések

- ▶ Ha $\mathbf{C} = \mathbf{D}$ -nek van $\mathbf{A}\mathbf{A}^\top$ alakú felbontása, akkor mivel ez nem egyértelmű: ha $\mathbf{U}$ egy ortogonális mátrix, akkor $\mathbf{B} = \mathbf{A}\mathbf{U}^\top$ -ra $\mathbf{A}\mathbf{A}^\top = \mathbf{B}\mathbf{B}^\top$, sokszor további feltételeket is tesznek $\mathbf{A}$ -ra.
- ▶ Közelítő megoldásokat szoktak adni (ML; legkisebb négyzetek módszere; főkomponensanalízis módosított változata; ...).

A faktorsúlyok értelmezése és a faktorok rotációja

$\text{Cov}(X_i, Y_j) = E((X_i - m_i)Y_j) = E((\sum_{j=1}^k a_{i,j} Y_j + W_i) Y_j) = a_{i,j}$
a faktorsúlyok korrelálatlansága miatt.

- ▶ X_i az Y_j faktorral akkor áll szoros kapcsolatban, ha $|a_{i,j}|$ nagy.
- ▶ A faktorok értelmezése szempontjából az a jó, ha minden i -re kevés (esetleg egy) $|a_{i,j}|$ nagy, a többi kicsi.
- ▶ A **rotáció** célja olyan $\mathbf{U}$ $k \times k$ -s ortogonális mátrix keresése, amelyre az

$$\mathbf{X} = \mathbf{X} = \mathbf{A}^* \mathbf{Y}^* + \mathbf{W} + \mathbf{m}$$

modell, ahol $\mathbf{A}^* = \mathbf{A}\mathbf{U}$ és $\mathbf{Y}^* = \mathbf{U}^\top \mathbf{Y}$, teljesíti a fentieket.

Diszkriminanciaanalízis

Legyen $\mathbf{X}$ változó $N(\mathbf{m}_1, \mathbf{C})$ vagy $N(\mathbf{m}_2, \mathbf{C})$ eloszlású. Ezt az **osztálybatartozást** szeretnénk eldönteni a tér egy $\mathcal{X}_1 \cup \mathcal{X}_2$ partícionálásával. Legyen f_i az i -edik osztályhoz tartozó sűrűségfüggvény. (Általában $\mathbf{m}_i$ -t és $\mathbf{C}$ -t a mintából becsüljük.) Két megközelítést tekintünk.

I. Az

$$\int_{\mathcal{X}_1} f_2(\mathbf{x}) d\mathbf{x} + \int_{\mathcal{X}_2} f_1(\mathbf{x}) d\mathbf{x}$$

veszteségfüggvényt minimalizáló partícionálást keresünk.

Belátható, hogy az átlagos veszteség akkor minimális, ha $\mathbf{x}$ osztálya j , amire $\mathbf{m}_j^\top \mathbf{C}^{-1} \mathbf{x} - \frac{1}{2} \mathbf{m}_j^\top \mathbf{C}^{-1} \mathbf{m}_j$ nagyobb. Ez alapján

$$L(\mathbf{x}) = (\mathbf{m}_1 - \mathbf{m}_2)^\top \mathbf{C}^{-1} \mathbf{x} + c$$

alakú **Fisher-féle diszkriminancia-függvényt** használjuk: $\mathbf{x}$ pontosan akkor tartozik az első osztályba, ha $L(\mathbf{x}) > 0$.

II. Legyen $\mathbf{A} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^\top$. Tekintsük a **Rayleigh-hányadost**:

$$R(\mathbf{v}) = \frac{\mathbf{v}^\top \mathbf{A} \mathbf{v}}{\mathbf{v}^\top \mathbf{C} \mathbf{v}}$$

Heurisztika: R maximumhelyének irányában jól tudunk szeparálni. Legyenek $\mathbf{C}^{-1}\mathbf{A}$ sajátértékei $\lambda_1 \geq \dots \geq \lambda_d$ és $\mathbf{v}_1, \dots, \mathbf{v}_d$ hozzájuk tartozó sajátvektorok a **kanonikus főirányok**.

► Belátható, hogy R maximumhelye $\mathbf{v}_1$.

► Egy $L(\mathbf{x}) = \mathbf{v}_1^\top \mathbf{x} + b$

alakú **diszkriminanciafüggvény** segítségével döntünk: $\mathbf{x}$ akkor esik az első osztályba, ha $L(\mathbf{x}) > 0$.

Megjegyzések

- A két diszkriminanciafüggvény pozitív konstans faktorban tér el, ugyanazt az osztályozást eredményezik.
- A szeparálás jóságát a **Wilks-féle** $\Lambda = (1 + \hat{\lambda}_1)^{-1}$ statisztikával jellemezzük. Ha az osztályok jól szeparálhatók, akkor Λ kicsi.

Logisztikus regresszió

Legyen η ismeretlen p paraméterű Bernoulli-változó és $\mathbf{x}$ egy d -dimenziós vektor. Tegyük fel, hogy $p = p(\mathbf{x})$, ezt a függést a következő **logisztikus regressziós modell (logitmodell)** írja le:

$$\ln \frac{p}{1-p} = \mathbf{a}^\top \mathbf{x} + b.$$

- ▶ Tekintsük egy n elemű mintát. Meghatározzuk $\hat{\mathbf{a}}, \hat{b}$ ML becsléseket (kihasználva, hogy az $\mathbf{x}_i$ vektorhoz tartozó n_i mintaelem összege binomiális eloszlású n_i és $p(\mathbf{x}_i)$ paraméterekkel). Ebből kapjuk p becslését:

$$\hat{p} = \frac{1}{1 + e^{-\hat{\mathbf{a}}^\top \mathbf{x} - \hat{b}}}.$$

- ▶ η értékei szerinti osztálybatartozás előrejelzése ennek segítségével úgy történik, hogy rögzítünk egy $0 < p_0 < 1$ értéket, és $\hat{\eta} = 1$, ha $\hat{p} > p_0$.

Különböző eloszlásból vett minták esetén nem tudjuk, melyik mintaelem melyik **osztályba (klaszterbe)** tartozik, esetleg az osztályok száma is ismeretlen. Célunk az osztályok meghatározása.

Tekintsük az $\mathbf{x}_1, \dots, \mathbf{x}_n$ realizációt. Tegyük fel, hogy van k klaszter: $\mathcal{C}_1, \dots, \mathcal{C}_k$. Legyen $\bar{\mathbf{x}}_{\mathcal{C}} = \frac{1}{|\mathcal{C}|} \sum_{\mathbf{x}_j \in \mathcal{C}} \mathbf{x}_j$ a $\mathcal{C}$ klaszterközéppontja.

Legyen $SST = \sum_{i=1}^n \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2$, $SSW = \sum_{i=1}^k \sum_{\mathbf{x}_j \in \mathcal{C}_i} \|\mathbf{x}_j - \bar{\mathbf{x}}_{\mathcal{C}_i}\|^2$, $SSB = \sum_{i=1}^k |\mathcal{C}_i| \cdot \|\bar{\mathbf{x}}_{\mathcal{C}_i} - \bar{\mathbf{x}}\|^2$. Most is $SST = SSW + SSB$.

A klasztereket úgy szeretnénk meghatározni, hogy a

$$\frac{SSW}{SSB} = \frac{SSW}{SST - SSW}$$

kifejezést **minimalizálják**, ami egyenértékű SSW minimalizálásával.

SSW minimumának meghatározása már két klaszter esetén is $2^{n-1} - 1$ eset vizsgálatát jelenti az n elemű mintán, ezért heurisztikákat alkalmaznak.

A klaszterező eljárásoknak két alapvető fajtája van aszerint, hogy a klaszterek száma ismert-e.

k -közép módszer

Veszünk k kezdeti klaszterközéppontot, minden mintaelemet ahhoz a klaszterhez soroljuk, amelyik középpontjához közelebb van, majd a középpontokat újraszámítjuk. Függ a kezdeti középpontoktól.

Hierarchikus módszerek

Vannak a klaszterszámot csökkentő (agglomeratív vagy összevonó) és növelő (divizív vagy felosztó) módszerek. Pl. előbbi esetében kezdetben minden pont külön klaszterbe tartozik, minden lépésben a két legközelebbi klasztert vonjuk össze, míg egy klaszter nem lesz. Függ a klaszterek távolságának definíciójától.

Klaszterek távolsága ($d(\mathcal{C}_i, \mathcal{C}_j)$)

- ▶ Legközelebbi szomszéd (nearest neighbor vagy single linkage):

$$\min_{\mathbf{x} \in \mathcal{C}_i, \mathbf{y} \in \mathcal{C}_j} \|\mathbf{x} - \mathbf{y}\|.$$

- ▶ Legtávolabbi szomszéd (furthest neighbor vagy complete linkage):

$$\max_{\mathbf{x} \in \mathcal{C}_i, \mathbf{y} \in \mathcal{C}_j} \|\mathbf{x} - \mathbf{y}\|.$$

- ▶ Átlagos távolság (between-groups linkage vagy average linkage):

$$\frac{1}{|\mathcal{C}_i| \cdot |\mathcal{C}_j|} \sum_{\mathbf{x} \in \mathcal{C}_i, \mathbf{y} \in \mathcal{C}_j} \|\mathbf{x} - \mathbf{y}\|.$$

- ▶ Unió pontjainak átlagos távolsága (within-groups linkage):

$$\frac{1}{\binom{|\mathcal{C}_i| + |\mathcal{C}_j|}{2}} \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{C}_i \cup \mathcal{C}_j} \|\mathbf{x} - \mathbf{y}\|.$$

- ▶ Klaszterközéppontok távolsága (centroid): $\|\bar{\mathbf{x}}_{\mathcal{C}_i} - \bar{\mathbf{x}}_{\mathcal{C}_j}\|$.

- ▶ Medián távolság: $\|\tilde{\mathbf{x}}_{\mathcal{C}_i} - \tilde{\mathbf{x}}_{\mathcal{C}_j}\|$,

ahol $\tilde{\mathbf{x}}_{\mathcal{C}}$ a $\mathcal{C}$ klaszter súlyozott közepe, amit rekurzívan számolunk a következő módon: ha $\mathcal{C}$ klaszter $\mathcal{C}' \cup \mathcal{C}''$ unióként jött létre, akkor $\tilde{\mathbf{x}}_{\mathcal{C}} = \frac{1}{2}(\tilde{\mathbf{x}}_{\mathcal{C}'} + \tilde{\mathbf{x}}_{\mathcal{C}''})$.

- ▶ Ward távolság:

$$\begin{aligned} & \sum_{\mathbf{x} \in \mathcal{C}_i \cup \mathcal{C}_j} \|\mathbf{x} - \bar{\mathbf{x}}_{\mathcal{C}_i \cup \mathcal{C}_j}\|^2 - \sum_{\ell \in \{i,j\}} \sum_{\mathbf{x} \in \mathcal{C}_\ell} \|\mathbf{x} - \bar{\mathbf{x}}_{\mathcal{C}_\ell}\|^2 \\ &= \frac{|\mathcal{C}_i| \cdot |\mathcal{C}_j|}{|\mathcal{C}_i| + |\mathcal{C}_j|} \|\bar{\mathbf{x}}_{\mathcal{C}_i} - \bar{\mathbf{x}}_{\mathcal{C}_j}\|^2. \end{aligned}$$

(Tehát a Ward módszer azt a két klasztert vonja össze, amelyek összevonásával SSW legkevésbé nő.)

Bolla Marianna, Krámlí András: *Statisztikai következtetések elmélete*, Typotex

Fazekas István (szerk.): *Bevezetés a matematikai statisztikába*, Kossuth Egyetemi Kiadó

Móri F. Tamás, Szeidl László, Zempléni András: *Matematikai statisztika példatár*, ELTE Eötvös Kiadó

Nagy-György Judit, Osztényiné Krauczi Éva, Székely László: *Valószínűségszámítás és statisztika példatár*, Polygon Jegyzettár

Viharos László: *A sztochasztika alapjai*, Polygon Jegyzettár