Alkalmazott statisztika jegyzetvázlat

Kevei Péter

2018. november 30.

Tartalomjegyzék

1. Többdimenziós normális eloszlás **2**
1.1. Véletlen vektorváltozók 2
1.2. Többdimenziós normális eloszlás 3
1.3. Paraméterek ML becslése 6
1.4. Hotelling-féle T^2 eloszlás 10
1.5. Várható érték tesztelése 13
1.5.1. Ismert kovarianciamátrix esete 13
1.5.2. Ismeretlen kovarianciamátrix esete 13
1.6. Többdimenziós CHT 14

2. Lineáris módszerek **15**
2.1. Főkomponensanalízis 15
2.2. Merőleges vetítés 17
2.3. Lineáris regresszió véletlen regresszorral 20
2.4. Determinisztikus változók 21
2.5. Varianciaanalízis 22

1. Többdimenziós normális eloszlás

1.1. Véletlen vektorváltozók

Az 1.1 és 1.2 fejezetek nagyrészt a Csörgő jegyzetből valók ([2, 33. fejezet]).

Legyen $\mathbf{X}^\top = (X_1, \ldots, X_k)$ véletlen vektor az $(\Omega, \mathcal{A}, \mathbf{P})$ valószínűségi mezőn. Ha $\mathbf{E}(|X_j|) < \infty$ és $m_j = \mathbf{E}(X_j)$, $1 \le j \le k$, akkor az $\mathbf{m}^\top = (m_1, \ldots, m_k)$ jelöléssel, a várható értéket komponensenként véve $\mathbf{E}(\mathbf{X}) = \mathbf{m}$ az $\mathbf{X}$ véletlen vektor *várható érték vektora.* Ha a szórások is végesek, azaz $\mathbf{E}(X_j^2) < \infty$, $1 \le j \le k$, akkor a $\sigma_{jl} = \mathbf{Cov}(X_j, X_l)$ kovariancia definiált minden $1 \le j, l \le k$ párra. A kovarianciákból képzett $k \times k$-as

$$\Sigma = \Sigma_X = \begin{pmatrix} \sigma_{11} & \ldots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \ldots & \sigma_{kk} \end{pmatrix} = \mathbf{E}\big((\mathbf{X}-\mathbf{m})(\mathbf{X}-\mathbf{m})^\top\big) = \mathbf{Cov}(\mathbf{X})$$

szimmetrikus mátrixot $\mathbf{X}$ *kovarianciamátrixának* nevezzük.

A $k \times k$-as A mátrix *pozitív szemidefinit*, ha a

$$\mathbf{x}^\top A \mathbf{x} = \sum_{i,j} a_{i,j} x_i x_j \ge 0$$

minden $\mathbf{x} \in \mathbb{R}^k$ esetén, és *pozitív definit*, ha az $x^T A x$ kvadratikus alak pozitív minden $\mathbf{x} \ne 0^\top = (0, \ldots, 0)$ vektor esetén. Egy nemnegatív definit mátrix pontosan akkor pozitív definit, ha nemszinguláris, azaz $\det A \ne 0$.

1.1. Állítás. *A kovarianciamátrix szimmetrikus, pozitív szemidefinit mátrix.*

Ha $\sigma_j^2 > 0$, $1 \le j \le k$, akkor definiálhatjuk a $\varrho_{jl} = \varrho(X_j, X_l) = \sigma_{jl}/(\sigma_j \sigma_l)$, $1 \le j,\ l \le k$ korrelációs együtthatókat. Legyen

$$R = \begin{pmatrix} \varrho_{11} & \ldots & \varrho_{1k} \\ \vdots & \ddots & \vdots \\ \varrho_{k1} & \ldots & \varrho_{kk} \end{pmatrix} \quad \text{és} \quad D_0 = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_k \end{pmatrix}.$$

A szimmetrikus nemnegatív definit $R = R_X$ mátrix az $\mathbf{X}$ *korrelációmátrixa.* Most is $\det(R) \ge 0$, és $\varrho_{11} = \cdots = \varrho_{kk} = 1$. A D_0 diagonális mátrix a $\Sigma = D_0 R D_0$ azonosság miatt hasznos.

1.2. Definíció. Legyenek $X_1, \ldots, X_k$ olyan véletlen változók, melyeknek Σ kovarianciamátrixa létezik. Akkor mondjuk, hogy az $X_1, \ldots, X_k$ változók *lineárisan függetlenek*, ha a $\mathbf{P}\big\{ \sum_{j=1}^{k} x_j(X_j - m_j) = 0\big\} = 1$ egyenlőség esetén $x_1^2 + \cdots + x_k^2 = 0$, azaz $(x_1, \ldots, x_k) = (0, \ldots, 0)$.

Gondoljuk meg, hogy a definíció szemléletesen azt jelenti, hogy $X_1, \ldots, X_k$ pontosan akkor függőek, ha az $\mathbf{X} = (X_1, \ldots, X_k)^\top$ véletlen vektor 1 valószínűséggel egy hipersíkra koncentrált, azaz az eloszlás degenerált.

1.3. Állítás. *Legyen $\mathbf{X}^\top = (X_1, \ldots, X_k)$ véletlen vektor, melyre $\mathbf{E}(X_j^2) < \infty$, $1 \le j \le k$, $\mathbf{E}(\mathbf{X}) = \mathbf{m}$ és $\mathbf{Cov}(\mathbf{X}) = \Sigma$.*

1. *A következő öt állítás ekvivalens: Σ pozitív definit; R pozitív definit; $\det \Sigma > 0$; $\det R > 0$; $X_1, \ldots, X_k$ lineárisan függetlenek.*

2. *Ha $r \in \mathbb{N}$, $A \in \mathbb{R}^{r \times k}$ és $\mathbf{b} \in \mathbb{R}^r$, akkor $\mathbf{E}(A\mathbf{X} + \mathbf{b}) = A\mathbf{m} + \mathbf{b}$ és $\mathbf{Cov}(A\mathbf{X} + \mathbf{b}) = A\Sigma A^\top$.*

1.4. Feladat. Bizonyítsuk be az állítást!

1.2. Többdimenziós normális eloszlás

Legyenek $Z_1, \ldots, Z_k$ független standard normális eloszlású véletlen változók egy $(\Omega, \mathcal{A}, \mathbf{P})$ valószínűségi mezőn. Ekkor $\mathbf{Z}^\top = (Z_1, \ldots, Z_k)$ kovarianciamátrixa $I = (\delta_{jl})_{j,l=1}^{k}$ a $k \times k$-as egységmátrix. Ekkor *Z k-dimenziós standard normális eloszlású véletlen vektorváltozó, jelben $Z \sim \mathcal{N}_k(0, I)$*. A függetlenség miatt Z sűrűségfüggvénye

$$f_{0,I}(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}}\, e^{-\frac{1}{2}\,|\mathbf{x}|^2} = \frac{1}{(2\pi)^{k/2}}\, e^{-\frac{1}{2}\,\langle x, x\rangle}\,, \quad \mathbf{x} \in \mathbb{R}^k\,,$$

Persze ha egy Z véletlen vektor $Z \sim \mathrm{N}_k(0, I)$ eloszlású, akkor komponensei szükségképpen független standard normálisok.

Legyenek $A \in \mathbb{R}^{k \times k}$ és $m \in \mathbb{R}^k$ tetszőlegesek, $Z \sim \mathrm{N}_k(0, I)$. Tekintsük az $\mathbf{X} = A\mathbf{Z} + m$ k-dimenziós véletlen vektort. Az 1.3. Állítás szerint $\mathbf{E}(\mathbf{X}) = \mathbf{m}$, $\Sigma := \mathbf{Cov}(\mathbf{X}) = AIA^\top = AA^\top$.

Legyen $\Sigma \in \mathbb{R}^{k \times k}$ szimmetrikus nemnegatív definit mátrix, és tekintsük λ_j sajátértékeit és $x_j \in \mathbb{R}^k$ sajátvektorait, azaz $\Sigma x_j = \lambda x_j$, $j = 1, \ldots, k$. Alapvető eredmény lineáris algebrából, hogy az $x_1, \ldots, x_k$ sajátvektorokból ortogonális U mátrix képezhető, és a $D = U^T \Sigma U$ mátrix, ill. ennek D_0 négyzetgyöke a következő alakú:

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{pmatrix} \quad \text{és} \quad D_0 = \begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_k} \end{pmatrix}.$$

Mivel Σ nemnegatív definit, ezért $\lambda_1 \geq 0, \ldots, \lambda_k \geq 0$, azaz D_0 valóban definiálható valós mátrix. Bevezetve az $A = UD_0$ mátrixot látjuk, hogy

$$AA^T = UD_0(UD_0)^T = UD_0D_0^TU^T = UD_0D_0U^T = UDU^T,$$

azaz $AA^T = \Sigma$. [Emlékeztetünk, hogy egy $U = (u_{jl})_{j,l=1}^k$ mátrix akkor ortogonális, ha $\sum_{j=1}^k u_{lj}u_{mj} = \delta_{lm}$, $1 \leq l, m \leq k$, vagyis oszlopai (sorai) merőlegesek egymásra és normáltak. Ekkor persze $U^{-1} = U^T$.]

Ezzel beláttuk, a következőt.

1.5. Állítás. *Adott $\Sigma \in \mathbb{R}^{k\times k}$ szimmetrikus pozitív szemidefinit mátrixhoz és $\mathbf{m} \in \mathbb{R}^k$ vektorhoz létezik olyan véletlen, melynek Σ a kovarianciamátrixa, és $\mathbf{m}$ a várható érték vektora.*

1.6. Definíció. A k-dimenziós $\mathbf{X}$ véletlen vektor *normális eloszlású*, ha van olyan $\mathbf{m} \in \mathbb{R}^k$, $A \in \mathbb{R}^{k\times k}$, és $\mathbf{Z}$ k dimenziós standard normális vektorváltozó, hogy

$$\mathbf{X} = A\mathbf{Z} + \mathbf{m}.$$

Ekkor $\mathbf{X} \sim \mathcal{N}_k(m, \Sigma)$, ahol $\mathbf{m} \in \mathbb{R}^k$ és $\Sigma = AA^\top$ szimmetrikus nemnegatív definit mátrix. Ekkor $\mathbf{E}(X) = \mathbf{m}$ és $\mathbf{Cov}(X) = \Sigma$.

Ha Σ szinguláris, azaz $\det \Sigma = 0$, akkor az 1.3. Állítás szerint $\mathbf{X} \sim \mathcal{N}_k(m, \Sigma)$ koordinátái lineárisan függőek, és így $\mathbf{X}$ degenerált abban az értelemben, hogy majdnem biztosan egy k-nál kisebb dimenziós hipersíkra koncentrált. Ekkor persze eloszlása nem lehet folytonos.

A nemdegenerált esetben az $\mathcal{N}_k(m, \Sigma)$ eloszlás folytonos és sűrűségét is meg tudjuk határozni.

1.7. Állítás. *Jelölje $\mathbf{X}$ véletlen vektor $\mathcal{N}_k(\mathbf{m}, \Sigma)$ normális eloszlású, ahol $\mathbf{m} \in \mathbb{R}^k$ és $\Sigma \in \mathbb{R}^{k\times k}$ szimmetrikus* pozitív definit *kovarianciamátrix. Ekkor $\mathbf{X}$ folytonos és sűrűségfüggvénye*

$$f_{\mathbf{m},\Sigma}(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}\sqrt{\det \Sigma}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^\top \Sigma^{-1}(\mathbf{x}-\mathbf{m})}, \quad \mathbf{x} \in \mathbb{R}^k.$$

Bizonyítás. Legyen $\mathbf{Z} \sim \mathcal{N}_k(0, I)$. Ekkor $AZ + m \sim \mathcal{N}_k(m, \Sigma)$, ahol A az a fent definiált $k \times k$-as mátrix, melyre $\Sigma = AA^\top$. Ekkor A is nemszinguláris, hiszen $0 \neq \det(\Sigma) = \det(A)\det(A^\top)$. Legyen $M : \mathbb{R}^k \to \mathbb{R}^k$ az a lineáris transzformáció, melyre $M(\mathbf{x}) = A^{-1}(\mathbf{x}-\mathbf{m})$, $\mathbf{x} \in \mathbb{R}^k$. Tetszőleges B Borel-halmazra

$$\begin{aligned}
\mathbf{P}(\mathbf{x} \in B) &= \mathbf{P}\{A\mathbf{Z} + \mathbf{m} \in B\} = \mathbf{P}\{\mathbf{Z} \in M(B)\} \\
&= \int_{M(B)} \frac{1}{(2\pi)^{k/2}}\, e^{-\frac{1}{2}|y|^2} \lambda^k(\mathrm{d}y) \\
&= \int_B \frac{1}{(2\pi)^{k/2}}\, e^{-\frac{1}{2}|A^{-1}(\mathbf{x}-\mathbf{m})|^2} |\det(A^{-1})| \lambda^k(\mathrm{d}x) \\
&= \int_B \frac{1}{(2\pi)^{k/2}|\det(A)|}\, e^{-\frac{1}{2}\langle A^{-1}(\mathbf{x}-\mathbf{m}), A^{-1}(\mathbf{x}-\mathbf{m})\rangle} \lambda^k(\mathrm{d}x) \\
&= \int_B \frac{1}{(2\pi)^{k/2}\sqrt{|\det(\Sigma)|}}\, e^{-\frac{1}{2}\langle (A^{-1})^T A^{-1}(\mathbf{x}-\mathbf{m}), \mathbf{x}-\mathbf{m}\rangle} \lambda^k(\mathrm{d}x) \\
&= \int_B \frac{1}{(2\pi)^{k/2}\sqrt{\det(\Sigma)}}\, e^{-\frac{1}{2}\langle \Sigma^{-1}(\mathbf{x}-\mathbf{m}), \mathbf{x}-\mathbf{m}\rangle} \lambda^k(\mathrm{d}x) \\
&= \int_B f_{\mathbf{m},\Sigma}(x)\, \mathrm{d}x\,,
\end{aligned}$$

ahol felhasználtuk azt az egyszerű tényt, hogy $(A^{-1})^\top = (A^\top)^{-1}$. □

Tekintsük a $\Sigma = UDU^\top$ spektrálfelbontást, ahol D diagonális, U pedig ortogonális. Bevezetve a $\mathbf{z} = U^\top(\mathbf{x}-\mathbf{m})$ változót (vegyük észre, hogy ez a transzformáció egy eltolás majd egy forgatás), a sűrűségfüggvényben az exponens

$$\begin{aligned}
(\mathbf{x}-\mathbf{m})^\top \Sigma^{-1}(\mathbf{x}-\mathbf{m}) &= (\mathbf{x}-\mathbf{m})^\top U D^{-1} U^\top (\mathbf{x}-\mathbf{m}) \\
&= \mathbf{z}^\top D^{-1} \mathbf{z}^\top \\
&= \sum_{i=1}^{k} \frac{1}{\lambda_i} z_i^2,
\end{aligned}$$

alakra hozható. Innen látjuk, hogy a sűrűségfüggvény szintvonalai ellipszoidok.

A sűrűségfüggvény explicit alakjából azonnal adódik a következő állítás.

1.8. Állítás. *Az $X_1, X_2, \ldots, X_k$ együttesen normális eloszlású véletlen változók pontosan akkor függetlenek, ha kovarianciamátrixuk diagonális.*

A következő állítás szerint a többváltozós normális eloszlások osztálya zárt a lineáris transzformációkra nézve.

1.9. Állítás. *Ha $\mathbf{X} \sim \mathcal{N}_k(\mathbf{m}, \Sigma)$, $\mathbf{b} \in \mathbb{R}^r$, $r \in \mathbb{N}$, és $A \in \mathbb{R}^{r\times k}$, akkor $A\mathbf{X} + \mathbf{b} \sim \mathcal{N}_r(A\mathbf{m} + \mathbf{b}, A\Sigma A^\top)$.*

Az 1.9. Állítás szerint az egydimenziós esetben megszokott módon lehet standardizálni.

1.10. Következmény. *Ha* $\mathbf{X} \sim \mathcal{N}_k(\mathbf{m}, \Sigma)$*, és* $\Sigma = AA^\top$*, akkor*

1. $\mathbf{Y} = A^{-1}(\mathbf{X} - \mathbf{m}) \sim \mathcal{N}_k(0, I_k)$.
2. $(\mathbf{X} - \mathbf{m})^\top \Sigma^{-1}(\mathbf{X} - \mathbf{m}) \sim \chi^2(k)$.

Bizonyítás. Az első állítás az 1.9. Állítás következménye, míg a második következik a χ^2-eloszlás definíciójából és az elsőből. □

Az 1.9. Állítás speciális esete a következő.

1.11. Következmény. *Ha* $\mathbf{X} \sim \mathcal{N}_k(\mathbf{m}, \Sigma)$ *, akkor minden* $\mathbf{t} \in \mathbb{R}^k$ *esetén* $\langle t, X\rangle \sim \mathcal{N}(\langle \mathbf{t}, \mathbf{m}\rangle, \mathbf{t}^\top \Sigma t)$*, ha* $\mathbf{t}^\top \Sigma \mathbf{t} > 0$ *és* $\langle \mathbf{t}, \mathbf{X}\rangle$ *degenerált – majdnem biztosan* $\langle \mathbf{t}, \mathbf{m}\rangle$ *– ha* $\mathbf{t}^\top \Sigma \mathbf{t} = 0$.

A többdimenziós normális eloszlás számos szép tulajdonsága ismert. Itt csak további kettőt említünk: Az 1.7. Állítás megfordítása is igaz, nevezetesen ha $\mathbf{X} \sim \mathrm{N}_k(\mathbf{m}, \Sigma)$ és $\mathbf{X}$ eloszlása abszolút folytonos, akkor Σ pozitív definit. Az utóbbi következmény megfordítása is igaz, azaz a következményben megfogalmazott tulajdonság karakterizálja a normális eloszlást.

1.3. Paraméterek ML becslése

Itt Johnson, Wichern [3, Chapter 4.3] jegyzetet követjük.

Legyenek $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ független $N_k(\mathbf{m}, \Sigma)$ eloszlású véletlen vektorok. A következőkben megadjuk $(\mathbf{m}, \Sigma)$ maximum likelihood becslését. A függetlenség miatt az együttes sűrűségfüggvény

$$\begin{aligned} f(\mathbf{x}_1, \ldots, \mathbf{x}_n) &= \prod_{j=1}^n \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}_i - \mathbf{m})^\top \Sigma^{-1}(\mathbf{x}_i - \mathbf{m})} \\ &= \frac{1}{(2\pi)^{nk/2}|\Sigma|^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})^\top \Sigma^{-1}(\mathbf{x}_i - \mathbf{m}) \right\}. \end{aligned}$$

Az alábbi jól ismert lineáris algebrából.

1.12. Feladat. Egy négyzetes mátrix *nyoma* a főátlóban lévő elemeinek összege, azaz $\mathrm{tr}(A) = \sum_{i=1}^k a_{ii}$. Igazoljuk, hogy

1. $\mathrm{tr}(BC) = \mathrm{tr}(CB)$, ahol $B \in \mathbb{R}^{k \times \ell}$, $C \in \mathbb{R}^{\ell \times k}$ (azaz a nyom *ciklikus*);

2. $\mathbf{x}^\top A\mathbf{x} = \mathrm{tr}(\mathbf{x}^\top A\mathbf{x}) = \mathrm{tr}(A\mathbf{x}\mathbf{x}^\top)$, ahol $A \in \mathbb{R}^{k\times k}$, $\mathbf{x} \in \mathbb{R}^k$;
3. $\mathrm{tr}A = \sum_{i=1}^k \lambda_i$, ahol λ_i-k az A sajátértékei.

A Steiner-formula magasabb dimenziós megfelelője az alábbi. A bizonyítás az egydimenziós esethez hasonlóan egyszerű.

1.13. Feladat. *Többdimenziós Steiner-formula.* Igazoljuk, hogy tetszőleges $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^k$, $\mathbf{v} \in \mathbb{R}^k$ esetén

$$\sum_{i=1}^n (\mathbf{x}_k - \mathbf{v})(\mathbf{x}_k - \mathbf{v})^\top = \sum_{i=1}^n (\mathbf{x}_k - \overline{\mathbf{x}})(\mathbf{x}_k - \mathbf{x})^\top + n(\overline{\mathbf{x}} - \mathbf{v})(\overline{\mathbf{x}} - \mathbf{v})^\top,$$

ahol $\overline{\mathbf{x}} = n^{-1}\sum_{i=1}^n \mathbf{x}_i$.

A levezetésnél szükségünk lesz az alábbi egyszerű állításra.

1.14. Feladat. Legyen Σ pozitív definit szimmetrikus mátrix. Igazoljuk, hogy ha $\Sigma\mathbf{x} = \lambda\mathbf{x}$, akkor $\Sigma^{-1}\mathbf{x} = \lambda^{-1}\mathbf{x}$, és Σ^{-1} pozitív definit!

A nyom ciklikusságát és a Steiner-formulát felhasználva, némi számolás után kapjuk, hogy

$$\begin{aligned}\sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})^\top \Sigma^{-1}(\mathbf{x}_i - \mathbf{m}) = \mathrm{tr}&\left(\Sigma^{-1}\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top\right)\\ &+ n(\overline{\mathbf{x}} - \mathbf{m})^\top \Sigma^{-1}(\overline{\mathbf{x}} - \mathbf{m}),\end{aligned}$$

ahol

$$\overline{\mathbf{x}}_n = \overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i. \tag{1}$$

A formulát beírva a sűrűségfüggvényre, és áttérve a likelihood függvényre (ami persze ugyanaz, csak az argumentum változott)

$$\begin{aligned}L(\mathbf{m}, \Sigma) = \frac{1}{(2\pi)^{nk/2}|\Sigma|^{n/2}}&\times\\ \times \exp&\left\{ -\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top\right)\right.\\ &\left. - n(\overline{\mathbf{x}} - \mathbf{m})^\top \Sigma^{-1}(\overline{\mathbf{x}} - \mathbf{m})\right\}.\end{aligned} \tag{2}$$

Mivel pozitív definit mátrix inverze is pozitív definit (lásd 1.14. Feladat), ezért $\mathbf{y}^\top\Sigma^{-1}\mathbf{y} \geq 0$, tehát a L maximuma $\mathbf{m}$-ben

$$\mathbf{m} = \overline{\mathbf{x}}.$$

Mivel ez nem függ Σ-tól, ezért egyszerűen beírhatjuk a formulába, és maximalizálhatunk Σ-ban, ami érdekesebb feladat, hiszen egy pozitív definit mátrixban keressük a maximumot. Kapjuk, hogy

$$L(\overline{\mathbf{x}}, \Sigma) = \frac{1}{(2\pi)^{nk/2}|\Sigma|^{n/2}} \exp\left\{ -\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top \right) \right\}. \quad (3)$$

A szélsőérték-feladatot a következő állítás segítségével oldjuk meg.

1.15. Lemma. *Legyen $B \in \mathbb{R}^{k\times k}$ szimmetrikus pozitív definit mátrix, $b > 0$. Tetszőleges $\Sigma \in \mathbb{R}^{k\times k}$ pozitív definit mátrixra*

$$\frac{1}{|\Sigma|^b} \exp\left\{ -\frac{1}{2}\mathrm{tr}(\Sigma^{-1}B) \right\} \le \frac{1}{|B|^b}(2b)^{kb}e^{-bk}.$$

Továbbá pontosan akkor teljesül egyenlőség, ha $\Sigma = \frac{1}{2b}B$.

Bizonyítás. Tekintsük a $B = UDU^\top$ spektrálfelbontást, és legyen $B^{1/2} = UD^{1/2}U^\top$. Legyenek $\lambda_1, \ldots, \lambda_k$ a $B^{1/2}\Sigma^{-1}B^{1/2}$ pozitív definit mátrix sajátértékei. A tr ciklikussága miatt

$$\mathrm{tr}(\Sigma^{-1}B) = \mathrm{tr}(B^{1/2}\Sigma^{-1}B^{1/2}) = \sum_{i=1}^{k} \lambda_i.$$

Mivel $|B^{1/2}\Sigma^{-1}B^{1/2}| = \prod_{i=1}^k \lambda_i = |B|/|\Sigma|$, így

$$\frac{1}{|\Sigma|^b} \exp\left\{ -\frac{1}{2}\mathrm{tr}(\Sigma^{-1}B) \right\} = \frac{\left(\prod_{i=1}^k \lambda_i\right)^b}{|B|^b} e^{-\frac{1}{2}\sum \lambda_i} = \frac{1}{|B|^b} \prod_{i=1}^{k} \lambda_i^b \, e^{-\frac{1}{2}\sum \lambda_i}.$$

Egyszerű számolás mutatja, hogy a $x^b e^{-x/2}$ függvény a maximumát az $x = 2b$ helyen veszi fel, ahonnan az egyenlőtlenség már adódik. Egyenlőség pontosan akkor van, ha

$$\lambda_1 = \ldots = \lambda_k = 2b,$$

ami pedig éppen azt jelenti, hogy $\Sigma = \frac{1}{2b}B$. □

A lemmából következik, hogy (3) formulában a maximum a

$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top$$

helyen vétetik fel. Ezzel beláttuk a következőt.

1.16. Tétel. *A többdimenziós normális eloszlásnál az* $(\mathbf{m}, \Sigma)$ *pár maximum likelihood becslése*

$$\widehat{\mathbf{m}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i = \overline{\mathbf{X}}_n,$$

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^\top = \frac{1}{n}S.$$

A várható érték becslése torzítatlan, míg a kovarianciamátrix becslése csak aszimptotikusan torzítatlan.

1.17. Feladat. Igazoljuk, hogy

$$\mathbf{E}\overline{\mathbf{X}} = \mathbf{m}, \quad \mathbf{E}\frac{1}{n}S = \frac{n-1}{n}\Sigma.$$

A konzisztencia egyszerűen következik a nagy számok törvényéből.

1.18. Feladat. Igazoljuk, hogy

$$\overline{\mathbf{X}}_n \to \mathbf{m}, \quad \text{és } \frac{1}{n}S \to \Sigma, \quad \text{amint } n \to \infty,$$

egy valószínűséggel, és (persze ebből már következik) gyengén.

A továbbiakban a Bolla és Krámli [1, 5.4 fejezet] jegyzetét követjük.

1.19. Definíció. Egy $W \in \mathbb{R}^{k\times k}$ véletlen mátrix n szabadsági fokú és Σ kovarianciájú Wishart-mátrix, jelben $W \sim \mathcal{W}_k(n, \Sigma)$, ha $W = XX^\top$, ahol $X \in \mathbb{R}^{k\times n}$ véletlen mátrix oszlopai független $\mathcal{N}_k(0, \Sigma)$ eloszlású normálisok. Ha $\Sigma = I_k$ akkor standard Wishart-eloszlásról beszélünk.

Az alábbi az 1.9. Állítás egyszerű következménye.

1.20. Állítás. *Legyen* $\Sigma \in \mathbb{R}^{k\times k}$ *pozitív definit. Ekkor* $W \sim \mathcal{W}_k(n, \Sigma)$ *pontosan akkor, ha* $\Sigma^{-1/2}W\Sigma^{-1/2} \sim \mathcal{W}_k(n, I_k)$*.*

1.21. Feladat. Lássuk be az állítást!

A következő tétel, mely Lukács Jenő tételének többdimenziós változata, a Hotelling-féle T-eloszlás tulajdonságainál lesz fontos.

1.22. Tétel. *Legyenek* $\mathbf{X}_1, \dots, \mathbf{X}_n$ *független* $\mathcal{N}_k(\mathbf{m}, \Sigma)$ *eloszlású véletlen változók. Ekkor* $\overline{\mathbf{X}} \sim \mathcal{N}_k(\mathbf{m}, \Sigma/n)$*,* $S \sim \mathcal{W}_k(n-1, \Sigma)$*, és* $\overline{\mathbf{X}}$ *és* S *függetlenek.*

Bizonyítás. Legyen $V \in \mathbb{R}^{n\times n}$ egy olyan ortogonális mátrix, melynek utolsó sora $(1/\sqrt{n}, \dots, 1/\sqrt{n})$, különben tetszőleges. Legyen

$$\mathbf{Y}_i = \sum_{j=1}^{n} v_{ij}\mathbf{X}_j \in \mathbb{R}^k.$$

Az $Y = (\mathbf{Y}_1, \dots, \mathbf{Y}_n) \in \mathbb{R}^{k\times n}$ jelöléssel

$$Y^\top = VX^\top.$$

Világos, hogy $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ együttesen normálisok, és egyszerű számolás mutatja, hogy

$$\mathbf{Cov}(\mathbf{Y}_i, \mathbf{Y}_j) = \delta_{ij}\Sigma,$$

azaz $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ függetlenek közös Σ kovarianciamátrixszal. Mivel U ortogonális, és az utolsó sor minden komponense egyenlő, ezért a többi sorösszeg 0, tehát

$$\mathbf{E}\mathbf{Y}_i = \delta_{in}\sqrt{n}\mathbf{m}.$$

Vegyük észre, hogy

$$\sum_{i=1}^{n} \mathbf{Y}_i\mathbf{Y}_i^\top = YY^\top = XV^\top VX = X^\top X = \sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i^\top,$$

és $\mathbf{Y}_n = \sqrt{n}\overline{\mathbf{X}}$. Tehát

$$\begin{aligned} S &= \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^\top \\ &= \sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i^\top - n\overline{\mathbf{X}}\overline{\mathbf{X}}^\top \\ &= \sum_{i=1}^{n} \mathbf{Y}_i\mathbf{Y}_i^\top - \mathbf{Y}_n\mathbf{Y}_n^\top \\ &= \sum_{i=1}^{n-1} \mathbf{Y}_i\mathbf{Y}_i^\top, \end{aligned}$$

amiből már adódik az állítás. □

1.4. Hotelling-féle T^2 eloszlás

Emlékeztetünk, hogy a Fisher-féle F-eloszlás független χ^2 eloszlású változók hányadosa. Azaz $F \sim \mathcal{F}(m, n)$, ha $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$, X és Y függetlenek, és

$$F = \frac{X/m}{Y/n};$$

informálisan $\mathcal{F}(m,n) = \frac{\chi^2(m)/m}{\chi^2(n)/n}$.

1.23. Tétel. *[Hotelling] Legyenek $W \sim W_k(n,\Sigma)$, $\boldsymbol{\xi} \sim \mathcal{N}_k(0,\Sigma)$ függetlenek. Ekkor $T^2 = \boldsymbol{\xi}^\top W^{-1}\boldsymbol{\xi}$ jelöléssel*

$$\frac{n-k+1}{k}T^2 \sim \mathcal{F}(k, n-k+1).$$

A bizonyítás a következő két segédtételen alapszik.

1.24. Lemma. *Legyen $X \in \mathbb{R}^{k\times n}$ független standard normálisokból álló mátrix, és $Q \in \mathbb{R}^{n\times n}$ egy X-től független véletlen ortogonális mátrix. Ekkor $XQ \in \mathbb{R}^{k\times n}$ elemei független standard normálisok, melyek függetlenek Q-tól.*

Bizonyítás. Ez determinisztikus Q-ra világos, és mivel Q és X függetlenek, ezért Q-ra feltételesen minden működik. □

1.25. Lemma. *Legyen $X \in \mathbb{R}^{k\times n}$, $n > k$, független standard normálisokból álló mátrix. Legyen $S = XX^\top \in \mathbb{R}^{k\times k}$, és $S_1 = (s_{ij})_{i,j\le k-1} \in \mathbb{R}^{(k-1)\times(k-1)}$, azaz S_1 az S utolsó sorának és oszlopának elhagyásával kapott mátrix. Ekkor*

$$\frac{|S|}{|S_1|} \sim \chi^2(n-k+1).$$

Bizonyítás. Legyen $Q \in \mathbb{R}^{n\times n}$ egy olyan *véletlen* ortogonális mátrix, melynek első oszlopa $Q_{\cdot 1} = R^{-1}X_{1\cdot}$, ahol $R = |X_{1\cdot}| = \sqrt{\sum_{i=1}^n X_{1i}^2}$. Ekkor

$$S = XX^\top = XQQ^\top X^\top = \begin{pmatrix} R & 0 & \dots & 0 \\ \widetilde{X}_{21} & \widetilde{X}_{22} & \dots & \widetilde{X}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{X}_{k1} & \widetilde{X}_{k2} & \dots & \widetilde{X}_{kn} \end{pmatrix} \begin{pmatrix} R & \widetilde{X}_{21} & \dots & \widetilde{X}_{k1} \\ 0 & \widetilde{X}_{22} & \dots & \widetilde{X}_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \widetilde{X}_{2n} & \dots & \widetilde{X}_{kn} \end{pmatrix}. \tag{4}$$

Másrészt

$$XQ = \begin{pmatrix} R & 0 & \dots & 0 \\ \widetilde{X}_{21} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{X}_{k1} & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \widetilde{X}_{22} & \dots & \widetilde{X}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \widetilde{X}_{kn} & \dots & \widetilde{X}_{kn} \end{pmatrix}, \tag{5}$$

ahol $\widetilde{X} = (\widetilde{X}_{ij})_{1\le i\le k-1; 1\le j\le n-1} \in \mathbb{R}^{(k-1)\times(n-1)}$ független standard normálisokból álló mátrix az előző lemma szerint. Ekkor a (4) és (5) formulák szerint $|S| = R^2|\widetilde{X}\widetilde{X}^\top|$. Ugyanezt el lehet játszani S_1-re, hiszen őt úgy kapjuk X-ből mint S-et, csak töröljük X utolsó sorát. Tehát $|S_1| = R^2|(\widetilde{X}\widetilde{X}^\top)_{i,j\le k-2}|$. Azaz

$$\frac{|S|}{|S_1|} = \frac{|\widetilde{X}\widetilde{X}^\top|}{|(\widetilde{X}\widetilde{X}^\top)_{i,j\le k-2}|},$$

ami éppen olyan hányados, mint a definícióban szereplő, csak független standard normálisokból álló $\widetilde{X} \in \mathbb{R}^{(k-1)\times(n-1)}$ véletlen mátrix $(k-1)\times(n-1)$-es.

Tehát elég $k = 2$-re igazolni az állítást, utána működik az indukció. Na de ekkor, az S_1 mátrix egy szám, és $\widetilde{X} \in \mathbb{R}^{n-1}$, ezért $|\widetilde{X}\widetilde{X}^\top| \sim \chi^2(n-1)$ így

$$\frac{|S|}{|S_1|} = \frac{R^2|\widetilde{X}\widetilde{X}^\top|}{R^2} = |\widetilde{X}\widetilde{X}^\top| \sim \chi^2(n-1),$$

amint állítottuk. □

Az 1.23. tétel bizonyítása.. A

$$T^2 = \boldsymbol{\xi}^t op\Sigma^{-1/2}(\Sigma^{-1/2}W\Sigma^{-1/2})^{-1}\Sigma^{-1/2}\boldsymbol{\xi}$$

előállításban $\Sigma^{-1/2}\boldsymbol{\xi} \sim \mathcal{N}_k(0, I_k)$ és $\Sigma^{-1/2}W\Sigma^{-1/2} \sim \mathcal{W}_k(n, I_k)$. Emiatt feltehető, hogy $\Sigma = I_k$. Legyen Q olyan $\boldsymbol{\xi}$-től függő ortogonális mátrix, hogy

$$\boldsymbol{\xi}^\top Q = (0, 0, \ldots, 0, |\boldsymbol{\xi}|)^\top.$$

Ezzel a transzformációval

$$T^2 = \boldsymbol{\xi}^\top QQ^\top W^{-1}QQ^\top\boldsymbol{\xi} = ((Q^\top WQ)^{-1})_{kk}|\boldsymbol{\xi}|^2,$$

ahol persze $|\boldsymbol{\xi}|^2 = \sum_{i=1}^k \xi_i^2 \sim \chi^2(k)$. Vegyük észre, hogy az egész $(Q^\top WQ)^{-1}$ nekünk csak a jobb alsó sarokban levő elem kell a $\boldsymbol{\xi}^\top Q$ vektor miatt. A W definíció szerint $W = XX^\top$ alakú, amit beírva

$$((Q^\top WQ)^{-1})_{kk} = ((Q^\top XX^\top Q)^{-1})_{kk}.$$

Az inverzmátrix jobb alsó elem, az inverzmátrix alakjából következően, a $(k-1)\times(k-1)$-es bal felső aldeterminháns és a determináns hányadosa. Az 1.25. Lemma szerint ennek reciproka éppen $\chi^2(n-k+1)$ eloszlású, és az 1.24. Lemma szerint ez független $\boldsymbol{\xi}$-től. Ezzel az állítást igazoltuk. □

1.5. Várható érték tesztelése

1.5.1. Ismert kovarianciamátrix esete

Egymintás eset. Legyenek $\mathbf{X}_1, \ldots, \mathbf{X}_n$ független $\mathcal{N}_k(\mathbf{m}, \Sigma)$ eloszlású vektorok, ahol a Σ pozitív definit kovarianciamátrix ismert. A következő hipotézist vizsgáljuk:

$$H_0 : \ \mathbf{m} = \mathbf{m}_0 \qquad H_A : \ \mathbf{m} \neq \mathbf{m}_0.$$

Ekkor H_0 fennállása esetén $\overline{\mathbf{X}}_n \sim \mathcal{N}_k(\mathbf{m}_0, n^{-1}\Sigma)$, és így az 1.10. Következmény szerint

$$t_1 = n \left(\overline{\mathbf{X}} - \mathbf{m}_0\right)^\top \Sigma^{-1} \left(\overline{\mathbf{X}} - \mathbf{m}_0\right) \sim \chi^2(k), \tag{6}$$

azaz a próba $\chi^2(k)$ eloszlású. Ez az egymintás u-próba többdimenziós változata.

Kétmintás eset. Legyenek $\mathbf{X}_1, \ldots, \mathbf{X}_n$ független $\mathcal{N}_k(\mathbf{m}, \Sigma)$ eloszlású vektorok és $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ az $\mathbf{X}$-ektől független $\mathcal{N}_k(\mathbf{m}', \Sigma)$ eloszlású vektorok, ahol a Σ pozitív definit kovarianciamátrix (közös!) ismert. A következő hipotézist vizsgáljuk:

$$H_0 : \ \mathbf{m} = \mathbf{m}' \qquad H_A : \ \mathbf{m} \neq \mathbf{m}'.$$

Ekkor az egymintás esethez hasonlóan H_0 fennállása esetén $\overline{\mathbf{X}}_n - \overline{\mathbf{Y}}_m \sim \mathcal{N}_k(0, (n^{-1} + m^{-1})\Sigma)$, és így az 1.10. Következmény szerint

$$t_2 = \frac{nm}{n+m} \left(\overline{\mathbf{X}}_n - \overline{\mathbf{Y}}_m\right)^\top \Sigma^{-1} \left(\overline{\mathbf{X}}_n - \overline{\mathbf{Y}}_m\right) \sim \chi^2(k),$$

azaz a próba $\chi^2(k)$ eloszlású. Ez az kétmintás u-próba többdimenziós változata.

1.5.2. Ismeretlen kovarianciamátrix esete

Egymintás eset. Legyenek $\mathbf{X}_1, \ldots, \mathbf{X}_n$ független $\mathcal{N}_k(\mathbf{m}, \Sigma)$ eloszlású vektorok, ahol a Σ pozitív definit kovarianciamátrix nem ismert. A következő hipotézist vizsgáljuk:

$$H_0 : \ \mathbf{m} = \mathbf{m}_0 \qquad H_A : \ \mathbf{m} \neq \mathbf{m}_0.$$

A (6) próbastatisztikában most a kovarianciamátrixot a ML becslésével (1.16. Tétel) helyettesítjük. Ekkor a Σ helyére

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^\top = \frac{1}{n} S$$

írva

$$T_1^2 = n\left(\overline{\mathbf{X}} - \mathbf{m}_0\right)^\top S^{-1}\left(\overline{\mathbf{X}} - \mathbf{m}_0\right) \tag{7}$$

próbastatisztikát kapjuk. Az 1.22. Tétel szerint S és $\overline{\mathbf{X}}$ függetlenek, és ezért H_0 fennállása esetén az 1.23. Tétel szerint $(n-k)/k \cdot T_1^2 \sim \mathcal{F}(k, n-k)$. Ez az egymintás t-próba többdimenziós változata.

Kétmintás eset. Legyenek $\mathbf{X}_1, \ldots, \mathbf{X}_n$ független $\mathcal{N}_k(\mathbf{m}, \Sigma)$ eloszlású vektorok és $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ az $\mathbf{X}$-ektől független $\mathcal{N}_k(\mathbf{m}', \Sigma)$ eloszlású vektorok, ahol a Σ pozitív definit kovarianciamátrix (közös!) ismeretlen. A következő hipotézist vizsgáljuk:

$$H_0: \ \mathbf{m} = \mathbf{m}' \qquad H_A: \ \mathbf{m} \neq \mathbf{m}'.$$

Legyen

$$S = \sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^\top + \sum_{i=1}^{m}(\mathbf{Y}_i - \overline{\mathbf{Y}})(\mathbf{Y}_i - \overline{\mathbf{Y}})^\top.$$

Ekkor S és $(\overline{\mathbf{X}} - \overline{\mathbf{Y}}$ függetlenek, így az egymintás esethez hasonlóan H_0 fennállása esetén

$$T_2^2 = \frac{nm}{n+m}\left(\overline{\mathbf{X}}_n - \overline{\mathbf{Y}}_m\right)^\top S^{-1}\left(\overline{\mathbf{X}}_n - \overline{\mathbf{Y}}_m\right)$$

statisztikára, az 1.23. Tétel szerint

$$\frac{n+m-k-1}{k} T_2^2 \sim \mathcal{F}(k, n+m-k-1).$$

1.6. Többdimenziós CHT

Egydimenziós esetben is láttuk, hogy a normális eloszlás azért különösen fontos, mert független véletlen változók összegének a normált és centrált határeloszlása. Ez a centrális határeloszlás-tétel, speciális esetben a de Moivre–Laplace tétel. Ennek a többváltozós megfelelője is igaz. Először az eloszlásbeli konvergencia fogalmát vezetjük be.

1.26. Definíció. Az $\mathbf{X}_n \in \mathbb{R}^k$ véletlen vektorváltozók *eloszlásban konvergálnak* $\mathbf{X} \in \mathbb{R}^k$ *véletlen vektorhoz*, ha tetszőleges olyan $\mathbf{x} = (x_1, \ldots, x_k) \in \mathbb{R}^k$ pontra, ami folytonossági pontja az

$$F(\mathbf{x}) = F(x_1, \ldots, x_k) = \mathbf{P}(\mathbf{X} \leq \mathbf{x}) = \mathbf{P}(X_1 \leq x_1, \ldots, X_k \leq x_k)$$

eloszlásfüggvénynek, teljesül, hogy

$$\lim_{n\to\infty} \mathbf{P}(\mathbf{X}_n \leq \mathbf{x}) = F(\mathbf{x}).$$

Mivel a többdimenziós normális eloszlásfüggvénye folytonos, ezért a konvergencia minden pontban teljesül. Ezek után kimondjuk a többdimenziós CHT-t.

1.27. Tétel. *CHT független, azonos eloszlású véletlen vektorokra Legyenek* $\mathbf{X}_1$*,* $\mathbf{X}_2$*, ... független azonos eloszlású véletlen vektorok* $\mathbb{R}^k$*-ban véges* Σ *kovarianciamátrixszal és* $\mathbf{m}$ *várható érték vektorral. Ekkor az* $\mathbf{S}_n = \sum_{j=1}^n \mathbf{X}_j$ *,* $n \in \mathbb{N}$ *, részletösszegekre*

$$\frac{1}{\sqrt{n}}\Big(\mathbf{S}_n - n\mathbf{m}\Big) \xrightarrow{\mathcal{D}} \mathbf{Y}, \quad \textit{amint} \ \ n \to \infty\,, \quad \textit{ahol} \ \ \mathbf{Y} \sim \mathcal{N}_k(0, \Sigma)\,.$$

Emiatt a tétel miatt a többdimenziós normálisra kapott próbák érvényesek tetszőleges olyan véletlen vektorra, melynek létezik kovarianciamátrixa. Ekkor persze csak nagy mintaelemszám esetén ($n \to \infty$) érvényes a kapott eredmény.

2. Lineáris módszerek

2.1. Főkomponensanalízis

Legyen $\mathbf{X} \sim \mathcal{N}_k(\mathbf{m}, \Sigma)$. Keressük $\mathbf{X}$ előállítását

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

alakban, ahol $V \in \mathbb{R}^{k\times k}$ ortogonális mátrix, $\mathbf{Y} \in \mathbb{R}^k$ független komponensekből álló k-dimenziós normális eloszlású véletlen vektor, 0 várható érték vektorral. Feltesszük továbbá, hogy $\mathbf{Y}$ komponenseinek a szórásnégyzete csökkenő. Mivel V ortogonális, így $\mathbf{Y} = V^\top(\mathbf{X} - \mathbf{m})$, és

$$\mathbf{E}\mathbf{Y}\mathbf{Y}^\top = V^\top U \Lambda U^\top V,$$

ahol a $\Sigma = U\Lambda U^\top$ spektrálfelbontást használtuk. Mivel $\mathbf{E}\mathbf{Y}\mathbf{Y}^\top$ diagonális, így $V = U$, és a $\mathbf{Y} = U^\top(\mathbf{X}-\mathbf{m})$ előállítást kapjuk. Az $\mathbf{Y}$ vektor komponenseit $\mathbf{X}$ *főkomponenseinek* nevezzük. Vegyük észre, hogy $\mathbf{Y}$ komponensei $\mathbf{X}-\mathbf{m}$ komponenseinek lineáris kombinációja; pontosabban, az $\mathbf{Y} = (Y_1, \ldots, Y_k)^\top$ jelöléssel

$$Y_i = \langle \mathbf{u}_i, \mathbf{X} - \mathbf{m}\rangle, \quad i = 1, \ldots, k,$$

ahol $\mathbf{u}_1, \ldots, \mathbf{u}_k$ a Σ sajátvektorai.

A főkomponensfelbontás az alábbi optimalitási tulajdonság miatt fontos.

2.1. Tétel. *Az Y_1 szórásnégyzete maximális az $\langle \mathbf{v}, \mathbf{X} - \mathbf{m} \rangle$ változók szórásnégyzetei között, ahol $|\mathbf{v}| = 1$, továbbá a maximum a $\mathbf{v} = \mathbf{u}_1$ sajátvektoron vétetik fel, értéke λ_1. Az Y_2 szórásnégyzete maximális az Y_1-től független $\langle \mathbf{v}, \mathbf{X} - \mathbf{m} \rangle$ változók szórásnégyzetei között, ahol $|\mathbf{v}| = 1$, továbbá a maximum a $\mathbf{v} = \mathbf{u}_2$ sajátvektoron vétetik fel, értéke λ_2. Általánosan, Y_ℓ, $\ell \in \{2, \ldots, k\}$, szórásnégyzete maximális az $Y_1, \ldots, Y_{\ell-1}$ változóktól független $\langle \mathbf{v}, \mathbf{X} - \mathbf{m} \rangle$ változók szórásnégyzetei között, ahol $|\mathbf{v}| = 1$, továbbá a maximum a $\mathbf{v} = \mathbf{u}_\ell$ sajátvektoron vétetik fel, értéke λ_ℓ.*

A bizonyítás a sajátvektorok következő maximumtulajdonságán múlik.

2.2. Tétel. *Legyen $A \in \mathbb{R}^{k \times k}$ szimmetrikus mátrix, és legyenek $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ az A sajátértékei, $\mathbf{u}_1, \ldots, \mathbf{u}_k$ pedig a hozzájuk tartozó sajátvektorok. Ekkor*

$$\max_{|\mathbf{v}|=1} \mathbf{v}^\top A \mathbf{v} = \lambda_1,$$

és a maximum a $\mathbf{u}_1$ vektoron vétetik fel. Továbbá, tetszőleges $\ell \in \{2, \ldots, k\}$ esetén

$$\max \left\{ \mathbf{v}^\top A \mathbf{v} \, : \, |\mathbf{v}| = 1, \mathbf{v} \perp \mathrm{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_{\ell-1}\} \right\} = \lambda_\ell,$$

és a maximum a $\mathbf{u}_\ell$ sajátvektoron vétetik fel.

Bizonyítás. Az egyszerűség kedvéért tegyük fel, hogy $\lambda_1 > \lambda_2 > \ldots > \lambda_k$, azaz minden sajátérték multiplicitása egy. A bizonyításból világos, hogy különben az egyértelműség nem igaz.

Tekintsük az $A = U \Lambda U^\top$ spektrálfelbontást. Legyen $\mathbf{v} \in \mathbb{R}^k$ tetszőleges, és tekintsük az $\mathbf{u}_1, \ldots, \mathbf{u}_k$ bázisban való kifejtését:

$$\mathbf{v} = \sum_{i=1}^{k} \alpha_i \mathbf{u}_i.$$

Ekkor, ha $|\mathbf{v}|^2 = \sum_{i=1}^{k} \alpha_i^2 = 1$, akkor

$$\mathbf{v}^\top A \mathbf{v} = \sum_{i=1}^{k} \lambda_i \alpha_i^2 \leq \lambda_1,$$

és egyenlőség pontosan akkor van, ha $\mathbf{v} = \mathbf{u}_1$. Ez persze csak akkor igaz, ha $\lambda_1 > \lambda_2$. Ha vannak egyenlő sajátértékek, akkor nincs teljes egyértelműség, a megfelelő sajátaltérben a bázis tetszőlegesen választható.

A $\mathbf{v}$ pontosan akkor merőleges az $\mathbf{u}_1, \ldots, \mathbf{u}_{\ell-1}$ vektorok által feszített altérre, ha $\alpha_1 = \ldots = \alpha_{\ell-1} = 0$. Ekkor, ha $|\mathbf{v}| = 1$,

$$\mathbf{v}^\top A\mathbf{v} = \sum_{i=\ell}^{k} \lambda_i \alpha_i^2 \leq \lambda_\ell,$$

és egyenlőség pontosan akkor van, ha $\mathbf{v} = \mathbf{u}_\ell$. Ezzel a tételt igazoltuk. $\square$

A 2.1. Tétel bizonyítása. Tetszőleges $\mathbf{v} \in \mathbb{R}^k$ esetén

$$\mathbf{E}(\langle \mathbf{v}, \mathbf{X} - \mathbf{m} \rangle^2) = \mathbf{v}^\top \Sigma \mathbf{v}.$$

Tehát a feladat a 2.2. Tételben tárgyalt maximumfeladatra egyszerűsödött. Azt kell még észrevenni, hogy az együttes normalitás miatt $\langle \mathbf{v}, \mathbf{X} - \mathbf{m} \rangle$ pontosan akkor független a $\langle \mathbf{u}_1, \mathbf{X} - \mathbf{m} \rangle$, ..., $\langle \mathbf{u}_{\ell-1}, \mathbf{X} - \mathbf{m} \rangle$ változóktól, ha

$$0 = \mathbf{E}(\langle \mathbf{u}_i, \mathbf{X} - \mathbf{m} \rangle \, \langle \mathbf{v}, \mathbf{X} - \mathbf{m} \rangle) = \lambda_i \langle \mathbf{v}, \mathbf{u}_i \rangle.$$

Azaz a függetlenségi megkötés éppen a 2.2. Tételben szereplő merőlegességi feltételt jelenti. Tehát minden következik a 2.2. Tételből. $\square$

2.2. Merőleges vetítés

2.3. Definíció. Legyen $\mathcal{H}$ egy lineáris vektortér a valós számtest felett. Ekkor $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ *belső szorzat*, ha tetszőleges $x, y, z \in \mathcal{H}$, $\alpha \in \mathbb{R}$ esetén

(i) $\langle x, y \rangle = \langle y, x \rangle$;

(ii) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$;

(iii) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$;

(iv) $\langle x, x \rangle \geq 0$, és pontosan akkor 0, ha $x = 0$.

2.4. Feladat. Legyen $\mathcal{H} = \mathbb{R}^k$, $k \geq 1$. Igazoljuk, hogy $\langle x, y \rangle = \sum_{i=1}^k x_i y_i$ belső szorzat!

Ekkor $(\mathcal{H}, \langle \rangle)$, vagy egyszerűen csak $\mathcal{H}$ belső szorzattér. Belső szorzattéren használjuk az euklideszi terekben megszokott terminológiát. Egy $x \in \mathcal{H}$ vektor *normája* $\|x\| = \sqrt{\langle x, x \rangle}$. Az x *merőleges* y-ra, ha $\langle x, y \rangle = 0$.

2.5. Állítás. *Norma tulajdonságai.*

(i) Teljesül a háromszög-egyenlőtlenség: $\|x + y\| \leq \|x\| + \|y\|$*;*

(ii) $\|\alpha x\| = |\alpha|\|x\|$*;*

(iii) $\|x\| \geq 0$ *és pontosan akkor 0, ha* $x = 0$*;*

(iv) Cauchy–Bunyakovszkij–Schwarz egyenlőtlenség: $\|\langle x, y\rangle \leq \|x\|\,\|y\|$*;*

(v) Paralelogramma azonosság: $\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2$.

Az első három tulajdonság a norma definiáló tulajdonsága.

2.6. Feladat. Bizonyítsuk be a fenti állítást!

A normából származik távolság, ami definiál egy konvergenciát. Akkor mondjuk, hogy $x_n \to x$ amint $n \to \infty$ ha $\|x_n - x\| \to 0$. Ekkor pedig van folytonosság.

2.7. Állítás. *A norma és a belső szorzat folytonosak. Azaz, ha* $x_n \to x$ *akkor* $\|x_n\| \to \|x\|$*, és ha* $x_n \to x$ *és* $y_n \to y$ *akkor* $\langle x_n, y_n\rangle \to \langle x, y\rangle$.

2.8. Feladat. Bizonyítsuk be a fenti állítást!

2.9. Definíció. Az (x_n) sorozat *Cauchy-sorozat*, ha minden $\varepsilon > 0$ esetén megadható n_0, hogy bármely $m, n \geq n_0$ esetén $\|x_n - x_m\| \leq \varepsilon$. Akkor mondjuk, hogy a $\mathcal{H}$ belső szorzattér *Hilbert-tér*, ha benne minden Cauchy-sorozat konvergens.

Analízisből tudjuk, hogy $\mathbb{R}^n$ Hilbert-tér.

2.10. Példa. Számunkra a legfontosabb Hilbert-tér a négyzetintegrálható véletlen változók tere. Legyen $(\Omega, \mathcal{A}, \mathbf{P})$ egy valószínűségi mező. Legyen $L^2 = \{X : \Omega \to \mathbb{R}; \mathbf{E}X^2 < \infty\}$, azaz a véges második momentummal rendelkező véletlen változók. Ekkor L^2 lineáris vektortér. Ezt láttuk valószínűségszámításból, hiszen csak annyi kell, hogy $X, Y \in L^2$ akkor $X + Y \in L^2$. Ez a Cauchy–Schwarz egyenlőtlenségből következik. Azt is könnyű látni, hogy $\langle X, Y\rangle = \mathbf{E}(XY)$ belső szorzat.

Az már lényegesen bonyolultabb, hogy L^2 Hilbert-tér. Ez majd mértékelméletből lesz.

2.11. Feladat. Tekintsük az előző példában látott L^2 teret. Legyenek X, X_1, X_2, $\ldots$ független standard normálisok, és legyen (a_n) determinisztikus sorozat. Igazoljuk, hogy $S_n = \sum_{i=1}^n a_i X_i$ pontosan akkor Cauchy-sorozat, ha $\sum_{i=1}^\infty a_i^2 < \infty$.

2.12. Definíció. Az $\mathcal{M} \subset \mathcal{H}$ lineáris altér *zárt*, ha minden torlódási pontját tartalmazza. Az $\mathcal{M}$ *ortogonális komplementere*

$$\mathcal{M}^{\perp} = \{x : x \in \mathcal{H}, \forall y \in \mathcal{M},\ \langle x, y \rangle = 0\}.$$

Ez az $\mathcal{M}$-re merőleges altér. A belső szorzat folytonosságából azonnal adódik, hogy $\mathcal{M}^{\perp}$ mindig zárt.

2.13. Példa. Nem minden lineáris altér zárt. Legyen $\mathcal{H}$ a $[0,1]$-en folytonos függvények tere. Ez Hilbert-tér a $\langle f, g \rangle = \int_0^1 f(x)g(x)\mathrm{d}x$ belső szorzattal. Ekkor a polinomok (szokásos véges fokszámú valós együtthatós) altere nem zárt. Hát persze, a Weierstrass-féle approximációs tétel szerint minden folytonos függvény közelíthető polinomokkal, azaz a polinomok alterének torlódási pontjai éppen a folytonos függvények halmaza, azaz az egész tér.

A következő tételt nem bizonyítjuk, viszont sokszor használjuk. A lényeg, hogy három dimenzióban értsük, hogy mi történik.

2.14. Tétel. *Merőleges vetítés. Legyen $\mathcal{M} \subset \mathcal{H}$ zárt altér, és legyen $x \in \mathcal{H}$. Ekkor létezik egy egyértelmű $\widehat{x} \in \mathcal{M}$, hogy $\|x - \widehat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|$. Továbbá*

$$\left(\widehat{x} \in \mathcal{M}, \|x - \widehat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\| \right) \Leftrightarrow \left(\widehat{x} \in \mathcal{M},\ x - \widehat{x} \in \mathcal{M}^{\perp} \right).$$

A második állítás mondja meg, hogy hogyan kell választani az $\widehat{x}$ vektort: a különbség vektor merőleges az altérre.

Tetszőleges $\mathcal{M}$ zárt altér esetén legyen $P_{\mathcal{M}} : \mathcal{H} \to \mathcal{M}; x \mapsto \widehat{x}$, az $\mathcal{M}$ altérre való merőleges vetítés. A merőleges vetítés néhány fontos tulajdonságait tartalmazza az alábbi állítás.

2.15. Állítás. *Legyen $\mathcal{M}$ zárt altér, és $P = P_{\mathcal{M}}$. Ekkor*

(i) $P(\alpha x + \beta y) = \alpha P(x) + \beta P(y)$;

(ii) Pitagorasz-tétel: $\|x\|^2 = \|Px\|^2 + \|(I - P)x\|^2$;

(iii) $\exists! u \in \mathcal{M}, v \in \mathcal{M}^{\perp}$, $x = u + v$, *és* $u = Px$, $v = (I - P)x$;

(iv) $Px_n \to Px$ *valahányszor* $x_n \to x$;

(v) $Px = x$ *pontosan akkor, ha* $x \in \mathcal{M}$;

(vi) $Px = 0$ *pontosan akkor, ha* $x \in \mathcal{M}^{\perp}$.

2.3. Lineáris regresszió véletlen regresszorral

Legyenek $Y, X_1, X_2, \ldots, X_k$, $k \geq 1$, véletlen változók. Keressük az Y független változó legjobb lineáris közelítését az $\mathbf{X} = (X_1, \ldots, X_k)^\top$ függő változókkal. Pontosabban keressük az $a_1, \ldots, a_k, b$ valós számokat, melyekre a

$$\mathbf{E}\left((Y - \sum_{i=1}^{k} a_i X_i - b)^2\right)$$

minimális. Mivel $\mathbf{E}(Z-b)^2$ minimuma a $b = \mathbf{E}Z$ helyen van, ezért $b = \mathbf{E}Y - \sum_i a_i \mathbf{E}X_i$. Vagyis az általánosság megszorítása nélkül feltehetjük, és fel is tesszük, hogy $\mathbf{E}Y = \mathbf{E}X_i = 0$ minden i-re.

Tekintsük az L^2 teret a szokásos $\mathbf{E}(UV)$ belső szorzattal, és legyen

$$\mathcal{M} = \{\sum_{i=1}^{k} \alpha_i X_i : \alpha_i \in \mathbb{R}\} = \operatorname{span}(X_1, \ldots, X_k).$$

Ekkor $\mathcal{M}$ zárt, hiszen véges dimenziós altér. A merőleges vetítés tétele szerint a legjobban közelítő vektor PY, az Y merőleges vetítése az $\mathcal{M}$ altérre. Tehát a feladatunk a PY meghatározása. A merőleges vetítés tétele szerint $Y - PY \perp \mathcal{M}$, ami azt jelenti, hogy

$$\mathbf{E}(Y - PY)X_j = 0, \quad \text{minden } j = 1, 2, \ldots, k \text{ esetén.}$$

Mivel $PY = \sum_{i=1}^{k} a_i X_i$ valamilyen $\mathbf{a}^\top = (a_1, \ldots, a_k)$ vektorra, ezért azt kapjuk, hogy

$$0 = \mathbf{E}\left[(Y - \sum_{i=1}^{k} a_i X_i)X_j\right] = \mathbf{Cov}(Y, X_j) - \mathbf{a}^\top \Sigma, \quad j = 1, 2, \ldots, k,$$

ahol $\Sigma = \mathbf{Cov}(\mathbf{X})$ az $\mathbf{X} = (X_1, \ldots, X_k)^\top$ kovarianciamátrix.

A $\mathbf{d} = (\mathbf{Cov}(Y, X_i))_{i=1}^{k} \in \mathbb{R}^k$ jelöléssel azt kapjuk, hogy $\mathbf{d} = \Sigma \mathbf{a}$, azaz, amennyiben Σ nemszinguláris,

$$\mathbf{a} = \Sigma^{-1} \mathbf{d}.$$

Ezzel beláttuk a következőt.

2.16. Tétel. *Legkisebb négyzetek módszere. Az* $\mathbf{E}\left((Y - \sum_{i=1}^{k} a_i X_i - b)^2\right)$ *négyzetes hiba pontosan akkor minimális, ha* $\mathbf{a} = \Sigma^{-1}\mathbf{d}$ *és* $b = \mathbf{E}Y - \mathbf{a}^\top \mathbf{E}\mathbf{X}$.

A vetítés miatt a közelítés hibája, az $\ell(\mathbf{X}) = \sum_i a_i X_i + b$ jelöléssel

$$\varepsilon = Y - \left(\sum_{i=1}^{k} a_i X_i + b \right) = Y - \ell(\mathbf{X}),$$

merőleges az $X_1, \dots, X_k$ változókra, azaz $\mathbf{Cov}(\varepsilon, X_i) = 0, i = 1, \dots, k$. Vagyis az $Y = \ell(\mathbf{X}) + \varepsilon$ előállításban a tagok korrelálatlanok, és így

$$\mathbf{D}^2(Y) = \mathbf{D}^2(\ell(\mathbf{X})) + \mathbf{D}^2(\epsilon).$$

Ebből következik, hogy

$$\mathbf{Cov}(\ell(\mathbf{X}), Y) = \mathbf{D}^2(\ell(\mathbf{X})).$$

2.17. Definíció. A Y független és az $\mathbf{X}^\top = (X_1, \dots, X_k)$ függő változók *többszörös korrelációs együtthatója*

$$r_{Y(X_1,\dots,X_k)} = \rho(Y, \ell(\mathbf{X})) = \frac{\mathbf{Cov}(Y, \ell(\mathbf{X}))}{\mathbf{D}(Y)\mathbf{D}(\ell(\mathbf{X}))}.$$

A $k = 1$ esetben ez éppen a két változó hagyományos korrelációja.

Némi számolással adódik, hogy

$$\mathbf{D}^2(\varepsilon) = \mathbf{D}^2(Y)\,(1 - r^2).$$

Ez megmagyarázza az r jelentését. Ha $r = 1$, akkor $\varepsilon \equiv 0$, azaz lineáris függvénykapcsolat van Y és $\mathbf{X}$ között. Ha pedig $r = 0$, akkor az $\mathbf{X}$ semmit nem magyaráz meg az Y-ből, a két változó között nincs kapcsolat. Ekkor $\mathbf{a} = (0, \dots, 0)$.

A következő állítás szerint a legkisebb négyzetes közelítés maximalizálja a korrelációt.

2.18. Állítás. *Az $X_1, \dots, X_k$ változók tetszőleges $h(\mathbf{X})$ lineáris kombinációjára*

$$|r_{Y(X_1,\dots,X_k)}| = \rho(Y, \ell(\mathbf{X})) \geq |\rho(Y, h(\mathbf{X}))|.$$

2.4. Determinisztikus változók

Vetítés. Gauss-féle normálegyenlet. Gauss–Markov tétel. Hipotézisvizsgálat.

2.5. Varianciaanalízis

Egyszempontos. Vetítés explicit formában. Kétszempontos, interakció. Hipotézisvizsgálat.

Hivatkozások

[1] Bolla Marianna, Krámli András: Statisztikai következtetések elmélete. Typotex, 2005.

[2] Csörgő Sándor: Fejezetek a valószínűségelméletből. Polygon, 2010.

[3] Richard A. Johnson, Dean W. Wichern: Applied Multivariate Statistical Analysis. Prentice-Hall, 1988.