Faktorová analýza

(přednáška v rámci předmětu Statistika a spolehlivost v lékařství)

Mirko Navara
Centrum strojového vnímání
katedra kybernetiky FEL ČVUT
Karlovo náměstí, budova G, místnost 104a
https://cw.felk.cvut.cz/doku.php/courses/a6m33ssl

25. prosince 2011

Cíl: Popsat *podstatné* závislosti mnoha náhodných veličin pomocí menšího počtu *společných faktorů*.
(Co nejjednodušší popis s malou ztrátou informace.)
$\boldsymbol{X} = (X_1, \dots, X_p)^\top$ = náhodný vektor tvořený p normovanými náhodnými veličinami, $\mathrm{E}X_i = 0$, $\mathrm{D}X_i = 1$,
$\Sigma_{\boldsymbol{X}} = \mathrm{E}(\boldsymbol{X}\,\boldsymbol{X}^\top)$ = kovarianční matice.
Lineární závislosti (=korelace) složek náhodného vektoru $\boldsymbol{X}$ chceme vysvětlit jako důsledek následujících náhodných veličin:
$F_1, \dots, F_m$, $m < p$, ... **společné faktory**, které vysvětlují korelace
$R_1, \dots, R_p$ ... **specifické faktory** (=*chybové faktory*=*reziduální složky*), které vysvětlují zbytkový rozptyl
Model:

$$\begin{aligned} X_1 &= w_{11}\,F_1 + w_{12}\,F_2 + \ldots + w_{1m}\,F_m + R_1\,, \\ X_2 &= w_{21}\,F_1 + w_{22}\,F_2 + \ldots + w_{2m}\,F_m + R_2\,, \\ &\dots \\ X_p &= w_{p1}\,F_1 + w_{p2}\,F_2 + \ldots + w_{pm}\,F_m + R_p\,, \end{aligned}$$

kde F_j jsou *normované*, $\mathrm{E}F_j = 0$, $\mathrm{D}F_j = 1$, a *nekorelované*, $\mathrm{E}(F_j\,F_k) = 0$,
R_i jsou *centrované*, $\mathrm{E}R_i = 0$, a nezávislé navzájem i na všech F_j, $\mathrm{E}(E_j\,E_k) = \mathrm{E}(E_j\,F_k) = 0$.
Maticový zápis pomocí $\boldsymbol{F} = (F_1, \dots, F_m)^\top$, $\boldsymbol{R} = (R_1, \dots, R_p)^\top$ a matice $\boldsymbol{W} \in \mathbb{R}^{p\times m}$:

$$\boldsymbol{X} = \boldsymbol{W}\,\boldsymbol{F} + \boldsymbol{R}\,,$$

kde $\mathrm{E}\boldsymbol{R} = \boldsymbol{O}_p = (0, \dots, 0) \in \mathbb{R}^p$,
$\Sigma_{\boldsymbol{R}} = \mathrm{E}(\boldsymbol{R}\,\boldsymbol{R}^\top)$ je *diagonální* kovarianční matice,
$\mathrm{E}\boldsymbol{F} = \boldsymbol{O}_m = (0, \dots, 0) \in \mathbb{R}^m$,
$\Sigma_{\boldsymbol{F}} = \mathrm{E}(\boldsymbol{F}\,\boldsymbol{F}^\top) = \boldsymbol{I}_m \in \mathbb{R}^{m\times m}$ je jednotková matice,
$\mathrm{E}(\boldsymbol{F}\,\boldsymbol{R}^\top) = \boldsymbol{O}_{m,p} \in \mathbb{R}^{m\times p}$ je nulová matice.
Vztah mezi kovariančními maticemi:

$$\begin{aligned} \Sigma_{\boldsymbol{X}} &= \mathrm{E}(\boldsymbol{X}\,\boldsymbol{X}^\top) = \mathrm{E}((\boldsymbol{W}\,\boldsymbol{F} + \boldsymbol{R})\,(\boldsymbol{W}\,\boldsymbol{F} + \boldsymbol{R})^\top) = \mathrm{E}((\boldsymbol{W}\,\boldsymbol{F}\,\boldsymbol{F}^\top\,\boldsymbol{W}^\top + \boldsymbol{W}\,\boldsymbol{F}\,\boldsymbol{R}^\top + \boldsymbol{R}\,\boldsymbol{F}^\top\,\boldsymbol{W}^\top + \boldsymbol{R}\,\boldsymbol{R}^\top) = \\ &= \mathrm{E}(\boldsymbol{W}\,\boldsymbol{F}\,\boldsymbol{F}^\top\,\boldsymbol{W}^\top) + \mathrm{E}(\boldsymbol{W}\,\boldsymbol{F}\,\boldsymbol{R}^\top) + \mathrm{E}(\boldsymbol{R}\,\boldsymbol{F}^\top\,\boldsymbol{W}^\top) + \mathrm{E}(\boldsymbol{R}\,\boldsymbol{R}^\top) = \\ &= \boldsymbol{W}\,\underbrace{\mathrm{E}(\boldsymbol{F}\,\boldsymbol{F}^\top)}_{\boldsymbol{I}_m}\,\boldsymbol{W}^\top + \boldsymbol{W}\,\underbrace{\mathrm{E}(\boldsymbol{F}\,\boldsymbol{R}^\top)}_{\boldsymbol{O}_{m,p}} + \underbrace{\mathrm{E}(\boldsymbol{R}\,\boldsymbol{F}^\top)}_{\boldsymbol{O}_{p,m}}\,\boldsymbol{W}^\top + \underbrace{\mathrm{E}(\boldsymbol{R}\,\boldsymbol{R}^\top)}_{\Sigma_{\boldsymbol{R}}} = \\ &= \boldsymbol{W}\,\boldsymbol{W}^\top + \Sigma_{\boldsymbol{R}}\,. \end{aligned}$$

První sčítanec vysvětlují *společné* faktory; má být co největší.
Druhý sčítanec vysvětlují *specifické* faktory; má být co nejmenší.
Kovarianční matici $\Sigma_{\boldsymbol{X}}$ tvoří prvky na diagonále,

$$(\Sigma_{\boldsymbol{X}})_{ii} = \mathrm{D}X_i = \sum_{j=1}^{m} w_{ij}^2 + \mathrm{D}R_i\,,$$

a prvky mimo diagonálu ($i \neq k$),

$$(\Sigma_{\boldsymbol{X}})_{ik} = \mathrm{Cov}(X_i, X_k) = \sum_{j=1}^{m} w_{ij}\, w_{kj}\,.$$

$$\boldsymbol{W}\,\boldsymbol{W}^\top = \Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{R}}$$

je **redukovaná kovarianční matice**. Její i-tý prvek na diagonále,

$$\left(\boldsymbol{W}\,\boldsymbol{W}^\top\right)_{ii} = \sum_{j=1}^{m} w_{ij}^2 = \mathrm{D}X_i - \mathrm{D}R_i\,,$$

se nazývá **komunalita** proměnné X_i; představuje část rozptylu $\mathrm{D}X_i$, vysvětlenou *společnými* faktory; zbytek, $\mathrm{D}R_i$, je **specifický rozptyl**.
Význam koeficientů w_{ij}:

$$w_{ij} = \mathrm{Cov}(X_i, F_j)$$

Důkaz:
A. Po složkách:

$$\begin{aligned}\mathrm{Cov}(X_i, F_j) &= \mathrm{E}(X_i\,F_j) = \mathrm{E}((w_{i1}\,F_1 + w_{i2}\,F_2 + \ldots + w_{im}\,F_m + R_i)\,F_j) = \\ &= \mathrm{E}(w_{i1}\,F_1\,F_j) + \mathrm{E}(w_{i2}\,F_2\,F_j) + \ldots + \mathrm{E}(w_{im}\,F_m\,F_j) + \mathrm{E}(R_i\,F_j) = \\ &= w_{i1}\,\mathrm{E}(F_1\,F_j) + w_{i2}\,\mathrm{E}(F_2\,F_j) + \ldots + w_{im}\,\mathrm{E}(F_m\,F_j) + \mathrm{E}(R_i\,F_j) = \\ &= w_{ij}\,.\end{aligned}$$

B. Maticový:

$$\begin{aligned}\mathrm{E}(\boldsymbol{X}\,\boldsymbol{F}^\top) &= \mathrm{E}((\boldsymbol{W}\,\boldsymbol{F} + \boldsymbol{R})\,\boldsymbol{F}^\top) = \mathrm{E}(\boldsymbol{W}\,\boldsymbol{F}\,\boldsymbol{F}^\top + \boldsymbol{R}\,\boldsymbol{F}^\top) = \\ &= \boldsymbol{W}\,\underbrace{\mathrm{E}(\boldsymbol{F}\,\boldsymbol{F}^\top)}_{\boldsymbol{I}_m} + \underbrace{\mathrm{E}(\boldsymbol{R}\,\boldsymbol{F}^\top)}_{\boldsymbol{O}_{p,m}} = \boldsymbol{W}\end{aligned}$$

Úloha: K dané (symetrické pozitivně definitní) kovarianční matici $\Sigma_{\boldsymbol{X}}$ hledáme diagonální matici $\Sigma_{\boldsymbol{R}}$ (s nezápornými prvky $\mathrm{D}R_i$ na diagonále) takovou, aby rozdíl $\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{R}}$ byl tvaru $\boldsymbol{W}\,\boldsymbol{W}^\top$, kde matice $\boldsymbol{W}$ je typu $p \times m$ (takže $\boldsymbol{W}\,\boldsymbol{W}^\top$ má hodnost nejvýše m).
Řešitelnost: Pomocí
$p\,m$ koeficientů matice $\boldsymbol{W}$ a
p koeficientů diagonální matice $\Sigma_{\boldsymbol{R}}$
chceme vysvětlit
$\frac{p\,(p-1)}{2}$ korelací (resp. koeficientů symetrické matice $\Sigma_{\boldsymbol{X}}$).
Pro $m = p$ řešení existuje a lze volit $\Sigma_{\boldsymbol{R}} = \boldsymbol{0}_{p,p}$.
Nás ale zajímá případ $m \ll p$; pro řešitelnost známe podmínky

- nutné
- nebo postačující,
- nikoli však nutné a postačující (otevřený problém).

Jednoznačnost: Pro $m > 1$ není řešení jednoznačné.
Je určen jen lineární podprostor (prostoru $\mathcal{N}$ všech náhodných veličin, které mají nulovou střední hodnotu a konečný rozptyl), který je lineárním obalem vektorů $F_1, \ldots, F_m$.
Za společné faktory můžeme volit i jakoukoli jinou jeho bázi (kterou dostaneme **rotací faktorů**).
Proto interpretace faktorů je problematická.
Jednoznačnost může obvykle zajistit **kanonický tvar faktorového modelu**: Faktory vyčerpávají postupně maximum celkového rozptylu (první faktor nejvíc atd.).
Problémy:

- Numerické řešení (např. metoda maximální věrohodnosti za předpokladu normálního rozdělení náhodných veličin).
- Počet faktorů (existují testy posuzující vhodnou složitost modelu podle jeho přesnosti).
- Interpretace faktorů (případně jejich rotace na lépe interpretovatelné veličiny).

Literatura

[Hebák] Hebák, P., Hustopecký, J.: *Vícerozměrné statistické metody s aplikacemi.* SNTL, Praha, 1987.

[Navara: PMS] Navara, M.: *Pravděpodobnost a matematická statistika.* Skriptum ČVUT, Praha, 2007.