Faktorová analýza

Ekonometrie

Jiří Neubauer

Katedra ekonometrie FVL UO Brno
kancelář 69a, tel. 973 442029
email:Jiri.Neubauer@unob.cz

Faktorová analýza – úvod

Na sledovaných objektech se často zjišťují hodnoty většího počtu statistických znaků nebo proměnných, ty jsou obvykle mezi sebou korelovány. Lze si představit, že korelace mezi jednotlivými proměnnými jsou způsobeny vlivem nějakého menšího počtu nepřímo měřitelných společných **faktorů**, které ovlivňují hodnoty sledovaných proměnných. Každý z těchto faktorů může ovlivnit hodnoty pozorování každé ze zkoumaných proměnných.

Cílem faktorové analýzy pak je tyto faktory odhadnout, odhadnout počet statisticky významných faktorů a konečně odhadnout hodnoty každého z faktorů pro každý sledovaný objekt, tedy popsat objekty pomocí nalezených faktorů. V tomto směru je faktorová analýza metodou pro snížení rozsahu dat. Navíc sledováním realizací společných faktorů u jednotlivých objektů mohou být identifikovány výjimečné objekty.

Ortogonální faktorový model

Ve faktorovém modelu předpokládáme, že p-rozměrný sloupcový náhodný vektor sledovaných proměnných $\boldsymbol{X} = (X_1, \ldots, X_p)'$ se střední hodnotou $\boldsymbol{\mu}$ a varianční maticí $\boldsymbol{\Sigma}$ lineárně závisí na m-rozměrném náhodném vektoru společných faktorů $\boldsymbol{F} = (F_1, \ldots, F_m)'$ a p-rozměrném vektoru specifických faktorů $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)'$. Maticově lze faktorový model zapsat

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\epsilon},$$

kde $\boldsymbol{L}$ je matice typu $p \times m$. Speciálně pro i-tou proměnnou X_i platí

$$X_i - \mu_i = l_{i1}F_1 + l_{i2}F_2 + \cdots + l_{im}F_m + \epsilon_i.$$

V uvedeném modelu popisuje koeficient l_{ij} vliv faktoru F_j na proměnnou X_i, a proto se matice $\boldsymbol{L}$ nazývá **matice faktorových zátěží**.

Ortogonální faktorový model

Vektor společných faktorů $\boldsymbol{F}$ a vektor specifických faktorů $\boldsymbol{\epsilon}$ nelze přímo pozorovat. Aby bylo možné je odhadnout, jsou na ně kladeny následující předpoklady.

1. Střední hodnoty obou těchto vektorů jsou nulové, tj. $\mathrm{E}(\boldsymbol{F}) = \boldsymbol{0}$, $\mathrm{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$.
2. Variační matice vektoru $\boldsymbol{F}$ je jednotková, tj. $\mathrm{var}(\boldsymbol{F}) = \boldsymbol{I}$, a varianční matice vektoru $\boldsymbol{\epsilon}$ je diagonální, tj. $\mathrm{var}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \ldots, \psi_p)$.
3. Vektory $\boldsymbol{F}$ a $\boldsymbol{\epsilon}$ jsou nekorelované, tedy jejich kovarianční matice $\mathrm{cov}(\boldsymbol{\epsilon}, \boldsymbol{F}) = \boldsymbol{0}$.

Za uvedených předpokladů lze snadno odvodit, že

$$\begin{aligned}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' = (\boldsymbol{LF} + \boldsymbol{\epsilon})(\boldsymbol{LF} + \boldsymbol{\epsilon})' = (\boldsymbol{LF} + \boldsymbol{\epsilon})((\boldsymbol{LF})' + \boldsymbol{\epsilon}') = \\ = \boldsymbol{LF}(\boldsymbol{LF})' + \boldsymbol{\epsilon}(\boldsymbol{LF})' + \boldsymbol{LF}\boldsymbol{\epsilon}' + \boldsymbol{\epsilon}\boldsymbol{\epsilon}',\end{aligned}$$

odkud pro varianční matici vektoru sledovaných proměnných $\boldsymbol{X}$ platí

$$\begin{aligned}\mathrm{var}(\boldsymbol{X}) = \mathrm{E}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' = \boldsymbol{L}\mathrm{E}(\boldsymbol{FF}')\boldsymbol{L}' + \mathrm{E}(\boldsymbol{\epsilon F}')\boldsymbol{L}' + \boldsymbol{L}\mathrm{E}(\boldsymbol{F\epsilon}') + \mathrm{E}(\boldsymbol{\epsilon\epsilon}') = \\ = \boldsymbol{LL}' + \boldsymbol{\Psi},\end{aligned}$$

neboli

$$\begin{aligned}\mathrm{D}(X_i) &= l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2 + \psi_i, \quad i = 1, 2, \ldots, p, \\ \mathrm{C}(X_i, X_k) &= l_{i1}l_{k1} + l_{i2}l_{k2} + \cdots + l_{im}l_{km}, \quad i, k = 1, 2, \ldots, p, \ i \neq k.\end{aligned} \tag{1}$$

Ortogonální faktorový model

Dále platí, že kovarianční matice vektorů $\boldsymbol{X}$ a $\boldsymbol{F}$ je rovna matici faktorových zátěží $\boldsymbol{L}$, tj.

$$\begin{aligned}\operatorname{cov}(\boldsymbol{X},\boldsymbol{F}) &= \mathrm{E}[(\boldsymbol{X}-\boldsymbol{\mu})\boldsymbol{F}'] = \mathrm{E}[(\boldsymbol{L}\boldsymbol{F}+\boldsymbol{\epsilon})\boldsymbol{F}'] = \mathrm{E}[\boldsymbol{L}\boldsymbol{F}\boldsymbol{F}'+\boldsymbol{\epsilon}\boldsymbol{F}'] = \\ &= \boldsymbol{L}\mathrm{E}(\boldsymbol{F}\boldsymbol{F})' + \mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{F}') = \boldsymbol{L},\end{aligned}$$

neboli

$$\mathrm{C}(X_i, F_j) = l_{ij}, \quad i = 1, 2, \ldots p,\ j = 1, 2, \ldots m.$$

V této souvislosti pak říkáme, že uvedená vyjádření varianční matice $\operatorname{var}(\boldsymbol{X})$ a kovarianční matice $\operatorname{cov}(\boldsymbol{X},\boldsymbol{F})$ popisují kovarianční strukturu ortogonálního faktorového modelu. Model $\boldsymbol{X}-\boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F}+\boldsymbol{\epsilon}$ je lineární vzhledem ke společným faktorům. Je-li vztah mezi $\boldsymbol{X}$ a faktory jiný než lineární, popis kovarianční struktury $\boldsymbol{L}\boldsymbol{L}'+\boldsymbol{\Psi}$ uvedený v (1) nebude odpovídající.

Ortogonální faktorový model

Ta část rozptylu proměnné X_i, která je vysvětlená pomocí m společných faktorů, se nazývá **komunalita** („communality“), příspěvek specifického faktoru k této variabilitě je označován jako **specifický rozptyl**, tedy

$$\begin{array}{rcccl} \underbrace{\sigma_{ii}} & = & \underbrace{l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2} & + & \underbrace{\psi_i} \\ \mathrm{D}(X_i) = & & \text{komunalita} & + & \text{specifický rozptyl} \end{array} \tag{2}$$

Pro i-tou komunalitu h_i^2 lze psát

$$h_i^2 = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2,$$

a rozptyl $\mathrm{D}(X_i)$ vyjádřit ve tvaru

$$\sigma_{ii} = h_i^2 + \psi_i, \quad i = 1, 2, \ldots, p.$$

Je-li dán ortogonální faktorový model, je třeba pomocí opakovaných nezávislých pozorování odhadnout matici faktorových zátěží $\boldsymbol{L}$ a dále najít odhady hodnot společných faktorů $F_1 \ldots F_p$ pro každou statistickou jednotku.

Ortogonální faktorový model

Matice faktorových zátěží $\boldsymbol{L}$ a vektor společných faktorů $\boldsymbol{F}$ nejsou v daném ortogonálním modelu určeny jednoznačně. Pro libovolnou ortogonální transformací vektoru společných faktorů $\boldsymbol{F}$ získáme nový vektor společných faktorů $\boldsymbol{F}^* = \boldsymbol{T}'\boldsymbol{F}$ (zde $\boldsymbol{T}$ je ortogonální matice typu $m \times m$, tj. $\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}$) a novou matici faktorových zátěží $\boldsymbol{L}^* = \boldsymbol{L}\boldsymbol{T}$, které opět splňují předpoklady modelu, neboť

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{F} + \boldsymbol{\epsilon} = \boldsymbol{L}^*\boldsymbol{F}^* + \boldsymbol{\epsilon},$$

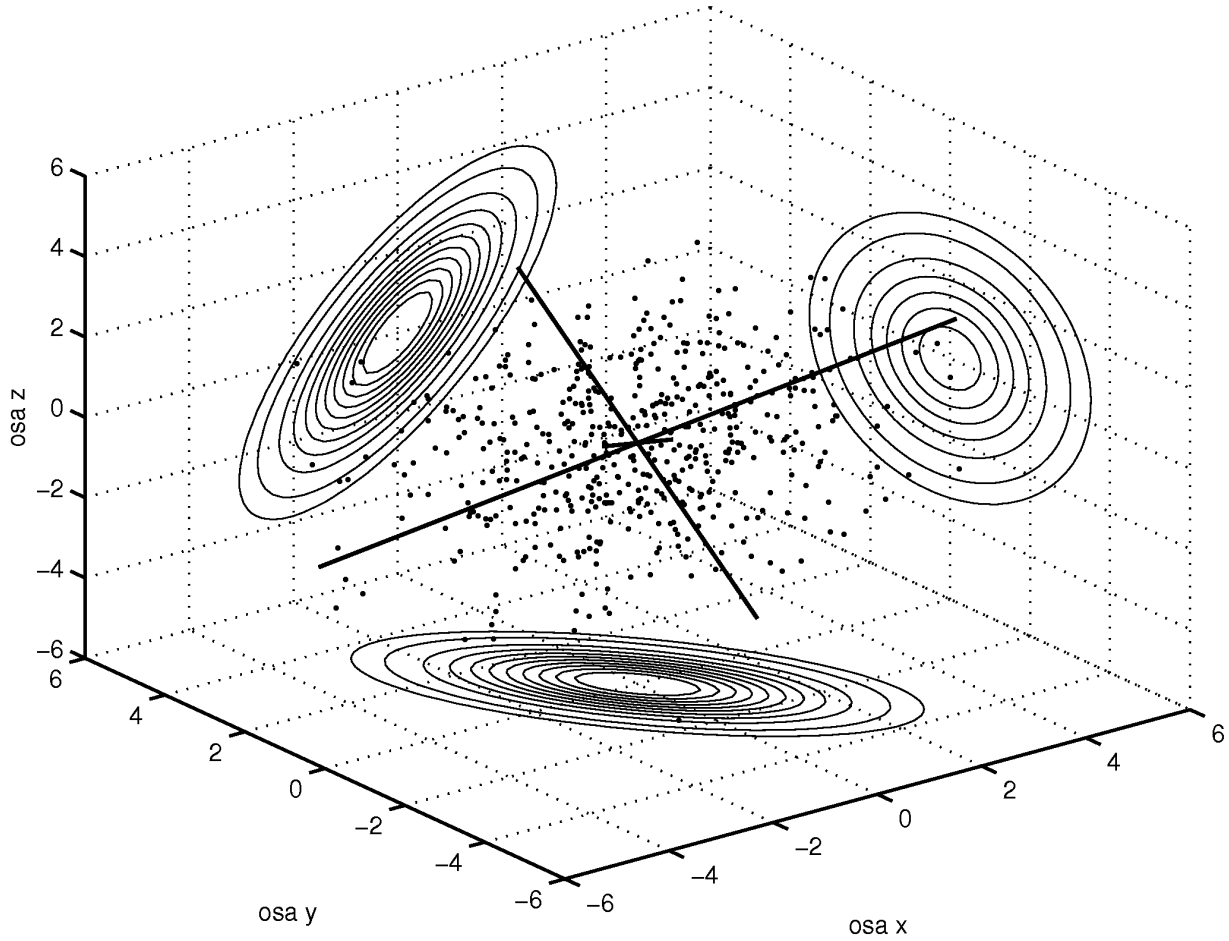
$$\text{var}(\boldsymbol{X}) = \boldsymbol{L}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{L}' + \boldsymbol{\Psi} = \boldsymbol{L}^*\boldsymbol{L}^{*\prime} + \boldsymbol{\Psi}.$$

Jak bude dále naznačeno, lze uvedenou nejednoznačnost modelu s výhodou použít při interpretaci výsledků.

Geometrická interpretace

Bylo nasimulováno 500 pozorování trojrozměrného náhodného vektoru $\boldsymbol{X}$ z normálního rozdělení s nulovou střední hodnotou a nediagonální varianční maticí $\boldsymbol{\Sigma}$. Na obr. 1 jsou jednotlivá pozorování znázorněna body. Z obrázku je vidět, jaká je variabilita dat.
V každém směru se souřadnice jednotlivých pozorování nacházejí přibližně v rozmezí od -4 do 4. Úsečky v 1 představují směry vlastních vektorů $\boldsymbol{e}_1$, $\boldsymbol{e}_2$ a $\boldsymbol{e}_3$ varianční matice $\boldsymbol{\Sigma}$. Jejich délky jsou úměrné vlastním číslům λ_1, λ_2 a λ_3 varianční matice $\boldsymbol{\Sigma}$, $\lambda_1 > \lambda_2 > \lambda_3$. Elipsy v rovinách os znázorňují vrstevnice hustot příslušných dvourozměrných marginálních rozdělení.

Geometrická interpretace

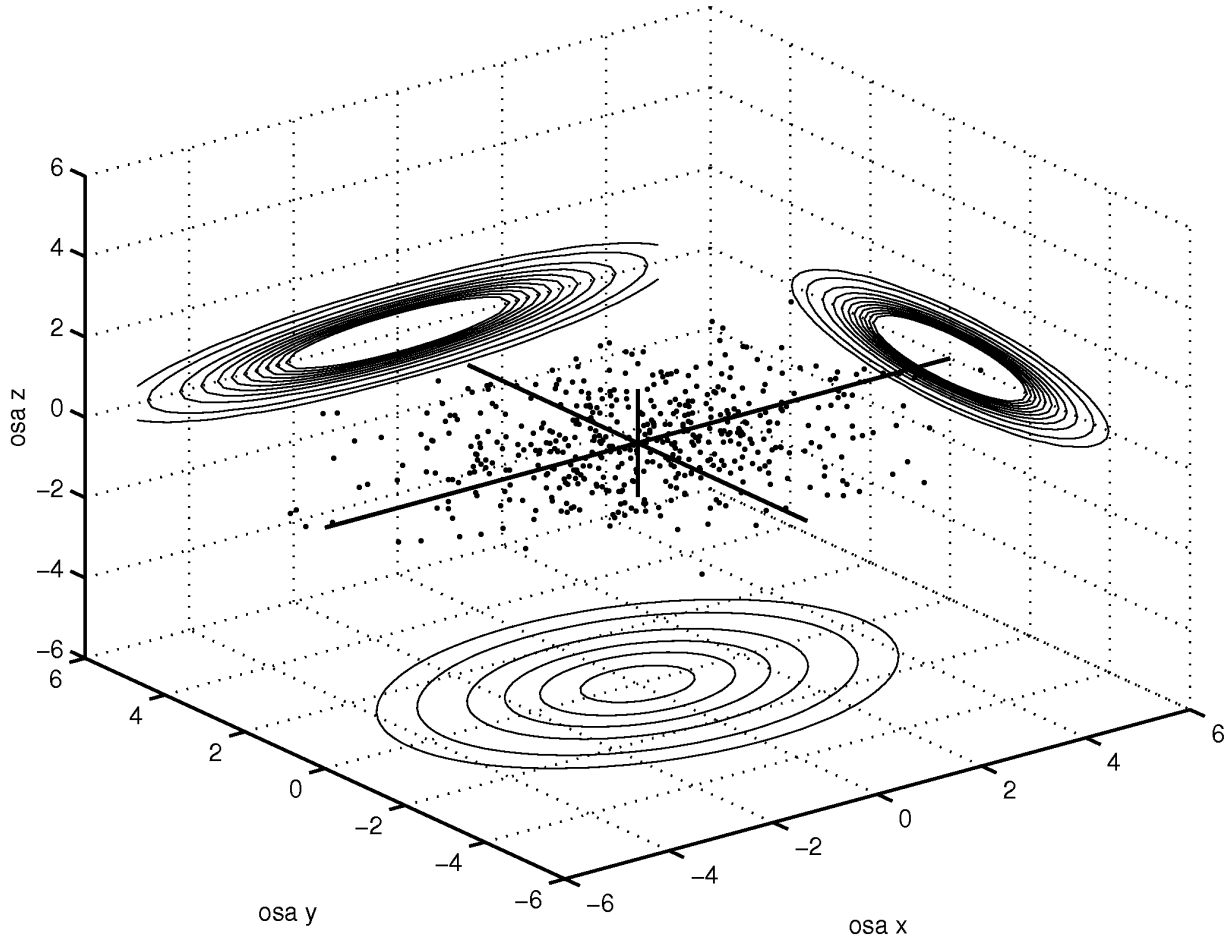


Obrázek: Simulovaná data před transformací

Geometrická interpretace

Když zvolíme novou souřadnou soustavu tak, že osa x bude ve směru vlastního vektoru $\boldsymbol{e}_1$, osa y ve směru vlastního vektoru $\boldsymbol{e}_2$ a osa z ve směru vlastního vektoru $\boldsymbol{e}_3$ a všech 500 simulovaných bodů znázorníme v této nové souřadné soustavě, dostaneme situaci znázorněnou na obr. 2. Z tohoto obrázku je dobře patrné, že největší variabilita znázorněných bodů je ve směru nové osy x. Při tom není možné najít jiný směr, v němž by byla větší variabilita dat, než ve směru osy x, tedy ve směru určeném vlastním vektorem $\boldsymbol{e}_1$. Vektor $\boldsymbol{e}_2$, který udává směr nové osy y, zároveň udává směr, kolmý na osu x, v němž je opět největší variabilita ze všech těchto kolmých směrů. Konečně se ukazuje, že ve směru osy z, tedy ve směru vlastního vektoru $\boldsymbol{v}_3$, je variabilita ze všech těchto tří směrů nejmenší. Variabilita znázorněných bodů ve směru vlastního vektoru $\boldsymbol{v}_i$ je úměrná vlastnímu číslu λ_i a pro celkovou variabilitu vektoru $\boldsymbol{X}$ platí
$\sigma_T^2(\boldsymbol{X}) = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 = \lambda_1 + \lambda_2 + \lambda_3$, zde $\sigma_i^2 = D(X_i)$.

Geometrická interpretace



Obrázek: Simulovaná data po transformaci do báze tvořené vlastními vektory varianční matice

Odhadování parametrů modelu

Předpokládejme, že je předem znám pevný počet faktorů m. Při odhadování parametrů modelu vycházíme z náhodného výběru $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ z rozdělení náhodného vektoru $\boldsymbol{X}$. Tento výběr je často před zpracováním standardizován, takže předpokládáme, že výsledný výběr je pak z rozdělení standardizované náhodné veličiny $\boldsymbol{Y} = (\text{Diag}\,(\boldsymbol{\Sigma}))^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$, pro kterou lze faktorový model přepsat následovně

$$\boldsymbol{Y} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\epsilon},$$

$$\text{var}(\boldsymbol{Y}) = \text{cor}(\boldsymbol{X}) = \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi}.$$

Z tohoto náhodného výběru nejdříve stanovíme výběrový průměr $\overline{\boldsymbol{X}}$ jako odhad vektoru střední hodnoty $\boldsymbol{\mu}$ dále odhadneme korelační matici $\text{cor}(\boldsymbol{X})$ výběrovou korelační maticí $\boldsymbol{R}$. Pokud mimodiagonální prvky matice $\boldsymbol{R}$ nejsou malé, tj. pokud jsou složky náhodného vektoru $\boldsymbol{X}$ silně korelované, má smysl hledat společné faktory. V opačném případě by totiž hlavní roli hrály pouze specifické faktory.

Z kovarianční struktury ortogonálního faktorového modelu vyplývá, že při faktorové analýze hledáme rozklad varianční matice $\boldsymbol{\Sigma}$ náhodného vektoru $\boldsymbol{X}$ na symetrickou, pozitivně definitní matici $\boldsymbol{L}\boldsymbol{L}'$ a diagonální matici $\boldsymbol{\Psi}$.

Metoda založená na hlavních komponentách

Mějme varianční matici $\boldsymbol{\Sigma}$ a jí odpovídající dvojice vlastních čísel a vlastních vektorů $(\lambda_i, \boldsymbol{e}_i)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, potom ji lze vyjádřit jako

$$\boldsymbol{\Sigma} = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1' + \lambda_2 \boldsymbol{e}_2 \boldsymbol{e}_2' + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p' =$$

$$= \left(\sqrt{\lambda_1}\boldsymbol{e}_1 \,\middle|\, \sqrt{\lambda_2}\boldsymbol{e}_2 \,\middle|\, \cdots \,\middle|\, \sqrt{\lambda_p}\boldsymbol{e}_p \right) \begin{pmatrix} \sqrt{\lambda_1}\boldsymbol{e}_1' \\ \hline \sqrt{\lambda_2}\boldsymbol{e}_2' \\ \hline \vdots \\ \hline \sqrt{\lambda_p}\boldsymbol{e}_p' \end{pmatrix}$$

Uvedený rozklad se nazývá **spektrální dekompozice**.

Metoda založená na hlavních komponentách

Tímto lze popsat faktorový model mající tolik faktorů kolik proměnných ($m = p$) a specifické rozptyly $\psi_i = 0$ pro všechna i, lze tedy psát

$$\underset{(p\times p)}{\boldsymbol{\Sigma}} = \underset{(p\times p)(p\times p)}{\boldsymbol{L}\ \boldsymbol{L}'} + \underset{(p\times p)}{\mathbf{0}} = \boldsymbol{L}\boldsymbol{L}' \tag{3}$$

Faktorové zátěže („factor loadings“) j-tého faktoru jsou až na $\sqrt{\lambda_j}$ rovny j-té hlavní komponentě.
Reprezentace matice $\boldsymbol{\Sigma}$ popsaná v (3) je sice přesná, nicméně obsahuje stejný počet faktorů kolik je proměnných. Cílem bude najít model, který vystihuje kovarianční strukturu pomocí několika málo faktorů. Jednou z možností je zanedbání posledních $p - m$ členů ve spektrálním rozkladu, pokud odpovídající vlastní čísla jsou malá.

$$\boldsymbol{\Sigma} \doteq \left(\sqrt{\lambda_1}\boldsymbol{e}_1 \,\middle|\, \sqrt{\lambda_2}\boldsymbol{e}_2 \,\middle|\, \cdots \,\middle|\, \sqrt{\lambda_m}\boldsymbol{e}_m\right) \begin{pmatrix} \sqrt{\lambda_1}\boldsymbol{e}_1' \\ \hline \sqrt{\lambda_2}\boldsymbol{e}_2' \\ \hline \vdots \\ \hline \sqrt{\lambda_m}\boldsymbol{e}_m' \end{pmatrix} = \underset{(p\times m)(m\times p)}{\boldsymbol{L}\ \boldsymbol{L}'} \tag{4}$$

Metoda založená na hlavních komponentách

Přibližné vyjádření popsané v (4) předpokládá, že specifické faktory ϵ lze zanedbat. Pokud tyto specifické faktory do modelu zahrneme, lze psát

$$\boldsymbol{\Sigma} \doteq \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi} = \left(\sqrt{\lambda_1}\boldsymbol{e}_1 \,\middle|\, \sqrt{\lambda_2}\boldsymbol{e}_2 \,\middle|\, \cdots \,\middle|\, \sqrt{\lambda_m}\boldsymbol{e}_m \right) \begin{pmatrix} \sqrt{\lambda_1}\boldsymbol{e}_1' \\ \hline \sqrt{\lambda_2}\boldsymbol{e}_2' \\ \hline \vdots \\ \hline \sqrt{\lambda_m}\boldsymbol{e}_m' \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{pmatrix} \tag{5}$$

kde pro rozptyly ψ_i specifických faktorů platí $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2$ pro $i = 1, 2, \ldots, p$.

Metoda založená na hlavních komponentách

Předtím, než uvedený postup aplikujeme na data (náhodný výběr), obvykle se provádí centrování odečtením výběrového průměru, případně standardizace proměnných.
Výběrová kovarianční matice $\boldsymbol{S}$ takto standardizovaných proměnných je rovna výběrové korelační matici původních dat $\boldsymbol{R}$.
Mějme výběrovou kovarianční matici $\boldsymbol{S}$, $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), (\hat{\lambda}_2, \hat{\boldsymbol{e}}_2), \ldots, (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$, jsou dvojice vlastních čísel a vektorů této matice splňující $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. Předpokládejme, že $m < p$ je počet společných faktorů. Odhad matice faktorových zátěží je dán výrazem

$$\hat{\boldsymbol{L}} = \left(\sqrt{\hat{\lambda}_1}\hat{\boldsymbol{e}}_1 \,\middle|\, \sqrt{\hat{\lambda}_2}\hat{\boldsymbol{e}}_2 \,\middle|\, \cdots \,\middle|\, \sqrt{\hat{\lambda}_m}\hat{\boldsymbol{e}}_m \right).$$

Odhady specifických rozptylů jsou diagonální prvky matice $\boldsymbol{S} - \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}'$

$$\hat{\boldsymbol{\Psi}} = \begin{pmatrix} \hat{\psi}_1 & 0 & \cdots & 0 \\ 0 & \hat{\psi}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\psi}_p \end{pmatrix}, \quad \hat{\psi}_i = s_{ii} - \sum_{j=1}^{m} \hat{l}_{ij}.$$

Komunality lze odhadnou ze vztahu

$$\hat{h}_i^2 = \hat{l}_{i1}^2 + \hat{l}_{i2}^2 + \cdots + \hat{l}_{im}^2.$$

Odhady metodou maximální věrohodnosti

Jestliže předpokládáme, že společné faktory $\boldsymbol{F}$ a specifické faktory $\boldsymbol{\epsilon}$ mají normální rozdělení, lze pro jejich odhady použít metodu maximální věrohodnosti. Věrohodnostní funkce má tvar

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\boldsymbol{x}_j-\overline{\boldsymbol{x}})(\boldsymbol{x}_j-\overline{\boldsymbol{x}})'+n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})(\overline{\boldsymbol{x}}-\boldsymbol{\mu})'\right)\right]} = \\ &= (2\pi)^{-\frac{(n-1)p}{2}} |\boldsymbol{\Sigma}|^{-\frac{n-1}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\boldsymbol{x}_j-\overline{\boldsymbol{x}})(\boldsymbol{x}_j-\overline{\boldsymbol{x}})'\right)\right]} \times \\ &\quad \times (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \mathrm{e}^{-\frac{n}{2}(\overline{\boldsymbol{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\boldsymbol{x}}-\boldsymbol{\mu})}. \end{aligned} \tag{6}$$

Věrohodnostní funkce (6) závisí na $\boldsymbol{L}$ a $\boldsymbol{\Psi}$ prostřednictvím $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi}$. Tento model není zcela jednoznačně definovaný vzhledem k možným volbám matice $\boldsymbol{L}$. K modelu se obvykle přidá podmínka

$$\boldsymbol{L}'\boldsymbol{\Psi}^{-1}\boldsymbol{L} = \boldsymbol{\Delta}$$

je diagonální matice. Maximálně věrohodné odhady $\hat{\boldsymbol{L}}$ a $\hat{\boldsymbol{\Psi}}$ se získají numerickou maximalizací (6). Odhady komunalit jsou tvaru

$$\hat{h}_i^2 = \hat{l}_{i1}^2 + \hat{l}_{i2}^2 + \cdots + \hat{l}_{im}^2.$$

Odhady metodou maximální věrohodnosti

Pokud jsou proměnné standardizované, podobně jako tomu bylo u metody hlavních komponent, máme $\boldsymbol{Z} = \boldsymbol{V}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$, potom kovarianční matice je rovna matici korelační a platí

$$\boldsymbol{\rho} = \boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{V}^{-1/2} = (\boldsymbol{V}^{-1/2}\boldsymbol{L})(\boldsymbol{V}^{-1/2}\boldsymbol{L})' + \boldsymbol{V}^{-1/2}\boldsymbol{\Psi}\boldsymbol{V}^{-1/2}.$$

Matice faktorových zátěží odpovídající $\boldsymbol{\rho}$ má tvar $\boldsymbol{L}_{\boldsymbol{z}} = \boldsymbol{V}^{-1/2}\boldsymbol{L}$ a matice specifických rozptylů je rovna $\boldsymbol{\Psi}_{\boldsymbol{z}} = \boldsymbol{V}^{-1/2}\boldsymbol{\Psi}\boldsymbol{V}^{-1/2}$. Maximálně věrohodný odhad $\boldsymbol{\rho}$ je potom dán výrazem

$$\begin{aligned}\hat{\boldsymbol{\rho}} &= (\hat{\boldsymbol{V}}^{-1/2}\hat{\boldsymbol{L}})(\hat{\boldsymbol{V}}^{-1/2}\hat{\boldsymbol{L}})' + \hat{\boldsymbol{V}}^{-1/2}\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{V}}^{-1/2} = \\ &= \hat{\boldsymbol{L}}_{\boldsymbol{z}}\hat{\boldsymbol{L}}_{\boldsymbol{z}}' + \hat{\boldsymbol{\Psi}}_{\boldsymbol{z}}\end{aligned}$$

Poznámka: Počítačové programy obvykle provádějí standardizaci proměnných, faktorizovaná je tedy výběrová korelační matice $\boldsymbol{R}$. Díky tomu získáme maximálně věrohodné odhady $\hat{\boldsymbol{L}}_{\boldsymbol{z}}$ a $\hat{\boldsymbol{\Psi}}_{\boldsymbol{z}}$. Maximálně věrohodné odhady matice faktorových zátěží a specifických rozptylů odpovídající matici $\frac{n-1}{n}\boldsymbol{S}$ jsou $\hat{\boldsymbol{L}} = \hat{\boldsymbol{V}}^{1/2}\hat{\boldsymbol{L}}_{\boldsymbol{z}}$ a $\hat{\boldsymbol{\Psi}} = \hat{\boldsymbol{V}}^{1/2}\hat{\boldsymbol{\Psi}}_{\boldsymbol{z}}\hat{\boldsymbol{V}}^{1/2}$, neboli

$$\hat{l}_{ij} = \hat{l}_{\boldsymbol{z},ij}\sqrt{\hat{\sigma}_{ii}} \quad \text{a} \quad \hat{\psi}_i = \hat{\psi}_{\boldsymbol{z},i}\hat{\sigma}_{ii},$$

kde $\hat{\sigma}_{ii}$ v tomto případě neznačí výběrový rozptyl, ale rozptyl momentový.

Stanovení počtu faktorů

Obecně počet faktorů není známý. Uvažujme **reziduální matici**

$$\boldsymbol{S} - (\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}' + \hat{\boldsymbol{\Psi}})$$

získanou aproximací výběrové kovarianční matice $\boldsymbol{S}$. Diagonální prvky jsou nulové, a pokud jsou i ostatní prvky matice malé, lze usoudit, že zvolený počet faktorů m je dostatečný. Pro jeho stanovení se při metodě hlavních komponent užívá podíl variability vysvětlené pomocí zvoleného počtu faktorů a celkové variability. Podíl j-tého faktoru na celkové výběrové variabilitě je

$$\frac{\hat{\lambda}_j}{s_{11} + s_{22} + \cdots + s_{pp}} = \frac{\hat{\lambda}_j}{\hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p},$$

kde $\hat{\lambda}_i$, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ jsou vlastní čísla matice $\boldsymbol{S}$. Máme-li m faktorů, pak jejich relativní příspěvek k celkové variabilitě je

$$\frac{\sum_{j=1}^{m} \hat{\lambda}_j}{\sum_{i=1}^{p} s_{ii}} = \frac{\sum_{j=1}^{m} \hat{\lambda}_j}{\sum_{i=1}^{p} \hat{\lambda}_i}.$$

Snahou je nalézt takový počet faktorů m, při kterém je tento podíl dostatečně blízký 1.

Stanovení počtu faktorů

U metody maximální věrohodnosti je možné podíl j-tého faktoru na celkové variabilitě vyjádřit zlomkem

$$\frac{\hat{l}_{1j}^2 + \hat{l}_{2j}^2 + \cdots + \hat{l}_{pj}^2}{s_{11} + s_{22} + \cdots + s_{pp}}$$

Za předpokladu normality lze odvodit test adekvátnosti modelu s m společnými faktory. Jedná se o test věrohodnostním poměrem a je založen a testovací statistice

$$-2 \ln \left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\boldsymbol{S}_n|} \right)^{-n/2} + n \left[\mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{S}_n) - p \right]$$

kde $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}' + \hat{\boldsymbol{\Psi}}$, $\hat{\boldsymbol{L}}, \hat{\boldsymbol{\Psi}}$ jsou odhady matic $\boldsymbol{L}$ a $\boldsymbol{\Psi}$ získané metodou maximální věrohodnosti a $\boldsymbol{S}_n = \frac{n-1}{n}\boldsymbol{S}$. Tato testovací statistika má asymptoticky rozdělení $\chi^2(\nu)$ se stupni volnosti $\nu = \frac{1}{2}[(p-m)^2 - (p+m)]$. Lze dokázal, že výraz $\mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_n) - p = 0$, za předpokladu, že $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}' + \hat{\boldsymbol{\Psi}}$ je maximálně věrohodný odhad $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi}$. Dostáváme tedy testovací statistiku

$$n \ln \left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\boldsymbol{S}_n|} \right). \tag{7}$$

Stanovení počtu faktorů

Je-li realizace této statistiky větší než příslušný kvantil χ^2 rozdělení, pak zamítáme hypotézu o dostatečném počtu faktorů. Aproximaci rozdělení testovací statistiky χ^2 rozdělením lze zpřesnit nahrazením n v testovací statistice (7) hodnotou $n - 1 - (2p + 4m + 5)/6$. Vzhledem k tomu, že stupně volnosti jsou kladné, pro použití testu musí platit nerovnost

$$m < \frac{1}{2}(2p + 1 - \sqrt{8p + 1}).$$

Rotace faktorů

Lepší interpretaci faktorů je možno získat po provedení ortogonální transformace společných faktorů. Takovou ortogonální transformací (rotací) neporušíme předpoklady modelu.
Je-li $\hat{\boldsymbol{L}}$ odhad $p \times m$ matice faktorových zátěží, potom

$$\hat{\boldsymbol{L}}^* = \hat{\boldsymbol{L}}\boldsymbol{T}, \quad \text{kde } \boldsymbol{T}\boldsymbol{T}' = \boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I},$$

je $p \times m$ matice „rotovaných“ faktorových zátěží. Navíc odhadnuté kovarianční (nebo korelační) matice zůstávají nezměněny, neboť

$$\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}' + \hat{\boldsymbol{\Psi}} = \hat{\boldsymbol{L}}\boldsymbol{T}\boldsymbol{T}'\hat{\boldsymbol{L}}' + \hat{\boldsymbol{\Psi}} = \hat{\boldsymbol{L}}^*\hat{\boldsymbol{L}}^{*\prime} + \hat{\boldsymbol{\Psi}}. \tag{8}$$

Z rovnice (8) lze usoudit, že reziduální matice $\boldsymbol{S}_n - \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}' - \hat{\boldsymbol{\Psi}} = \boldsymbol{S}_n - \hat{\boldsymbol{L}}^*\hat{\boldsymbol{L}}^{*\prime} - \hat{\boldsymbol{\Psi}}$ zůstávají nezměněny. Totéž platí pro specifické rozptyly $\hat{\psi}_i$ a komunality $\hat{h}_i^2$.

Rotace faktorů

Původní matice faktorových zátěží nemusí být snadno interpretovatelná. V praxi je obvyklé provádět takovou rotaci, která umožní snadnější interpretaci. V ideálním případě docílit toho, aby každá proměnná byla silně zastoupena v jednom faktoru a v ostatních faktorech se již téměř nevyskytovala.
Jedním z nejčastěji používaných kritérií optimální ortogonální transformace je **varimax** kritérium. Definujme $\tilde{l}_{ij}^{*} = \hat{l}_{ij}^{*} / \hat{h}_i$ jako rotované koeficienty škálované pomocí druhých odmocnin komunalit. Cílem je najít takovou ortogonální transformaci $\boldsymbol{T}$, která maximalizuje výraz

$$V = \frac{1}{p} \sum_{j=1}^{m} \left[\sum_{i=1}^{p} \tilde{l}_{ij}^{*4} - \frac{1}{p} \left(\sum_{i=1}^{p} \tilde{l}_{ij}^{*2} \right)^2 \right].$$

Po nalezení matice $\boldsymbol{T}$ přenásobíme získané zátěže $\tilde{l}_{ij}^{*}$ konstantami $\hat{h}_i$ a získáme potřebné zátěže $\hat{l}_{ij}^{*}$. Varimax kritérium tedy minimalizuje počet proměnných vysvětlovaných jedním faktorem.

Odhad společných faktorů

Při faktorové analýze se pozornost obvykle zaměřuje na parametry faktorového modelu, nicméně odhady hodnot společných faktorů, které se nazývají **faktorové skóry**, mohou být také užitečné. Tyto hodnoty jsou často používány pro diagnostické účely, případně jako vstupy pro následující analýzy.

Faktorové skóry nejsou odhady neznámých parametrů v obvyklém smyslu, jsou to odhady hodnot nepozorovaných náhodných faktorů $\boldsymbol{F}_j$, $j = 1, 2, \ldots, n$. Faktorové skóry $\hat{\boldsymbol{f}}_j$ tedy jsou odhady $\boldsymbol{f}_j$ získané pro $\boldsymbol{F}_j$.

Považujeme-li nyní získané odhady matic $\boldsymbol{L}$ a $\boldsymbol{\Psi}$ za skutečné pevné hodnoty, můžeme odhadnout společné faktory buď váženou metodou nejmenších čtverců, nebo metodou regresní.

Odhad společných faktorů – vážená metoda nejmenších čtverců

Předpokládejme, že vektor středních hodnot $\boldsymbol{\mu}$, matice faktorových zátěží $\boldsymbol{L}$ a matice specifických rozptylů jsou ve faktorovém modelu známé, tedy

$$\underset{(p\times 1)}{\boldsymbol{X}} - \underset{(p\times 1)}{\boldsymbol{\mu}} = \underset{(p\times m)(m\times 1)}{\boldsymbol{L}\ \boldsymbol{F}} + \underset{(p\times 1)}{\boldsymbol{\epsilon}}$$

Považujme specifické faktory za chybové složky $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_p)'$. Protože rozptyly $\mathrm{D}(\epsilon_i) = \psi_i$, $i = 1, 2, \ldots, p$, nemusí být stejné, použijeme pro odhad společných faktorů váženou metodu nejmenších čtverců. Součet čtverců vážený převrácenými hodnotami rozptylů je

$$\sum_{i=1}^{p} \frac{\epsilon_i^2}{\psi_i} = \boldsymbol{\epsilon}'\boldsymbol{\Psi}^{-1}\boldsymbol{\epsilon} = (\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{L}\boldsymbol{f})'\boldsymbol{\Psi}(\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{L}\boldsymbol{f}). \tag{9}$$

Minimalizací výrazu (9) lze získat odhad

$$\hat{\boldsymbol{f}} = \left(\hat{\boldsymbol{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{L}}\right)^{-1}\hat{\boldsymbol{L}}'\hat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{X}_j - \boldsymbol{\mu}). \tag{10}$$

Odhad společných faktorů – vážená metoda nejmenších čtverců

Dosazením odhadů do (10) získáme faktorové skóry ve tvaru

$$\hat{\boldsymbol{f}}_j = \left(\hat{\boldsymbol{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{L}}\right)^{-1}\hat{\boldsymbol{L}}'\hat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{X}_j - \overline{\boldsymbol{X}}),\ j = 1, 2\ldots, n.$$

V případě, že byla faktorizována korelační matice, jsou faktorové skóry dány vztahem

$$\hat{\boldsymbol{f}}_j = \left(\hat{\boldsymbol{L}}_z'\hat{\boldsymbol{\Psi}}_z^{-1}\hat{\boldsymbol{L}}_z\right)^{-1}\hat{\boldsymbol{L}}_z'\hat{\boldsymbol{\Psi}}_z^{-1}\boldsymbol{z}_j,\ j = 1, 2\ldots, n,$$

kde $\boldsymbol{z}_j$ jsou standardizované proměnné a $\hat{\boldsymbol{\rho}} = \hat{\boldsymbol{L}}_z\hat{\boldsymbol{L}}_z' + \hat{\boldsymbol{\Psi}}_z$. Faktorové skóry mají nulovou střední hodnotu a nulovou výběrovou kovarianci.

Odhad společných faktorů – regresní metoda

Regresní metoda vede k odhadu

$$\hat{\boldsymbol{f}}_j = \hat{\boldsymbol{L}}' \left(\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}' + \hat{\boldsymbol{\Psi}} \right)^{-1} (\boldsymbol{X}_j - \overline{\boldsymbol{X}}), \quad j = 1, 2, \ldots, n.$$

Z důvodu snížení efektu možného nesprávného určení počtu faktorů v modelu, je někdy místo matice $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}' + \hat{\boldsymbol{\Psi}}$ použita matice výběrová kovarianční matice $\boldsymbol{S}$. Odhady potom mají tvar

$$\hat{\boldsymbol{f}}_j = \hat{\boldsymbol{L}}' \boldsymbol{S}^{-1} (\boldsymbol{X}_j - \overline{\boldsymbol{X}}), \quad j = 1, 2, \ldots, n.$$

Byla-li faktorizována korelační matice, potom

$$\hat{\boldsymbol{f}}_j = \hat{\boldsymbol{L}}' \boldsymbol{R}^{-1} \boldsymbol{z}, \quad j = 1, 2, \ldots, n.$$