Analýza hlavních komponent

Ekonometrie

Jiří Neubauer

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

Metoda hlavních komponent

Metoda hlavních komponent (PCA – „Principal component analysis“) je metodou, pomocí níž lze vysvětlit variančně kovarianční strukturu množiny proměnných pomocí několika lineárních kombinací těchto proměnných. K základním úkolům metody patří redukce dat (snížení dimenze) a interpretace. Máme-li p proměnných, pomocí nichž lze vysvětlit celkovou variabilitu, v řadě případů je možné většinu této variability popsat pomocí mnohem menšího počtu k hlavních komponent. Těchto k hlavních komponent potom může nahradit původních p proměnných, tedy původní datový soubor obsahující n měření p proměnných je redukován na n měření k hlavních komponent.

Metoda hlavních komponent

- Algebraicky jsou hlavní komponenty jisté lineární kombinace původních p náhodných veličin $X_1, X_2, \ldots, X_p$.
- Z geometrického hlediska reprezentují tyto lineární kombinace výběr nového systému souřadnic získaného rotací původního souřadnicového systému s osami $X_1, X_2, \ldots, X_p$. Nové osy reprezentují směry s největší variabilitou a nabízí jednodušší a více parsimonní popis kovarianční struktury.

Příklad

The weekly rates of return for five stocks (JPMorgan, Citibank, WellsFargo, RoyalDutchShell, ExxonMobil) listed on the New Yourk Stock Exchange were determined for the period January 2004 through December 2005. The weekly rates od return are defined

$$\frac{\text{current week closing price} - \text{previous weeek closing price}}{\text{previous weeek closing price}},$$

adjusted for stick splits and dividents. The observations in 103 succesive weeks appear to be independently distibuted, but the rates of return across stock are correlated, because as on expects, stocks tend to move together in response to general economic conditions.

Johnson, R. and Wichern, D. *Applied Multivariate Statistical Analysis*. 6th ed. Pearson 2014.

Metoda hlavních komponent

Hlavní komponenty závisí výhradně na varianční matici $\boldsymbol{\Sigma}$ (nebo na korelační matici $\boldsymbol{\rho}$) proměnných $X_1, X_2, \ldots, X_p$. Nechť náhodný vektor $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$ má varianční matici $\boldsymbol{\Sigma}$ s vlastními čísly $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Uvažujme lineární kombinace

$$
\begin{aligned}
Y_1 &= \boldsymbol{a}_1'\boldsymbol{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 &= \boldsymbol{a}_2'\boldsymbol{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
&\vdots \\
Y_p &= \boldsymbol{a}_p'\boldsymbol{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p.
\end{aligned}
$$

Pro proměnné $Y_1, Y_2, \ldots, Y_p$ dostáváme

$$
\mathrm{D}(Y_i) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_i, \qquad i = 1, 2, \ldots, p \tag{1}
$$

$$
\mathrm{C}(Y_i, Y_k) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_k, \quad i, k = 1, 2, \ldots, p \tag{2}
$$

Hlavní komponenty jsou takové nekorelované kombinace $Y_1, Y_2, \ldots, Y_p$, jejichž rozptyly (4) jsou co možná největší.

Metoda hlavních komponent

První hlavní komponenta je lineární kombinace s největším rozptylem, tedy maximalizující výraz $\mathrm{D}(Y_1) = \boldsymbol{a}_1'\boldsymbol{\Sigma}\boldsymbol{a}_1$. Je zřejmé, že vynásobením $\boldsymbol{a}_1$ nějakou konstantou lze hodnotu rozptylu měnit. Je třeba přidat podmínku jednotkové délky vektoru $\boldsymbol{a}_1$, tedy $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$.

Druhá hlavní komponenta je lineární kombinace $\boldsymbol{a}_2'\boldsymbol{X}$ jež maximalizuje výraz $\boldsymbol{a}_2'\boldsymbol{\Sigma}\boldsymbol{a}_2$ při platnosti $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$ a $\mathrm{C}(\boldsymbol{a}_1'\boldsymbol{X}, \boldsymbol{a}_2'\boldsymbol{X}) = 0$. Podobně se postupuje dále, což znamená, že i-tá hlavní komponenta je lineární kombinace $\boldsymbol{a}_i'\boldsymbol{X}$ jež maximalizuje výraz $\boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_i$ při platnosti $\boldsymbol{a}_i'\boldsymbol{a}_i = 1$ a $\mathrm{C}(\boldsymbol{a}_i'\boldsymbol{X}, \boldsymbol{a}_k'\boldsymbol{X}) = 0$ pro $k < i$.

Určení hlavních komponent souvisí s vlastními čísly a vektory matice $\boldsymbol{\Sigma}$.

Pozn. Nechť $\boldsymbol{A} = (a_{ij})$ je matice řádu n. Číslo λ se nazývá vlastní nebo charakteristické číslo matice $\boldsymbol{A}$, jestliže existuje nenulový vektor $\boldsymbol{u}$ tak, že $\boldsymbol{A}\boldsymbol{u} = \lambda\boldsymbol{u}$. Vektor $\boldsymbol{u}$ se nazývá vlastní nebo charakteristický vektor příslušný k λ.
Rovnici $\boldsymbol{A}\boldsymbol{u} = \lambda\boldsymbol{u}$ lze přepsat do tvaru $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{u} = \boldsymbol{0}$. Tato soustava má netriviální řešení, právě když $|\boldsymbol{A} - \lambda\boldsymbol{I}| = 0$.

Příklad: Určete vlastní čísla a vlastní vektory matice

$$\boldsymbol{A} = \begin{pmatrix} 4 & -2 \\ 1 & 1 \end{pmatrix}$$

Metoda hlavních komponent

Mějme náhodný vektor $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$ s varianční maticí $\boldsymbol{\Sigma}$ a dvojicemi vlastních čísel a vlastních vektorů $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), \ldots, (\lambda_p, \boldsymbol{e}_p)$ splňující $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Potom i-tá hlavní komponenta je dána vztahem

$$Y_i = \boldsymbol{e}_i \boldsymbol{X} = e_{i1} X_1 + e_{i2} X_2 + \cdots + e_{ip} X_p, \quad i = 1, 2, \ldots, p, \tag{3}$$

přičemž

$$\mathrm{D}(Y_i) = \boldsymbol{e}_i' \boldsymbol{\Sigma} \boldsymbol{e}_i = \lambda_i, \quad i = 1, 2, \ldots, p, \tag{4}$$

$$\mathrm{C}(Y_i, Y_k) = \boldsymbol{e}_i' \boldsymbol{\Sigma} \boldsymbol{e}_k = 0, \quad i \neq k. \tag{5}$$

Označíme-li prvky matice $\boldsymbol{\Sigma}$ jako σ_{ij}, $i, j = 1, 2 \ldots, p$, potom platí

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} \mathrm{D}(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \mathrm{D}(Y_i).$$

Metoda hlavních komponent

Podíl k-té hlavní komponenty na celkové variabilitě je dán podílem

$$\frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}.$$

Pokud lze většinu celkové variability (např. 80–90, %) pro velké p popsat několika hlavními komponentami (např jednou, dvěma nebo třemi), potom lze původní proměnné „nahradit“ těmito komponentami bez velké ztráty informace.

Kromě vlastních čísel matice $\boldsymbol{\Sigma}$ si pozornost zaslouží i koeficienty vlastních vektorů $\boldsymbol{e} = (e_{i1}, \ldots, e_{ik}, \ldots, e_{ip})'$. Velikost hodnoty e_{ik} určuje význam k-té proměnné v i-té hlavní komponentě. Hodnota e_{ik} je úměrná korelačnímu koeficientu mezi Y_i a X_k, platí

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{ii}}} \quad i, k = 1, 2, \ldots, p.$$

Metoda hlavních komponent pro standardizované proměnné

Standardizované proměnné získáme následovně

$$Z_1 = \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}}$$
$$Z_2 = \frac{X_2 - \mu_1}{\sqrt{\sigma_{22}}}$$
$$\vdots$$
$$Z_p = \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}$$

v maticovém vyjádření

$$\boldsymbol{Z} = (\boldsymbol{V}^{1/2})^{-1}(\boldsymbol{X} - \boldsymbol{\mu}), \tag{6}$$

kde $\boldsymbol{V}^{1/2}$ je diagonální matice směrodatných odchylek. Je zřejmé, že $\mathrm{E}(\boldsymbol{Z}) = \boldsymbol{0}$ a

$$\mathrm{var}(\boldsymbol{Z}) = (\boldsymbol{V}^{1/2})^{-1}\boldsymbol{\Sigma}(\boldsymbol{V}^{1/2})^{-1} = \boldsymbol{\rho}.$$

Metoda hlavních komponent pro standardizované proměnné

Hlavní komponenty lze získat pomocí vlastních vektorů korelační matice $\boldsymbol{\rho}$. Uvažujme dvojice vlastních čísel a vektorů $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), \ldots, (\lambda_p, \boldsymbol{e}_p)$ korelační matice $\boldsymbol{\rho}$, kde $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Pro proměnné $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_p)'$ s varianční maticí $\mathrm{var}(\boldsymbol{Z}) = \boldsymbol{\rho}$, jsou hlavní komponenty dána vztahem

$$Y_i = \boldsymbol{e}_i' \boldsymbol{Z} = \boldsymbol{e}_i' (\boldsymbol{V}^{1/2})^{-1} (\boldsymbol{X} - \boldsymbol{\mu}), \quad i = 1, 2, \ldots, p$$

Dále platí

$$\sum_{i=1}^{p} \mathrm{D}(Y_i) = \sum_{i=1}^{p} \mathrm{D}(Z_i) = p$$

a

$$\rho_{Y_i, Z_k} = e_{ik} \sqrt{\lambda_i}, \quad i, k = 1, 2, \ldots, p.$$

Podíl k-té hlavní komponenty na celkové variabilitě je dán vztahem

$$\frac{\lambda_k}{p}, \quad k = 1, 2, \ldots, p.$$

Příklad

Mějme kovarianční matici

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1{,}5 \\ 1{,}5 & 4 \end{pmatrix}.$$

Odpovídající korelační matice je

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0{,}75 \\ 0{,}75 & 1 \end{pmatrix}.$$

Vlastní čísla a vlastní vektory matice $\boldsymbol{\Sigma}$ jsou

$$\begin{aligned} \lambda_1 &= 4{,}621, \quad \boldsymbol{e}_1 = (0{,}383; 0{,}924)' \\ \lambda_2 &= 0{,}379, \quad \boldsymbol{e}_2 = (-0{,}924; 0{,}383)' \end{aligned}$$

Podobně vlastní čísla a vlastní vektory matice $\boldsymbol{\rho}$ jsou

$$\begin{aligned} \lambda_1^* &= 1 + \rho = 1{,}75, \quad \boldsymbol{e}_1^* = (0{,}707; 0{,}707)' \\ \lambda_2^* &= 1 - \rho = 0{,}25, \quad \boldsymbol{e}_2^* = (-0{,}707; 0{,}707)' \end{aligned}$$

Příklad

Hlavní komponenty odpovídající matici $\boldsymbol{\Sigma}$ potom jsou

$$Y_1 = 0{,}383X_1 + 0{,}924X_2$$
$$Y_2 = -0{,}924X_1 + 0{,}383X_2$$

a pro matici $\boldsymbol{\rho}$

$$Y_1^* = 0{,}707Z_1 + 0{,}707Z_2 = 0{,}707\left(\frac{X_1 - \mu_1}{1}\right) + 0{,}707\left(\frac{X_2 - \mu_1}{2}\right)$$
$$Y_2^* = -0{,}707Z_1 + 0{,}707Z_2 = -0{,}707\left(\frac{X_1 - \mu_1}{1}\right) + 0{,}707\left(\frac{X_2 - \mu_1}{2}\right)$$

Rozptyl proměnné X_2 je větší než rozptyl X_1, k první hlavní komponentě přispívá více. Navíc tato první hlavní komponenta vysvětluje podíl

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{4{,}621}{4{,}621 + 0{,}379} = 0{,}924$$

na celkovém rozptylu.

Příklad

Podíváme-li se na standardizované proměnné, pak příspěvky obou proměnných Z_1 a Z_2 k hlavním komponentám jsou stejné. Dále platí

$$\rho_{Y_1,Z_1} = e_{11}^* \sqrt{\lambda_1^*} = 0{,}707\sqrt{1{,}75} = 0{,}935$$
$$\rho_{Y_1,Z_2} = e_{21}^* \sqrt{\lambda_1^*} = -0{,}707\sqrt{1{,}75} = -0{,}935$$

První hlavní komponenta určená ze standardizovaných veličin vysvětluje podíl

$$\frac{\lambda_1}{p} = \frac{1{,}75}{2} = 0{,}875$$

na celkovém rozptylu. Z uvedeného příkladu je vidět, že standardizace ovlivňuje výsledky metody hlavních komponent, hlavní komponenty získané z varianční matice $\boldsymbol{\Sigma}$ se liší od těch získaných z korelační matice $\boldsymbol{\rho}$.

Výběrové hlavní komponenty

Předpokládejme, že máme náhodný výběr $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ ze základního souboru se střední hodnotou $\boldsymbol{\mu}$ a varianční maticí $\boldsymbol{\Sigma}$. Určíme výběrový průměr $\overline{\boldsymbol{x}}$, výběrovou varianční matici $\boldsymbol{S}$ a výběrovou korelační matici $\boldsymbol{R}$.

Cílem je najít takové nekorelované lineární kombinace, které budou schopny popsat většinu variability ve výběru.

- První výběrová hlavní komponenta je taková lineární kombinace $\boldsymbol{a}_1'\boldsymbol{x}_j$, která maximalizuje výběrový rozptyl $\boldsymbol{a}_1'\boldsymbol{x}_j$ za podmínky $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$.
- Druhá výběrová hlavní komponenta je lineární kombinace $\boldsymbol{a}_2'\boldsymbol{x}_j$, která maximalizuje výběrový rozptyl $\boldsymbol{a}_2'\boldsymbol{x}_j$ za podmínek $\boldsymbol{a}_2'\boldsymbol{a}_2 = 1$ a nulové kovariance dvojice $(\boldsymbol{a}_1'\boldsymbol{x}_j, \boldsymbol{a}_2'\boldsymbol{x}_j)$.
- Podobným způsobem se postupuje dále, tudíž i-tá výběrová hlavní komponenta je lineární kombinace $\boldsymbol{a}_i'\boldsymbol{x}_j$ maximalizující výběrový rozptyl $\boldsymbol{a}_1'\boldsymbol{x}_j$ při splnění podmínek $\boldsymbol{a}_i'\boldsymbol{a}_i = 1$ a nulové kovariance párů $(\boldsymbol{a}_1'\boldsymbol{x}_j, \boldsymbol{a}_k'\boldsymbol{x}_j)$, $k < i$.

Výpočet výběrových hlavních komponent je úzce spojen s vlastními čísly a vektory výběrové varianční matice $\boldsymbol{S}$.

Výběrové hlavní komponenty

Je-li $\boldsymbol{S} = \{s_{ik}\}$ výběrový varianční matice typu $p \times p$ s vlastními čísly a vektory tvořící dvojice $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), (\hat{\lambda}_2, \hat{\boldsymbol{e}}_2), \ldots, (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$, pak i-tá výběrová hlavní komponenta je dána vztahem

$$\hat{y}_i = \hat{\boldsymbol{e}}_i' \boldsymbol{x} = \hat{e}_{i1} x_1 + \hat{e}_{i2} x_2 + \cdots + \hat{e}_{ip} x_p, \quad i = 1, 2, \ldots, p,$$

kde $\boldsymbol{x}$ je libovolné pozorování proměnných $X_1, X_2, \ldots, X_p$. Dále platí, že výběrový rozptyl $\hat{y}_k$ je roven $\hat{\lambda}_k$, $k = 1, 2, \ldots, p$, výběrová kovariance dvojic $(\hat{y}_i, \hat{y}_k)$ je pro $i \neq k$ rovna nule. Celkový výběrový rozptyl je potom dán vztahem

$$\sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p$$

a korelační koeficienty výrazem

$$r_{\hat{y}_i, x_k} = \frac{\hat{e}_{ik} \sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}, \quad i, k = 1, 2, \ldots, p.$$

Při analýze dat je často používáno „centrování“ odečtením výběrového průměru $\overline{\boldsymbol{x}}$. Toto centrování nemá vliv na výběrovou varianční matici. Pro libovolný vektor pozorování $\boldsymbol{x}$ je i-tá hlavní komponenta dána

$$\hat{y}_i = \hat{\boldsymbol{e}}_i'(\boldsymbol{x} - \overline{\boldsymbol{x}}), \quad i = 1, 2, \ldots, p.$$

Grafické zobrazení hlavních komponent

S využitím jednoduchých grafických nástrojů lze nalézt „podezřelá“ pozorování, lze také provést ověření předpokladu normality. Obvykle se provádí ověření normality pro několik prvních hlavních komponent. To lze udělat například pomocí QQ-plotu.

Na druhé straně poslední hlavní komponenty mohou pomoci při odhalení neobvyklých pozorování. Každé pozorování lze vyjádřit jako lineární kombinaci

$$\boldsymbol{x}_j = (\boldsymbol{x}_j'\hat{\boldsymbol{e}}_1)\hat{\boldsymbol{e}}_1 + (\boldsymbol{x}_j'\hat{\boldsymbol{e}}_2)\hat{\boldsymbol{e}}_2 + \cdots + (\boldsymbol{x}_j'\hat{\boldsymbol{e}}_p)\hat{\boldsymbol{e}}_p = \hat{y}_{j1}\hat{\boldsymbol{e}}_1 + \hat{y}_{j2}\hat{\boldsymbol{e}}_2 + \cdots + \hat{y}_{jp}\hat{\boldsymbol{e}}_p$$

všech vlastních vektorů $\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \ldots, \hat{\boldsymbol{e}}_p$ matice $\boldsymbol{S}$. Velikost posledních hlavních komponent poukazuje na to, jak dobře prvních několik hlavních komponent propisují data. Takže $\hat{y}_{j1}\hat{\boldsymbol{e}}_1 + \hat{y}_{j2}\hat{\boldsymbol{e}}_2 + \cdots + \hat{y}_{j,q-1}\hat{\boldsymbol{e}}_{q-1}$ se liší od $\boldsymbol{x}_j$ o $\hat{y}_{jq}\hat{\boldsymbol{e}}_q + \cdots + \hat{y}_{jp}\hat{\boldsymbol{e}}_p$, tedy o délku $\sqrt{\hat{y}_{jq}^2 + \cdots + \hat{y}_{jp}^2}$. Neobvyklé pozorování obvykle takové, jehož souřadnice $\hat{y}_{jq}, \ldots, \hat{y}_{jp}$ jsou velké.

Grafické zobrazení hlavních komponent

Dalším typem grafu, který se obvykle v analýze hlavních komponent používá, je bodový graf výběrových hlavních komponent (např. na vodorovnou osu vyneseme souřadnice první výběrové hlavní komponenty určené z analyzovaných dat, na osu svislou potom druhou výběrovou hlavní komponentu).
Předpokládejme, že analyzovaná data jsou výběrem z vícerozměrného náhodného rozdělení se střední hodnotou $\boldsymbol{\mu}$ a varianční maticí $\boldsymbol{\Sigma}$. Uvažujme první dvě hlavní komponenty. Výběrový rozptyl první hlavní komponenty $\hat{y}_1$ je $\hat{\lambda}_1$, výběrový rozptyl druhé hlavní komponenty $\hat{y}_2$ je $\hat{\lambda}_2$. Vykreslíme-li bodový graf hodnot $(\hat{y}_{j1}), \hat{y}_{j2}$ pro $j = 1, 2, \ldots, n$, měly tyto body ležet uvnitř elipsy o rovnici

$$\frac{\hat{y}_1^2}{\hat{\lambda}_1} + \frac{\hat{y}_2^2}{\hat{\lambda}_2} \leq \chi^2_{1-\alpha}(2).$$

Vynesené body by měly mít přibližně eliptický tvar.