Analýza hlavních komponent

Lenka Přibylová

27. března 2014

Opakování pojmů z lineární algebry

Matice A se nazývá **symetrická**, jestliže pro všechny její prvky platí

$$a_{ij} = a_{ji}.$$

Symetrická matice A se nazývá **pozitivně semidefinitní**, jestliže pro každý sloupcový vektor x platí:

$$x^T Ax \geq 0.$$

Reálná (nebo komplexní) čísla λ_i, pro které platí

$$Au_i = \lambda_i u_i, \ \ i = 1, \ldots, n.$$

se nazývají **vlastní čísla** matice A a vektory $u_i \neq 0$ se nazývají **vlastní vektory** matice A.

Opakování pojmů z lineární algebry

Věta - spektrální rozklad reálné symetrické matice A:

Nechť $A \in R^{n,n}$ je symetrická matice. Pak existují ortogonální matice $U \in R^{n,n}$ a diagonální matice $D \in R^{n,n}$ tak, že platí:

$$A = UDU^T.$$

Pozn. Matice D je diagonální matice, jejímiž prvky jsou vlastní čísla $\lambda_1, \lambda_2, \ldots, \lambda_n$ matice A a U je ortogonální matice normovaných vlastních vektorů psaných do sloupců, odpovídajících po řadě příslušným vlastním číslům $\lambda_1, \lambda_2, \ldots, \lambda_n$.

Motivace

Analýza hlavních komponent (Principal Component Analysis, PCA)

Motivace: Nahrazení velkého počtu vstupních proměnných mnohem menším počtem nových proměnných tzv. **komponent** bez větší ztráty podstatné informace o vstupních datech.

Od nových proměnných se požaduje, aby maximálně reprezentovaly původní proměnné. U metody hlavních komponent se požaduje, aby nové proměnné (linerní kombinace původních proměnných) co nejvíce vysvětlovaly **variabilitu** původních proměnných.

Historie vs. současnost

Historie

1901 - Karl Pearson ... popisná statistická metoda, sloužící k redukci dat

1933 - Harold Hotelling ... zobecnění postupu na náhodné vektory a návrh použití pro rozbor kovarianční struktury proměnných

Současnost:

1. součást explorační analýzy dat
2. pomocník jiných metod analýzy vícerozměrných pozorování
3. samostatná metoda vícerozměrné analýzy proměnných

Definice hlavní komponenty

Uvažujme populaci s p-rozměrným náhodným vektorem $X = (X_1, \ldots, X_p)$, přičemž náhodné veličiny $X_1, \ldots, X_p$ mají p-rozměrné normální rozdělení s vektorem středních hodnot $\mu = (\mu_1, \ldots, \mu_p)$ a kovarianční maticí

$$\Sigma = \begin{pmatrix} DX_1 & \ldots & cov(X_1, X_p) \\ \vdots & \ddots & \vdots \\ cov(X_1, X_p) & \ldots & DX_p \end{pmatrix}.$$

Předpokládejme, že známe μ, Σ. Označme vlastní čísla matice Σ po řadě $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ a příslušné ortonormální vlastní vektory $v_1, v_2, \ldots, v_p$.

Definice hlavní komponenty

Náhodnou veličinu Y_1 definovanou vztahem

$$Y_1 = v_{11}(X_1 - \mu_1) + v_{12}(X_2 - \mu_2) + \cdots + v_{1p}(X_p - \mu_p)$$

nazýváme **první hlavní komponenta** náhodného vektoru X.

Analogicky definujeme $Y_2, \ldots, Y_r$, kde $r \leq p$.

Pozn: Koeficienty LK u první hlavní komponenty jsou tvořeny souřadnicemi vlastního vektoru $v_1 = (v_{11}, v_{12}, \ldots, v_{1p})$ odpovídajícímu největšímu vlastnímu číslu λ_1. Obdobně u ostatních hlavních komponent.

Vlastnosti hlavních komponent

Rozptyl DY_r r-té hlavní komponenty Y_r je roven r-tému vlastnímu číslu λ_r kovarianční matice Σ, tj

$$DY_r = \lambda_r.$$

Pro **variabilitu**, tj. součet rozptylů, $r \leq p$ hlavních komponent platí

$$DY_1 + \cdots + DY_r = \lambda_1 + \cdots + \lambda_r$$

Míra významu r-té hlavní komponenty Y_r

Mírou významu r-té hlavní komponenty Y_r, z hlediska vysvětlované celkové variability veličin $Y_1, Y_2, \cdots, Y_r$, je podíl

$$\frac{DY_r}{DY_1 + DY_2 + \cdots + DY_r} = \frac{\lambda_r}{tr\Lambda} = \frac{\lambda_r}{\lambda_1 + \lambda_2 + \cdots + \lambda_r}.$$

Pozn. Obecně lze takto dosáhnout stavu, ve kterém úplný systém všech $r = p$ komponent $Y_1, Y_2, \cdots, Y_r$ jednoznačně (beze zbytku) vysvětlí celkový rozptyl výchozích náhodných veličin $X_1, X_2, \cdots, X_r$. V praxi však preferujeme stav, kdy počet hlavních komponent $r << p$, tj. r bude mnohem nižší než p. Obvykle se za ideální považuje 2 - 7 hlavních komponent, i když rozměr p zadaných dat je mnohem vyšší.

Komponentní skóre (komponentní váhy)

Pro účely prakticky zaměřených statistických analýz zavádíme pojem **komponentní skóre (komponentní váha)**. Počítáme hodnoty hlavních komponent zvlášť pro každou výběrovou jednotku. Označme jako x_i vektor hodnot i-té jednotky souboru $(i = 1, \cdots, p)$.

Komponentní skóre (komponentní váha) r-té hlavní komponenty u i-té jednotky je definováno následujicím vztahem:

$$y_{ir} = v_r^T (x_i - \mu), \;\; r = 1, 2, \cdots, R, \;\; i = 1, 2, \cdots, p.$$

kde vlastní vektory v_i splňují ortonormalizační podmínku:

$$v_r^T v_r = 1, \;\; r = 1, 2, \cdots, R.$$

Teoretický příklad 1

Mějme náhodný vektor $X = (X_1, X_2)$ s vektorem středních hodnot $\mu = (2, 3)$ a kovariační maticí

$$\mathbf{\Sigma} = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}.$$

Najděme hlavní komponenty náhodného vektoru $X = (X_1, X_2)$.

Teoretický příklad 1

Řešení: Nejprve najdeme vlastní čísla kovarianční matice Σ:

$$0 = det(\Sigma - \lambda I) = (5 - \lambda)(1 - \lambda) - 4 = \lambda^2 - 6\lambda + 1.$$

Vlastní čísla: $\lambda_1 \doteq 5,83,\ \lambda_2 \doteq 0,17$. Vypočtené podíly

$$\frac{DY_1}{DY_1 + DY_2} = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{\lambda_1}{st\Sigma} \doteq \frac{5,83}{6} \doteq 97\%,$$

$$\frac{DY_2}{DY_1 + DY_2} = \frac{\lambda_2}{\lambda_1 + \lambda_2} = \frac{\lambda_2}{st\Sigma} \doteq \frac{0,17}{6} \doteq 3\%$$

ukazují v tomto případě na vhodnost volby jedné hlavní komponenty.

Teoretický příklad 1

Určeme tvar této hlavních komponenty. Normovaný vlastní vektor odpovídající vlastnímu číslu λ_1 je roven $v_1 = (0,38; 0,92)$, tedy

$$Y_1 = (0,38; 0,92)^T \cdot [(X_1, X_2) - (2,3)] \doteq 0,38 \cdot (X_1 - 2) + 0,92 \cdot (X_2 - 3).$$

První hlavní komponenta:

$$Y_1 = 0,38 \cdot (X_1 - 2) + 0,92 \cdot (X_2 - 3).$$

Na závěr určeme ještě komponentí skóre y_1 první hlavní komponenty Y_1 pro jednu konkrétní realizaci např. $x = (2,1; 3,2)$ náhodného vektoru $X = (X_1,\ X_2)$:

$$y_1 = 0,38 \cdot (2,1 - 2) + 0,92 \cdot (3,2 - 3) \doteq 0,22.$$

Teoretický příklad 2

Mějme náhodný vektor $X = (X_1, X_2)$ s vektorem středních hodnot $\mu = (2, 3)$ a kovariační maticí

$$\mathbf{\Sigma} = \begin{pmatrix} 10 & 1 \\ 1 & 9 \end{pmatrix}.$$

Najděme hlavní komponenty náhodného vektoru $X = (X_1, X_2)$.

Teoretický příklad 2

Řešení: Najdeme opět nejprve vlastní čísla kovarianční matice Σ:

$$0 = det(\Sigma - \lambda I) = (10 - \lambda)(9 - \lambda) - 1 = \lambda^2 - 19\lambda + 89.$$

Vlastní čísla: $\lambda_1 \doteq 10,62,\ \lambda_2 = 8,38$. Na vhodnost volby dvou hlavních komponent v tomto případě ukazují vypočtené podíly

$$\frac{DY_1}{DY_1 + DY_2} = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{\lambda_1}{st\Sigma} \doteq \frac{10,62}{19} \doteq 56\%,$$

$$\frac{DY_2}{DY_1 + DY_2} = \frac{\lambda_2}{\lambda_1 + \lambda_2} = \frac{\lambda_2}{st\Sigma} \doteq \frac{8,38}{19} \doteq 44\%.$$

Teoretický příklad 2

Vlastní vektory: $v_1 = (0,85; 0,53)$, $v_2 = (0,53; -0,85)$.

Vektory v_1, v_2 jsou ortogonální (kolmé).

Výpočet první hlavní komponenty Y_1:

$$Y_1 = (0,85; 0,53)^T \cdot [(X_1, X_2) - (2,3)] \doteq 0,85 \cdot (X_1 - 2) + 0,53 \cdot (X_2 - 3).$$

Výpočet druhé hlavní komponenty Y_2:

$$Y_2 = (0,53, -0,85)^T \cdot [(X_1, X_2) - (2,3)] \doteq 0,53 \cdot (X_1 - 2) - 0,85 \cdot (X_2 - 3).$$

Teoretický příklad 2

První a druhá hlavní komponenta:

$$Y_1 \doteq 0,85 \cdot (X_1 - 2) + 0,53 \cdot (X_2 - 3),$$

$$Y_2 \doteq 0,53 \cdot (X_1 - 2) - 0,85 \cdot (X_2 - 3).$$

Komponentní skóre y_1, y_2 komponent Y_1, Y_2 pro $x = (2,1; 3,2)$:

$$y_1 = 0,85 \cdot (2,1 - 2) + 0,53 \cdot (3,2 - 3) \doteq 0,19,$$

$$y_2 = 0,53 \cdot (2,1 - 2) - 0,85 \cdot (3,2 - 3) \doteq -0,12.$$

PCA ve výběru

Při praktických úlohách neznáme kovarianční matici Σ, ale pouze její protějšek ve výběru, tj. výběrovou kovarianční matici S. Doplníme-li předpoklady o vícerozměrném rozdělení $X_1, X_2, \ldots X_p$ o předpoklad vícerozměrné normality, lze dokázat, že vlastní čísla a vlastní vektory vypočtené z výběrové kovarianční matice jsou **maximálně věrohodnými odhady** svých protějšků ze základního souboru. Bez předpokladů normality půjde zpravidla jen o **konzistentní odhady**. Příslušným χ^2 testem na závěr obvykle testujeme hypotézu o rovnosti "zbývajících" vlastních čísel (podstatných je prvních r komponent), tj. hypotézu $H_0 : \ \lambda_{r+1} = \cdots = \lambda_p$.

Geometrický význam PCA

Projekce p-rozměrného eukleidovského prostoru do prostoru nižší dimenze r (projekce p proměnných do r komponent).

Pokud nastane případ, kdy $r = p$, nedochází k redukci počtu proměnných, tedy k projekci do nižší dimenze, ale k rotaci původní souřadnicové soustavy do směru maximálního rozptylu shluku bodů.

Transformace p souřadnic systému do r komponent reprezentovaná maticí V^T, jejíž řádky tvoří prvních r vlastních vektorů výběrové kovarianční matice S, tedy umožňuje zachytit na několika prvních osách maximum informace o prostorové struktuře souboru vícerozměrných pozorování.

Příklad řešený pomocí Statgraphicsu

Formulace problému Máme k dispozici záznam hodnot krevního tlaku od 177-mi pacientů, kteří jsou sledováni v kardiologické ambulanci. Hodnoty krevního tlaku byly měřeny každých 30 minut, u každého pacienta bylo naměřeno 48 hodnot v průběhu 24 hodin. Dále máme k dispozici informaci, zda pacient užívá či neužívá beta-blokátory.

Příklad řešený pomocí Statgraphicsu

Příklad řešený pomocí Statgraphicsu

Příklad řešený pomocí Statgraphicsu

Řešení problému pomocí software Statgraphics, verze 5.

Vstupní datová matice 177×48 (48 údajů u 177-mi pacientů).

Statgraphics vyhodnotil v tomto případě jako reprezentativní počet **osm** hlavních komponent (pro $\lambda > 1$) vysvětlujících 69% variability proměnných. **Čtyři** dominantní hlavní komponenty (pro $\lambda > 2$) vysvětlují 58% rozptylu proměnných. **První** hlavní komponenta vysvětuje variabilitu proměnných ze 40%.

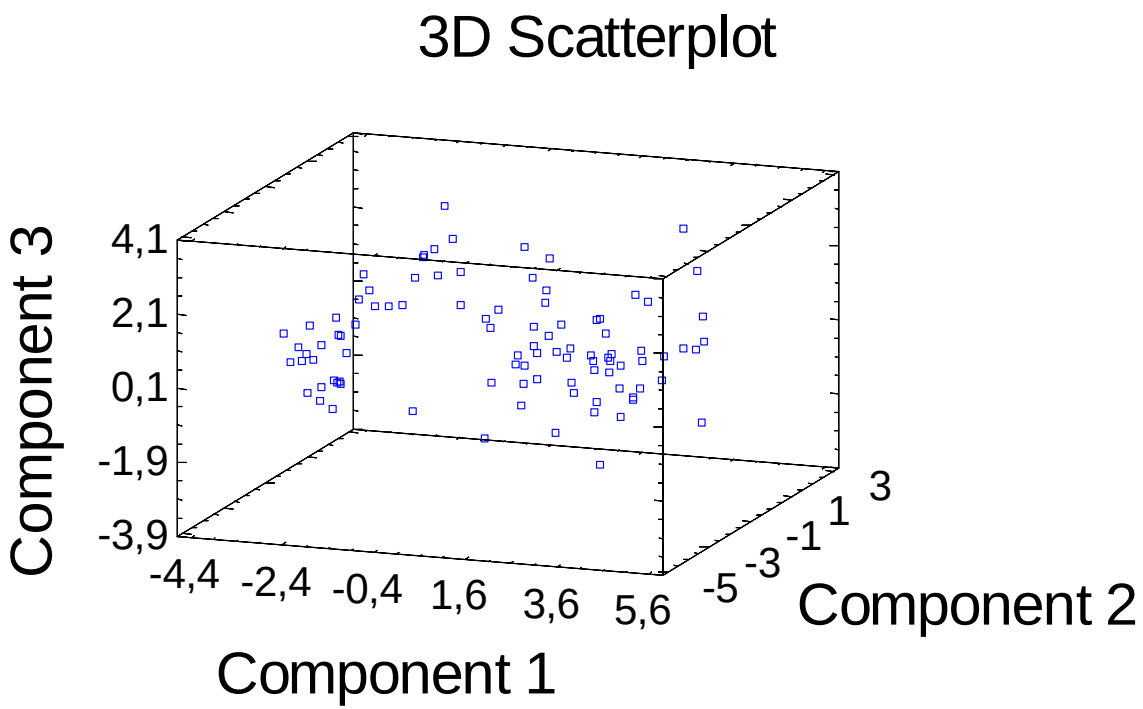
Vstupní datová matice 177×48 je tedy redukována na matici 177×8, resp 177×4.

Výstupy ze Statgraphicsu

Table of PCA

C.n.	Eigenvalue	% of Variance	Cumulative %
1	19,32	40,25	40,25
2	3,71	7,72	47,97
3	2,50	5,21	53,18
4	2,32	4,83	58,00
5	1,73	3,60	61,60
6	1,31	2,74	64,34
7	1,23	2,57	66,90
8	1,04	2,17	69,07

Výstupy ze Statgraphicsu



Výstupy ze Statgraphicsu

U výstupu je podstatný počet hlavních komponent. Pro doplnění uveďme např. tvar první hlavní komponenty (X_i jsou uvedeny ve standardizované podobě):

$$Y_1 = 0,17 * h0 + 0,17 * h0,30 + 0,18 * h1 + 0,17 * h1,30 + 0,12 * h10 + 0,12 * h10,30 +$$
$$+0,12 * h11 + 0,09 * h11,30 + 0,13 * h12 + 0,12 * h12,30 + 0,13 * h13 + 0,13 * h13,30 +$$
$$+0,13 * h14 + 0,13 * h14,30 + 0,14 * h15 + 0,15 * h15,30 + 0,15 * h16 + 0,15 * h16,30 +$$
$$+0,11 * h17 + 0,12 * h17,30 + 0,15 * h18 + 0,14 * h18,30 + 0,14 * h19 + 0,15 * h19,30 +$$
$$+0,17 * h2 + 0,17 * h2,30 + 0,14 * h20 + 0,15 * h20,30 + 0,17 * h21 + 0,16 * h21,30 +$$
$$+0,16 * h22 + 0,16 * h22,30 + 0,16 * h23 + 0,17 * h23,30 + 0,17 * h3 + 0,16 * h3,30 +$$
$$+0,16 * h4 + 0,16 * h4,30 + 0,16 * h5 + 0,16 * h5,30 + 0,13 * h6 + 0,13 * h6,30 +$$
$$+0,11 * h7 + 0,09 * h7,30 + 0,11 * h8 + 0,13 * h8,30 + 0,11 * h9 + 0,11 * h9,30.$$

Porovnání výsledků s údaji známými z medicínské praxe

Z medicínské praxe je známo, že u údajů postupně měřených tlaků během dne je podstatný vliv léku, tzv. beta-blokátoru, který upravuje hodnoty krevního tlaku. Provedli jsme tedy následující srovnání:

První hlavní komponenta - vliv beta-blokátoru, vysvětluje variabilitu ze 40%. Jedná se tedy o velice **silný latentní faktor** ovlivňující hodnoty krevního tlaku.

Porovnání výsledků s údaji známými z medicínské praxe

Dalších sedm hlavních komponent (celkem vysvětlují variabilitu z 29%), nemají již tak silný vliv na hodnoty krevního tlaku:

1. zdravý/nemocný
2. sekundární hypertenze - vliv jiného onemocnění
3. obezita
4. dědičné faktory
5. chronický stres
6. nadměrné solení,
7. nadměrná konzumace kávy a čaje

Shrnutí získaných výsledků

Analyzováním vstupního datového souboru metodou hlavních komponent jsme zjistili počet dominantních latentních vlivů, které mohou ovlivňovat data, tedy hodnoty naměřeného krevního tlaku v průběhu 24 hodin tzv. holterovským typem měření. PCA odhalí počet skrytých vlivů, nikoli přesně jejích podobu. Výše uvedené srovnání se znalostmi z medicínské praxe je jen jednou z možných interpretací výsledku této PCA analýzy.

Nevýhody a úskalí PCA

1. Subjektivní apriorní volba počtu komponent. Neexituje přesné kritérium pro apriorní stanovení počtu komponent, tedy jak velký reziduální rozptyl je již zanedbatelný (10 %, 5% anebo 1 % ?).
2. Software automaticky vyhodnotí počet komponent podle počtu vlastních čísel větších než 1 (lze uživatelem upravit).
3. Komponenta je neměřitelná (skrytá) veličina, není snadné jí přiřadit konkrétní praktický význam (měřitelné jednotky, ...) a často to ani nelze.
4. Pokud proměnné nejsou zadány ve srovnatelných jednotkách, je nutné použít korelační matici místo kovarianční matice.

Poznámky k použití metody hlavních komponent

Někteří autoři mylně ztotožňují pojmy SVD (singular value decomposition) a metoda hlavních komponent. Vhodnější by bylo uvést, že metoda hlavních komponent je statistická metoda jejímž algebraickým základem je SVD, přesněji spektrální rozklad symetrické pozitivně semidefinitní kovarianční matice.

Při vyhodnocování datového souboru pomocí této metody proměnné nerozdělujeme na vysvětlované a vysvětlující, ale bereme je rovnocenně. Ptáme se z kolika procent lineární kombinace proměnných vysvětluje celkovou variabilitu proměnných.

Poznámky k použití metody hlavních komponent

Komponentní (PCA) i faktorová analýza (FA) řeší situaci kdy výchozí počet statistických proměnných je relativně vysoký a datový soubor je nepřehledný. Pro zpřehlednění je dobré zkoumat, zda by původní proměnné bylo možno nahradit menším počtem jiných proměnných, shrnujících poznatky o výchozích proměnných, aniž by došlo k velké ztrátě podstatných informací z původního datového souboru. Od nových proměnných se požaduje, aby maximálně reprezentovaly původní proměnné. U metody hlavních komponent požadujeme, aby nové proměnné co nejvíce vysvětlovaly variabilitu původních proměnných.

Poznámky k použití metody hlavních komponent

Výsledky PCA a FA bývají u praktických příkladů de facto stejné. Počet nalezených faktorů při FA bývá roven počtu hlavních komponent při PCA. Některé brožury uvádějí, že PCA je pouze matematický aparát pro FA, což také není přesné. Jedná se o dvě rozdílné statistické metody, které však dávají rovnocenné výsledky.

Souvislost mezi SVD a metodou hlavních komponent

Pokusme se na závěr dokázat následující tvrzení:

Koeficienty $v_{11},\ v_{12}, \ldots,\ v_{1p}$ lineární kombinace původních náhodných veličin figurujících ve vyjádření **první** hlavní komponenty

$$Y_1 = v_{11}(X_1 - \mu_1) + v_{12}(X_2 - \mu_2) + \cdots + v_{1p}(X_p - \mu_p),$$

odpovídají souřadnicím ortonormálního vlastního vektoru kovarianční matice Σ, který přísluší největšímu vlastnímu číslu λ_1.

Souvislost mezi SVD a metodou hlavních komponent

Výše uvedený problém optimalizace budeme řešit metodou Lagrangeových multiplikátorů. Hledáme maximum funkce

$$DY_1 = v^T \mathbf{\Sigma} v,$$

která charakterizuje rozptyl náhodné veličiny Y_1 za podmínky, že $v^T v = 1$.
Označme v_1 hodnotu, ve které nabývá funkce DY_1 maxima, tj.

$$v_1 = argmax\ v^T \mathbf{\Sigma} v, \text{ kde } v^T v = 1.$$

Zvolme za Lagrangeovu funkci L s Lagrangeovým multiplikátorem λ funkci

$$L(v, \lambda) = v^T \mathbf{\Sigma} v - \lambda(v^T v - 1).$$

Souvislost mezi SVD a metodou hlavních komponent

Hodnota maxima se nalézá v bodě, kde je gradient Lagrangeovy funkce roven 0:

$$\nabla L(v, \lambda) = 2\Sigma v - 2\lambda v = o.$$

neboli

$$(\Sigma - \lambda I)v = o.$$

Abychom získali netriviální řešení této rovnice, je třeba volit hodnotu Lagrangeova multiplikátoru λ tak, že splňuje rovnici

$$det(\mathbf{\Sigma} - \lambda \mathbf{I}) = 0,$$

která je charakteristickou rovnicí a tedy λ je vlastním číslem matice Σ.

Souvislost mezi SVD a metodou hlavních komponent

Abychom dostali maximální rozptyl, musí být multiplikátor λ roven největšímu vlastnímu číslu λ_1. Vektor v_1 je odpovídající vlastní vektor. Rozptyl hlavní komponenty Y_1 je pak roven přímo vlastnímu číslu λ_1:

$$DY_1 = v_1^T \Sigma v_1 = \lambda_1.$$

Analogicky budeme postupovat u druhé hlavní komponenty. Budeme maximalizovat rozptyl DY_2 pomocí Langrangeovy funkce $L(v, \lambda, m)$ s Lagrangeovými multiplikátory λ, m, přičemž musí být splněny následující podmínky: $v^T v = 1$, $v_1^T v = o$

$$L(v, \lambda, m) = v^T \mathbf{\Sigma} v - \lambda(v^T v - 1) + m v_1^T v.$$

Souvislost mezi SVD a metodou hlavních komponent

Gradient této funkce položíme roven 0:

$$\nabla L(v, \lambda, m) = 2\Sigma v - 2\lambda v + m v_1 = o.$$

Jelikož platí, že $v_1^T v_2 = 0$, protože v_1 a v_2 jsou ortogonální, bude 2. Lagrangeův multiplikátor m vždy roven 0. Tedy dostáváme analogickou rovnici, jako v 1. případě:

$$(\Sigma - \lambda I)v = o,$$

která má netriviální řešení, jestliže λ splňuje charakteristickou rovnici

$$det(\mathbf{\Sigma} - \lambda \mathbf{I}) = 0.$$

Souvislost mezi SVD a metodou hlavních komponent

Řešením je vlastní vektor v_2 odpovídající druhému největšímu vlastnímu číslou λ_2.

Vztahy pro další hlavní komponenty lze obdobně odvodit pomocí ostatních vlastních vektorů matice Σ.

Pozn. Vektory koeficientů j-té a k-té hlavní komponenty jsou pro $j \neq k$ nutně ortogonální.

Souvislost mezi SVD a metodou hlavních komponent

Označme dále jako V ortogonální matici, jejíž sloupce jsou tvořeny vlastními vektory matice Σ. Komponentní analýza je založena na transformaci nánodného vektoru X na náhodný vektor Y, přičemž platí, že $Y = V^T X$. Matice V je tedy transformační maticí. Kovarianční matice vektoru komponent $Y = (Y_1, Y_2, \ldots, Y_R)$ má tvar

$$Cov(Y) = V^T \Sigma V = \Lambda,$$

kde Λ je diagonální matice s vlastními čísly λ_i na diagonále.

Souvislost mezi SVD a metodou hlavních komponent

Kovarianční matice vektoru původních náhodných veličin X a komponent Y je obdobně dána jako

$$Cov(X, Y) = \Sigma V,$$

kde k-tý sloupec této matice tvořený vektorem Σv_k udává kovariance původních náhodných veličin X_i s k-tou komponentou. Vzhledem k tomu, že vektor v_k je vlastním vektorem matice Σ příslušný vlastnímu číslu λ_k, platí

$$\Sigma v_k = \lambda_k v_k.$$

Souvislost mezi SVD a metodou hlavních komponent

Kovarianci j-té náhodné veličiny X_j s k-tou komponentou Y_k lze tedy vyjádřit hodnotou $\lambda_k v_{jk}$. Korelační koeficienty původních náhodných veličin s k-tou komponentou tvoří souřadnice vektoru $\mathbf{v}_k\sqrt{\lambda_k}$. Korelační koeficienty původních náhodných veličin s komponentami jsou obvykle základem pro interpretaci hlavních komponent. Tyto koeficienty korelace jsou vlastně komponentními vahami nalezených komponent. Vektory $\sqrt{\lambda_k}\mathbf{v}_k$ mají důležitý vztah ke kovarianční, resp. korelační matici, ze které byly odvozeny.

Souvislost mezi SVD a metodou hlavních komponent

Pomocí matic V a Λ můžeme zkonstruovat spektrální rozklad symetrické pozitivně semidefinitní kovarianční matice Σ:

$$\Sigma = V \Lambda V^T,$$

kde Λ je diagonální matice vlastních čísel a V je ortogonální matice vlastních vektorů. Výše uvedená rovnice ukazuje, že analýza hlavních komponent je ekvivaletní spektrálnímu rozkladu matice Σ. Rozklad kovarianční matice je rovněž cílem faktorové analýzy, ovšem na rozdíl od faktorové analýzy je v komponentní analýze tento rozklad pro různá vlastní čísla jednoznačný (díky normalizační podmínce pro vektory v_i).

Souvislost mezi SVD a metodou hlavních komponent

Zavedeme-li navíc matici $A = V\Lambda^{\frac{1}{2}}$, pak sloupce matice A reprodukují matici Σ na základě vztahu

$$\Sigma = \sum_{k=1}^{p} \lambda_k \mathbf{v}_k \mathbf{v}_k^T = AA^T.$$

Jsou-li z matice Σ postupně získávány hlavní komponenty, lze tvořit matice $\lambda_k \mathbf{v}_k \mathbf{v}_k^T$ a srovnávat jejich postupný součet s maticí Σ. Toto srovnání ukazuje, do jaké míry je kovarianční matice Σ reprodukována dosud vytvořenými hlavními komponentami. Analýza hlavních komponent je tedy ekvivalentní spektrálnímu rozkladu matice A, resp. rozkladu matice Σ na součin matice A a k ní transponované. Tento rozklad je (narozdíl např. od faktorové analýzy) jednoznačný.

Literatura

1. Eldén L. : Matrix Methods in Data Mining and Pattern Recognition, SIAM, Philadelphia 2007
2. Dostál Z. : Lineární algebra, VŠB-TU Ostrava 2004
3. Hebák P., Hustopecký J., Malá I. : Vícerozměrné statistické metody I., II., III., Informatorium, Praha 2005
4. Hebák P., Hustopecký J. : Vícerozměrné statistické metody s aplikacemi, SNTL/ALFA, Praha 1987
5. Laub J. A. : Matrix analysis for scientists and engineers, Davis, California 2005

Závěr

Děkuji za pozornost.