

Univerzita Pardubice

8. licenční studium chemometrie

Statistické zpracování dat při managementu jakosti

Semestrální práce

**Metody s latentními proměnnými a klasifikační
metody**

**Ing. Jan Balcárek, Ph.D.
vedoucí Centrálních laboratoří
Precheza a.s., Přerov**

Přerov, 20.11.2000

Příklad 1

Vypočtete algoritmem NIPALS první latentní proměnnou z matice A:

$$\begin{bmatrix} 1 & 1 & 6 \\ 2 & 2 & 4 \\ 3 & 0 & 2 \end{bmatrix}$$

Matici před zpracováním standardizujte.

Prvním krokem je výpočet vektoru aritmetických průměrů $\bar{x}^T$ a vektoru směrodatných odchylek s^T .

$$\bar{x}^T = [(1+2+3)/3 \quad (1+2+0)/3 \quad (6+4+2)/3] = [2 \quad 1 \quad 4]$$

$$s^T = [s_1 \quad s_2 \quad s_3] = [1 \quad 1 \quad 2]$$

Provedením standardizace se získá matice A:

$$\begin{bmatrix} -1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & -1 & -1 \end{bmatrix}$$

Variabilita všech sloupců je stejná, proto lze vzít za odhad hlavní komponenty libovolný sloupec, např. první:

$$t_1^T = [-1 \quad 0 \quad 1]$$

Po dosazení do vzorce

$$p_1^T = (t_1^T t_1)^{-1} t_1^T A$$

a znormování podle vztahu

$$p_1^N = (p_1^T p_1)^{-1/2} p_1$$

se získá počáteční odhad vektoru vyjadřujícího míru příspěvků odhadu hlavní komponenty t_1 . Následujícím dosazením do vztahu:

$$t_1^T = (p_1^T p_1)^{-1} p_1^T A^T$$

se získá odhad hlavní komponenty t_1 . Opakováním postupu se získá stabilní rozklad vektorů t_1 a p_1 . Jako konvergenční kritérium se použije vztah:

$$d = (t_{\text{nové}} - t_{\text{staré}})^T (t_{\text{nové}} - t_{\text{staré}}) (t_{\text{nové}}^T t_{\text{nové}})^{-1}$$

K ukončení výpočtu je třeba dosáhnou $d/N < 10^{-10}$.

K dosažení této podmínky je třeba devíti kroků. Získá se následující stabilní rozklad vektorů:

$$\begin{array}{l} p_1^T = [0,62797 \quad -0,45969 \quad -0,62797] \\ t_1^T = [-1,25594 \quad -0,45969 \quad 1,71562] \end{array}$$

Příklad 2

S použitím vhodných kritérií určete nezbytný počet latentních proměnných, bylo-li z dat určeno: PRESS(0) = S(0) = 100; PRESS(1) = 20; S(1) = 10; PRESS(2) = 3,5; S(2) = 3,4; PRESS(3) = 3,45; S(3) = 3,39.

Nezbytný počet latentních proměnných se určí s pomocí Woldova kritéria. To je založeno na poměru:

$$\text{PRESS}(P) / S_R(P - 1)$$

Je-li hodnota kritéria větší než 0,95, je zařazení další (P+1) latentní proměnné nevhodné.

Po dosazení dostaneme:

$$\begin{array}{ll} P = 1: & \text{PRESS}(1) / S_R(0) = 20 / 100 = 0,2 \\ P = 2: & \text{PRESS}(2) / S_R(1) = 3,5 / 10 = 0,35 \\ P = 3: & \text{PRESS}(3) / S_R(2) = 3,45 / 3,4 = 1,01 \end{array}$$

Pro $P = 3$ je již hodnota vyšší než 0,95, čtvrtá latentní proměnná již není významná.

Příklad 3

Odhadněte hodnotu chybějícího prvku A [2,2], jestliže výpočtem z nekompletní matice byly určeny vektory p: 0,541 0,423 0,514 0,514 t: -1,340 -0,735 2,076

Prvky zdrojové matice odpovídající p-té latentní proměnné lze získat ze vzorce:

$$A_p^{\text{pred}} = t_p p_p^T$$

Pro odhad hodnoty prvku A[2,2] takto platí: $A[2,2] = t[2] p[2] = -0,311$

Příklad 4

Výpočtem metodou PCA byly určeny vektory:

$$p_1^T: \quad 0,012 \quad 0,458 \quad -0,352 \quad 0,987$$

$$p_2^T: \quad 0,926 \quad -0,238 \quad 0,872 \quad -0,115$$

Vyberte reprezentativní sloupce, které nejlépe charakterizují celou matici.

Míra příspěvku příslušné latentní proměnné k popisu variability sloupců zdrojové matice je představována vektory. Vektor p_1^T ukazuje, že pro zdrojovou matici je reprezentativním sloupcem sloupec čtvrtý $p_1^T [4] = 0,987$. V případě vektoru p_2^T je reprezentativním sloupcem sloupec první $p_2^T [1] = 0,926$ a sloupec třetí $p_2^T [3] = 0,872$.

Příklad 5

Vysvětlete, proč vysvětlená variabilita je při výpočtu metodou FA vždy nižší, než při výpočtu metodou PCA.

Metoda PCA je z hlediska faktorové analýzy považována za úplnou komponentní analýzu. Pomocí hlavních komponent je možné přesně reprodukovat variabilitu zdrojové matice. Metoda FA je z tohoto pohledu neúplnou komponentní analýzou, protože připouští existenci matice jedinečností, která zahrnuje část variability, která se nedá vysvětlit společnými faktory.

Příklad 6

Výpočtem metodou kanonických korelací bylo zjištěno:

$$0,297 X_1 + 0,298 X_2 + 0,050 X_3 + 0,256 X_4 = 0,493 Y_1 - 0,213 Y_2 \quad r_1 = 0,830$$

$$0,006 X_1 - 0,115 X_2 + 0,950 X_3 + 0,056 X_4 = 0,493 Y_1 + 0,213 Y_2 \quad r_2 = 0,512$$

Vypočtete skupinový korelační koeficient a interpretujte výsledky.

Skupinový korelační koeficient se vypočte ze vztahu:

$$R = [1 - (1 - r_1^2)(1 - r_2^2)]^{1/2}$$

$$R = [1 - (1 - 0,830^2)(1 - 0,512^2)]^{1/2} = 0,878$$

$$R^2 = 0,770, \text{ což znamená, že výsledek popisuje 77 \% variability dat}$$

Parametr X_3 má vzhledem k ostatním parametrům zanedbatelný vliv na součet parametrů Y_1 a Y_2 , ostatní parametry mají vliv zhruba stejný. Na rozdíl parametrů Y_1 a Y_2 má parametr X_3 výrazný vliv, zatímco vliv ostatních parametrů je nevýrazný.

Příklad 7

Uved'te nějaký konkrétní případ vhodný pro zpracování metodou PLS.

Metoda projekce latentních struktur PLS (partial least square) je zobecněným postupem pro získání popisu vztahu mezi závislým náhodným vektorem y a vysvětlujícím náhodným vektorem x prostřednictvím latentních proměnných. Tímto postupem je možné opakováním kroků až do konvergence získat vektor latentních proměnných a vektor zátěží a nakonec i z residuální matice i matici latentních proměnných a zátěží. Jako praktickou aplikaci lze uvést vliv obsahu prvkových příměsí na optické vlastnosti pigmentového oxidu titaničitého. Souřadnice L^* vyjadřuje jas pigmentu, souřadnice b^* jeho žlutost.

Obsah příměsí	souřadnice L^*	souřadnice b^*	barvivost	podtón
železo	*****	*****	*****	*****
chrom	*****	*****	*****	*****
niob	*****	*****	*****	*****
draslík	*****	*****	*****	*****
fosfor	*****	*****	*****	*****
zirkonium	*****	*****	*****	*****

Příklad 8

Jeden objekt je charakterizován metrickými znaky (2,10), druhý (3,8), třetí (4,9), čtvrtý (10,4) a pátý (11,5). Vypočtete matici vzdálenosti v Euklidově metrice a dokumentujte výpočet shlukování některou z používaných metod. Výsledky interpretujte graficky.

Matice vzdálenosti znaků má tvar:

1	0,00				
2	2,24	0,00			
3	2,24	1,41	0,00		
4	10,00	9,75	7,81	0,00	
5	10,30	8,54	8,06	1,41	0,00

Nejmenší vzdálenost mají prvky 3-2 a 5-4. Z nich je možné vytvořit první shluky, spočítat těžiště nových shluků a spočítat těžiště nových shluků a opět vypočítat matici vzdáleností:

1	0,00	
2-3	2,12	0,00
4-5	10,12	8,06

Nejmenší vzdálenost má shluk 2-3 a znak 1. Je tedy možné tyto prvky spojit do shluku a spočítat těžiště nového shluku a opět vypočítat matici vzdáleností:

1-2-3 0,00
4-5 8,75 0,00

Tento proces je možné shrnout do následujícího dendogramu:

