Fakulta chemicko technologická
Katedra analytické chemie
licenční studium Management systému jakosti

Metody s latentními proměnnými

Autor: Ing. Radek Růčka

Přednášející: Prof. Ing. Oldřich Pytela, DrS

Otázka 1.

Vypočtěte algoritmem NIPALS 1. latentní proměnnou z matice A[řádek,sloupec]: A[1,1]=1, A[2,1]=2, A[3,1]=3, A[1,2]=1, A[2,2]=2, A[3,2]=0, A[1,3]=6, A[2,3]=4, A[3,3]=2. Matici před zpracováním standardizujte.

Výpočet vektoru aritmetických průměrů $\overline{x}^T$:

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

Výpočet vektoru směrodatných odchylek $\overline{s}^T$:

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

Provedením standardizace se získá matice A:

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

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

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

Po dosazení do vzorce:

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

a znormování podle vztahu:

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

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

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

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

$$d = (t_{nové} - t_{staré})^T \cdot (t_{nové} - t_{staré}) \cdot (t^T{}_{nové} \cdot t_{nové})^{-1}$$

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

K dosažení této podmínky je potřeba 9-ti kroků. Získá se následující stabilní rozklad vektorů:

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

Otázka 2.
S použitím vhodných kriterií 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.

V tomto případě se nezbytný počet latentních proměnných určí pomocí Waldem navrženého kritéria, které je definováno poměrem:

$$\frac{PRESS(P)}{S_R(P-1)}$$

Výpočtem dostaneme pro P=1…… $\frac{PRESS(3)}{S_R(2)} = \frac{20}{100} = 0,20$

P=2…… $\frac{PRESS(2)}{S_R(1)} = \frac{35}{10} = 0,35$

P=3…… $\frac{PRESS(3)}{S_R(2)} = \frac{3,45}{3,4} = 1,01$

Pokud je hodnota Waldova kritéria vyšší než 0,95 je zahrnutí další (P+1) ní proměnné nevhodné.
V našem případě (P=3) je hodnota vyšší než 0,95, proto čtvrtá latentní proměnné již není významná.

Otázka 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:

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

Pro odhad hodnoty prvku A[2, 2] takto platí:

A[2, 2] = t[2] p[2] = -0,311

Otázka 4.
Výpočtem metodou PCA byly určeny vektory p1: 0.012, 0.458, -0.352, 0.987 a p2: 0.926, -0.238, 0.872, -0.115 Vypočtěte komunality a vyberte sloupec, který nejlépe charakterizuje 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, tj. $p_2^T[1] = 0{,}926$ a sloupec třetí $p_2^T[3] = 0{,}872$.

Otázka 5.
Vysvětlete, proč vysvětlená variabilita je při výpočtu metodou FA vždy nižší, než při výpočtu metodou PCA.

Metoda PCA je považována z hlediska faktorové analýzy za úplnou komponentní analýzu. Je to z důvodu, že prostřednictvím přesně vypočítaných hlavních komponent lze přesně reprodukovat variabilitu zdrojové matice.
Metoda FA je považována za neúplnou komponentní analýzu, s modelem, u kterého nevysvětlená část variability pouze hrubě aproximuje definované rozptyly a kovariace a není odhadem jedinečnosti. K reprodukci zdrojové matice tedy slouží P hlavních komponent, které reprodukují podstatnou část variability manifestních proměnných.

Otázka 6.
Výpočtem metodou kanonických korelací bylo zjištěno: 0.297 X1 + 0.298 X2 + 0.050 X3 + 0.256 X4 = 0.493 Y1 - 0.213 Y2, r1 = 0.830
0.006 X1 - 0.115 X2 + 0.950 X3 + 0.056 X4 = 0.493 Y1 + 0.213 Y2, r1 = 0.512. Vypočtěte skupinový korelační koeficient a interpretujte výsledky.

Skupinový korelační koeficient se kanonickou korelací vypočte ze vztahu:

$$R_{XY} = \left[1 - \frac{|C|}{|C_{XX}||C_{YY}|}\right]^{1/2}$$

kde C.....celková kovariační matice
C_{XX}.....kovariační matice náhodného vektoru x
C_{YY}.....kovariační matice náhodného vektoru y

$$R_{XY} = \left[1 - (1 - r_1^2)(1 - r_2^2)\right]^{1/2}$$

$$R_{XY} = \left[1 - (1 - 0{,}830^2)(1 - 0{,}512^2)\right]^{1/2} = 0{,}878$$

$$R^2 = 0,770$$

Hodnota skupinového korelačního koeficientu je rovna 0,770, znamená to, 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 aY_2 má parametr X_3 výrazný vliv, zatímco vliv ostatních parametrů je nevýrazný.

Otázka 7.

Uveďte nějaký konkrétní příklad vhodný pro zpracování metodou PLS.

Při anorganické povrchové úpravě se pro jednotlivé rutilové druhy používají postupy povrchové úpravy, lišící se druhem požívaných činidel, velikostí povrchové úpravy a průběhem srážení. Všechny tyto faktory ovlivňují vlastnosti konečného produktu dané charakterem povrchu jeho částic jako je měrný povrch, spotřeba oleje, spotřeba vody, dispergovatelnost.

typ TB/parametr	měrný povrch	spotřeba oleje	spotřeba vody	dispergace
RGU	xxxxxxxx	xxxxxxxx	xxxxxxxx	xxxxxxxx
RG18	xxxxxxxx	xxxxxxxx	xxxxxxxx	xxxxxxxx
RGZW	xxxxxxxx	xxxxxxxx	xxxxxxxx	xxxxxxxx
RGX	xxxxxxxx	xxxxxxxx	xxxxxxxx	xxxxxxxx
RXI	xxxxxxxx	xxxxxxxx	xxxxxxxx	xxxxxxxx

Otázka 8.

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

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

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

Nejmenší vzdálenost má shluk 2-3 a znak 1. Je tedy možné spojit tyto prvky 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 lze shrnout do tohoto dendogramu: