Univerzita Pardubice
FAKULTA CHEMICKO TECHNOLOGICKÁ

METODY S LATENTNÍMI PROMĚNNÝMI A KLASIFIKAČNÍ METODY

SEMINÁRNÍ PRÁCE LICENČNÍHO STUDIA
Statistické zpracování dat při kontrole jakosti

Ing. Karel Drápela, CSc.

Brno 2001

Příklad 1:

Vypočtěte algoritmem NIPALS 1. latentní proměnnou z matice $A = \begin{vmatrix} 1 & 1 & 6 \\ 2 & 2 & 4 \\ 3 & 0 & 2 \end{vmatrix}$.

Nejprve vypočteme vektor aritmetických průměrů $x^T = |2 \quad 1 \quad 4|$

a vektor směrodatných odchylek $s^T = |1 \quad 1 \quad 2|$.

Standardizací obdržíme matici $A = \begin{vmatrix} -1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & -1 & -1 \end{vmatrix}$

Protože všechny sloupce mají stejnou variabilitu, vezmeme za první odhad hlavní komponenty t_1 např. první sloupec:

$$t_1^T = |-1 \quad 0 \quad 1|$$

Po dosazení do vztahu

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

a znormováním podle

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

získáme počáteční odhad vektru míry příspěvků odhadu vektoru hlavní komponenty t_1. Dosazením pak do vztahu

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

získáme odhad hlavní komponenty t_1. Iteračním opakováním tohoto postupu dostaneme stabilní rozklad vektorů t_1 a p_1. Konvergenční kritérium má tvar

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

Výpočet se ukončí, je-li $d/N < 10^{-10}$.
Jednotlivé kroky udává tabulka 1.

Tabulka č. 1 - Vypočítané vektory latentních proměnných a zátěží

Číslo opakování	Vektor zátěží (p1)	Normovaný vektor zátěží	Vektor hlavní komponenty
1	1 -0,5 -1	0,6667 -0,3333 -0,6667	-1,3333 -0,3334 1,6667
2	0,6750 -0,4500 -0,6750	0,6396 -0,4264 -0,6396	-1,2792 -0,4264 1,7056
3	0,6314 -0,4510 -0,6314	0,6312 -0,4508 -0,6312	-1,2623 -0,4508 1,7132
4	0,6289 -0,4574 -0,6289	0,6288 -0,4573 -0,6288	-1,2577 -0,4573 1,7150
5	0,6282 -0,4591 -0,6282	0,6282 -0,4591 -0,6282	-1,2564 -0,4591 1,7154
6	0,6280 -0,4595 -0,6280	0,6280 -0,4595 -0,6280	-1,2561 -0,4596 1,7157
7	0,6279 -0,4596 -0,6279	0,6280 -0,4597 -0,6280	-1,2559 -0,4597 1,7156
8	0,6280 -0,4597 -0,6280	0,6280 -0,4597 -0,6280	-1,2559 -0,4597 1,7156

Metodou NIPALS jsme získali po 8. opakování hodnoty shodné s údaji po 7. opakování (viz tabulka). Určili jsme tedy stabilní vektory (vektor první latentní proměnné i vektor zátěže). První latentní proměnná má následující hodnotu:

$$t_1^T = \| -1{,}2559 \quad 0{,}4597 \quad 1{,}7156 \|$$

Vektor zátěže této latentní proměnné je:

$$p_1^T = \| 0{,}6280 \quad -0{,}4597 \quad -0{,}6280 \|$$

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) = 3.4, PRESS (3) = 3.45, S (3) = 3.39.

K určení nezbytného počtu latentních proměnných použijeme Woldovo kritérium založené na poměru hodnoty PRESS (P) a hodnoty S_R (P-1)

PRESS (P) / S_R (P-1).

Zařazení další (P+1) latentní proměnné je nevhodné, je-li hodnota kritéria větší než 0.95.

P = 1: PRESS (1) / S_R (0) = 0.2
P = 2: PRESS (2) / S_R (1) = 0.35
P = 3: PRESS (3) / S_R (2) = 1.01

Protože pro P = 3 je hodnota kritéria větší než 0.95, není čtvrtá latentní proměnná již významná.

Příklad 3:
Odhadněte hodnotu chybějícího prvku A [2,2], jestliže výpočtem z nekompletní matice byly určeny vektory p^T = *0.541 0.423 0.514 0.514 a* t^T = *-1.340 -0.735 2.076* .

Prvky zdrojové matice odpovídající p-té latentní proměnné lze rekonstruovat podle vzorce

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

Odtud odhad hodnoty prvku A [2,2] je

$$\begin{Vmatrix} -1{,}340 \\ -0{,}735 \\ 2{,}076 \end{Vmatrix} \cdot \| \; 0{,}541 \quad 0{,}423 \quad 0{,}514 \quad 0{,}514 \; \| = \begin{Vmatrix} -0{,}725 & -0{,}567 & -0{,}589 & -0{,}589 \\ -0{,}398 & \mathbf{-0{,}311} & -0{,}378 & -0{,}378 \\ 1{,}123 & 0{,}878 & 1{,}067 & 1{,}067 \end{Vmatrix}$$

Pomocí „krátkého cyklu“ byla stanovena hodnota prvku A [2,2] = -0,311.

Příklad 4:
Výpočtem metodou PCA byly určeny vektory p_1^T = *0.012 0.458 -0.352 0.987 a* p_2^T *0.926 -0.238 0.872 -0.115 . Vyberte reprezentativní sloupce charakterizující nejlépe zdrojovou matici.*

Vektory p_1^T a p_2^T představují míru příspěvku příslušné latentní proměnné k popisu variability sloupců zdrojové matice. Vektor p_1^T nám říká, že reprezentativním sloupcem zdrojové matice je sloupec čtvrtý p_1^T [illegible] = 0.987 a vektor p_2^T, že reprezentativním sloupcem je sloupec první p_2^T [illegible] = 0.926 a třetí p_2^T [illegible] = 0.872 .

Příklad 5:

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

Je to způsobeno tím, že z hlediska faktorové analýzy je metoda PCA považována za úplnou komponentní analýzu (neboť pomocí hlavních komponent lze přesně reprodukovat variabilitu zdrojové matice) a metoda FA za neúplnou komponentní analýzu (neboť připouští existenci matice jedinečností, která představuje část variability nevysvětlitelné 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 $r_1 = 0.830$
0.006 X_1 *- 0.115* X_2 *+ 0.950* X_3 *+ 0.056* X_4 *= 0.493* Y_1 *+ 0.213* Y_2 *r2 = 0.512*
Vypočtěte skupinový korelační koeficient a interpretujte výsledky.

Skupinový korelační koeficient R vypočteme ze vztahu

$$R^2 = 1 - (1 - r_1^2)(1 - r_2^2)$$

tj.

$$R^2 = 0.770, \quad R = 0.878$$

tj. výsledek popisuje 77 % variability dat.
Z první rovnice vidíme, že na parametry Y_1 a Y_2 má parametr X_3 zanedbatelný vliv vzhledem k ostatním parametrům, ostatní parametry mají vliv stejný.
Z druhé rovnice vidíme, že na parametry Y_1 a Y_2 má výrazný vliv parametr, vliv ostatních parametrů je nevýrazný. Znaménka mají stejný smysl jako v „klasické“ regresi.

Příklad 7:
Uveďte typický konkrétní příklad vhodný pro zpracování metodou PLS.

1) Při rozlišování příbuznosti jedlí byly použity morfologické znaky na semenech (váha, tvar a rozměry semene, počet semen v šišce, tvar šišky) a chemické vlastnosti silic v semenech (jejich procentuální zastoupení)

Bylo zjištěno, že příbuznost mezi jedlemi lze sledovat nejen morfologickými znaky, ale i chemicky a byla zde nalezena závislost.

2) Při výzkumu mechanických vlastností dřeva byly také zkoumány fyzikální vlastnosti dřeva v rámci letokruhu (šířka letního a jarního dřev, jejich hustota, síla buněčných stěn). Tyto fyzikální vlastnosti dřeva závisí mimo jiné na klimatických faktorech. Metodou PLS je tedy možné zjistit závislost mezi klimatickými charakteristikami (teploty, srážky, sluneční svit, apod.) a fyzikálními vlastnostmi dřeva letokruhů.

Příklad 8:
Jeden objekt je charakterizován metrckými znaky [A](2,10), druhý [B](3,8), třetí [C] (4,9), čtvrtý [D](10,4) a pátý [E](11,5). Vypočtěte matici vzdáleností v Euklidově metrice a proveďte shlukování metodou průměrné vazby. Výsledky interpretujte graficky.

Matice vzdáleností znaků má tvar

A	0				
B	2.24	0			
C	2.24	1.41	0		
D	10	8.06	7.81	0	
E	10.3	8.94	8.06	1.41	0
	A	B	C	D	E

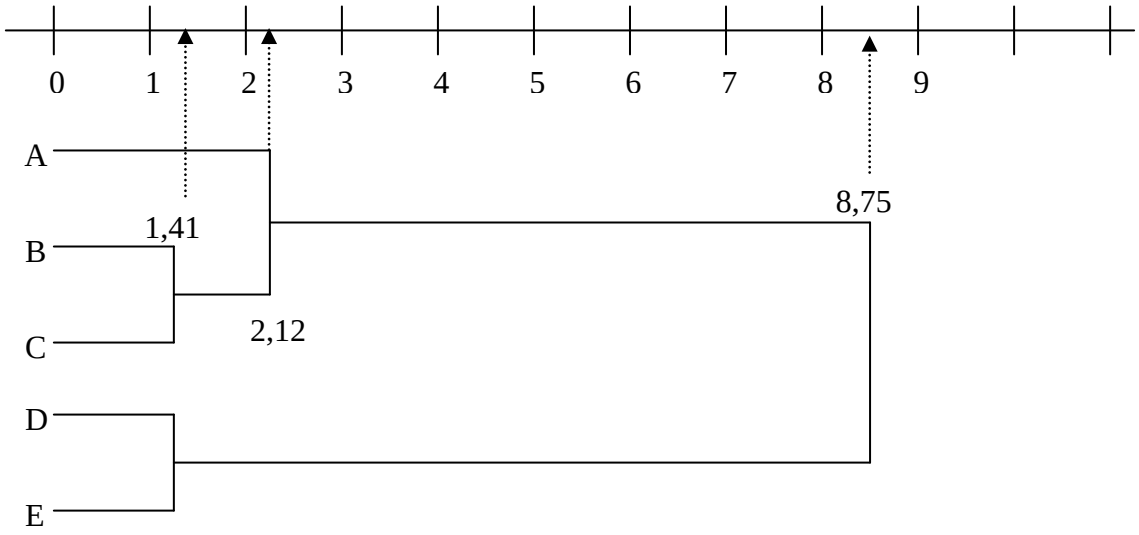
Nejmenší vzdálenost mají prvky B-C a D-E. Vytvořme tak první shluky, spočteme těžiště nových shluků a vypočteme matici vzdáleností:

A	0		
B-C	2.12	0	
D-E	10.12	8.06	0

Zde má nejmenší vzdálenost shluky B-C a A. Vytvoříme z nich shluk, spočteme těžiště nového shluku a vypočteme matici vzdáleností:

A-B-C	0	
D-E	8.75	0

Výsledný dendrogram je uveden na obrázku 1. Vyplývá z něho, že prvně se sloučí body C a B a D a E na vzdálenosti 1,41, poté se ke shluku BC přidá bod A na vzdálenosti 2,12 a vytvoří se výsledné shluky ABC a DE. Ty si jsou značně nepodobné, k jejich sloučení dojde až na vzdálenosti 8.75.



Obrázek 1 – Dendrogram shlukování metodou průměrné vazby

Příklad 9

Toto je pouze názor nechemika, který v životě v chemické výrobě nebyl:
Vzhledem k tomu, že kontrola kvality v chemické výrobě je značně náročná a drahá, bylo by možné prozkoumat možnost redukce kontrol kvality pomocí vícerozměrné analýzy, která by ukázala, zda některé zkoušky netestují podobné vlastnosti, tedy metodou PCA se pokusit snížit počet sledovaných proměnných a tím i prováděných testů kvality. Je nutné sledovat optimální poměr komponent, protože to má vliv na kvalitu výsledného produktu. Dále je možné prozkoumat, zda nejsou rozdíly v kvalitě výroby mezi směnami, v případě, že ano, zjistit příčiny.