SIEMENS VDO

UNIVERZITA PARDUBICE

Fakulta chemicko-technologická
Katedra analytické chemie

3.2 Metody s latentními proměnnými a klasifikační metody

Vypracoval: Ing. Tomáš Nekola

Studium: licenční

Datum: 21. 1. 2008

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.

Řešení :

Zdrojová matice: A = Vektor aritmetických průměrů: X^T =

Vektor směrodatných odchylek: s^T =

Standardizace matice: A =

Jako první odhad vektoru t_1 byl zvolen první sloupec standardizované matice :

t_1^T =

Dosazením do vztahu: $p_1^T = (t_1^T t_1)^{-1}\ t_1^T$ A a znormováním podle $p_1^N = (p_1^T p_1)^{-1/2}\ p_1$ získáme počáteční odhad vektoru míry příspěvků odhadu vektoru hlavní komponenty t_1. Dosazením za $t_1^T = (p_1^T p_1)^{-1}\ p_1^T\ A^T$ získáme odhad hlavní komponenty t_1. Iteračním opakováním tohoto postupu dostaneme stabilní rozklad vektorů t_1 a p_1.

Konverg. kritérium má tvar : $d = (\ t_{nové} - t_{staré}\ )^T\ (\ t_{nové} - t_{staré}\ )(\ t^T_{nové} - t_{nové}\ )^{-1}$

Výpočet se ukončí když $d / N < 10^{-10}$

Výpočet vektoru p_1:

$t_1^T t_1$ = = =

p_1^T =

Normování vektoru p_1:

$(p_1^T p_1)^{-1/2}$ = = =0,667 0,667

p_1^N =

Výpočet t_1:

$p_1^T p_1 =$

$t_1^T =$

Výpočet se opakuje. Jednotlivé vypočítané vektory jsou uvedeny v tabulce č.1.

Tabulka č. 1 - Vypočítané vektory latentních proměnných a zátěží.

Číslo opakování	Vektor zátěží (p_1)	Normovaný vektor zátěží (p_1N)	Vektor latentních hodnot
1			
2			
3			
4			
5			
6			
7			
8			

Po 8 krocích obdržíme zátěžový vektor $p_1^T =$

vektor latentních proměnných $t_1^T =$

Pronásobením a odečtením dle $E_1 = A - t_1 p_1^T$ získáme matici nevysvětlené variability

$E =$

Závěr:

Algoritmem NIPALS byla spočtena první latentní proměnná.

Otázka 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.

Řešení : Výpočtem se má stanovit nejmenší signifikantní počet latentních proměnných, které popisují variabilitu zdrojové matice bez zahrnutí experimentální chyby. Vzhledem k znalosti hodnot PRESS(P) a S(P) můžeme pro výpočet použít test navržený Woldem. Zařazení další (P+1) latentní proměnné je nevhodné, je-li hodnota podílu
PRESS(P)/S(P-1) větší než 0,95.

P = 1 : PRESS(1)/S(0) = 0,2

P = 2 : PRESS(2)/S(1) = 0,35

P = 3 : PRESS(3)/S(2) = 1,01 (1,01>0,95)

Závěr:

Woldovým kritériem byl stanoven počet latentních proměnných na základě výše uvedených hodnot na 3. Čtvrtá latentní proměnná není třeba, protože hodnota kritéria vyšla větší (1,01) než 0,95.

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

Řešení :

Na rekonstrukci zdrojové matice X se používají první latentní proměnné popisující podstatnou část variability zdrojové matice bez zahrnutí experimentální chyby. Tato metoda se nazývá Krátký cyklus.

Prvky zdrojové matice odpovídající p-té latentní proměnné lze rekonstruovat podle vzorce $A_p^{pred} = t_p\, p_p^T$

Metodou Krátký cyklus byla stanovena hodnota chybějícího prvku na pozici

A[2,2] = -0,311

Otázka 4.

Výpočtem metodou PCA byly určeny vektory p_1: 0,012 0,458 -0,352 0,987 p_2: 0,926 -0,238 0,872 -0,115.

Vypočtěte komunality a vyberte sloupec, který nejlépe charakterizuje celou matici.

Řešení :

Prvek matice zátěží pip příslušejícímu i-tému sloupci zdrojové matice je mírou variability tohoto sloupce popsané p-tou latentní proměnnou. Podíl variability daného sloupce popsané společnými latentními proměnnými lze pakvyjádřit jako součet příspěvků jednotlivých latentních proměnných, tedy

kde h_i^2 je komunalita pro i-tý sloupec

P je počet latentních proměnných

M je počet sloupců zdrojové matice

p_{ip}^N jsou zátěže normované podle vztahu: ———

Výpočet normovaných zátěží

vektor p_1: 0,012 0,458 -0,352 0,987

vektor p_2: 0,926 -0,238 0,872 -0,115

Vypočítáme normované zátěže pro jednotlivé sloupce při dvou latentních proměnných.

——— = ——————————— ———

——— = ——————————— ———

——— = ——————————— ———

——— = ——————————— ———

——— = ——————————— ———

——— = ——————————— ———

——— = ——————————— ———

_____ = _____ _____

Výpočet komunalit

= =

= =

= =

Závěr:

Platí pravidlo, že čím je komunalita příslušného sloupce větší, tím má sloupec vlastnosti společné s ostatními sloupci zdrojové matice. V našem případě čtvrtý sloupec nejlépe charakterizuje zdrojovou matici.

Otázka 5.

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

Řešení :

Rozdíl spočívá ve způsobu rozkladu variability zdrojové matice.

V případě **metody PCA** je model postaven na S2 = H2, (S2 - diagonální matice rozptylů manifestních proměnných, H2 - diagonální matice komunalit), který předpokládá, že lze variabilitu zdrojové matice reprodukovat přesně .

U **metody FA** platí, že variabilita je rozložena do dvou složek S2 = H2 + L2, (L2 - matice jedinečností, která představuje část variability nevysvětlitelné společnými faktory). Tím, že u metody FA předpokládáme existence nevysvětlené variability dospíváme vždy k nižším hodnotám vysvětlené variability než u metody PCA.

Otázka 6.

Výpočtem metodou kanonických korelací bylo zjištěno:

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

$0{,}006X_1 - 0{,}115X_2 + 0{,}950X_3 + 0{,}056X_4 = 0{,}493Y_1 + 0{,}213Y_2$ $r_1 = 0{,}512$

Vypočtěte skupinový korelační koeficient a interpretujte výsledky.

Řešení :

Skupinový korelační koeficient R vypočteme ze vztahu

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

$R^2 = 1 - (1 - 0{,}830^2)(1 - 0{,}512^2) = 0{,}7704$

$R^2 = 0{,}7704$

Závěr:

Matice X je tvořena 4 sloupci a matice Y dvěmi. Skupinový korelační koeficient má hodnotu 0,7704, což znamená, že 77 % variability dat jsme vysvětlili kanonickými korelačním koeficienty. S růstem proměnných X_1, X_2 a X_4, které mají stejný vliv, klesá význam proměnné Y_2 a roste významně proměnná Y_1. S nadbytkem proměnné X_3 se zvyšuje proměnná Y_2.

Otázka 7.

Uveďte nějaký konkrétní příklad vhodný pro zpracování metodou PLS.

Metoda projekce latentních struktur PLS 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 vector latentních proměnných a vektor zátěží a nakonec i z residuální matice i matici latentních proměnných a zátěží.

Praktickým příkladem může být sledování chyb na tester IPTE v čase u různých typů BMW radii:

Chyby na IPTE

	BMW RCD 114	BMW RCD 121	BMW RCD 114	BMW RCD 121	BMW RCD 114	BMW RCD 121
	REFERENCE VAL. FM FL	REFERENCE VAL. FM FL	PROCESS DATA START	PROCESS DATA START	DIFF. F L/R	DIFF. F L/R
40/2007	0	3	1	2	1	0
41/2007	1	1	7	1	0	1
42/2007	0	2	1	0	1	3
43/2007	1	3	1	0	0	0
44/2007	2	4	3	2	0	0
45/2007	3	2	1	2	0	1
46/2007	1	6	5	2	0	2
47/2007	1	2	3	1	1	0
48/2007	2	0	0	0	0	0
49/2007	3	31	2	0	0	13
50/2007	3	5	1	0	3	0
51/2007	2	2	5	0	0	3
52/2007	0	0	0	0	0	0
1/2008	53	2	1	0	0	0
2/2008	3	4	22	0	15	1
3/2008	1	14	7	1	0	15

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 vzdálenosti v Euklidově metrice a dokumentujte výpočet shlukování některou z používaných metod. Výsledky interpretujte graficky.

Řešení :

Postupy aglomerativní shlukové analýzy lze neuspořádanou množinu objektů charakterizovaných náhodnými vektory postupně uspořádat do tříd až ke konečnému stavu, kdy se všechny objekty spojí do jediné třídy. Výše uvedené vlastnosti objektů (blízkost či podobnost) posuzujeme podle vzdálenosti objektů v p-rozměrném prostoru znaků nejjednodušším typem Euklidové metriky, která je definovaná vztahem $d_E(X_k, X_l) =$, kde d_E je vzdálenost mezi objekty X_k a X_l, jsou souřadnice objektů v P-rozměrném prostoru.

Vzdálenost určená podle zásady nejbližšího souseda (nejbližší jsou ty třídy, které mají nejmenší vzdálenost mezi dvěma nejbližšími objekty (objekt můžeme chápat jako třídu)) se vypočte podle vzorce $D_{NN} = (S_r, S_s) = \min. d(X_k, X_l)$.

SIEMENS VDO

Matice vzdáleností

Hierarchical Cluster Analysis

Complete Linkage Euclidean Distance

Amalgamation Steps

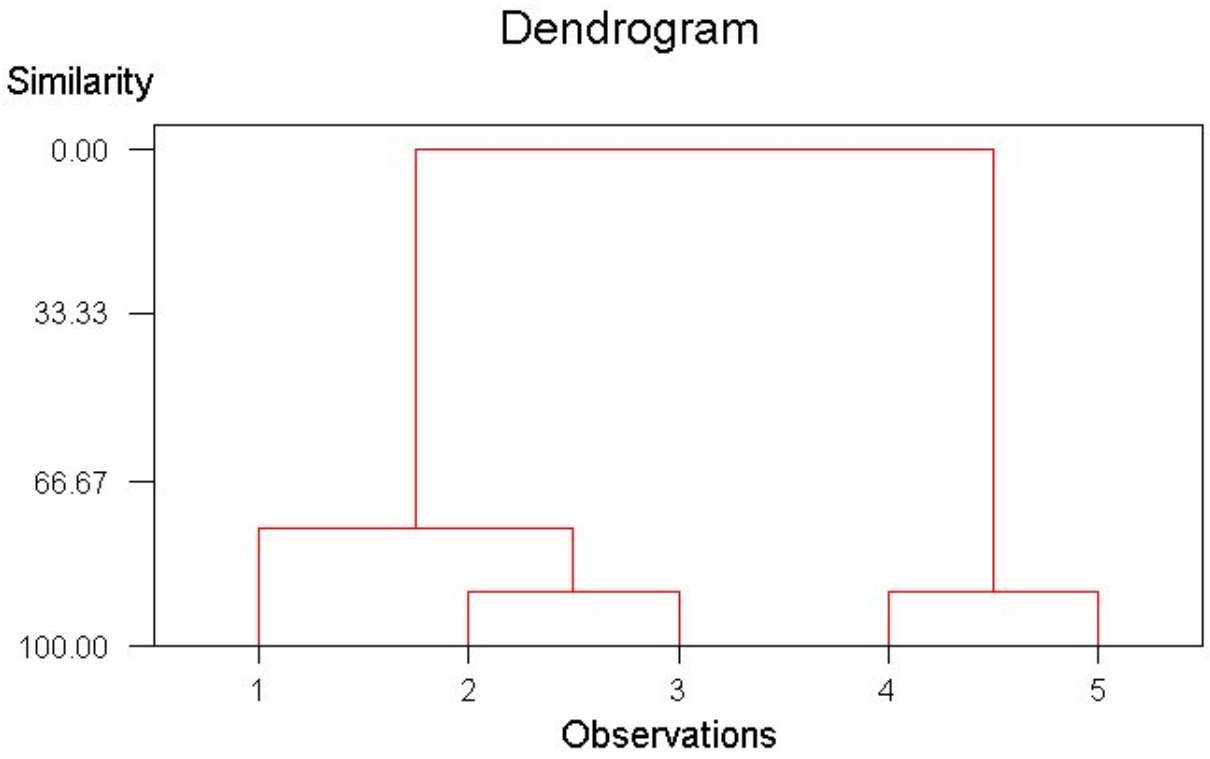
Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	4	89.07	2.072	2	3	2	2
2	3	88.96	2.093	4	5	4	2
3	2	76.14	4.523	1	2	1	3
4	1	0.00	18.960	1	4	1	5

Final Partition

Number of clusters: 1

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	5	377.749	8.512	10.697

Grafické znázornění jednotlivých objektů



Z grafického pohledu vyplývá, že objekty postupem nejbližšího souseda vytvářejí nejdříve dva shluky a to 2-3 a 4-5 dalším postupem je objekt č.1 připojen ke shluku 2-3 a následně tento shluk vytvoří jeden shluk se shlukem 4-5.

Závěr:

Byly stanoveny vzdálenosti mezi jednotlivými objekty formou matice vzdálenosti a metodou shlukové analýzy klasifikovány objekty postupem nejbližšího souseda.

Otázka 9.

Popište slovně postup aplikace metod s latentními proměnnými nebo klasifikačních metod na nějakém konkrétním příkladu ze své praxe.

Postup při analýze dat :

Praktickým příkladem může být sledování chyb na tester IPTE v čase u různých typů BMW radii (viz Otázka 7.):

1. Nejdříve provedeme průzkumovou analýzu dat. Tím odstraníme případné chyby.
2. Analýza korelační matice a matice parciálních korelačních koeficientů mezi vlastnostmi.
3. Zvolíme metodu. V našem případě zvolíme faktorovou analýzu a metodu hlavních komponent.
4. Data standardizujeme, protože nejsou vyjádřena ve stejných jednotkách.
5. Vypočteme latentní proměnné
6. Určíme počet signifikantních latentních proměnných.
7. Analyzujeme matici zátěží a matici latentních proměnných.
8. Analyzujeme fyzikálně-chemický smysl latentních proměnných.
9. Interpretujeme výsledky vzhledem k cíli analýzy.