Univerzita Pardubice
Chemicko-technologická fakulta
Katedra analytické chemie

12. licenční studium PYTHAGORAS
Statistické zpracování dat

3.2 Metody s latentními proměnnými a klasifikační metody

Semestrální práce
2010

RNDr. Markéta Vaňková, Ph.D.
Endokrinologický ústav
Národní 8, 116 94 Praha 1

Otázka 1

Vypočtěte algoritmem NIPALS první 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:

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

1. Standardizace zdrojové matice

Vektor aritmetických průměrů sloupců matice

$X^{-T} = (2,1,4)$

Vektor směrodatných odchylek

$S^T = (1,1,2)$

Standardizovaná matice X_S

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

2. Určení první latentní proměnné

Zvolí se vektor matice T(t1) a vypočte se vektor zátěží (p1)

$P_1^T = (t_1^T t_1)^{-1}\ t_1^T\ X$

Vektor p_1 se normuje podle vztahu

$P_1^N = (p_1^T p_1)^{-1/2}\ p_1$

Normovaný vektor se dosadí do následujícího vztahu a vypočte se stabilnější vektor t_1

$T_1^T = (p_1^T p_1)^{-1}\ p_1^T\ X^T$

Postup se opakuje až do získání stabilní hodnoty vektorů. To je v momentě, kdy je podíl $d/N < 10^{-10}$.

N je počet normování

d je konvergenční kritérium vypočítané ze vztahu

$d = (t_{nové} - t_{staré})^T (t_{nové} - t_{staré})(t^T_{nové}\, t_{nové})^{-1}$).

$p_1^T = (2/3 \;\; -1/3 \;\; -2/3)$

$t_1^T = (-4/3 \;\; 1/3 \;\; 5/3)$

dále

$p_1^T = (0{,}6396 \;\; -0{,}4264 \;\; -0{,}6396)$

$t_1^T = (-1{,}2792 \;\; -0{,}4264 \;\; 1{,}7056)$

$d = 0{,}00277$

po splnění konvergenčního kritéria získáme tyto vektory

první latentní proměnná má následující hodnotu

$t_1^T = (-1{,}2559 \;\; -0{,}4569 \;\; 1{,}7156)$

vektor zátěže této latentní proměnné je

$p_1^T = (0{,}6279 \;\; -0{,}4596 \;\; -0{,}6279)$

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:

Cílem je určení co nejmenšího počtu latentních proměnných, které popisují variabilitu zdrojové matice bez zahrnutí experimentální chyby. Vzhledem ke znalosti hodnot PRESS(P) a S(P-1), lze použít test navržený Woldem. Je-li PRESS(P)/S(P-1) > 0,95, tak další latentní proměnnou již vylučujeme.

PRESS(1)/S)0) = 20/100 = 0,2
PRESS(2)/S(1) = 3,5/10 = 0,35
PRESS(3)/S(2) = 63,45/3,4 = 1,015

Hodnota třetího podílu je vyšší než 0,95, čtvrtá latentní proměnná je tedy již nevýznamná. Nejmenší signifikantní počet proměnných je 3.

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 použije první latentní proměnná popisující podstatnou část variability zdrojové matice. Tato metoda se nazývá "krátký cyklus". Rekonstrukce zdrojové matice je popsána vztahem

$X_p^{pred} = t_p\, p_p^T$.

po dosazení

$$\begin{vmatrix} -1{,}340 \\ -0{,}735 \\ 2{,}076 \end{vmatrix} * \begin{vmatrix} 0{,}541 & 0{,}423 & 0{,}514 & 0{,}514 \end{vmatrix} = \begin{vmatrix} -0{,}725 & -0{,}567 & -0{,}589 & -0{,}589 \\ -0{,}398 & -0{,}311 & -0{,}378 & -0{,}378 \\ 1{,}398 & 0{,}878 & 1{,}067 & 1{,}067 \end{vmatrix}$$

Metodou „krátký cyklus" byla stanovena hodnota 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 p_{ip} 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 vyjádřit jako součet příspěvků jednotlivých latentních proměnných

$$h_i^2 = p^N (p^N)^T = \sum_{p=1}^{P} (p_{ip}^N)^2$$

kde h_i^2 …komunalista pro i-tý sloupec

P …..počet latentních proměnných

M …..počet sloupců zdrojové matice

p_{ip}^N jsou zátěže normované podle vztahu $p_{ip}^N = \dfrac{p_{ip}}{\sum_{i=1}^{M} p_{ip}^2}$

Výpočet normovaných zátěží

Vektory

p_1 = (0,012 0,458 -0,352 0,987)

p_2 = (0,926 -0,238 0,872 -0,115)

M = 4, P = 2

Normované zátěže pro jednotlivé sloupce při dvou latentních proměnných.

$$p_{11}^N = \frac{p_{11}}{\sum_{i=1}^{M} p_{i1}^2} = \frac{0{,}012}{0{,}012^2 + 0{,}458^2 + (-0{,}352)^2 + 0{,}987^2} = \frac{0{,}012}{1{,}308} = 0{,}0092$$

$$p_{21}^N = \frac{p_{21}}{\sum_{i=1}^{M} p_{i1}^2} = \frac{0{,}458}{0{,}012^2 + 0{,}458^2 + (-0{,}352)^2 + 0{,}987^2} = \frac{0{,}458}{1{,}308} = 0{,}3502$$

$$p_{31}^N = \frac{p_{31}}{\sum_{i=1}^{M} p_{i1}^2} = \frac{-0{,}352}{0{,}012^2 + 0{,}458^2 + (-0{,}352)^2 + 0{,}987^2} = \frac{-0{,}352}{1{,}308} = -0{,}2691$$

$$p_{41}^N = \frac{p_{41}}{\sum_{i=1}^{M} p_{i1}^2} = \frac{0{,}987}{0{,}012^2 + 0{,}458^2 + (-0{,}352)^2 + 0{,}987^2} = \frac{0{,}987}{1{,}308} = 0{,}7546$$

$$p_{12}^N = \frac{p_{12}}{\sum_{i=1}^{M} p_{i2}^2} = \frac{0{,}926}{0{,}926^2 + (-0{,}238)^2 + 0{,}872^2 + (-0{,}115)^2} = \frac{0{,}926}{1{,}6877} = 0{,}5487$$

$$p_{22}^N = \frac{p_{22}}{\sum_{i=1}^{M} p_{i2}^2} = \frac{-0{,}238}{0{,}926^2 + (-0{,}238)^2 + 0{,}872^2 + (-0{,}115)^2} = \frac{-0{,}238}{1{,}6877} = -0{,}141$$

$$p_{32}^N = \frac{p_{32}}{\sum_{i=1}^{M} p_{i2}^2} = \frac{0{,}872}{0{,}926^2 + (-0{,}238)^2 + 0{,}872^2 + (-0{,}115)^2} = \frac{0{,}872}{1{,}6877} = 0{,}5167$$

$$p_{42}^N = \frac{p_{42}}{\sum_{i=1}^{M} p_{i2}^2} = \frac{-0{,}115}{0{,}926^2 + (-0{,}238)^2 + 0{,}872^2 + (-0{,}115)^2} = \frac{-0{,}115}{1{,}6877} = -0{,}0681$$

Výpočet komunalit

$$h_1^2 = \sum_{p=1}^{P} (p_{1p}^N)^2 = p_{11}^2 + p_{12}^2 = 0{,}0092^2 + 0{,}54487^2 = 0{,}2970$$

$$h_2^2 = \sum_{p=1}^{P} (p_{2p}^N)^2 = p_{21}^2 + p_{22}^2 = 0{,}3501^2 + (-0{,}141)^2 = 0{,}143$$

$$h_4^2 = \sum_{p=1}^{P} (p_{4p}^N)^2 = p_{41}^2 + p_{42}^2 = 0{,}7546^2 + (-0{,}0681)^2 = 0{,}5741$$

Obecně platí, že sloupec s nejvyšší hodnotou komunality nejlépe charakterizuje zdrojovou matici. V tomto konkrétním případě je to čtvrtý sloupec s hodnotou 0,5741.

Otázka 5

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

Řešení:

V případě metody PCA jde o ortogonální transformaci, která zachovává veškerou původní informaci. Faktorová analýza, díky předem zvolenému počtu hlavních komponent, vysvětlí podstatnou, ale ne veškerou variabilitu proměnných. V případě metody FA mluvíme o neúplné komponentní analýze, zatímco u metody PCA mluvíme o úplné komponentní analýze.

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.

Řešení:

Skupinový korelační koeficient se vypočítává ze vztahu

$R^2_{xy} = 1 - (1-r_1^2)(1-r_2^2)....(1-r_p^2)$

kde r_p jsou kanonické korelační koeficienty

po zadání do vzorce

$R^2_{xy} = 1 - (1-0{,}830^2)(1-0{,}512^2) = 0{,}77$

$R_{xy} = 0{,}878$

Skupinový korelační koeficient má hodnotu 0,77. Znamená to, že kanonickými korelačními koeficienty lze vysvětlit 77 % variability dat.

Otázka 7

Uveďte příklad vhodný pro zpracování metodou PLS.

Metoda PLS kombinuje metodu hlavních komponent a mnohonásobné lineární regrese. Cílem je objasnit vztahy mezi mnohorozměrnými daty v databázích a využít znalostí těchto vztahů k vysvětlení hodnot jedné skupiny veličin z hodnot veličin jiné skupiny. Např.v lékařském výzkumu by metoda PLS mohla odhalovat biochemické parametry, které mají souvislost s konkrétním onemocněním a např. předpovídat směr léčby pro definovanou skupinu pacientů.

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.

Řešení:

Zvolená metoda – metoda nejbližšího souseda

Je založená na principu, že se objekty shlukují podle nejmenší vzdálenosti mezi dvěmi nejbližšími sousedy.

Matice vzdáleností (v bodě 1)

0	2,24	2,24	10	10,3
2,24	0	1,41	8,06	8,54
2,24	1,41	0	7,81	8,06
10	8,06	7,81	0	1,41
10,3	8,54	8,06	1,41	0

1. krok

Nejmenší vzdálenost mají body 2-3 a 4-5 (hodnota 1,41). Vznikly 3 shluky (1, 2-3, 4-5).

2. krok

výpočet matice vzdáleností

1	0	2,42	10
2 + 3	2,42	0	7,81
4 + 5	10	7,81	0

3. krok

Nejmenší vzdálenost je mezi shlukem 1 a 2+3 (hodnota 2,24). Vznikly 2 shluky (1-2-3, 4-5).

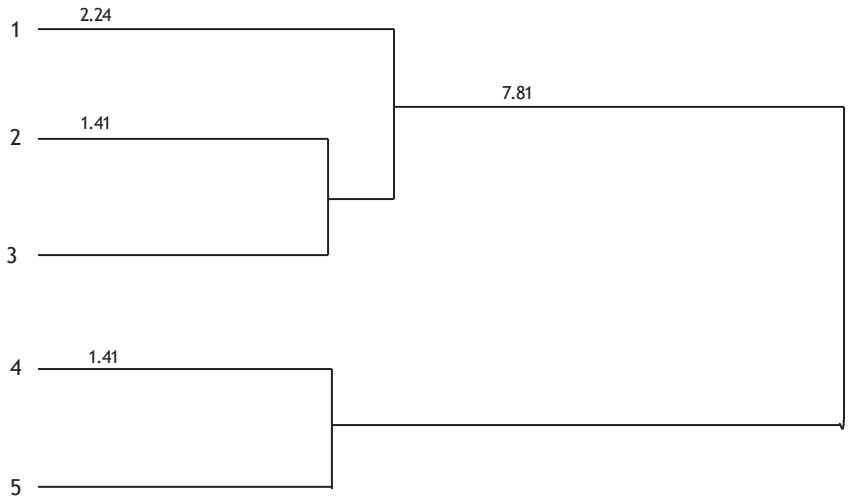
4. krok

výpočet matice vzdáleností

$$\begin{array}{c|cc} 1+2+3 & 0 & 7{,}81 \\ 4+5 & 7{,}81 & 0 \end{array}$$

5. krok

V tomto kroku jsou již všechny objekty seskupené v jednom shluku. Seskupení shluků se graficky znázorní dendogramem.



Jednou z metod shlukové analýzy, metodou nejbližšího souseda, byly vytvořeny shluky znázorněné v dendogramu. Podobné si jsou body 2 a 3 a pak 4 a 5. Bod 1 je podobný shluku bodů 2 a 3. Shluky bodů 1,2,3 a 4,5 si nejsou podobné.

Otázka 9

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

Tyto metody jsou využitelné v biomedicínckém výzkumu, protože pomáhají odhalit skryté proměnné mezi stanovovanými biochemickými, antropometrickými nebo epidemiologickými daty u pacientů.

Konkrétní příklad využití je např. u polygenních onemocnění jako je diabetes mellitus 2. typu, syndrom polycystických ovárií, obezita nebo osteoporóza. Tato onemocnění studujeme na našem pracovišti již mnoho let a k dispozici máme stovky sledovaných parametrů, včetně genetických dat.

Všechna tato onemocnění mají svojí genetickou složku a složku vlivu prostředí, kam patří vlivy výživy, jídelních zvyklostí, stresu, pohybové aktivity, pracovního prostředí a mnoho dalších.

Přes usilovnou snahu nejsou přesné příčiny těchto onemocnění dosud známé. Je tedy třeba stále hledat nové skryté parametry, případně upřesňovat (překlasifikovávat) skupiny pacientů.