Univerzita Pardubice

Fakulta chemicko – technologická

Katedra analytické chemie

Licenční studium chemometrie

Statistické zpracování dat

Metody s latentními proměnnými a klasifikační metody

Zdravotní ústav se sídlem v Ostravě
Odbor hygienických laboratoří Karviná

V Karviné dne 20.5.2006 **Ing. Miluše Galuszková**

Předmět: 3.2 Metody s latentními proměnnými a klasifikační metody

Přednášející: Prof. Ing. Oldřich Pytela, DrSc.

Zadání: Do příštího soustředění předložte ke klasifikaci následující písemnou práci s vypracovanými odpověďmi na předložené otázky. Text s tabulkami napište editorem..

Obsah

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 =

1	**1**	**6**
2	**2**	**4**
3	**0**	**2**

Standardizace zdrojové matice

- vypočteme vektor aritmetických průměrů a vektor směrodatných odchylek

$\bar{x}^T$ = [2 1 4]

s^T = [1 1 2]

- od každého prvku zdrojové matice odečteme aritmetický průměr příslušného sloupce a podělíme směrodatnou odchylkou příslušného sloupce

$$a_{k,i,s} = \frac{a_{k,i} - \bar{x}}{s}$$

Standardizovaná matice A =

-1	**0**	**1**
0	**1**	**0**
1	**-1**	**-1**

Variabilita sloupců je stejná. za odhad hlavní komponenty t_1 vezmeme první sloupec.

t_1^T = [-1 0 1]

- Dosadíme do vztahu

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

- Provedeme normování

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

Získáme počáteční odhad vektoru míry příspěvků odhadu vektoru hlavní komponenty t_1.

- Dosadíme do vztahu

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

Získáme odhad hlavní komponenty t_1.

- Opakováním postupu získáme stabilní rozklad vektorů t_1 a p_1.
- Konvergenční kritérium má vztah

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

- Když $\frac{d}{N} < 10^{-10}$, ukončíme výpočet.

N je počet normování, d je konvergenční kritérium.

Výpočet byl ukončen po 9 krocích. Získali jsme stabilní rozklad vektorů p_1^T a t_1^T:

p_1^T =[0,627967 -0,45969 -0,627997]

t_1^T =[-1,25594 -0,45969 1,71562]

Závěr:

Metodou NIPALS jsme určili stabilní vektory první latentní proměnné

t_1^T =[-1,25594 -0,45969 1,71562]

a její vektor zátěže p_1^T =[0,627967 -0,45969 -0,627997].

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.

Řešení:

K určení nezbytného počtu latentních proměnných použijeme Woldovo kritérium:

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

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

P=1 $\quad \frac{\mathrm{PRESS}(1)}{\mathrm{S_R}\,(0)} = \frac{20}{100} = 0{,}20$

P=2 $\quad \frac{\mathrm{PRESS}(2)}{\mathrm{S_R}\,(1)} = \frac{3{,}5}{10} = 0{,}35$

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

Závěr:

Pro P=3 je hodnota Woldova kritéria větší než 0,95, proto není čtvrtá latentní proměnná významná. Nezbytný počet latentních 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:

Prvky zdrojové matice odpovídající p-té proměnné odhadneme pomocí metody „Krátký cyklus“
Rekonstrukci provedeme podle vztahu:

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

$$A[2,2] = t[2]p[2 = -0{,}311$$

Závěr:

Metodou „Krátký cyklus“ jsme odhadli hodnotu chybějícího prvku A[2,2]= -0,311.

Otázka 4.

Výpočtem metodou PCA byly určeny vektory

p1	0,012	0,458	-0,352	0,987
p2	0,926	-0,238	0,872	-0,115

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

Řešení:

Prvek matice zátěží p_{ip} příslušející 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, tedy **komunalitou**.

- Výpočet komunality h_i^2 podle vztahu:

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

- Výpočet normovaných zátěží p_{ip}^N

$$p_i^N = \frac{p_{ip}}{\sum_{i=1}^{M} p_{ip}^2}$$

kde *p je počet latentních proměnných*
M je počet sloupců zdrojové matice

Jsou určeny vektory zátěží

p1	0,012	0,458	-0,352	0,987
p2	0,926	-0,238	0,872	-0,115

M = 4
P =2

Normované zátěže $p_i^N = \frac{p_{ip}}{\sum_{i=1}^{M} p_{ip}^2}$				Komunality $h_i^2 = \sum_{p=1}^{p} (p_{ip}^N)^2$	
p_{11}^N	0,0092	p_{12}^N	0,5487	h_1^2	0,2970
p_{21}^N	0,3502	p_{22}^N	-0,1410	h_2^2	0,1430
p_{31}^N	-0,2691	p_{32}^N	0,5167	h_3^2	0,3394
p_{41}^N	0,7546	p_{42}^N	-0,068	h_4^2	**0,5741**

Závěr:

Vypočetli jsme komunality. Čím je komunalita příslušného sloupce větší, tím má sloupec vlastnosti společné s ostatními sloupci zdrojové matice. Největší komunalitu má čtvrtý sloupec $h_4^2 = 0{,}5741$, který proto nejlépe charakterizuje zdrojovou matici.

Otázka 5.

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

Řešení:

Vysvětlená variabilita je podíl variability zdrojové matice popsané latentní proměnnou nebo příslušným počtem latentních proměnných.

Vysvětlená variabilita metodou faktorové analýzy FA je vždy nižší než vysvětlená variabilita vypočtená metodou hlavních komponent PCA .

Při výpočtu pomocí FA předem volíme počet hlavních komponent, které reprodukují podstatnou, ale ne všechnu variabilitu proměnných. Z tohoto pohledu je metoda FA neúplnou komponentní analýzou.

Při výpočtu pomocí PCA si předem počet hlavních komponent neurčujeme, ale počítáme je. Vypočtené hlavní komponenty reprodukují variabilitu přesně. PCA je považována za úplnou komponentní analýzu.

Závěr:

Důvodem nerovnosti vysvětlené variability je při výpočtem metodou PCA a FA rozdílný počet hlavních komponent.

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:

- Výpočet skupinového korelačního koeficientu podle vzorce:

$$R_{XY}^2 = 1 - (1 - r_1^2)(1 - r_2^2).....(1 - r_p^2)$$

$$R_{XY}^2 = 1 - (1 - 0{,}830^2)(1 - 0{,}512^2) = 0{,}7704$$
$$R_{XY} = \sqrt{0{,}7704} = 0{,}878$$

Závěr:

Skupinový korelační koeficient má hodnotu 0,878, což znamená, že 77% variability dat bylo vysvětleno kanonickými koeficienty.
První rovnice:
Parametr X3 má velmi malou zátěž 0,050, proto má velmi malý vliv na růst parametru Y1 a pokles parametru Y2.
Druhá rovnice:
U parametru X1 je velmi malá zátěž 0,006, proto je parametr zanedbatelný. Také parametry X2 a X4 mají malou zátěž. Významný vliv je parametru X3 se zátěží 0,950 a podporuje růst obou parametrů Y.

Otázka 7.

Uveďte nějaký konkrétní příklad vhodný pro zpracování metodou PLS.

Řešení:

V laboratoři řízení jakosti důlní firmy jsou prováděny analýzy černého vlaku uhlí (vagony).

Příklad zkoušky tuhých paliv:
- Stanovení spalného tepla kalorimetrickou metodou v tlakové nádobě a výpočet výhřevnosti
- Stanovení prchavé hořlaviny
- Zrychlené stanovení celkového obsahu vody v uhlí
- Třídící zkouška proséváním
- Stanovení skutečné hustoty uhlí
- Stanovení veškeré síry metodou Eschka
- Stanovení popela spálením na vzduchu

Tabulka:

č.	popel	spalné teplo	prchavá hořlavina	celkový obsah uhlí	zkouška prosévánim	hustota uhlí	veškerá síra
1							
2							
3							
...							
...							
...							
30							

Na datech uvedených v tabulce se pokuste najít vztah mezi chemickým složením a užitnými vlastnostmi, aby bylo možné urychleně expedovat uhlí již nasypané do expedovaného vlaku

- Provede se standardizace dat
- Metoda PLS

Závěr:

Využitím metody PLS zjistíme vazby mezi skupinami veličin, t.j. závislost mezi obsahy jednotlivých složek.

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:

Stanovení znaků určujících podobnost

Podobnost mezi objekty je užita jako kritérium tvorby shluků objektů.Jedním z typů podobnosti vyjádřené vzdáleností pro metricky proměnné je eukleidovská vzdálenost.
Platí vztah:

$$d_E(x_k, x_l) = \sqrt{\sum_{j'1}^{m} (x_{kj} - x_{lj})^2}$$

$d_E(x_k,x_l)$ *vzdálenost mezi objekty*
x_{kj} a x_{lj} *jsou souřadnice objektů v M rozměrném prostoru*

Dosazením metrických znaků objektů do vzorce vypočteme vzdálenosti mezi jednotlivými objekty.

$d_E(x_k, x_l) = \sqrt{\sum_{j'1}^{m} (x_{kj} - x_{lj})^2}$	
$d_E(1,2)$	2,24
$d_E(1,3)$	2,24
$d_E(1,4)$	10,0
$d_E(1,5)$	10,3
$d_E(2,3)$	1,41
$d_E(2,4)$	8,06
$d_E(2,5)$	8,54
$d_E(3,4)$	7,81
$d_E(3,5)$	8,06
$d_E(4,5)$	1,41

Uspořádání matice vzdáleností

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

Shlukovací metody

Metoda průměrné vzdálenosti

1.krok
Nejmenší vzdálenosti mají body 2-3 a 4-5.
- hodnota vzdálenosti 1,41

V prvém kroku vzniknou 3 shluky 1, 2-3, 4-5

2.krok
Určení metrických znaků pro vzniklé shluky.

- u bodu 1 zůstanou 2,10
- u shluku 2-3, 4-5 se metrické znaky vypočtou pomocí průměru

znak 1	2; 10
znak 2-3	3,5; 8,5
znak 4-5	10,5; 4,5

3.krok
Vypočtení matice vzdáleností

1	0	2,12	10,12
2+3	2,12	0	8,06
4+5	10,12	8,06	0

4.krok
Nejkratší vzdálenost mají bod 1 a shluk 2-3. Hodnota vzdálenosti je 2,12
Vytvořily se 2 shluky 1-2-3 a 4-5.

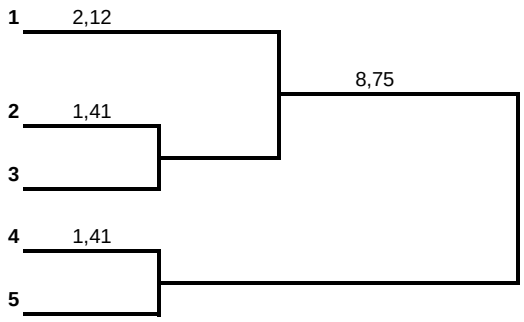
5.krok
Určení metrických znaků pro vzniklé shluky.

znak 1-2-3	3; 9
znak 4-5	10,5; 4,5

6.krok
Vypočtení matice vzdáleností

1+2+3	0	8,75
4+5	8,75	0

7.krok
Objekty jsou seskupeny do jediného shluku.
Znázornění seskupení objektů je provedeno pomocí dendrogramu.



Závěr:

Byla vypočtena matice vzdáleností pomocí eukleidovské vzdálenosti.Shlukovací metodou průměrné vzdálenosti byly objekty roztříděny do shluků. Postup shlukování byl znázorněn pomocí dendrogramu.

Otázka 9.

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

Řešení:

Obecný postup
1.Průzkumová analýza vícerozměrných dat
2.Analýza korelační matice a matice parciálních korelačních koeficientů
3.Volba metody
4.Předzpracování dat - standardizace
5.Výpočet latentních proměnných
6.Určení počtu signifikantních latentních proměnných
7.Analýza matice zátěží a matice latentních proměnných
8.Analýza fyzikálně chemického smyslu latentních proměnných
9.Interpretace výsledků vzhledem k cíli analýzy

Zadání:

Na území ovlivněném důlní činností je prováděn monitoring vod. Pomocí zvolených ukazatelů je sledován průsak podzemní minerální vody do povrchových vod. Vzorkování a analýzy v laboratoři jsou finančně a časově náročné. Je nutné redukovat počet odběrových míst a stanovovat omezené množství ukazatelů.

Matice vstupních dat obsahuje 120 objektů a 11 znaků.

Průzkumovou analýzou jednorozměrných dat vyšetříme vstupní data, pro vícerozměrnou analýzu dat budou využity znaky s největší proměnlivostí.Za účelem vyšetření vztahů mezi jednotlivými páry znaků bude provedena lineární regrese a vypočteny korelační koeficienty. Maticové diagramy znázorní rozptylové diagramy jednotlivých dvojic znaků. Pokud je znázorněn mrak bodů znamená to, že mezi znaky není korelace.Pro vícerozměrnou analýzu dat budou využity znaky s největší proměnlivostí a silnou korelací.

Pro analýzu vícerozměrných dat bude využita zdrojová matice obsahující 120 objektů a 6 znaků.

Vícerozměrná statistická analýza:

- ze vstupních dat sestavit zdrojovou matici (120 objektů, 6 znaků)
- matici je nutné standardizovat (různé jednotky znaků)
- odhad parametrů polohy, rozptýlení, tvaru a intenzity vztahu mezi proměnnými
- exploratorní analýza dat
  - podobnost objektů (rozptylové diagramy, symbolové grafy, profilové grafy)
  - vybočující objekty nevhodné k analýze
  - předpoklad lineárních vazeb
  - testování předpokladů o datech

- **volba metody hlavních komponent PCA**

- určení počtu latentních proměnných
  - pomocí Cattelova indexového grafu úpatí vlastních čísel
- určení struktury proměnných a vzájemných vazeb (graf komponentních vah, rozptylový diagram komponentního skore, dvojný graf Biplot))

- určení struktury a vzájemných vazeb v objektech
  - analýza shluků CLU. Znázornění pomocí dendrogramu.

- **Interpretace výsledků vzhledem k cíli analýzy:**

  - Metodou analýzy hlavních komponent PCA byla zjednodušena skupina korelovaných znaků. Pomocí grafu úpatí vlastních čísel byl určen počet hlavních komponent. První hlavní dvě komponenty popsaly data z *80,4*%.Byla snížena rozměrovost zdrojové matice Monitoring vod ze 6 znaků na 2 latentní proměnné.Byl nalezen shluk vzájemně podobných objektů a objekty odlišné od ostatních objektů.

  - Metodou analýzy shluků CLU byla zkoumána podobnost vícerozměrných objektů a objekty byly roztříděny do skupin. Výsledek třídění byl zobrazen dendrogramem .

Závěr:

Statistická analýza dat potvrdila podezření na průsak vod minerálních vlivem důlních činností.
Pro sledování v dalších letech je možné zredukovat sledované znaky z 11 na 6 a snížit počet objektů, které jsou vtěsnány do velkého shluku zřetelného v grafu komponentních vah a/nebo v dendrogramu.