Statistické zpracování dat

11.licenční studium rok 2007

VI.soustředění září 2007

Téma úkolů: Matematické principy analýzy vícenásobných dat
Metoda s latentními proměnnými

Vypracovala: Ivana Kulhánková

Otázka 1. Najděte vlastní (charakteristická) čísla a vlastní vektory, determinant, stopu a odmocninu od této jednoduché matice

1 2 3
2 8 2
3 2 10

Vlastní čísla: -0,143295, 6.76570, 12,37758

Vektory: u1 = c1[-0,953204, 0,173258, 0,247759]^T
u2 = c2*[-0,014312, -0,844442, 0,535456]^T
u3 = c3*[0,301999, 0,506852, 0,807405] ^T

Determinant: -12

Stopa: součet prvků na diagonále = 19

Odmocnina: 0,32138 + 0,34394i 0,56994 – 0,06252i 0,83790 – 0,06940i
0,56994 – 0,06252i 2,75861 + 0,01136i 0,26365 + 0,01625i
0,83790 – 0,08940i 0,26365 + 0,01625i 3,03928 + 0,02324i

Příklad 1
Vypočtěte algoritmem NIPALS 1 latentní proměnnou z matice A:

1 1 6
2 2 4
3 0 2

Vektor aritmetických průměrů: x^{-T} = [2 1 4]
Vektor směrodatných odchylek: s^{T}= [1 1 2]
Standardizace: -1 0 1
0 1 0
1 -1 -1
Odhad hlavní komponenty z kteréhokoliv sloupce vzhledem ke stejné variabilitě všech sloupců:
$t_i = (p_i^T - p_i^T)^{-1} p_i^T . A^T$

Konvergační kritérium: $d = (t_{nové} - t_{staré})^T (t_{nové} - t_{staré}) (t^T_{nové} . t_{nové})^{-1}$

$d/N < 10^{-10}$

Pro případ prvního sloupce je třeba osmi kroků.
Vektor zátěží: p_1^T = [0,62797 -0,45969 -0,62797]
Vektor latentních proměnných: t_1^T = [-1,25594 -0,45969 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.

Woldovo kritérium je založeno na poměru PRESS(P) / S_R(P-1), pak platí

P=1: PRESS(1) / S_R (1) = 20/100 = 0,2
P=2: PRESS(2) / S_R (2) = 3,5/10 = 0,35
P=3: PRESS(3) / S_R (3) = 3,45/3,4 = 1,01

Je-li hodnota kritéria vyšší jak 0,95, další latentní proměnná je nevhodná.

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

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

A[2,2] = t [2] p[2]

A[2,2] = - 0,3109

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.

Výpočet komunality: $h_i^2 = p^N (p^N)^T = \sum^{P} (p^N_{ip})^2$

$h_1^2 = 0{,}0092^2 + 0{,}54487^2 = 0{,}2970$
$h_3^2 = 0{,}3501^2 + (-0{,}1410)^2 = 0{,}1430$
$h_4^2 = 0{,}7546^2 + (-0{,}0681)^2 = 0{,}5741$

Čtvrtý sloupec nejlépe charakterizuje zdrojovou matici.

p=1

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

Metoda PCA je metoda kompletní komponentní analýzy, která předpokládá přesné reprodukování variability zdrojové matice, kdežto metoda faktorová FA připouští existenci nevysvětlitelné variability.

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.

Výpočet skupinového korelačního koeficientu R:
$R^2 = 1 - (1- r_1^2)(1- r_2^2)$

$R^2 = 0{,}77$

R = 0,878

Výsledek popisuje 77% variability dat. S nárůstem proměnné X3 roste proměnná Y2.
S nárůstem X1, X2 a X4 roste proměnná Y1.

Otázka 7. Uveďte nějaký konkrétní příklad vhodný pro zpracování metodou metodou PLS.

Metodou PLS ověřte, zda lze popsat tepelné vlastnosti spalovaného uhlí, štěpek, obilí pomocí spalného tepla a výhřevnosti.

Materiál	Spalné teplo	Výhřevnost
Uhlí	**********	**********
Štěpky	**********	**********
Obilí	**********	**********