Stockholms universitet

STATISTISKA INSTITUTIONEN HT-2014

Written exam in Multivariate Methods, 7.5 ECTS credits
Thursday, 23th October 2014, 12:00 – 17:00
Time allowed: FIVE hours
Examination Hall: Uggleviksalen

You are required to answer all **6 (six)** questions as well as motivate your solutions. The total amount of points is 80. In order to pass this part, you need to get at least 40 points. Points from this exam will be added to your results from the home assignment. The final grades are assigned as follows: **A** (91+), **B** (75-90), **C** (66-74), **D** (58-65), **E** (50-57), **Fx** (30-49), and **F** (0-29).

You are **allowed** to use a pocket calculator, a language dictionary, and a list of formulas (attached). In addition, you are **allowed** to use a one-sided A4 containing your own formulae, but excluding proofs and solutions. The A4 must be approved (signed) by the teacher; and, it must be submitted along with your solutions. If the A4 is not signed by the teacher and discovered, student might be accused in cheating on exam.

The teacher reserves the right to examine the students **orally** on the questions in this examination.

---

1. (10 points) The coordinates of three points with respect to the orthogonal basis vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ are as follows:

   A = (2,-1) B = (1, 0.7) C = (-3, 7.5)

   Show that A, B and C lie on a straight line.
   The set of oblique basis vectors f_1 and f_2 are related to vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ as follows:

   $$\mathbf{e}_1 = .804\mathbf{f}_1 + .629\mathbf{f}_2$$
   $$\mathbf{e}_2 = .273\mathbf{f}_1 - .588\mathbf{f}_2$$

   Compute the coordinates of A, B and C with respect to $\mathbf{f}_1$ and $\mathbf{f}_2$. Do they still lie on a straight line? Justify your answer.

2. (18 points) Use the data in the contingency Table below to answer the following questions:

Blood Cholesterol (mg/100 cc)	Heart Disease	
	Present	Absent
<200	6	5
200-219	10	6
220-259	30	5
>259	45	7

   a. What is the probability that
      (i) Heart disease is present?
      (ii) Blood cholesterol is less than 219 mg/100 cc?

b. What are the odds that

(i) Heart disease is present given that blood cholesterol is less than 200 mg/100 cc?

(ii) Blood cholesterol is greater than 219 mg/100 cc given that heart disease is present?

c. Compute the log of odds that

(i) Heart disease is present given that blood cholesterol is greater than 259 mg/100 cc.

(ii) Heart disease is present given that blood cholesterol is between 200 and 219 mg/100 cc.

3. (15 points) Consider the two-indicator two-factor model represented by the following equations:

$$X_1 = 0.104F_1 + 0.824F_2 + U_1$$
$$X_2 = 0.065F_1 + 0.959F_2 + U_2$$
$$X_3 = 0.065F_1 + 0.725F_2 + U_3$$
$$X_4 = 0.906F_1 + 0.134F_2 + U_4$$
$$X_5 = 0.977F_1 + 0.116F_2 + U_5$$
$$X_6 = 0.827F_1 + 0.016F_2 + U_6$$

The usual assumptions hold for the above model. Answer the following questions assuming that the correlation between the common factors F_1 and F_2 is given by
Corr(F_1, F_2)= $\Phi_{12} = 0.3$

(a) What are the pattern loadings of indicators X_1, X_4 and X_6 on the factors F_1 and F_2? (3 points)

(b) Compute the correlation between the indicators X_1 and X_2. (4 points)

(c) What percentage of the variance of indicators X_1 and X_2 is not accounted for by the common factors F_1 and F_2?

4. (12 points) Perform principal component analysis on the following data by hand. In other words, determine the angle (with best precision you can: 1%, 5%, 10%: your calculators should be of help) between the new axis and the old axis that would give a new variable, which accounts for the maximum variance in the data. What conclusions can you draw?

X_1	X_2
1	4
1	1
2	2
2	3
3	2
3	2
4	1
4	4

5. (15 points) Consider the following single-factor model

$$\begin{aligned} x_1 &= \lambda_1 \xi + \delta_1 \\ x_2 &= \lambda_2 \xi + \delta_2 \\ x_3 &= \lambda_3 \xi + \delta_3 \end{aligned}$$

If the sample covariance matrix of the indicators is given by:

$$S = \begin{pmatrix} 1.20 & 0.93 & 0.45 \\ 0.93 & 1.56 & 0.27 \\ 0.45 & 0.27 & 2.15 \end{pmatrix}$$

Compute the estimates of the model parameters $(\lambda_1, \lambda_2, \lambda_3, Var(\delta_1), Var(\delta_2), Var(\delta_3))$ using hand calculations. Are the parameter estimates unique?

6. (10 points)
   (i) What problems cluster analysis is intended to solve? What types of cluster analysis you know: name them. What assumptions on data are imposed in order to apply "cluster analysis"? How strict you should be with those assumptions: speculate and exemplify "why?" to the best of your knowledge.
   (ii) Cluster the following hypothetical data set into two groups using similarity matrixes. Intermediate similarity matrixes should be stated clearly. It is your choice to choose the type of distance used in calculations.

	Hypothetical Data	
Subject ID	Income in EUR	Education (in years)
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19

Two-Factor Model: $x_1 = \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \varepsilon_1$

$x_2 = \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \varepsilon_2$

$\vdots$

$x_p = \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \varepsilon_p$

The variance of x: $E(x^2) = E(\lambda_1\xi_1 + \lambda_1\xi_2 + \varepsilon_1)^2;\ Var(x) = \lambda_1^2 + \lambda_2^2 + Var(\varepsilon) + 2\lambda_1\lambda_2\phi$

The correlation between any indicator and any factor (the structure loading):

$E(x\xi_1) = E[(\lambda_1\xi_1 + \lambda_1\xi_2 + \varepsilon_1)\xi_1]\ ;\ Corr(x\xi_1) = \lambda_1 + \lambda_2\phi$

The shared variance between the factor and an indicator: $Shared\ variance = (\lambda_1 + \lambda_2\phi)^2$

The correlation between two indicators:

$$E(x_j x_k) = E\left[(\lambda_{j1}\xi_1 + \lambda_{j2}\xi_2 + \varepsilon_j)(\lambda_{k1}\xi_1 + \lambda_{k2}\xi_2 + \varepsilon_k)\right]$$

$$Corr(x_j x_k) = \lambda_{j1}\lambda_{k1} + \lambda_{j2}\lambda_{k2} + (\lambda_{j1}\lambda_{k2} + \lambda_{j2}\lambda_{k1})\phi$$

Confirmatory Factor Analysis

The covariance matrix (one-factor model, two indicators): $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$

Evaluating model fit: χ^2-test H_0: $\boldsymbol{\Sigma = \Sigma(\theta)}$ H_a: $\boldsymbol{\Sigma \neq \Sigma(\theta)}$ (test whether the difference between the sample and the estimated covariance matrix is a zero matrix)

$\chi^2 = \sum_{i=1}^{k} \frac{|n_i - E(n_i)|^2}{E(n_i)}$

Cluster Analysis

Measure of similarity – squared Euclidean distance between two points

Hierarchical clustering:

Centroid method – each group is replaced by centroid

Nearest-neighbor or single-linkage method – the distance between two clusters is represented by the minimum of the distance between all possible pair of subjects in the two clusters

Farthest-neighbor or complete-linkage method - … the maximum of the distances…

Average-linkage method - … the average distance…

Ward's method – does not compute distances between clusters. Method tries to minimize the total within-group sums of squares.

Discriminant Analysis

Assumptions: multivariate normality, equality of covariance matrices

Discriminant function: $Z = w_1x_1 + w_2x_2$

$\lambda = \frac{between\ -group\ sum\ of\ squares}{within-group\ sum\ of\ squares}$

$\boldsymbol{\Sigma}$-variance-covariance matrix, **T**-total SSCP matrix. $\boldsymbol{\gamma}$-vector of weights.

Discriminant function $\boldsymbol{\xi = X'\gamma}$. **B** and **W** are between-groups and within-group SSCP matrices.

Maximize $\lambda = \frac{\gamma' B\gamma}{\gamma' W\gamma}$

$\boldsymbol{|W^{-1}B - \lambda I| = 0;\ \gamma = \Sigma^{-1}(\mu_1 - \mu_2)}$ - Fisher's discriminant function

Logistic regression

$odds = \frac{p}{1-p}$

$\ln odds = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$

$p = \frac{1}{1+e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$

Maximum likelihood estimation: $P(Y = 1) = p = \frac{e^{\beta X}}{1+e^{\beta X}}$

$L = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$

Quadratic equations: $ax^2 + bx + c = 0; x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

Cubic equations:

$$y^3 + ay^2 + by + c = 0; y = x - \frac{a}{3};\ x^3 + px + q = 0; x_1 = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}$$

Formula Sheet, Multivariate Methods

Matrices

Transpose – exchange rows and columns
Identity (I) – diag (1,1…) of order n*n
Inverse of A (A^{-1}): $AA^{-1} = A^{-1}A = I$
$A + B = B + A$; $x(A + B) = xA + xB$; $AB \neq BA$ $(in\ general)$;
If order (A)=m*n, order (B)=n*p, then C=AB is of order m*p
D=det A=$\begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nn} \end{vmatrix}$
detA=$a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots + a_{in}A_{in}$ where cofactor $A_{ij} = (-1)^{i+j} D_{ij}$ (i-row, j-column of D)
Cramer's rule: $x_j = D_j / D$ where D=detA and D_j is the determinant that arises when the j column of D is replaced by the column elements $b_1, \ldots, b_n$. **(Ax=b)**

Vectors

$\boldsymbol{a} = (a_1 a_2 \ldots a_p)$
A right-angle triangle: α - angle between a and c; c – hypotenuse; $\cos\alpha = \frac{a}{c}$, $\sin\alpha = \frac{b}{c}$
Length of vector **a**=$\|\boldsymbol{a}\| = \sqrt{a_1^2 + a_2^2}$
Basis vectors $\boldsymbol{e_1} = (1\ 0), \boldsymbol{e_2} = (0\ 1)$
$\boldsymbol{a} = a_1\boldsymbol{e_1} + a_2\boldsymbol{e_2}$
Scalar product $\boldsymbol{ab} = a_1b_1 + a_2b_2 + \cdots + a_pb_p$; $\boldsymbol{ab} = \|\boldsymbol{a}\|\|\boldsymbol{b}\|\cos\alpha$
Length of the projection: $\|\boldsymbol{a_p}\| = \|\boldsymbol{a}\|\cos\alpha$
Variance of x_i: $s_1^2 = \frac{\|x_i\|^2}{n-1}$; Generalized variance: $GV = \left(\frac{\|x_1\|\cdot\|x_2\|}{n-1} \cdot \sin\alpha\right)^2$

Distances

Euclidean: $D_{AB} = \sqrt{\sum_{j=1}^{p}(a_j - b_j)^2}$
Statistical: $SD_{ij}^2 = \left(\frac{x_i - x_j}{s}\right)^2$, s-standard deviation
Manalanobis: $MD_{ik}^2 = \frac{1}{1-r^2}\left[\frac{(x_{i1}-x_{k1})^2}{s_1^2} + \frac{(x_{i2}-x_{k2})^2}{s_2^2} - \frac{2r(x_{i1}-x_{k1})(x_{i2}-x_{k2})}{s_1 s_2}\right]$

Variance, Sum of Squares, and Cross Products

Variance: $s_j^2 = \frac{\sum_{i=1}^{n} x_{ij}^2}{n-1} = \frac{SS}{df}$ (sum of squares/degrees of freedom)
Covariance: $s_{jk} = \frac{\sum_{i=1}^{n} x_{ij}x_{ik}}{n-1} = \frac{SCP}{df}$ (sum of the cross products/degrees of freedom)
SSCP – sum of squares and cross products matrix $\begin{pmatrix} SSX_1 & SCP \\ SCP & SSX_2 \end{pmatrix}$
S – covariance matrix $S_t = \frac{SSCP_t}{df}$
Within-Group Analysis: $SSCP_w = SSCP_1 + SSCP_2$ (pooled SSCP matrix) $S_w = \frac{SSCP_w}{n_1+n_2-2}$ (pooled cov m)
Between-Group Analysis: $SS_j = \sum_{g=1}^{G} n_g(\bar{x}_{jg} - \bar{x}_j)^2$; $SCP_{jk} = \sum_{g=1}^{G} n_g(\bar{x}_{jg} - \bar{x}_j)(\bar{x}_{kg} - \bar{x}_k)$
$SSCP_t = SSCP_w + SSCP_b$

Principal Components Analysis

$x_1^* = \cos\theta * x_1 + \sin\theta * x_2$; $x_2^* = -sin\theta * x_1 + cos\theta * x_2$
$\boldsymbol{\Sigma}$ covariance matrix; $\boldsymbol{\lambda}$-eigenvalues; $|\boldsymbol{\Sigma} - \boldsymbol{\lambda I}| = \mathbf{0}$; $\boldsymbol{\gamma}$-eigenvector; $(\boldsymbol{\Sigma} - \boldsymbol{\lambda I})\boldsymbol{\gamma} = \mathbf{0}$; $\boldsymbol{\gamma}'\boldsymbol{\gamma} = \mathbf{1}$;

Factor Analysis

Assumptions: 1. Means of indicators, common factor, unique factors are zero.
2. Variances of indicators and common factors are one. 3.$E(\xi_i\varepsilon_i) = 0$ and $E(\varepsilon_i\varepsilon_j) = 0$

MULTIVARIATE METHODS

PCA

$$\xi_1 = W_{11}X_1 + W_{12}X_2 + \dots + W_{1p}X_p$$
$$\xi_2 = W_{21}X_1 + W_{22}X_2 + \dots + W_{2p}X_p$$
$$\vdots$$
$$\xi_p = W_{p1}X_1 + W_{p2}X_2 + \dots + W_{pp}X_p$$

- The weights: $W_{i1}^2 + W_{i2}^2 + \dots + W_{ip}^2 = 1 \quad , i=1,\dots,p$

$$W_{i1}W_{j1} + W_{i2}W_{j2} + \dots + W_{ip}W_{jp} = 0 \quad \forall i \neq j$$

- Characteristic equation: $\det(\lambda I - A) = 0$

- Loadings = correlation between the original and the new variables

$$l_{ij} = \frac{W_{ij}}{\hat{s}_j}\sqrt{\lambda_i}$$

FA

$$X_1 = \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \dots + \lambda_{1m}\xi_m + \varepsilon_1$$
$$X_2 = \lambda_{12}\xi_1 + \lambda_{22}\xi_2 + \dots + \lambda_{2m}\xi_m + \varepsilon_2$$
$$\vdots$$
$$X_p = \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \dots + \lambda_{pm}\xi_m + \varepsilon_p$$

- Assumptions:

1) Means of indicators, common factors and unique factors are zero.
2) Variances of indicators and common factors are one.
3) The unique factors are not correlated among themselves or with the common factors. $\Rightarrow E(\xi_i\varepsilon_j) = E(\varepsilon_i\varepsilon_j) = 0 \quad \forall i \neq j$

Model	Equations	The variance of ing indicator x_j	Structure loading	Shared variance = (structure loading)^2	Correlation between indicators
One-factor model	$X_1 = \lambda_1\xi + e_1$ $\vdots$ $X_p = \lambda_p\xi + \varepsilon_p$	$V(x_j) = \lambda_j^2 + V(\varepsilon_j)$	$Cor(x_j, \xi) = \lambda_j$	λ_j^2	$Corr(x_j, x_k) = \lambda_j\lambda_k$
Two-factor model	$X_1 = \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \varepsilon_1$ $\vdots$ $X_p = \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \varepsilon_p$	$V(x_j) = \lambda_{j1}^2 + \lambda_{j2}^2 + 2\lambda_{j1}\lambda_{j2} + V(\varepsilon_j)$	$Cor(x_j, \xi_2) = \lambda_{j2} + \lambda_{j1}\phi$ $Cor(x_j, \xi_1) = \lambda_{j1} + \lambda_{j2}\phi$	$(\lambda_{j2} + \lambda_{j1}\phi)^2$ $(\lambda_{j1} + \lambda_{j2}\phi)^2$	$Cor(x_j, x_k) = \lambda_{j1}\lambda_{k1} + \lambda_{j2}\lambda_{k2} + [\lambda_{j1}\lambda_{k2} + \lambda_{j2}\lambda_{k1}]\phi$

$[Cor(\xi_1, \xi_2) = \phi]$

CFA

Underidentified model: # equations < # variables
Just-identified model: # equations = # variables
Overidentified model: # equations > # variables

$$P(X_1, X_2, X_3) = P(X_1)\cdot P(X_2|X_1)\cdot P(X_3|X_1, X_2)$$

CA

$$D_{AB}^2 = \sum_{j=1}^{p}(a_j - b_j)^2 \qquad SD_{ik}^2 = \sum_{j=1}^{p}\left(\frac{x_{ij} - x_{kj}}{s_j}\right)^2$$

$$MD_{ik}^2 = \frac{1}{1-r^2}\left[\frac{(x_{i1}-x_{k1})^2}{s_1^2} + \frac{(x_{i2}-x_{k2})^2}{s_2^2} - \frac{2r(x_{i1}-x_{k1})(x_{i2}-x_{k2})}{s_1 s_2}\right]$$

DA (two groups)

$$\lambda = \frac{SS_B}{SS_W} \to \max \qquad \text{If } \xi = \bar{X}^T\bar{\gamma} \Rightarrow \text{The estimation of } \bar{\gamma}: \ \bar{\gamma}^T = (\bar{\mu}_1 - \bar{\mu}_2)^T\Sigma^{-1}$$

$$SSCP_W = SSCP_1 + SSCP_2$$
$$SSCP_T = SSCP_W + SSCP_B$$

LOG-REG

$$odds = \frac{p}{1-p} \qquad \ln\frac{p}{1-p} = \beta_0 + \beta_1X_1 + \dots + \beta_kX_k$$

$$p = \frac{odds}{1+odds} \qquad P(Y=1) = p = \frac{1}{1 + e^{-(\beta_0 + \beta_1X_1 + \dots + \beta_kX_k)}}$$

If $Y = \beta_0 + \beta_1X$:

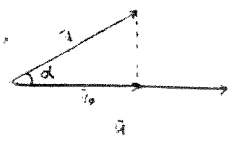
$$\hat{\beta}_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

- Rotation $\Rightarrow$ $a_1^r = \cos\theta a_1 + \sin\theta a_2$
$a_2^r = -\sin\theta a_1 + \cos\theta a_2$

$$S^2 = \frac{SS}{df} = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{\Sigma x_i^2 - n\bar{x}^2}{n-1}$$

$$S_{xy} = \frac{SCP}{df} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{n-1}$$

Projection: of $\bar{v}$ onto $\bar{u}$



- The projection vector: $\bar{v}_p = \frac{\|\bar{v}_p\|}{\|\bar{u}\|}\bar{u}$

- $\|\bar{v}_p\| = \|\bar{v}\|\cos\alpha = \frac{\bar{v}\cdot\bar{u}}{\|\bar{u}\|}$

Direction cosines = The cosines of the angle between a vector and the axes

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 23/10 -2014

Sal: Uggleviksalen

Tenta: Multivariata metoder

Kurs: Multivariata metoder

ANONYMKOD: MM-0022

☒ Jag godkänner att min tenta får läggas ut anonymt på hemsidan som studentsvar.

OBS! SKRIV ÄVEN PÅ BAKSIDAN AV SKRIVBLADEN

Markera besvarade uppgifter med kryss

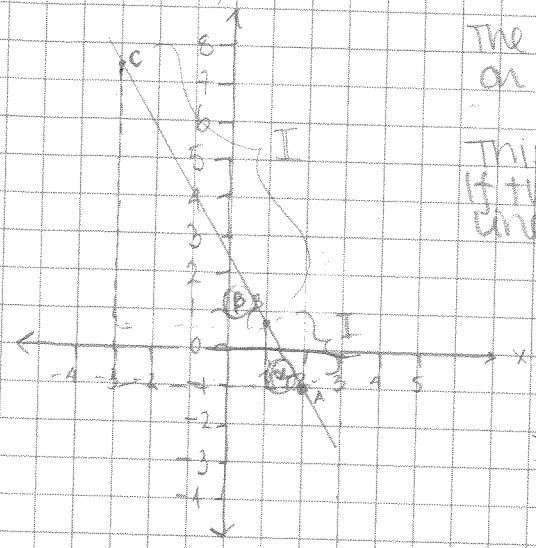
1	2	3	4	5	6	7	8	9	Antal inl. blad
X	X	X	X	X	X				6 2k
Lär.ant.									
10	18	15	12	14,	10				

POÄNG	BETYG	Lärarens sign.
79 AA+		

SU, STATISTIK

Skrivsal: UGGLEVIK Anonymkod: MMI-0022 Blad nr: 1

① $A = (2, -1)$ $B = (1, 0.7)$ $C = (-3, 7.5)$



The graphic solution shows that they are on a straight line.

This can also be proved analytically. If the points are in a straight line, then $\cos(\alpha) = \cos(\beta)$.

$\cos(\beta) = \frac{(3+1)}{II}$

$II = D_{BC} = \sqrt{4^2 + 6{,}8^2} = 7{,}8892$

$\cos(\beta) = \frac{4}{7{,}8892} = 0{,}5070$

$\cos(\alpha) = \frac{5}{I+II}$ $I+II = \sqrt{5^2 + 8{,}5^2} = 9{,}8615$

$\cos\alpha = \frac{5}{9{,}8615} = 0{,}5070$ $\cos(\alpha) = \cos(\beta)$!

We can also show this by calculating the linear equations:

$-1 = 2x + m$ $0{,}7 = x + m \rightarrow x = 0{,}7 - m$

$-1 = 2(0{,}7 - m) + m = 1{,}4 - 2m + m \rightarrow m = 2{,}4$ $x = -1{,}7$

$y = -1{,}7x + m$ → BY CALCULATING Y FROM THE LINEAR EQUATION:

$x_A = 2 \rightarrow y_A = -1$ ✓

$x_B = 1 \rightarrow y_B = 0{,}7$ ✓

$x_C = -3 \rightarrow y_C = 7{,}5$ ✓

THEY ARE ON A STRAIGHT LINE!

(1). cont. $e_1 = 0,804 f_1 + 0,629 f_2$

$e_2 = 0,273 f_1 - 0,588 f_2$

$\begin{cases} A = 2e_1 - e_2 \\ B = e_1 + 0,7 e_2 \\ C = -3e_1 + 7,5 e_2 \end{cases}$

$A = f(f_1, f_2)$?

$A = 2e_1 - e_2 = 2(0,804 f_1 + 0,629 f_2) - (0,273 f_1 - 0,588 f_2)$

$A = 1,608 f_1 + 1,258 f_2 - 0,273 f_1 + 0,588 f_2$

$A = 1,335 f_1 + 1,846 f_2$ $\quad A = (1.335, 1.846)$

$B = f(f_1, f_2)$?

$B = e_1 + 0,7 e_2 = 0,804 f_1 + 0,629 f_2 + 0,7(0,273 f_1 - 0,588 f_2)$

$B = 0,804 f_1 + 0,629 f_2 + 0,1911 f_1 - 0,4116 f_2$

$B = 0,9951 f_1 + 0,2174 f_2$ $\quad B = (0.9951, 0.2174)$

$C = f(f_1, f_2)$?

$C = -3e_1 + 7,5 e_2 = -3(0,804 f_1 + 0,629 f_2) + 7,5(0,273 f_1 - 0,588 f_2)$

$C = -2,412 f_1 - 1,887 f_2 + 2,0475 f_1 - 4,41 f_2$

$C = -0,3645 f_1 - 6,297 f_2$ $\quad C = (-0,3645, -6,297)$

To test if these are on a linear line I've chosen the method of finding one linear regression to see if it is suitable for A, B and C.

$1,846 = 1,335 x + m$ $\quad 0,2174 = 0,9951 x + m$

$m = 0,2174 - 0,9951 x \rightarrow 1,846 = 1,335 x + 0,2174 - 0,9951 x$

$1,6286 = 0,3399 x$

$x = 4,7914$ $\quad m = -4,5505$ $\quad \Big\} \; y = 4,7914 x - 4,5505$

BY CALCULATING Y FROM EQUATION ↗

A: $x_A = 1,335 \rightarrow y_A = 1,846$ ✓

B: $x_B = 0,9951 \rightarrow y_B = 0,2174$ ✓

C: $x_C = -0,3645 \rightarrow y_C = -6,297$ ✓

THEN (ON A) ARE LINEAR LINE!

We know that this is true since the axis have changed from e_1 and e_2 to f_1 and f_2, it does not change that the relationship between A, B and C has changed. They are still on a straight line.

DK

(2)

Blood Cholesterol (mg/100cc)	Heart Disease Present	Heart Disease Absent	
<200	6	5	= 11
200-219	10	6	= 16
220-259	30	5	= 35
>259	45	7	= 52
	91	23	114

a) i) $P(\text{present}) = \frac{91}{114} = 0,7982$ +

ii) $P(BC < 219) = \frac{16 + 11}{114} = 0,2368$ +

b) i) $P(\text{present} \mid BC < 200) = \frac{6}{11}$

$\text{odds}(\text{present} \mid BC < 200) = \frac{\left(\frac{6}{11}\right)}{1 - \left(\frac{6}{11}\right)} = 1,20$ +

ii) $P(BC > 219 \mid \text{present}) = \frac{75}{91}$

$\text{odds}(BC > 219 \mid \text{present}) = \frac{\left(\frac{75}{91}\right)}{1 - \left(\frac{75}{91}\right)} = 4,6875$ +

c) i) $P(\text{present} \mid BC > 259) = \frac{45}{52}$ +

$\text{odds}(\text{present} \mid BC > 259) = \frac{\left(\frac{45}{52}\right)}{1 - \left(\frac{45}{52}\right)} = 6,429$

$\ln(\text{odds}) = \ln(6,429) \approx 1,861$ +

ii) $P(\text{present} \mid 200 < BC < 219) = \frac{10}{16}$

$\text{odds}(\text{present} \mid 200 < BC < 219) = \frac{\left(\frac{10}{16}\right)}{1 - \left(\frac{10}{16}\right)} = 1,667$

$\ln(\text{odds}) = \ln(1,667) = 0,511$ +

(3)

$X_1 = 0,104F_1 + 0,824F_2 + U_1$
$X_2 = 0,065F_1 + 0,959F_2 + U_2$
$X_3 = 0,065F_1 + 0,725F_2 + U_3$
$X_4 = 0,906F_1 + 0,134F_2 + U_4$
$X_5 = 0,977F_1 + 0,116F_2 + U_5$
$X_6 = 0,827F_1 + 0,016F_2 + U_6$

Usual assumptions: 1) Mean indicator, common factor, unique factor = 0
2) Variance indicator, common factor = 1
3) $E(X,U) = 0$
4) $E(U_i, U_j) = 0$

$\text{corr}(F_1, F_2) = \phi_{12} = 0,3$

a) Pattern loadings: F_1 F_2

X_1 $\lambda_{11} = 0,104$ $\lambda_{12} = 0,824$

X_4 $\lambda_{41} = 0,906$ $\lambda_{42} = 0,134$

X_6 $\lambda_{61} = 0,827$ $\lambda_{62} = 0,016$ ok

b) $\text{corr}(X_1, X_2) = \lambda_{11}\cdot\lambda_{21} + \lambda_{12}\cdot\lambda_{22} + (\lambda_{11}\cdot\lambda_{22} + \lambda_{12}\cdot\lambda_{21})\phi_{12}$

$= 0,104 \cdot 0,065 + 0,824 \cdot 0,959 + (0,104 \cdot 0,959 + 0,065 \cdot 0,824) \cdot 0,3$

$\text{corr}(X_1, X_2) = 0,843$ ok

The correlation between the 2 indicators x_1 and x_2 is 0,8430.

c) $\text{Var}(X) = \lambda_1^2 + \lambda_2^2 + \underbrace{\text{Var}(U)}_{} + 2\lambda_1\lambda_2\phi$

The variance in X that is not accounted for by the common factors F_1 and F_2

- $\text{Var}(X_1) = 0,104^2 + 0,824^2 + \text{Var}(U_1) + 2 \cdot 0,104 \cdot 0,824 \cdot 0,3$

$\text{Var}(X_1) = 0,7412 + \text{Var}(U_1)$! $\text{Var}(X) = 1$ from assumption ② !

% of the variance not accounted for by F_1 and F_2 =

$\text{Var}(U_1) = 1 - 0,7412 = 0,2588 = \underline{25,9\%}$

- $\text{Var}(X_2) = 0,065^2 + 0,959^2 + \text{Var}(U_2) + 2 \cdot 0,065 \cdot 0,959 \cdot 0,3$

$\text{Var}(X_2) = 0,9613 + \text{Var}(U_2)$

% of the variance not accounted for by F_1 and F_2

$\text{Var}(U_2) = 1 - 0,9613 = 0,0387 = \underline{3,87\%}$

ok

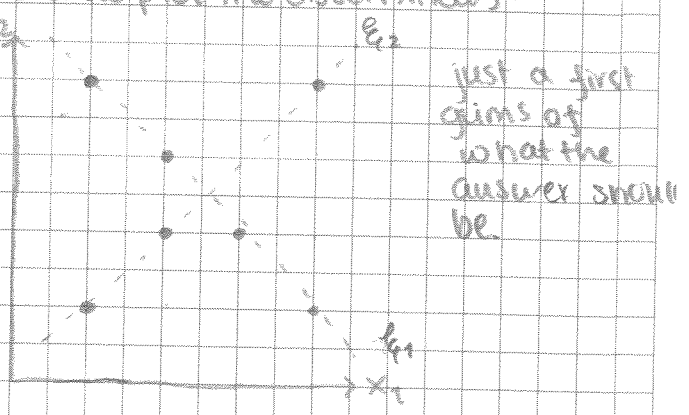
SU, STATISTIK

Skrivsal: Uppsala Anonymkod: MM-0022 Blad nr: 3

(A)

	X_1	X_2
1	1	4
2	1	1
3	2	2
4	3	3
5	3	2
6	3	2
7	1	1
8	1	4
Mean:	2,5	2,375

First I want to plot the observations:



just a first opinion of what the answer should be

I first want to create an SSCP-matrix:

$$SS_1^2 = \sum_{i=1}^{8}(X_{i1}-\bar{X}_1)^2 = \sum_{i=1}^{8}(X_{i1}-2,5)^2 = 10 \qquad SS_2^2 = \sum_{i=1}^{8}(X_{i2}-\bar{X}_2)^2 = \sum_{i=1}^{8}(X_{i2}-2,375)^2 = 9,875$$

$$SCP_{12} = \sum_{i=1}^{8}(X_{i1}-\bar{X}_1)(X_{i2}-\bar{X}_2) = -0,5$$

$$SSCP_{12} = \begin{pmatrix} 10 & -0,5 \\ -0,5 & 9,875 \end{pmatrix}$$

covariance matrix

$$\Sigma = \frac{SSCP}{df}_{n-1=7} = \begin{pmatrix} 1,43 & -0,07 \\ -0,07 & 1,41 \end{pmatrix}$$

I want to calculate the λ_i (eigenvalue). But how?

$$\xi_i = \gamma' X \qquad \Sigma = E(XX')$$

In principal component analysis I want to maximize the variance.

$$Var(\xi_i) = E(\xi_i'\xi_i) = E(\gamma' XX' \gamma) = \gamma' \Sigma \gamma$$

$$\max \gamma'\Sigma\gamma \quad s.t\ \gamma'\gamma = 1 \longrightarrow Z = \gamma\Sigma\gamma - \lambda(\gamma'\gamma - 1)$$

$$\frac{\partial Z}{\partial \gamma} = 2\gamma\Sigma - 2\lambda\gamma = 0$$

$$\gamma(\Sigma - I\lambda) = 0$$

$$|\Sigma - I\lambda| = 0$$ The characteristic equation!

→

4) $|\Sigma - \lambda I| = 0$

$$\left| \begin{pmatrix} 1,43 & -0,07 \\ -0,07 & 1,41 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \lambda \right| = 0$$

$$\begin{vmatrix} 1,43 - \lambda & -0,07 \\ -0,07 & 1,41 - \lambda \end{vmatrix} = (1,43 - \lambda)(1,41 - \lambda) + 0,0049 = 0$$

$$2,0163 - 1,43\lambda - 1,41\lambda + \lambda^2 + 0,0049 = 0$$

$$\lambda^2 - 2,84\lambda + 2,0114 = 0$$

$$\lambda_{1,2} = \left(\frac{2,84}{2}\right) \pm \sqrt{\left(\frac{2,84}{2}\right)^2 - 2,0114} = \left(\frac{2,84}{2}\right) \pm 0,0707$$

$\lambda_1 = 1,49$ $\lambda_2 = 1,35$ = THE EIGENVALUES.

we want to find the eigenvector for the PC that accounts for the maximum variance, i.e. $\lambda_1 = 1,49$.

$$(\Sigma - 1,49\, I)\mathbf{x} = 0$$

$$\begin{pmatrix} 1,43 & -0,07 \\ -0,07 & 1,41 \end{pmatrix} - \begin{pmatrix} 1,49 & 0 \\ 0 & 1,49 \end{pmatrix} \mathbf{x} = 0$$

$$\begin{pmatrix} -0,06 & -0,07 \\ -0,07 & -0,08 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$-0,06x_1 - 0,07x_2 = 0 \rightarrow x_1 = -\frac{0,07}{0,06} x_2 \rightarrow \underline{x_1 = 1 \quad x_2 = -0,8571}$

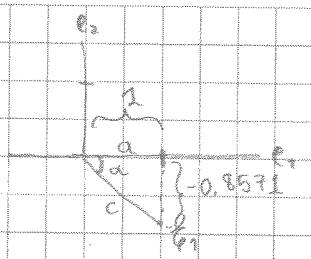
$-0,07x_1 - 0,08x_2 = 0$

$e_1 = (1, -0.8571)$

The eigenvector for the principal component that accounts for the maximum variance is $e_1 = (1; 0.8571)$

So, what is the angle between the new axis and the old axis?

(A). $e_1 = (1, -0.8571)$



Angle α?

$$\cos\alpha = \frac{a}{c} = \frac{1}{\sqrt{1^2 + 0{,}8571^2}} = 0{,}75927$$

$\alpha = 40{,}6°$

The angle between the new axis and the old axis is 40,6°.

This means that the new values will be:

$x_1^* = \cos\alpha \cdot x_1 + \sin\alpha \cdot x_2$

$x_2^* = -\sin\alpha \cdot x_1 + \cos\alpha \cdot x_2$

	X_1^*	X_2^*
1	3,36	2,39
2	1,41	0,11
3	2,82	0,22
4	3,47	0,98
5	3,58	-0,93
6	3,58	-0,43
7	3,69	-1,84
8	5,64	0,43

see covariance matrix

Total variance in the dataset = 1,43 + 1,41 = 2,84

The new axis explains (1,41 | 2,84) x 100 = 52,5 % of the total variance.

excellent

(5)

$$\begin{aligned} x_1 &= \lambda_1 \xi + \delta_1 \\ x_2 &= \lambda_2 \xi + \delta_2 \\ x_3 &= \lambda_3 \xi + \delta_3 \end{aligned} \qquad S = \begin{pmatrix} 1,2 & 0,93 & 0,45 \\ 0,93 & 1,56 & 0,27 \\ 0,45 & 0,27 & 2,15 \end{pmatrix}$$

$\lambda_1^2 + \text{Var}(\delta_1) = 1,2$ $\quad \lambda_2\lambda_1 = \lambda_1\lambda_2 = 0,93$

$\lambda_2^2 + \text{Var}(\delta_2) = 1,56$ $\quad \lambda_3\lambda_2 = \lambda_2\lambda_3 = 0,27$

$\lambda_3^2 + \text{Var}(\delta_3) = 2,15$ $\quad \lambda_3\lambda_1 = \lambda_1\lambda_3 = 0,45$

6 equations = 6 unknowns → JUST-IDENTIFIED! OK

We know the equations above since the covariance matrix is:

$$S = \begin{pmatrix} \underbrace{\lambda_1^2 + \text{Var}(\delta_1)}_{\text{var}(x_1)} & \underbrace{\lambda_1\lambda_2}_{\text{cov}(x_1,x_2)} & \underbrace{\lambda_1\lambda_3}_{\text{cov}(x_1,x_3)} \\ \underbrace{\lambda_2\lambda_1}_{\text{cov}(x_2,x_1)} & \underbrace{\lambda_2^2 + \text{Var}(\delta_2)}_{\text{var}(x_2)} & \underbrace{\lambda_2\lambda_3}_{\text{cov}(x_2,x_3)} \\ \underbrace{\lambda_3\lambda_1}_{\text{cov}(x_3,x_1)} & \underbrace{\lambda_3\lambda_2}_{\text{cov}(x_3,x_2)} & \underbrace{\lambda_3^2 + \text{Var}(\delta_3)}_{\text{var}(x_3)} \end{pmatrix}$$

where $\lambda_i\lambda_j$ are covariances ($\text{Cov}(i,j)$) and $\lambda_i + \text{Var}(\delta_i)$ is the variance of indicator i or ($\text{Cov}(i,i)$).

$$\left.\begin{aligned} \lambda_1 &= \frac{0,93}{\lambda_2} \\ \lambda_1 &= \frac{0,45}{\lambda_3} \end{aligned}\right\} \quad \frac{0,93}{\lambda_2} = \frac{0,45}{\lambda_3} \rightarrow 0,45\lambda_2 = 0,93\lambda_3 \rightarrow \lambda_2 = \frac{0,93\lambda_3}{0,45}$$

$\lambda_2\lambda_3 = 0,27 \rightarrow \left(\frac{0,93\lambda_3}{0,45}\right)\lambda_3 = 0,27 \rightarrow \frac{0,93}{0,45}\cdot\lambda_3^2 = 0,27 \rightarrow \lambda_3^2 = \frac{81}{620} \approx 0,1306$

$\lambda_3 \approx 0,3614$ OK $\quad \lambda_2 = \frac{0,93 \cdot 0,3614}{0,45} = 0,74690$ OK $\quad \lambda_1 = 1,2450$ OK

$1,2450^2 + \text{Var}(\delta_1) = 1,2$ $\quad 0,74690^2 + \text{Var}(\delta_2) = 1,56$ $\quad 0,3614^2 + \text{Var}(\delta_3) = 2,15$

$\text{Var}(\delta_1) = -0,3500$ $\quad \text{Var}(\delta_2) = 1,002$ $\quad \text{Var}(\delta_3) = 2,019$

→ How it can be negative?

$\lambda_1 = 1,2450$ $\quad \text{Var}(\delta_1) = -0,3500$

$\lambda_2 = 0,74690$ $\quad \text{Var}(\delta_2) = 1,002$

$\lambda_3 = 0,3614$ $\quad \text{Var}(\delta_3) = 2,019$

The solutions are unique since the equation is just-identified. (Estimates of the parameters.)

(6) i) Cluster analysis can be used when we want to reduce data. It is done by clustering observations that are closest to each other. CA can also be used if we want to see which observations are the most similar to each other by grouping them.

There is a hierarchical analysis which is what will be used in the calculations below. The clusters are computed by either centroid method, nearest-neighbor method, furthest-neighbor method or Ward's method. There's also non-hierarchical analysis where you know the amounts of clusters a priori.

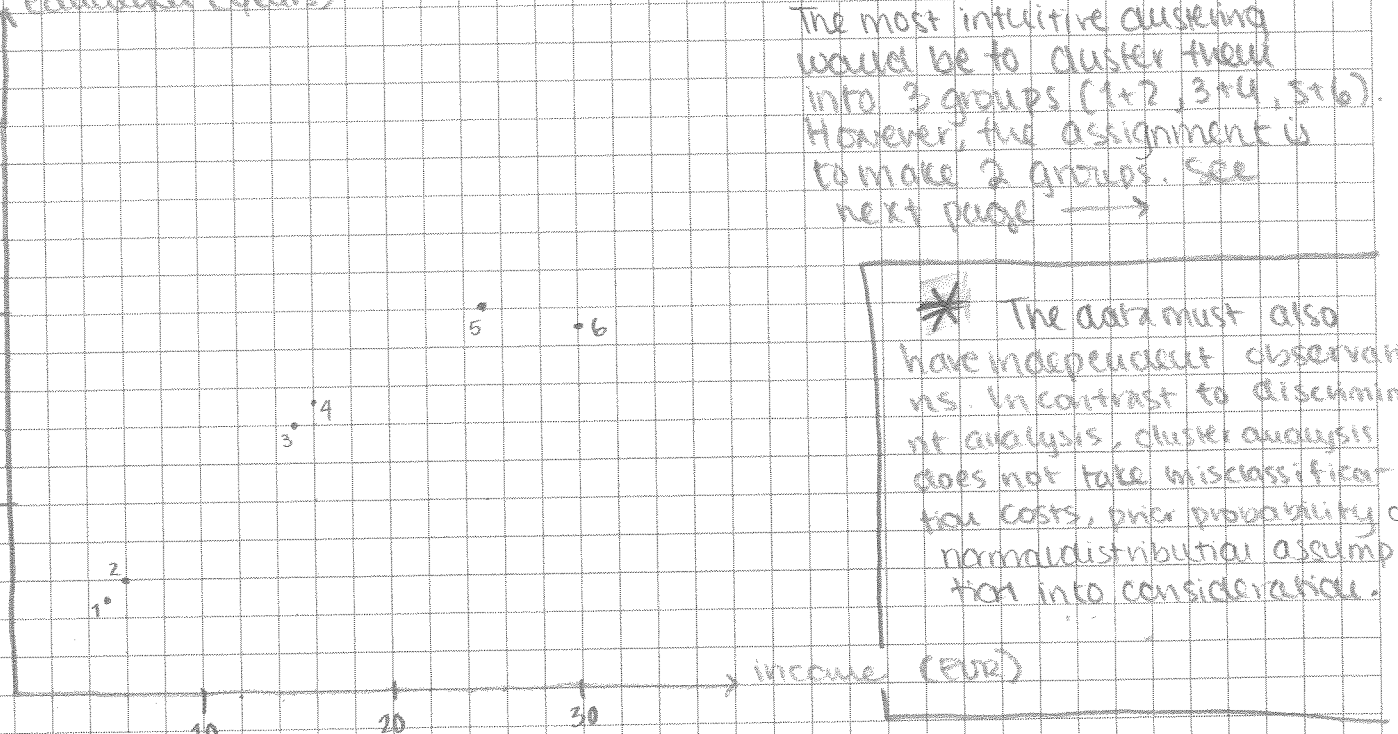
Assumptions for cluster analysis: *

1) Each group/cluster is homogeneous with respect to a specific characteristic

2) Each group/cluster is different from the other groups with respect to the same characteristic.

These assumptions should be very strict in my opinion. When clustering observations there should not be other factors that can bias the data since that would mean that one observation might be misplaced in the wrong cluster due to this unobservable.
If we are not strict, then the analysis fails since we can't say that the clusters differ significantly with respect to the observed variables. This would give poor conclusions and sample tests.

(ii) I first want to plot the data to get a visual of what answers I should get.



The most intuitive clustering would be to cluster them into 3 groups (1+2, 3+4, 5+6). However, the assignment is to make 2 groups. See next page →

* The data must also have independent observations. In contrast to discriminant analysis, cluster analysis does not take misclassification costs, prior probability or normal distributional assumption into consideration.

(cont.)

(6.) The distances are calculated by the squared Euclidean distance:

$$D^2_{ab} = \sum_{j=1}^{p} (a_j - b_j)^2$$

Similarity Matrix I

	S1	S2	S3	S4	S5	S6
S1	0	·	·	·	·	·
S2	(2)	0	·	·	·	·
S3	181	145	0	·	·	·
S4	221	181	(2)	0	·	·
S5	625	557	136	106	0	·
S6	821	745	250	212	26	0

The smallest distance is a tie between S1S2 and S3S4 where the distance is 2. These two observations will be clustered by the centroid method ⇒ the coordinates of the cluster = the average of the 2 clustered observations.

cluster S12: Income = $\frac{(6+5)}{2} = 5,5$ Education = $\frac{(6+5)}{2} = 5,5$

S34: Income = $\frac{(16+15)}{2} = 15,5$ Education = $\frac{(15+14)}{2} = 14,5$

New table

	Income	Educ
S12	5,5	5,5
S34	15,5	14,5
S5	25	20
S6	30	19

Similarity matrix II

	S12	S34	S5	S6
S12	0	·	·	·
S34	181	0	·	·
S5	590,5	120,5	0	·
S6	787,5	230,5	(26)	0

The smallest distance now is 26 so S5 and S6 will be clustered.

S56: Income: $\frac{(30+25)}{2} = 27,5$ Education: $\frac{(19+20)}{2} = 19,5$

→

(6.) New table:

	Income	Education
S12	5,5	5,5
S34	15,5	14,5
S56	29,5	19,5

Similarity Matrix

III :

	S12	S34	S56
S12	0	•	•
S34	181	0	•
S56	680	169	0

The smallest distance now is between S34 and S56.

Now we have the 2 clusters:

Cluster 1 = S1 and S2

Cluster 2 = S3, S4, S5, S6.

OK

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 23/10 -2014

Sal: Uggleviksalen

Tenta: Multivariata metoder

Kurs: Multivariata metoder

ANONYMKOD: MM-0016

☒ Jag godkänner att min tenta får läggas ut anonymt på hemsidan som studentsvar.

OBS! SKRIV ÄVEN PÅ BAKSIDAN AV SKRIVBLADEN

Markera besvarade uppgifter med kryss

1	2	3	4	5	6	7	8	9	Antal inl. blad
X	X	X	X	X	X				15
Lär.ant. 7	18	15	12	14	9				

POÄNG	BETYG	Lärarens sign.
75 MA+		

SU, STATISTIK

Skrivsal: Vaggeviksalen Anonymkod: MM-0016 Blad nr: 1

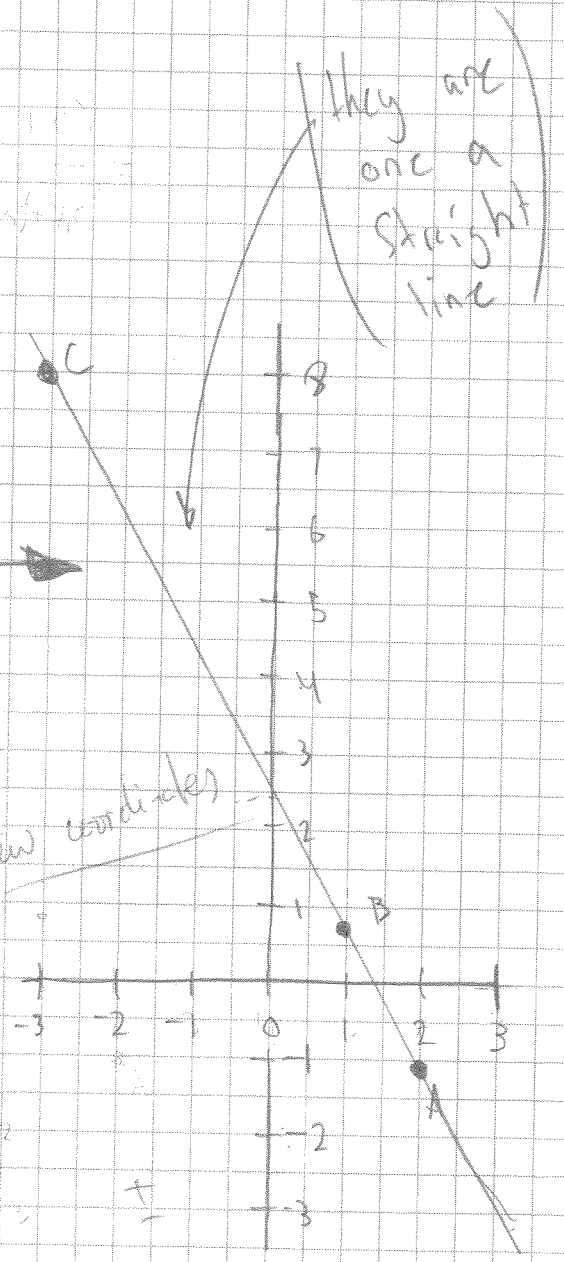
Question 1:

a) $A=(2;1)$ $B=(1;0,7)$ $C=(-3;7.5)$

$$\cos(\theta_{AB}) = \frac{a \cdot b}{\|a\| \cdot \|b\|} = \frac{(2 \cdot 1 + 0,7 \cdot 1)}{\sqrt{2^2+1^2}\sqrt{1^2+0,7^2}} = 1$$

$$\theta_{AB} = \cos^{-1}(1) = 0$$

Thus, A and B liz on a straight line, ther is no angel. Fom graph we see that they do.



b) If the distanz (Euclidian) do not change with the new coordinates then they lie on a straight line! No mor time to pawr...

SU, STATISTIK

Skrivsal: Vegetarianen Anonymkod: MM-006 Blad nr: 2

Question 2:

Blood. Ch	Heart disease P	A	Σ
< 200	6	5	11
200-219	10	6	16
220-254	30	5	35
≥ 254	45	7	52
Σ	91	23	114

a) What is the probability that:

(i) $P(P) = \frac{91}{114} \approx 0,7982 = 79,82\ \%$

(ii) $P(<219) = \frac{(11+16)}{114} \approx 0,2368 = 23,68\ \%$

b) What are the odds that:

$$\text{odds} = \frac{P}{1-P}$$

(i) $\text{odds}(P \mid <200) = \frac{6}{5} = 1,2$

(ii) $\text{odds}(>219 \mid P) = \frac{(30+45)}{(6+10)} = 4,6875$

c) Compute the log odds:

(i) $\ln[\text{odds}(P \mid >254)] = \ln\left(\frac{45}{7}\right) \approx 1,8609$

(ii) $\ln[\text{odds}(P \mid 200-219)] = \ln\left(\frac{10}{6}\right) \approx 0,5108$

Question 3:

6-indicator 2-factor model!

$$X_1 = 0,104 \times F_1 + 0,824 \times F_2 + U_1$$
$$X_2 = 0,065 \times F_1 + 0,459 \times F_2 + U_2$$
$$X_3 = 0,065 \times F_1 + 0,725 \times F_2 + U_3$$
$$X_4 = 0,906 \times F_1 + 0,134 \times F_2 + U_4$$
$$X_5 = 0,977 \times F_1 + 0,116 \times F_2 + U_5$$
$$X_6 = 0,827 \times F_1 + 0,016 \times F_2 + U_6$$

$\text{Corr}(F_1, F_2) = \phi_{12} = 0,3$

a) What are the pattern loadings of indicators X_1, X_4 and X_6 on the factors F_1 and F_2?

The pattern loading (λ_{ij}) for a common factor (F_j) is simply the coefficient for that common factor.

Patter loading of indicator (i) and common factor (j) = λ_{ij}

X_1: $\lambda_{1,1} = 0,104$ $\lambda_{1,2} = 0,824$

X_4: $\lambda_{4,1} = 0,906$ $\lambda_{4,2} = 0,134$

X_6: $\lambda_{6,1} = 0,827$ $\lambda_{6,2} = 0,016$

Next page

b) Compute the correlation between X_1 and X_2.

$$Corr(X_1, X_2) = \lambda_{1,1}\lambda_{2,1} + \lambda_{1,2}\lambda_{2,2} + \left[\lambda_{1,1}\lambda_{2,2} + \lambda_{1,2}\lambda_{2,1}\right]\phi_{1,2} =$$

$$= 0{,}104 \times 0{,}065 + 0{,}824 \times 0{,}959 + \left[0{,}104 \times 0{,}959 + 0{,}824 \times 0{,}065\right] \times 0{,}3$$

$$\approx 0{,}8430$$

c) What percentage of the variance of indicators X_1 and X_2 is not accounted for by the common factor F_1 and F_2?

$$Var(X_i) = \underbrace{\lambda_{i,1}^2 + \lambda_{i,2}^2 + 2(\lambda_{i,1})(\lambda_{i,2})\phi_{1,2}}_{\text{Variance from factors}} + \underbrace{Var(U_i)}_{\text{variance of unique factor}}$$

$$\Leftrightarrow$$

$$Var(U_i) = \underbrace{Var(X_i)}_{\text{is equal to 1 by assumption 2.}} - \left[\lambda_{i,1}^2 + \lambda_{i,2}^2 + 2(\lambda_{i,1})(\lambda_{i,2})\phi_{1,2}\right]$$

$$\Leftrightarrow$$

$$Var(U_i) = 1 - \left[\lambda_{i,1}^2 + \lambda_{i,2}^2 + 2(\lambda_{i,1})(\lambda_{i,2})\phi_{1,2}\right]$$

X_1: $Var(U_1) = 1 - \left[(0{,}104)^2 + (0{,}824)^2 + 2(0{,}104)(0{,}824)(0{,}3)\right] \approx 0{,}2588 = 25{,}88\%$

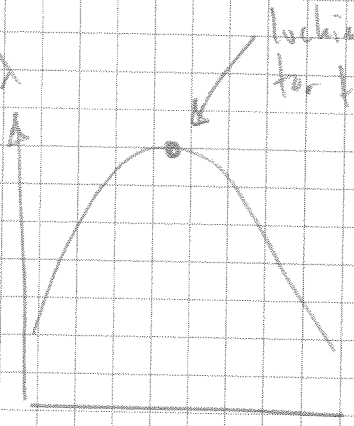
X_2: $Var(U_2) = 1 - \left[(0{,}065)^2 + (0{,}959)^2 + 2(0{,}065)(0{,}959)(0{,}3)\right] \approx 0{,}0387 = 3{,}87\%$

Question 4:

Finding the angle θ, so that the first principal component (new variable) accounts for the maximum variance of the total variance of the data can be obtained by the equation:

(1) $\boxed{X_1^* = \cos(\theta) \times X_1 + \sin(\theta) \times X_2}$

Then one changes the angle θ until a maximum of λ is reached, where $\lambda = \frac{\sigma^2_{X_1^*}}{\sigma^2_{X_1} + \sigma^2_{X_2}}$. The problem here is to do all the computations.



I will instead compute the covariance matrix, and compute the eigenvalues and eigen vectors to get the θ directly and confirm this point is the maximum by using equation (1) two times for angels θ close to θ_{max}.

a) Realize that covariance matrix is sufficient due to the fact that the data is in the same units.

Obs	X_1	X_2
1	1	4
2	1	1
3	2	2
4	2	3
5	3	2
6	3	2
7	4	1
8	4	4

$$\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ik} \qquad \sigma_i^2 = \frac{1}{(n-1)}\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)^2$$

$$\sigma_{ij} = \frac{1}{(n-1)}\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)$$

$$\bar{X}_1 = \frac{1}{8}\left[2\times1+2\times2+2\times3+2\times4\right] = 2{,}5$$

$$\sigma_1^2 = \frac{1}{7}\left[2(1-2{,}5)^2+2(2-2{,}5)^2+2(3-2{,}5)^2+2(4-2{,}5)^2\right] \approx 1{,}43$$

$$\bar{X}_2 = \frac{1}{8}\left[2\times4+2\times1+3\times2+3\right] = 2{,}375$$

$$\sigma_2^2 = \frac{1}{7}\left[2(4-2{,}375)^2+2(1-2{,}375)^2+3(2-2{,}375)^2+(3-2{,}375)^2\right] = 1{,}41$$

$$\sigma_{12} = \frac{1}{7}\left[(1-2{,}5)(4-2{,}375)+(1-2{,}5)(1-2{,}375)+(2-2{,}5)(2-2{,}375)+\right.$$
$$+(2-2{,}5)(3-2{,}375)+(3-2{,}5)(2-2{,}375)+(3-2{,}5)(2-2{,}375)+$$
$$\left.+(4-2{,}5)(1-2{,}375)+(4-2{,}5)(4-2{,}375)\right] = -0{,}07$$

$$\hat{\Sigma} = \begin{pmatrix} 1{,}43 & -0{,}07 \\ -0{,}07 & 1{,}41 \end{pmatrix}$$

- Variances are similar which justify the use of the covariance matrix in PCA.

Really low covariance and thereby correlation between X_1 and X_2. →

SU, STATISTIK

Skrivsal: Ugglevägen Anonymkod: MM-006 Blad nr: 7

Find eigenvalues λ_i:

$$|\Sigma - \lambda I| = 0$$

$$\left| \begin{pmatrix} 1,43 & -0,07 \\ -0,07 & 1,41 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \right| = 0$$

$$\begin{vmatrix} 1,43-\lambda & -0,07 \\ -0,07 & 1,41-\lambda \end{vmatrix} = 0$$

$$(1,43-\lambda)(1,41-\lambda) - (-0,07)^2 = 0$$

$$\lambda^2 - 2,84\lambda + 2,0114 = 0$$

$$\left(\lambda - \frac{2,84}{2}\right)^2 = \left(\frac{2,84}{2}\right)^2 - 2,0114$$

$$(\lambda - 1,42)^2 = 0,0050$$

$$\lambda = 1,42 \pm \sqrt{0,0050}$$

$$\boxed{\begin{aligned} \lambda_1 &\approx 1,49 \\ \lambda_2 &\approx 1,35 \end{aligned}}$$

Principal component 1 accounts for

$$\frac{1,49}{(1,49+1,35)} = 0,5246 \quad 52,46\ \%$$

of the total variance in the data.

This is really low, not good for PCA analysis.

Find the eigen vectors for λ_1:

$$(\Sigma - \lambda_1 I)\vec{w}_1 = \vec{0}$$

$$\left[\begin{pmatrix} 1,43 & -0,07 \\ -0,07 & 1,41 \end{pmatrix} - \begin{pmatrix} 1,429 & 0 \\ 0 & 1,429 \end{pmatrix}\right]\begin{bmatrix} w_{11} \\ w_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\left[\begin{array}{cc|c} 0,06 & -0,07 & 0 \\ -0,07 & 0,08 & 0 \end{array}\right] \times \left(-\frac{7}{6}\right) \rightsquigarrow \left[\begin{array}{cc|c} -0,06 & -0,07 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

$$\Rightarrow \begin{cases} -0,06 w_{11} - 0,07 w_{12} = 0 \\ w_{12} = t_1, \text{ where } t_1 \in \mathbb{R} \end{cases}$$

(all real numbers)

(thus there are infinite numbers of solutions).

- Set $t_1 = \frac{1}{\sqrt{2}}$

$$w_{11} = -\frac{7}{6} w_{12} = -\frac{7}{6} \times \frac{1}{\sqrt{2}} \approx -0,825$$

$$w_{12} = \frac{1}{\sqrt{2}} \approx 0,707$$

$$\cos(\theta) = -0,825$$

$$\theta = \frac{\cos^{-1}(-0,825) \times 180}{\pi} \approx 34,4^\circ \text{ C}$$

Klar!

Conformation of 34,4 as max θ.

Case 1: $\theta = 30°$

$$X_{i1}^* = \cos\left(\frac{30° \times \pi}{180}\right) \times X_1 + \sin\left(\frac{30° \times \pi}{180}\right) \times X_2 =$$

$$= 0{,}866 \times X_1 + 0{,}5 X_2 = \begin{bmatrix} 2{,}87 \\ 1{,}37 \\ 2{,}73 \\ 3{,}23 \\ 3{,}60 \\ 3{,}60 \\ 3{,}96 \\ 5{,}46 \end{bmatrix}$$

$$\bar{X}_{i1}^* = \frac{1}{8}\left[2{,}87 + 1{,}37 + 2{,}73 + 3{,}23 + 3{,}60 + 3{,}60 + 3{,}96 + 5{,}46\right] = 3{,}3525$$

$$\sigma_{\theta}^2 = \frac{1}{7}\left((2{,}87 - \bar{X}_1^*)^2 + \ldots + (5{,}46 - \bar{X}_1^*)^2\right) = 1{,}36$$

Case 2: $\theta = 40$

$$X_{i2}^* = \cos\left(\frac{40° \pi}{180}\right) X_1 + \sin\left(\frac{40° \pi}{180}\right) X_2 = 0{,}766 \times X_1 + 0{,}643 \times X_2 = \begin{bmatrix} 3{,}34 \\ 1{,}41 \\ 2{,}82 \\ 3{,}46 \\ 3{,}88 \\ 3{,}58 \\ 3{,}70 \\ 5{,}62 \end{bmatrix}$$

$$\bar{X}_{i2}^* = \frac{1}{8}\left[3{,}34 + \ldots + 5{,}62\right] = 3{,}48375$$

$$\sigma_{i2}^2 = \frac{1}{7}\left[(3{,}34 - \bar{X}_{i2}^*) + \ldots + (5{,}62 - \bar{X}_{i2}^*)\right] = 1{,}34$$

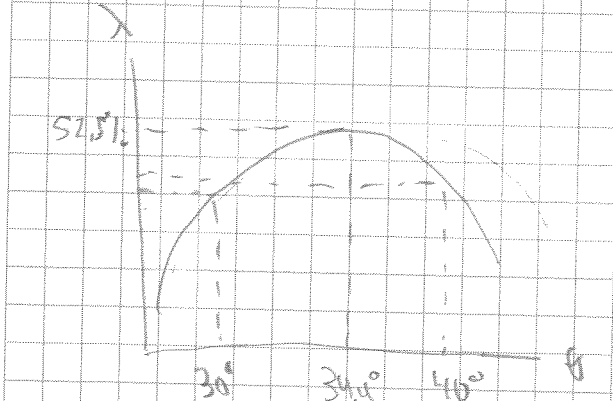
Nästa sida

Thus, we have obtained three points to locate the maximum angle θ:

1) $\theta = 30°$: $\frac{\sigma_{c1}^*}{\sigma_1^2 + \sigma_2^2} = \frac{1,36}{(1,43+1,41)} \approx 0,479 = 47,9\%$

2) $\theta^{max} = 34,4°$: $\frac{\sigma_{max}}{\sigma_1^2 + \sigma_2^2} = \frac{\lambda_1}{\sigma_1^2 + \sigma_2^2} = \frac{1,449}{(1,43+1,41)} \approx 0,525 = 52,5\%$

3) $\theta = 40°$: $\frac{\sigma_{c2}^*}{\sigma_1^2 + \sigma_2^2} = \frac{1,34}{1,43+1,41} = 0,472 = 47,2\%$



However we can conclude that only accounting for 52,5% is not a good principal component. In this data PCA should not be used due to low correlation between x_1 and x_2.

TG!

Question 5:

$$\begin{cases} x_1 = \lambda_1 \xi + \delta_1 \\ x_2 = \lambda_2 \xi + \delta_2 \\ x_3 = \lambda_3 \xi + \delta_3 \end{cases} \qquad S = \begin{pmatrix} 1,20 & 0,43 & 0,45 \\ 0,43 & 1,56 & 0,27 \\ 0,45 & 0,27 & 2,15 \end{pmatrix}$$

Are the parameter estimates unique?

$$\hat{S} = \begin{bmatrix} \lambda_1^2 + \mathrm{Var}(\delta_1) & & \\ \lambda_1\lambda_2 & \lambda_2^2 + \mathrm{Var}(\delta_2) & \\ \lambda_1\lambda_2 & \lambda_2\lambda_3 & \lambda_3^2 + \mathrm{Var}(\delta_3) \end{bmatrix}$$

This matrix is symmetric and I only need to consider the unique elements

We can conclude that there are 6 unique equations in the system and there is also 6 unique parameters to estimate. Because these are equal ($6 = 6$) there will exist one solution for the system that is unique (the degrees of freedom are equal to zero, $6 - 6 = 0$), the system is just-identified!!

The unique parameter estimates are obtained on the next page!

(1) $Var(x_i) = \lambda_i^2 + Var(\delta_i)$

(2) $Cov(x_i, x_j) = \lambda_i \lambda_j$

The elements of these are in the matrices S and Λ.

$$\begin{cases} 1,20 = \lambda_1^2 + Var(\delta_1) \\ 1,56 = \lambda_2^2 + Var(\delta_2) \\ 2,15 = \lambda_3^2 + Var(\delta_3) \\ 0,93 = \lambda_1 \lambda_2 \\ 0,27 = \lambda_2 \lambda_3 \\ 0,415 = \lambda_1 \lambda_3 \end{cases} \Rightarrow \begin{cases} Var(\delta_1) = 1,20 - \lambda_1^2 & (1) \\ Var(\delta_2) = 1,56 - \lambda_2^2 & (2) \\ Var(\delta_3) = 2,15 - \lambda_3^2 & (3) \\ \lambda_1 = \frac{0,93}{\lambda_2} & (4) \\ \lambda_2 = \frac{0,27}{\lambda_3} & (5) \\ \lambda_3 = \frac{0,415}{\lambda_1} & (6) \end{cases}$$

Step 1: Sub. (5) and (6) into (4).

$$\lambda_1 = \frac{0,93}{\lambda_2} = \frac{0,93}{\left(\frac{0,27}{\lambda_3}\right)} = \frac{0,93}{\left(\frac{0,27}{\left(\frac{0,415}{\lambda_1}\right)}\right)} = \frac{1,55}{\lambda_1} \Leftrightarrow \lambda_1^2 = 1,55$$

$$\lambda_1 = \sqrt{1,55} \approx 1,2450$$

Step 2: Solve (6).

$$\lambda_3 = \frac{0,415}{\lambda_1} = \frac{0,415}{\sqrt{1,55}} \approx 0,3614$$

Step 3: Solve (5).

$$\lambda_2 = \frac{0,27}{\lambda_3} = \frac{0,27}{0,3614} \approx 0,74711$$

Step 4: Solve (1), (2) and (3)

$Var(\delta_1) = 1,20 - 1,55 = -0,35$

$Var(\delta_2) = 1,56 - (0,74711)^2 \approx 1,0018$

$Var(\delta_3) = 2,115 - (0,3614)^2 \approx 1,5418$

Solutions:

$\lambda_1 = 1,2450$

$\lambda_2 = 0,74711$

$\lambda_3 = 0,3614$ negative loading?

$Var(\delta_1) = (-0,3500)$

$Var(\delta_2) = 1,0018$

$Var(\delta_3) = 1,5418$

(OK)

Question 6:

i) Cluster analysis intend accomplish data reduction by grouping observation into groups ("clusters") that share similar characteristics, and then only analyze these groups. The similarity of characteristics is measured by distance between variables.

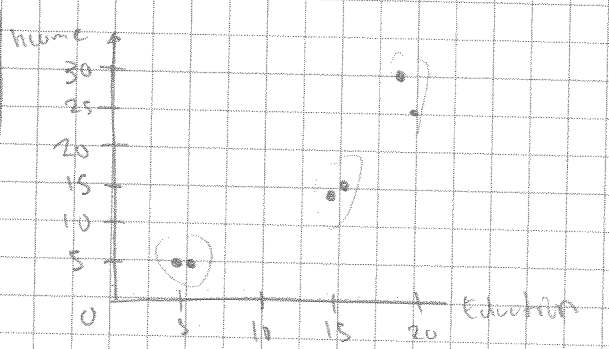
There are two types of clustering categories; hierarchical and non-hierarchical. Within the hierarchical there are the centroid, single linkage, complete linkage, average method and Ward's method to combine observations to clusters.

The assumptions are that; the data is multivariate normal, the groups have equal covariance matrices and that the observations are independent.

One have to have all these fulfilled to be able to trust and draw conclusions from the clusters, and if one wants to compute test statistics. It might be easier to cluster a variable on a categorical variable but the data should still be multivariate normal.

9. 11) Squared Euclidian distance: $D^2_{AB} = \sum_{k=1}^{p} (a_k - b_k)^2$

The intermediate similarity matrix consist of the squared Euclidian distances between observations, and thereby a zero diagonal.



- Visual inspection do data suggest three clusters.

$D^2_{S1S2} = (5-6)^2 + (5-6)^2 = 1$

$D^2_{S1S3} = (5-15)^2 + (5-14)^2 = 181$

$D^2_{S1S4} = (5-16)^2 + (5-15)^2 = 221$

$D^2_{S1S5} = (5-25)^2 + (5-20)^2 = 625$

$D^2_{S1S6} = (5-30)^2 + (5-14)^2 =$

First column

$D^2_{S2S3} = (6-15)^2 + (6-14)^2 = 145$

$D^2_{S2S4} = (6-16)^2 + (6-15)^2 = 181$

$D^2_{S2S5} = (6-25)^2 + (6-20)^2 = 557$

$D^2_{S2S6} = (6-30)^2 + (6-14)^2 = 745$

Second column, and so on...

Until we get the intermediate similarity matrix

$S^{(1)} =$

	S_1	S_2	S_3	S_4	S_5	S_6
S_1	0					
S_2	1	0				
S_3	181	145	0			
S_4	221	181	1	0		
S_5	625	557	136	106	0	
S_6	221	745	250	212	26	0

Symmetry

Now we start the hierarchical clustering by the single linkage method, i.e. we choose the minimum distance of the two possibilities in the similarity matrix when forming a cluster. But we face a problem at the first step, should we start to cluster S_1S_2 or S_3S_4? They have the same squared Euclidian distance, I will start by S_1S_2 (it is convinient and they did it in the book).

$$S_{(step\,1)} = \begin{array}{c} \\ S_1S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} \begin{array}{c} \begin{array}{ccccc} S_1S_2 & S_3 & S_4 & S_5 & S_6 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 145 & 0 & & & \\ 181 & (1) & 0 & & \\ 557 & 136 & 106 & 0 & \\ 745 & 250 & 212 & 26 & 0 \end{bmatrix} \end{array}$$

"Note that 1 have then the minimum value of the "red" lines of 5 rows/cols ex $145 = \text{Min}(181, 145)$"

$$\Rightarrow \quad S_{(step\,2)} = \begin{array}{c} \\ S_1S_2 \\ S_3S_4 \\ S_5 \\ S_6 \end{array} \begin{array}{c} \begin{array}{ccccc} S_1S_2 & S_3S_4 & S_5 & S_6 & \end{array} \\ \begin{bmatrix} 0 & & & & \\ 145 & 0 & & & \\ 557 & 106 & 0 & & \\ 745 & 212 & (26) & 0 & \end{bmatrix} \end{array}$$

$$\Rightarrow \quad S_{(step\,3)} = \begin{array}{c} \\ S_1S_2 \\ S_3S_4 \\ S_5S_6 \end{array} \begin{array}{c} \begin{array}{ccc} S_1S_2 & S_3S_4 & S_5S_6 \end{array} \\ \begin{bmatrix} 0 & & \\ 145 & 0 & \\ 557 & 106 & 0 \end{bmatrix} \end{array}$$

I would stop here and use three clusters, as the plot and the increase in distance, from [1,26] to 106, indicates but it if it is required.

$$\Rightarrow \quad S_{(step\,4)} = \begin{array}{c} \\ S_1S_2 \\ (S_3S_4)(S_5S_6) \end{array} \begin{array}{c} \begin{array}{cc} S_1S_2 & (S_3S_4)(S_5S_6) \end{array} \\ \begin{pmatrix} 0 & \\ 145 & 0 \end{pmatrix} \end{array}$$

2 cluster solution: S_1S_2 and $S_3S_4S_5S_6$

OK, some calculation mistakes