Stockholms universitet

STATISTISKA INSTITUTIONEN HT-2017

Written exam in Multivariate Methods, 7.5 ECTS credits
Tuesday, 26th September 2017, 10:00 – 15:00
Time allowed: FIVE hours
Examination Hall: Uggleviksalen

You are expected 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 computer lab assignment. The final grades are assigned as follows: **A** (91+), **B** (81-90), **C** (71-80), **D** (61-70), **E** (51-60), **Fx** (30-49), and **F** (0-29)

You are **allowed** to use a pocket (graphical) calculator, a language dictionary, and a list of formulas (attached)

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

1. (15 points)
Let us analyse the following 3-variate dataset with 6 (six) observations. Each observation consists of 3 measurements and recorded in the following matrix

Kolumn1	Kolumn2	Kolumn3
8	5	7
7	2	9
5	3	3
9	5	8
7	4	5
8	2	2

What portion of total variance each variable accounts for? Compute the correlation matrix. Next, find eigenvalues of the correlation matrix and interpret them in style of PCA. How much of each of the three variables, the two first principal components "explain"? Have you mean-adjusted and/or standardized the original data set before the analysis: why yes/no?

2. (10 points)

(a) Points A and B have the following coordinates with respect to orthogonal axes X_1 and X_2: $A=(3,-2)$; $B=(5,1)$. If the axes X_1 and X_2 are rotated 50^o clockwise to produce a new set of orthogonal axes X_1^* and X_2^*, find the coordinates of A and B with respect to X_1^* and X_2^*.

(b) Coordinates of a point A with respect to an orthogonal set of axes X_1 and X_2 are $(5,2)$. The axes X_1 and X_2 are rotated clockwise by an angle θ. If the new coordinates of the point A with respect to the rotated axes are $(3.69, 3.939)$, find θ.

3. (10 points)
Perform principal component analysis on the following data by hand. In other words, determine the angle (with precision up to 5%: 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? Discuss assumptions behind PCA applied to below given data

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

4. (15 points)
Cluster the following hypothetical data set into two groups using complete linkage method and the associated similarity matrixes. Would you get the same answer if the data were reported in EUR and exchange rate 10SEK for 1EUR? Explain your answer.

Moreover. perform one step of Ward's method as well as Algorithm I (II) of non-hierarchical clustering to show that you know the procedures but, at the same time, trying to keep calculations as minimal as you can using below stated data set.

Subject ID	Income (tSEK)	Education (in years)
S1	250	12
S2	300	14
S3	280	16
S4	330	18
S5	360	24
S6	400	20

5. (15 points) The correlation matrix for a hypothetical data set is given in the following table:

	X_1	X_2	X_3	X_4
X_1	1.000			
X_2	0.7	1.000		
X_3	0.3	0.25	1.000	
X_4	0.35	0.2	0.6	1.000

The following estimated factor loadings were extracted by the principal axis factoring procedure:

Variable	F_1	F_2
X_1	0.90	0.20
X_2	0.70	0.15
X_3	0.20	0.90
X_4	0.20	0.70

Compute and discuss the following: (a) specific variances: what high specific variance indicates? Explain using data above; (b) communalities and % of shared variance; interpret both; (c) proportion of variance explained by each factor, what can you say about chosen factors? (d) Estimated or reproduced correlation matrix; how good is the estimate? Discuss; and (e) residual matrix, compute RMSR and interpret. Can you suggest how to improve (if necessary) the model on the basis of your findings?

6. (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) = Φ_{12} = - 0.2. Repeat all your calculations in assumption that correlation changed to Corr(F_1, F_2) = Φ_{12} = 0.2 and discuss the differences in detail. Try to provide intuition for at least some of your answers: without calculating, what you would expect in case correlation is 1, 0.9, 0.5, 0, -0.5, -0.9, -1?

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

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

(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?

GOOD LUCK

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=$\begin{vmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{vmatrix}$
detA=$a_{i1}A_{i1} + a_{i2}A_{i2} + \dots + a_{in}A_{in}$ 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 $b_1, \dots, b_n$. (**Ax=b**)

Vectors

$\boldsymbol{a} = (a_1 a_2 \dots a_p)$
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 $\boldsymbol{ab} = a_1b_1 + a_2b_2 + \dots + a_pb_p$; $\boldsymbol{ab} = \|\boldsymbol{a}\|\|\boldsymbol{b}\|\cos\alpha$
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
Mahalanobis: $MD_{ik}^2 = \frac{1}{1-r^2}\left[\frac{(x_{i1}-x_{k1})^2}{s_1^2} + \frac{(x_{i2}-x_{k2})^2}{s_2^2} - \frac{2r(x_{i1}-x_{k1})(x_{i2}-x_{k2})}{s_1 s_2}\right]$

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; λ-eigenvalues; $|\boldsymbol{\Sigma} - \lambda\boldsymbol{I}| = \boldsymbol{0}$; γ-eigenvector; $(\boldsymbol{\Sigma} - \lambda\boldsymbol{I})\boldsymbol{\gamma} = \boldsymbol{0}$; $\boldsymbol{\gamma}'\boldsymbol{\gamma} = \boldsymbol{1}$;

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 $E(\varepsilon_i\varepsilon_j) = 0$

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: $\Sigma = \Sigma(\theta)$ H_a: $\Sigma \neq \Sigma(\theta)$ (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: $Z = w_1 x_1 + w_2 x_2$

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

Σ-variance-covariance matrix, **T**-total SSCP matrix. γ-vector of weights.

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

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

$|W^{-1}B - \lambda I| = 0;\ \gamma = \Sigma^{-1}(\mu_1 - \mu_2)$ - 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:

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

Added formulas:

RMSR = $\sqrt{\frac{\sum_{i=1}^{p}\sum_{j=i}^{p} res_{ij}^2}{p(p-1)/2}}$, where res_{ij} is the correlation matrix between the ith and jth var, p is number of variables.

Cubic equations:

There is an analogous formula for polynomials of degree three:
$ax^3 + bx^2 + cx + d = 0$ is:

$$x = \sqrt[3]{\left(\frac{-b^3}{27a^3} + \frac{bc}{6a^2} - \frac{d}{2a}\right) + \sqrt{\left(\frac{-b^3}{27a^3} + \frac{bc}{6a^2} - \frac{d}{2a}\right)^2 + \left(\frac{c}{3a} - \frac{b^2}{9a^2}\right)^3}} + \sqrt[3]{\left(\frac{-b^3}{27a^3} + \frac{bc}{6a^2} - \frac{d}{2a}\right) - \sqrt{\left(\frac{-b^3}{27a^3} + \frac{bc}{6a^2} - \frac{d}{2a}\right)^2 + \left(\frac{c}{3a} - \frac{b^2}{9a^2}\right)^3}} - \frac{b}{3a}.$$

Department of Statistics

Stockholms universitet

Correction sheet

Date: 26/09/2017

Room: Uggleviksal

Exam: Multivariate methods (eng)

Course: Multivariate methods (eng)

Anonymous code: MM0006

☒ I authorise the anonymous posting of my exam, in whole or in part, on the department homepage as a sample student answer.

NOTE! ALSO WRITE ON THE BACK OF THE ANSWER SHEET

Mark answered questions

1	2	3	4	5	6	7	8	9	Total number of pages
X	X	X	X	X	X				08 12
Teacher's notes 15	6	7	9	6	10				

Points	Grade	Teacher's sign.
53 +13 (66)	1	AA

SU, DEPARTMENT OF STATISTICS

Room: D61 Anonymous code: MM0006 Sheet number: 1

(1) Let X = Kolumn 1
Y = Kolumn 2
Z = Kolumn 3

$\bar{X} = (8+7+\dots+8)/16 = 7,\overline{33}$

$\bar{Y} = (5+2+\dots+9)/16 = 3,5$

$\bar{Z} = (7+9+\dots+2)/16 = 5,\overline{66}$

$X-\bar{X}$	$Y-\bar{Y}$	$Z-\bar{Z}$
0,66	1,5	1,33
-0,33	-1,5	3,33
-2,33	-0,5	-2,66
1,66	1,5	2,33
-0,33	0,5	-0,66
0,66	-1,5	-3,66

$var(X) = \sum_i^6 (x_i - \bar{x})^2 / 6 = 1,866$

$var(Y) = \quad = 1,9$ OK

$var(Z) = \quad = 7,866$

$tot\ var = var(X) + var(Y) + var(Z) = 11,633$

* Now that we have each of variances, we can calculate the portion of total variance

$X: \frac{var(X)}{tot\ var} = \frac{1,866}{11,633} = 16,0\ \%$

$Y: \frac{var(Y)}{tot\ var} = \frac{1,9}{11,633} = 16,3\ \%$

$Z: \frac{var\ Z}{tot\ var} = \frac{7,866}{11,633} = 67,6\ \%$ OK

→

* Correlation Matrix

$$SSCP = \begin{pmatrix} \sum (X_i - \bar{X})^2 & \sum (X_i - \bar{X})(Y_i - \bar{Y}) & \sum (X_i - \bar{X})(Z_i - \bar{Z}) \\ \sum (X_i - \bar{X})(Y_i - \bar{Y}) & \sum (Y_i - \bar{Y})^2 & \sum (Y_i - \bar{Y})(Z_i - \bar{Z}) \\ \sum (X_i - \bar{X})(Z_i - \bar{Z}) & \sum (Y_i - \bar{Y})(Z_i - \bar{Z}) & \sum (Z_i - \bar{Z})^2 \end{pmatrix}$$

$$SSCP = \begin{pmatrix} 9,333 & 24 & 7,666 \\ 24 & 9,5 & 7 \\ 7,666 & 7 & 39,333 \end{pmatrix}$$

$$\Sigma = SSCP / (df = n-1) = \begin{pmatrix} 1,866 & 0,8 & 1,533 \\ 0,8 & 1,9 & 1,4 \\ 1,533 & 1,4 & 7,8666 \end{pmatrix}$$

$$C = \begin{pmatrix} 1 & 0,4248 & 0,40 \\ 0,4248 & 1 & 0,3621 \\ 0,40 & 0,3621 & 1 \end{pmatrix} \leftarrow \text{Correlation Matrix}$$

or

* Calculate the eigen values

$$\det(C - \lambda I) = 0 \rightarrow \det \begin{pmatrix} 1-\lambda & 0,4248 & 0,4 \\ 0,4248 & 1-\lambda & 0,3621 \\ 0,40 & 0,3621 & 1-\lambda \end{pmatrix} = 0$$

$$\det 0,4248 = A$$
$$0,40 = B$$
$$0,3621 = C$$

$$1-\lambda\left((1-\lambda)^2 - C^2\right) - A(A(1-\lambda) - BC) + B(AC - B(1-\lambda)) = 0$$
$$(1-\lambda)^3 - (1-\lambda)(C^2 + B^2 + A^2) + 2ABC = 0$$
$$(1-\lambda)^3 - 0,4716(1-\lambda) + 0,123056$$

By solving third qubic equation, we get following result:

$\lambda_1 = 0,567$
$\lambda_2 = 0,641$
$\lambda_3 = 1,79$

$\lambda_1 + \lambda_2 + \lambda_3 = 2,998$

λ_3 explain 59,7 %

λ_2 explain 21,4 % →

SU, DEPARTMENT OF STATISTICS

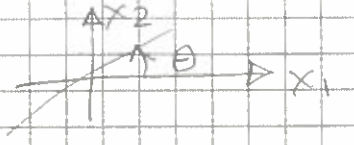
Room: 106 Anonymous code: XM0006 Sheet number: 2

Commenting ex 1

I mean-adjusted and standardized the data because I wanted that each observation affected the outcome equally.

EM

(2) (a) $A = (3, -2)$ $B = (5, 1)$ with respect to X_1 and X_2 and we know that $X_1 \perp X_2$. We rotate the cordinat system $(360° - 50°) = \theta$. We want to know the new positions



Point A

$X_1^* = 3 \cos 310° - 2 \sin 310° = 3,46$

$X_2^* = -3 \sin 310° - 2 \cos 310° = 1,01$

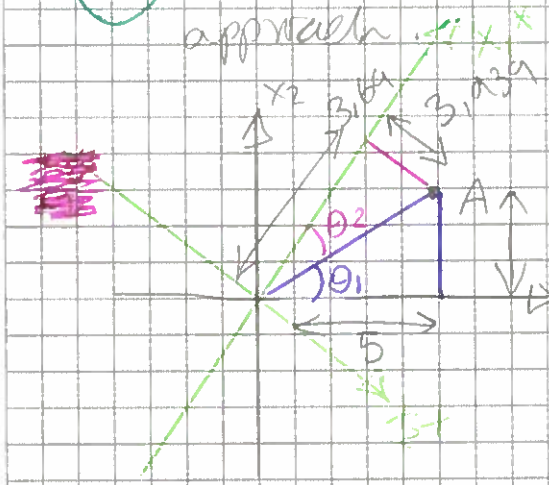
Point B

$X_1^* = 5 \cos 310° + 1 \sin 310° \cong 2,45$

$X_2^* = -5 \sin 310° + 1 \cos 310° \cong 4,27$

Answer the new cordinates for Point A & B in the X_1^* an X_2^* cordinat system are $(3,46 ; 1,01)_{A^*}$ and $(2,45 ; 4,27)_{B^*}$

(b) we can solve this problem by having a geometrical approach



innovative but wrong

$\theta = \theta_1 + \theta_2 =$ that is not what is asked !

$\arctan\left(\frac{2}{5}\right) + \arctan\left(\frac{3,939}{3,69}\right)$

$= 68°$

Answer we have rotate the cordinat system by $68°$ counter clockwise

SU, DEPARTMENT OF STATISTICS

Room: 06 Anonymous code: MM0006 Sheet number: 3

(3) To use PCA there should be a correlation between the independet variables in this case X_1 and X_2. We want to project the values from a 2D space to a 1D (line) by maximizing the variance of the new values.

* Calculate the var(X_1) & var(X_2)

$\bar{X}_1 = (1 + \dots + 24)/8 = 2.5$

$\bar{X}_2 = (4 + \dots + 21)/8 = 2.375$

$\text{var}(X_1) = \sum (X_1 - \bar{X}_1)^2 / (8-1) = 1.4286$

$\text{var}(X_2) = \sum (X_2 - \bar{X}_2) / (8-1) = 1.4107$

tot var = var(X_1) + var(X_2) = 2.8393

* Projection to a line with angle θ.

We use the following equation to calculate the projection

$X_1^* = X_1 \cos\theta + X_2 \sin\theta$

(To save space I just present the result and not the exact calculations)

θ	var(X_1^*)	θ	var(X_1^*)	θ	var(X_1^*)
0°	1,4286	70°	1,3669	130°	1,4884
10°	1,4036	80°	1,3860	140°	1,4915
20°	1,3806	90°	1,4107	150°	1,4286
30°	1,3622	100°	1,4357	160°	1,4724
40°	1,3508	110°	1,4587	170°	1,4525
50°	1,3477	120°	1,4477	180°	1,4286
60°	1,3533				

Max

If we rotate more than 180° we will exact the same answers because we have just changed the direction of the axis.

We see from table that the maximum is between 130° and 150°. This time we calculate the variance of X_1^* for each 5° between 130° & 150°

θ	var(X_1^*)
130°	1,4684
135°	1,4911
140°	1,4915
145°	1,4898
150°	1,4886

The maximum variance we could get is at 140° ± 5°

and we also can calculate the lost of information by this projection

$$\frac{var(X_1^*)}{var(X_1) + var(X_2)} = \frac{1,4915}{2,8393} = 0,5253$$

±

SU, DEPARTMENT OF STATISTICS

Room: UG1 Anonymous code: MM0006 Sheet number: 4

(4)

ID	USER	Edu
S1	250	12
S2	300	14
S3	280	16
S4	330	18
S5	360	24
S6	400	20

Linkage Method

For each of the combination S1 — S6 we calculate the euclidean distance

	S1	S2	S3	S4	S5	S6
S1	0					
S2	2504	0				
S3	916	404	0			
S4	6436	916	2504	0		
S5	12244	3700	6464	936	0	
S6	22564	10036	14416	4404	1616	0

$D(S_1, S_2) = (250-300)^2 + (12-14)^2$

S2 and S3 closest.

	S1	S2 S3	S4	S5	S6
S1	0				
S2 S3	1710	0			
S4	6436	1710	0		
S5	12244	5082	936	0	
S6	22564	12226	4404	1616	0

S4 and S5 closest

	S_1	S_2 S_3	S_4 S_5	S_6
S_1	0			
S_2 S_3	1710	0		
S_4 S_5	9340	33916	0	
S_6	22564	12226	3260	0

S_1 and S_2 S_3 closest

	S_1 S_2 S_3	S_4 S_5	S_6
S_1 S_2 S_3	0		
S_4 S_5	6368	0	
S_6	17396	7743	0

S_1 S_2 S_3 and S_4 S_5 are closest

Answer: By using linkage method we get the following 2 clusters $\{S_1, S_2, S_3, S_4, S_5\}$ and $\{S_6\}$

Ward's method:

In ward's method we calculate the distance with the mean values, for example $D(S_1, S_2) = \left(250 - \frac{300+250}{2}\right)^2 + \left(300 - \frac{300+250}{2}\right)^2 + \left(12 - \frac{14+12}{2}\right)^2 + \left(14 - \frac{14+12}{2}\right)^2$

OBS: As you can see in the equation above the difference in salary is much higher the the difference in education. Without lost of generality from here on we just focus on the Salary and not the education.

S_1 S_2	1250
S_1 S_3	1800
S_4 S_4	12800
S_1 S_5	↓
S_1 S_6	↓ increase

S_2 S_3	200
S_2 S_4	1810
S_2 S_5	↓ increase
S_2 S_6	↓
S_3 S_4	1800

S_3 S_5	↓ increase
S_3 S_6	↓
S_4 S_5	1800
S_4 S_6	↓ increase
S_5 S_6	3200

SU, DEPARTMENT OF STATISTICS

Room: 0G Anonymous code: MM0006 Sheet number: 5

By performing wards method and using the Regnet approximation that education difference is less significant we get the following 5 clusters. $\{S_1\}$ $\{S_2, S_3\}$ $\{S_4\}$, $\{S_5\}$ and $\{S_6\}$ (±)

Algorithm I

In algorithm I which is a non-hierarchical method we take the kth first observation to be the kth first cluster. Then we calculate the distance between each of the observation and put the observation with least distance to that cluster. We repeat this procedure but this time we use mean value of the that cluster that has 2 observations.

Let assume that we want to cluster the data in 3 groups.

	Cl 1 = 250,112	Cl 2 = 300,114	Cl 3 = 280,116
S_1	0	2504	916
S_2	2504	0	4014
S_3	916	404	0
S_4	64136	916	2504
S_5	122441	3700	64164
S_6	22564	10036	141116
Cluster	I	II	III

(±)

Cluster I $\{S_1\}$ Cluster II $\{S_2\}$ Cluster III $\{S_4, S_5\}$ $\{S_3\}$

→

By using Algorithm 1 we get the following 3 clusters and remember that we choosed 3 by ourselves, it wasn't mentioned in the text $\{s_1\}$ $\{s_2 s_4 s_5 s_6\}$ and $\{s_3\}$. For the next run we have to calculate the mean value for the second cluster.

c) Would you get same answer if we were the salary in SEK instead of EUR.

7. D) I'm not sure because than the difference in salary wouldn't be as significant as it was now. On the other side if we changed it from SEK to JPY or any other currency that is less worth than 1 SEK we will get exactly the same result.

(5)

variable	F_1	F_2
x_1	0,9	0,20
x_2	0,7	0,15
x_3	0,2	0,90
x_4	0,2	0,70

a) Specific variances

$\text{mean}(F_1) = (0,9 + \text{---} + 0,2)/4 = 0,5$

$\text{mean}(F_2) = (0,2 + \text{---} + 0,7)/4 = 0,4875$

$\text{var}(F_1) = \left((0,9-\bar{F}_1)^2 + \text{---} + (0,7-\bar{F}_1)^2\right)/(4-1)$

$= 0,1266$

$\text{var}(F_2) = 0,1373$

(?) -

Answer: The specific variances are 0,1266 and 0,1373 for F_1 and F_2.

b) Communalities and % of share variance

F_1	F_2	F_1 % share variance	F_2 % share variance
0,81	0,04	$\frac{0,81}{0,81+0,04} = 95\%$	5 %
0,49	0,0225	96%	4 % (?)
0,04	0,81	5%	95%
0,04	0,49	7,5 %	92,5%

(?)

(c) proportion of variance explained by each factor.
if we assume that $Var(X_i) = 1$ for $i \in \{1, \dots, 5\}$
we can calculate the variance of the unq factors.

$Var(\varepsilon_1) = \widetilde{Var(X_1)} - 0,9^2 - 0,2^2 = 0,15$

$Var(\varepsilon_2) = 0,4875$

$Var(\varepsilon_3) = 0,15$

$Var(\varepsilon_4) = 0,47$

d) Estimated or reproduced correlation matrix

$$C = \begin{pmatrix} 1 & & & \\ 0,166 & 1 & & \\ 0,36 & 0,275 & 1 & \\ 0,32 & 0,245 & 0,167 & 1 \end{pmatrix}$$

$Corr(X_1, X_2) = 0,9 \cdot 0,7 + 0,2 \cdot 0,15 = 0,66$

e) RMSR

$$\begin{pmatrix} 1 & & & \\ 0,7 & 1 & & \\ 0,3 & 0,25 & 1 & \\ 0,35 & 0,2 & 0,6 & 1 \end{pmatrix} - C =$$

$$\begin{pmatrix} 0 & & & \\ 0,024 & 0 & & \\ -0,06 & -0,025 & 0 & \\ 0,03 & 0,045 & -0,07 & 0 \end{pmatrix}$$

→

Communality ex. 5

$$RMSR = \sqrt{\frac{0{,}04^2 + 0{,}06^2 + 0{,}03^2 + 0{,}025^2 + 0{,}045^2 + 0{,}07^2}{6}}$$

$$= \sqrt{\frac{0{,}0566\,25}{6}} = 0{,}097$$

mistake in calculation ? ±

We see that RMSR is really close to zero which implies that the estimated matrix is close to the correlation matrix.

6)

$X_1 = 0,104\ F_1 + 0,824\ F_2 + U_1$

$X_2 = 0,065 F_1 + 0,959\ F_2 + U_2$

$X_3 =$

$X_4 =$

X_6

a) Pattern loadings

pattern loading are the coefficient behind F_1 and F_2. Their value is independent φ for all $\varphi = \{0,9\ 0,5\ 0,2\ 0\ -0,2\ -0,5\ -0,9\}$

X_1: $F_1 \rightarrow 0,104$, $F_2 \rightarrow 0,824$

X_2: $F_1 \rightarrow 0,906$, $F_2 \rightarrow 0,134$

X_6: $F_1 \rightarrow 0,827$, $F_2 \rightarrow 0,016$

ok

b) correlation between X_1 and X_2

$$\text{corr}(X_1 X_2) = \underbrace{0,104 \cdot 0,065 + 0,824 \cdot 0,959}_{A} + \varphi(\underbrace{0,104 \cdot 0,959 + 0,824 \cdot 0,065}_{=B})$$

As you can see the first part A, is independent of φ but the other part depend on if φ. For higher φ the correlation beetwe X_i and X_j increase).

if $\varphi = 0,2$ $\text{corr}(X_1 X_2) = 0,8276$

$\varphi = -0,2$ $\text{corr}(X_1 X_2) = 0,7663$

ok

→

Continuing ex 6 (±) → discuss other com

(c) Procentage of variance of X_1 and X_2 is not account by F_1 and F_2.

To be able to answer this question we have to first calculate $\text{var}(U_1)$ and $\text{var}(U_2)$ c

If we assume that $\text{var}(X_1)$ & $\text{var}(X_2)$ is equal to 1

$$\text{var}(U_i) = \text{var}(X_i) - \lambda_1^2 - \lambda_2^2 - 2\lambda_1\lambda_2\varphi$$

for $\varphi = 0{,}2$

$$\text{var}(U_1) = 1 - 0{,}104^2 - 0{,}824^2 - 2 \cdot 0{,}104 \cdot 0{,}824 \cdot 0{,}2$$
$$= 0{,}2759$$

$$\text{var}\, U_2 = 1 - 0{,}065^2 - 0{,}959^2 - 2 \cdot 0{,}065 \cdot 0{,}959 \cdot 0{,}2$$
$$= 0{,}05116$$

for $\varphi = -0{,}2$

$$\text{var}\, U_1 = 1 - 0{,}104^2 - 0{,}824^2 + 2 \cdot 0{,}104 \cdot 0{,}824 \cdot 0{,}2$$
$$= 0{,}3444$$

$$\text{var}\, U_2 = 1 - 0{,}065^2 - 0{,}959^2 + 2 \cdot 0{,}065 \cdot 0{,}224 \cdot 0{,}2$$
$$= 0{,}1010$$

higher correlation implies higher variance of the unique factors. (±)

For the case $\varphi = 0{,}2$ the procentage that was not accounted by F_1 and F_2 is 27.6% for X_1 and 5% for X_2