Written exam in Multivariate Methods, 7.5 ECTS credits
Tuesday, 15th February 2016, 9:00 – 14:00
Time allowed: FIVE hours
Examination Hall: Laduvikssalen

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 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 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. (12 points) Let us analyse the following 3-variate dataset with 10 observations. Each observation consists of 3 measurements and recorded in the following matrix

Kolumn1	Kolumn2	Kolumn3
7	4	3
4	1	8
6	3	5
8	6	1
8	5	7
7	2	9
5	3	3
9	5	8
7	4	5
8	2	2

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"? Provide detailed calculations for all three numbers recorded in per cents.

2. (12 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 20° counter 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. (12 points) Do the following for the data given below:

a) Assume that data is transformed into mean corrected form. Will the results of the statistical techniques (e.g. factor analysis, principal component analysis) be affected by the transformation? Why or why not?

b) Assume that data is transformed into standardized form. Will the results of the statistical techniques (e.g. factor analysis, principal component analysis) be affected by standardizing the data? Why or why not?

c) Compute the total, between-group, and within-group SSCP matrices. What conclusions can you draw from these matrices?

Financial Data for Failed and non-Failed firms

Observations (Failed Firms)	EBITASS	ROTC	Observations (non-failed)	EBITASS	ROTC
1	0.16	0.18	1	-0.01	-0.03
2	0.21	0.2	2	-0.05	-0.11
3	0.23	0.3	3	0.09	0.12
4	0.16	0.19	4	0.03	0.05
5	0.28	0.17	5	0.04	0.06
6	0.15	0.13			

4. (15 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.

(ii) Cluster the following hypothetical data set into two groups using average linkage method and the associated similarity matrixes. Moreover, cluster the same data into 4 clusters using Ward's method. Discuss your findings.

(iii)

Subject ID	Income in tEUR	Education (in years)
S1	17	10
S2	23	12
S3	25	14
S4	28	15
S5	30	20
S6	35	18

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.

6. (14 points) Consider the following single-factor model

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

Assume that three students give three different sample covariance matrixes of the indicators:

$$
S_1 = \begin{pmatrix} 1.20 & 0.93 & 0.45 \\ 0.93 & 1.56 & 0.27 \\ 0.45 & 0.27 & 2.15 \end{pmatrix}; \quad S_2 = \begin{pmatrix} 1.20 & -0.93 & -0.45 \\ -0.93 & 1.56 & -0.27 \\ -0.45 & -0.27 & 2.15 \end{pmatrix}; \; S_3 = \begin{pmatrix} 1.20 & -0.93 & -0.45 \\ -0.93 & 1.56 & 0.27 \\ -0.45 & 0.27 & 2.15 \end{pmatrix}
$$

Note that the difference is only in the sign of selected covariances. Compute the estimates of the model parameters *($\lambda_1, \lambda_2, \lambda_3, Var(\delta_1), Var(\delta_2), Var(\delta_3)$)* by hand for all three covariance matrixes. Are the parameter estimates unique? After doing the calculations, explain the difference in estimates the best you can and argue how/why the change of sign in the covariance matrix has influenced the estimates. Use intuition if calculations go beyond real numbers. You can also use intuition directly if calculations become too complicated or too long.

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 $b_1, \dots, b_n$. (**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 $\boldsymbol{ab} = a_1b_1 + a_2b_2 + \cdots + a_pb_p$; $\boldsymbol{ab} = \|\boldsymbol{a}\|\|\boldsymbol{b}\|\cos\alpha$

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

Mahalanobis: $MD_{ik}^2 = \frac{1}{1-r^2}\left[\frac{(x_{i1}-x_{k1})^2}{s_1^2} + \frac{(x_{i2}-x_{k2})^2}{s_2^2} - \frac{2r(x_{i1}-x_{k1})(x_{i2}-x_{k2})}{s_1s_2}\right]$

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

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:

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

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

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: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ H_a: $\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}(\boldsymbol{\theta})$ (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 $\boldsymbol{\xi} = \boldsymbol{X}'\boldsymbol{\gamma}$. **B** and **W** are between-groups and within-group SSCP matrices.

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

$|\boldsymbol{W}^{-1}\boldsymbol{B} - \boldsymbol{\lambda I}| = \boldsymbol{0};\ \boldsymbol{\gamma} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ - Fisher's discriminant function

Logistic regression

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

$\ln odds = \beta_0 + \beta_1X_1 + \cdots + \beta_kX_k$

$p = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\cdots+\beta_kX_k)}}$

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:

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

Stockholms universitet

Department of Statistics

Correction sheet

Date: 16/2 - 2016

Room: Laduvikssalen

Exam: Multivariate Methods

Course: Multivariate Methods

Anonymous code: MME-0012

[X] 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				6 / 17
Teacher's notes 8	12	12	9	15	10				

Points	Grade	Teacher's sign.
66		AA

SU, DEPARTMENT OF STATISTICS

Room: Ludwikssalen Anonymous code: MME-0012 Sheet number: 1

1

K1	K2	K3	K1m	K2m	K3m	K1std	K2std	K3std
7	4	3	✓ 0,1	0,5	-2,1	0,0656	0,3162	-0,7482
4	1	8	✓ -2,9	-2,5	2,9	-1,9030	-1,8342	1,0332
6	3	5	✓ -0,9	-0,5	-0,1	-0,5906	-0,3162	-0,0479
8	6	1	✓ 1,1	2,5	-4,1	0,7218	1,8342	-1,4608
8	5	7	✓ 1,1	1,5	1,9	0,7218	0,9487	0,9105
7	2	9	✓ 0,1	-1,5	3,9	0,0656	-0,9487	1,3895
5	3	3	✓ -1,9	-0,5	-2,1	-1,2468	0,3162	-0,7482
9	5	8	✓ 2,1	1,5	2,9	1,3780	0,9487	1,0332
7	4	5	✓ 0,1	0,5	-0,1	0,0656	0,3162	-0,0356
8	2	2	1,1	-1,5	-3,1	0,7218	-0,9487	-1,1045

ok

M 6,9 3,5 8,1

Var 2,32 2,5 7,8778

Std = 1,5239 1,5811 2,8067

$R = \cancel{\begin{pmatrix} 1 & & \\ & & \\ & & 1 \end{pmatrix}}$ $\times \left(\frac{1}{n-1}\right)$

$SSX_1 = 20,9$

$SSX_2 = 22,5$

$SSX_3 = 70,9$

$SCP = 24,4$

$cov(X_1 X_2) = 14,5$

$cov(X_1 X_3) = -3,9$

$cov(X_2 X_3) = -11,5$

$COV\ \cancel{SSCP} = \begin{pmatrix} 20,9 & 14,5 & 3,9 \\ 14,5 & 22,5 & -11,5 \\ 3,9 & -11,5 & 70,9 \end{pmatrix} = \begin{pmatrix} 2,22 & 1,61 & 0,43 \\ 1,61 & 2,5 & -1,28 \\ 0,43 & -1,28 & 7,88 \end{pmatrix}$

$R = \begin{pmatrix} 1 & 0,669 & 0,101 \\ 0,669 & 1 & -0,288 \\ 0,101 & -0,288 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0,617 & \\ & 1 & \\ & & 1 \end{pmatrix}$

$SS = \frac{1}{9} \cdot \begin{pmatrix} 20,9 & 14,5 & 3,9 \\ 14,5 & 22,5 & -11,5 \\ 3,9 & -11,5 & 70,9 \end{pmatrix} = \begin{pmatrix} 2,32 & 1,61 & 0,43 \\ 1,61 & 2,5 & -1,28 \\ 0,43 & -1,28 & 7,88 \end{pmatrix}$

$R = \begin{pmatrix} 1 & 0,669 & 0,792 \\ 0,669 & 1 & -0,299 \\ 0,792 & -0,299 & 1 \end{pmatrix}$? → mistakes

±

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

$\left| \begin{pmatrix} 1 & 0,669 & 0,792 \\ 0,669 & 1 & -0,299 \\ 0,792 & -0,299 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \lambda \right| = 0$

$\begin{vmatrix} 1-\lambda & 0,7 & 0,8 \\ 0,7 & 1-\lambda & -0,3 \\ 0,8 & -0,3 & 1-\lambda \end{vmatrix}$ $\begin{vmatrix} 1-\lambda & 0,669 & 0,792 \\ 0,669 & 1-\lambda & -0,299 \\ 0,792 & -0,299 & 1-\lambda \end{vmatrix}$ $\Longrightarrow$

① $(1-\lambda)\left[(1-\lambda)(1-\lambda)-(-0{,}299)(-0{,}299)\right]=(1-\lambda)\left[1-\lambda-\lambda+\lambda^2-0{,}089401\right]=(1-\lambda)(\lambda^2-2\lambda+0{,}910599)$

$=\lambda^2-2\lambda+0{,}910599-\lambda^3+2\lambda^2-0{,}910599\lambda=-\lambda^3+3\lambda^2-2{,}910599\lambda+0{,}910599$

② $0{,}669\left[0{,}669(1-\lambda)-0{,}792(-0{,}299)\right]=0{,}669(0{,}669-0{,}669\lambda+0{,}236808)=0{,}447561-0{,}447561\lambda+0{,}15842$

③ $0{,}792\left[0{,}669(-0{,}299)-0{,}792(1-\lambda)\right]=0{,}792(-0{,}209001-0{,}792+0{,}792\lambda)=-0{,}165528792-$
$0{,}627264+0{,}627264\lambda$

①+②+③ $=-\lambda^3+3\lambda^2-2{,}910599\lambda+0{,}910599+0{,}447561-0{,}44561\lambda+0{,}158424552-0{,}165528792$
$-0{,}627264+0{,}627264\lambda=-\lambda^3+3\lambda^2-3{,}356209\lambda+0{,}72379176+0{,}627264\lambda=$
$\approx-\lambda^3+3\lambda^2-2{,}73\lambda+0{,}724$

① $(1-\lambda)\left[(1-\lambda)(1-\lambda)-(-0{,}3)(-0{,}3)\right]=(1-\lambda)\left[1-\lambda-\lambda+\lambda^2-0{,}09\right)=(1-\lambda)\left[1-2\lambda+\lambda^2-0{,}09\right]=$
$=1-2\lambda+\lambda^2-0{,}09-\lambda+2\lambda^2-\lambda^3+0{,}09\lambda=-\lambda^3+3\lambda^2-3{,}09\lambda+1$

② $0{,}7(0{,}7(1-\lambda)-0{,}8(-0{,}3)]=0{,}7(0{,}7-0{,}7\lambda+0{,}24)=0{,}49-0{,}49\lambda+0{,}168$

③ $0{,}8\left[0{,}7(-0{,}3)-0{,}8(1-\lambda)\right]=0{,}8\left[-0{,}21-0{,}8+0{,}8\lambda\right)=-0{,}168+0{,}64+0{,}64\lambda$

①+②+③ $=-\lambda^3+3\lambda^2-3{,}09\lambda+1+0{,}49-0{,}49\lambda+0{,}168-0{,}168+0{,}64+0{,}64\lambda$

$=-\lambda^3+3\lambda^2-2{,}94\lambda+2{,}13=0$

I struggled to long with my correlation matrix which caused me to not being able to find my eigenvalues, which make it impossible to find my eigenvectors, which I would have used to calculate the angle θ, probably using $\cos\alpha=\frac{a}{c}$ and thereafter calculate the new points for my first axis. I would then have calculated the variance of that one which I would have divided by the total variance to find how much variance the new axis explained.

I wasn't fast enough on the other questions, sorry!

Direction is ok

(+)

SU, DEPARTMENT OF STATISTICS

Room: Laduvikssalen Anonymous code: NME-0012 Sheet number: 2

2 a) $X_1^* = \cos\theta \cdot X_1 + \sin\theta \cdot X_2$ $A = (3, -2)$

$X_2^* = -\sin\theta \cdot X_1 + \cos\theta \cdot X_2$ $B = (5, 1)$

For A: $X_1^* = 3\cos(20) - 2\sin(20) = 2,135$

$X_2^* = -3\sin(20) + 2\cos(20) = -2,905$

$\Rightarrow A^* = (2,135, -2,905)$ ok

For B: $X_1^* = 5\cos(20) + \sin(20) = 5,040$

$X_2^* = -5\sin(20) + \cos(20) = -0,770$

$B^* = (5,040, -0,770)$ ok

b) (1) $3,69 = 5\cos\theta + 2\sin\theta \longrightarrow 5\cos\theta = 3,69 - 2\sin\theta$

(2) $3,939 = -5\sin\theta + 2\cos\theta$ $\cos\theta = 0,738 - 0,4\sin\theta$ into (2)

$\Rightarrow 3,939 = -5\sin\theta + 2(0,738 - 0,4\sin\theta)$

$3,939 = -5\sin\theta + 1,476 - 0,8\sin\theta$

$2,463 = -5,8\sin\theta$

$\sin\theta = -0,42465551724$

$\theta = \sin^{-1}(-0,42465551724)$

$\theta \approx -25,129°$

Which gives:

25,129° counter clockwise ?

(360 - 25,129) for clockwise

= 334,871° clockwise

ok

SU, DEPARTMENT OF STATISTICS

Room: Laduvikssalen   Anonymous code: MME-0012   Sheet number: 3

① $\boxed{3}$

Failed firms	EBITASS	ROTC	$EBITASS_M$	$ROTC_M$	$EBITASS_{std}$	$ROTC_{std}$
1	0,16	0,18	-0,038	-0,015	-14,5216	-4,6440
2	0,21	0,2	0,012	0,005	4,5858	1,5480
3	0,23	0,3	0,032	0,105	12,2287	32,5077
4	0,16	0,19	-0,038	-0,005	-14,5216	-1,5480
5	0,28	0,17	0,082	-0,025	31,3360	-7,740
6	0,15	0,13	-0,048	-0,065	-18,3430	-20,124

$\mu_E^F \approx 0,198$ $V(E)^F = 0,0026168$

$\mu_R^F \approx 0,195$ $V(R)^F = 0,00323$

②

Non-failed	EBITASS	ROTC	$EBITASS_M$	$ROTC_M$	$EBITASS_{std}$	$ROTC_{std}$
1	-0,01	-0,03	-0,03	-0,048	-10,7143	-6,0226
2	-0,05	-0,11	-0,07	-0,128	-25	-16,0602
3	0,09	0,12	0,07	0,102	25	12,7980
4	0,03	0,05	0,01	0,032	3,5714	4,0151
5	0,04	0,06	0,02	0,042	7,1429	5,2698

$\mu_E^N = 0,02$ $V(E)^N = 0,0028$

$\mu_R^N = 0,018$ $V(R)^N = 0,00797$

$SS_{1E} = \sum x_{ij}^2 = 0,013084$

$SS_{1R} = \cancel{\sum} = 0,01615$

$SCP_{1ER} = 0,00525$

$$\Rightarrow SSCP_1 = \begin{pmatrix} 0,013084 & 0,00525 \\ 0,00525 & 0,01615 \end{pmatrix}$$

$SS_{2E} = 0,0112$

$SS_{2R} = 0,03188$

$SCP_{2ER} = 0,0187$

$$\Rightarrow SSCP_2 = \begin{pmatrix} 0,0112 & 0,0187 \\ 0,0187 & 0,03188 \end{pmatrix}$$

$$\underline{SSCP_W} = SSCP_1 + SSCP_2 = \begin{pmatrix} 0,024284 & 0,02395 \\ 0,02395 & 0,04803 \end{pmatrix} \text{ ok}$$

$\mu_{ET} = 0,1173$, $\mu_{RT} = 0,1145$ $\Rightarrow$ $SSX_{ET} = 0,11102$, $SSX_{RT} = 0,13347$, $SCP_T = 0,11003635$

$$\Rightarrow \underline{SSCP_T} = \begin{pmatrix} 0,11102 & 0,11003635 \\ 0,11003635 & 0,13347 \end{pmatrix} \text{ ok}$$

$SSCP_T = SSCP_W + SSCP_B \Rightarrow SSCP_B = SSCP_T - SSCP_W$

$$\Rightarrow SSCP_B = \begin{pmatrix} 0,085916 & 0,08608635 \\ 0,08608635 & 0,08544 \end{pmatrix} \text{ ok}$$

a) Calculations on previous side!

No, the result of using mean-corrected data in FA and PCA will not affect the result.

ok

b) Yes, the result will change using standardized data on FA and PCA because these methods are not scale invariant.

ok

c) The $SSCP_W$, $SSCP_B$ and $SSCP_T$ gives information on the variance within and between groups, as well as the total sum of squares for the data set. Judging from my calculations on the previous page, the variance ~~betw~~ within groups looks similar, but the variance between groups seem to differ a bit more. Although I suspect that the failed- and non-failed labels should switch places because it might be more intuitive with non-failed firms that have higher ~~Earnings Before Interest~~ EBITASS and ROTC than failed firms given that it is a rather common measure of financial health.

ok

SU, DEPARTMENT OF STATISTICS

Room: Laduvikssalen Anonymous code: MME-0012 Sheet number: 4

4 (i) Cluster analysis is used to reduce data and maybe analyze differences between clusters, how similar observations are to each other, or how far apart they are

Types of cluster analysis:
- centroid method
- nearest-neighbour
- furthest-neighbour
- Average-linkage method
- Ward's method

} hierarchical clustering. The difference between these is how you calculate the clusters. Replacing by centroid, minimum/maximum distance, average distance or minimiza total within-group sums of squares.

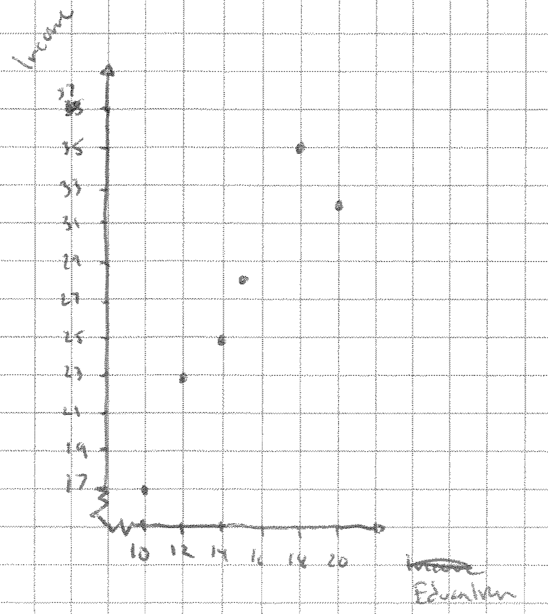
There is also non-hierarchical clustering.

Assumptions that should be met are that each group is similar (homogenous) with respect to some desired characteristic and that each group is different from the others with respect to the same characteristic.

These assumptions should, in my opinion, be strict in order to not involve even more subjectivity to this method of analysis. Also, by relaxing these assumptions, different people might get different clustering

OK

(ii)



Just by eye-balling the data, clustering will be pretty hard

⟹

Average-linkage method - two groups

$$D_{AB} = \sqrt{\sum_{j=1}^{p} (a_j - b_j)^2}$$

$\frac{23+25}{2} = 24 \qquad \frac{12+14}{2} = 13$

	S_1	S_2	S_3	S_4	S_5	S_6
S_1	0	6,32	8,94	12,08	16,4	
S_2	6,32	0	2,83	5,83		
S_3	8,94	2,83	0	3,16		
S_4	12,08	5,83	3,16	0		
S_5	16,40	10,63	7,81	5,39	0	
S_6	18,11	13,42	10,77	7,62	5,39	0

$\frac{24+28}{2} = 26 \qquad \frac{13+15}{2} = 14$

	S_1	$S_{2,3}$	S_4	S_5	S_6
S_1	0				
$S_{2,3}$	7,62	0			
S_4	12,08	2	0		
S_5	16,40	9,22	5,39	0	
S_6	19,70	12,08	7,62	5,39	0

$\frac{30+35}{2} = 32,5 \qquad \frac{20+18}{2} = 19$

	S_1	$S_{2,3,4}$	S_5	S_6
S_1	0			
$S_{2,3,4}$	9,85	0		
S_5	16,40	7,21	0	
S_6	18,11	9,85	5,39	0

$\frac{26+32,5}{2} = 29,25 \qquad \frac{14+19}{2} = 16,5$

	S_1	$S_{2,3,4}$	$S_{5,6}$
S_1	0		
$S_{2,3,4}$	~~7,9~~ 9,85	0	
$S_{5,6}$	17,92	8,20	0

The two clusters: $C_1 = S_1$

$C_2 = S_{2,3,4,5,6}$

which might make sense because the first observation might be considered an outlier and seem relatively different from the rest of the obs.

iii) ⊖

good

SU, DEPARTMENT OF STATISTICS

Room: Ladugårdssalen Anonymous code: MME-0012 Sheet number: 5

$\boxed{5}$ a) $Var(X_1) = \lambda_1^2 + \lambda_2^2 + Var(\varepsilon_1)$ ⟸ Assumes uncorrelated factors

$\Rightarrow Var(\varepsilon_1) = 1 - 0{,}9^2 - 0{,}2^2 = \underline{0{,}15}$

$Var(\varepsilon_2) = 1 - 0{,}7^2 - 0{,}15^2 = \underline{0{,}4875}$

$Var(\varepsilon_3) = 1 - 0{,}2^2 - 0{,}9^2 = \underline{0{,}15}$

$Var(\varepsilon_4) = 1 - 0{,}2^2 - 0{,}7^2 = \underline{0{,}47}$

OK

High specific variances indicate that there is a large proportion of variance that is not explained by the factors. The lower the specific variance, the better because specific variance is variance not shared with the fact

b)

	F_1	F_2
X_1	$0{,}9^2 = \underline{0{,}81}$	$0{,}2^2 = \underline{0{,}04}$
X_2	$0{,}7^2 = \underline{0{,}49}$	$0{,}15^2 = \underline{0{,}0225}$
X_3	$0{,}2^2 = \underline{0{,}04}$	$0{,}9^2 = \underline{0{,}81}$
X_4	$0{,}2^2 = \underline{0{,}04}$	$0{,}7^2 = \underline{0{,}49}$
Tot:	$1{,}38$ +	$1{,}3625$ = $2{,}7425$

Communalities are the shared variance between an indicator and a factor. Therefore, high communality means much shared variance and thus a good indicator.

Shared variance is 1 − unique variance: $\Rightarrow$

$\% X_1 = 1 - 0{,}15 = \underline{85\%}$

$\% X_2 = 1 - 0{,}4875 = \underline{51{,}25\%}$

$\% X_3 = 1 - 0{,}15 = \underline{85\%}$

$\% X_4 = 1 - 0{,}47 = \underline{53\%}$

% shared variance is the % of an indicators total variance that is shared with the factors. And like the communalities case, the higher the shared variance, the better the indicator is.

ok

c) See calculations in table above. Proportion of variance explained by each factor is:

For F_1: $\dfrac{1{,}38}{2{,}7425} \approx 50{,}32\%$

For F_2: $\dfrac{1{,}3625}{2{,}7425} \approx 49{,}68\%$

Thus, the factors explain approximately the same amount of variance and are therefore almost equally good.

OK

Assumes uncorrelated factors

Estimated correlation matrix

d) $corr(X_1, X_2) = 0{,}9 \cdot 0{,}7 + 0{,}2 \cdot 0{,}15 = 0{,}66$

$corr(X_1, X_3) = 0{,}9 \cdot 0{,}2 + 0{,}2 \cdot 0{,}9 = 0{,}36$

$corr(X_1, X_4) = 0{,}9 \cdot 0{,}2 + 0{,}2 \cdot 0{,}7 = 0{,}32$

$corr(X_2, X_3) = 0{,}7 \cdot 0{,}2 + 0{,}15 \cdot 0{,}9 = 0{,}275$

$corr(X_2, X_4) = 0{,}7 \cdot 0{,}2 + 0{,}15 \cdot 0{,}7 = 0{,}245$

$corr(X_3, X_4) = 0{,}2 \cdot 0{,}2 + 0{,}9 \cdot 0{,}7 = 0{,}67$

	X_1	X_2	X_3	X_4
X_1	1			
X_2	0,66	1		
X_3	0,36	0,275	1	
X_4	0,32	0,245	0,67	1

Compared to the given correlation matrix, the error seem pretty small which would mean that the model is pretty good.

e)
$$\begin{pmatrix} 1 & & & \\ 0{,}7 & 1 & & \\ 0{,}3 & 0{,}25 & 1 & \\ 0{,}35 & 0{,}2 & 0{,}6 & 1 \end{pmatrix} - \begin{pmatrix} 1 & & & \\ 0{,}66 & 1 & & \\ 0{,}36 & 0{,}275 & 1 & \\ 0{,}32 & 0{,}245 & 0{,}67 & 1 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & & & \\ 0{,}04 & 0 & & \\ -0{,}06 & -0{,}025 & 0 & \\ 0{,}03 & -0{,}045 & -0{,}07 & 0 \end{pmatrix}}$$

Residual matrix show relatively small errors, which concur with ~~statement~~ argument in d)

$$RMSR = \sqrt{\frac{\sum_{i=1}^{\hat{n}} e_{ij}^2}{n}} = \sqrt{\frac{0{,}04^2 + (-0{,}06)^2 + 0{,}03^2 + (-0{,}025)^2 + (-0{,}045)^2 + (-0{,}07)^2}{6}} \approx \underline{0{,}0477}$$

RMSR of 0,0477 is relatively small and shows that the initial hypothetical data set is pretty good. Because the error is small...

ok

SU, DEPARTMENT OF STATISTICS

Room: Ladvikssalen Anonymous code: MME-0012 Sheet number: 6

6

For S_1: $S_1 = \begin{pmatrix} 1,2 & 0,93 & 0,45 \\ 0,93 & 1,56 & 0,27 \\ 0,45 & 0,27 & 2,15 \end{pmatrix} \Rightarrow$

$$\underbrace{\begin{matrix} \lambda_1^2 + Var(\delta_1) = 1,2 & \lambda_1\lambda_2 = 0,93 \\ \lambda_2^2 + Var(\delta_2) = 1,56 & \lambda_1\lambda_3 = 0,45 \\ \lambda_3^2 + Var(\delta_3) = 2,15 & \lambda_2\lambda_3 = 0,27 \end{matrix}}$$

six equations, six unknown $\Rightarrow$ just-identified.

$\lambda_1 = \frac{0,45}{\lambda_3}$, $\lambda_2 = \frac{0,27}{\lambda_3}$, $\lambda_1\lambda_2 = 0,93$

$\Rightarrow \frac{0,45 \cdot 0,27}{\lambda_3^2} = 0,93$

$0,1215 = 0,93\,\lambda_3^2$

$\lambda_3^2 = 0,1306451613$

$\boxed{\lambda_3 \approx 0,3614}$

$\boxed{\lambda_2 = \frac{0,27}{0,3614} \approx 0,7471}$ $\boxed{\lambda_1 = \frac{0,45}{0,3614} \approx 1,2452}$

$\boxed{Var(\delta_1) = 1,2 - 1,2452^2 = -0,3505}$ → Negative variance something wrong with model.

$\boxed{Var(\delta_2) = 1,56 - 0,7471^2 = 1,002}$ Calculations ok

$\boxed{Var(\delta_3) = 2,15 - 0,3614^2 = 2,019}$

For S_2: $S_2 = \begin{pmatrix} 1,2 & -0,93 & -0,45 \\ -0,93 & 1,56 & -0,27 \\ -0,45 & -0,27 & 2,15 \end{pmatrix} \Rightarrow$

$$\underbrace{\begin{matrix} \lambda_1^2 + Var(\delta_1) = 1,2 & \lambda\lambda_2 = -0,93 \\ \lambda_2^2 + Var(\delta_2) = 1,56 & \lambda_1\lambda_3 = -0,45 \\ \lambda_3^2 + Var(\delta_3) = 2,15 & \lambda_2\lambda_3 = -0,27 \end{matrix}}$$

just-identified

$\lambda_1 = \frac{-0,45}{\lambda_3}$, $\lambda_2 = \frac{-0,27}{\lambda_3}$, $\lambda_1\lambda_2 = -0,93$

$\Rightarrow \frac{(-0,45)(-0,27)}{\lambda_3^2} = -0,93$

$0,1215 = -0,93\,\lambda_3^2$

$\lambda_3 = \sqrt{\frac{0,1215}{-0,93}} \Rightarrow$ solution without real numbers. for all lambda's

For S_3: $S_3 = \begin{pmatrix} 1,2 & -0,93 & -0,45 \\ -0,93 & 1,56 & 0,27 \\ -0,45 & 0,27 & 2,15 \end{pmatrix} \Rightarrow$

$$\underbrace{\begin{matrix} \lambda_1^2 + Var(\delta_1) = 1,2 & \lambda_1\lambda_2 = -0,93 \\ \lambda_2^2 + Var(\delta_2) = 1,56 & \lambda_1\lambda_3 = -0,45 \\ \lambda_3^2 + Var(\delta_3) = 2,15 & \lambda_2\lambda_3 = 0,27 \end{matrix}}$$

just-identified.

$\lambda_1 = \frac{-0,45}{\lambda_3}$, $\lambda_2 = \frac{0,27}{\lambda_3}$, $\lambda_1\lambda_2 = -0,93$

$\frac{(-0,45)\cdot 0,27}{\lambda_3^2} = -0,93$

$-0,1215 = -0,93\,\lambda_3^2$

$\lambda_3^2 = 0,1306...$

$\boxed{\lambda_3 \approx 0,3614}$

$\boxed{\lambda_1 = \frac{-0,45}{0,3614} \approx -1,2451}$

$\boxed{\lambda_2 = \frac{0,27}{0,3614} \approx 0,7471}$

$\boxed{Var(\delta_2) = 1,56 - 0,7471^2 = 1,002}$

$\boxed{Var(\delta_3) = 2,15 - 0,3614^2 = 2,019}$

$\boxed{Var(\delta_1) = 1,2 - (-1,2451)^2 \approx -0,3503}$

negative variance

As can be seen on the previous page, calculations for S_2 leads to solutions without real numbers. Also, calculations for S_1 and S_3 leads to variances with negative sign, which is not correct due to the definition of variance says it should be positive. In all S's we have six equations and six unknowns (just-identified) which leads to unique parameter estimates. However, we want the correlation between indicators and factors to be positive in order to have real solutions. Since we don't have this ~~...~~ for all S's, there should be something wrong with our model specification. In our case, most wrong should be S_2 and S_3 even though they will have unique parameters estimates.

$\pm$