Stockholms universitet

STATISTISKA INSTITUTIONEN HT-2015

Written exam in Multivariate Methods, 7.5 ECTS credits
Thursday, 14th December 2015, 16:00 – 21:00
Time allowed: FIVE hours
Examination Hall: Brunnssvikssalen

You are required to answer all **7 (seven)** 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 two lists 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)

(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.

2. (10 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	Observation (non-Failed)	EBITASS	ROTC
1	0.158	0.182	13	-0.012	-0.031
2	0.210	0.206	14	0.036	0.053
3	0.207	0.188	15	0.038	0.036
4	0.280	0.236	16	-0.063	-0.074
5	0.197	0.193	17	-0.054	-0.119
6	0.227	0.173	18	0.000	-0.005
7	0.148	0.196	19	0.005	0.039
8	0.254	0.212	20	0.091	0.122
9	0.079	0.147	21	-0.036	-0.072
10	0.149	0.128	22	0.045	0.064

3. (10 points) This question belongs to the two group discriminant analysis. Show that

$B = \frac{n_1 n_2}{n_1 + n_2}(\bar{\mu}_1 - \bar{\mu}_2)(\bar{\mu}_1 - \bar{\mu}_2)'$, where B is between-groups **SSCP** matrix for p variables, $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ are the $p*1$ vectors of means for group 1 and group 2, and n_1 and n_2 are the number of observations in group 1 and group 2. Hint: start with the case of only one variable, say X and then generalize your calculations to the multivariate case.

4. (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} = -0.9. Repeat all your calculations in assumption that correlation changed to Corr(F_1, F_2)= Φ_{12} = 0.9 and discuss the differences in detail. Try to provide intuition for at least some of your answers.

(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?

5. (10 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:

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

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.

6. (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.

7. (10 points) Describe assumptions on data/observations you will be checking before applying PCA (principal component analysis), FA (factor analysis), CA (cluster analysis), two group DA (discriminant analysis) and LogR (logistic regression). Briefly describe one example (remember first page of each chapter?) of a suitable problem per method. For each example you mention, indicate which other (if any) of the above mentioned methods is applicable.

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

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; $|\boldsymbol{\Sigma} - \boldsymbol{\lambda I}| = \boldsymbol{0}$; $\boldsymbol{\gamma}$-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 $\lambda = \frac{\boldsymbol{\gamma}'\boldsymbol{B}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\boldsymbol{W}\boldsymbol{\gamma}}$

$|\boldsymbol{W}^{-1}\boldsymbol{B} - \boldsymbol{\lambda I}| = \boldsymbol{0}$; $\boldsymbol{\gamma} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ - 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:

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

Madeleine Hagberg

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: $w_{i1}^2 + w_{i2}^2 + \dots + w_{ip}^2 = 1$, $i = 1, \dots, p$
  $w_{i1}w_{j1} + w_{i2}w_{j2} + \dots + w_{ip}w_{jp} = 0 \quad \forall i \neq j$
- Characteristic equation: $\det(\lambda I - A) = 0$
- Loadings = correlation between the original and the new variables
  $l_{ij} = \frac{w_{ij}}{s_j}\sqrt{\lambda_i}$

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 any indicator x_j	Structure loading	Shared variance = (structure loading)^2	Correlation between indicators
One-factor model	$x_1 = \lambda_1\xi + \varepsilon_1$ ⋮ $x_p = \lambda_p\xi + \varepsilon_p$	$V(x_j) = \lambda_j^2 + V(\varepsilon_j)$	$Cor(x_j, \xi) = \lambda_j$	λ_j^2	$Corr(x_j, x_k) = \lambda_j\lambda_k$
Two-factor model	$x_1 = \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \varepsilon_1$ ⋮ $x_p = \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \varepsilon_p$	$V(x_j) = \lambda_{j1}^2 + \lambda_{j2}^2 + 2\lambda_{j1}\lambda_{j2} + V(\varepsilon_j)$	$Cor(x_j, \xi_2) = \lambda_{j2} + \lambda_{j1}\phi$; $Cor(x_j, \xi_1) = \lambda_{j1} + \lambda_{j2}\phi$	$(\lambda_{j2} + \lambda_{j1}\phi)^2$; $(\lambda_{j1} + \lambda_{j2}\phi)^2$	$Cor(x_j, x_k) = \lambda_{j1}\lambda_{k1} + \lambda_{j2}\lambda_{k2} + [\lambda_{j1}\lambda_{k2} + \lambda_{j2}\lambda_{k1}]\phi$

$\boxed{Cor(\xi_1, \xi_2) = \phi}$

CFA

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

$P(x_1, x_2, x_3) = P(x_1) \cdot P(x_2 \mid x_1) \cdot P(x_3 \mid x_1, x_2)$

CA

$$D_{AB}^2 = \sum_{j=1}^{p} (a_j - b_j)^2 \qquad SD_{ik}^2 = \sum_{j=1}^{p} \left(\frac{x_{ij} - x_{kj}}{s_j}\right)^2 \qquad MD_{ik}^2 = \frac{1}{1-r^2}\left[\frac{(x_{i1} - x_{k1})^2}{s_1^2} + \frac{(x_{i2} - x_{k2})^2}{s_2^2} - \frac{2r(x_{i1} - x_{k1})(x_{i2} - x_{k2})}{s_1 s_2}\right]$$

DA
(two groups)

$\lambda = \frac{SS_b}{SS_w} \to \max$

If $\xi = \bar{X}^T\hat{\gamma} \Rightarrow$ The estimation of $\hat{\gamma}$: $\hat{\gamma}^T = (\bar{\mu}_1 - \bar{\mu}_2)^T \Sigma^{-1}$

$SSCP_w = SSCP_1 + SSCP_2$
$SSCP_T = SSCP_w + SSCP_b$

LOG-REG

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

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

$\ln\frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$

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

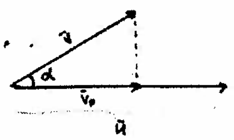
If $Y = \beta_0 + \beta_1 X$: $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

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

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

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

- Projection: of $\vec{v}$ onto $\vec{u}$



- The projection vector: $\vec{v}_p = \frac{\|\vec{v}_p\|}{\|\vec{u}\|}\vec{u}$
- $\|\vec{v}_p\| = \|\vec{v}\|\cos\alpha = \frac{\vec{v} \cdot \vec{u}}{\|\vec{u}\|}$

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

Department of Statistics

Stockholms universitet

Correction sheet

Date: 14/12 - 2015

Room: Brunnsvikssalen

Exam: Multivariate Methods

Course: Multivariate Methods

Anonymous code: MME-0004

☒ 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

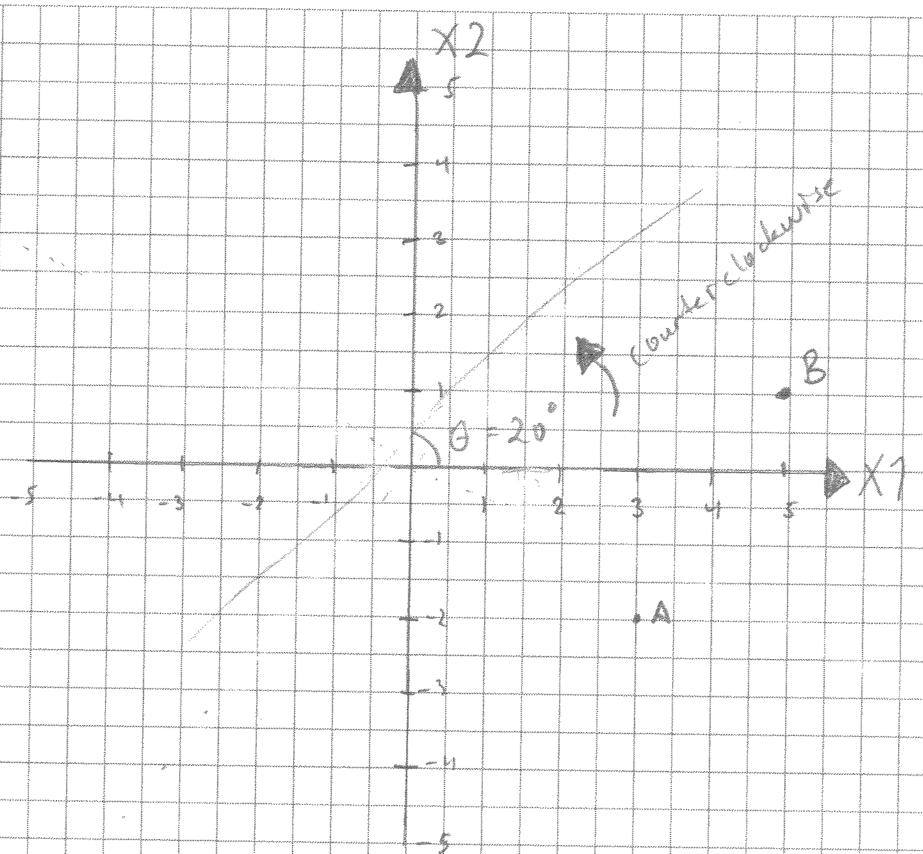
1	2	3	4	5	6	7	8	9	Total number of pages
X	X	X	X	X	X	X			7 17
Teacher's notes 9	4	5	14	9	15	10			

Points	Grade	Teacher's sign.
66 + 13 79	C	AA

SU, DEPARTMENT OF STATISTICS

Room: Brunnsviksalen Anonymous code: MME-0004 Sheet number: 1

1. a)



$X_1^* = \cos\theta x_1 + \sin\theta x_2$ Counter-clockwise means

$X_2^* = -\sin\theta x_1 + \cos\theta x_2$ 20° the unit circle way

$X_{1A}^* = \cos 20° \cdot 3 + \sin 20° \cdot (-2) = 2{,}14$

$X_{2A}^* = -\sin 20° \cdot 3 + \cos 20° \cdot (-2) = -2{,}90$ $\Big\}$ $A^* = (2.14, -2.9)$

$X_{2A} = \cos 20° \cdot 5 + \sin 20° \cdot 1 = 5{,}04$

$X_{2B} = -\sin 20° \cdot 5 + \cos 20° \cdot 1 = -0{,}77$ $\Big\}$ $B^* = (5, -0{,}77)$

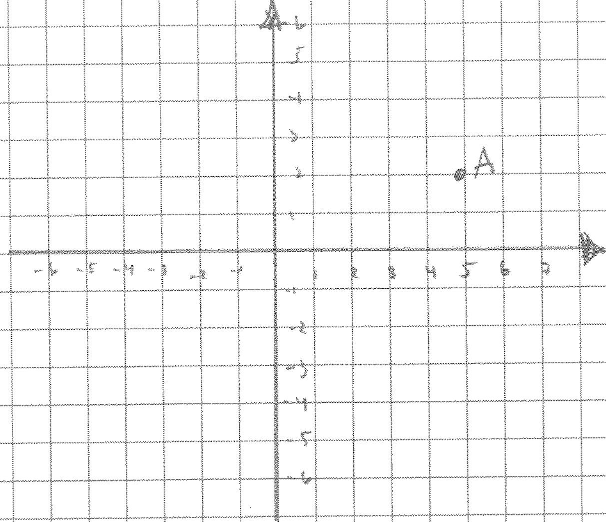
Check:

$D_{AB} = \sqrt{(3-5)^2 + (-2-1)^2} = 3{,}605$ $D_{A^*B^*} = \sqrt{(5-2{,}14)^2 + (-0{,}77+2{,}9)^2}$
$= 3{,}605$

Which means they are positioned to eachother the same as before the rotation

on +

1. b)



$X^{*}_{1A} = \cos\theta \cdot 5 + \sin\theta \cdot 2 = 3.69$

$X^{*}_{1B} = -\sin\theta \cdot 5 + \cos\theta \cdot 2 = 3.939$

$$\left.\begin{aligned} 5\cos\theta + 2\sin\theta &= 3.69 \\ \times -2.5 \quad 2\cos\theta - 5\sin\theta &= 3.939 \end{aligned}\right\}$$ Equation to be solved!

$14.5 \sin\theta = -6.1575$ $\quad \theta = \cos^{-1}\left(-\frac{6.1575}{14.5}\right) = 115^\circ$

$2\cos\theta - 5\sin\theta = 3.939$

correct

Answer $\theta = 115^\circ$ clockwise

$\theta_{counterclockwise} = 360 - 115^\circ = \underline{245}$

I am not sure how you got to this answer

$\theta = -25^\circ$

is correct answer

clockwise rotation

Somehow you got mixed with $90^\circ + 25^\circ =$

$= 115^\circ$

+

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvikssalen **Anonymous code:** MME-0004 **Sheet number:** 2

It is hard to follow this solution.

2. Standardized data only calculated for first row of each group!

	Nr	EBITDAS	ROTC	Mean(ci) EBITASS	Mean(ci) ROTC	Std EBITAS	Std ROTC
Mean 1	1	0,158	0,182	-0,0329	-0,041	-0,52	-0,65
	2	0,216	0,206	0,0191	0,02	$\frac{x-u}{s}$	$\frac{x_i-u}{s}$
	3	0,207	0,188	0,016	0,0019		
	4	0,280	0,236	0,089	0,05		
	5	0,197	0,193	0,06	-0,1261		
	6	0,227	0,173	0,037	-0,013		
	7	0,148	0,196	-0,012	0,01		
	8	0,254	0,212	0,064	0,026		
	9	0,079	0,147	-0,111	-0,04		
	10	0,149	0,128	-0,041	-0,06		
Mean 2	11	-0,012	-0,031	-0,017	-0,0321	-0,38	-0,85
	12	0,036	0,053	0,031	0,05	$\frac{x_i-u}{s}$	$\frac{x_i-u}{s}$
	13	0,038	0,036	0,033	0,034		
	14	-0,063	-0,074	-0,068	-0,076		
	15	-0,054	-0,119	-0,059	-0,12		
	16	0	0,005	-0,005	0,003		
	17	0,005	0,039	0	0,037		
	18	0,091	0,122	0,086	0,12		
	19	-0,036	-0,072	-0,041	0,073		
	20	0,045	0,064	0,04	0,062		

$Mean_{EBITDAS1} = \frac{1,909}{10} = 0,1909$

$Mean_{EBITDAS2} : \frac{0,05}{10} = 0,005$

$Mean_{ROTC1} : \frac{1,861}{10} = 0,1861$

$Mean_{ROTC2} : \frac{0,0186}{10} = 0,00186$

$\left(\text{Standardize} = \frac{X-u}{\sigma} \right)$

$S^2_{EBITASS1} = \frac{\Sigma(x-u)^2}{n-1} = \frac{0,034}{9} = 0,004$

$S^2_{EBITASS2} = \frac{0,02}{9} = 0,002$

$S^2_{ROTC1} = \frac{0,037}{9} = 0,004$

$S^2_{ROTC2} = \frac{0,05}{9} = 0,005$

$Mean_{Etot} = 0,098$

$Mean_{Rtot} = 0,094$

$S^2_{Etot} = \frac{0,039 + 0,02}{19} = 0,003$ $S^2_{Rtot} = 0,005$

$S^2_{E1} = 0,004$ $\quad$ $Cov_i = \sum(x_i - u_i)_{\text{scaled}}(x_j - u_j)$

$S^2_{E2} = 0,002$

$S^2_{R1} = 0,004$

$S^2_{R2} = 0,005$

$\sum(x_i - u)_{E1} = 0,0603$

$\sum(x_i - u)_{E2} = 0$

$\sum(x - u)_{R1} = -0,1462$

$\sum(x - u)_{R2} = 0,15$

$\sum(x_i - u_i)(x_j - u_j)_1 = 0,0603 \cdot -0,1462 = -0,009$

$\sum(x_i - u_i)(x_j - u_j)_2 = 0$

$\sum(x_i - u_i)(x_j - u_j)_{tot} = 0,0603 \cdot (0,15 - 0,1462) = 2,41E-5$

$$S_T = \begin{bmatrix} 0,003 & 2,41E-5 \\ 2,41E-5 & 0,005 \end{bmatrix} \quad S_{W1} = \begin{bmatrix} 0,004 & 0,009 \\ 0,009 & 0,004 \end{bmatrix}$$

$$S_{W2} = \begin{bmatrix} 0,002 & 0 \\ 0 & 0,005 \end{bmatrix} \quad S_{W tot} = \begin{bmatrix} 0,006 & 0,009 \\ 0,009 & 0,009 \end{bmatrix}$$

$$S_B = S_T - S_W = \begin{bmatrix} -0,003 & 0,009 - 2,41E-5 \\ 0,009 - 2,41E-5 & -0,004 \end{bmatrix} ?$$

a) The data will not be affected by mean corrected, the data will be scale dependent.

b) * The data will be affected since each variable will have the same weight of variance. So called scale indifferent.

c) Small variances between groups, but hard to know if calc are correct. If correct not to high var between, but som within each group.

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvikssalen Anonymous code: MME-0004 Sheet number: 3

3. Let's start from the beginning:

1. DM function:

$z = x'y$ SS of DM $= z'z = (x'y)'(x'y) = y'xx'y$

$= y'Ty$

$T = B + W$ where B is the between variance and W is the within variance

2. SSCP om $= y'By + y'Wy$

Goal is to estimate $\lambda = \frac{(y'By)}{(y'Wy)}$

3. $\frac{d\lambda}{dy} = \frac{2By\,y'Wy - 2y'By\,Wy}{(y'Wy)^2} = 0$

$\downarrow$

$\frac{2By - \lambda Wy}{y'Wy} = 0$

$\downarrow$

$(B - \lambda W)y = 0 \rightarrow (W^{-1}B - \lambda I)y = 0$

? $\boxed{B = \frac{n_1 n_2}{n_1 + n_2}(\bar{u}_1 - \bar{u}_2)(\bar{u}_1 - \bar{u}_2)'}$ Since $\frac{n_1 n_2}{n_1 + n_2}$ is a constant we can write $B = C(u_1 - u_2)(u_1 - u_2)' \rightarrow$ Apply $W^{-1}B - \lambda I = 0$

4. $C\,W^{-1}(\bar{u}_1 - \bar{u}_2)(\bar{u}_1 - \bar{u}_2)'y = \lambda y$

$\frac{C}{\lambda} W^{-1}(\bar{u}_1 - \bar{u}_2)(\bar{u}_1 - \bar{u}_2)'y = y$

Since $(\bar{u}_1 - \bar{u}_2)'y$ is a scalar we can write it as
$y = KW^{-1}(\bar{u}_1 - \bar{u}_2)$ and since W is proportional to
varcov-matrix Σ we can write it as:
$y = K\Sigma^{-1}(\bar{u}_1 - \bar{u}_2) \Rightarrow y' = (\bar{u}_1 - \bar{u}_2)'\Sigma^{-1}$

Different values of K gives different values of y, therefore the absolute weight are not unique, only relatively unique

1

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvikssalen Anonymous code: MME-0004 Sheet number: 4

4. a) Pattern loadings are:

	F1	F2
X_1	0,104	0,824
X_4	0,906	0,134
X_6	0,827	0,016

ok

b) 1. $\text{corr}(F_1, F_2) = \phi = -0.9 \rightarrow l_{i1} l_{k1} + l_{i2} l_{k2} + (l_{i1} l_{k2} + l_{i2} l_{k1})\phi$

$= 0{,}104 \cdot 0{,}065 + 0{,}824 \cdot 0{,}959 + (0{,}109 \cdot 0{,}959 + 0{,}824 \cdot 0{,}065) \cdot -0{,}9$

$= 0{,}65$

ok

2. $\text{corr}(F_1, F_2) = \phi = 0{,}9$

$= 0{,}104 \cdot 0{,}065 + 0{,}824 \cdot 0{,}959 + (0{,}109 \cdot 0{,}959 + 0{,}824 \cdot 0{,}065) \cdot 0{,}9$

$= 0{,}939$

ok

Answer: Since the factors are positive intuively a positive correlation among the factors should be a better fit. We can just by looking at the factor loadings of indicators X_1 and X_2 and conclude that they are almost the same, which indicates a high positive correlation between them.

ok

4c) $Var(x_i) = l_1^2 + l_2^2 + Var(\varepsilon) + 2\, l_1 l_2 \phi$

1. Corr = -0.9: $\Rightarrow 1-(0{,}104^2 + 0{,}824^2 + 2 \cdot 0{,}104 \cdot 0{,}824 \cdot (-0{,}9)) = Var(\varepsilon_1)$

$\Rightarrow Var(\varepsilon_1) = 0{,}464 \approx 46\%$ ok

$Var(\varepsilon_2) = 1-(0{,}065^2 + 0{,}959^2 + 2 \cdot 0{,}065 \cdot 0{,}959 \cdot -0{,}9)$

$= 0{,}188 \approx 19\%$ ok

2. Corr = 0,9: $Var(\varepsilon_1) = 1-(0{,}104^2 + 0{,}824^2 + 2 \cdot 0{,}104 \cdot 0{,}824 \cdot 0{,}9)$

$= 0{,}1559 \approx 16\%$ ok

$\boxed{Var(\varepsilon_2)} = 1-(0{,}065^2 + 0{,}959^2 + 2 \cdot 0{,}065 \cdot 0{,}959 \cdot 0{,}9)$

$= -0{,}036 \approx \boxed{-4\%}$ minus how?

Conclusion: Just like in answer 4b we can conclude that the positive correlation fit better. The positive corr explaines more of the model data.

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvikssalen Anonymous code: MME-0004 Sheet number: 5

5. We have 6 parameters:

- λ_1
- λ_2
- λ_3
- $Var(\delta_1)$
- $Var(\delta_2)$
- $Var(\delta_3)$

And $\frac{p(p+1)}{2}$ equations $\rightarrow \frac{3\cdot(3+1)}{2} = 6$

This means are model have same nr of parameters as nr of equations and that are model is just-identified. We do not release any df.

Equations:

$\lambda_1 \lambda_2 = 0{,}93$

$\lambda_1 \lambda_3 = 0{,}45$

$\lambda_2 \lambda_3 = 0{,}27$

$Var(x_1) = 1{,}20$

$Var(x_2) = 1{,}56$

$Var(x_3) = 2{,}15$

δ_1

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

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

$\rightarrow \lambda_1^2 = 1{,}55 \quad \lambda_1 = \pm 1{,}244 = 1{,}244$

$\lambda_1 = 1{,}244$

$\lambda_2 = 0{,}93 / 1{,}244 = 0{,}747$

$\lambda_3 = 0{,}45 / 1{,}244 = 0{,}36$

$Var(\delta_1) = 1{,}2 - 1{,}244^2 = -0{,}347$ negative, how?

$Var(\delta_2) = 1{,}56 - 0{,}747^2 = 1$

$Var(\delta_3) = 2{,}15 - 0{,}36^2 = 2$

→

S_2:

$$\left.\begin{aligned} \lambda_1 \lambda_2 &= -0{,}93 \\ \lambda_1 \lambda_3 &= -0{,}45 \\ \lambda_2 \lambda_3 &= -0{,}27 \end{aligned}\right\} \rightarrow$$

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

$$\lambda_1 \cdot (-0{,}27) = -0{,}93 \cdot \frac{-0{,}45}{\lambda_1}$$

$$\lambda_1^2 = -1{,}55 \quad \lambda_1 = \pm \sqrt{1{,}55}\, i$$

Not a unique solution !

Since it's not unique for λ_1 I can expect the same results for λ_2 and λ_3 and therefore I don't proceed with S_2 unique

ok

S_3:

$$\left.\begin{aligned} \lambda_1 \lambda_2 &= -0{,}93 \\ \lambda_1 \lambda_3 &= -0{,}45 \\ \lambda_2 \lambda_3 &= 0{,}27 \end{aligned}\right\}$$

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

$$\rightarrow \lambda_1 \cdot 0{,}27 = -0{,}93 \cdot \frac{-0{,}45}{\lambda_1}$$

Why results for S_1 and S_3 are the same?

$$\lambda_1 = \sqrt{1{,}55} = \pm 1{,}244$$

$$\lambda_2 = -0{,}747$$

$$\lambda_3 = -0{,}36$$

$$Var(\delta_1) = 1{,}2 - 1{,}244^2 = 0{,}347$$

$$Var(\delta_2) = 1{,}56 - (-0{,}747)^2 = 1$$

$$Var(\delta_3) = 2{,}15 - (-0{,}36^2) = 2$$

The parameter estimates for S1 and S3 are unique since we have the same nr of equations as parameters. S2 don't have a solution with real numbers. S1 is the cov matrix which fits best since the covariances are all positive with the positive factors. It is also jus-identified with only one solution. All the variances are unique.

4

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvikssalen Anonymous code: MNE-0004 Sheet number: 6

6 a) Specific variances:

X_1 $Var(\varepsilon_1) = 1 - 0{,}9^2 - 0{,}2^2 = 0{,}15$ +

X_2 $Var(\varepsilon_2) = 1 - 0{,}7^2 - 0{,}15^2 = 0{,}4875$ +

X_3 $Var(\varepsilon_3) = 1 - 0{,}2^2 - 0{,}9^2 = 0{,}15$ +

X_4 $Var(\varepsilon_4) = 1 - 0{,}2^2 - 0{,}7^2 = 0{,}47$ +

High specific variance indicates a bad fit, which means a lot of the variance is left unexplained.

b-c) communalities!!

	F_1	F_2	F_1^2	F_2^2	Shared F1	Shared F2
X_1	0,90	0,2	0,81	0,04	58,7%	2,9%
X_2	0,7	0,15	0,49	0,0225	35,5%	1,7%
X_3	0,2	0,9	0,04	0,81	2,9%	59,4%
X_4	0,2	0,7	0,04	0,49	2,9%	35,9%
		Σ	1,38	1,3625	≈ 100%	≈ 100%

Communalities is the square of the loadings and also the variance explained by each factor of an indicator.

Proportion explained by each factor: $F1 = \frac{1{,}38}{1{,}38 + 1{,}3625}$

$= 50{,}3\%$

$F2 = 49{,}7\%$

ok

→

6d) Estimated corr matrix:

$Corr(X_i, X_j) = l_{i1} l_{j1} + l_{i2} l_{j2}$

$Corr(X_1, X_2) = 0,4 \cdot 0,7 + 0,2 \cdot 0,15$

$Corr(X_3, X_4) = 0,2 \cdot 0,2 + 0,9 \cdot 0,7$

	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

The estimates seems to be quite close to the correlations of the sample corr matrix, which means it's a good fit.

Residual matrix is sampled corr matrix − est. corr matrix

	X_1	X_2	X_3	X_4
X_1	0			
X_2	0,04	0		
X_3	−0,06	0,025	0	
X_4	0,03	0,025	0,07	0

ok

ok

e) $RMSR = \sqrt{\dfrac{0,04^2 + 0,06^2 + 0,025^2 + 0,025^2 + 0,03^2 + 0,07^2}{4}} = 0,05$

Don't remember if it was n-1 or just n

With n-1 RMSR = 0,06. Anyway it's a low nr which is good and it means our estimated model fits quite well.

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvikssalen Anonymous code: MMG-0004 Sheet number: 7

7. Assumptions:

- Var uncorr among variables
- Prin 1 accounts for max variance, prin 2 for second m... etc.
- Seen as data reduction

PCA: Will a rotation of axis help in explaining the variance of the linear combinations?

- Data should have correlated variables which can be reduced to only a few that could explain max of the variance.

Ex: Create an index of crime for cities in Sweden

FA: Is there corr between variables and unexplained factors? More focused on if the data is suitable rather than specific assumptions. In EFA we explore. In CFA we set hypothesis and test.

Ex: Measure intelligence of people based on their grade - exams

CA: Could we cluster the data? Can clusters make any sense? Homogenous with respect to certain characteristics within each group. Each group diff with respect to some characteristics.

Ex: Defining clusters for different types of purchases i.e. gender, income etc.

DA: - Multivariate normally distributed
- Independent
- Equal variance of variables among groups

Ex: We want to divide firms into groups depending on if they are successful or not based on financial status. See question 2. Could also use MANOVA.

LogR: - No distribution assumption
- More if the problem could be solved by a binary outcome.
- Is the binary cut/variables
- Do we have several uncorrelated variables that may be used in the regression.

Ex: Explain risks associated with heartattack based on lifestyle.

ok

Department of Statistics

Stockholms universitet

Correction sheet

Date: 14/12 - 2015

Room: Brunnsvikssalen

Exam: Multivariate Methods

Course: Multivariate Methods

Anonymous code: MME-0009

[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 7F
Teacher's notes 9↑	10		14	9 ↘	14~	10			

Points	Grade	Teacher's sign.
66 + 12 = 78	C	AA

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvik Anonymous code: MME-0009 Sheet number: 1

7.

a) $A = (3, -2)$
$B = (5, 1)$

for A:

$X_1^* = \cos(20) \cdot 3 + \sin(20) \cdot -2 = 2{,}135$ ✓

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

For B:

$X_1^* = \cos(20) \cdot 5 + \sin(20) \cdot 1 = 5{,}04$ ✓

$X_2^* = -\sin(20) \cdot 5 + \cos(20) \cdot 1 = -0{,}77$ ✓

New coordinates are: $A = (2{,}135, -2.905)$

$B = (5.04, -0.77)$

ok

b) $A = (5, 2)$ $A^* = (3.69, 3.939)$ Find θ:

$\cos(\theta) \cdot 5 + \sin(\theta) \cdot 2 = 3{,}69$ (1)

$-\sin(\theta) \cdot 5 + \cos(\theta) \cdot 2 = 3{,}939$ (2)

Solve for (1):

$5\cos(\theta) + 2\sin(\theta) = 3{,}69 \Leftrightarrow 5\cos(\theta) = 3{,}69 - 2\sin(\theta)$

$\Leftrightarrow \cos(\theta) = 0{,}738 - 0{,}4\sin(\theta)$

Solve for (2):

$-5\sin(\theta) + 2(0{,}738 - 0{,}4\sin(\theta)) = 3{,}939 \Leftrightarrow -5\sin(\theta) + 1{,}476 - 0{,}8\sin(\theta)$

$= 3{,}939 \Leftrightarrow -5{,}8\sin(\theta) = 2{,}463 \Leftrightarrow \sin(\theta) = -0{,}424655$

$\sin^{-1}(-0{,}424655) = 25{,}13^\circ \Rightarrow \underline{\theta \approx 25^\circ}$ (1)

ok till here → which direction happens rotation?

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvik Anonymous code: MME-0009 Sheet number: 2

2.

a) Mean corrected = $(X_i - \bar{X})$

NO, the result will not be affected for FA, PCA when data is in mean corrected form since you are using that formula when calculating for example PCA. ok

b)

Standardized form = $\frac{\text{mean corrected}}{\text{Standard Deviation}}$

Yes, transforming the data into standardized form will affect FA or PCA since you would get new values using that formula for standardizing data. With mean corrected formula you would still get the same values in the end for covariance matrix for example. ok

c) ① = Failed Firms ② = Non-Failed Firms

Obs	EBITASS₁	ROTC₁	Obs	EBITASS₂	ROTC₂
1.	0,158	0,182	11.	-0,012	-0,031
2.	0,210	0,206	12.	0,036	0,053
3.	0,207	0,188	13.	0,038	0,036
4.	0,280	0,236	14.	-0,063	-0,074
5.	0,197	0,193	15.	-0,054	-0,119
6.	0,227	0,173	16.	0,000	-0,005
7.	0,148	0,196	17.	0,005	0,039
8.	0,254	0,212	18.	0,091	0,122
9.	0,079	0,147	19.	-0,036	-0,072
10	0,149	0,128	20	0,045	0,064

M(EBITASS₁) = 0,1909 M(EBITASS2) = 0,005 M(EBITASS) = 0,098

M(ROTC₁) = 0,1861 M(ROTC2) = 0,0013 M(ROTC) = 0,089

SSCP₁ $SS = \sum_{i=1}^{n} (X_i - \bar{X}_i)^2$ $SCP = \sum_{i=1}^{n} (X_i - \bar{X})(X_j - \bar{X})$

$SS_{E1} = 0{,}031 \rightarrow$ EBITASS₁

$SS_{R1} = 0{,}009 \rightarrow$ ROTC₁

$SCP_1 = 0{,}013$

$SSCP_1 = \begin{pmatrix} 0{,}031 & 0{,}013 \\ 0{,}013 & 0{,}009 \end{pmatrix}$

SSCP₂

$SS_{E2} = 0{,}021 \rightarrow$ EBITASS₂

$SS_{R2} = 0{,}050 \rightarrow$ ROTC₂

$SCP_2 = 0{,}031$

$SSCP_2 = \begin{pmatrix} 0{,}021 & 0{,}031 \\ 0{,}031 & 0{,}050 \end{pmatrix}$ ok.

$SSCP_W = SSCP_1 + SSCP_2$

$\boxed{SSCP_W = \begin{pmatrix} 0{,}052 & 0{,}044 \\ 0{,}044 & 0{,}059 \end{pmatrix}}$

MD

$SSCP_t = SSCP_W + SSCP_b \iff SSCP_b = SSCP_t - SSCP_W$

SSCPt

$SS_t = 0,3609$ → EBITASS TOT

$SS_R = 0,2304$ → ROTC TOT

$SCP = 0,2158$

$$SSCP_t = \begin{pmatrix} 0,361 & 0,216 \\ 0,216 & 0,230 \end{pmatrix}$$ ⇒ Total sscp matrix

$$SSCP_W = \begin{pmatrix} 0,052 & 0,044 \\ 0,044 & 0,059 \end{pmatrix}$$ ⇒ within-group SSCP matrix

$$SSCP_b = \begin{pmatrix} 0,309 & 0,172 \\ 0,172 & 0,171 \end{pmatrix}$$ → between-group sscp matrix

From these matrices it can be seen that the within-group variance is quite small comparing to the variance for between and total matrices.

Oh

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvik Anonymous code: MME-0009 Sheet number: 3

4.

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

① $Corr(F_1, F_2) = \phi_{12} = -0{,}9$
② $Corr(F_1, F_2) = \phi_{12} = 0{,}9$

a)

	F_1	F_2
X_1	0,104	0,824
X_4	0,906	0,134
X_6	0,827	0,016

ok

The pattern loadings will not change depending on the correlation.

b)

For ① $\phi_{12} = -0{,}9$:

$Corr(X_1, X_2) = 0{,}104 \cdot 0{,}065 + 0{,}824 \cdot 0{,}959 + (0{,}104 \cdot 0{,}959 + 0{,}824 \cdot 0{,}065) \cdot -0{,}9$
$= 0{,}659$

ok

For ② $\phi_{12} = 0{,}9$

$Corr(X_1, X_2) = 0{,}104 \cdot 0{,}065 + 0{,}824 \cdot 0{,}959 + (0{,}104 \cdot 0{,}959 + 0{,}824 \cdot 0{,}065) \cdot 0{,}9$
$= 0{,}935$ ok

When the correlation between the factors is positive, the correlation between the variables gets higher than when the correlation between the factors is negative. This is because the pattern loading for each factor already is positive. ok

c) $Var(U_i)$ is the variance not accounted by the common factor. Looking at the formula sheet, $Var(U_i)$ can be calculated by: (The variance of a variable is 1)

For $\phi_{12} = -0{,}9$ $Var(U_1) = 1 - 0{,}104^2 - 0{,}824^2 - 2 \cdot 0{,}104 \cdot 0{,}824 \cdot -0{,}9 = 0{,}464 = 46{,}4\%$
$Var(U_2) = 1 - 0{,}065^2 - 0{,}959^2 - 2 \cdot 0{,}065 \cdot 0{,}959 \cdot -0{,}9 = 0{,}188 = 18{,}8\%$

ok

For $\phi_{12} = 0{,}9$
$Var(U_1) = 1 - 0{,}104^2 - 0{,}824^2 - 2 \cdot 0{,}104 \cdot 0{,}824 \cdot 0{,}9 = 0{,}156 = 15{,}6\%$ ~~ok~~
$Var(U_2) = 1 - 0{,}065^2 - 0{,}959^2 - 2 \cdot 0{,}065 \cdot 0{,}959 \cdot 0{,}9 = -0{,}036 = -3{,}6\%$ ok

minus, how?

When $\phi_{12} = -0{,}9$ the variance for U_i is higher than for $\phi_{12} = 0{,}9$. This means that $\phi_{12} = 0{,}9$ is a better fit since you want the specific variance to be as small as possible.

+

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvik Anonymous code: MME-0009 Sheet number: 4

5.
$$X_1 = \lambda_1 \xi + \delta_1$$
$$X_2 = \lambda_2 \xi + \delta_2$$
$$X_3 = \lambda_3 \xi + \delta_3$$

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

$$S_1 = \begin{pmatrix} 1,20 & 0,93 & 0,45 \\ 0,93 & 1,56 & 0,27 \\ 0,45 & 0,27 & 2,15 \end{pmatrix} \quad S_2 = \begin{pmatrix} 1,20 & -0,93 & -0,45 \\ -0,93 & 1,56 & -0,27 \\ -0,45 & -0,27 & 2,15 \end{pmatrix} \quad S_3 = \begin{pmatrix} 1,20 & -0,93 & -0,45 \\ -0,93 & 1,56 & 0,27 \\ -0,45 & 0,27 & 2,15 \end{pmatrix}$$

For S_1:

$\lambda_1\lambda_2 = 0,93 \Leftrightarrow \lambda_1 = \frac{0,93}{\lambda_2}$

$\lambda_1\lambda_3 = 0,45 \Leftrightarrow \lambda_1 = \frac{0,45}{\lambda_3}$

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

$\lambda_3\lambda_2 = 0,27 \Leftrightarrow \lambda_3 = \frac{0,27}{\lambda_2}$

$\frac{0,27}{\lambda_2} = \frac{0,45\lambda_2}{0,93} \Leftrightarrow 0,45\lambda_2^2 = 0,2511$

$\lambda_2^2 = 0,558 \Rightarrow \lambda_2 = 0,747$ ok

$\lambda_1 = \frac{0,93}{0,747} = 1,245$ $\lambda_3 = \frac{0,27}{0,747} = 0,361$ ok ok

$Var(\delta_1) = 1,20 - 1,245^2 = -0,35$ negative, how?
$Var(\delta_2) = 1,56 - 0,747^2 = 1,00$
$Var(\delta_3) = 2,15 - 0,361^2 = 2,02$

For S_2:

$\lambda_1\lambda_2 = -0,93 \Leftrightarrow \lambda_1 = \frac{-0,93}{\lambda_2}$

$\lambda_1\lambda_3 = -0,45 \Leftrightarrow \lambda_1 = \frac{-0,45}{\lambda_3}$

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

$\lambda_3\lambda_2 = -0,27 \Leftrightarrow \lambda_3 = \frac{-0,27}{\lambda_2}$ $\quad \frac{-0,27}{\lambda_2} = \frac{-0,45\lambda_2}{-0,93} \Leftrightarrow -0,45\lambda_2^2 = 0,2511 \Leftrightarrow \lambda_2^2 = -0,558$

$\lambda_2^2 = -0,558$ is not possible. If I would proceed calculating the variance for the error term would be the same as for S_1. ok.

For S_3:

$\lambda_3\lambda_2 = 0,27$ $\quad \frac{0,27}{\lambda_2} = \frac{-0,45\lambda_2}{-0,93} \Leftrightarrow -0,45\lambda_2^2 = -0,2511 \Leftrightarrow \lambda_2^2 = 0,558 \Rightarrow \lambda_2 = 747$

$\lambda_1 = -\frac{0,93}{0,747} = -1,245$ $\quad \lambda_3 = -0,361$

The error variance is the same for S_1 and S_3

$Var(\delta_1) = 1,20 - (-1,245^2) = -0,35$
$Var(\delta_2) = 1,56 - 0,747^2 = 1,00$
$Var(\delta_3) = 2,15 - (-0,361^2) = 2,02$

The covariance matrix S_2 goes beyond real numbers.

any interpretation, explanation?

ok.

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvik Anonymous code: MMTE-0009 Sheet number: 5

6. Corr matrix for a hypothetical dataset:

	X_1	X_2	X_3	X_4
X_1	1,0			
X_2	0,7	1,0		
X_3	0,3	0,25	1,0	
X_4	0,35	0,2	0,6	1,0

Estimated factor loadings:

	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

a)

$Var(U_1) = 1 - 0,90^2 - 0,20^2 = 0,15$
$Var(U_2) = 1 - 0,70^2 - 0,15^2 = 0,49$
$Var(U_3) = 1 - 0,20^2 - 0,90^2 = 0,15$
$Var(U_4) = 1 - 0,20^2 - 0,70^2 = 0,47$

ok

Specific variance is the variance not accounted by the common factor F_1 and F_2. A high specific variance means that F_1 & F_2 does not explain much of the variance which is a bad thing. Here can be seen that model 1 and 3 have a lower specific variance, they are a better fit.

+

b) Communalities:

	F_1	F_2
X_1	$0,90^2$	$0,20^2$
X_2	$0,70^2$	$0,15^2$
X_3	$0,20^2$	$0,90^2$
X_4	$0,20^2$	$0,70^2$

Communalities is the squared factor loadings and is the same as the shared variance for uncorrelated models +

misunderstanding of question

Shared variance:

$X_1: 0,90^2 + 0,20^2 = 0,85 = 85\%$
$X_2: 0,70^2 + 0,15^2 = 0,51 = 51\%$
$X_3: 0,20^2 + 0,90^2 = 0,85 = 85\%$
$X_4: 0,20^2 + 0,70^2 = 0,53 = 53\%$

Shared variance is 1 - specific variance. Here you want the shared variance to be as high as possible.

ok

should add to 100%

c)

Variance of $F_1 = 0,90^2 + 0,70^2 + 0,20^2 + 0,20^2 = 1,38$ ok

Variance of $F_2 = 0,20^2 + 0,15^2 + 0,90^2 + 0,70^2 = 1,13625$ ok

Total variance: $1,38 + 1,13625 = 2,7425$

Proportion:

$F_1 = \frac{1,38}{2,7425} = 0,503$ $F_2 = \frac{1,13625}{2,7425} = 0,497$ +

The proportion of variance is almost the same for F_1 and F_2.

d)

$Corr(X_1, X_2) = 0{,}90 \cdot 0{,}70 + 0{,}20 \cdot 0{,}15 = 0{,}66$
$Corr(X_1, X_3) = 0{,}90 \cdot 0{,}20 + 0{,}20 \cdot 0{,}90 = 0{,}36$
$Corr(X_1, X_4) = 0{,}90 \cdot 0{,}20 + 0{,}20 \cdot 0{,}70 = 0{,}32$
$Corr(X_2, X_3) = 0{,}70 \cdot 0{,}20 + 0{,}15 \cdot 0{,}90 = 0{,}275$
$Corr(X_2, X_4) = 0{,}70 \cdot 0{,}20 + 0{,}15 \cdot 0{,}70 = 0{,}245$
$Corr(X_3, X_4) = 0{,}20 \cdot 0{,}20 + 0{,}90 \cdot 0{,}70 = 0{,}67$

ok

Correlation matrix:

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

ok

how close are estimated and original correlation matrices.

To see how good the estimates are, you must look at the residual matrix. you want the residual matrix to be as small as possible. ok.

e)

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

ok

Residual Matrix

$$RMSR = \sqrt{\frac{\sum_{i=1}^{6} res_i^2}{\left(\frac{p(p-1)}{2}\right)}} = \sqrt{\frac{0{,}04^2 + (-0{,}06^2) + 0{,}03^2 + (-0{,}025^2) + (-0{,}045^2) + (-0{,}07^2)}{6}}$$

$= 0{,}0477$ ok

RMSR is quite small, which means that the estimates are good, i.e. they are close to the hypothetical data set.

and F_1, F_2 are able to explain ...

f.

VG.

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvik Anonymous code: MME-0009 Sheet number: 6

7.

PCA:

* Assumptions: With a change in axis, new variables can explain the variance better.
* Method: If you want to look at crime based index and see which city in a certain country that has the highest crime rate.

FA:

* Assumptions: Means of indicators, common factors and unique factors are zero.
  - Variance of indicators and common factors are one
  - The unique factors are not correlated among themselves or with the common factors.
* Method: We want to see the correlation with a child's grade and the interest of the subject.

CA:

* Assumptions: - Each group/cluster are homogeneous with respect to certain characteristic
  - Each group/cluster are different with respect to certain characteristic.
* Method: We want to see how people have voted similar in a certain subject.

DA:

* Assumptions: Multivariate normality, equality of covariance matrices
* Method: We want to divide firms into those who have failed by certain financial variables.

LogR:

* Assumptions: No distribution assumption
* Method: We want to see the probability of a firm being successful given a certain financial ratios.

VG.