Stockholms universitet

STATISTISKA INSTITUTIONEN HT-2015

Written exam in Multivariate Methods, 7.5 ECTS credits
Thursday, 29th October 2015, 16:00 – 21:00
Time allowed: FIVE hours
Examination Hall: Laduvikssalen

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° 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 counter 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) Represent the data in mean corrected form. Will the results of the statistical techniques (e.g. factor analysis, principal component analysis) be affected by mean correcting the data? Why or why not?

b) Represent the data in 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
11	0.200	0.150	23	-0.026	-0.024
12	0.187	0.191	24	0.016	0.026

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

$B = \frac{n_1 n_2}{n_1 + n_2}(\vec{\mu}_1 - \vec{\mu}_2)(\vec{\mu}_1 - \vec{\mu}_2)'$, where B is between-groups **SSCP** matrix for p variables, μ_1 and μ_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} = -0.4. Repeat all your calculations in assumption that correlation changed to Corr(F_1, F_2)= Φ_{12} = 0.4 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 two students give two different sample covariance matrixes of the indicators:

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

Note that the difference is only in the sign of covariance. 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 both 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 one entry of the covariance matrix has influenced the estimates. Use intuition if calculations go beyond real numbers.

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	0.80	0.20
X_2	0.70	0.15
X_3	0.10	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= $\begin{vmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{vmatrix}$

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

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 $H_0\colon \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ $H_a\colon \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}(\boldsymbol{\theta})$ (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 $\lambda = \frac{\boldsymbol{\gamma}'\boldsymbol{B}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\boldsymbol{W}\boldsymbol{\gamma}}$

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

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

MULTIVARIATE METHODS

PCA

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

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

FA

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

- 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	$X_1 = \lambda_1\xi + e_1$ $\vdots$ $X_p = \lambda_p\xi + \varepsilon_p$	$V(x_j) = \lambda_j^2 + V(\varepsilon_j)$	$Cor(x_j,\xi) = \lambda_j$	λ_j^2	$Corr(x_j,x_k) = \lambda_j\lambda_k$
Two-factor model	$X_1 = \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \varepsilon_1$ $\vdots$ $X_p = \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \varepsilon_p$	$V(x_j) = \lambda_{j1}^2 + \lambda_{j2}^2 + 2\lambda_{j1}\lambda_{j2}\phi + V(\varepsilon_j)$	$Cor(x_j,\xi_2) = \lambda_{j2} + \lambda_{j1}\phi$ $Cor(x_j,\xi_1) = \lambda_{j1} + \lambda_{j2}\phi$	$(\lambda_{j2} + \lambda_{j1}\phi)^2$ $(\lambda_{j1} + \lambda_{j2}\phi)^2$	$Cor(x_j,x_k) = \lambda_{j1}\lambda_{k1} + \lambda_{j2}\lambda_{k2} + [\lambda_{j1}\lambda_{k2} + \lambda_{j2}\lambda_{k1}]\phi$

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

CFA

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

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

CA

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

DA (two groups)

$\lambda = \frac{SS_B}{SS_w} \rightarrow max$

If $\xi = \bar{X}^T\gamma \Rightarrow$ 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

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

$\ln\frac{p}{1-p} = \beta_0 + \beta_1X_1 + \ldots + \beta_kX_k$

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

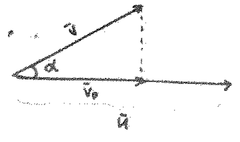
If $Y = \beta_0 + \beta_1X$:
$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$

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

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

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

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



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

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

Department of Statistics

Stockholms universitet

Correction sheet

Date: 29/10 - 2015

Room: Laduvikssalen

Exam: Multivariate methods

Course: Multivariate methods

Anonymous code: MME-0014

[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			7
Teacher's notes 10	9	0	13	8	15	8			

Points	Grade	Teacher's sign.
63 + 19	B	AA

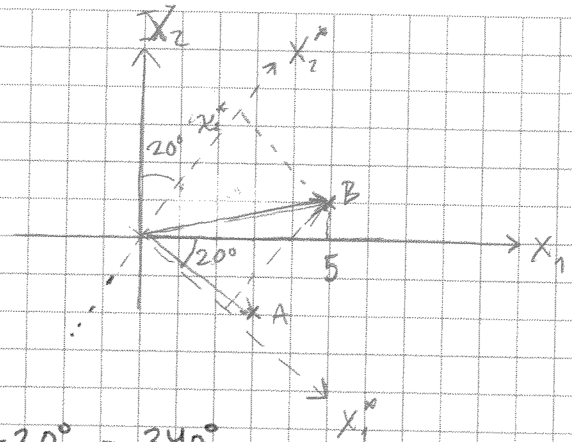
82

SU, DEPARTMENT OF STATISTICS

Room: Laduviksalen Anonymous code: MME-0014 Sheet number: 1

1) $A = (3, -2)$

$B = (5, 1)$



$X_1^* = 20° \rightarrow 360° - 20° = 340°$

~~$X_1^* = \cos 340° \cdot 3 + \sin 340° \cdot (-2) = 1.92$~~

~~$X_2^* = -\sin 340° \cdot 5 + \cos 340° \cdot 1 =$~~

A: $X_1^* = \cos 340° \cdot 3 + \sin 340° \cdot (-2) = 3.503$

$X_2^* = -\sin 340° \cdot 3 + \cos 340° \cdot (-2) = -0.853$

B: $X_1^* = \cos 340° \cdot 5 + \sin 340° \cdot 1 \approx 4.356$

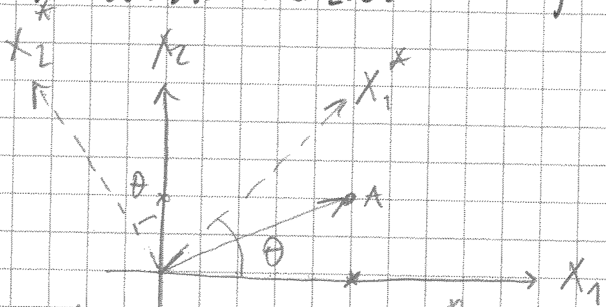
$X_2^* = -\sin 340° \cdot 5 + \cos 340° \cdot 1 \approx 2.650$

Answer: A: $(3.503, -0.853)$

B: $(4.356, 2.650)$

ok +

b) $A = (5, 2)$



A with respect to X_1^* and X_2^* $(3.69, 3.939)$ find A

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

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

$3.69 = \cos\theta \cdot 5 + \sin\theta \cdot 2 \quad (1)$

$3.939 = -\sin\theta \cdot 5 + \cos\theta \cdot 2 \quad (2)$

$5\cos\theta + 2\sin\theta = 3.69 \Rightarrow 5\cos\theta = 3.69 - 2\sin\theta$

$\cos\theta = \dfrac{3.69 - 2\sin\theta}{5}$ in (2): $\Rightarrow$

$3.939 = -5\sin\theta + 2\left(\dfrac{3.69 - 2\sin\theta}{5}\right) \Rightarrow 3.939 - \dfrac{7.38}{5} = -5\sin\theta - \dfrac{4 - 4\sin\theta}{5}$

$$2.463 = -5\sin\theta - \frac{4\sin\theta}{5} \Rightarrow 2.463 = -5.8\sin\theta$$

$$\sin\theta = -\frac{2.463}{5.8} \Rightarrow \theta = \sin^{-1}\left(\frac{-2.463}{5.8}\right) \quad \theta = -25.128^\circ$$

$$360^\circ + (-25.128^\circ) = 334.87^\circ \text{ clockwise}$$

? counter?

ok

(+)

Answer: 334.87° clockwise

---

② a)

	EBITASS	~~EBIT~~ EBITASS_m		~~EBI~~TASS_st	ROTC	ROTC_m	ROTC_st
1	0.158	−0.18533 =	−0.02733 /V	−12.9526	0.182	−0.0015	−1.6426
2	0.210	−〃−	0.02467 −〃−	11.6919	0.206	0.0225	24.64
3	0.207	−〃−	0.02167 :	10.2701	0.188	−0.0045	−4.93
4	0.208	−〃−	0.02267 :	10.744	0.236	0.0525	57.49
5	0.197	−〃−	0.01167 :	5.5308	0.193	0.0095	10.40
6	0.227	−〃−	0.04167 :	19.7488	0.173	−0.0105	−11.50
7	0.148	−〃−	−0.03733 :	−17.6919	0.196	0.0125	13.69
8	0.254	:	0.06867 :	32.5450	0.212	0.0285	31.21
9	0.079	:	−0.10633 :	−50.3933	0.147	−0.0365	−39.97
10	0.149	:	−0.03633	−17.21800	0.128	−0.0555	−60.78
11	0.200	:	0.01467 :	6.9526	0.150	−0.0335	−36.68
12	0.187		0.00167 .	0.7915	0.191	0.0075	8.213

$\bar{\mu} = 0.18533$ mean corrected Standardized

$V(EBITASS) = \frac{\sum(x_i - \bar{x})^2}{df} = 0.00219$

$\bar{\mu} = 0.1835$

$V(ROTC) = 9.1318 \cdot 10^{-4}$

	EBITASS			EBITASS_m	Ebitass_st		ROTC	ROTC_m	ROTC_st
13	−0.012	−0.0153	0.051	−0.1973	~~−19.95~~ −7.65		−0.031	−0.03225	−6.86
14	0.036	0.0387	0.099	−0.1493	~~14.04~~ 16.35		0.053	0.05175	11.01
15	0.038	0.0347	0.101	−0.1002	17.35	−0.119	0.036	0.03475	7.39
16	−0.063	−0.063	0	−0.2483	−31.65	−0.005	[illegible]	−0.07525	−16.01
17	−0.054	−0.0573	0.099	−0.2393	−28.65	0.039	[illegible]	−0.12025	−25.59
18	0.00	−0.0033	0.063	−0.1853	−1.65	0.122	[illegible]	−0.00625	−1.33
19	0.005	0.0017	0.068	−0.1803	0.85		−0.072	0.03775	8.0319
20	0.091	0.0877	0.154	−0.0943	43.85		0.064	0.12075	25.69
21	−0.036	−0.0393	0.027	−0.2213	−19.65		−0.024	−0.07325	−15.56
22	0.045	0.0417		−0.1403	20.85			0.06275	13.3510
23	−0.026	−0.0293		−0.2113	−14.65			−0.02525	−5.372
24	0.016	0.0127		−0.1693	6.35			0.02475	5.2659

$\bar{\mu} =$ ~~0.18533~~ ~~−0.2063~~ 0.0033

V(EBITASS) = 0.0020

~~−0.031~~
~~0.053~~
~~0.036~~
~~−0.074~~
~~−0.119~~
~~−0.005~~
~~0.039~~
~~0.122~~
~~−0.072~~
~~0.064~~
−0.024
0.026

⟹

$\bar{\mu} = 0.0025 \quad V(ROTC_2) = 0.0047$

2) a) See calcs on other sheet. No, results of the stat. techniques such as FA & principal component analysis will not be affected by using mean corrected data. ok

b) See calcs on previous sheet. ~~No ves~~ Yes results of FA and PCA will be effected and you will get different results using standardized data. The stat methods ~~[illegible]~~ are not scale invariant. ok

c) Compute $SSCP_T$, $SSCP_W$ and $SSCP_B$: What conclusions may be drawn from these matrices?

$SS_{1E} = \sum x_{ij}^2 = 0.02167$ (EBITASS)

$SS_{1RO} = \sum x_{ik}^2 = 0.010045$ (ROTC

$SCP_1 = \sum x_{ij}x_{ik} = 0,0083$

$$SSCP_1 = \begin{pmatrix} 0.02167 & 0.0083 \\ 0.0083 & 0.010045 \end{pmatrix}$$

$SS_{2E} = \sum x_{ij}^2 = 0.02219468$

$SS_{2RO} = \sum x_{ik}^2 = 0.05164$

$SCP_2 = \sum x_{ij}x_{ik} = 0.03222$

$$SSCP_2 = \begin{pmatrix} 0,02219 & 0,0322 \\ 0,0322 & 0,05164 \end{pmatrix} \Rightarrow$$

2) c)

$$SSCP_w = SSCP_1 + SSCP_2 = \begin{pmatrix} 0{,}04386 & 0{,}0405 \\ 0{,}0405 & 0{,}061685 \end{pmatrix}$$

$$\Rightarrow SSCP_t = SSCP_w + SSCP_b \qquad SSCP_b = SSCP_t - SSCP_w$$

$$SSCP_t = \begin{pmatrix} 0{,}1630 & 0{,}04146 \\ 0{,}04146 & 0{,}09116 \end{pmatrix}$$

$$SS_E = \sum x_{ij}^2 = 0{,}1630$$

$$SS_R = \sum x_{ik}^2 = 0{,}09116$$

$$SCP = \sum (x_j \cdot x_k) = 0{,}04146$$

$$SSCP_B = \begin{pmatrix} 0{,}11914 & -9{,}6\cdot 10^{-4} \\ -9{,}6\cdot 10^{-4} & 0{,}02931 \end{pmatrix}$$

The $SSCP_W$, $SSCP_P$, and $SSCP_B$ gives us information on the within group and between group variance as well as the total sum of squares. ~~We~~ The conclusions I can draw is that the variance within the two groups are similar and between groups are larger. However, I worry that w. the number of calculations I have the wrong numbers...

ok

yes, there is mistake

& have to reduce amount of data to see concepts more clearly +-

SU, DEPARTMENT OF STATISTICS

Room: Ladunhssalen Anonymous code: MME-0014 Sheet number: 3

(4)

$X_1 = 0.104F_1 + 0.824F_2 + u_1$
$X_2 = 0.065F_1 + 0.959F_2 + u_2$
$X_3 = 0.065F_1 + 0.725F_2 + u_3$
$X_4 = 0.906F_1 + 0.134F_2 + u_4$
$X_5 = 0.977F_1 + 0.116F_2 + u_5$
$X_6 = 0.827F_1 + 0.016F_2 + u_6$
~~X_7~~
~~X_8~~

$Corr(F_1, F_2) = \phi_{12} = -0.4$ repeat for $\phi_{12} = 0.4$

(a): X_1 : 0.104 on F_1 0.824 on F_2

X_4 : 0.906 on F_1 0.134 on F_2

X_6 : 0.827 on F_1 0.016 on F_2

Cannot redo "calc" for $\phi_{12} = 0.4$

(b): $Corr(X_1, X_2) = (0.104 \cdot 0.065) + (0.824 \cdot 0.959) + (0.104 \cdot 0.959 + 0.824 \cdot 0.065)\phi$

with $\phi_{12} = -0.4$ $Corr(X_1, X_2) = 0.796976 + (-0.0613184)$

$= 0.7356576 \approx 0.7$

with $\phi = 0.4$ $Corr(X_1, X_2) = 0.796976 + 0.0613184$

$= 0.8582944 \approx 0.9$

ok

Since all ~~factor~~ pattern loadings are positive → we know that we have positive correlation between the indicators → with negative correlation between the factors, the correlation between indicators will be smaller than with a positive correlation. A somewhat strange situation.

→

c) $Var(X_1) = \lambda_1^2 + \lambda_2^2 + Var(u_1) + 2\lambda_1\lambda_2\phi_{12}$

$\Rightarrow Var(X_1) = 1$

$\Rightarrow 1 - \lambda_1^2 - \lambda_2^2 - 2\lambda_1\lambda_2\phi_{12} = Var(u_1)$

$= 1 - (0.104)^2 - (0.824)^2 - 2 \cdot 0.104 \cdot 0.824 \cdot \phi_{12}$

w. $\phi_{12} = -0.4$ $\quad \underline{Var(u_1) = 0.3787648 \approx 0.38}$

w. $\phi_{12} = 0.4$ $\quad \underline{Var(u_1) = 0.2416512 \approx 0.24}$

$Var(u_2) = 1 - (0.065)^2 - (0.959)^2 - 2(0.065 \cdot 0.959)\phi_{12}$

w. $\phi_{12} = -0.4$ $\quad Var(u_2) = 0.125962 \approx 0.13$

w. $\phi_{12} = 0.4$ $\quad Var(u_2) = 0.026226 \approx 0.0262$

$\cancel{\phi(X_1, F_1) = 0.104 + 0.824 \cdot \phi_{12}}$

$\cancel{\phi(X_2, F_2) = 0.824 + 0.104 \cdot \phi_{12}}$

ok

Answer: For $\phi = -0.4$. For indicator X_1 38 % of the variance is not accounted for. The corresponding percentage for X_2 is 13 %
← mistake in calculations ~24%
For $\phi = 0.4$ 13 % is not accounted for by X_1 and the corresponding number for X_2 is 2.6 %.

Since the pattern loadings are all positive we would account for more variance if the two factors were positively correlated. With negative correlation between factors and positive correlation between indicators we explain less ⇒ The factors are supposed to be the reason for the correlation between indicators

ok

SU, DEPARTMENT OF STATISTICS

Room: Ladunkssalen Anonymous code: MME-0014 Sheet number: 4

(5) $x_1 = \lambda_1 \xi + \delta_1$

$x_2 = \lambda_2 \xi + \delta_2$

$x_3 = \lambda_3 \xi + \delta_3$

Two different sample covariance matrices of the indicators

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

For S_1:

$\lambda_1^2 + var(\delta_1) = 1.20$ $\quad \lambda_1 \lambda_2 = 0.93$

$\lambda_2^2 + var(\delta_2) = 1.56$ $\quad \lambda_1 \lambda_3 = 0.45$ $\quad \lambda_1 = 0.45/\lambda_3$

$\lambda_3^2 + var(\delta_3) = 2.15$ $\quad \lambda_2 \lambda_3 = 0.27$ $\quad \lambda_2 = 0.27/\lambda_3$

$\Rightarrow \lambda_1 \lambda_2 = 0.93 \Rightarrow \dfrac{0.45 \cdot 0.27}{\lambda_3^2} = 0.93 \quad \lambda_3^2 = (0.93)^{-1} \cdot 0.45 \cdot 0.27$

$\lambda_3^2 \approx 0.1306 \quad \lambda_3 = \underline{\underline{0.3614}} \quad \lambda_1 = 0.45/0.3614 \approx \underline{\underline{1.2450}}$

$\lambda_2 = 0.27/0.3614 = 0.74709 \approx \underline{\underline{0.7471}}$

$Var(\delta_1) = 1.20 - (\overset{1.2450}{\cancel{0.3614}})^2 = -0.3500 \Rightarrow$ Negative, something wrong with this model.

$var(\delta_2) = 1.56 - (0.7471)^2 = 1.0018$ What?

$var(\delta_3) = 2.15 - (0.3614)^2 = 2.019$

For S_1: $\lambda_1 = 1.2450 \quad \lambda_2 = 0.7471 \quad \lambda_3 = 0.3614$

$Var(\delta_1) = \underline{-0.3500} \quad Var(\delta_2) = 1.0018$ $\longrightarrow$

$Var(\delta_3) = 2.019$

We have 6 unknown and 6 equations. ~~The covarianc~~ The problem is just-identified and the parameter estimates are unique for S_2 $\lambda_1^2 + var(\delta_1) = 1.20$ $\lambda_1 \lambda_2 = -0.93$

$\lambda_2^2 + var(\delta_2) = 1.56$ $\lambda_1 \lambda_3 = -0.45$

$\lambda_3^2 + var(\delta_3) = 2.15$ $\lambda_2 \lambda_3 = -0.27$

$\Rightarrow$ $\lambda_1 = \frac{-0.45}{\lambda_3}$ $\lambda_2 = \frac{-0.27}{\lambda_3}$

$\lambda_1 \lambda_2 \Rightarrow \frac{-0.45 \cdot -0.27}{\lambda_3^2} = -0.93$

$\lambda_3^2 = \frac{0.1215}{-0.93} \Rightarrow \lambda_3^2 = -0.1306...$

$\Rightarrow$ this would lead to an imaginary number. i.e calculations would go beyond real numbers

$\Rightarrow$ This situation should not arise as you want the correlation between indicators positive correlated with eachother and the factors in order to have a real solution. There is a problem with model specification for the S_2 covariance matrix. Although both will have unique estimates only S_1 would make any sense. (just identified).

What exactly is wrong?

SU, DEPARTMENT OF STATISTICS

Room: Ladugårdssalen Anonymous code: MME-0014 Sheet number: 5

(6) d) Estimated correlation matrix:

	X1	X2	X3	X4
X1	1.0	0.59	0.26	0.30
X2	0.59	1.0	0.205	0.245
X3	0.26	0.205	1.0	0.65
X4	0.30	0.245	0.65	1.0

$Corr(X_j, X_k) = \lambda_{j1}\lambda_{k1} + \lambda_{j2}\lambda_{k2} \cdot$

$(\lambda_{j1}\lambda_{k2} + \lambda_{j2}\lambda_{k1})\phi$

assume 0

$CORR(X_2, X_1) = 0.70 \cdot 0.80 + 0.15 \cdot 0.20 = 0.59$

$CORR(X_3, X_1) = 0.10 \cdot 0.80 + 0.90 \cdot 0.20 = 0.26$

$CORR(X_4, X_1) = 0.20 \cdot 0.80 + 0.70 \cdot 0.20 = 0.30$

$CORR(X_3, X_2) = 0.10 \cdot 0.70 + 0.90 \cdot 0.15 = 0.205$

$CORR(X_4, X_2) = 0.20 \cdot 0.70 + 0.70 \cdot 0.15 = 0.245$

$CORR(X_4, X_3) = 0.20 \cdot 0.10 + 0.70 \cdot 0.90 = 0.65$

You want the estimated correlation matrix to be close to sample correlation matrix. Comparing the corr. matrix to the one given for the hypothetical data set the numbers appear to ~~ble~~ be close

→ easier to see in ~~an~~ residual matrix or RMSR.

→

e) Computing the residual matrix

$$\begin{bmatrix} 1 & - & - & - \\ 0.7 & 1 & - & - \\ 0.3 & 0.25 & 1 & - \\ 0.35 & 0.20 & 0.6 & 1.0 \end{bmatrix} - \begin{bmatrix} 1 & - & - & - \\ 0.59 & 1 & \cdot & \cdot \\ 0.26 & 0.205 & 1 & \cdot \\ 0.30 & 0.245 & 0.65 & 1 \end{bmatrix} = \begin{bmatrix} 0 & & & \\ 0.11 & 0 & & \\ 0.04 & 0.045 & 0 & \\ 0.05 & -0.045 & -0.05 & \end{bmatrix}$$

$$RMSR : \sqrt{\frac{\sum \underbrace{\ell_{ij}}_{\text{Residual matrix}}{}^2}{6}} = \sqrt{\frac{0.11^2 + 0.04^2 + 0.045^2 + 0.05^2 + (-0.045)^2 + (0.05)^2}{6}}$$

$\Rightarrow RMSR = 0.0616$

The residul matrix show small 'residuals' as discussed in d. The RMSR confirms this with a small value (0.0616). i.e. the ~~residul matrix~~ estimated matrix is close to the given ~~covariance~~ correlation matrix for the data. The factor model explains the correlation in the data in a satisfactory way!

7

METHOD	ASSUMPTIONS TO CHECK
PCA	Will a rotation of axis, i.e. PCA help in explaining the variance. Will interpretation of linear combinations of variables make any sense?
FA	Is there correlation between variables that could be used to explain a factor(s) $\Rightarrow$ More focused on if it is a suitable dataset for FA than on specific assumption of the data

$\Rightarrow$

SU, DEPARTMENT OF STATISTICS

Room: Laduvikssalen Anonymous code: MME-0014 Sheet number: 6

(6)

	X1	X2	X3	X4
X1	1.00			
X2	0.7	1.00		
X3	0.3	0.25	1.00	
X4	0.35	0.2	0.6	1.0

	F1	F2
X1	0.80	0.20
X2	0.70	0.15
X3	0.10	0.90
X4	0.20	0.70

(a) compute specific variance; what does high specific variance indicate

Assume F1 & F2 are uncorrelated

$VAR(X_1) = 0.80^2 + 0.20^2 + VAR(u_1)$

$VAR(u_1) = 1 - 0.8^2 - 0.2^2 = 0.32$

$VAR(u_2) = 1 - 0.7^2 - 0.15^2 = 0.4875$

$VAR(u_3) = 1 - 0.10^2 - 0.90^2 = 0.18$

$VAR(u_4) = 1 - 0.20^2 - 0.70^2 = 0.47$ ok

Specific variance

X_1	0.32
X_2	0.4875
X_3	0.18
X_4	0.47

Specific variance is variance not shared with the factors (F1, F2) and thus the indicator with high specific variance is not good at describing the latent factors, i.e. they do not share ~~variance~~ substantial variance with the factors

(b) Communalities:

	F1	F2
X1	$0.80^2 = 0.64$	$0.20^2 = 0.04$
X2	$0.70^2 = 0.49$	$0.15^2 = 0.0225$
X3	$0.10^2 = 0.01$	$0.90^2 = 0.81$
X4	$0.20^2 = 0.04$	$0.70^2 = 0.49$
total:	1.18	1.325

Total variance: $1.18 + 1.325 = 2.5425$

$\Rightarrow$

The communalities are the shared variance between an indicator and a factor → a high communality would mean a good indicator (such as X_3 for F_2 and X_1 for F_1). ~~The proportion % shared variance is simply how much of the variance for each factor that is explained by the variables. In our case this would be $\frac{1.18}{2.5425}$ = 0.4641 ≈ 46% and $\frac{1.3625}{2.5425}$ ≈ 53.6%. The % shared variance of the variables for both factors are similar.~~

To calculate the % of shared variance we look at the unique variance for each indicator.

$VAR(u_1) = 0.32 \rightarrow \% X_1 = 1 - 0.32 \Rightarrow 68\%$ shared variance w. F1 & F2

$VAR(u_2) = 0.4875 \Rightarrow \% X_2 = 1 - 0.4875 \Rightarrow 51{,}25\%$ — " —

$VAR(u_3) = 0.18 \Rightarrow \% X_3 = 1 - 0.18 \Rightarrow 82\%$ — " —

$VAR(u_4) = 0.47 \Rightarrow \% X_4 = 1 - 0.47 \Rightarrow 53\%$ — " —

X.

% of shared variance is thus the % of an indicators total variation that is shared with the factors. The higher the shared variance the better the indicator.

c) Proportion explained by each factors is the sum of communalities for that factor divided in the total variance for the factors (see calculations on previous page). For $F_1 \Rightarrow \frac{1.18}{2.5425} = 0.46411 \approx 46.4\%$ ok

for $F_2 \Rightarrow \frac{1.3625}{2.5425} = 0.53588 \approx 53.59\%$. I.e the two factors approximately explain equal amount of the total variance.

⇒

SU, DEPARTMENT OF STATISTICS

Room: Ladviksalen Anonymous code: MME-0014 Sheet number: 7

for PCA ?
FA

7 +

METHOD	Assumptions
CA	Could one suspect that the data may be clustered, i.e. would the existence of clusters make sense?
Two group DA	~~Is their~~ Are data multivariate normally distributed? Could we distinguish groups based on means/variances { compute test statistics.
LogR	No distributional assumption. More if the problem may be solved with regression with a binary outcome (categorical yes/no) Do we have several uncorrelated variable that may be used in the regression?

Example problem	Method of choice	alternative 1	2	3
We want to explain risks associated with heart attack based on life-style (habits, smoking etc.	LogR	—	-	—
We want to seperate molecules in two groups based on physical ~~car~~ properties, drug-like and non-drug like	CA	PCA	FA	DA if properties Normally distr.
We want to create a crime based index for swedish cities on crime reports	PCA	FA	—	—
We want to divide firms into groups ~~depending~~ on if they are successful or not based on variables regarding finacial status	DA	CA		

(7). Example

We want to measure intelligence of children based on their grades in a number of school subjects.

Method	1	2	3 ..
	$\boxed{FA}$	—	—