Stockholms universitet

STATISTISKA INSTITUTIONEN

VT-2017

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

You are asked to answer all questions as well as to 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 <u>**allowed**</u> to use a pocket calculator, a language dictionary, and a list of formulas (attached)

The teacher reserves the right to examine the students <u>**orally**</u> on the questions in this examination

1. (12 points) Let us analyse the following 3-variate dataset with 5 observations. Each observation consists of 3 measurements and recorded in the following matrix

$$\begin{matrix} 7 & 4 & 3 \\ 4 & 1 & 8 \\ 6 & 3 & 5 \\ 8 & 5 & 7 \\ 7 & 2 & 9 \end{matrix}$$

What portion of total variance each variable accounts for? Compute the correlation matrix. Next, find the eigenvalues/eigenvectors of the correlation matrix and interpret them in style of PCA. If you get an answer which is a complex number, how you interpret it for the purposes of PCA? How much of total variance the first two principal components "explain"? Have you mean-adjusted and/or standardized the original data set before the analysis: why yes/no?

2. (12 points)

(a) Points A and B have the following coordinates with respect to orthogonal axes X_1 and X_2: $A=(3,-3)$; $B=(7,1)$. If the axes X_1 and X_2 are rotated 200° 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 $(2,2)$. 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 $(2.8284,\ 0)$, find θ. Provide geometric motivation.

3. (8 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, 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. (12 points) Consider the two-indicator two-factor model represented by the following equations:

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

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} = - 0.1. Repeat all your calculations in assumption that correlation changed to Corr(F_1, F_2)= Φ_{12} = 0.1 and discuss the differences in detail. Try to provide intuition for at least some of your answers: without calculating, what you would expect in case correlation is 0.9, 0, -0.9?

(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. (12 points)

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. Analyse and discuss your findings.

Subject ID	Income in tEUR	Education (in years)
S1	17	10
S2	23	12
S3	25	14
S4	28	15
S5	32	20
S6	35	18

6. (12 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.

7. (12 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	0.4	1.000

You cannot be sure in the above calculated correlation matrix as it has been calculated by a novice student to MM techniques. You can assume that the absolute values of the entries have been calculated correctly but the sign of correlations might contain mistake.

Further, the following estimated factor loadings were extracted by the principal axis factoring procedure, starting from the initial data set. It has been done by an expert in the field and the result that has been reported is as follows:

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

Try to reconcile above two matrixes by computing and discussing 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 speculate about the correlation matrix calculated by student (matrix 1); and (e) calculate residual matrix, compute RMSR and interpret.

GOOD LUCK

Formula Sheet, Multivariate Methods

Matrices

Transpose – exchange rows and columns

Identity (I) – diag (1,1...) of order n*n

Inverse of A (A^{-1}): $AA^{-1} = A^{-1}A = I$

$A + B = B + A; x(A + B) = xA + xB; AB \neq BA \ (in\ general);$

If order (A)=m*n, order (B)=n*p, then C=AB is of order m*p

D=det A=$\begin{vmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{vmatrix}$

detA=$a_{i1}A_{i1} + a_{i2}A_{i2} + \dots + a_{in}A_{in}$ where cofactor $A_{ij} = (-1)^{i+j}D_{ij}$ (i-row, j-column of D)

Cramer's rule: $x_j = D_j/D$ where D=detA and D_j is the determinant that arises when the j column of D is replaced by the column elements $b_1, \dots, b_n$. (Ax=b)

Vectors

$\boldsymbol{a} = (a_1 a_2 \dots a_p)$

A right-angle triangle: α - angle between a and c; c – hypotenuse; $\cos\alpha = \frac{a}{c}$, $\sin\alpha = \frac{b}{c}$

Length of vector a=$\|\boldsymbol{a}\| = \sqrt{a_1^2 + a_2^2}$

Basis vectors $\boldsymbol{e_1} = (1\ 0), \boldsymbol{e_2} = (0\ 1)$

$\boldsymbol{a} = a_1\boldsymbol{e_1} + a_2\boldsymbol{e_2}$

Scalar product $\boldsymbol{ab} = a_1b_1 + a_2b_2 + \dots + a_pb_p$; $\boldsymbol{ab} = \|\boldsymbol{a}\|\|\boldsymbol{b}\|\cos\alpha$

Length of the projection: $\|\boldsymbol{a_p}\| = \|\boldsymbol{a}\|\cos\alpha$

Variance of x_i: $s_1^2 = \frac{\|x_i\|^2}{n-1}$; Generalized variance: $GV = \left(\frac{\|x_1\|\cdot\|x_2\|}{n-1} \cdot \sin\alpha\right)^2$

Distances

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

Statistical: $SD_{ij}^2 = \left(\frac{x_i - x_j}{s}\right)^2$, s-standard deviation

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

Variance, Sum of Squares, and Cross Products

Variance: $s_j^2 = \frac{\sum_{i=1}^{n} x_{ij}^2}{n-1} = \frac{SS}{df}$ (sum of squares/degrees of freedom)

Covariance: $s_{jk} = \frac{\sum_{i=1}^{n} x_{ij}x_{ik}}{n-1} = \frac{SCP}{df}$ (sum of the cross products/degrees of freedom)

SSCP – sum of squares and cross products matrix $\begin{pmatrix} SSX_1 & SCP \\ SCP & SSX_2 \end{pmatrix}$

S – covariance matrix $S_t = \frac{SSCP_t}{df}$

Within-Group Analysis: $SSCP_w = SSCP_1 + SSCP_2$ (pooled SSCP matrix) $S_w = \frac{SSCP_w}{n_1+n_2-2}$ (pooled cov m)

Between-Group Analysis: $SS_j = \sum_{g=1}^{G} n_g(\bar{x}_{jg} - \bar{x}_j)^2$; $SCP_{jk} = \sum_{g=1}^{G} n_g(\bar{x}_{jg} - \bar{x}_j)(\bar{x}_{kg} - \bar{x}_k)$

$SSCP_t = SSCP_w + SSCP_b$

Principal Components Analysis

$x_1^* = \cos\theta * x_1 + \sin\theta * x_2$; $x_2^* = -\sin\theta * x_1 + \cos\theta * x_2$

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

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

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

$y' * x = (y_a \quad y_b \quad y_c)\begin{pmatrix} x_a \\ x_b \\ x_c \end{pmatrix} = y_a x_a + y_b x_b + y_c x_c, \quad x * y' = \begin{pmatrix} x_a \\ x_b \\ x_c \end{pmatrix}(y_a \quad y_b \quad y_c) = \begin{matrix} x_a y_a & x_a y_b & x_a y_c \\ x_b y_a & x_b y_b & x_b y_c \\ x_c y_a & x_c y_b & x_c y_c \end{matrix}$

Formula for standardized data: $x_j = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)}{\sigma_j}$, for the jth variable

In Mahalanobis distance (page 1): S_1 and S_2 is the variance in variable 1 and 2 and r is the correlation

In the variance and covariance formula on page 1, x_{ij} *and* $x_{ij}x_{ik}$ are mean corrected.

$Mean: \ \bar{x}_i = \sum_{i=1}^{n} \frac{x_{ij}}{n}$, for the jth variable

PCA Loadings: $l_{ij} = \frac{w_{ij}}{\hat{S}_j} * \sqrt{\lambda_i}$, where l_{ij} is the loading of the jth variable for the ith principal component, w_{ij} is the weight of the jth variable for the ith principal component, λ_i is the eigenvalue of the ith components and $\hat{S}_j$ is the standard deviation of the jth variable.

Factor analysis: The RMSR are given by the recidual matrix $RMSR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=i}^{p} res_{ij}^2}{p(p-1)/2}}$, where res_{ij} is the residual correlation between the ith and the jth factor and p is the number of factors. This calculation does not include the diagonal (corr between two of the same factors). RMSR should be as low as possible. The residual matrix is produced by subtracting the original correlation matrix from the reproduced correlation matrix.

Confirmatory factor analysis: $\sigma_1^2 = \lambda_1^2 + V(\delta_1), \ldots, \sigma_n^2 = \lambda_n^2 + V(\delta_n) \ \ \sigma_{12} = \lambda_1\lambda_2, \ \sigma_{13} = \lambda_1\lambda_3,$ $\sigma_{23} = \lambda_2\lambda_3$ *and so on*

If a factor is correlated Φ is added between the correlated λ:s

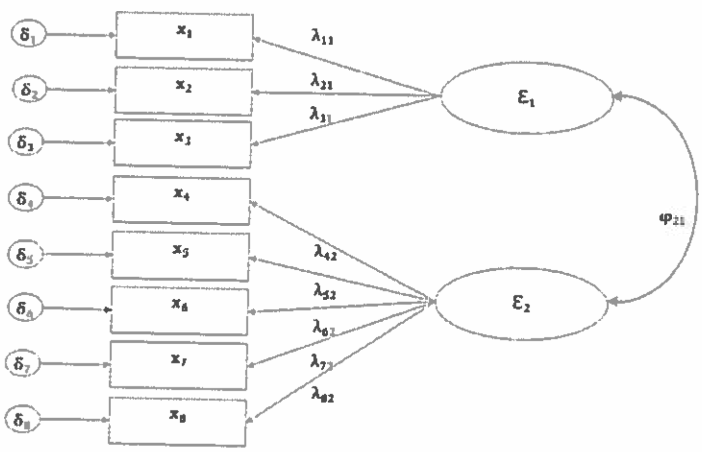
Under-identified: #equations<#unknowns

just-identified: #equations=#unknowns

over-identified: #equations>#unknowns

$$\#equations = \frac{x(x+1)}{2}$$

$$\#unknowns = \xi x + x + \binom{\xi}{2}$$



Example of CFA | Measurement Model

if all correlations between factors are known to be non-zero. If there is no correlation between factors #unknowns $= \xi x + x$

Where x= #indications, ξ= #factors and ϕ = #correlations

Wards method: maximizes within-cluster homogeneity. Clusters are created from the lowest within cluster SS.

$$x^{-1} = \begin{pmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \end{pmatrix}^{-1} = \frac{1}{x_{11}x_{22} - x_{12}x_{21}}\begin{pmatrix} x_{22} & -x_{12} \\ -x_{21} & x_{11} \end{pmatrix}$$

$$y' * x = (y_a \quad y_b \quad y_c)\begin{pmatrix} x_a \\ x_b \\ x_c \end{pmatrix} = y_a x_a + y_b x_b + y_c x_c, \quad x * y' = \begin{pmatrix} x_a \\ x_b \\ x_c \end{pmatrix}(y_a \quad y_b \quad y_c) = \begin{matrix} x_a y_a & x_a y_b & x_a y_c \\ x_b y_a & x_b y_b & x_b y_c \\ x_c y_a & x_c y_b & x_c y_c \end{matrix}$$

Cubic equations:

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

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

Discriminate analysis: is used to find the angel that best separates the means of two groups.

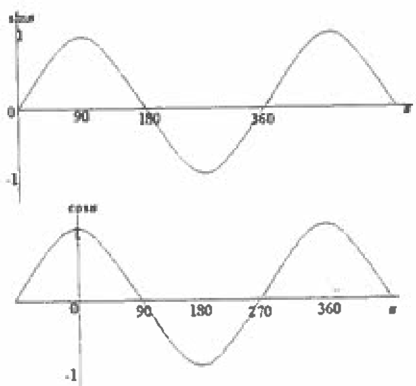
Fishers: $(w^{-1}B - \lambda 1)\gamma = 0 \rightarrow \gamma = \Sigma^{-1}(\mu_1 - \mu_2)$, where Σ is the variance-covariance matrix, γ is $p * 1$ vector of weights and μ_1 *and* μ_2 are the $p * 1$vectors of means for group 1 and 2.

Log-regression: create a contingency table and then calculate the conditional probability of the event you are looking for.

Relative size	Case B	Case $\bar{B}$	Total
Condition A	w	x	w+x
Condition $\bar{A}$	y	z	y+z
Total	w+y	x+z	w+x+y+z

$$P(A|B) \times P(B) = \frac{w}{w+y} \times \frac{w+y}{w+x+y+z} = \frac{w}{w+x+y+z}$$

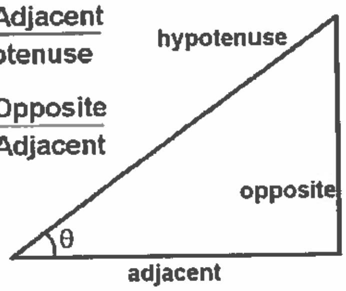
$$P(B|A) \times P(A) = \frac{w}{w+x} \times \frac{w+x}{w+x+y+z} = \frac{w}{w+x+y+z}$$



$$\text{SIN}\,\theta = \frac{\text{Side Opposite}}{\text{Hypotenuse}}$$

$$\text{COS}\,\theta = \frac{\text{Side Adjacent}}{\text{Hypotenuse}}$$

$$\text{TAN}\,\theta = \frac{\text{Side Opposite}}{\text{Side Adjacent}}$$



Department of Statistics

Stockholms universitet

Correction sheet

Date: 15/2 - 2017

Room: Värtasalen

Exam: Multivariate Methods

Course: Multivariate Methods

Anonymous code: MM-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 82
Teacher's notes 10	12	6	12	12	10,	10			

Points	Grade	Teacher's sign.
72 +14	B	AA

86 (B)

SU, DEPARTMENT OF STATISTICS

Room: VÄ Anonymous code: MM-0004 Sheet number: 1

1

X_1	X_2	X_3	$X_1-\bar{X}_1$	$X_2-\bar{X}_2$	$X_3-\bar{X}_3$
7	4	3	0,6	1	-3,4
4	1	8	-2,4	-2	1,6
6	3	5	-0,4	0	-1,4
8	5	7	1,6	2	0,6
7	2	9	0,6	-1	2,6
			V=2,3	V=2,5	V=5,8

$\bar{x}_1 = 6,4$
$\bar{x}_2 = 3$
$\bar{x}_3 = 6,4$

% Variance accounted for by variable

$x_1 = \frac{2,3}{10,6} \cdot 100 = 21,7\%$

$x_2 = 23,58\%$

$x_3 = 54,72\%$

$$SSCP = \begin{bmatrix} 9,2 & 8 & -2,8 \\ 8 & 10 & -8 \\ -2,8 & -8 & 23,2 \end{bmatrix} / n-1 \qquad n=5$$

$$\Sigma = \begin{bmatrix} 1,84 & 1,6 & -0,56 \\ 1,6 & 2 & -1,6 \\ -0,56 & -1,6 & 4,64 \end{bmatrix} / \sqrt{V(x_i)V(x_j)}$$

$$C = \begin{bmatrix} 1 & 0,834 & -0,192 \\ 0,834 & 1 & -0,525 \\ -0,192 & -0,525 & 1 \end{bmatrix}$$ ok

$$|C - I\lambda| = \begin{vmatrix} 1-\lambda & 0,834 & -0,192 \\ 0,834 & 1-\lambda & -0,525 \\ -0,192 & -0,525 & 1-\lambda \end{vmatrix} \begin{matrix} 1-\lambda & 0,834 \\ 0,834 & 1-\lambda \\ -0,192 & -0,525 \end{matrix}$$

$= (1-\lambda)^3 + 2(0,834 \cdot -0,192 \cdot -0,525) - (0,834^2 \cdot (1-\lambda)) - -0,525^2 \cdot (1-\lambda)$
$- -0,192^2 \cdot (1-\lambda)$

$= (1-\lambda)^3 + 0,1681344 - 0,695556 + 0,695556\lambda - 0,275625 + 0,275625\lambda$
$- 0,036864 + 0,036864\lambda$

$= (1-\lambda)^3 + 1,008045\lambda - 0,839906$

$\lambda_3 = 0,093 \qquad \lambda_2 = 0,828 \qquad \lambda_3 = 2,08$ ok

$\lambda_1 + \lambda_2 + \lambda_3 = 2{,}08 + 0{,}821 + 0{,}093 = 3$

% of variance explained by principle component

1: $\frac{2{,}08}{3} \cdot 100 = 69{,}33\%$

2: $\frac{0{,}828}{3} \cdot 100 = 27{,}6\%$. eigen vectors ..

I used mean correction at the beginning. I could have used mean corrected standardized data in the start also, but I chose to first calculate the SSCP, then the covariance matrix and then standardize the cov matrix to get the corr matrix.

Standardized data removes heavy weights if some variable have a much higher variance than the others. In this case variable x_3 had a bit higher variance than x_1 and x_2 making it affect more than it should if not standardized.

ok.

SU, DEPARTMENT OF STATISTICS

Room: V'A Anonymous code: MM-0004 Sheet number: 2

2

a) $A = (3, -3)$ $B = (7, 1)$ 200° clockwise $= 160^\circ$ counter clockwise

$\theta = 160$

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

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

A:

$x_1^* = \cos 160 \cdot 3 + \sin 160 \cdot -3 = -3,845$

$x_2^* = -\sin 160 \cdot 3 + \cos 160 \cdot -3 = 1,793$

$A^* = (-3,845; 1,793)$

ok

B:

$x_1^* = \cos 160 \cdot 7 + \sin 160 = -6,236$

$x_2^* = -\sin 160 \cdot 7 + \cos 160 = -3,334$

$B^* = (-6,236; -3,334)$

b) $A = (2, 2)$

$A^* = (2,8284; 0)$ find θ

$2\cos\theta + 2\sin\theta = 2,8284$

$-2\sin\theta + 2\cos\theta = 0$

$2\cos\theta = 2\sin\theta$

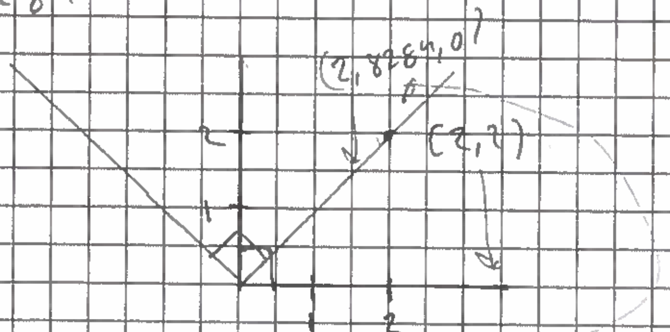
$2\cos\theta + 2\cos\theta = 2,8284$

$4\cos\theta = 2,8284$

$\cos\theta = 0,7071$

$\theta = 45^\circ$

ok



if 45°, $\sqrt{2^2+2^2} = \sqrt{8} = 2,8284$

so 45° is right

SU, DEPARTMENT OF STATISTICS

Room: VÄ **Anonymous code:** MM-0004 **Sheet number:** 3

3

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 for p variable,

μ_1 and μ_2 are p·1 vectors of means for group 1 and 2

n_1 and n_2 are the number of observations in group 1 and 2

$$SSCP_B = \sum_{g=1}^{G} n_g (\bar{x}_{jg} - \bar{x}_j)^2 \quad \text{for one variable } j$$

$$SSCP_B = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2$$

$$n_1(\bar{x}_1^2 - 2\bar{x}_1\bar{x} + \bar{x}^2) + n_2(\bar{x}_2^2 - 2\bar{x}_2\bar{x} + \bar{x}^2)$$

$$n_1\bar{x}_1^2 - 2n_1\bar{x}_1\bar{x} + n_1\bar{x}^2 + n_2\bar{x}_2^2 - 2n_2\bar{x}_2\bar{x} + n_2\bar{x}^2 =$$

$\bar{x} = \frac{\bar{x}_1 + \bar{x}_2}{n_1 + n_2}$

$$n_1\bar{x}_1^2 - 2n_1\bar{x}_1\left(\frac{\bar{x}_1+\bar{x}_2}{n_1+n_2}\right) + n_1\left(\frac{\bar{x}_1+\bar{x}_2}{n_1+n_2}\right)^2 + n_2\bar{x}_2^2 - 2n_2\bar{x}_2\left(\frac{\bar{x}_1+\bar{x}_2}{n_1+n_2}\right) + n_2\left(\frac{\bar{x}_1+\bar{x}_2}{n_1+n_2}\right)^2 =$$

$$= n_1\bar{x}_1^2 + n_2\bar{x}_2^2 + \left(\frac{\bar{x}_1+\bar{x}_2}{n_1+n_2}\right)\left(-2n_1\bar{x}_1 + n_1\left(\frac{\bar{x}_1+\bar{x}_2}{n_1+n_2}\right) - 2n_2\bar{x}_2 + n_2\left(\frac{\bar{x}_1+\bar{x}_2}{n_1+n_2}\right)\right) =$$

$$= n_1\bar{x}_1^2 + n_2\bar{x}_2^2 + \left(\frac{\bar{x}_1+\bar{x}_2}{n_1+n_2}\right)\left(-2n_1\bar{x}_1 + \frac{n_1\bar{x}_1}{n_1+n_2} + \frac{n_1\bar{x}_2}{n_1+n_2} - 2n_2\bar{x}_2 + \frac{n_2\bar{x}_1}{n_1+n_2} + \frac{n_2\bar{x}_2}{n_1+n_2}\right)$$

= Time was running out and I'm sure my way is not the fastest way, which makes it ok that time is running out, should have learnt a faster way.

+1

SU, DEPARTMENT OF STATISTICS

Room: VÄ Anonymous code: MM-0004 Sheet number: 4

4

$X_1 = 0{,}104 F_1 + 0{,}824 F_2 + U_1$
$X_2 = 0{,}065 F_1 + 0{,}459 F_2 + U_2$
$X_3 = 0{,}065 F_1 + 0{,}725 F_2 + U_3$
$X_4 = 0{,}906 F_1 + 0{,}134 F_2 + U_4$
$X_5 = 0{,}977 F_1 + 0{,}116 F_2 + U_5$
$X_6 = 0{,}877 F_1 + 0{,}016 F_2 + U_6$

a) Pattern loading of indicators on factors are not affected by correlation, so the pattern loading are for both $\phi_{12} = -0{,}1$ and $\phi_{12} = 0{,}1$ the following:

X_1: $F_1 = 0{,}104$, $F_2 = 0{,}824$

X_4: $F_1 = 0{,}906$, $F_2 = 0{,}134$

X_6: $F_1 = 0{,}827$, $F_2 = 0{,}016$

ok

b) Correlation between X_1 and X_2

$$\text{Corr}(X_1, X_2) = \lambda_{11} \cdot \lambda_{21} + \lambda_{12} \cdot \lambda_{22} + (\lambda_{11}\lambda_{22} + \lambda_{12}\lambda_{21}) \cdot \phi$$

so without calculating we can see that positive correlation between the common factors will give a higher correlation between the indicators than a negative correlation between the common factor and higher pos correlationship between common factor will give a higher correlation between the indicators than a low correlation between common factor.

+

for $\phi_{12} = -0{,}1$

$$\text{Corr}(X_1, X_2) = \underbrace{0{,}104 \cdot 0{,}065 + 0{,}824 \cdot 0{,}459}_{0{,}796976} + \underbrace{(0{,}104 \cdot 0{,}954 + 0{,}824 \cdot 0{,}065)}_{0{,}153296} \cdot -0{,}1$$

$= 0{,}7816464$ ok

and for $\phi_{12} = 0{,}1$:

$= 0{,}796976 + 0{,}153296 \cdot 0{,}1 = 0{,}8123056$ ok

and for example $\phi_{12} = 0{,}9 = 0{,}796976 + 0{,}153296 \cdot 0{,}9 \approx 0{,}935$
$\phi_{12} = 0 = 0{,}796976$
$\phi_{12} = -0{,}9 \approx 0{,}659$. ok

proving my intuition was right.

c)

What percentage of variance of indicator X_1 and X_2 are not accounted for by the common factors F_1 and F_2.

$V(U_1)$ and $V(U_2)$

$V(U_i) = \underbrace{V(X_i)}_{=1} - \lambda_{i1}^2 - \lambda_{i2}^2 - 2\lambda_{i1}\lambda_{i2}\phi$

$V(U_1) = \underbrace{1}_{B} - \underbrace{(\lambda_{i1}^2 + \lambda_{i2}^2 + 2\lambda_{i1}\lambda_{i2}\phi)}_{A}$

this means, a positive ϕ will make A bigger, making B smaller. A negative ϕ will make A smaller making B bigger. A high ϕ will make A bigger than a small ϕ, and therefor making B smaller.

So, a high positive correlation between the common factors will make the variance of the indicators that are not accounted for by the common factors smaller.

for $\phi_{12} = -0,1$

$V(U_1) = 1 - (\underbrace{0,104^2 + 0,824^2}_{0,689792} + \underbrace{2 \cdot 0,104 \cdot 0,824}_{0,171392} \cdot -0,1) = 0,3273472 \approx 0,327$

$V(U_2) = 1 - (\underbrace{0,065^2 + 0,959^2}_{0,923906} + \underbrace{2 \cdot 0,065 \cdot 0,959}_{0,12467} \cdot -0,1) = 0,08856 \approx 0,089$

% not explained by factors for X_1: $\frac{0,327}{0,690 + 0,327} \cdot 100 = 32,15\%$ X_2: $\frac{0,089}{0,923906 + 0,089} \cdot 100 = 8,7$

for $\phi_{12} = 0,1$ $V(U_1) = 1 - (0,689792 + 0,171392 \cdot 0,1) = 0,293$

% not explained by factors: $\frac{0,293}{0,690 + 0,293} \cdot 100 = 29,81\%$ OK

$V(U_2) = 1 - (0,923906 + 0,12467 \cdot 0,1) = 0,064$

% not explained by factors: $\frac{0,064}{0,924 + 0,064} \cdot 100 = 6,48\%$.

meaning my intuition was right. higher corr $(0,1 > -0,1)$ gives lower variance of indicators not accounted by factors

SU, DEPARTMENT OF STATISTICS

Room: VÄ   Anonymous code: MM-0004   Sheet number: 5

5

Average linkage

	1	2	3	4	5	6
1	0					
2	40	0				
3	80	8	0			
4	146	34	10	0		
5	325	145	85	41	0	
6	388	180	116	58	13	0

$D_{12}^2 = (17-23)^2 + (10-12)^2 = 40$

$D_{13}^2 = (17-25)^2 + (10-14)^2 = 80$

and so on

D_{23} is the lowest distance and form a cluster

	1	23	4	5	6
1	0				
23	60	0			
4	146	22	0		
5	325	115	41	0	
6	388	148	58	13	0

$D_{1\,23} = \frac{D_{12} + D_{13}}{2} = \frac{40+80}{2} = 60$

$D_{4\,23} = \frac{D_{42} + D_{43}}{2} = \frac{34+10}{2} = 22$

and so on

D_{56} is lowest and form a cluster

	1	23	4	56
1	0			
23	60	0		
4	146	22	0	
56	356,5	131,5	49,5	0

$D_{1\,56} = \frac{D_{15} + D_{16}}{2} = \frac{325+388}{2} = 356,5$

$D_{23\,56} = \frac{D_{23\,5} + D_{23\,6}}{2} = \frac{115+148}{2} = 131,5$

$D_{23\,4}$ is lowest and form a cluster

	1	234	56
1	0		
234	103	0	
56	356,5	90,5	0

$D_{1\,234} = \frac{D_{1\,23} + D_{14}}{2} = \frac{60+146}{2} = 103$

$D_{234\,56} = \frac{D_{56\,23} + D_{56\,4}}{2} = \frac{131,5+49,5}{2} = 90,5$

$D_{234\,56}$ is lowest and cluster

	1	23456
1	0	
23456	229,75	0

The 2 groups/clusters using average linkage is

C1 = 1

C2 = 2, 3, 4, 5, 6

Wards method :

SSw

12 = $\left(17 - \frac{17+23}{2}\right)^2 + \left(23 - \frac{17+23}{2}\right)^2 + \left(10 - \frac{10+12}{2}\right)^2 + \left(12 - \frac{10+12}{2}\right)^2 = 20$

13 = $(17-21)^2 + (25-21)^2 + (10-12)^2 + (14-12)^2 = 40$

14 =

15 = All these are further away than 2 from 1. (14,5,6)

16

23 = $(23-24)^2 + (25-24)^2 + (12-13)^2 + (14-13)^2 = 4$

24

25 All these are further away than 3 from 2 (4,5,6)

26

34 = $(25-26,5)^2 + (28-26,5)^2 + (14-14,5)^2 + (15-14,5)^2 = 5$

35

36 Further away than 4 from 3

45 = $(28-30)^2 + (32-30)^2 + (15-17,5)^2 + (20-17,5)^2 = 20,5$

46 = $(28-31,5)^2 + (35-31,5)^2 + (15-16,5)^2 + (18-16,5)^2 = 29$

56 = $(32-33,5)^2 + (35-33,5)^2 + (20-19)^2 + (18-19)^2 = 6,5$

2 and 3 lowest within group sum of squares and are grouped

23 1 = $(24-20,5)^2 + (17-20,5)^2 + (13-11,5)^2 + (10-11,5)^2 = 29$

23 4 = $(24-26)^2 + (28-26)^2 + (13-14)^2 + (15-14)^2 = 10$

23 5 =

23 6 = Further away than 4 from 23

23 + 14 = $4 + (28-22,5)^2 + (17-22,5)^2 + (15-12,5)^2 + (10-12,5)^2 = 73$

23 + 15 = Further away than 4 from 1

23 + 16 =

23 + 45 = 4 + 20,5 = 24,5

23 + 46 = 4 + 29 = 33

23 + 56 = 4 + 6,5 = 10,5

23 and 4 lowest within sum of squares and are grouped

C1 = 1
C2 = 2, 3, 4
C3 = 5
C4 = 6

The two methods got different results the second time they clustered, but for Average linkage, 2,3 and 4 was the second closest distances, which is what group Wards' method clustered.

And Wards method's second cluster had 23 and 56 as the second lowest within sum of squares, which is the what group Average linkage clustered.

The two methods might have gotten the same answer if we kept using Ward's method till there only was 2 groups.

Looking at the plot both methods are reasonable.

SU, DEPARTMENT OF STATISTICS

Room: VA   Anonymous code: MM-0004   Sheet number: 6

6

$x_1 = \lambda_1 \xi + \delta_1$, $x_2 = \lambda_2 \xi + \delta_2$, $x_3 = \lambda_3 \xi + \delta_3$.

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

λ_1 λ_2 λ_3 $V(\delta_1)$ $V(\delta_2)$ $V(\delta_3)$.

$V(\delta_1) = V(x_1) - \lambda_1^2$ $\text{cov}(x_1, x_2) = \lambda_1 \lambda_2$

$V(\delta_2) = V(x_2) - \lambda_2^2$ $\text{cov}(x_1, x_3) = \lambda_1 \lambda_3$

$V(\delta_3) = V(x_3) - \lambda_3^2$ $\text{cov}(x_2, x_3) = \lambda_2 \lambda_3$

6 unknown and 6 equations. $\Rightarrow$ parameter estimates unique

S_1

$\lambda_1 = \frac{0,93}{\lambda_2} = \frac{0,45}{\lambda_3} \Rightarrow 0,93\lambda_3 = 0,45\lambda_2$; $2,07\lambda_3 = \lambda_2$

$\lambda_3 = \frac{0,27}{\lambda_2} = \frac{0,27}{2,07\lambda_3}$ $2,07\lambda_3^2 = 0,27$ $\lambda_3^2 \approx 0,13$; $\lambda_3 \approx 0,36$

$\lambda_1 = \frac{0,45}{0,36} = 1,25$ $\lambda_2 = \frac{0,93}{1,25} \approx 0,74$ ok

$V(\delta_1) = 1,2 - 1,25^2 \approx -0,36$

$V(\delta_2) = 1,56 - 0,74^2 \approx 1,01$ ok

$V(\delta_3) = 2,15 - 0,36^2 \approx 2,02$

S_3

$\lambda_1 = \frac{-0,93}{\lambda_2} = \frac{-0,45}{\lambda_3} \Rightarrow -0,93\lambda_3 = -0,45\lambda_2$; $2,07\lambda_3 = \lambda_2$

$\lambda_3 = \frac{0,27}{\lambda_2} = \frac{0,27}{2,07\lambda_3}$ $2,07\lambda_3^2 = 0,27$ $\lambda_3 = 0,36$

$\lambda_1 = \frac{-0,45}{0,36} = -1,25$ $\lambda_2 = \frac{-0,93}{-1,25} = 0,74$

$V(\delta_1) = 1,2 - (-1,25^2) \approx -0,36$

$V(\delta_2) = 1,56 - 0,74^2 \approx 1,01$

$V(\delta_3) = 2,15 - 0,36^2 \approx 2,02$

S_2

$$\lambda_1 = \frac{-0{,}43}{\lambda_2} = \frac{-0{,}45}{\lambda_3} \Rightarrow 2{,}07\lambda_3 = \lambda_2$$

$$\lambda_3 = \frac{-0{,}27}{\lambda_2} = \frac{-0{,}27}{2{,}07\lambda_3} = \quad 2{,}07\lambda_3^2 = -0{,}27$$

$$\boxed{\lambda_3^2 = -0{,}13}$$

Since you can't square root a negative number. We have a problem here.

Looking at the matrix shows that something is off.

There is not posible for student 2 to have gotten negative on everything except the diagonal.

$$S = \begin{bmatrix} \lambda_1^2 & \lambda_1\lambda_2 & \lambda_1\lambda_3 \\ \lambda_2\lambda_1 & \lambda_2^2 & \lambda_2\lambda_3 \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3^2 \end{bmatrix}$$

It is not possible to get $\lambda_2 \cdot \lambda_1$, $\lambda_3 \cdot \lambda_1$ and $\lambda_2 \cdot \lambda_3$ all negative. If one λ is negative the one where it is not included will be positive. If two λ are negative the one where they are multiplied will be positive.

If all 3 λ are the same sign all will be positive.

So in order to get the matrix student 2 got there has to be a different model with more factors that might be correlated and cause the matrix to be possible. but it is not reasonable to get that matrix from a single factor model.

Model misspecification!!

Student 1 and 3 get the same estimates with the exception of student 3 getting λ_1 negative, but $|\lambda_1|$ is same

Which makes sence since the matrices are the same absolute value

and for student 3 $\lambda_1\lambda_2$ and $\lambda_1\lambda_3$ are negative but not $\lambda_2\lambda_3$, indicating λ_1 should be negative.

+

SU, DEPARTMENT OF STATISTICS

Room: VÄ **Anonymous code:** MM-0004 **Sheet number:** 7

7

a) specific variances.

$V(v) = V(X_i) - \lambda_1^2 - \lambda_2^2 - 2\lambda_1\lambda_2\phi$

high specific variance indicates that a lot of the variance is not explained by the factors, which is not good. you want the factors to explain as much variance as possible

If I assume $V(X) =$ and $\phi_{12} = 0$

$V(v_1) = 1 - 0{,}9^2 - 0{,}2^2 = 0{,}15$

$V(v_2) = 1 - 0{,}7^2 - 0{,}15^2 = 0{,}4875$

$V(v_3) = 1 - 0{,}2^2 - 0{,}9^2 = 0{,}15$

$V(v_4) = 1 - 0{,}7^2 - 0{,}7^2 = 0{,}47$

ok

b) Communalities $= \lambda_1^2 + \lambda_2^2$

shared variance $= (\lambda_1 + \lambda_2\phi)^2$ $\phi_{12} = 0$

	Communalities	shared variance F_1	F_2	% of shared variance F_1	F_2
X_1	$0{,}9^2 + 0{,}2^2 = 0{,}85$	$0{,}9^2 = 0{,}81$	$0{,}2^2 = 0{,}04$	95,29%	4,71%
X_2	0,5125	0,49	0,0225	95,61%	4,39%
X_3	0,85	0,04	0,81	4,71%	95,29%
X_4	0,53	0,04	0,49	7,55%	92,45%

ok

Communalities is the total variance explained by the factors and % of shared variance is how much % of the variance of the factor ... explained by each factor

c) Proportion of variance explained

F_1 $0{,}81 + 0{,}49 + 0{,}04 + 0{,}04 = 1{,}38$

F_2 $0{,}04 + 0{,}0225 + 0{,}81 + 0{,}49 = 1{,}3625$

Proportion of $F_1 = \frac{1{,}38}{1{,}38 + 1{,}3625} = 0{,}5032 = 50{,}32\%$

$F_2 = \frac{1{,}3625}{1{,}38 + 1{,}3625} = 0{,}4968 = 49{,}68\%$

not a total

±

they approximately explain the same amount of variance

F_1 a little more tho.

d) Estimated correlation matrix

	X_1	X_2	X_3	X_4
X_1	0,85			
X_2	0,66	0,5125		
X_3	0,36	0,275	0,85	
X_4	0,32	0,245	0,67	0,53

If we assume the novice student messed up all correlations for X_1 and they all should have been positive.

The estimate looks rather good, it is hard to say by just looking at it the so we will calculate Residual matrix an RSMR assuming the signs in the origin matrix all should be positive.

e) Residual

$$\begin{bmatrix} 1 & & & \\ 0,7 & 1 & & \\ 0,3 & 0,25 & 1 & \\ 0,35 & 0,2 & 0,4 & 1 \end{bmatrix} - \begin{bmatrix} 0,85 & & & \\ 0,66 & 0,5125 & & \\ 0,36 & 0,275 & 0,85 & \\ 0,32 & 0,245 & 0,67 & 0,53 \end{bmatrix} = \begin{bmatrix} 0,15 & & & \\ 0,04 & 0,4875 & & \\ -0,06 & -0,025 & 0,15 & \\ 0,03 & -0,045 & -0,27 & 0,4 \end{bmatrix}$$

$$RSMR = \sqrt{\frac{0,04^2 + (-0,06)^2 + 0,03^2 + (-0,025)^2 + (-0,045)^2 + (-0,27)^2}{6}}$$

$\approx 0,116655$.

the RSMR looks rather low but it is hard to say anything without the RSMR from an other estimated correlation matrix.

the lower RSMR the better