Stockholms
universitet

STATISTISKA INSTITUTIONEN

HT-2016

Written exam in Multivariate Methods, 7.5 ECTS credits
Thursday, 1st December 2016, 16:00 – 21:00
Time allowed: FIVE hours
Examination Hall: Värtasalen

You are required to answer all **6 (six)** questions as well as motivate your solutions. The total amount of points is 80. In order to pass this part, you need to get at least 40 points. Points from this exam will be added to your results from the computer lab assignment. The final grades are assigned as follows: **A** (91+), **B** (81-90), **C** (71-80), **D** (61-70), **E** (51-60), **Fx** (30-49), and **F** (0-29).

You are <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. (15 points) Let us analyse the following 3-variate dataset with 5 observations. Each observation consists of 3 measurements and recorded in the following matrix

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

What portion of total variance each variable accounts for? Compute the correlation matrix. Next, find eigenvalues of the correlation matrix and interpret them in style of PCA. How much of 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 300° 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 $(2,2)$. 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 $(2.8284,\ 0)$, find θ. Provide geometric motivation.

3. (12 points)
   a) What problems cluster analysis is intended to solve? What types of cluster analysis you know: name them. What assumptions on data are imposed in order to apply "cluster analysis"? How strict you should be with those assumptions: speculate and exemplify.
   b) 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	30	20
S6	35	18

4. (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	0.6	1.000		
X_3	0.3	0.25	1.000	
X_4	0.35	0.3	0.5	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	0.20	0.80
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.

5. (14 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? (2p)
   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? (2p)
   c) Compute the total, between-group, and within-group SSCP matrices. What conclusions can you draw from these matrices? Are there all assumptions fulfilled to apply discriminant

analysis? (out of 10 points)

Financial Data for Failed and non-Failed firms

Observations (Failed Firms)	EBITASS	ROTC	Observations (non-failed)	EBITASS	ROTC
1	0.1	0.3	1	-0.01	-0.03
2	0.2	0.2	2	-0.05	-0.11
3	0.2	0.3	3	0.09	0.12
4	0.1	0.2	4	0.03	0.05
5	0.3	0.2	5	0.04	0.06
6	0.2	0.1			

6. (12 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} = - 0.6. Repeat all your calculations in assumption that correlation changed to Corr(F_1, F_2)= Φ_{12} = 0.6 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?

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; $|\boldsymbol{\Sigma} - \boldsymbol{\lambda I}| = \boldsymbol{0}$; γ-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: $x_1 = \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \varepsilon_1$

$$\begin{aligned} x_2 &= \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \varepsilon_2 \\ &\vdots \\ x_p &= \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \varepsilon_p \end{aligned}$$

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

$|W^{-1}B - \lambda I| = 0$; $\gamma = \Sigma^{-1}(\mu_1 - \mu_2)$ - Fisher's discriminant function

Logistic regression

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

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

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

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:

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

Added formulas:

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

Cubic equations:

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

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

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