STOCKHOLM UNIVERSITY
Department of Statistics
Johan Koskinen

RESIT EXAM IN MULTIVARIATE METHODS
24 March 2023

Time: 5 hours

Allowed aids: Pocket calculator, language dictionary

The exam consists of five questions. To score maximum points on a question solutions need to be clear, detailed and well motivated.

Results will be announced no later than April 14
GOOD LUCK!

Question 1. (16 Points)

(a) You have data on peoples' preference for ice-cream, with values 'chocolate', 'vanilla', and 'strawberry' (everyone likes ice-cream so you need to provide a preference - and NO mixing of flavours). Can you provide a meaningful distance measure and the accompanying coding of the variable (-s)?

(b) What is the difference between Nominal scale and Ordinal scale?

(c) Briefly describe the difference between statistical inference and prediction (it might be helpful to explain this in terms of an example or a task, such as classification).

Question 2. (20 Points)
After some particularly severe bushfires in Australia, respondents in an affected community provided answers to questions on how they had been affected, took some psychological tests on their wellbeing as well as 'attachment style'. The variables measure are described in Table 1.

The eigenvalues for the correlation matrix $\mathbf{R}$ of these variables are

$$\boldsymbol{\Lambda} = \begin{bmatrix} 2.315 \\ 1.124 \\ 0.698 \\ 0.476 \\ 0.386 \end{bmatrix}.$$

Table 1: Summary of variables

Variable	Description
X_1	The extent to which they had suffered loss of property, on a scale from none (0) to everything (10)
X_2	Their rating on a post-traumatic stress disorder (PTSD) symptom scale, from none (0) to all symptoms (4)
X_3	Their depression state on a scale from not depressed (0) to depressed (8)
X_4	Whether their 'attachment style' was anxious, on a scale from not at all (0) to very (6)
X_5	Whether their 'attachment style' was avoidant, on a scale from not at all (0) to very (6)

and the matrix with eigenvectors as columns is

$$\mathbf{W} = \begin{bmatrix} -0.180 & 0.779 & 0.575 & -0.025 & 0.171 \\ -0.508 & 0.291 & -0.353 & 0.398 & -0.612 \\ -0.515 & 0.122 & -0.521 & -0.264 & 0.615 \\ -0.495 & -0.305 & 0.334 & -0.654 & -0.351 \\ -0.446 & -0.448 & 0.402 & 0.587 & 0.307 \end{bmatrix}.$$

(a) For an observation i, write the first principal component ξ_{i1} as a linear combination of the variable values x_{i1}, x_{i2}, x_{i3}, and x_{i4} in terms of the eigenvalues and eigenvectors.

(b) Draw the scree plot and comment and interpret it.

(c) How many principal components would you need in order to account for 80% of the variance?

(d) Compute the loadings of the variables on the first two PCs, i.e. compute the correlation of the variables with the PCs.

(e) Try to come up with some interpretation of the first PCs (Since you are unlikely to know what the variables actually stand for you may have to be creative).

Question 3. (20 Points)
For the variables in Question 2, an *orthogonal* 2-factor model is fitted to $\mathbf{R}$ which yields the following pattern loadings

$$\mathbf{\Lambda} = \begin{bmatrix} -0.038 & 0.361 \\ 0.358 & 0.791 \\ 0.445 & 0.525 \\ 0.718 & 0.162 \\ 0.734 & 0.040 \end{bmatrix}.$$

The usual assumptions are made and the rotation is varimax.

(a) What is the correlation between the two (latent) factors?

(b) What are the usual assumptions?

(c) Provide an estimate of the residual ψ_3^2 for X_3.

(d) How much of the variance of X_3 is explained by F_1, the first factor?

(e) A researcher wants to test a model where $\lambda_{11} = \lambda_{21} = 0$ and $\lambda_{42} = \lambda_{52} = 0$, and fits a model using CFA. State formally the χ^2-test of fit for testing the constraints of the researcher.

Question 4. (12 Points)

For the same dataset as in Questions 2 and 3, a variable Y is available for each respondent, which is 1 if the respondent 'feared for their life', and 0 otherwise. Using the variables X_1 and X_2, a researcher wants to use LDA to classify people based on Y.

The sample means for $Y = 0$ ($n_0 = 61$) and $Y = 1$ ($n_1 = 67$) are

$$\bar{\mathbf{x}}_0 = \begin{bmatrix} 5.344 \\ 0.770 \end{bmatrix}, \text{ and } \bar{\mathbf{x}}_1 = \begin{bmatrix} 6.209 \\ 1.179 \end{bmatrix},$$

respectively, and the sum of squares matrices are

$$\mathbf{SSCP}_0 = \begin{bmatrix} 897.77 & 78.82 \\ 78.82 & 78.79 \end{bmatrix}, \mathbf{SSCP}_1 = \begin{bmatrix} 745.07 & 72.49 \\ 72.49 & 119.85 \end{bmatrix}.$$

(a) Calculate the pooled covariance matrix $\mathbf{S}_{\text{pool}}$.

(b) Calculate Fisher's linear discriminant function for these data (assume equal weights and costs).

(c) How would you classify a new observation with $x_{i1} = 6$ and $x_{i2} = 1$.

Question 5. (12 Points)

For the same bushfire dataset as above, logistic regression of Y on X_1, X_2, X_3, X_4, and X_5, yields the parameter estimates

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} -0.681 \\ 0.052 \\ 0.003 \\ 0.202 \\ 0.024 \\ 0.084 \end{bmatrix}, \text{ and standard errors } s.e.(\boldsymbol{\beta}) = \begin{bmatrix} 0.452 \\ 0.053 \\ 0.189 \\ 0.116 \\ 0.181 \\ 0.163 \end{bmatrix},$$

(a) Can you tell from the estimates and standard errors which variables do the best job of discriminating between $Y = 0$ and $Y = 1$?

(b) Test if the coefficient for PTSD is statistically significantly different from zero.

(c) For a new observation with values

$$\mathbf{x} = \begin{bmatrix} 6.2 & 1.2 & 1.9 & 1.7 & 2.1 \end{bmatrix}$$

what is the predicted probability that the observation has $Y = 1$?

(d) For a new observation with values identical to $\mathbf{x}$ in (c) for variables $j = 2, \ldots, 5$, for what value of X_1 is $\Pr(Y = 1)$ equal to 0.5?