STOCKHOLM UNIVERSITY
Department of Statistics
Johan Koskinen

EXAM IN MULTIVARIATE METHODS
27 September 2022

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 October 11

Question 1. (16 Points)
Define and describe the following:

(a) Multivariate data

(b) Euclidean distance

(c) False positive rate

(d) Orthogonal vectors

Question 2. (16 Points)
The following correlation matrix is given

$$\mathbf{R} = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}.$$

(a) Compute the eigenvalues and unnormalised eigenvectors of the correlation matrix (start with the eigenvalues).

(b) Provide the normalised eigenvectors. How do they differ from the unnormalised eigenvectors and which ones do you use in principal components analysis? Why?

(c) What proportion of the variance is accounted for by the first principal component?

(d) Calculate the principal components scores for the observation $\mathbf{x} = (\frac{1}{2}, \frac{1}{4})$?

Question 3. (16 Points)
For $p = 4$ variables and $M = 3$ factors, the following model is assumed:

$$\mathbf{X} = \mathbf{\Lambda}\mathbf{F} + \boldsymbol{\epsilon}.$$

Furthermore, the following assumptions are made $E(\mathbf{F}) = \mathbf{0}$, $E(\mathbf{F}\mathbf{F}^\top) = \mathbf{I}$, $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $E(\boldsymbol{\epsilon}\mathbf{F}^\top) = \mathbf{0}$, and $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) = \mathbf{I}$.

(a) What is $E(\mathbf{X})$?

(b) Draw the graph of the model (be sure to include all variables and sources of variation).

(c) With the observed correlation matrix decomposed as, $\mathbf{R} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{I}$, and with given

$$\mathbf{\Lambda} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & 0 \\ \frac{1}{6} & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} \end{bmatrix},$$

what is the residual variance for X_1?

(d) What are the factor loadings for observation 2 if the solution is rotated by

$$\mathbf{Q} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Question 4. (16 Points)
For n households (Netflix accounts), Netflix has denoted by x_{ij} if household i has watched more than 10 minutes of film/series j, $x_{ij} = 1$, or not $x_{ij} = 0$, for $j = 1, \ldots, m$. The following data is provided for a subset of $n = 5$ households and $m = 6$ shows

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

(a) Cluster the *households* using nearest neighbour hierarchal clustering

(b) Draw the dendrogram

(c) How many clusters would you have decided upon - provide a reasoned motivation

(d) What, if any, would be the reason for choosing nearest neighbour rather than centroid in this example?

Question 5. (16 Points)
A cross-sectional, medical dataset from Iran is provided. For a subset of $n = 10$ individuals, data are provided for a number of risk factors of coronary heart decease. Individuals are furthermore classified by whether they have been diagnosed with coronary artery disease or not. Data are provided in Table 1 and plotted in Figure 1

For the data we calculate

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 27.8 \\ 124.2 \end{bmatrix}, \ \bar{\mathbf{x}}_2 = \begin{bmatrix} 29.6 \\ 130.2 \end{bmatrix}, \ \mathbf{SSCP}_1 = \begin{bmatrix} 12.8 & -0.8 \\ -0.8 & 16.8 \end{bmatrix}, \ \mathbf{SSCP}_2 = \begin{bmatrix} 11.2 & 21.4 \\ 21.4 & 66.8 \end{bmatrix}.$$

No diagnosis		Diagnosis	
x_1: BP	x_2: BMI	x_1: BP	x_2: BMI
27	126	29	125
27	124	27	127
27	126	30	132
27	121	31	135
31	124	31	132

Table 1: Systolic blood pressure and body mass index for individuals with diagnosed coronary artery disease and with no diagnosed coronary artery disease

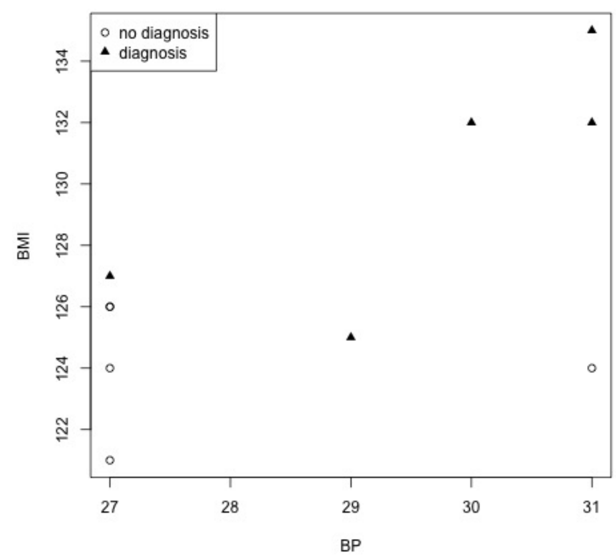

Figure 1: Systolic blood pressure and body mass index for individuals with diagnosed coronary artery disease and with no diagnosed coronary artery disease

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

(b) Calculate Fisher's linear discriminant function for these data.

(c) In addition to these two variables, gender (male: 0; female: 1), and smoking (yes: 1; no: 0), are included for all individuals. The (estimated) conditional probability that you will be diagnosed with coronary artery disease, given that you are a male smoker is much higher than for a non-smoking female. Could you use this information in your discriminant function? If not, how would you use it?