STOCKHOLM UNIVERSITY
Department of Statistics
Johan Koskinen

EXAM IN MULTIVARIATE METHODS
13 February 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 March 6
GOOD LUCK!

Question 1. (16 Points)

(a) Is cluster analysis a dependence or an independence method?

(b) Can we calculate the arithmetic mean for a variable on the nominal scale?

(c) We have three variables, two of them, say X_1 and X_2, are perfectly correlated, and the third variable, say X_3, is uncorrelated with the two others. What is the determinant of the correlation matrix? (*Hint: Do you really need to do the calculation?*)

(d) Let $\mathbf{R}$ be the correlation matrix in (c), provide an example of a vector $\mathbf{b}$ for which we cannot find values $\mathbf{x}$ in $\mathbf{Rx} = \mathbf{b}$.

Question 2. (20 Points)
In the US General Social Survey 2021 (GSS21), respondents were asked

> On a scale of 0 to 10, how much do you personally trust each of the following institutions? 0 means you do not trust an institution at all, and 10 means you trust it completely

Variables X_1, X_2, and X_3 record the answers for three institutions. Respondents were also asked

> On a scale from 0 to 10, how bad or good do you think the impacts of climate change will be for the world as a whole? 0 means extremely bad, 10 means extremely good.

Their responses were recorded as the variable X_4

Table 1 summarises the data.

Table 1: Summary of variables

Variable Name	Description	Sample mean	Sample sd
X_1	University research centers (TRRESRCH)	6.33	2.36
X_2	The news media (TRMEDIA)	3.71	2.71
X_3	The U.S. Congress (TRLEGIS)	3.31	2.41
X_4	Impact on climate (CLMTWRLD)	2.87	2.36

The (rounded) correlation matrix for $n = 1675$ observations on the variables is

$$\mathbf{R} = \begin{bmatrix} 1 & 0.51 & 0.33 & -0.31 \\ 0.51 & 1 & 0.53 & -0.29 \\ 0.33 & 0.53 & 1 & -0.05 \\ -0.31 & -0.29 & -0.05 & 1 \end{bmatrix}.$$

The correlation matrix $\mathbf{R}$ has (rounded) eigenvalues $\lambda_1 = 2.06$, $\lambda_2 = 0.97$, $\lambda_3 = 0.57$, and $\lambda_4 = 0.40$.

The first *three* elements of the (normalised) eigenvector for the first eigenvalue of $\mathbf{R}$ are -0.54 -0.60, and -0.48.

In the following, make calculations based on the assumption that we standardise data. As all reported quantities have been rounded, the eigenvalues and vectors may not be precise.

(a) You are going to treat the variables as metric, but what is the true measurement scale?

(b) What is the (sample) variance of the first principal component?

(c) Compute the normalised eigenvector for the largest eigenvalue

(d) What proportion of variance does the first principal component account for?

(e) What is the sample correlation between the first principal component and the second principal component?

(f) What is the sample correlation between the first principal component and the first and second variables, respectively (i.e. the loadings)? Provide a brief interpretation.

(g) What is the sample variance-covariance matrix for the principal components?

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

$$\mathbf{\Lambda} = \begin{bmatrix} 0.628 & -0.165 \\ 0.823 & 0.034 \\ 0.627 & 0.394 \\ -0.371 & 0.462 \end{bmatrix}.$$

A second 3-factor model is estimated which yields the pattern loadings

$$\mathbf{\Lambda} = \begin{bmatrix} 0.723 & -0.007 & 0.420 \\ 0.719 & 0.145 & -0.032 \\ 0.624 & 0.471 & -0.282 \\ -0.540 & 0.727 & 0.193 \end{bmatrix}.$$

The usual assumptions are made and in addition $\phi_{jk} = 0$, for all $j \neq k$ in both models.

(a) Based on the two models, what are the residuals for X_2, i.e. provide the estimates of ϕ_2^2.

(b) For the first model, what is the correlation between the first factor and the variables X_1 and X_2, respectively?

(c) For the 3-factor model, $\mathbf{F}$ is rotated using

$$\mathbf{Q} = \begin{bmatrix} 0.648 & 0.556 & -0.520 \\ -0.437 & 0.831 & 0.343 \\ 0.623 & 0.005 & 0.782 \end{bmatrix},$$

i.e. we construct new factors $\mathbf{QF}$. Will this rotation change $cor(X_1, X_4)$? What is $E(X_4, F_2)$ for the model with the rotated solution?

(d) Is the rotation in (c) orthogonal or oblique? (*You only need to check unit length of one vector and orthogonality for one pair of vectors. Take into account that there may be rounding errors so round your final numbers to 2 or three decimals*)

(e) What model do you prefer? Provide a brief motivation based on two different reasons. (*You may draw on all available information in Questions 2 and 3*)

Question 4. (12 Points) We will now do cluster analysis on the (standardised) data of Question 2. It one stage, clustering the $n = 1675$ observations using euclidian distance and the *centroid method*, we arrive at three clusters with the following centroids

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 0.000 \\ -0.002 \\ 0.000 \\ -0.003 \end{bmatrix}, \bar{\mathbf{x}}_2 = \begin{bmatrix} 1.559 \\ 1.213 \\ -1.373 \\ 3.016 \end{bmatrix}, \bar{\mathbf{x}}_3 = \begin{bmatrix} -2.262 \\ 2.320 \\ 1.118 \\ 1.748 \end{bmatrix},$$

Using a different hierarchical clustering, with the *nearest neighbour* (single linkage) method, when three clusters are obtained, the cluster sizes are 1673, 1, and 1, respectively.

(a) Based on these three centroids, which two clusters are next to merge based on the centroid method?

(b) For the hierarchical clustering with the nearest neighbour (single linkage) method, how many pairwise distances will you need to compute to determine what clusters, out of the 3, will be merged?

(c) With k-means, for $k = 3$, the centroids are

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 0.009 \\ 0.361 \\ 0.850 \\ 0.879 \end{bmatrix}, \bar{\mathbf{x}}_2 = \begin{bmatrix} -0.735 \\ -0.991 \\ -0.859 \\ 0.416 \end{bmatrix}, \bar{\mathbf{x}}_3 = \begin{bmatrix} 0.601 \\ 0.626 \\ 0.260 \\ -0.805 \end{bmatrix},$$

and for $k = 2$

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 0.516 \\ 0.666 \\ 0.503 \\ -0.373 \end{bmatrix}, \bar{\mathbf{x}}_2 = \begin{bmatrix} -0.657 \\ -0.848 \\ -0.640 \\ 0.475 \end{bmatrix},$$

It can be shown that for observations $\mathbf{x}_i$ and $\mathbf{x}_j$, being two rows in the data matrix, and their principal components $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$, for the (squared) distances

$$(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top = (\boldsymbol{\xi}_i - \boldsymbol{\xi}_j)(\boldsymbol{\xi}_i - \boldsymbol{\xi}_j)^\top$$

(note that these are row vectors, which is why the inner product is written this way). In light of this, would you choose the $k = 2$ or $k = 3$ solution and how would you characterise the clusters based on the model in Questions 2 and 3? Keep your answer to at most half a page.

Question 5. (12 Points)
The *unstandardised* variable X_4 from Q2 is binarised into

$$Y_i = \begin{cases} 1, & \text{if } X_4 > \bar{X}_4 \text{ for } i \\ 0, & \text{if } X_4 \leq \bar{X}_4 \text{ for } i \end{cases}.$$

Logistic regression of Y on X_1, X_2, and X_3, all not standardised, yields the parameter estimates

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 0.157 \\ -0.591 \\ -0.552 \\ 0.307 \end{bmatrix},$$

Pooled within-group sum of squares and cross products is

$$\mathbf{SSCP}_{within} = \begin{bmatrix} 8299.351 & 4354.038 & 2871.664 \\ 4354.038 & 11190.132 & 5478.219 \\ 2871.664 & 5478.219 & 9647.824 \end{bmatrix}$$

with the inverse is

$$\mathbf{SSCP}^{-1}_{within} = \begin{bmatrix} 0.000153 & -0.000052 & -0.000016 \\ -0.000052 & 0.000141 & -0.000065 \\ -0.000016 & -0.000065 & 0.000145 \end{bmatrix}$$

Assume that we have two new observations. The first observation (A) has X_1, X_2, and X_3 according the centroid of the first cluster in the k-means solution for $k = 2$; and the second observation (B) has X_1, X_2, and X_3 according the centroid of the *second* cluster in the k-means solution for $k = 2$. Take care to make sure that you translate these values to the original, non-standardised variables

(a) How would you classify A and B using Fisher's linear discriminant analysis? Assume equal weights and costs.

(b) How would you classify A and B using logistic regression?