1 Student 32 points

Mult - 21 / Mult-21 omtenta A
(17/12 2015)

UPPSALA UNIVERSITY
Department of Mathematics
Silvelyn Zwanzig

MATHEMATICAL STATISTICS
Multivariate Methods 1MS003
Aug 15, 2022

Permitted aids: two pages with handwritten notes

Time: 5 hours. For a pass (mark 3) the requirement is at least 18 points. For the mark 4, 25-31 points are necessary. For an excellent test (mark 5) the requirement is at least 32 points. Every problem is worth 5 points. For the international ECTS the following main rules are valid: A: 36-40 points, B: 28-35 points, C: 23-27 points, D: 20-22 points, E: 18-19 points.

OBS: *Please explain and interpret your approach carefully. Don't try to write more than really needed, but what you write must be clear and well argued. Solutions without any explanation will not be accepted!!!*

1. Suppose

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim N_3\left(\begin{pmatrix} a \\ b \\ c \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix}\right), \ \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix}$$

Furthermore we know

$X \sim N_1(0,1)$, $Cov(X,Z) = \frac{1}{8}$, $X|(Y = y) \sim N_1(y-1, 0.5)$, $Y|(Z = z) \sim N_1(z-1, 0.25)$.

(a) Identify $\mu = (a,b,c)^T$ and Σ.

(b) Determine the joint distribution of (X, Z).

(c) Determine the distribution $Y|(2X + 4Z)$.

(d) Determine the best linear prediction of Y by X.

(e) Determine an arbitrary function of X which predict Y best.

2. Suppose we have a balanced data set of 30 "old" oaks without replications.

measured: wood density, height, tree girth, crown diameter, average weight of the acorn, tannin concentration in the acorn

location: North America, Europe, Asia

species: Hungarian oak, white oak, red oak

(a) Write down the two way MANOVA model equation without interaction and the identification conditions.

(b) Derive the decomposition of the sample covariance matrix.

(c) Formulate the testing problem, that the geographical location is significant.

(d) Formulate the model under the null hypothesis.

(e) Assume normal distribution. Derive the likelihood ratio statistic. Compare it with Wilk's Lamda.

(f) Give the definition of the p-value.

(g) Suppose the p-value=0.0001. What is your conclusion?

3. Suppose we have a data set of 100 "old" oaks.

measured: wood density, height, tree girth, crown diameter, average weight of the acorn, tannin concentration in the acorn

determined species: white oak, red oak

Additional there are data of 10 oaks of unknown species.

(a) Formulate different research questions, that can be answered by using classification and factor analysis.

(b) Formulate the related models for every method and decide how the measured or determined variables can be included.

(c) What latent factors could be recovered by factor analysis?

(d) Give the main R-command for every method.

4. Given the covariance matrix Σ of a random vector $(X_1, X_2, X_3, X_4, X_5)$.

$$\begin{pmatrix} 1 & 0.5 & 0.3 & 0 & 0 \\ 0.5 & 1 & 0.9 & 0 & 0 \\ 0.3 & 0.9 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.8 \\ 0 & 0 & 0 & 0.8 & 1 \end{pmatrix}$$

$$\text{eigenvectors:} \quad \begin{pmatrix} .18465 \\ -.73421 \\ .65333 \\ 0 \\ 0 \end{pmatrix} \leftrightarrow 7.3384 \times 10^{-2}, \quad \begin{pmatrix} .43659 \\ .65685 \\ .61476 \\ 0 \\ 0 \end{pmatrix} \leftrightarrow 2.1747,$$

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ .70711 \\ .70711 \end{pmatrix} \leftrightarrow 1.8, \quad \begin{pmatrix} .8805 \\ -.17172 \\ -.44184 \\ 0 \\ 0 \end{pmatrix} \leftrightarrow .75195, \quad \begin{pmatrix} 0 \\ 0 \\ 0 \\ -.70711 \\ .70711 \end{pmatrix} \leftrightarrow .2$$

(a) Sign (roughly) a scree plot.

(b) How many principal components, would you include in your study? Discuss the structure of the correlation matrix.

(c) Calculate the first three principal components.

(d) Calculate the covariance matrix of the principal components.

5. Assume we have two bivariate normal populations. The first is the standard normal distribution. The second is the bivariate normal distribution with expected values $\mu_1 = 4$ and $\mu_2 = -4$ and

$$\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

An optimal discriminant rule is searched, where the error of wrong classification to population 1 is twice of the other error. The prior distribution is uniform.

(a) Determine eigenvalues and eigenvectors of Σ.

(b) Plot the contours sets of both distributions in one picture.

(c) Formulate the TPM.

(d) Determine the best regions with respect to TPM. Sign the regions in the plot above.

6. Let us compare languages by the following table.

Eng	Ger	Se	Fin
house	Haus	hus	talo
mother	Mutter	mor	äiti
tree	Baum	träd	puu
father	Vater	far	isä
dog	Hund	hund	koira

(a) Which main conditions should a similarity relation fulfill?

(b) Define a similarity relation, for the comparison above.

(c) Calculate the similarity matrix.

(d) Illustrate the distances between these languages in a plane.

7. Given a similarity matrix

$$\begin{array}{ccccc} & A & B & C & D \\ A & 10 & & & \\ B & 8 & 10 & & \\ C & 2 & 1 & 10 & \\ D & 1 & 5 & 4 & 10 \end{array}$$

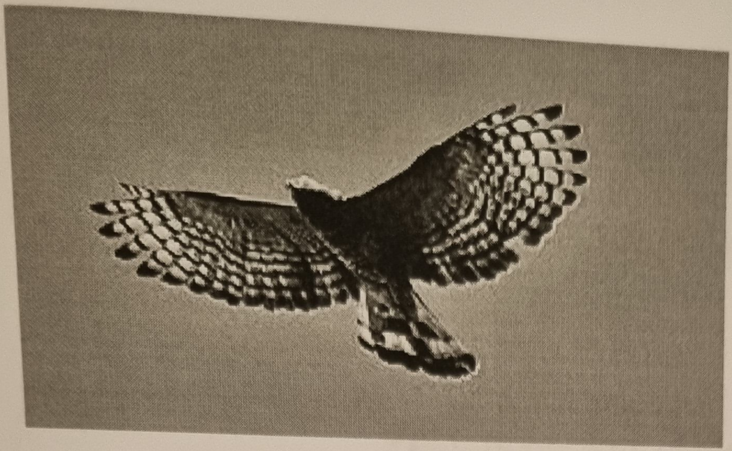
(a) Carry out a cluster analysis:

i. hierarchical with single linkage

ii. hierarchical with complete linkage

(b) Plot the dendrogram for both methods.

(c) Compare the results.



8. Lets us consider the data set in the book on female hooke-billed kites, compare the plot and the picture. The following R- code was used.

```
library(mvnormtest)
mshapiro.test(t(bird))
library(ICSNP)
HotellingsT2(bird,mu=c(200,300))
HotellingsT2(bird,mu=c(193,279))
```

The following R results are given.

```
Shapiro-Wilk normality test
W = 0.9806, p-value = 0.6441

        Hotelling's one sample T2-test
data:  bird
T.2 = 56.6479, df1 = 2, df2 = 43, p-value = 8.906e-13
alternative hypothesis: true location is not equal to c(200,300)

Hotelling's one sample T2-test
data:  bird
T.2 = 0.0762, df1 = 2, df2 = 43, p-value = 0.9268
alternative hypothesis: true location is not equal to c(193,279)
```

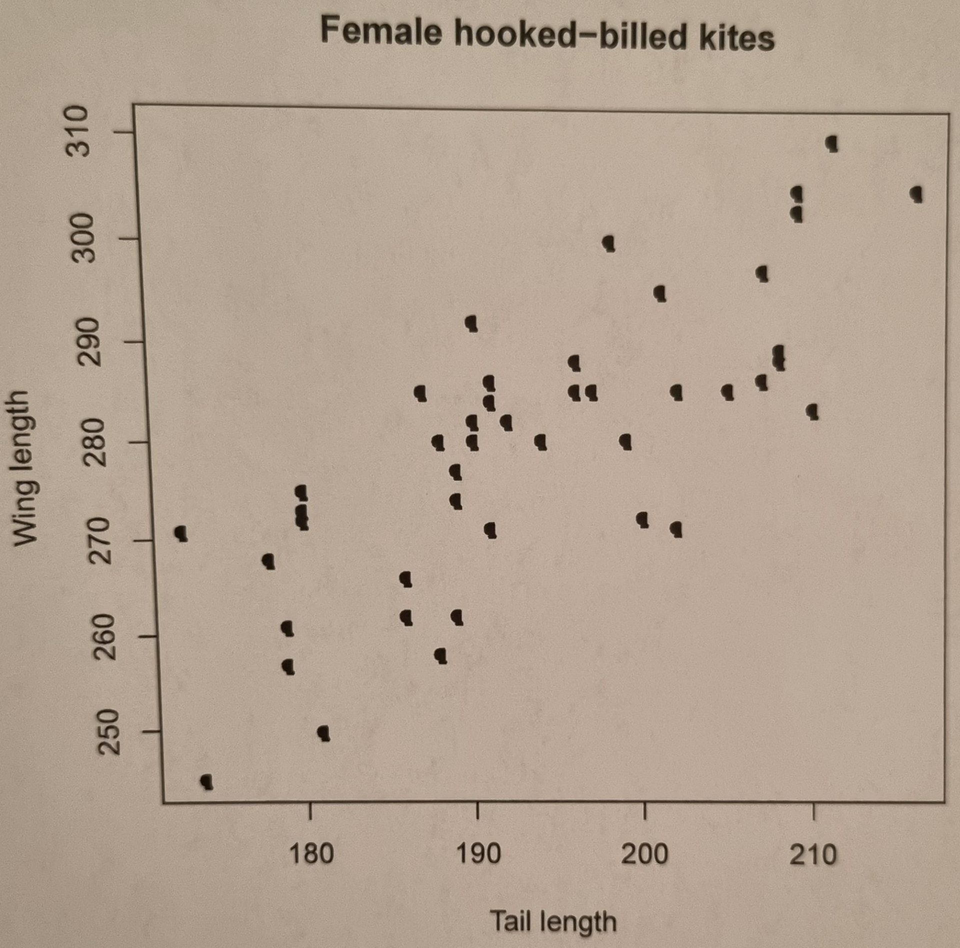
(a) Formulate the model assumptions for applying this Hotelling's test.

(b) Which testing problems are considered?

(c) How is the Hotelling's T test statistic defined? Give the definition of the p-value.

(d) Is the distribution assumption realistic?

(e) Interpret the result of the tests. Compare it with plot of the data.



Good Luck! Lycka till!! Viel Glück!!!