2021

Mult20

UPPSALA UNIVERSITY
Department of Mathematics
Silvelyn Zwanzig

Multivariate Methods
1MS003
Examination 2021-01-11

Permitted aids: Pocket calculator, handwritten sheet of formulae, dictionaries from any language to English and vice versa.

Time: 5 hours. For a pass = mark 3 the requirement is at least 18 points. For mark 5 (an excellent test) the requirement is at least 32 points and for mark 4 the requirement is at least 25 points. Every problem is worth 5p.

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.*

Instructions

- Your solutions and (!) your handwritten sheet of formularies have to be scanned and uploaded to studentportalen, latest at 13:30.
- In case of any problem contact me via email: zwanzig@math.uu.se
- You have to carry out the exam personally without any other additional aid.

1. Suppose

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \frac{1}{4} & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{4} & 0 & 2 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 2 \end{pmatrix} \right)$$

   (a) Determine the joint distribution of (X_1, X_2).
   (b) Determine the conditional distribution $(X_1, X_2)|(X_3, X_4)$.
   (c) Determine the best linear prediction of X_3 by X_2.
   (d) Determine an arbitrary function of X_3, X_4 which predict $3X_2 + 2$ best.

2. Suppose a balanced coffee data set with equal number of replications for each coffee sort.

   measured variables: opacity, viscosity, coffein concentration, bitter substances concentration
   sort: Lindvalls Mörkrost, Lindvalls Brygg, Lindvalls Kokkaffe, Lindvalls Brazil

   (a) Define a useful MANOVA model.

(b) Formulate the testing problem.

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

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

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

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

3. Suppose two sample:

$$X_1, \ldots, X_n \text{ i.i.d. } N_p(\mu_x, \Sigma)$$

and

$$Y_1, \ldots, Y_m \text{ i.i.d. } N_p(\mu_y, \Sigma)$$

The first sample includes measured variables: opacity, viscosity, coffein concentration, bitter substances concentration of n cups of Lindvalls Mörkrost. The second sample includes measured variables: opacity, viscosity, coffein concentration, bitter substances concentration of m cups of Jacobs Krönung. A German student states that there is no difference between the two coffee sorts.

(a) Formulate the testing problem.

(b) Propose an estimate S_{pool} of the covariance matrix Σ using the both samples. Which distribution has the estimator S_{pool}?

(c) Propose an estimate S_1 of the covariance matrix Σ using the first sample only. Which distribution has the estimator S_1?

(d) Let

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where Σ_{11} is a 2×2 matrix. Propose an estimate S_{11} of the covariance matrix Σ_{11} using the both samples. Which distribution has the estimator S_{11}?

(e) Formulate the testing problem, when you are only interested in the taste of the coffee. Which test statistic do you propose? Which distribution hast the test statistic under the null hypothesis?

4. Suppose the coffee data set of Problem 2. Now we consider only the sort "Lindvalls Mörkrost". We are interested in a study of the dependence between the physical and chemical quantities.

(a) Explain the set up of a CCA (Canonical correlation analysis) Define the canonical variates and the canonical correlation.

(b) Suppose the sample correlation of the data set above is

$$\begin{pmatrix} 1 & 0.4 & 0.5 & 0.6 \\ 0.4 & 1 & 0.3 & 0.4 \\ 0.5 & 0.3 & 1 & 0.2 \\ 0.6 & 0.4 & 0.2 & 1 \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Then the matrix $\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}$ is

$$\begin{pmatrix} 0.4371 & 0.2178 \\ 0.2178 & 0.1096 \end{pmatrix}.$$

(c) Calculate the first pair of canonical variates.

(d) Is it useful to include the second pair of canonical variates in the study?

(e) Give the main R-command this method.

5. Given the covariance matrix M of a random vector $(X_1, \dot{X}_2, X_3, X_4, X_5)$and the following R-result:

```
>M
     [,1] [,2] [,3] [,4] [,5]
[1,]  1.0  0.9  0.0  0.0    0
[2,]  0.9  1.0  0.0  0.0    0
[3,]  0.0  0.0  3.0  0.5    0
[4,]  0.0  0.0  0.5  3.0    0
[5,]  0.0  0.0  0.0  0.0    5
> eigen(M)
$values
[1] 5.0 3.5 2.5 1.9 0.1

$vectors
     [,1]      [,2]       [,3]      [,4]       [,5]
[1,]    0 0.0000000  0.0000000 0.7071068 -0.7071068
[2,]    0 0.0000000  0.0000000 0.7071068  0.7071068
[3,]    0 0.7071068  0.7071068 0.0000000  0.0000000
[4,]    0 0.7071068 -0.7071068 0.0000000  0.0000000
[5,]    1 0.0000000  0.0000000 0.0000000  0.0000000
```

(a) Give the scree plot.

(b) How many principal components, would you include in your study?

(c) Calculate the first two principal components.

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

6. Assume we have two bivariate normal populations. The first is the standard normal distribution. The second is the bivariate normal distribution with expected values $\mu_1 = 1$ and $\mu_2 = -1$ 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. Sign the regions in the plot above.

7. Let us compare languages by the following table.

Ger	Se	Fin	Est
Haus	hus	talo	maja
Mutter	mor	äiti	ema
Baum	träd	puu	puu
Vater	far	isä	isa
Hund	hund	koira	koer

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

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

(c) Calculate the similarity matrix.

(d) Carry out a cluster analysis, hierarchical with single linkage.

(e) Plot the dendrogram.

8. In order to compare the behavior of the inhabitants during two waves of Corvid-19. The following data set is studied: the relative reduction of passengers in public transport in week 14 and week 48 (T14, T48) the gasoline consumption per capita in week 14 and week 48 (G14, G48) taken in 11 different towns Göteborg, Linöping, Malmö, Jonk¨ping, Mora, Gävle, Umeå, Kiruna, Stockholm, Uppsala, Falun The R-code:

```
names(waves)<-c("T14","T48","G14","G48")
mshapiro.test(t(waves))
diff.waves<-data.frame(waves$T14-waves$T48,waves$G14-waves$G48)
names(diff.waves)<-c("T","G")
mshapiro.test(t(diff.waves))
M<-mvOutlier(diff.waves)
Mnew<-M$newData
```

```
mshapiro.test(t(Mnew))
HotellingsT2(diff.waves,mu=c(0,0))
HotellingsT2(Mnew,mu=c(0,0))
```

Results

```
 Shapiro-Wilk normality test
data: waves
W = 0.7487, p-value = 0.001997

Shapiro-Wilk normality test
data: diff.waves
W = 0.9663, p-value = 0.847

Shapiro-Wilk normality test
data: Mnew
W = 0.8539, p-value = 0.1333

Hotelling's one sample T2-test
data:  diff.waves
T.2 = 6.3718, df1 = 2, df2 = 9, p-value = 0.01888
alternative hypothesis: true location is not equal to c(0,0)

Hotelling's one sample T2-test
data:  Mnew
T.2 = 1.724, df1 = 2, df2 = 5, p-value = 0.2695
alternative hypothesis: true location is not equal to c(0,0)
```

(a) Which testing problem is considered?

(b) Formulate the model assumptions.

(c) What are the results of the 3 different Shapiro-Wilk tests?

(d) How is Mnew defined?

(e) Make it sense to apply Mnew instead of diff.waves?

(f) Compare the result of the Hotelling tests. Which result is more reliable?

Good Luck!