2022

Mult21

UPPSALA UNIVERSITY
Department of Mathematics
Silvelyn Zwanzig

MATHEMATICAL STATISTICS
Multivariate Methods 1MS003
2022-01-05

Permitted aids: pocket calculator, two pages with handwritten notes (one sheet)

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.*

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 replication number m for each combination of coffee and milk.

   measured variables: opacity, viscosity, caffeine concentration, bitter substances concentration
   coffee: Lindvalls Mörkrost, Lindvalls Brygg, Lindvalls Kokkaffe, Lindvalls Brazil
   milk: without, coffee cream, oat milk

   (a) Formulate the possible most general MANOVA model including all model assumptions.

   (b) Formulate the testing problem related to an interaction between coffee sort and use of milk.

   (c) Formulate the MANOVA model under the null hypothesis.

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

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

(f) Suppose the p-value=0.01. What is your conclusion?

3. Suppose the coffee data set of Problem 2. Now we consider the measured variables: opacity, viscosity, caffeine concentration, bitter substances concentration only for the sort "Lindvalls Mörkrost" and for two different types of milk use: coffee cream and oat milk. A vegetarian states that there is no difference in taste between the two milk sorts.

   (a) Formulate all model assumptions of this two sample problem.

   (b) Formulate the testing problem for an Hotelling T2- test. Give the definition of the test statistics.

   (c) How is the p-value defined? Which conclusion do you can derive from a p-value=0.67.

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

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

4. Suppose the coffee data set of Problem 2. We consider only the measurements of the sort "Lindvalls Mörkrost" without milk. We are interested in a study of the dependence between the physical and chemical quantities.

   Suppose the sample correlation of the data set above is

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

R-code

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

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

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

$$\begin{pmatrix} 0.2077 & 0.2644 \\ 0.2644 & 0.3389 \end{pmatrix}.$$

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

   (b) Calculate the first pair of canonical variates.

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

5. Given the covariance matrix $\Sigma = M$ of a random vector (X_1, X_2, X_3, X_4): Given the following R- result:

```
> M
     [,1] [,2] [,3] [,4]
[1,]  1.0  0.9  0.8  0.7
[2,]  0.9  1.0  0.9  0.8
[3,]  0.8  0.9  1.0  0.9
[4,]  0.7  0.8  0.9  1.0
eigen(M)
$values
[1] 3.50293864 0.34142136 0.09706136 0.05857864

$vectors
          [,1]       [,2]       [,3]       [,4]
[1,] 0.4850974  0.6532815  0.5144711 -0.2705981
[2,] 0.5144711  0.2705981 -0.4850974  0.6532815
[3,] 0.5144711 -0.2705981 -0.4850974 -0.6532815
[4,] 0.4850974 -0.6532815  0.5144711  0.2705981
```

(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 all principal components.

6. Assume two populations. First population is the two dimensional Cauchy distribution with density

$$f_1(x,y) = f(x,y;x_0,y_0,\gamma) = \frac{1}{2\pi}\frac{\gamma}{((x-x_0)^2+(y-y_0)^2+\gamma^2)^{\frac{3}{2}}}$$

the second is the two dimensional t distribution with ν degrees of freedom

$$f_2(x,y) = f(x,y;\nu) = \frac{1}{2\pi}\left(1+\frac{x^2+y^2}{\nu}\right)^{-\frac{(\nu+2)}{2}}.$$

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

(a) Formulate the ECM.

(b) Determine the best regions. Set $\nu = 1$.

(c) Set $\nu = \gamma = 1$. Suppose there are a trainings sample for each population. Estimate the optimal region. (Obs! Both are Cauchy distributions!)

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) Illustrate the distances between these languages in a plane by using multivariate scaling.

8. The production of high quality paper depends mainly on the quality of the fibers, measured by the variables AFF, FFF, ZST, LFF. The quality of the paper is determined by: BL (Breaking length), EL (Elastic modulus), SF (Stress at failure) and BL (Burst length). Given the following Rcode:

```
> paper.res<-cbind(paper$BL,paper$SF)
> M1<-lm(paper.res~paper$AFL+paper$LFF+paper$ZST)
> summary(M1)

lm(formula = Y1 ~ paper$AFL + paper$LFF + paper$ZST)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -55.92157   12.05091  -4.640 2.03e-05
paper$AFL    -4.15731    1.97136  -2.109  0.03929
paper$LFF     0.09690    0.03368   2.877  0.00561
paper$ZST    69.15191   11.60638   5.958 1.60e-07

Residual standard error: 1.575 on 58 degrees of freedom
Multiple R-squared:  0.7159,     Adjusted R-squared:  0.7012
F-statistic: 48.72 on 3 and 58 DF,  p-value: 7.418e-16

lm(formula = Y2 ~ paper$AFL + paper$LFF + paper$ZST)
Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.38538    5.32244  -6.836 5.52e-09
paper$AFL    -2.01678    0.87068  -2.316  0.02409
paper$LFF     0.04668    0.01487   3.138  0.00267
paper$ZST    37.64230    5.12610   7.343 7.77e-10

Residual standard error: 0.6957 on 58 degrees of freedom
Multiple R-squared:  0.785,     Adjusted R-squared:  0.7739
F-statistic: 70.58 on 3 and 58 DF,  p-value: < 2.2e-16
#############
> paper.res2<-cbind(paper$EM,paper$SF)
> M2<-lm(paper.res2~paper$LFF+paper$ZST-1)
> summary(M2)

lm(formula = Y1 ~ paper$LFF + paper$ZST - 1)
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
paper$LFF 0.018667   0.004514   4.135 0.000112
paper$ZST 6.133471   0.176495  34.751  < 2e-16

Residual standard error: 0.4945 on 60 degrees of freedom
Multiple R-squared:  0.9956,    Adjusted R-squared:  0.9954
F-statistic:  6728 on 2 and 60 DF,  p-value: < 2.2e-16

lm(formula = Y2 ~ paper$LFF + paper$ZST - 1)
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
paper$LFF 0.069056   0.008391   8.230 1.99e-11
paper$ZST 2.767216   0.328071   8.435 8.91e-12

Residual standard error: 0.9191 on 60 degrees of freedom
Multiple R-squared:  0.9759,    Adjusted R-squared:  0.9751
F-statistic:  1213 on 2 and 60 DF,  p-value: < 2.2e-16
```

(a) Formulate both fitted models, including all assumptions.

(b) Give the matrix of estimates for model M1.

(c) Which testing problems are considered for model M2?

(d) Which model do you would prefer?

(e) Propose simpler model, which is reasonable.

Good Luck! Lycka till!! Viel Glück!!!