MAM5220: Statistical Techniques for Computational Biology
MA35210: Topics in Biological Statistics

MANOVA workbook 2017

You will be assessed on your answers to questions 2 to 8.

The basic principles of MANOVA

Suppose we have a multivariate response variable, ***Y***, i.e. a vector-response, across a number of groups. As we saw in the MANOVA lecture, if we wish to test the null hypothesis that the population mean vectors are the same across the groups against the alternative hypothesis that at least two of the population mean vectors are different, we construct a one-way MANOVA table:

Source	d.f.	SSP matrix	Wilks' criterion
Between samples	g-1	$\boldsymbol{B} = \sum_{j=1}^{g} \boldsymbol{n_j}(\overline{\boldsymbol{x}}_j - \overline{\boldsymbol{x}})\,(\overline{\boldsymbol{x}}_j - \overline{\boldsymbol{x}})^T$	$\lvert\boldsymbol{W}\rvert / \lvert\boldsymbol{W}+\boldsymbol{B}\rvert$
Within samples	N-g	$\boldsymbol{W} = \boldsymbol{T} - \boldsymbol{B}$	
Total	N-1	$\boldsymbol{T} = \sum_{j=1}^{g}\sum_{i=1}^{n_j}(\boldsymbol{x_{ij}} - \overline{\boldsymbol{x}})(\boldsymbol{x_{ij}} - \overline{\boldsymbol{x}})^T$	

In the above table, the superscript T indicates the transpose of a column vector to make a row vector and |A| is the determinant of the matrix A. N is the total number of observations and g is the number of groups.

Question 1

This is a practice question to get you used to the `manova()` function in R. The question uses the anteater data set that was used in the lecture. The data are stored in the file `anteaters.csv` and consist of the logarithms of three skull measurements on each anteater, with six anteaters from Minas Graes, four from Matto Grosso and three from Santa Cruz

Read in the data

```
anteaters<-read.csv("anteaters.csv")
```

a) Carry out a one-way MANOVA using the following code

```
anteaters.man<-
manova(cbind(logx1,logx2,logx3)~location,data=ant
eaters)
```

Look at the coefficients, `anteaters.man$coef`, and interpret what they mean.

b) Check the residuals for the three variables across the three sites

```
par(mfrow=c(2,2))  # a 2 x 2 grid of plots

plot(as.numeric(anteaters$location),anteaters.man
$resid[,1],xlab="location
(code)",ylab="Residual",main="logx1 residuals")

plot(as.numeric(anteaters$location),anteaters.man
$resid[,2],xlab="location(code)",ylab="Residual",
main="logx2 residuals")

plot(as.numeric(anteaters$location),anteaters.man
$resid[,3],xlab="location
(code)",ylab="Residual",main="logx3 residuals")
```

c) Use `summary(anteaters.man)` to obtain a p-value to test the null hypothesis that the mean vector of log measurements does not depend on the site from which the anteaters come. The default is the Pillai trace. To obtain the other types (Wilks' lambda, Hotelling-Lawley trace, Roy's statistic), use e.g. `summary(anteaters.man,test="W")`. What are the conclusions of the four tests?

Question 2

This question makes use of the data in the data frame `Sales.scores.csv`. The data are gathered from a company who were interested in assessing the effect of training seminars on

knowledge and motivation amongst their sales staff. The first few rows of the data frame are given below

Salesperson ID	Knowledge	Motivation	Treatment group
1	8.599	10.364	T1
2	10.222	9.836	T1
3	9.134	11.190	T1
…	…	…	…

The treatment group entry is one of:

T1: Attended 5-day seminar
T2 (control): Did not attend seminar
T3: Attended 3-day seminar

a) Use one-way ANOVA to test the null hypothesis that Knowledge is the same across all three treatment groups (5 marks)

b) Repeat part a) for the Motivation variable (5 marks)

c) Use MANOVA to test the null hypothesis that the mean vector (Knowledge, Motivation) is the same across all three treatment groups. Report your findings and state your conclusion. (5 marks)

Question = (15 marks)

Hints

- For one-way ANOVA you need first to build a linear model – what is the factor in this case?
- For one-way MANOVA, you need to have a formula with a matrix on the left and the factor on the right (either side of the ~ sign)

Question 3

In this question we use the `iris` data set that is available as part of the base package (i.e. with no extra libraries installed) of `R`. The data set consists of sepal length, sepal width, petal length and petal width for 150 irises (50 of each of three species).

a) Carry out four one-way ANOVAs with Tukey HSD posthoc tests to test whether each of the four measurements (sepal length, sepal width, petal length and petal width) are different for irises of different species

(8 marks)

b) What is the problem with carrying out four independent one-way ANOVAs in terms of the probability of a type I error? Explain how you could use a Bonferroni correction to account for multiple testing

(4 marks)

c) Use MANOVA on the four variables simultaneously and clearly interpret the output. Explain the difference between MANOVA and repeated one-way ANOVAs in the case of multivariate data.

(8 marks)

Question = (20 marks)

Question 4

The following data table includes the measurements of three amino acids (alanine, aspartic acid and tyrosine) in the haemolymph (like blood) of male and female centipedes

Male centipedes			Female centipedes		
Alanine	Aspartic acid	Tyrosine	Alanine	Aspartic acid	Tyrosine
7.0	17.0	19.7	7.3	17.4	22.5
7.3	17.2	20.3	7.7	19.8	24.9
8.0	19.3	22.6	8.2	20.2	26.1
8.1	19.8	23.7	8.3	22.6	27.5
7.9	18.4	22.0	6.4	23.4	28.1

Enter the data into `R` such that each row of the 10×3 matrix corresponds to the three amino acid measurements for one of the centipedes. Let the first five rows correspond to the males and the last five rows to the females.

Define a variable called `gender` that gives the genders of each of the ten centipedes

```
gender<-rep(c("male","female"),rep(5,2))
```

a) Perform a one-way MANOVA with the matrix of amino acid measurements as the response and `gender` as the factor

```
aa.man<-manova(aadata~gender)
```

What does the MANOVA reveal in this case? (Use `summary()` to explore the MANOVA fit. Does it make any difference which of the four tests you use? The default is Pillai trace, but you can also use Wilks' Lambda (`test="W"`), Hotelling-Lawley (`test="H"`) or Roy's statistic (`test="R"`))

(6 marks)

b) Use `summary.aov(aa.man)` to perform individual one-way ANOVAs on the three amino acids. Where does this analysis suggest the differences lie between the male and female centipedes?

(4 marks)

(10 marks)

Question 5

In this question we consider the extension of MANOVA to a two-way design. The data are bivariate, with plasma calcium and water loss measured in male and female birds, with and without hormone treatment. The data are shown in the following table.

No hormone treatment			
Female		Male	
Plasma Ca	Water loss	Plasma Ca	Water loss
16.5	76	14.5	80
18.4	71	11.0	72
12.7	64	10.8	77
Hormone treatment			
Female		Male	
Plasma Ca	Water loss	Plasma Ca	Water loss
39.1	71	32.0	65
26.2	70	23.8	69
21.3	63	28.8	67

Enter the data using the following code

```
Plasma.Ca<-
c(16.5,18.4,12.7,14.5,11.0,10.8,39.1,26.2,21.3,
32.0,23.8,28.8)
Water.loss<-
c(76,71,64,80,72,77,71,70,63,65,69,67)
Sex<-
as.factor(rep(rep(c("Female","Male"),rep(3,2)),
2))
Hormone<-as.factor(rep(c("No
hormone","Hormone"),rep(6,2)))
```

a) What are the two factors of interest here?

(2 marks)

b) Make a matrix of response variables using the `cbind()` function and carry out a two-way MANOVA (with interactions) using the formula `Y~Sex*Hormone`, where `Y` is the name of the matrix of response variables. Use the `summary()` to report the p-values for the main effects and the interaction effect. What is your conclusion here? What are the fitted means for the different groups? Use `summary.aov()` to perform one-way ANOVAs on the variables separately. Report your findings.

(11 marks)

Question = (13 marks)

Question 6

This question involves one-way MANOVA for three (simulated) variables measured on agricultural plots. The measurements (labeled M1, M2, M3) and the treatment (Control, T1, T2, T3) are given in the file called `fields.csv`

Read in the data:

```
fields<-read.csv("fields.csv",row.names=1)
```

a) What is the model for a one-way MANOVA in this case?

(2 marks)

b) Fit the model using the `manova()` function in `R` and quote the p-values for fitting the model using Wilk's lambda, Pillai trace, Hotelling-Lawley statistic and Roy's statistic

(4 marks)

c) The following code enables the comparison of Control and T1 using subsetting of the data set. Modify the code and quote the p-values for all pairwise comparisons of treatment levels, including an explanation of the application of Bonferroni's method to account for multiple testing.

```
Cont.T1.man<-
manova(cbind(M1,M2,M3)~Treatment,data=fields,subs
et=Treatment %in% c("Control","T1"))
```

(9 marks)

d) What does the output of `summary.aov()` applied to the MANOVA object tell you in this case?

(3 marks)

Question = (18 marks)

Question 7

This question makes use of a data set consisting of four measurements on 150 male skulls from Egypt. The skulls are from 5 epochs (c4000BC, c3300BC, c1850BC, c200BC and cAD150). The measurements are maximum breadth (mb), basibregmatic height (bh), basialiveolar length (bl) and nasal height (nh). The question is whether the measurements change over time. Non-constant skull measurements over time would indicate interbreeding with immigrant populations.

Type in the following code to load the library "HSAUR" and access the help file for the `skulls` data set

```
library(HSAUR)
help(skulls)
```

Copy and paste the R code at the bottom of the help page for skulls to produce pairwise plots of the means for the four variables according to epoch. (If you have trouble accessing this package, there is a file called skulls.csv on the Blackboard page, and an R script called skulls.R that contains the example from the bottom of the help page for the skulls data set.)

Comment on the pairwise plots, in light of the question of whether the measurements change over time.

Type

```
means
```

to look at the matrix of means you have just calculated. Explain the differences that you can see in the table (at this stage we do not know whether those differences are significant).

Carry out an investigation of the skull data set, using a combination of MANOVA and the `summary.aov()` function. Include details of the pairwise multivariate tests comparing c4000BC with each progressively later time point (modify the code of question 7c).) You can address the multiple comparison issue by using a significance level of α=0.15 and carry out each test at the α/m level, where m is the number of tests carried out.

Question = (18 marks)

Question 8 (contrasts)

A linear combination of the population means,

$$\boldsymbol{\Psi} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{g}} \boldsymbol{c_i \mu_i},$$

where the sum of the coefficients is 0, i.e. $\sum_{i=1}^{g} c_i = 0$, is a ***contrast***. For example if we wanted to compare population means for the first two groups, we would set $c_1 = 1$ and $c_2 = -1$ and look at the contrast

$$\boldsymbol{\Psi} = \boldsymbol{\mu_1} - \boldsymbol{\mu_2}.$$

We estimate a contrast by estimating the population means with the relevant sample means.

$$\widehat{\boldsymbol{\Psi}} = \sum_{i=1}^{g} c_i \bar{Y}_i.$$

Then

$$Var(\widehat{\Psi}) = \left(\sum_{i=1}^{g} \frac{c_i^2}{n_i}\right)\Sigma,$$

where Σ is the covariance matrix of the original data. We estimate

$$Var(\widehat{\Psi}) = \left(\sum_{i=1}^{g} \frac{c_i^2}{n_i}\right)\frac{W}{N-g},$$

where W is the within-sample sum of squares and products matrix.

Hypothesis testing

Let

$$\boldsymbol{\Psi} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{g}} \boldsymbol{c_i \mu_i}$$

with $\sum_{i=1}^{g} c_i = 0$ be a contrast. Wish to test $H_0: \Psi = 0$ against $H_A: \Psi \neq 0$.

Let

$$B_\Psi = \frac{\widehat{\Psi}\widehat{\Psi}^T}{\sum_{i=1}^{g} c_i^2/n_i},$$

where the superscript T indicates the transpose of the column vector to make a row vector.

The Wilks' lambda statistic in this case is given by

$$\Lambda_\Psi^* = \frac{|\,W\,|}{|W + B_\Psi|},$$

where | A | is the determinant of the matrix A. The corresponding F approximation is given by

$$F = \left(\frac{1-\Lambda_\Psi^*}{\Lambda_\Psi^*}\right)\left(\frac{N-g-p+1}{p}\right),$$

where N is the total number of observations, g is the number of groups and p is the number of variables.

We reject H_0 in favour of H_A at level α if $F > F_{p,N-g-p+1,\alpha}$, i.e. the value for which a proportion α of the area under an F distribution on p and N-g-p+1 degrees of freedom lies to the right of that value.

Consider the data set in the file Q8.csv. There are four columns, representing three measurements (X1, X2 and X3), and a grouping variable which takes values 1, 2, 3 and 4. There are six observations per group.

a) Use MANOVA to verify that there are differences between the population means for the four groups.
b) Set up and test three contrasts corresponding to the following three comparisons:
   i) Group 1 against group 4
   ii) Group 2 against group 3
   iii) Groups 1 and 4 against groups 2 and 3.
   Present your findings clearly, including the sample means, the estimates of the contrasts, the denominators used in calculating the B_Ψ matrices, the Lambda values, the F-values (with the corresponding degrees of freedom for the relevant F distributions) and the p-values.

(In the file Q8.R you will find sample code for calculating the Between SSP, the Within SSP and the Total SSP for these data. You should be able to modify this code to carry out the hypothesis tests as described above. Note that %*% means matrix multiplication in R and t(x) is the transpose of x.)

(16 marks)

Total = 110 marks