



Practical Workbook

Dr. Lee Fawcett and Dr. Joe Matthews,
Newcastle University

Exploratory Multivariate Data Analysis

The R code used throughout this session can be found in plain text format (for copy-and-pasting) at:

<https://www.mas.ncl.ac.uk/~nlf8/cpact/Rcode.R>

Problem Session 1

Question: A hidden outlier?

Suppose that you're the production supervisor at a company making a cancer drug. The fermentation process used to make the active ingredient has been running smoothly until, suddenly, alarms are sounding, and you need to decide what to do.

Your production relies on observations on two main variables – temperature (°C) and acidity (pH) – both of which are recorded once per hour.

The data are stored on a webpage, and can be read into R using the following code:

```
1 #Reads the data into R from a webpage:  
2 experiment = read.table("https://www.mas.ncl.ac.uk/~nlf8/cpact/experiment.txt")
```

Load the `experiment` data into RStudio. The data on acidity are stored in the first column; data on temperature are stored in the second column.

(a) Produce time series plots for both acidity and temperature. For example:

```
1 plot(ts(experiment[,1]), main="Plot of acidity values", xlab="Time", ylab="pH")
```

Can you detect anything unusual from your time series plots?

(b) Now produce a scatterplot of acidity against temperature, and comment.

Problem Session 2

Question: Scatterplot matrices

Let's consider the construction of simple scatterplot matrices in R.

- (a) First, import the `heptathlon` dataset and take a look around:

```
1 #Reads the data into R from a webpage:
2 heptathlon = read.table("https://www.mas.ncl.ac.uk/~nlf8/cpact/heptathlon.txt", header=
  TRUE)
3 #Look at the head of the dataset:
4 head(heptathlon)
5
6 #Extract performances for the 100 metres hurdles:
7 heptathlon$H100M
```

- (b) Use R's `apply` command to produce histograms of performances on all 7 variables. Use `par(mfrow=c(2,4))` to partition your plotting window!
- (c) Use the `pairs` command to re-produce the scatterplot matrix from the lecture.
- (i) What is the nature of the association between performances in the 100 metres hurdles and the 800 metres? What about the relationship between performances in the 100 metres hurdles and the long jump?
 - (ii) How are performances in the shot put and the javelin throw associated?

Question: Correlation matrices

Let's consider the construction of correlation matrices in R. Execute the following code in R:

```
1 cor(heptathlon)
```

- (a) Can you explain how the correlation matrix supports your answers in part (c) of the last question?
- (b) Why do all the cells down the diagonal of the matrix contain 1.00000?
- (c) Which pair of variables seem to have the strongest association? What is the *direction* of this association?
- (d) For which pairs of variables might you say there is *no* association? Why should you be careful here?

Problem Session 3

Question: Principal Components Analysis

The primary focus of the lecture was on the use of Principal Components Analysis (PCA) as a way of exploring multivariate datasets. Recall that, for easier interpretation of the principal components, we should make sure the direction of each variable has the same interpretation. In the lecture, we re-scaled data on the running events to make high values correspond to better performances. The following code reads the heptathlon dataset into R (you would have already used this in an earlier question), and then re-scales the running variables so that – for all variables – larger values correspond to better performances:

```
1 #First, copy the original dataset into a new dataframe:
2 heptathlon2 = heptathlon
3
4 #Now re-scale, as we saw in the lecture:
5 heptathlon2$H100M = max(heptathlon2$H100M) - heptathlon2$H100M
6 heptathlon2$R200M = max(heptathlon2$R200M) - heptathlon2$R200M
7 heptathlon2$R800M = max(heptathlon2$R800M) - heptathlon2$R800M
```

Now try the following code to re-produce the analysis from the lecture – making sure you understand the output!

```
1 (pca = prcomp(x=heptathlon2, scale=TRUE))
```

In the final part of the lecture, we will produce a *score plot*. Use the following code to produce the score plot for the heptathlon dataset, then return to the lecture to find out what it shows and what it tells us!

```
1 plot(pca$x[,1], pca$x[,2], xlab="First PC", ylab="Second PC")
2 text(pca$x[,1], pca$x[,2], labels=rownames(heptathlon2))
```

Problem Session 4

Question: Principal Components Analysis

This is an additional question for you to try on Principal Components Analysis.

Load the USArrests dataset into R by typing:

```
1 > USA = read.table("https://www.mas.ncl.ac.uk/~nlf8/cpact/USArrests.txt", header=TRUE)
2
3 #Look at the first 6 rows:
4 > head(USA)
5      State Murder Assault UrbanPop Rape Region
6 1  Alabama   13.2    236      58 21.2  South
7 2   Alaska   10.0    263      48 44.5  North
8 3  Arizona    8.1    294      80 31.0  South
9 4  Arkansas    8.8    190      50 19.5  South
10 5 California    9.0    276      91 40.6  West
11 6  Colorado    7.9    204      78 38.7  West
```

You should see arrests, per 100,000 residents, for assault, murder and rape in each of the 50 US states in 1983. Also given is the percent of the population living in urban areas, and a region identifier. Perform a PCA on this dataset, and use a score plot to detect any clustering.

Analysis of Variance

The R code used throughout this session can be found in plain text format (for copy-and-pasting) at:

<https://www.mas.ncl.ac.uk/~nlf8/cpact/Rcode.R>

Problem Session 1

Question: One-way ANOVA

The following data are from an experiment to compare the effects of four catalysts on the yield of a chemical process:

Catalyst	Observations				
A	4.85	4.89	5.60	5.08	3.42
B	9.43	6.77	5.39	7.39	
C	4.67	5.49	5.96	3.71	5.51
D	5.50	4.90	6.00	5.81	4.68

Type the data manually into R in the following way:

```
1 yield = c(4.85, 4.89, 5.60, 5.08, ..., 4.68)
2 catalyst = c(rep("A", 5), rep("B", 4), rep("C", 5), rep("D", 5))
```

- Using any techniques you have learnt in this course so far, check the assumption that the four samples are Normally distributed.
- Also, informally check the underpinning assumption of equal variances in the four populations from which these samples have been drawn.
- Write down your null hypothesis.
- Now perform a one-way Analysis of Variance using the following code, and interpret the output.

```
1 fit = aov(yield ~ catalyst)
2 summary(fit)
3 (tukeytest = TukeyHSD(fit))
4 plot(tukeytest)
```

Question: Randomised blocks ANOVA

Recall the suncream example from the lecture:

In an experiment to determine the relative effectiveness of four brands of sun cream (Miksa *et al.*, 2016), six volunteers each tried four creams (A, B, C and D) on various parts of their legs. They were then subjected to ultra-violet radiation and an estimate of the degree of burning was made (higher figures indicate greater burning).

The results are given below, along with some summaries:

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Total	mean	s.d.
A	8.6	8.8	7.9	6.4	6.8	8.1	46.6	7.77	0.969
B	5.9	6.2	6.6	6.6	7.4	6.3	39.0	6.50	0.514
C	6.7	9.0	8.4	8.6	6.9	7.2	46.8	7.80	0.982
D	8.2	7.4	8.4	6.9	7.5	7.5	45.9	7.65	0.554

Read the `suncream` data used in the lecture into R:

```
1 #Read the suncream data into R from a webpage:  
2 suncream = read.table("https://www.mas.ncl.ac.uk/~nlf8/cpact/suncream.txt", header=TRUE)
```

- Repeat the analysis from the lecture, checking for differences in the performances of the four suncreams.
- Add an extra term into the model to check the significance of a blocking factor (the 'subject' could be a blocking factor here), and perform the randomised blocks ANOVA in R. What do you conclude?

Problem Session 2

Question: Factorial experiments

Run the following code in R to re-produce the analysis on the `rats` data we have just examined in the lecture. Make sure you understand what's going on each step of the way!

```
1 # Read the dataset in from a webpage:
2 rats = read.table("https://www.mas.ncl.ac.uk/~nlf8/cpact/rats.txt", header=TRUE)
3
4 #Make sure R knows that the source, and level, of protein are both factors:
5 rats$Source = as.factor(rats$Source)
6 rats$Level = as.factor(rats$Level)
7
8 #Produce some summary plots – checking for Normality along the way!
9 par(mfrow=c(1,2))
10 #Boxplots for the data by source of protein:
11 boxplot(rats$Gains~rats$Source)
12 #And now be level of protein:
13 boxplot(rats$Gains~rats$Level)
14
15 #How about Normal Q-Q plots from last week? Easy!
16 par(mfrow=c(2,3))
17 qqnorm(rats$Gains[rats$Source==1], main="Source = Beef")
18 qqline(rats$Gains[rats$Source==1])
19 qqnorm(rats$Gains[rats$Source==2], main="Source = Cereal")
20 qqline(rats$Gains[rats$Source==2])
21 qqnorm(rats$Gains[rats$Source==3], main="Source = Pork")
22 qqline(rats$Gains[rats$Source==3])
23
24 qqnorm(rats$Gains[rats$Level==1], main="Level = High")
25 qqline(rats$Gains[rats$Level==1])
26 qqnorm(rats$Gains[rats$Level==2], main="Level = Low")
27 qqline(rats$Gains[rats$Level==2])
28
29 #Let's now produce some numerical summaries – we can use any function we like, but let's
   check the assumption of equal variances:
30 aggregate(rats$Gains, list(rats$Source), FUN = sd)
31 aggregate(rats$Gains, list(rats$Level), FUN = sd)
32
33 #Now let's do the ANOVA!
34 fit = aov(rats$Gains~rats$Source+rats$Level+rats$Source*rats$Level)
35 summary(fit)
36
37 #And finally, an interaction plot:
38 par(mfrow=c(1,1))
39 interaction.plot(x.factor = rats$Source, #x-axis variable
40                 trace.factor = rats$Level, #variable for lines
41                 response = rats$Gains, #y-axis variable
42                 fun = mean, #metric to plot
43                 ylab = "Weight gain (g)",
44                 xlab = "Source of protein",
45                 col = c("pink", "blue"),
46                 lty = 1, #line type
47                 lwd = 2, #line width
48                 trace.label = "Level of protein")
```

Problem Session 3

Question: ANOVA

A pilot study was developed to investigate the clinical and psychological effects of marijuana use in human subjects. Thirty male subjects, all of whom claimed never to have used marijuana before, received doses in the form of cigarettes of uniform size. Each subject received each of the three treatments: low dose, high dose and placebo, in that order [placebo=0, low=1, high=2].

Load the data into R and then test the hypothesis that heart rate is not affected by smoking marijuana. You can load the data using the following code:

```
1 marijuana = read.table("https://www.mas.ncl.ac.uk/~nlf8/cpact/marijuana.txt", header=TRUE)
```