Data provided: Graph Paper

**MAS6011**

The University Of Sheffield.

**SCHOOL OF MATHEMATICS AND STATISTICS** **Spring Semester 2018–2019**

**Dependent Data** **3 hours**

*Marks will be awarded for your best* ***five*** *answers.*

*RESTRICTED OPEN BOOK EXAMINATION*
*Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator that conforms to University regulations.*
*There are 100 marks available on the paper.*

**Please leave this exam paper on your desk**
**Do not remove it from the hall**

Registration number from U-Card (9 digits)
to be completed by student

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |

Blank

**1** The heptathlon is an athletics event for women consisting of seven disciplines: two throwing events (the shot put and the javelin), two jumping events (the high jump and the long jump), all measured in metres, and three running events (the 200m, the 800m and the 100m hurdles), measured in seconds. Points are awarded for every discipline, with more points awarded for faster (i.e., *lower*) times in running events, and for *larger* distances in throwing and jumping events, and added to give an overall score.

Data were collected from the Olympic Games in Rio in 2016, where the competitors are listed in order of how they finished, with the winner being given first.

Below is part of an R transcript of an analysis on this data.

(i) In the command `princomp(heptathlon[,-8],cor=TRUE)`, discuss why

(a) `heptathlon[,-8]` is used, and not `heptathlon` ***(2 marks)***

(b) `cor=TRUE` is necessary. ***(1 mark)***

(ii) The data set gives the results of the events prior to converting them to points. If the data for each discipline had been presented after converting to points, how might your answers to (i)(b) change? What would you do to decide whether to include `cor=TRUE`? ***(2 marks)***

(iii) In the PCA analysis, it is decided to use only the first few components. Using an informal graphical technique or otherwise, say how many components would you choose, justifying your answer. ***(2 marks)***

(iv) Interpret the first principal component, justifying your answer briefly. There is an outlier on PC1; what are you able to predict about the performances of this athlete? ***(4 marks)***

(v) Interpret the second principal component. The lowest score on PC2 corresponds to the winner of the competition, Thiam. Explain what it is about her performance that leads to such a low score on PC2. ***(3 marks)***

(vi) Interpret the third principal component. One competitor has a very low score for PC3: what can you suggest about her performance? ***(2 marks)***

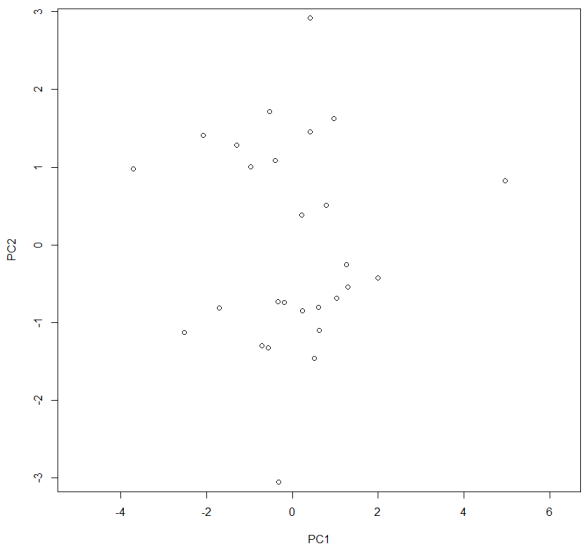
(vii) There is also an outlier on PC4. Which events are likely to give this result? ***(2 marks)***

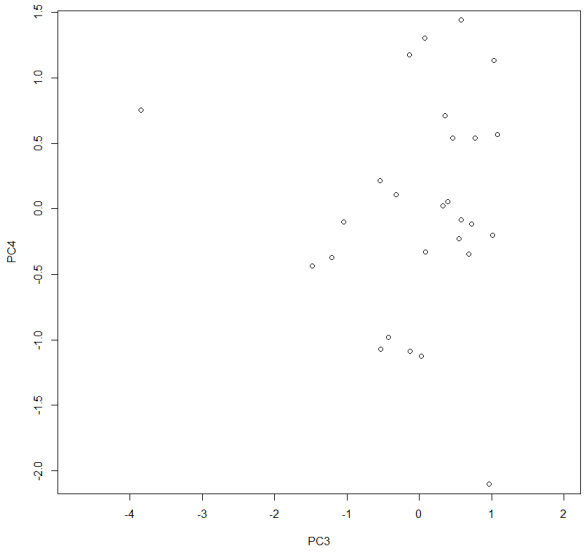
(viii) After projecting the data onto the principal components, suppose that each principal component is scaled to have standard deviation equal to 1. What is the variance matrix of the resulting set? Justify your answer. ***(2 marks)***

**1** (continued)

```
> heptathlon[1:3,]
              hurdles highjump  shot run200m longjump javelin run800m score
        Thiam   13.56     1.98 14.91   25.10     6.58   53.13  136.54  6810
   Ennis-Hill   12.84     1.89 13.86   23.49     6.34   46.06  129.07  6775
Theisen Eaton   13.18     1.86 13.45   24.18     6.48   47.36  129.50  6653

> summary(heptathlon[,-8])
    hurdles   highjump     shotput     run200m   longjump     javelin     run800m
 Min.:12.84  Min.:1.65  Min.:11.21  Min.:23.26  Min.:5.51  Min.:36.36  Min.:126.8
 Q1  :13.26  Q1  :1.77  Q1  :12.94  Q1  :24.11  Q1  :6.09  Q1  :41.47  Q1  :131.7
 Mean:13.54  Mean:1.80  Mean:13.52  Mean:24.55  Mean:6.19  Mean:45.78  Mean:135.8
 Q3  :13.80  Q3  :1.86  Q3  :14.34  Q3  :24.98  Q3  :6.33  Q3  :48.59  Q3  :137.1
 Max.:14.24  Max.:1.98  Max.:14.91  Max.:26.32  Max.:6.58  Max.:55.93  Max.:161.1

> hep.pca<-princomp(heptathlon[,-8],cor=TRUE)
> summary(hep.pca)
Importance of components:
                       Comp.1 Comp.2 Comp.3  Comp.4  Comp.5  Comp.6  Comp.7
Standard deviation     1.5805 1.3001 1.0145 0.82022 0.67692 0.58875 0.55226
Proportion of Variance 0.3568 0.2415 0.1470 0.09611 0.06546 0.04952 0.04357
Cumulative Proportion  0.3568 0.5983 0.7453 0.84145 0.90691 0.95643 1.00000

> loadings(hep.pca)
Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
hurdles   0.416  0.222  0.451  0.476  0.181  0.507  0.241
highjump -0.352 -0.457         0.498  0.491 -0.263  0.324
shotput   0.202 -0.652                0.187  0.321 -0.620
run200m   0.515                0.460 -0.326 -0.603 -0.225
longjump -0.470 -0.216  0.169  0.342 -0.717  0.270
javelin   0.326 -0.504  0.239 -0.405 -0.237         0.597
run800m   0.263 -0.112 -0.837  0.172 -0.122  0.359  0.212

> hep.pc<-predict(hep.pca)
> hep.pc[1:3,]
               Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7
        Thiam -0.3227 -3.0565 0.58052  1.4410 -0.4802 -0.2576  0.5640
   Ennis-Hill -2.5232 -1.1267 0.02755 -1.1242  0.4298 -0.3661 -0.1308
Theisen Eaton -1.7094 -0.8105 0.54783 -0.2297 -0.5016 -0.3994  0.1491

> eqscplot(hep.pc[,1:2])
> eqscplot(hep.pc[,3:4])
```

**1** (continued)



**2** Measurements were taken on two samples of butterflies. The overall sizes of the butterflies were assessed by two measurements, the wing span and the body length. There were 31 butterflies of type A and 25 of type B. The mean measurements obtained are as follows:

| | Wing span (mm) | Body length (mm) |
|---|---|---|
| Type A | 41.63 | 29.78 |
| Type B | 40.40 | 29.14 |

The variance matrix for the group of 31 butterflies of type A is $S_A = \begin{pmatrix} 6.43 & 3.02 \\ 3.02 & 2.09 \end{pmatrix}$ (so the variance of the wing span is 6.43 and the variance for the body length is 2.09), while the variance matrix for the group of 25 type B butterflies is $S_B = \begin{pmatrix} 5.63 & 2.98 \\ 2.98 & 2.16 \end{pmatrix}$, so that the pooled variance matrix (on 54 d.f.) is

$$S = \frac{1}{54}[(31-1)S_A + (25-1)S_B] = \begin{pmatrix} 6.07 & 3.00 \\ 3.00 & 2.12 \end{pmatrix}.$$

You may use R's calculation `qt(0.95,54)=1.674`.

(i) Discuss the correlation between the wing span and the body length for each of the two types. ***(2 marks)***

(ii) Perform a $t$-test to test the hypothesis that the two types have the same wing span. ***(3 marks)***

(iii) Perform a $t$-test to test the hypothesis that the two types have the same body length. ***(3 marks)***

(iv) Using a $T^2$-test, test the bivariate hypothesis that the pair of measurements of the two types are the same.

You are reminded that the inverse of the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is $\frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

You may also use that `qf(0.95,2,n)` is at least 2.5 for all $n$.

Compare your answers with parts (ii) and (iii), and summarise your conclusions. ***(7 marks)***

(v) In fact, there were further observations where the type of the butterfly was not recorded.

Compute Fisher's linear discriminant function for classifying an observation with wing span `ws` and body length `bl` as type A. ***(5 marks)***

**3** (i) We perform a discriminant analysis for two groups in a bivariate situation, where group 1 is modelled as having mean $(-1, 0)$ and variance $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$, and group 2 as having mean $(1, 0)$ and variance $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

(a) Why is this a *quadratic* discriminant analysis, rather than a *linear* discriminant anaysis? ***(1 mark)***

(b) Compute the decision boundary between the two groups, and verify that it is a circle. ***(4 marks)***

(c) Which group would a new reading of $(2, 2)$ be classified as? ***(2 marks)***

(ii) In a quadratic discriminant analysis, the command `qda(data,CV=TRUE)` was used. Describe what the option `CV=TRUE` refers to, and the difference in the output when the option is included. ***(3 marks)***

(iii) Consider the following simple data frame, consisting of 4 observations `Ob1`, `Ob2`, `Ob3` and `Ob4` of 3 variables `X1`, `X2`, `X3` and a response variable `Y`:

| | X1 | X2 | X3 | Y |
|---|---|---|---|---|
| Ob1 | 0 | 1 | 0 | 1 |
| Ob2 | 0 | 1 | 2 | 4 |
| Ob3 | 1 | 0 | 1 | 7 |
| Ob4 | 1 | 1 | 2 | 3 |

Construct a regression tree for this data. ***(5 marks)***

(iv) Five observations with dissimilarity matrix

$$\begin{pmatrix} 0 & 10 & 6 & 8 & 5 \\ 10 & 0 & 9 & 7 & 11 \\ 6 & 9 & 0 & 4 & 13 \\ 8 & 7 & 4 & 0 & 12 \\ 5 & 11 & 13 & 12 & 0 \end{pmatrix}$$

are to be clustered into two classes using hierarchical clustering with single linkage. Find the clusters.

Are the clusters the same if complete linkage is used? ***(5 marks)***

**4** (i) Consider that 200 observations of a time series $\{y_t\}$ gave values of the sample partial autocorrelation function (PACF) and sample autocorrelation function (ACF) tabulated below:

| Lag $h$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| ACF ($r_h$) | 0.4 | 0.3 | 0.1 | 0.1 |
| PACF ($a_h^{(h)}$) | $\star$ | $\star\star$ | 0.01 | 0.04 |

(a) Find the values of $\star$ and $\star\star$. ***(3 marks)***

(b) Test whether $\{y_t\}$ is consistent with moving average models: MA(1) and MA(2). ***(4 marks)***

(c) Test whether $\{y_t\}$ is consistent with autoregressive models: AR(1) and AR(2). ***(3 marks)***

(ii) Consider the time series $\{y_t\}$, defined as

$$y_t = (-1)^t x_t,$$

where $x_t$ is a time series generated by the autoregressive model:

$$x_t = \frac{1}{2}x_{t-1} - \frac{1}{3}x_{t-2} + \epsilon_t, \tag{1}$$

with $\epsilon_t$ being a white noise process with variance 1.

(a) Show that the time series $\{x_t\}$ is causal. ***(4 marks)***

(b) Show that the time series $\{y_t\}$ is stationary. ***(6 marks)***

**5** (i) Consider the non-stationary time series $\{y_t\}$ so that the time series

$$x_t = (1 - B^4)y_t$$

is stationary.

Suppose that $x_t$ is modelled with the autoregressive model

$$x_t = 0.7x_{t-1} + \epsilon_t,$$

where $\epsilon_t$ is white noise with variance 2.

8 observations of $y_t$ are recorded in the table below

| $t$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $y_t$ | 12 | 8 | 14 | 18 | 11 | 9 | 16 | 20 |

(a) Based on the data above, calculate the 1-step and 5-step ahead forecasts of the observations $y_9$ and $y_{13}$, respectively. ***(4 marks)***

(b) Calculate the 1-step and 5-step ahead forecast variances of the observations $y_9$ and $y_{13}$. ***(7 marks)***

(c) Provide a 95% (5-step ahead) prediction interval for the observation $y_{13}$. ***(1 mark)***

(ii) Consider that a time series $\{y_t\}$ is generated by the moving average model (MA):

$$y_t = \epsilon_t + \beta_1\epsilon_{t-1} + \beta_2\epsilon_{t-2},$$

where $\epsilon_t$ is a white noise sequence with variance $\sigma^2$ and the MA parameters $\beta_1$ and $\beta_2$ are assumed known.

Suppose this model is fitted to data $y_{1:n} = (y_1, \ldots, y_n)$, for some given value of $n$. Based on these data, show that a 95% (2-step ahead) prediction interval for the observation $y_{n+2}$ is given by

$$\beta_2\epsilon_n \pm 2\sigma\sqrt{1 + \beta_1^2}.$$

***(8 marks)***

**6** It is well known that an economy's growth as measured by gross domestic product (GDP) is related to unemployment rate. Arthur Okun studied how much GDP is likely to fall, if unemployment increased by a certain level. Let $y_t$ denote the GDP growth of the UK economy at time $t$ and let $x_t$ denote the unemployment rate at time $t$. A simple regression between the two can reveal the likely decrease of the GDP growth, for an increase of the unemployment rate. However, a more elaborate analysis, considers the following dynamic regression model:

$$y_t = \alpha + \gamma_t x_t + \epsilon_t,$$

where $\alpha$ is a static intercept, $\epsilon_t$ is a Gaussian white noise with variance 1 and $\gamma_t$ is a time-varying slope, which follows the autoregressive model

$$\gamma_t = 0.3\gamma_{t-1} + \nu_t,$$

with $\nu_t$ a Gaussian white noise with variance 2. It is further assumed that $\epsilon_t$ and $\nu_s$ are independent for any $t, s$ and that $\gamma_0$ is independent of $\nu_t$, for any $t$.

(i) Define the state vector

$$\beta_t = \begin{bmatrix} \alpha \\ \gamma_t \end{bmatrix}.$$

Write the above model in state space form,

$$y_t = L_t^T \beta_t + \kappa_t,$$
$$\beta_t = F\beta_{t-1} + \zeta_t,$$

and determine the design vector $L_t$ and the evolution matrix $F$. Write down $\kappa_t$ and $\zeta_t$ and obtain their distributions. ***(4 marks)***

(ii) The above state space model is fitted to data of length $n$. The posterior distribution of $\beta_n$, given information $y_{1:n} = \{y_1, \ldots, y_n\}$ is

$$\beta_n \mid y_{1:n} \sim N\left\{ \begin{bmatrix} -1.5 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix} \right\}.$$

(a) If $x_{n+1} = 0.5$ and $y_{n+1} = 1$, then obtain the posterior distribution $p(\beta_{n+1} \mid y_{1:n+1})$ of $\beta_{n+1}$, given information $y_{1:n+1}$. ***(14 marks)***

(b) Provide a 95% credible interval for $\gamma_{n+1}$, given $y_{1:n+1}$. ***(2 marks)***

## End of Question Paper