The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2019–2020**

Dependent Data **1.5 hours**

This is an open book exam.
*Answer **all** questions. The allocation of marks is shown in brackets.*
There are 75 marks available on the paper.

The submission deadline is 10 am (BST), twenty-four hours after it is released. Late submission will not be considered without extenuating circumstances. It is expected that you will be able to complete this exam in approximately one and a half hour and it is recommended that you submit the work within four hours. You will not be penalised for taking longer, however.

Unless it is explicitly stated otherwise, it is intended that calculations are performed by hand (possibly with the aid of a calculator). To gain full marks, you will need to show your working. You will not get full marks if you simply write down output from a computer package.

By uploading your solutions you declare that your submission consists entirely of your own work, that any use of sources or tools other than material provided for this module is cited and acknowledged and that no unfair means have been used.

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 times and longer distances, and summed to give an overall score.

Data were collected from the Olympic Games in Seoul in 1988, where the competitors are listed in order of how they finished, with the winner being given first.

Below is part of an R transcript on the results; parts (i)–(vii) of this question refer to 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, how many components would you choose? ***(3 marks)***

(iv) Interpret the first principal component, justifying your answer briefly. What does the competition winner score on this component? ***(3 marks)***

(v) Interpret the second principal component. The data set omits the competitors who did not complete the event, but, of those who finished, the competitor who came last overall did very well at one event. Which event was this? ***(3 marks)***

(vi) Interpret the third principal component. A Swiss athlete appears to be an outlier on PC3. How might her performance have been? ***(3 marks)***

(vii) A Chinese athlete did rather well at the high jump, but was a slow 800m runner. Where might one expect to see her on the last graph? ***(2 marks)***

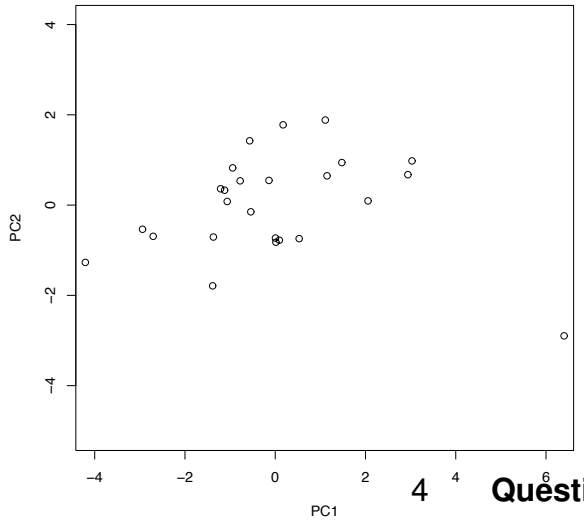
(viii) Suppose that two correlation matrices S_1 and S_2 satisfy $S_1 = tS_2 + (1-t)I_p$. Explain that the *ratios* between the correlations of pairs of distinct variables are the same for S_1 and S_2. Show that Principal Components Analysis (using the correlation matrix) only depends on these ratios in the sense that the principal components for the matrix S_1 are the same as those for S_2. ***(3 marks)***

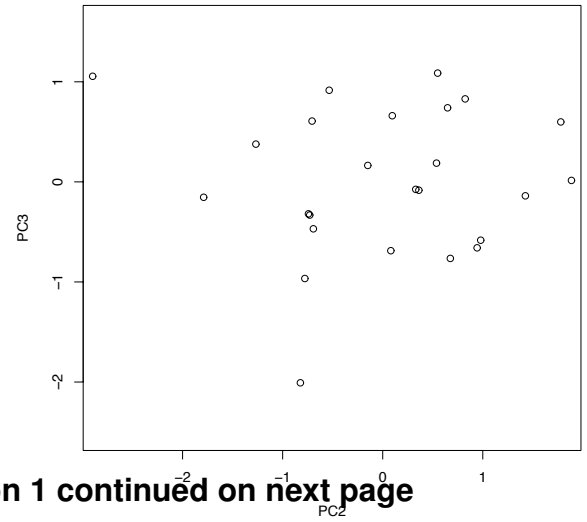
(ix) If PCA is applied to a $p \times p$ correlation matrix R with $R_{ij} = \rho$ for all $i \neq j$ (as is often the case in biological situations), then the first principal component is (proportional to) $(1, \ldots, 1)'$. What proportion of the total variance does this explain? ***(3 marks)***

```
> attach(heptathlon)
```

1 (continued)

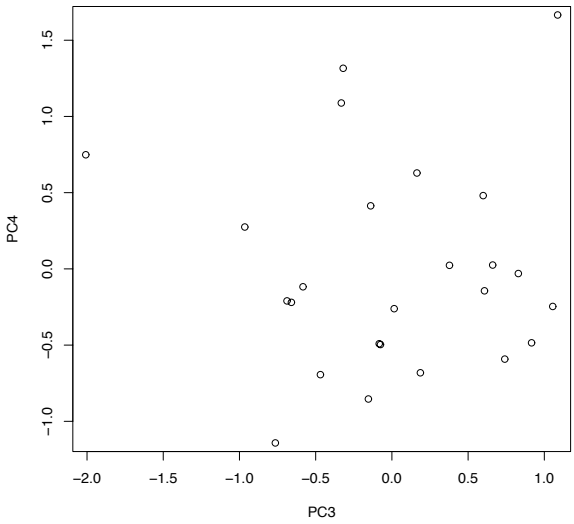
```
> heptathlon[1:3,]
              hurdles highjump  shot run200m longjump javelin run800m score
Joyner-Kersee   12.69     1.86 15.80   22.56     7.27   45.66  128.51  7291
John            12.85     1.80 16.23   23.65     6.71   42.56  126.12  6897
Behmer          13.20     1.83 14.20   23.10     6.68   44.54  124.20  6858

> 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     2.11194 1.09285 0.721813 0.67614 0.495244 0.270103 0.2213617
Proportion of Variance 0.63718 0.17062 0.074431 0.06531 0.035038 0.010422 0.0070001
Cumulative Proportion  0.63718 0.80780 0.882230 0.94754 0.982578 0.993000 1.0000000
> loadings(hep.pca)
Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
hurdles   0.453 -0.158                 0.783 -0.380
highjump -0.377  0.248 -0.368  0.680               -0.434
shot     -0.363 -0.289  0.676  0.124 -0.512        -0.218
run200m   0.408  0.260         0.361 -0.650         0.453
longjump -0.456         0.139  0.111  0.184  0.590  0.612
javelin         -0.842 -0.472  0.121 -0.135         0.173
run800m   0.375 -0.224  0.396  0.603  0.504 -0.156
>
> hep.pc<-predict(hep.pca)
> hep.pc[1:3,]
               Comp.1   Comp.2   Comp.3    Comp.4   Comp.5   Comp.6   Comp.7
Joyner-Kersee -4.2064 -1.26802  0.37754  0.023476  0.43479  0.34633  0.35510
John          -2.9416 -0.53453  0.91592 -0.485256 -0.71756 -0.24300  0.14699
Behmer        -2.7043 -0.69276 -0.46865 -0.693643  0.10770  0.24412 -0.13232

> eqscplot(hep.pc[,1],hep.pc[,2])
> eqscplot(hep.pc[,2],hep.pc[,3])
> eqscplot(hep.pc[,3],hep.pc[,4])
> eqscplot(hep.pc[,4],hep.pc[,5])
```

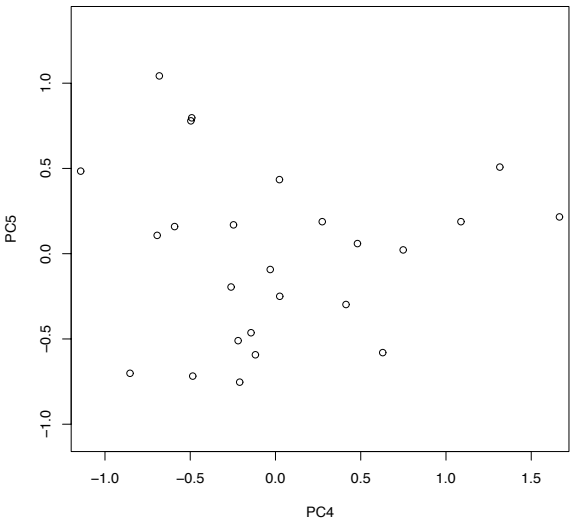




 Question 1 continued on next page

1(continued)


PC4
PC3
1.5
1.0
0.5
0.0
−0.5
−1.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0


PC5
PC4
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
1.5

2 (i) A time series $\{y_t\}$ consisting of 8 observations on quarterly average of daily temperatures (in degrees Celsius) over two years is shown in the table below.

	Year 1				Year 2			
Quarter	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
t	1	2	3	4	5	6	7	8
y_t	8	13	16	10	6	11	15	12

In order to analyse this data the decomposition method is considered, according to which the time series y_t is decomposed into trend m_t, seasonal s_t and residual components r_t, so that

$$y_t = m_t + s_t + r_t.$$

(a) Using a 3-point moving average calculate estimates of m_t, for $t = 2, 3, \ldots, 7$. ***(2 marks)***

(b) Provide estimates of the seasonals s_{Q1}, s_{Q2}, s_{Q3} and s_{Q4}, for quarters Q1, Q2, Q3 and Q4. ***(6 marks)***

(c) Calculate estimates of the residual term r_t, for $t = 2, 3, \ldots, 7$. ***(5 marks)***

(ii) Consider the time series y_t is generated from the time series model

$$y_t = \epsilon_t + \frac{1}{2}\epsilon_{t-1} - \frac{1}{4}\epsilon_{t-2},$$

where ϵ_t is a white noise sequence with variance 2.

(a) Show that y_t is invertible. ***(2 marks)***

(b) Calculate the variance of y_t. ***(1 mark)***

(c) Provide the autocorrelation function (ACF) of y_t. ***(5 marks)***

(d) Calculate the first two values $a_1^{(1)}$ and $a_2^{(2)}$ of the partial autocorrelation function (PACF) of y_t. ***(4 marks)***

3 The price p_t of an asset traded in the stock market follows the evolution given below

$$p_t = p_{t-1} \exp(r_t),$$

where r_t denotes the logarithmic return of the asset at time t.

A time series model for the returns r_t is adopted given by

$$r_t = 0.8 r_{t-1} + \kappa_t,$$

where κ_t is a white noise process, following a normal distribution with variance 4.

Define the state vector

$$\beta_t = \begin{bmatrix} \log p_{t-1} \\ r_t \end{bmatrix}.$$

(i) Find a state-space model for the log-price $\log p_t$ of A, i.e.

$$\log p_t = x_t^T \beta_t + \epsilon_t$$
$$\beta_t = F \beta_{t-1} + \zeta_t$$

and determine x_t and F and state the distributions of ϵ_t and ζ_t.

(8 marks)

(ii) After 100 trading days the posterior distribution of β_{100}, given data $p_{1:100} = \{p_1, p_2, \ldots, p_{100}\}$ is:

$$\beta_{100} \mid p_{1:100} \sim N\left\{ \begin{bmatrix} 5 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix} \right\}.$$

(a) Provide the 2-step ahead forecast distribution of $\log p_{102}$, given data $p_1, p_2, \ldots, p_{100}$. ***(15 marks)***

(b) Obtain a 95% predictive interval for the return r_{102}, given data $p_1, p_2, \ldots, p_{100}$. ***(2 marks)***

End of Question Paper