The University Of Sheffield.

PLEASE LEAVE THIS EXAM PAPER ON YOUR DESK.
DO NOT REMOVE IT FROM THE HALL.

Data Provided:
Neaves Tables
Graph Paper

SCHOOL OF MATHEMATICS AND STATISTICS

MAS6011

Session 2010-2011

3 Hours

Dependent Data

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.
*All answers will be marked but credit will be given for only the best **FIVE** answers.*
All questions carry equal marks. Total marks 100.

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

(This page is left blank)

1 Measurements (in mmHg) of systolic and diastolic blood pressure were made on 22 subjects, 11 of whom were treated with a placebo and 11 with a beta-blocker drug. The means of the systolic and diastolic blood pressures in the placebo readings were 134.1 and 88.7, respectively, with variances 46.0 and 55.8. The covariance between systolic and diastolic readings in this group was 45.4. The corresponding five summary results in the beta-blocker treated group were 126.9, 84.1, 49.5, 60.6 and 46.2.

(a) Do these data provide evidence of an overall difference in mean blood pressure between the placebo and beta-blocker groups?

(8 marks)

(b) What linear combination of systolic and diastolic pressures exhibits the greatest difference between the groups?

(3 marks)

(c) After the analysis had been performed the clinician involved in the study congratulated the statistician on using the multivariate test performed in part (a), saying that often studies only looked at systolic and diastolic measures separately, but said that ideally she would like to see the conventional mean summary measure of (systolic + $2 \times$ diastolic) /3 included as a third variable in the multivariate test. What response should you give to this request and what justification would you give?

(3 marks)

(d) The clinician also mentioned that there were in fact only 11 subjects in all, not 22. All 11 subjects had first been treated with the placebo and then later with the beta-blocker.

i) How would you wish to modify your analysis to take advantage of this extra information?

(3 marks)

ii) What further sample covariances would be required to enable you to perform the calculations for this further analysis?

(3 marks)

2 Collett (2003) describes the measurements of fibrinogen and gamma globulin in 32 subjects whose ESR (Erythrocytes Sedimentation Rate) had also been determined. Subjects with ESR below 20 are considered healthy. The mean values of fibrinogen and gamma globulin of the 26 subjects with ESR < 20 were 2.65 and 35.12 respectively. The corresponding values of those not classified as healthy by their ESR measurement were 3.39 and 38.00. The pooled within group variances of the two measurements were 0.330 and 20.360 with covariance –0.102. A key aim of the study was to see if it was possible to predict whether subjects could be classified as healthy without determining their ESR.

(a) Estimate Fisher's linear discriminant function for classifying a subject as healthy or not on the basis of measurements of fibrinogen and gamma globulin.

(8 marks)

(b) Assuming that these measurements are adequately modelled by bivariate Normal distributions with a common variance and that the classification of subjects uses Fisher's discriminant function, estimate the probability of misclassifying a randomly selected healthy subject as unhealthy.

(4 marks)

(c) Suppose, for a particular subject, the gamma globulin measurement is not available. What value of fibrinogen should be used as a lower limit to ensure that the probability of missing an unhealthy subject is the same as that using the rule determined in part (a)?

(5 marks)

(d) What proportion of healthy subjects will be falsely diagnosed as unhealthy by the rule in part (c)?

(3 marks)

3 Measurements were made on the dimensions of the wings of two sub-species of the *spiravolarens* butterfly, *atalanta* and *britannicus*. The first is an infrequent summer visitor to the south coast and the second is a native species. A scientific museum has provided measurements of the lengths and widths of left and right wings of a sample of 26 specimens of *s.atalanta* from its reference collection and an amateur entomologist has provided similar measurements of 21 specimens of *s.britannicus* from his personal collection. The length of the wing is taken as the maximum diameter parallel to the body and the width as the maximum diameter orthogonal to it. The initial aims of the analysis are to gain insight into the nature of the variation in sizes of the wings, investigate the differences between the wings of the two sub-species and ultimately the possibility of using wing measurements as an aid to identification. Given below is a record of an **R** session (edited in places) performing various preliminary analyses both separately on the measurements and on the two sets of measurements grouped together.

(a) Principal component analyses have been performed on the correlation rather than the covariance matrices. Why in all three cases is this choice likely to be more informative than the alternative?

(1 mark)

(b) What features of the wings in the first two analyses on the separate species do the first three principal components derived from the correlation matrices reflect?

(7 marks)

(c) Scatterplots have been produced from the scores on the four principal components calculated from the correlation matrix of the combined data. What recommendation would you give regarding the advisability of investigating further the use of these wing measurements for distinguishing *s.atalanta* from *s.britannicus*?

(3 marks)

(d) What feature of the principal component analysis on *s.britannicus* is unusual and why might this make people suspect that the data on this species have been faked (i.e. are artificial)?

(3 marks)

(e) The amateur entomologist (who is also an amateur statistician) suggests extending the Principal Component Analysis by feeding the scores on the four principal components referred to in part (c) back into the PCA routine of the statistical package in the optimistic hope that 'better results' will be produced. What advice should you give to the entomologist regarding this idea?

(3 marks)

(f) What will the eigenvalues be if the PCA in (d) is performed on the correlation matrix of the principal component scores?

(3 marks)

Question 3 continued on next page

Question 3 continued

Analysis of measurements of *spiravolarens* butterflies

```
***  Summary Statistics for data in:  butterflies ***
     $$$"Factor Summaries":
        species
    atalanta:26
 britannicus:21

     $$$"Numeric Summaries":
          left.length left.width right.length right.width
    Mean:        7.02       9.04         7.31       10.67
 Total N:       47.00      47.00        47.00       47.00
Std Dev.:        1.29       3.78         1.49        3.48
------------------------------------------------

species:atalanta
     $$$"Factor Summaries":
        species
    atalanta:26
 britannicus: 0
     species: 0

 $$$"Numeric Summaries":
          left.length left.width right.length right.width
    Mean:        6.99      10.67         7.16       10.79
 Total N:       26.00      26.00        26.00       26.00
Std Dev.:        1.45       3.08         1.38        3.71

------------------------------------------------
species:britannicus
     $$$"Factor Summaries":
        species
    atalanta: 0
 britannicus:21
     species: 0

     $$$"Numeric Summaries":
          left.length left.width right.length right.width
    Mean:        7.07       7.01         7.51       10.51
 Total N:       21.00      21.00        21.00       21.00
Std Dev.:        1.11       3.65         1.64        3.26
```

Question 3 continued on next page

Question 3 continued

Principal Component Analyses for *s.atlanta*:

```
> pca.atlantacor<- princomp(butterflies[1:26,-1],cor=T)
> summary(pca.atlantacor)
Importance of components:
                          Comp.1 Comp.2 Comp.3  Comp.4
    Standard deviation     1.48  1.263 0.4333 0.12881
Proportion of Variance     0.55  0.399 0.0469 0.00415
 Cumulative Proportion     0.55  0.949 0.9959 1.00000
> print(loadings(pca.atlantacor),cutoff=0.1)
             Comp.1 Comp.2 Comp.3 Comp.4
 left.length  0.530 -0.471 -0.342  0.617
  left.width  0.483  0.510 -0.611 -0.365
right.length  0.536 -0.456  0.396 -0.590
 right.width  0.445  0.558  0.594  0.372
```

Principal Component Analyses for *s.Britannicus*:

```
>
> pca.britancor<- princomp(butterflies[27:47,-1],cor=T)
> summary(pca.britancor)
Importance of components:
                          Comp.1 Comp.2 Comp.3 Comp.4
    Standard deviation     1.251  1.069  0.887  0.710
Proportion of Variance     0.391  0.286  0.197  0.126
 Cumulative Proportion     0.391  0.677  0.874  1.000
> print(loadings(pca.britancor),cutoff=0.1)
             Comp.1 Comp.2 Comp.3 Comp.4
 left.length -0.417 -0.683  0.214  0.560
  left.width  0.636  0.251  0.306  0.663
right.length -0.439  0.429  0.782 -0.102
 right.width  0.478 -0.535  0.499 -0.486
>
```

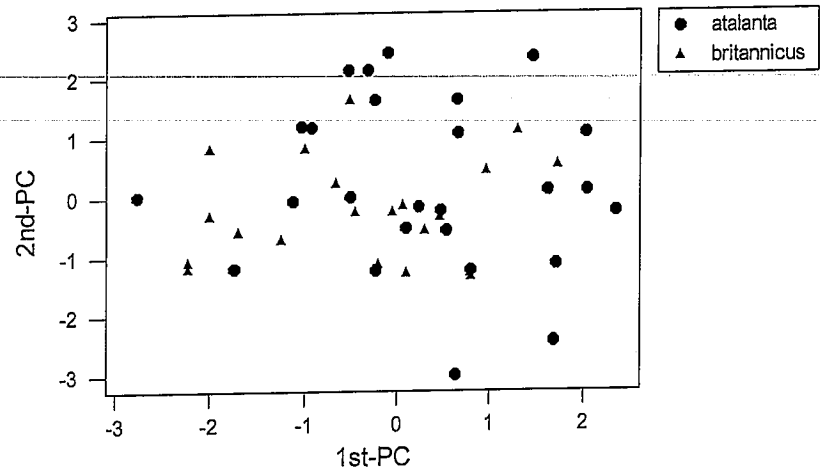
Principal Component Analyses for combined group:

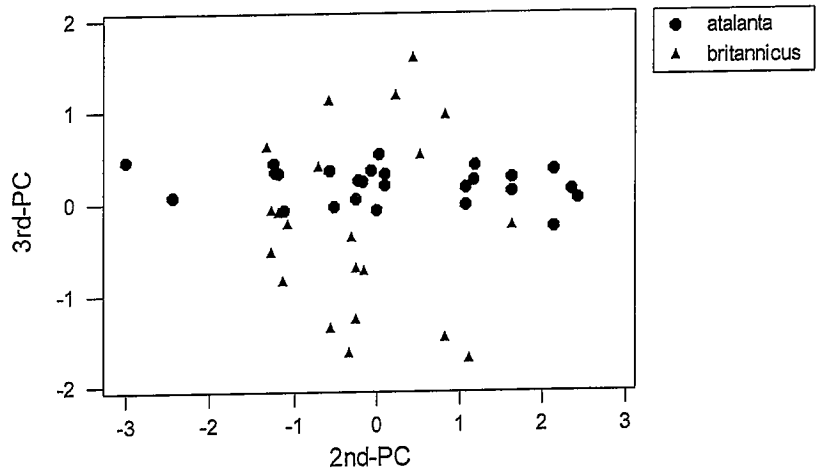
```
> pca.bothcor<- princomp(butterflies[,-1],cor=T)
> summary(pca.bothcor)
Importance of components:
                          Comp.1 Comp.2 Comp.3 Comp.4
    Standard deviation     1.261  1.226  0.700  0.646
Proportion of Variance     0.398  0.375  0.123  0.104
 Cumulative Proportion     0.398  0.773  0.896  1.000
> print(loadings(pca.bothcor),cutoff=0.1)
             Comp.1 Comp.2 Comp.3 Comp.4
 left.length -0.530 -0.472 -0.396 -0.583
  left.width  0.472 -0.523  0.584 -0.404
right.length -0.574 -0.407  0.469  0.534
 right.width  0.408 -0.582 -0.532  0.461
>
```

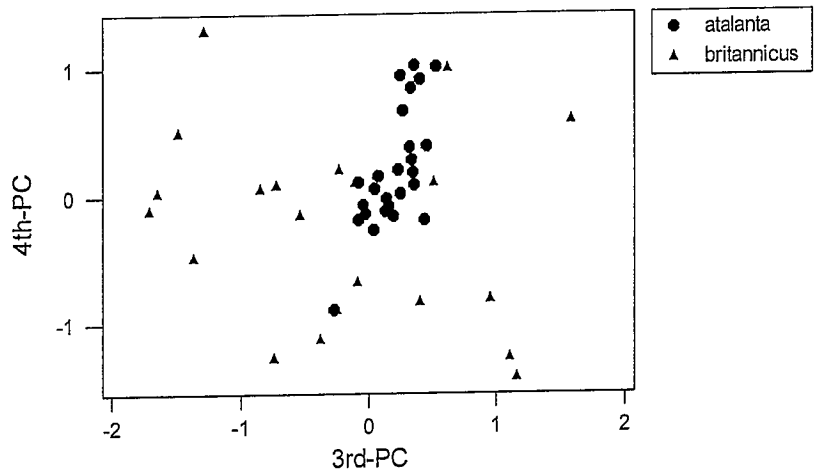
Question 3 continued on next page

Question 3 continued

Principal Component Score Plots for combined group:







4 (i) (a) In the context of descriptive analysis of time series x_t, briefly explain why a moving average for even span s is *not* defined as

$$\frac{1}{s}(x_{t-s/2} + x_{t-s/2-1} + \cdots + x_{t-1} + x_t + x_{t+1} + \cdots + x_{t+s/2-1}).$$

(1 mark)

(b) Consider the time series with values

$$x_1 = 5, \quad x_2 = 4, \quad x_3 = 6, \quad x_4 = 5, \quad x_5 = 7, \quad x_6 = 6, \quad x_7 = 3.$$

Using the *correct* definition of the even-span moving average, calculate moving averages of span 4, for the values x_3, x_4 and x_5. *(3 marks)*

(ii) A time series of length 70 gave values for the sample autocorrelation function (ACF), denoted by r_h and values for the partial ACF, denoted by a_h, according to the table below.

Lag h	1	2	3	4
r_h	0.58	0.43	0.37	0.22
a_h	*	*	0.19	0.21

(a) Using this table, find the values of a_1 and a_2, indicated in the table by stars. *(4 marks)*

(b) Test whether this time series is consistent with a white noise process, a moving average model and an autoregressive model. *(10 marks)*

(c) Suggest a model which you would expect to fit well to this time series data. *(2 marks)*

5 Consider the time series model

$$X_t = \frac{1}{2}X_{t-1} + \epsilon_t + \frac{1}{3}\epsilon_{t-1} + \frac{1}{4}\epsilon_{t-2}, \tag{1}$$

where ϵ_t is a white noise process with variance 3, i.e. $\epsilon_t \sim WN(0, 3)$.

(i) Give the abbreviated name of the model for X_t. *(1 mark)*

(ii) Write down model (1) in compact form, using the backward shift operator B. *(2 marks)*

(iii) Show that model (1) is causal and invertible. *(5 marks)*

(iv) Find the variance of X_t. *(12 marks)*

6 Consider the trend dynamic linear model, given by equations

$$X_t = [1, 0]\begin{bmatrix} \theta_{1t} \\ \theta_{2t} \end{bmatrix} + \epsilon_t = \boldsymbol{F}^T\boldsymbol{\theta}_t + \epsilon_t, \qquad (2)$$

$$\boldsymbol{\theta}_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t = \boldsymbol{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \qquad (3)$$

where $\boldsymbol{\theta}_t = [\theta_{1t}, \theta_{2t}]^T$ is a state vector, ϵ_t follows a normal distribution with zero mean and variance 50, and $\boldsymbol{\omega}_t$ follows a bivariate normal distribution with zero mean vector and covariance matrix

$$\boldsymbol{W} = \begin{bmatrix} 20 & 0 \\ 0 & 20 \end{bmatrix},$$

written as

$$\boldsymbol{\omega}_t \sim N_2\left\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 20 & 0 \\ 0 & 20 \end{bmatrix}\right\}.$$

It is also assumed that ϵ_t and $\boldsymbol{\omega}_t$ are mutually and individually independent, and they are independent of the initial state $\boldsymbol{\theta}_0$. Suppose that $x_1, x_2, \ldots, x_n$ values of the time series are observed and that the posterior distribution of $\boldsymbol{\theta}_n$, given information $x^n = (x_1, \ldots, x_n)$ is given by

$$\boldsymbol{\theta}_n | x^n \sim N_2\left\{\begin{bmatrix} 250 \\ 100 \end{bmatrix}, \begin{bmatrix} 10 & 0 \\ 0 & 33 \end{bmatrix}\right\}.$$

For some positive integer $k > 0$, define the new time series

$$S_n = X_{n+1} + X_{n+2} + \cdots + X_{n+k}.$$

(i) Show that the k-step forecast function of $\{X_t\}$ is $\widehat{X}_{n+k} = E(X_{n+k}|x^n) = 100k + 250$. *(4 marks)*

(ii) Find the posterior mean of S_n, given x^n, for $k = 2$. *(2 marks)*

(iii) For $k = 2$, show that, given x^n, the covariance of X_{n+1} and X_{n+2} is 96, and hence calculate the posterior variance of S_n, given x^n. *(13 marks)*

(iv) Derive the posterior distribution of S_n, given x^n, for $k = 2$. *(1 mark)*

End of Question Paper