Data provided: Neave's Tables, Graph Paper

MAS6011

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2013–2014**

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 As part of an investigation into determining possible locations of diamond deposits in Australia, data were collated giving the numbers of geographical micro-deposits in various categories found at 90 different sites. These sites included 11 sites (numbered 80 to 90) where diamonds have been found; no diamonds have been found in the other 79 sites (numbered 1 to 79). The five categories recorded were Igneous (`ign`), Igneous/Calcific (`ign.calc`), Sedimentary (`sed`), Metamorphic Sedimentary (`meta.sed`) and Amorphous (`amo`).

Given below is an edited record of various preliminary analyses of these data using R.

(i) The principal component analysis has been performed using the correlation matrix. Would you recommend instead using the variance matrix? Justify your recommendation. ***(2 marks)***

(ii) With the aid of an informal graphical technique, how many principal components would you recommend retaining for further exploratory analyses? ***(2 marks)***

(iii) What features of the sites do the three most important principal components reflect? ***(3 marks)***

(iv) What characteristics of the sites (in terms of the categories of deposits found at them) seem to be typical of the majority of the diamond sites? Explain your answers. ***(4 marks)***

(v) Two additional sites are under consideration for further intensive excavation in the hope of identifying diamond deposits, but resources are only sufficient for a single expedition to one of the sites. The numbers (respectively) of Igneous, Igneous/Calcific, Sedimentary, Metamorphic Sedimentary and Amorphous recorded at Site A are 6, 0, 4, 1 and 1. At Site B, they were 7, 3, 0, 1 and 0. Upon which site would you recommend concentrating the available resources? ***(4 marks)***

(vi) A colleague notices that the analysis uses the function `princomp`, and believes that `prcomp` is meant to have certain advantages numerically. Looking up the help page, he spots that `prcomp` uses the formula $S = \frac{1}{n-1}(X - \overline{X})(X - \overline{X})'$ for the variance matrix, whereas `princomp` uses $S = \frac{1}{n}(X - \overline{X})(X - \overline{X})'$. What differences, if any, would this make to the R analysis below? And would it have any effect on your answer to part (ii)? Justify your answers. ***(3 marks)***

1 (continued)

(vii) 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)

```
> attach(diamonds)
> library(MASS)
> apply(diamonds[1:79,-6],2,mean)
   ign ign.calc    sed meta.sed     amo
 5.443   0.4684 1.3544  0.40506 0.18987
> apply(diamonds[1:79,-6],2,sdev)
   ign ign.calc    sed meta.sed     amo
 9.316   1.3759 2.4075  0.75987 0.39471
```

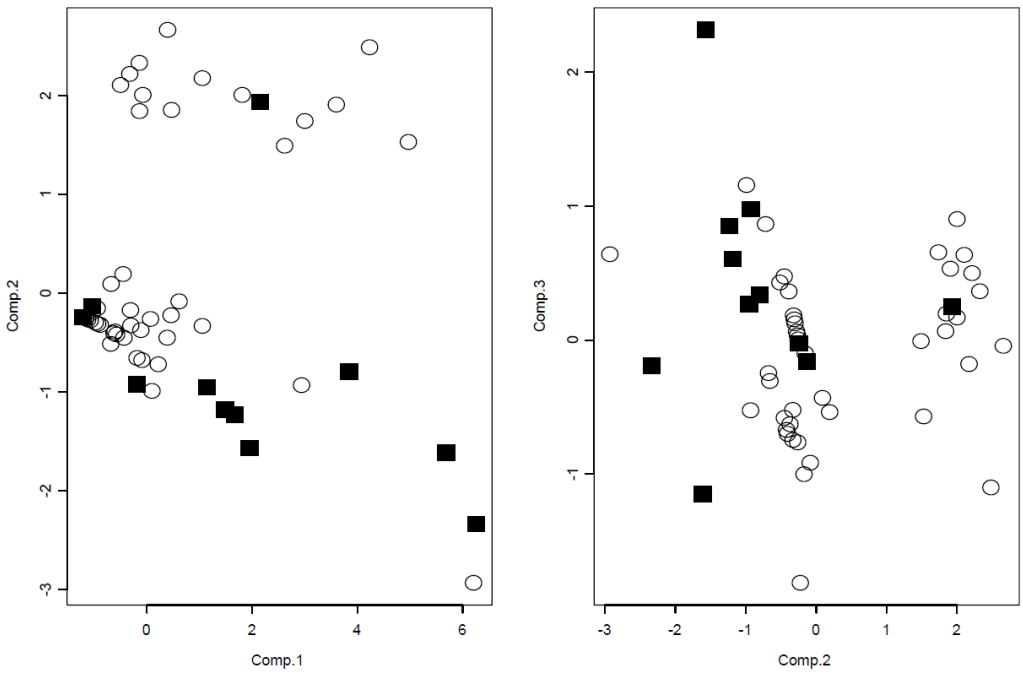
1 (continued)

```
> apply(diamonds[80:90,-6],2,mean)
   ign ign.calc    sed meta.sed     amo
20.182   3.7273 3.4545   1.1818 0.09091
> apply(diamonds[80:90,-6],2,sdev)
   ign ign.calc    sed meta.sed     amo
14.586   2.6867 3.5879   1.4709 0.30151

> dia.pca<-princomp(diamonds[-6],cor=T)
> summary(dia.pca)
Importance of components:
                        Comp.1  Comp.2   Comp.3   Comp.4   Comp.5
Standard deviation     1.79353 1.04910 0.536479 0.521141 0.351077
Proportion of Variance 0.64335 0.22012 0.057562 0.054318 0.024651
Cumulative Proportion  0.64335 0.86347 0.921031 0.975349 1.000000

> loadings(dia.pca)
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
     ign  0.522 -0.123  0.329         0.776
ign.calc  0.471 -0.387  0.520        -0.598
     sed  0.473  0.293 -0.357  0.734 -0.155
meta.sed  0.490 -0.137 -0.633 -0.581
     amo  0.205  0.855  0.304 -0.349 -0.114

> par(mfrow=c(1,2))
> plot(dia.pc[,1:2],type='n')
> points(dia.pc[1:79,1:2],pch=1)
> points(dia.pc[80:90,1:2],pch=15)
> plot(dia.pc[,2:3],type='n')
> points(dia.pc[1:79,2:3],pch=1)
> points(dia.pc[80:90,2:3],pch=15)
```



2 Measurements were taken on a sample of children in a European town on their second birthdays. The overall sizes of the children were assessed by two measurements, the height and the chest circumference. In total, the sample consisted of 31 boys and 25 girls. The mean lengths obtained are as follows:

	Height (cm)	Chest (cm)
Boys	83.26	59.55
Girls	80.79	58.28

The variance matrix for the group of 31 boys is $S_B = \begin{pmatrix} 25.72 & 12.09 \\ 12.09 & 8.36 \end{pmatrix}$ (so the variance of the height is 25.72 and the variance for the chest is 8.36), while the variance matrix for the group of 25 girls is $S_G = \begin{pmatrix} 22.32 & 11.91 \\ 11.91 & 8.65 \end{pmatrix}$, so that the pooled variance matrix (on 54 d.f.) is

$$S = \frac{1}{54}[(31-1)S_B + (25-1)S_G] = \begin{pmatrix} 24.21 & 12.01 \\ 12.01 & 8.49 \end{pmatrix}.$$

(i) Do the data provide evidence that the boys are taller than the girls? ***(3 marks)***

(ii) Do the data provide evidence that the boys have larger chest measurements than the girls? ***(3 marks)***

(iii) Test the hypothesis that the height and chest measurements of the group of boys is the same as that of the girls. Compare your answers with parts (i) and (ii), and summarise your conclusions. ***(7 marks)***

(iv) The experiment was partly conducted to compare the results with an earlier large study on a group of Australian boys on their second birthdays. The variance in the Australian study was found to be $\begin{pmatrix} 25.89 & 13.01 \\ 13.01 & 10.21 \end{pmatrix}$. Use a likelihood-ratio test to test the hypothesis that the variance of the European boys is the same as that found in the Australian study. You may assume any standard results on MLEs, and may also assume that the sample size is sufficiently large that Wilks's Theorem applies. ***(7 marks)***

3 Johnson and Wichern (2002) report on a study into potential haemophilia A carriers, consisting of a group of 30 subjects without the haemophilia gene (the *non-carrier group*), and a group of 22 subjects who were known haemophilia carriers (the *carrier group*). Measurements were made of two variables; X_1 is related to antihaemophiliac factor activity, and X_2 to antihaemophiliac-like antigens. (Since the quantities involved were recorded on a logarithmic scale, some of the entries are negative.)

The investigators provided information

$$\overline{x}_N = \begin{pmatrix} -0.0065 \\ -0.0390 \end{pmatrix}, \quad \overline{x}_C = \begin{pmatrix} -0.2483 \\ 0.0262 \end{pmatrix},$$

and

$$S^{-1} = \begin{pmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{pmatrix},$$

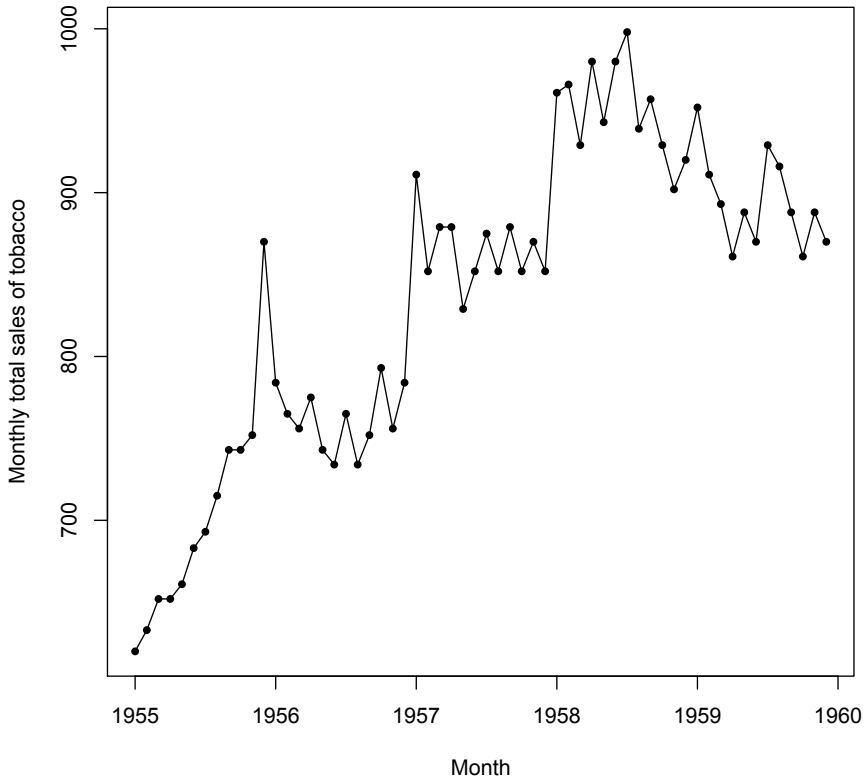
where $\overline{x}_N$ and $\overline{x}_C$ denote the sample mean for readings of X_1 and X_2 for the non-carrier group and carrier group respectively, and S is the pooled sample variance matrix.

(i) Estimate Fisher's linear discriminant function for classifying a subject as in the carrier group or not on the basis of the measurements of X_1 and X_2. ***(6 marks)***

(ii) Informal investigations suggest that the data for each group is reasonably well approximated by a bivariate normal distribution, and, further, that the variance matrices for both groups appear to be very similar, so that they may be assumed to be the same. Using your function from part (i) to classify observations, estimate the probability that a randomly selected noncarrier is misclassified as a carrier. ***(5 marks)***

(iii) The cost of measuring the variable X_2 is high, and it is hoped to develop a test using only the value of X_1. What value should be used as a lower limit to ensure that the probability of missing a carrier is the same as that using the rule determined in part (i)? ***(5 marks)***

(iv) What proportion of non-carriers will be falsely diagnosed as carriers by the rule in part (iii)? ***(4 marks)***



4 (i) The plot above shows data consisting of monthly total sales (in some standard scale) of UK tobacco and related products in the period 1955 to 1959[1]. Briefly describe the features of the data. ***(2 marks)***

[1]Source: West, M. and Harrison, P.J. (1997) Bayesian Forecasting and Dynamic Models, Springer

4 (continued)

(ii) For a new time series z_t with length 20, the sample ACF and the sample PACF are tabulated below:

Lag	1	2	3	4	5
ACF	0.62	0.57	0.30	0.10	0.05

and

Lag	1	2	3	4	5
PACF	$a_1^{(1)}$	$a_2^{(2)}$	0.28	0.15	0.01

(a) Determine whether z_t is stationary or not and give a brief explanation. ***(1 mark)***

(b) Find the values of $a_1^{(1)}$ and $a_2^{(2)}$. ***(4 marks)***

(c) Test whether z_t is a white noise. ***(3 marks)***

(d) Test whether z_t is consistent with autoregressive models. ***(3 marks)***

(e) Test whether z_t is consistent with moving average models. ***(5 marks)***

(f) Based on your analysis above, suggest a time series model for z_t that is likely to perform well when fitted to the data. ***(2 marks)***

5 Suppose that a model is set up for a seasonal time series y_t so that the transformed series x_t, defined by

$$x_t = (1 - B^4)^3 y_t,$$

where B is the backward shift operator, follows the AR(1) model

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

where ϵ_t is white noise with variance 1.

(i) Write down the abbreviated form of the model of y_t: SARIMA$(p, d, q) \times (P, D, Q)_s$, i.e. identify p, d, q, P, D, Q and s. ***(1 mark)***

(ii) Show that

$$y_t = x_t + 3y_{t-4} - 3y_{t-8} + y_{t-12},$$

for $t > 12$. ***(5 marks)***

(iii) If the first 13 observations of y_t were

t	1	2	3	4	5	6	7	8	9	10	11	12	13
y_t	10	12	15	16	9	13	15	17	8	12	14	15	9

find the one-step, two-step and three-step forecast means of y_{14}, y_{15} and y_{16} respectively. ***(10 marks)***

(iv) If y_{14}, y_{15} and y_{16} were respectively 13, 16 and 18, then calculate the forecast error in each of the forecasts in part (iii) above and briefly comment on the forecast performance, based on these 3 predictions. ***(4 marks)***

6 Consider the time series model

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + z_t \gamma_t + \eta_t,$$

where α_1, α_2 are some parameters, z_t is a known time-varying covariate, γ_t is a time-varying regression parameter and η_t is a white noise with variance 10. The modeller postulates that $\gamma_t \approx \gamma_{t-1}$ (γ_t has a slow evolution), hence she suggests using the evolution model

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

where ν_t is a white noise with variance 1. Assume that η_t and ν_t are independent for all t, each of them following a normal distribution.

(i) Define the state vector

$$\beta_t = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \gamma_t \end{bmatrix}.$$

Write the model of y_t in state-space form, i.e.

$$y_t = x_t^\top \beta_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2)$$
$$\beta_t = F\beta_{t-1} + \zeta_t, \quad \zeta_t \sim N(0, Z)$$

and determine the values of x_t, F, σ^2 and Z. ***(4 marks)***

(ii) If $z_3 = 2$, $y_1 = 5$, $y_2 = 7$, $y_3 = 4$, the posterior mean vector and the posterior covariance matrix of β_2 are

$$\hat{\beta}_{2|2} = \begin{bmatrix} 1 \\ 1/5 \\ 1 \end{bmatrix} \quad \text{and} \quad P_{2|2} = \begin{bmatrix} 20 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 20 \end{bmatrix},$$

then calculate the posterior mean vector and the posterior covariance matrix of β_3 at time $t = 3$. ***(12 marks)***

(iii) If $z_4 = 3$, find a 95% prediction interval for y_4. Based on this interval alone, comment on the forecast accuracy for the future observation y_4.
(4 marks)

End of Question Paper