Data provided: Graph Paper

MAS6011

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2017–2018**

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 70 randomly selected kernels of the Kama variety of wheat were measured using a soft X-ray technique (Charytanowicz et al, 2010). Several characteristics were measured:

x_1 a measure of compactness (got from the area and perimeter)
x_2 roundedness (the width as a percentage of length)
x_3 groove length as percentage of length

The means were given by $\overline{x} = (\overline{x}_1, \overline{x}_2, \overline{x}_3) = (88.01, 58.94, 92.35)$, with variance matrix $S = \begin{pmatrix} 2.621 & 4.107 & -0.090 \\ 4.107 & 7.901 & -0.670 \\ -0.090 & -0.670 & 5.817 \end{pmatrix}$. The data is believed to look approximately multivariate normal.

You may use the R output `qt(0.975,69)=1.994`, and the following values of the F-distribution (not all of which will be relevant):

`qf(0.95,2,67)=3.134   qf(0.95,2,68)=3.132   qf(0.95,2,69)=3.130`.

(i) Compute the correlation between x_1 and x_2. ***(1 mark)***

(ii) There is interest in comparing the Kama variety with another, ancient variety. For the ancient variety, the means for the compactness and roundedness variables x_1 and x_2 are estimated as 88.36 and 58.44 respectively.

(a) Perform a t-test to test the hypothesis that the mean compactness for the Kama variety is equal to 88.36. ***(2 marks)***

(b) Perform a t-test to test the hypothesis that the mean roundedness for the Kama variety is equal to 58.44. ***(2 marks)***

(c) Test the hypothesis that the compactness and roundedness of the Kama variety are simultaneously the same as the ancient variety. You may use the fact that $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$. ***(4 marks)***

(d) Comment on the results of (a)–(c). ***(2 marks)***

(iii) Work out the mean and variance for the difference $x_1 - x_3$ for the Kama variety. Without doing any calculations, say how you would test that x_1 and x_3 had equal means. ***(3 marks)***

(iv) The same measurements were taken for 70 samples of the Rosa variety. For this variety, the mean for x_1 is 88.35, and the mean for x_2 is 59.86; the variance matrix for these two variables is $S_2 = \begin{pmatrix} 2.403 & 3.907 \\ 3.907 & 7.903 \end{pmatrix}$. Test the hypothesis that the means of these two variables are the same for the Kama and the Rosa varieties. You may assume that $F_{2,k}(0.95) > 2.5$ for all k. ***(5 marks)***

(v) A third variety, of Canadian wheat, was tested in the same way. How would you test whether all three varieties had the same means? ***(1 mark)***

2 (i) If $X_1, \ldots, X_p$ denote the values of the original variables, and $Y_1, \ldots, Y_p$ denote the principal components, then $Y' = X'A$ where A is the matrix whose columns are the normalised eigenvectors in descending order of eigenvalue.

Show that $X = AY$, stating which properties of the matrix A you are using.

Explain why $A'\text{var}(X)A$ is diagonal, assuming any results from the course. ***(3 marks)***

(ii) The `state.x77` dataset contains data on the 50 US states with the following variables:

`Population`	population estimate as of July 1, 1975
`Income`	per capita income (1974)
`Illiteracy`	illiteracy (1970, percent of population)
`Life Exp`	life expectancy in years (1969–71)
`Murder`	murder rate per 100,000 population (1976)
`HS Grad`	percent high-school graduates (1970)
`Frost`	mean number of days below freezing (1931–1960)
`Area`	land area in square miles

A principal components analysis was carried out on the data, and an edited R transcript (using the `screeplot` function from the course) is given on the following two pages.

(a) Can you use the information to say which variable is most positively correlated with `HS Grad`? Which is the most negatively correlated? ***(2 marks)***

(b) What proportion of the variance is explained by the first two principal components? What would seem like a reasonable number of components to work with? Justify your answer. ***(4 marks)***

(c) What are the characteristics of a state which gets a high score on PC1? ***(3 marks)***

(d) Explain why Alaska (AK) and California (CA) should score so highly on PC2. ***(3 marks)***

(e) There is an outlier for PC3. What are its likely characteristics? Which of the states listed in the first R command is most likely to be the outlier? ***(3 marks)***

(f) Investigators are interested in how the `Murder` variable depends on the other variables. What would you use for this instead of PCA? Give the R command. ***(2 marks)***

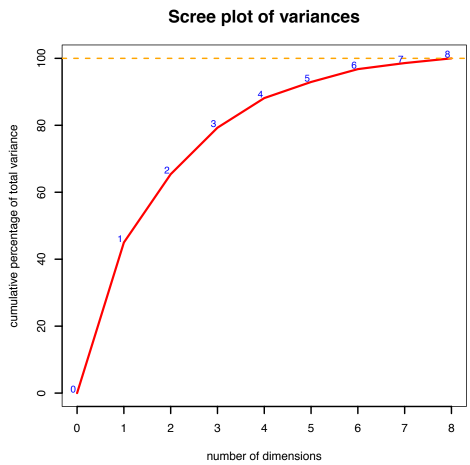
 Question 2 continued on next page

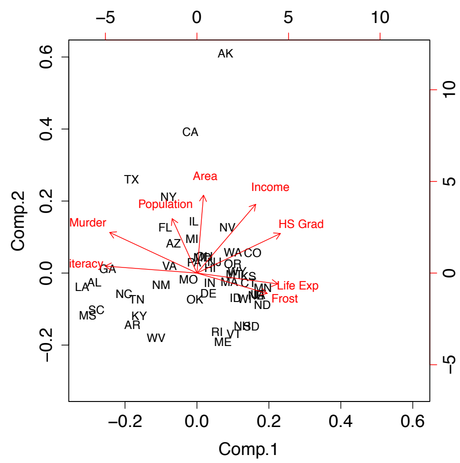
2 (continued)

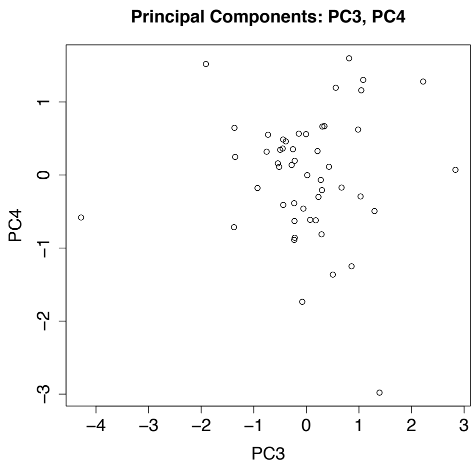
```
> state.x77[c(1:5,11),]
   Population Income Illiteracy Life Exp Murder HS Grad Frost   Area
AL       3615   3624        2.1    69.05   15.1    41.3    20  50708
AK        365   6315        1.5    69.31   11.3    66.7   152 566432
AZ       2212   4530        1.8    70.55    7.8    58.1    15 113417
AR       2110   3378        1.9    70.66   10.1    39.9    65  51945
CA      21198   5114        1.1    71.71   10.3    62.6    20 156361
> apply(state.x77,2,mean)
Population  Income  Illiteracy  Life Exp  Murder  HS Grad   Frost  Area
   4246.42  4435.80    1.17     70.88      7.38     53.11  104.46 70735.88
> apply(state.x77,2,sd)
Population Income Illiteracy Life Exp Murder HS Grad  Frost   Area
   4464.49  614.47  0.61     1.34    3.69     8.08    51.98  85327.30
> state.pca<-princomp(state.x77,cor=TRUE)
> summary(state.pca)
Importance of components:
               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
Standard       1.90   1.277  1.054  0.8411 0.6202 0.5545 0.3801 0.3364
deviation
Proportion     0.45   0.204  0.139  0.0884 0.0481 0.0384 0.0181 0.0141
of Variance
Cumulative     0.45   0.654  0.793  0.8813 0.9294 0.9678 0.9859 1.0000
Proportion
> loadings(state.pca)
Loadings:
           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
Population -0.126  0.411  0.656  0.409 -0.406                0.219
Income      0.299  0.519  0.100         0.638  0.462
Illiteracy -0.468               -0.353         0.387 -0.620  0.339
Life Exp    0.412         0.360 -0.443 -0.327  0.219 -0.256 -0.527
Murder     -0.444  0.307 -0.108  0.166  0.128 -0.325 -0.295 -0.678
HS Grad     0.425  0.299        -0.232        -0.645 -0.393  0.307
Frost       0.357 -0.154 -0.387  0.619 -0.217  0.213 -0.472
Area               0.588 -0.510 -0.201 -0.499  0.148  0.286
> screeplot(state.x77,cor=TRUE)
> state.pc<-predict(state.pca)
> biplot(state.pca)
> plot(state.pc[,3:4],xlab="PC3",ylab="PC4")
```

 Question 2 continued on next page

2 (continued)


Scree plot of variances
cumulative percentage of total variance
number of dimensions


Comp.1
Comp.2
Area
Income
Population
HS Grad
Murder
iteracy
Life Exp
Frost


Principal Components: PC3, PC4
PC3
PC4

3 (i) Suppose that we use linear discriminant analysis to classify *univariate* data between two groups. Write $p_1(x)$ for the probability that an observation with reading x is classified in group 1, and $p_2(x) = 1 - p_1(x)$ that the observation is classified into group 2. We suppose that group 1 is generated from a univariate normal distribution with mean μ_1 and variance σ^2, and that group 2 is generated from a univariate normal distribution with mean μ_2 and variance σ^2. Write

$$f_i(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_i)^2\right)$$

for the density of the group i distribution, and let

$$p_1(x) = \frac{f_1(x)}{f_1(x) + f_2(x)}, \qquad p_2(x) = \frac{f_2(x)}{f_1(x) + f_2(x)},$$

which could be interpreted as the probabilities of an observation x coming from the two groups. Show that

$$\log\left(\frac{p_1(x)}{1 - p_1(x)}\right) = c_0 + c_1 x$$

for some constants c_0 and c_1 whch you should give explicitly. ***(4 marks)***

(ii) Suppose that group 1 is generated from a bivariate normal distribution with mean $\mu_1 = (-1, \ -1)'$ and variance $\Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}$, and group 2 is generated from a bivariate normal distribution with mean $\mu_2 = (1, \ 1)'$ and variance $\Sigma_2 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix}$.

Show that the decision boundary between the two groups is in fact a pair of straight lines, and draw a picture of the plane explaining which regions would be classified as in group 1, and which as group 2. ***(5 marks)***

(iii) What is the main difference in modelling between linear and quadratic discriminant analysis? ***(2 marks)***

(iv) Lohweg et al (2012) recorded data on two measurements (x_1, x_2) (the *variance* and *skewness*) for 100 genuine banknotes and 100 forged banknotes. For the genuine banknotes, the means are $(2.297, 3.764)$, and for the forged notes, the means are $(-1.903, -1.118)$. The inverse pooled variance matrix is $S^{-1} = \begin{pmatrix} 0.28 & 0.01 \\ 0.01 & 0.04 \end{pmatrix}$.

(a) Give Fisher's linear discriminant function for this data. ***(4 marks)***

(b) Hence classify a note with measurements $(1, 0)$. ***(1 mark)***

 Question 3 continued on next page

3 (continued)

(v) Suppose that we have 4 observations, for which we compute dissimilarities as $\begin{pmatrix} 0.3 & & \\ 0.4 & 0.5 & \\ 0.7 & 0.8 & 0.45 \end{pmatrix}$.

(a) Complete this to a 4×4-matrix of distances between the observations. ***(1 mark)***

(b) On your graph paper, draw the dendrogram that results from hierarchical clustering using single linkage, plotting the heights at which clusters fuse. Do the same for clustering with complete linkage. ***(3 marks)***

4 (i) Consider that 100 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	5
PACF ($a_h^{(h)}$)	0.72	0.41	0.14	0.07	0.02
ACF (r_h)	$\star$	$\star\star$	0.70	0.62	0.44

(a) Find the values of $\star$ and $\star\star$. ***(4 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), AR(2) and AR(3). ***(2 marks)***

(d) Based on your analysis above, propose an overall model that you feel is expected to fit the data well. ***(1 mark)***

(ii) Consider the autoregressive time series model

$$y_t = \frac{1}{4}y_{t-1} + \epsilon_t - \frac{1}{2}\epsilon_{t-1} + \frac{1}{3}\epsilon_{t-2}, \qquad (1)$$

where ϵ_t is white noise with variance 1.

(a) Show that model (1) is causal and invertible. ***(3 marks)***

(b) Calculate the variance of y_t. ***(6 marks)***

5 Suppose that observations $y_1, y_2, \ldots, y_n$ are generated from the autoregressive (AR) model of order two:

$$y_t = \alpha y_{t-1} - (1-\alpha) y_{t-2} + \epsilon_t,$$

where α is a parameter and ϵ_t is a Gaussian white noise with variance σ^2.

(i) Write down the conditional likelihood function $L(\alpha, \sigma^2; y_{1:n})$ and the conditional log-likelihood function $\ell(\alpha, \sigma^2; y_{1:n})$ of the parameters α and σ^2, based on observation $y_{1:n} = \{y_1, y_2, \ldots, y_n\}$. ***(3 marks)***

(ii) Using (i) and adopting conditional least squares, show that the likelihood estimates of α and σ^2 satisfy

$$\hat{\alpha} = \frac{\sum_{t=3}^n y_t y_{t-1} + \sum_{t=3}^n y_t y_{t-2} + \sum_{t=3}^n y_{t-2}^2 + \sum_{t=3}^n y_{t-1} y_{t-2}}{\sum_{t=3}^n y_{t-1}^2 + \sum_{t=3}^n y_{t-2}^2 + 2\sum_{t=3}^n y_{t-1} y_{t-2}}$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{t=3}^n \left[y_t - \hat{\alpha} y_{t-1} + (1-\hat{\alpha}) y_{t-2} \right]^2$$

(11 marks)

(iii) (a) If $y_1 = 1$, $y_2 = 2$, $y_3 = 0$ and $y_4 = 1$, calculate the maximum likelihood estimate $\hat{\alpha}$ of α. ***(2 marks)***

(b) Based, on the observed data above, calculate the 2-step ahead forecast of the observation y_6. ***(3 marks)***

(c) If y_6 is observed to be equal to $y_6 = 0.1$, calculate the corresponding two-step ahead forecast error of the forecast in part (b) above. ***(1 mark)***

6 Consider that a time series $\{y_t\}$ is generated from an ARIMA(1,1,1) model, so that

$$y_t - y_{t-1} = \alpha(y_{t-1} - y_{t-2}) + \epsilon_t + \gamma\epsilon_{t-1},$$

where α is the AR parameter, γ is the MA parameter and $\{\epsilon_t\}$ is a Gaussian white noise sequence with variance equal to 1.

Define the state vector

$$\beta_t = \begin{bmatrix} y_t \\ y_{t-1} \\ \epsilon_t \end{bmatrix}.$$

(i) Write down a state space representation for y_t, i.e. express y_t as a state space model:

$$y_t = x^\top \beta_t + \delta_t,$$
$$\beta_t = F\beta_{t-1} + \zeta_t.$$

In your answer you should:

(a) specify the components x, F, δ_t and ζ_t; ***(2 marks)***

(b) write down the distributions of δ_t and ζ_t. ***(2 marks)***

(ii) Suppose that at time $t = 2$, the posterior distribution of β_2 is

$$\beta_2 \mid y_{1:2} \sim N\left\{ \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right\},$$

where $y_{1:2} = \{y_1 = 1, y_2 = 0\}$ denotes the information available at time $t = 2$.

(a) If $y_3 = 1$, perform a step of the Kalman filter and hence derive the posterior distribution of

$$\beta_3 \mid y_{1:3},$$

where $y_{1:3} = \{y_1 = 1, y_2 = 0, y_3 = 1\}$. ***(14 marks)***

(b) Given information $y_{1:3}$, find the posterior distribution of ϵ_3. ***(2 marks)***

End of Question Paper