



The
University
Of
Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS

**Spring Semester
2014–2015**

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 times in the running events, and for longer distances in the other events, 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 listed 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` *(1 mark)*
 - (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 . *(2 marks)*

`> attach(heptathlon)`

1 (continued)

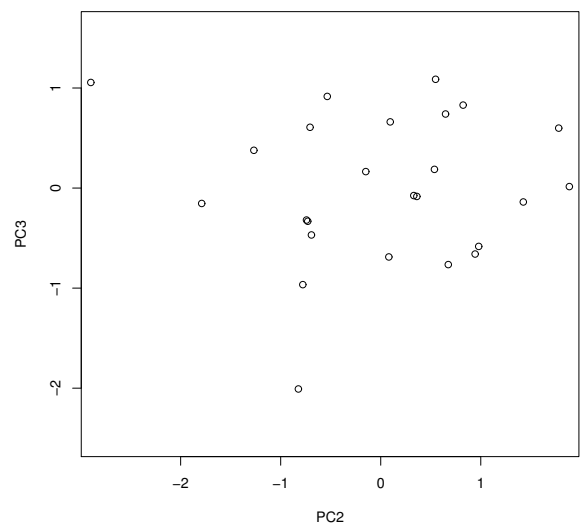
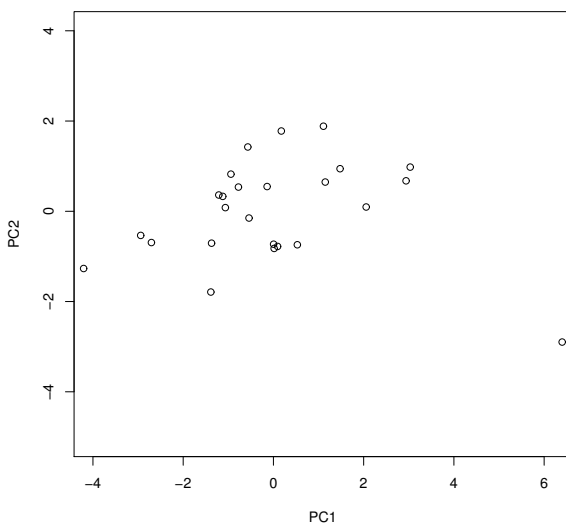
```
> heptathlon[1:3,]
      hurdles highjump shot run200m longjump javelin run800m score
Joyner-Kersee 12.69   1.86 15.80  22.56   7.27  45.66 128.51 7291
John          12.85   1.80 16.23  23.65   6.71  42.56 126.12 6897
Behmer        13.20   1.83 14.20  23.10   6.68  44.54 124.20 6858

> 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

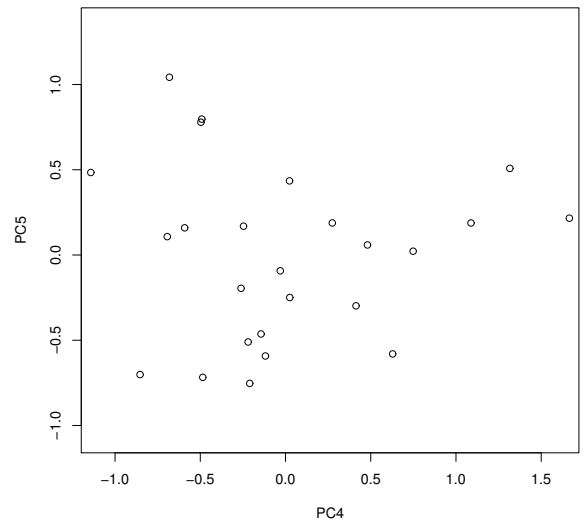
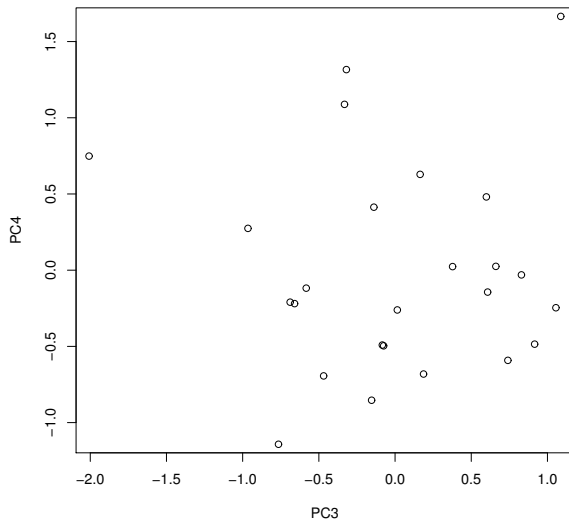
> loadings(hep.pca)
Loadings:
      Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7
hurdles  0.453 -0.158                0.783 -0.380
highjump -0.377  0.248 -0.368  0.680                -0.434
shot     -0.363 -0.289  0.676  0.124 -0.512                -0.218
run200m  0.408  0.260                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])
```



1 (continued)



2 (i) The diagram below is of a 99% confidence ellipse for a bivariate distribution around a mean μ , together with its marginal projection onto the axis representing the first variable (the x -axis):



(a) Explain why the 99% marginal confidence interval for the first variable is a subset of the marked projection of the ellipse. (2 marks)

(b) Copy the diagram, and sketch the simultaneous 99% confidence region for the simultaneous univariate tests for the two variables. (An accurate sketch is not expected, merely an approximate shape.) (2 marks)

(c) Mark on your sketch **either** a point which is significant (at the 1% level) for the multivariate test, but not for *either* univariate test, **or** a point which is significant for *both* univariate tests, but not for the multivariate test, saying which you have plotted. (1 mark)

2 (continued)

(ii) Measurements are taken on 6 approximately two-year old boys of height, chest circumference and mid-upper-arm circumference, all measured in centimetres (Height, Circ and MUAC in the R session below) in a highland area of a certain country.

```
> apply(boys,2,mean)
Height  Circ  MUAC
  82.0   60.2  14.5
> S<-var(boys)
> S
      Height  Circ MUAC
Height  31.6   8.0  0.5
Circ     8.0   3.2  1.3
MUAC     0.5   1.3  1.9
> solve(S)
      Height  Circ  MUAC
Height  0.2   -0.6  0.4
Circ   -0.6    2.6 -1.6
MUAC   0.4   -1.6  1.5
```

(a) The corresponding means for children in lowland areas are considered to be 88cm, 58.4cm and 16cm respectively. Compute the statistic which you should use to test the hypothesis that the means for the highland children are the same, stating the distribution of the test statistic under this hypothesis. *(5 marks)*

(b) Given the R output

```
qf(.99,2,2)=99,      qf(.99,2,3)=30.82,   qf(.99,2,4)=18
qf(.99,3,2)=99.17,  qf(.99,3,3)=29.46,   qf(.99,3,4)=16.69
qf(.99,4,2)=99.25,  qf(.99,4,3)=28.71,   qf(.99,4,4)=15.98,
```

explain that we can reject the null hypothesis at a 1% level of significance. *(1 mark)*

(c) By computing an appropriate expression involving the F -statistic, and taking projections of the multivariate confidence ellipsoid onto each variable, compute the simultaneous 99% confidence intervals for the actual mean of each variable around the observed mean.

Verify that these confidence intervals contain $\mu = (88.0, 58.4, 16.0)'$. *(5 marks)*

(d) Given that $t_5(0.005) = 4.032$, give the 99% univariate confidence intervals for each variable mean individually.

Verify that these confidence intervals contain μ . *(4 marks)*

3 Thomson and Randall-Maciver (1905) gave a data set of 150 Egyptian skulls from predynastic, 12th-13th dynasties and postdynastic periods (coded in the analyses below as periods 1, 2 and 3 respectively). The measurements available (all converted to cm) are maximum breadth (**mb**), basibregmatic height (**bh**), basialveolar length (**bl**) and nasal height (**nh**). Skulls 1–60 come from period 1, skulls 61–90 from period 2, and skulls 91–150 from period 3.

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

(i) The first question is whether the measurements change over time; nonconstant measurements would indicate interbreeding with immigrant populations.

How would you test the hypothesis that the three groups of skulls are the same sizes? Discuss the R command you would use, and the number of degrees of freedom involved. *(3 marks)*

(ii) Justifying your answer, what linear combination of the four measurements shows the greatest differences between the three periods? *(3 marks)*

(iii) From which period are skulls most difficult to classify correctly with this technique? *(1 mark)*

(iv) What proportion of the skulls used in the analysis does the linear discriminant analysis classify correctly? *(1 mark)*

(v) It is proposed to try to understand whether or not the skulls from 12th-13th dynasties were more like the predynastic or postdynastic ones. With this in mind, it is proposed to repeat the Linear Discriminant Analysis simply with the skulls from periods 1 and 3, and then using Fisher’s linear discriminant function to decide whether each skull from period 2 is closer to period 1 or to period 3.

(a) Compute Fisher’s linear discriminant function for the LDA on the skulls from periods 1 and 3, and hence give the condition that a new observation is classified as period 1. *(6 marks)*

(b) To assess the reliability of this method, the investigators wonder how it would work for a skull from period 1 or 3. Give a command that you would use in R to calculate the probability of misclassification.

What assumptions does this calculation make, and does this seem reasonable? (You are not expected to give any analysis.) *(4 marks)*

(c) If a skull from period 3 were to be misclassified as from period 1, what features might be the most likely cause? *(2 marks)*

*** Summary Statistics for data in: skulls ***

```

period:1
      mb      bh      bl      nh
Mean: 13.1867 13.3150  9.9117  5.0383
Std Dev.:  0.4956  0.4543  0.5129  0.2841
-----
period:2
      mb      bh      bl      nh
Mean: 13.4467 13.3800  9.6033  5.0567
Std Dev.:  0.3481  0.4979  0.4552  0.3549
-----

```

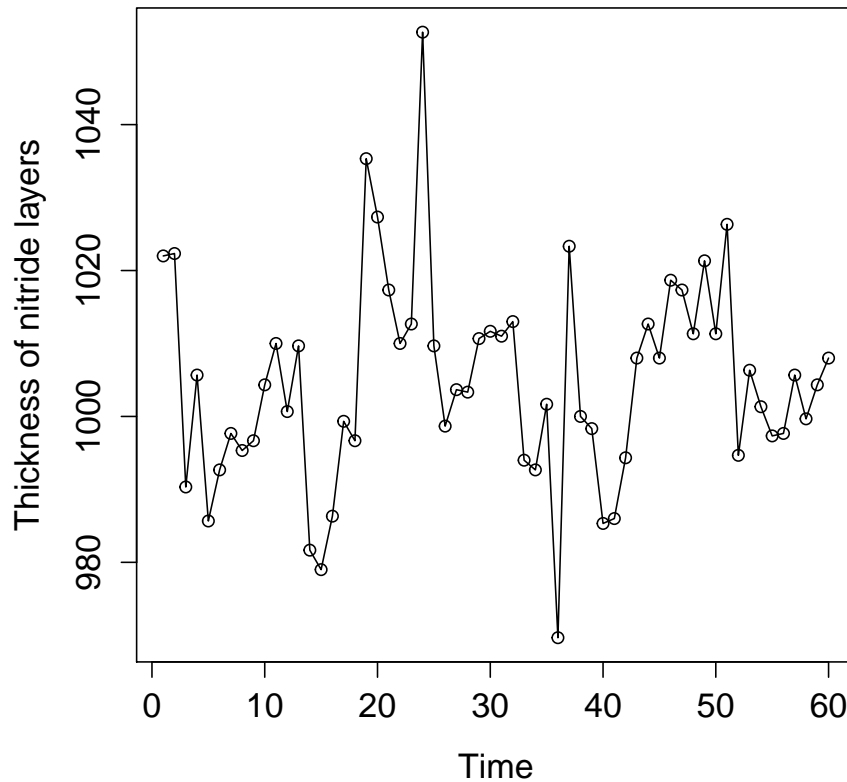

3 (continued)

```

period:3
      mb      bh      bl      nh
Mean: 13.5833 13.1317  9.4017  5.1667
Std Dev.:  0.4662 0.5107 0.4817 0.3287
>
> skull.lda<-lda(period~mb+bh+bl+nh,prior=c(1,1,1)/3)
> skull.lda
Coefficients of linear discriminants:
      LD1      LD2
mb -0.1256624 -0.0615404
bh  0.0277626 -0.1913655
bl  0.1485194  0.0592604
nh -0.0890593  0.1761597

Proportion of trace:
      LD1  LD2
0.882 0.118
>
> skull.pred<-predict(skull.lda,data.frame(cbind(mb,bh,bl,nh)))
>
> table(skull.pred$class,period)
  period
    1  2  3
1  36  7 11
2  12 16 14
3  12  7 35
>
> skulls1<-skulls[1:60,]
> skulls3<-skulls[91:150,]
> m1<-apply(skulls1,2,mean)
> m3<-apply(skulls3,2,mean)
> pooled<-(var(skulls1)+var(skulls3))/2
> pooled
      mb      bh      bl      nh
mb  0.23146328 -0.00429096 0.00187288 0.0215876
bh -0.00429096  0.23361299 0.05776554 0.0370240
bl  0.00187288  0.05776554 0.24755650 0.0106483
nh  0.02158757  0.03702401 0.01064831 0.0943658
> sinv<-solve(pooled)
> a<-t(m1-m3)%*%sinv
> a
      mb      bh      bl      nh
[1,] -1.5894 0.4797 2.0210 -1.4126
> a%*%m1
      [,1]
[1,] -1.6578
> a%*%m3
      [,1]
[1,] -3.5883

```



4 (i) The plot above shows time series data consisting of 60 observations of the thickness of nitride layers (unknown units); the data was part of a larger experiment on the manufacturing of a microelectronic device ¹.

(a) Describe the data by commenting on their structure, their variation and dynamics. *(2 marks)*

(b) Based on your answer in (a) or otherwise, suggest suitable time series model(s) that may be appropriate for this data. *(2 marks)*

¹Source: Triantafyllopoulos, K., Godolphin, J.D. and Godolphin, E.J., 2005, Process improvement in the microelectronic industry by state space modelling, Quality and Reliability Engineering International, 21, 465-475

4 (continued)

(ii) A model is to be fitted to a time series of length 100. Values of the sample autocorrelation function (ACF) and sample partial ACF (PACF) are tabulated below.

Lag (h)	1	2	3	4	5
ACF (r_h)	0.6	0.4	0.1	0.05	0.01
PACF ($\hat{a}_h^{(h)}$)	*	**	0.02	0.01	-0.02

- (a) Find the omitted values (* and **). *(4 marks)*
- (b) Check whether the time series is stationary. *(1 mark)*
- (c) Test whether the time series is consistent with white noise. *(2 marks)*
- (d) Test whether the time series is consistent with moving average models. *(4 marks)*
- (e) Test whether the time series is consistent with autoregressive models. *(3 marks)*
- (f) Based on your answer in (c)-(e) above, suggest a time series model that may be suitable to model the data. *(2 marks)*

5 (i) In the context of maximum likelihood estimation of ARMA models describe briefly what is meant by conditional least squares estimation. *(2 marks)*

(ii) Consider that y_t is generated by an autoregressive model of order 2 (AR(2))

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \epsilon_t,$$

where α_1, α_2 are the AR coefficients and ϵ_t is white noise with variance σ^2 .

(a) Write down the conditional likelihood and the conditional log-likelihood functions of the parameters α_1, α_2 and σ^2 , based on a collection of observations $y_{1:n} = (y_1, y_2, \dots, y_n)$. *(4 marks)*

(b) Using conditional least squares, show that the maximum likelihood estimates of α_1, α_2 and σ^2 are

$$\begin{aligned} \hat{\alpha}_1 &= \frac{\sum_{t=3}^n y_{t-2}^2 \sum_{t=3}^n y_t y_{t-1} - \sum_{t=3}^n y_{t-1} y_{t-2} \sum_{t=3}^n y_t y_{t-2}}{\sum_{t=3}^n y_{t-1}^2 \sum_{t=3}^n y_{t-2}^2 - (\sum_{t=3}^n y_{t-1} y_{t-2})^2} \\ \hat{\alpha}_2 &= \frac{\sum_{t=3}^n y_{t-1}^2 \sum_{t=3}^n y_t y_{t-2} - \sum_{t=3}^n y_t y_{t-1} \sum_{t=3}^n y_{t-1} y_{t-2}}{\sum_{t=3}^n y_{t-1}^2 \sum_{t=3}^n y_{t-2}^2 - (\sum_{t=3}^n y_{t-1} y_{t-2})^2} \\ \hat{\sigma}^2 &= \frac{1}{n-2} \sum_{t=3}^n (y_t - \hat{\alpha}_1 y_{t-1} - \hat{\alpha}_2 y_{t-2})^2. \end{aligned}$$

(14 marks)

6 A company trades 10 products, with the i th product projected to give a return r_{it} at time t , for $i = 1, 2, \dots, 10$. The company believes that each of these returns r_{it} follows an autoregressive process

$$r_{it} = 0.9r_{i,t-1} + \zeta_{it},$$

where ζ_{it} is a white noise with variance 1, $\zeta_{it} \sim N(0, 1)$, and ζ_{it} is independent of ζ_{jt} , for any $i \neq j$.

Due to a data recording error r_{it} is not available. However, the aggregate return can be observed subject to additive noise, according to the model

$$y_t = \sum_{i=1}^{10} r_{it} + \epsilon_t,$$

where ϵ_t is a white noise with variance 1, $\epsilon_t \sim N(0, 1)$, and it is assumed that ϵ_t is independent of ζ_{it} , for any t and for any i .

(i) Define the state

$$\beta_t = \sum_{i=1}^{10} r_{it}.$$

Show that y_t follows a state space model

$$\begin{aligned} y_t &= x\beta_t + \epsilon_t \\ \beta_t &= F\beta_{t-1} + \zeta_t \end{aligned}$$

and determine x , F , ζ_t and the variance of ζ_t . **(4 marks)**

(ii) A prior distribution for β_0 is set as

$$\beta_0 \sim N(0, 100).$$

If the first observation is $y_1 = 2$, perform the Kalman filter iteration for $t = 1$ and obtain the posterior distribution of

$$\beta_1 \mid \{y_1 = 2\}. \quad \textbf{(8 marks)}$$

(iii) Using the result in (ii) obtain a 95% predictive interval for y_2 . **(4 marks)**

(iv) Describe briefly what is the likely effect on the posterior distribution of β_t (for large t), if the prior distribution of β_0 changes from (a) $\beta_0 \sim N(0, 1)$ to (b) $\beta_0 \sim N(0, 1000)$. **(4 marks)**

End of Question Paper