UNIVERSITY OF DUBLIN

XST30071

TRINITY COLLEGE

FACULTY OF ENGINEERING, MATHEMATICS AND SCIENCE

School of Computer Science and Statistics

JS MSISS & JS-SS Mathematics **Trinity Term 2011**

Multivariate Linear Analysis and Applied Forecasting (ST3007)

Wednesday, May 25 2011 **GMB** **9.30 — 12.30**

Dr. Rozenn Dahyot & Dr. Brett Houlding

Attempt two questions out of three in each section A and B

All questions carry equal marks

Non-programmable calculators are permitted for this examination—please indicate the make and model of your calculator on each answer book used.

You may not start this examination until you are instructed to do so by the Invigilator.

Section A - Multivariate Linear Analysis

1. The percentage of the population in lifelong learning and the education spending of 20 European Countries were recorded.

Variable	Description
Country	The European Country the data relates to.
PLL	Percentage of population in lifelong learning.
GDP	The education spending of the Country as a percentage of its Gross Domestic Product.

A hierarchical cluster analysis was completed on a standardized version of the data using Euclidean dissimilarity and complete linkage. A dendrogram of the result is provided in Appendix A at the end of this question.

a) Choices have been made concerning the decision to standardize the data, the dissimilarity measure, and the linkage function. Explain what each of these do, the importance of the choice, and alternatives that could have been considered.

[10 marks]

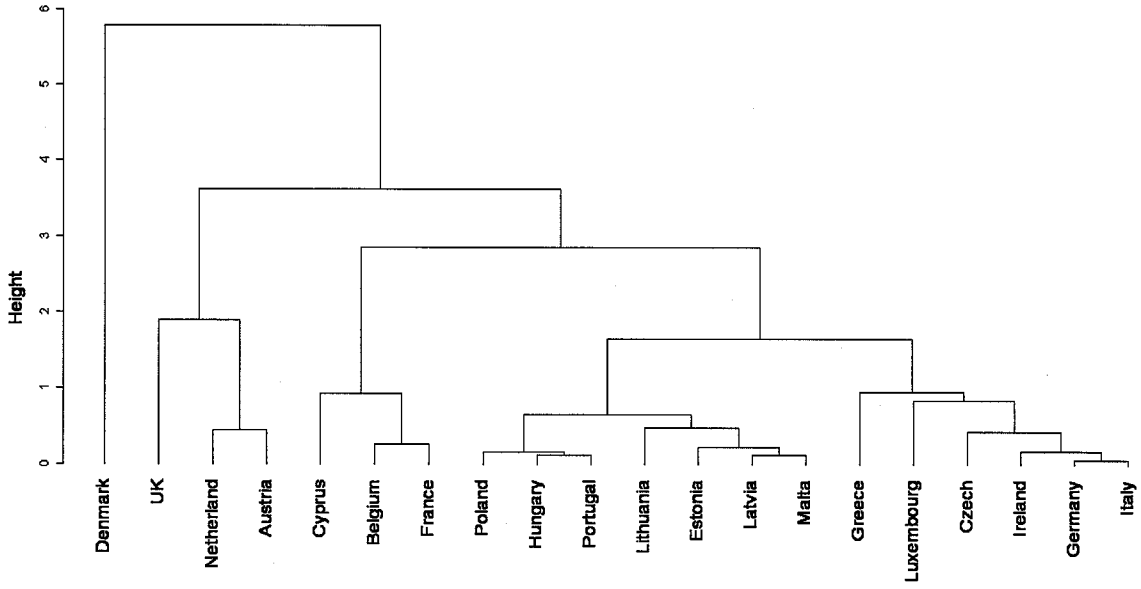
b) Making explicit reference to the dendrogram of Appendix A, explain how an analyst could use this to determine if there exist groups of similar countries. If such groups do exist, how can an analyst decide on the number of groups present? Provide, with explanation, your own suggestion as to the number of clusters and the countries included in them.

[7 marks]

c) Describe how the *k*-Means clustering algorithm could have also been used to cluster the data, and contrast the main differences with the hierarchical analysis of Appendix A. You should provide a description of the k-Means algorithm.

[8 marks]

Appendix A


Height
0
1
2
3
4
5
6
Denmark
UK
Netherland
Austria
Cyprus
Belgium
France
Poland
Hungary
Portugal
Lithuania
Estonia
Latvia
Malta
Greece
Luxembourg
Czech
Ireland
Germany
Italy

2. Data were recorded on body measurements for 100 crabs (50 of each sex).

Variable	Description
sex	"M" for Male and "F" for Female.
FL	Frontal lobe size (mm).
RW	Rear width (mm).
CL	Carapace length (mm).
CW	Carapace width (mm).
BD	Body depth (mm).

Of interest is the determination of whether a new crab is Male or Female based on the recorded body measurements. The output from a discriminant analysis is given in Appendix B at the end of this question.

a) Describe the modeling assumptions underlying linear and quadratic discriminant analysis and how these are used to classify a new observation. In your description, explain why one method is called 'linear' disciminant analysis and the other is called 'quadratic' discriminant analysis.

[8 marks]

b) Explain what is meant by cross-validation in the context of discriminant analysis.

[6 marks]

c) Using the linear discriminant output, classify the sex of a crab whose measurements are (FL=13, RW=12, CL=29, CW=33, BD=12).

[5 marks]

d) Describe two alternative classification techniques that could have been applied in this situation, and contrast their similarities and differences with linear and quadratic discriminant analysis.

[6 marks]

Appendix B

Call:
lda(Crab[, -1], grouping = Crab[, 1])

Prior probabilities of groups:

F	M
0.5	0.5

Group means:

	FL	RW	CL	CW	BD
F	13.3	12.1	28.1	32.6	11.8
M	14.8	11.7	32.0	36.8	13.4

Coefficients of linear discriminants:

	LD1
FL	-0.90
RW	-1.67
CL	0.91
CW	0.13
BD	-0.18

> colMeans(Crab[,-1])

FL	RW	CL	CW	BD
14.1	11.9	30.1	34.7	12.6

Note: Level orders are: F, M

3. Poverty and Inequality statistics were recorded for 91 different countries.

Variable	Description
BR	Live birth rate per 1,000 population.
DR	Death rate per 1,000 population.
ID	Infant deaths per 1,000 population under 1 year old.
LEM	Life expectancy at birth for males.
LEF	Life expectancy at birth for females.
GNP	Gross National Product per capita in U.S. dollars.

The covariance matrix for this data, and a principal component analysis applied on the correlation matrix, are provided in Appendix C at the end of this question.

a) Explain why it is not appropriate to perform the analysis on the covariance matrix.
[2 marks]

b) Calculate the correlation between the variables DR and ID.
[3 marks]

c) With reference to the output in Appendix C, explain whether a reduced set of variables will capture the variability in the data. Interpret the output of the analysis in Appendix C and provide an intuitive meaning behind any variable in the lower dimensional representation suggested.
[10 marks]

d) Describe two other methods of data reduction that could be used on this data. Clearly explain any choices that need to be made when applying these data reduction methods and how they differ from Principal components.
[10 marks]

Appendix C

> cov(poverty)

	BR	DR	ID	LEM	LEF	GNP
BR	187.7	32.4	543.3	-115.4	-136.4	-69747.8
DR	32.4	21.9	147.0	-34.4	-37.3	-11477.2
ID	543.3	147.0	2143.9	-421.2	-491.8	-225470.9
LEM	-115.4	-34.4	-421.2	94.6	106.3	50622.3
LEF	-136.4	-37.3	-491.8	106.3	123.9	58561.7
GNP	-69747	-11477.2	-225470.9	50622.3	58561.7	65507653.6

> prcomp(poverty,scale=TRUE)

Standard deviations:

[1] 2.08 0.99 0.72 0.34 0.24 0.12

Rotation:

	PC1	PC2	PC3	PC4	PC5	PC6
[1,]	0.43	-0.03	0.49	-0.71	0.24	0.07
[2,]	0.37	0.14	-0.84	-0.30	0.19	-0.03
[3,]	0.46	0.06	0.14	0.60	0.62	0.15
[4,]	-0.47	-0.03	0.00	-0.13	0.63	-0.60
[5,]	-0.48	-0.01	-0.10	-0.17	0.35	0.78
[6,]	0.08	-0.99	-0.12	0.02	0.03	0.01

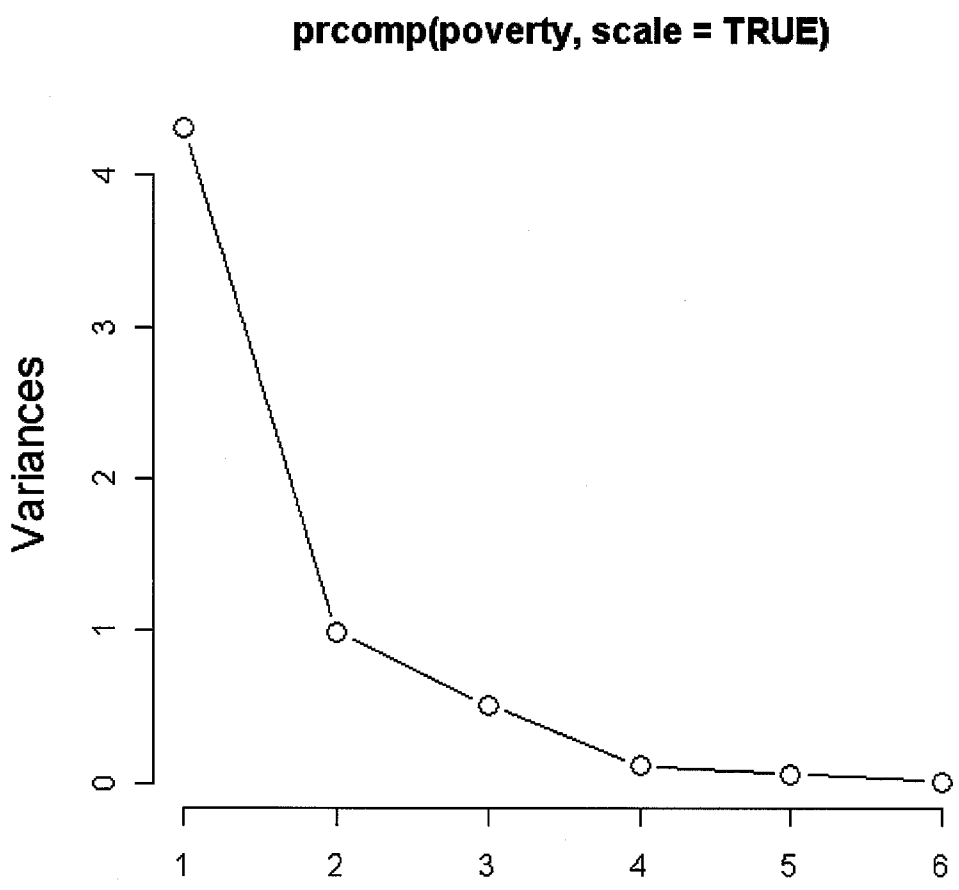
> summary(prcomp(poverty,scale=TRUE))

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.08	0.99	0.72	0.34	0.24	0.12
Proportion of Variance	0.718	0.165	0.086	0.020	0.010	0.002
Cumulative Proportion	0.718	0.882	0.969	0.988	0.998	1.00

APPENDIX C CONTINUED ON NEXT PAGE

Scree Plot:



Section B - Applied Forecasting

4. **Definitions.**

Write short notes (approx. 150 to 200 words) on FIVE of the following topics (5 marks each):

(a) Stationarity in time series.

(b) Autoregression versus Linear regression with indicator variables in seasonal time series models.

(c) What are the components of a time series?

(d) MAPE and RMSE

(e) AIC and BIC.

(f) Holt-Winters algorithms.

(g) Identifying seasonality in time series.

(25 marks)

5. (a) Consider the AR(1) model: $y_t = \phi_1 \, y_{t-1} + \phi_0 + \epsilon_t$

   (i) What is the acronym AR standing for?

   [2 marks]

   (ii) Explain the link between Linear regression and an AR(1) model.

   [2 marks]

   (iii) What are the assumptions made on ϵ_t in both linear regression and the AR(1) model?

   [2 marks]

   (iv) Indicate an algorithm which may be used to estimate the coefficients ϕ_0 and ϕ_1.

   [2 marks]

   (b) The Moving Average model of order 1 (or MA(1)) for time series is defined as:

$$y_t = \psi_1 \, \epsilon_{t-1} + \psi_0 + \epsilon_t \tag{1}$$

   (i) Explain why this model is called 'Moving Average'?

   [3 marks]

   (ii) Can we use the same algorithm to estimate the parameters (ψ_0, ψ_1) as the one we use for linear regression and AR(1)? Explain.

   [3 marks]

   (iii) Show that when $\psi_0 = 0$, with $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, $\forall t$ (i.e. the errors are independent and normally distributed with mean 0 and variance σ^2), then:

$$\begin{cases} \mathbb{E}[y_t] = 0 \\ \mathbb{V}\text{ar}[y_t] = (1 + \psi_1^2) \, \sigma^2 \end{cases} \tag{2}$$

   where $\mathbb{E}[y_t]$ is the expectation of y_t and $\mathbb{V}\text{ar}[y_t]$ is the variance of y_t.

   [3 marks]

   (c) The ACF and PACF plots are both used to analyse time series. Figures 1 and 2 shows typical ACF and PACF plots simulated for two time series named A and B.

(i) Define the ACF and PACF.

[4 marks]

(ii) Using Figures 1 and 2, identify the time series A and B as being AR(1) or MA(1). Comment on the sign of ϕ_1 or ψ_1 in both cases.

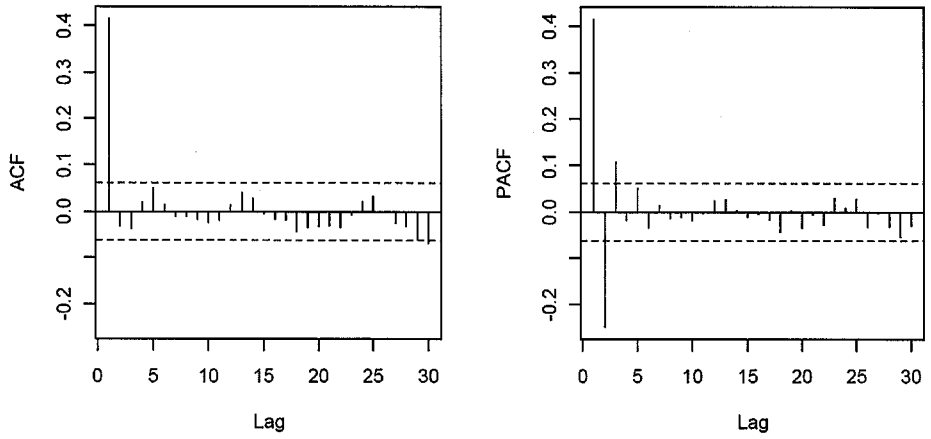


Figure 1: ACF and PACF of time series A.

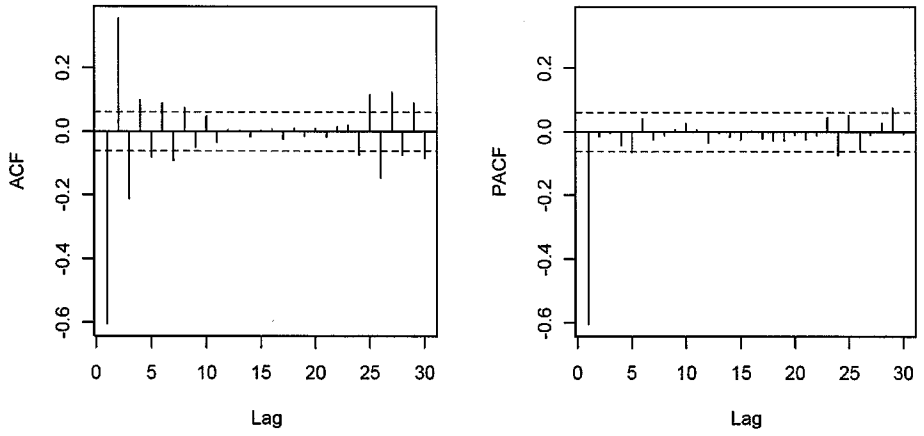


Figure 2: ACF and PACF of time series B.

[4 marks]

(25 marks)

6. (a) Consider the model $ARIMA(1,1,2)(2,1,1)_4$ for a time series X_t.

   (i) Indicate what is the meaning of each number in '$ARIMA(1,1,2)(2,1,1)_4$' and their association to corresponding time series models.

   [4 marks]

   (ii) Write this model with the backshift operator.

   [3 marks]

   (iii) Redefine this model algebraically (without the backshift operator).

   [3 marks]

   (b) Consider an AR(1) model applied to a time series X_t for which we have collected the values upto time T.

   (i) What is the prediction one step ahead, $\hat{X}_{T+1}$?

   [2 marks]

   (ii) What is the 95% prediction interval of this prediction one step ahead?

   [2 marks]

   (iii) What is the prediction k step ahead, $\hat{X}_{T+k}$?

   [3 marks]

   (iv) What is the 95% prediction interval of this prediction k step ahead?

   [2 marks]

   (c) Holt-Winters algorithms.

   (i) Define the Single Exponential Smoothing (SES) algorithm.

   [3 marks]

   (ii) What are the criteria for selecting the best the SES algorithm?

   [3 marks]

(25 marks)