UNIVERSITY OF DUBLIN

ST3007-2

TRINITY COLLEGE

FACULTY OF ENGINEERING, MATHEMATICS AND SCIENCE

School of Computer Science and Statistics

JS MSISS, MATHS, TSM **Trinity Term 2013**

Applied Forecasting & Multivariate Analysis (ST3007)

Tuesday, April 30, 2013 **GOLDHALL** **09:30—12:30**

Dr. Rozenn Dahyot, Dr. Brett Houlding

Do 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 a) Explain the benefits and problems of lower dimensional representation for Multivariate data.

[3 marks]

b) Briefly describe the objective of Multidimensional Scaling and the difference between its Metric and Non-Metric versions.

[3 marks]

c) Explain the meaning and role of the 'Stress' value for a Multidimensional Scaling analysis and explain how this value is found.

[4 marks]

d) Given a matrix $\boldsymbol{D}$ detailing the dissimilarities between any two multivariate observations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, explain how Classical Multidimensional Scaling can be used to obtain a lower-dimensional co-ordinate array $\boldsymbol{X}$.

[6 marks]

e) Making reference to the Procrustes Sum of Squares, explain the role and application of a Procrustes Analysis within the context of Multidimensional Scaling.

[4 marks]

f) Contrast the similarities and differences between the approaches of Classical Multidimensional Scaling and Principle Components.

[5 marks]

2. a) Explain the differences between hierarchical and iterative clustering methods.

[4 marks]

b) For both of the following, explain whether or not there will be an effect on the results of a hierarchical clustering algorithm, and if so, why this is the case:

i) Scaling the data so that each variable has equal standard deviation.
ii) Subtracting the mean vector of the data from each observation.

[3 marks]

c) The following is an extract of scaled demographic information concerning birth rates, death rates, life expectancy, inflation and Gross Domestic Product for four countries:

	Birth	Death	Life	Infla	GDP
Ireland	0.55	0.63	1.16	0.33	2.03
UK	0.43	0.96	1.13	0.11	1.89
USA	0.57	0.78	1.12	0.20	2.56
China	0.53	0.65	1.05	0.34	0.36

Using Maximum dissimilarity generate a dissimilarity matrix for this extract.

[5 marks]

d) Using your answer to part c), and complete linkage, produce a sketch of the resulting dendrogram for this extract (you should show your working for the values of heights at which clusters are merged).

[8 marks]

e) For your sketch dendrogram, suggest an appropriate 'cut-value', and hence confirm the clustering solution.

[5 marks]

3. a) Contrast clustering and classification algorithms.

[3 marks]

b) Describe the modeling assumptions underpinning Linear Discriminant Analysis and Quadratic Discriminant Analysis.

[4 marks]

c) Show that, in the case of Linear Discriminant Analysis with equal prior probability of group assignment, the probability that a point $\mathbf{x}$ belongs to group 1 is larger than the probability of belonging to another group 2 if:

$$(\mathbf{x} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2)^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0.$$

[6 marks]

d) Group 1 has mean vector $\boldsymbol{\mu}_1^T=(1,1)$, whilst group 2 has mean vector $\boldsymbol{\mu}_2^T=(1,0)$, with pooled precision matrix $\Sigma^{-1} = \begin{pmatrix} 0.6 & -0.2 \\ -0.2 & 0.4 \end{pmatrix}$. To which group would Linear Discriminant Analysis assign the point $\mathbf{x}^T=(0.9,0.6)$?

[4 marks]

e) Describe the alternative classification technique of Logistic Regression and contrast with Linear Discriminant Analysis.

[4 marks]

f) Describe the alternative classification technique of *k*-Nearest Neighbours and contrast with Linear Discriminant Analysis.

[4 marks]

Section B - Applied Forecasting

4. (a) The time series eggs (collecting the annual price of dozen eggs in US, between 1900 to 1993, in constant dollars) is analysed using R and the following lines are entered in the R console:

```
> require(fma)
> tsdisplay(eggs)
> HoltWinters(eggs,beta=FALSE,gamma=FALSE)$SSE
[1] 66362.79
> HoltWinters(eggs,gamma=FALSE)$SSE
[1] 79209.45
```

Explain each command line.

[4 marks]

(b) Using the R outputs (above), which algorithm is the best suited to fit the eggs time series? Explain the criterion used, and discuss other criteria that could have also been considered to select the best Holt-Winters algorithm.

[5 marks]

(c) Give the definition of the SES algorithm.

[4 marks]

(d) Would seasonal Holt-Winters algorithms be suitable to analyse this time series eggs? Justify your answer using figure 1.

[2 marks]

(e) Using the R function `auto.arima` applied to the time series eggs, the best selected model is ARIMA(0,1,0). Explain how the function `auto.arima` selects the best ARIMA model.

[2 marks]

(f) Using figure 1, explain why the model ARIMA(0,1,0) may be suitable to fit the time series eggs.

[2 marks]

(g) Write the equation associated with the model ARIMA(0,1,0).

[2 marks]

(h) What are the criteria that can be used to select the best ARIMA models? Explain your answer.

[4 marks]

(25 marks)

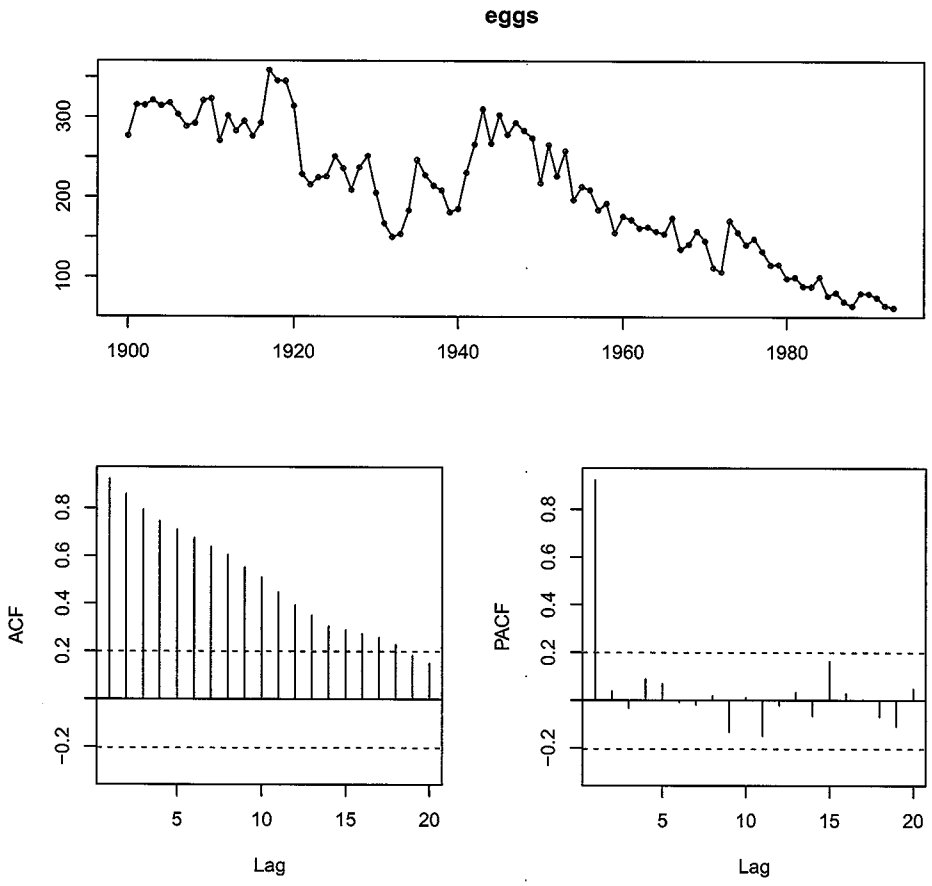

Figure 1: eggs time series.

5. The time series of interest is the quarterly earnings (dollars) per Johnson & Johnson share from 1960 to 1980 (cf. figure 2).

   (a) What are the components or patterns that can be found in a time series? Comment on the components of the Johnson & Johnson time series (cf. figure 2).

   [3 marks]

   (b) The auto.arima function identified the ARIMA$(2,1,2)(1,0,1)_4$ model as the best model. Write the equation of this model with the backshift operator.

   [4 marks]

   (c) What are the statistical assumptions about the errors that are made when one fits a seasonal arima model?

   [4 marks]

   (d) The time plot, ACF and PACF of the residuals after fitting the model found by auto.arima are shown in figure 3. Is it a good model? Explain your answer.

   [3 marks]

   (e) The expert decides then to analyse the logarithm of the Johnson & Johnson time series. Figure 4 shows the transformed time series and the residuals after using `auto.arima` are shown in figure 5.

      (i) Why are time series sometimes transformed before analysis? Explain.

      [4 marks]

      (ii) Using figures 4 and 5, is the choice of the log transformation justified for the Johnson & Johnson time series? Explain.

      [4 marks]

      (iii) Is there any criterion that can be used to select the best model between the ones fitted in figures 3 and 5? Explain.

      [3 marks]

   (25 marks)

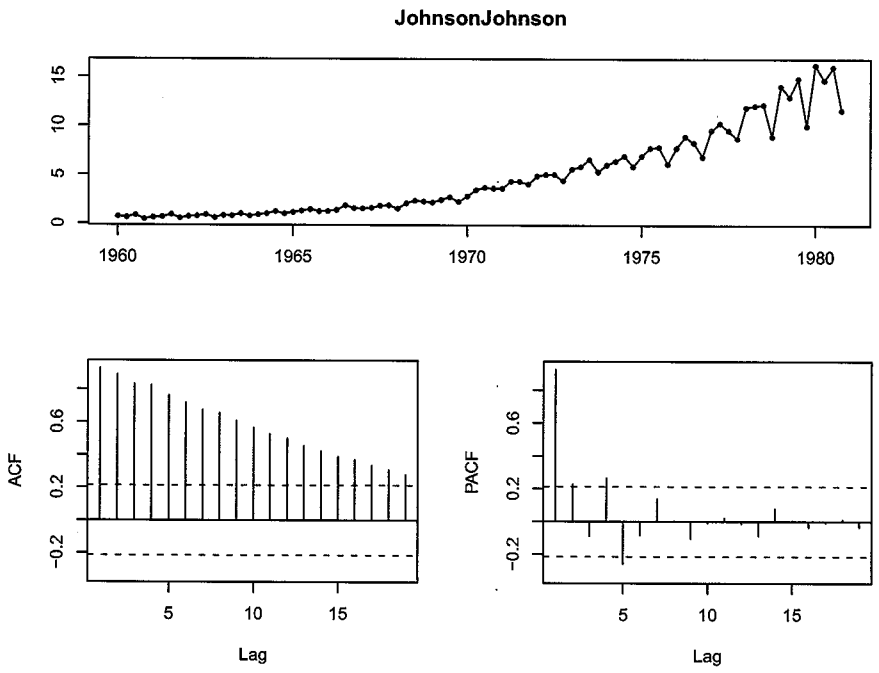

Figure 2: Johnson & Johnson time series.

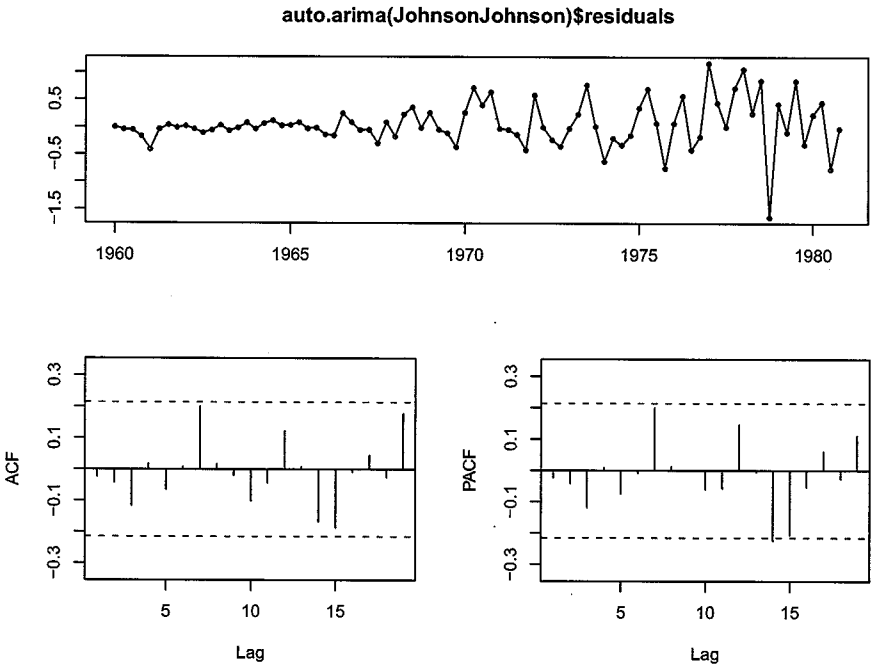

Figure 3: Residuals after using auto.arima on the Johnson & Johnson time series.

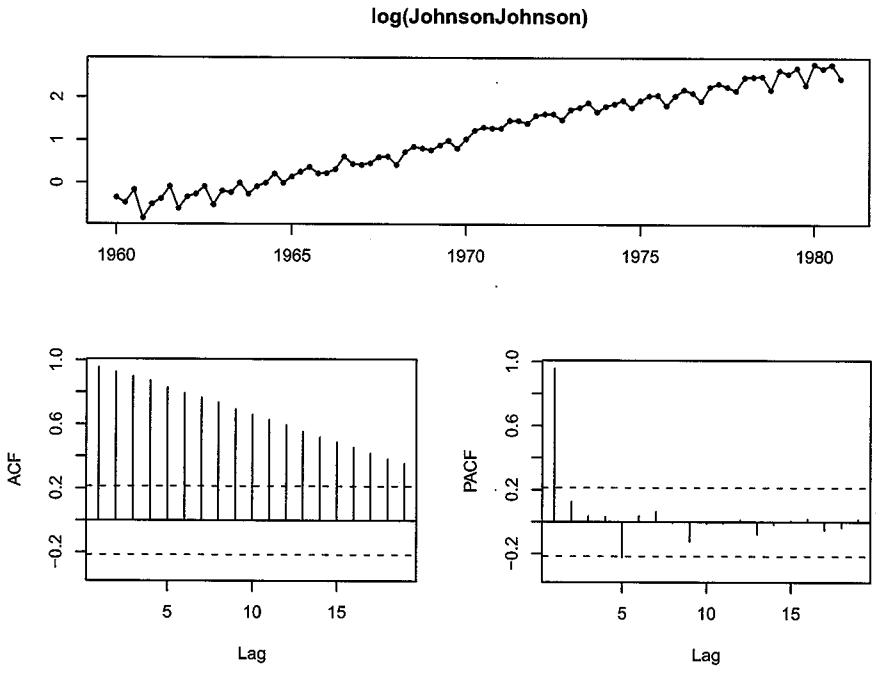

Figure 4: Logarithm of Johnson & Johnson time series.

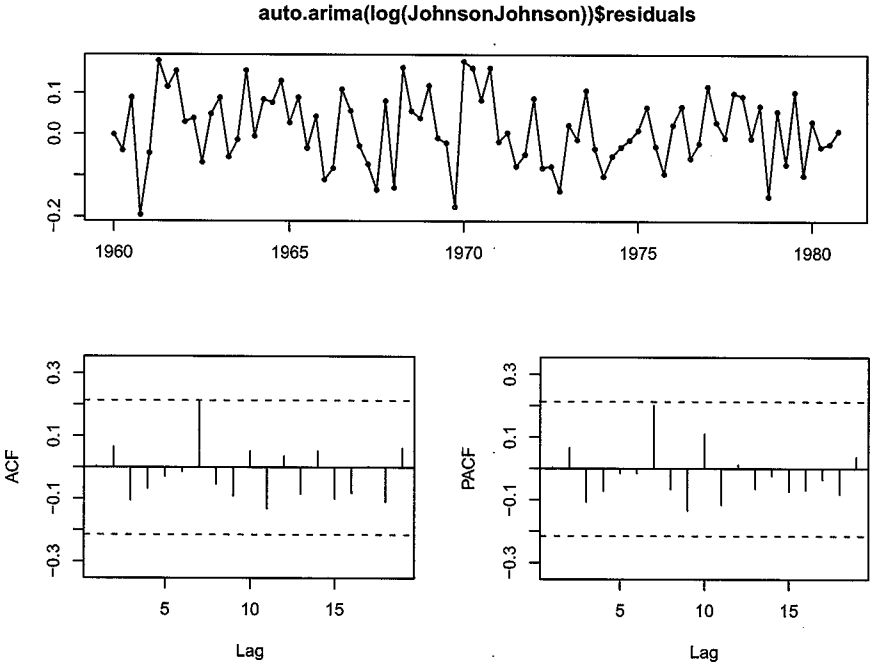

Figure 5: Residuals after using auto.arima on the logarithm of Johnson & Johnson time series.

6. Consider a time series $\{y_1, y_2, \cdots, y_n\}$, for which observations are collected up to time n, that follows an AR(1) model.

   (a) Explain the meaning of the acronym AR.

   [1 marks]

   (b) Define mathematically the AR(1) model.

   [2 marks]

   (c) What are the assumptions made for using an AR(1) model?

   [2 marks]

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

   [3 marks]

   (e) Indicate an algorithm which may be used to estimate the coefficients of the AR(1) model. Assume these coefficients to be known for the following questions.

   [2 marks]

   (f) What is the prediction and its 95% prediction interval at time $n+1$?

   [3 marks]

   (g) What is the prediction and its 95% prediction interval at time $n+2$?

   [3 marks]

   (h) Comment on the evolution of prediction intervals computed at $n+k$ (k step ahead).

   [4 marks]

   (i) Explain how seasonal ARIMA models are defined to tackle trends and seasonality.

   [5 marks]

   (25 marks)