UNIVERSITY OF DUBLIN
TRINITY COLLEGE

Faculty of Engineering, Mathematics and Science

School of Computer Science & Statistics

BA (Mod) JS MSISS

JS-SS MATHS & TSM **Trinity Term 2014**

Multivariate Linear Analysis

Dr. Brett Houlding, Prof. John Quigley

Wednesday 7 May 2014 Exam Hall 14:00 – 16:00

Instructions to Candidates:

Marks are awarded for the best **two** question solutions.

Materials permitted for this examination:

Non-programmable calculators are permitted for this examination.

Question 1

a) Explain the motivation behind, and usefulness of, the technique of Principal Components Analysis.

(4 Marks)

b) Using the method of Lagrange multipliers, show that for a Covariance matrix Σ, the expression $\boldsymbol{a}^T\Sigma\boldsymbol{a}$, subject to the constraint $\boldsymbol{a}^T\boldsymbol{a} - 1 = 0$, is maximised when $\boldsymbol{a}$ is an eigenvector of Σ.

(6 Marks)

c) 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 at the end of this question. Explain

i) Why it is inappropriate to perform the analysis on the non-standardized data.
ii) An appropriate suggestion for the number of principal components to be kept in a lower dimensional representation.
iii) A meaning or explanation of how the data are contrasted by each of the first two principal components.
iv) The co-ordinate location in the best two dimensional representation of a country with standardized values of (-0.3, -1.1, -0.5, 0.8, 0.9, -0.6)

(7 Marks)

d) Describe two other methods of dimension reduction that we have considered. Clearly explain any choices that need to be made when applying such methods and how they differ from Principal Components Analysis.

(8 Marks)

Output for Question 1 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

Question 2

a) What is meant by both a hierarchical and iterative clustering algorithm?

(4 Marks)

b) A subset of the poverty inequality data described in Question 1 c) is as follows:

Country	***BR***	***DR***	***ID***	***LEM***
UK	13.6	11.5	8.4	72.2
USA	16.7	8.1	9.1	71.5
Ireland	15.1	9.1	7.5	71.0
China	21.2	6.7	32.0	68.0

Using Maximum dissimilarity generate a dissimilarity matrix for this extract.

(5 Marks)

c) Using your answer to part b), and complete linkage, produce a sketch of the resulting dendrogram (you should show your working in calculating the values of heights at which clusters are merged), and suggest an appropriate 'cut value', hence confirming your clustering solution.

(6 Marks)

d) Explain the role/purpose of the Rand Index in cluster analysis and calculate its value for two cluster assignments with agreement as indicated by the table below (show the formula used for this specific table):

		Cluster A	
		Group 1	Group 2
Cluster B	Group 1	5	6
	Group 2	3	4
	Group 3	8	5

(4 Marks)

e) Explain a method for calculating the dissimilarity from data points that contain binary, numeric, and categorical data.

(3 Marks)

f) Describe a real life situation (both background and objective), in which a cluster technique, rather than any other statistical technique, would be appropriate.

(3 Marks)

Question 3

Data were recorded on the educational transition of 474 Irish school children aged 11 in 1967.

Variable	Description
lvcert	Indicator variable with value 1 if Leaving Certificate taken (0 otherwise).
DVRT	Drumcondra Verbal Reasoning Test Score of child.
sex	Sex of the child (value 2 for female and 1 for male).
fathocc	A prestige score for the father's occupation.

Logistic regression was used to determine which factors were good predictors of whether a child would take a Leaving Certificate. Output from the logistic regression is given at the end of this question.

a) Provide a motivation for the use of logistic regression, rather than linear regression, for this data.

(3 Marks)

b) Explain what the output (at the end of the question) tells us about the relationship in this model between whether a child takes a Leaving Certificate and the other variables.

(6 Marks)

c) Give the formula for the probability that *lvcert*=1 that results from the logistic model. Use this formula along with the output to predict the probability that a female with a *DVRT* score of 120 and a *fathocc* score of 30 will take a Leaving Certificate.

(6 Marks)

d) Using the above example, describe the role and use of interactions in logistic regression.

(5 Marks)

e) Briefly outline the difference between logistic regression and linear discriminant analysis as a classification tool.

(5 Marks)

Output for Question 3

Call: glm(formula = lvcert ~ DVRT + sex + fathocc, family = binomial(logit))

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.1557	-0.9151	-0.4497	0.9175	2.2907

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.648086	0.984865	-8.781	< 2e-16 ***
DVRT	0.060585	0.008155	7.429	1.09e-13 ***
sex	0.516547	0.215021	2.402	0.0163 *
fathocc	0.039139	0.007492	5.224	1.75e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	651.39 on 473 degrees of freedom
Residual deviance:	530.63 on 470 degrees of freedom
AIC:	538.63

Number of Fisher Scoring iterations: 4