Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Faculty of Engineering, Mathematics and Science

School of Computer Science & Statistics

BA (Mod) JS MSISS, JS-SS MATHS & TSM **Trinity Term 2016**

Multivariate Linear Analysis

Friday 20th May 2016 **Sports Centre** **09.30-11.30**

Prof. Brett Houlding and Prof Sally Brailsford

Instructions to Candidates:

Attempt **two** questions. All questions carry equal marks. Each question is scored out of a total of 25 marks.

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

Materials Permitted for this examination:

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

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

[5 marks]

(b) Using the method of Lagrange multipliers, show that for a n by n 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.

[7 marks]

(c) The United States' CIA publishes demographics of different countries through its World Factbook. In particular, the following information is made available for 219 countries:

Variable	Description
Birth	Annual number of births per 1,000 people.
Death	Annual number of deaths per 1,000 people.
Life	Life expectancy at birth (years).
Infla	Inflation rate.
GDP	GDP per capita.

Note that inflation is a measure of how the average cost of goods increase annually, whilst GDP is a measure of a country's overall economic output (per capita means that this value is divided by the country's population).

The data was saved as a 219 row and 5 column matrix named CIA. An R command to perform PCA on the raw data was:

```
> eigen(cov(CIA))
```

What would be an appropriate R command to perform a PCA on a scaled version of the data such that each variable had unit variance?

[3 marks]

(d) The covariance matrix for the CIA data, and a PCA applied on the scaled version of the data are provided on the next page. Explain i) why the analysis was performed on the scaled version, ii) a suggestion for the number of principal components to be kept, and for each that is kept, how the data are contrasted along that component, iii) the co-ordinate location of the best two-dimensional representation for a country with scaled values of
(-0.2, -0.9, -0.4, 0.7, 0.9)

[10 marks]

Output for Question 1 (d)

> cov(CIA)

	Birth	Death	Life	Infla	GDP
Birth	127	23	-107	23	-82874
Death	23	25	-48	15	-12163
Life	-107	-48	143	-40	81899
Infla	23	15	-40	114	-26835
GDP	-82874	-12163	81899	-26835	126698928

Principal Components applied on the Correlation Matrix:

Rotation:

	PC1	PC2	PC3	PC4	PC5
Birth	-0.50	0.35	-0.06	-0.71	-0.35
Death	-0.43	-0.41	0.59	0.29	-0.48
Life	0.57	0.02	-0.23	0.03	-0.79
Infla	-0.25	-0.72	-0.64	-0.11	-0.02
GDP	0.43	-0.45	0.44	-0.63	0.16

Importance of Components:

	PC1	PC2	PC3	PC4	PC5
St. Dev.	1.71	0.96	0.90	0.55	0.21
Prop. Var.	0.58	0.19	0.16	0.06	0.01
Cum. Prop.	0.58	0.77	0.93	0.99	1.00

2. (a) Explain the difference between classification and clustering techniques.

[2 marks]

(b) Specify the modelling assumptions of linear and quadratic discriminant analysis, and indicate for what type of data such techniques are not appropriate.

[3 marks]

(c) Bivariate data has mean $\boldsymbol{\mu}^T = (1,1)$ and covariance matrix:

$$\Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}, \text{ hence } \Sigma^{-1} = \begin{pmatrix} 1/3 & 0 \\ 0 & 1 \end{pmatrix}$$

Determine i) the Mahalanobis distance and ii) the Euclidean distance of the point $\boldsymbol{x}^T = (2,3)$ to the mean.

[6 marks]

(d) Data were recorded on the educational transition of 474 Irish school children aged 11 in 1967:

Variable	**Description**
lvcert	Indicator variable, 1 if Leaving Certificate taken, 0 otherwise.
DVRT	Drumcondra Verbal Reasoning Test Score of child.
sex	Gender of the child (2 for female, 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. To do so the following command was entered into R:

```
>glm(lvcert ~ DVRT + sex+ fathocc, family=binomial(logit))
```

How would you alter this R command to include an interaction term between the child's gender and the prestige score for their father, and why may you wish to do this?

[5 marks]

(e) The output of the R command given in part (d) is given on the next page. Explain what can be inferred from this output in terms of conclusions regarding the Irish school children, and (without any calculation being made), explain the use/role of the deviance and AIC information that is also provided.

[9 marks]

Output for Question 2 (e)

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

3. (a) Explain the differences between hierarchical and iterative clustering algorithms.

[3 marks]

(b) Detail the pseudo-code for *k*-Means clustering, explain what sort of clusters it is effective at finding, and why the algorithm should be run multiple times.

[5 marks]

(c) A (scaled) subset of the CIA Factbook data described in 1 (c) is as follows:

Country	Birth	Death	Life	Infla	GDP
China	-0.41	-0.32	-1.48	1.46	-1.41
Ireland	1.49	-1.15	0.65	-0.80	0.66
UK	-0.66	1.23	0.57	-0.40	-0.02
USA	-0.41	0.25	0.25	-0.27	0.76

Using Maximum dissimilarity generate a dissimilarity matrix for the subset.

[6 marks]

(d) Using your answer to part (b), and complete linkage, produce a sketch of the resulting dendrogram (you should show your working in calculating heights at which merges occur). Hence confirm your suggested clustering solution.

[7 marks]

(e) A student tries to perform a hierarchical clustering on the full CIA data, they enter the following into R resulting in the error message given below:

```
> hclust(scale(CIA))

Error in if (is.na(n) || n > 65536L) stop("size cannot be NA nor exceed
65536") :

  missing value where TRUE/FALSE needed
```

What did the student do wrong and what should they have entered into R?

[4 marks]