TRINITY COLLEGE DUBLIN
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 2015**

Multivariate Linear Analysis

Monday 11th May 2015 Luce Lower 14:00-16:00

Prof. Brett Houlding and Prof. John Quigley (External Examiner)

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

The location, depth and magnitude of 50 significant seismic events that have occurred near Fiji since 1964 are recorded.

Variable	Description
lat	Latitude of event.
long	Longitude of event.
depth	Depth of event (km).
mag	Richter magnitude of event.

Summary statistics are given at the end of the question. Geo-physicists are interested in determining if there exist groups of seismic events that exhibit similar properties. A Cluster Analysis was performed to establish if groups existed within the data. Outputs from three hierarchical cluster analyses and from a k-Means clustering are given after the question statements.

a) Describe the hierarchical clustering methods, making reference to the term “linkage”, and compare the output from the three hierarchical clustering methods.

[6 marks]

b) Provide a description of the k-Means clustering algorithm and describe what type of cluster structure k-Means clustering is effective at finding.

[5 marks]

c) Provide an analysis description of the k-Means clustering output. In particular, describe what the numerical summaries tell us about the clusters for an appropriate value of k. Justify your choice of k.

[8 marks]

Question 1 continues on next page.....

.......Question 1 continued from previous page.

d) Give a brief account on the use of standardization and data reduction in conjunction with cluster analysis methods.

[3 marks]

e) Describe a real problem in which a clustering algorithm, rather than any other statistical technique, would offer an appropriate solution.

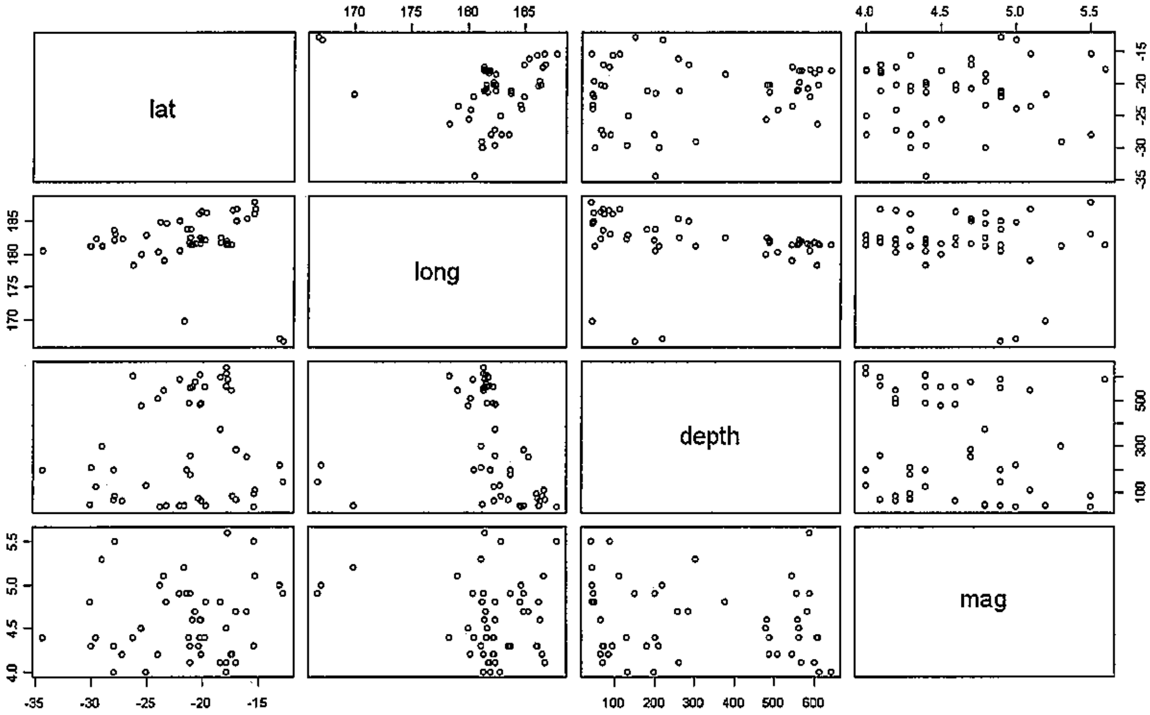
[3 Marks]

Material for this Question in subsequent pages.

Summary Statistics:

```
Variable     Mean        Standard Deviation
lat          -21.46           4.77
long         181.9            4.19
depth        306.2            222.3
mag          4.58             0.43
```

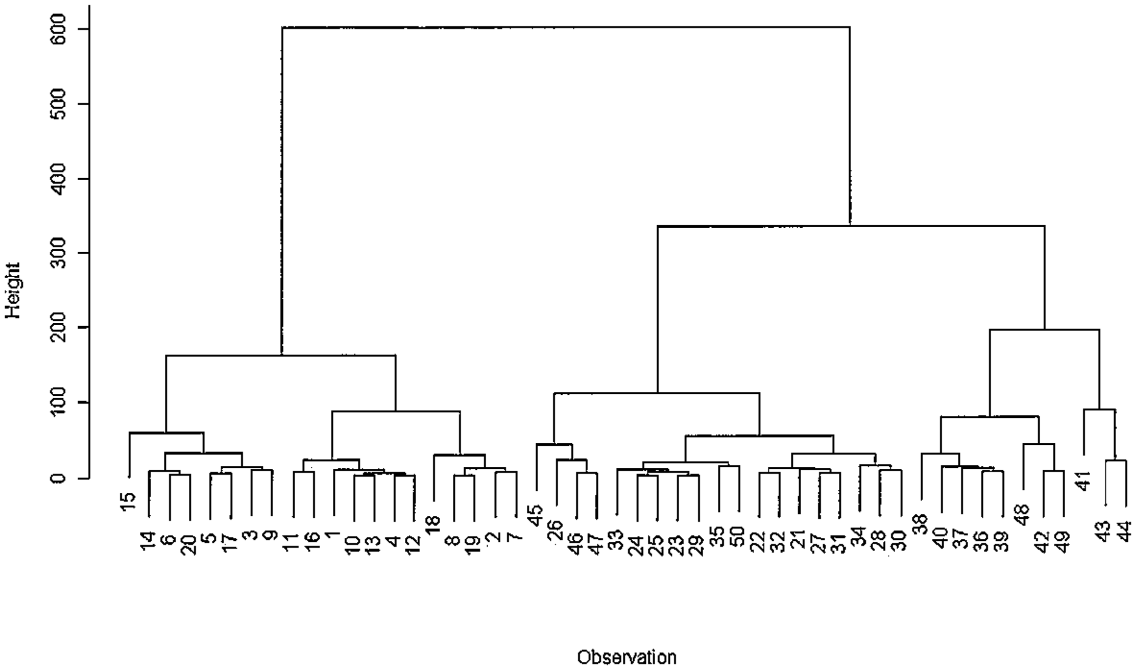
Matrix plot of lat, long, depth and mag



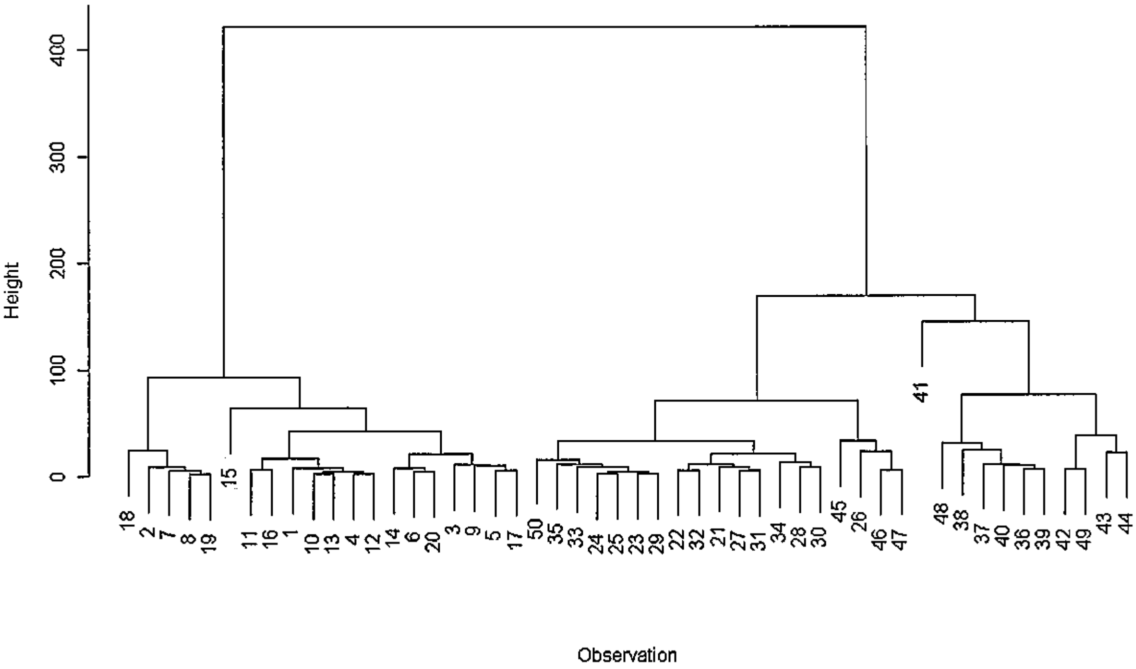
Continued on next page

Hierarchical Clustering

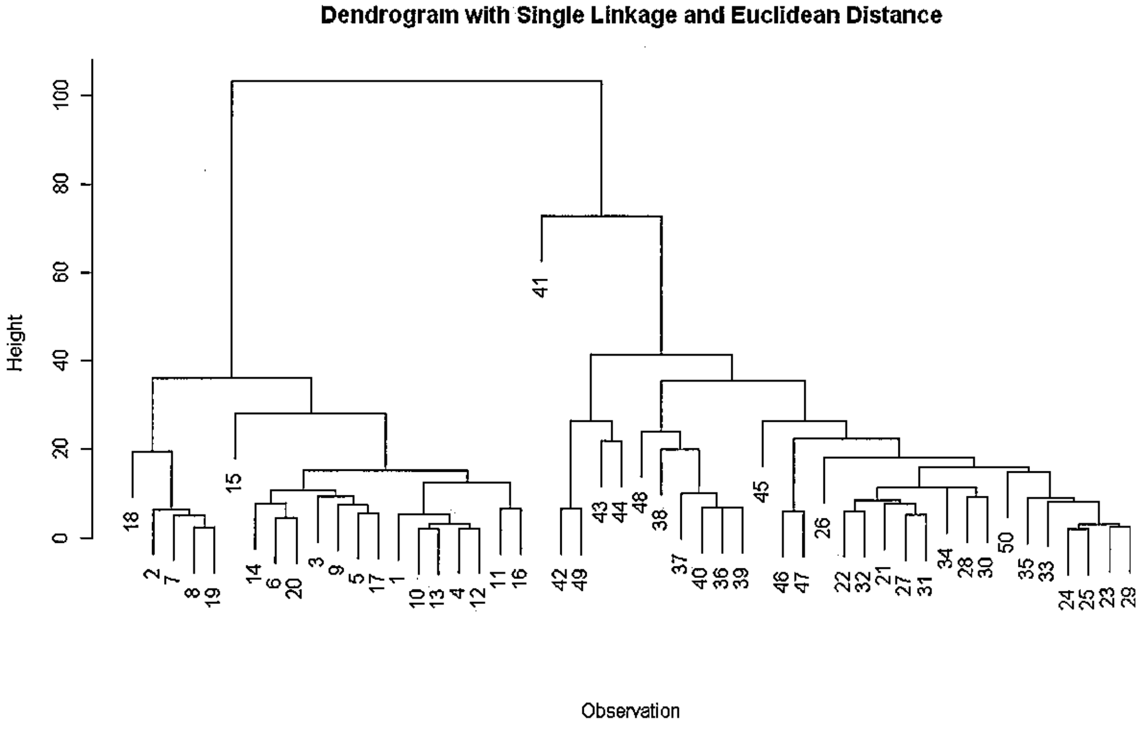
Dendrogram with Complete Linkage and Euclidean Distance



Dendrogram with Average Linkage and Euclidean Distance



Continued on next page



K-Means Clustering

Number of clusters: 1

```
            Number of Obs.         WSS                Avg Dist to Centroid
Cluster 1        50                2423669                  204.97
Sum              50                2423669

Cluster Centroids:

            Cluster 1  Total Data
lat         -21.47     -21.47
long        181.91     181.91
depth       306.2      306.2
mag         4.58       4.58

Distance Between Cluster Centroids:

            Cluster 1
Cluster 1   0.00
```

Continued on next page

Number of clusters: 2

```
            Number of Obs.        WSS               Avg Dist to Centroid
Cluster 1        21               76121                   46.77
Cluster 2        29               197092                  72.82
Sum              50               273213

Cluster Centroids:

            Cluster 1        Cluster 2        Total Data
lat         -20.45           -22.20           -21.47
long        181.20           182.43           181.91
depth       549.90           129.72           306.2
mag         4.50             4.64             4.58

Distance Between Cluster Centroids:

            Cluster 1  Cluster 2
Cluster 1   0.00       420.19
Cluster 2   420.19     0.00
```

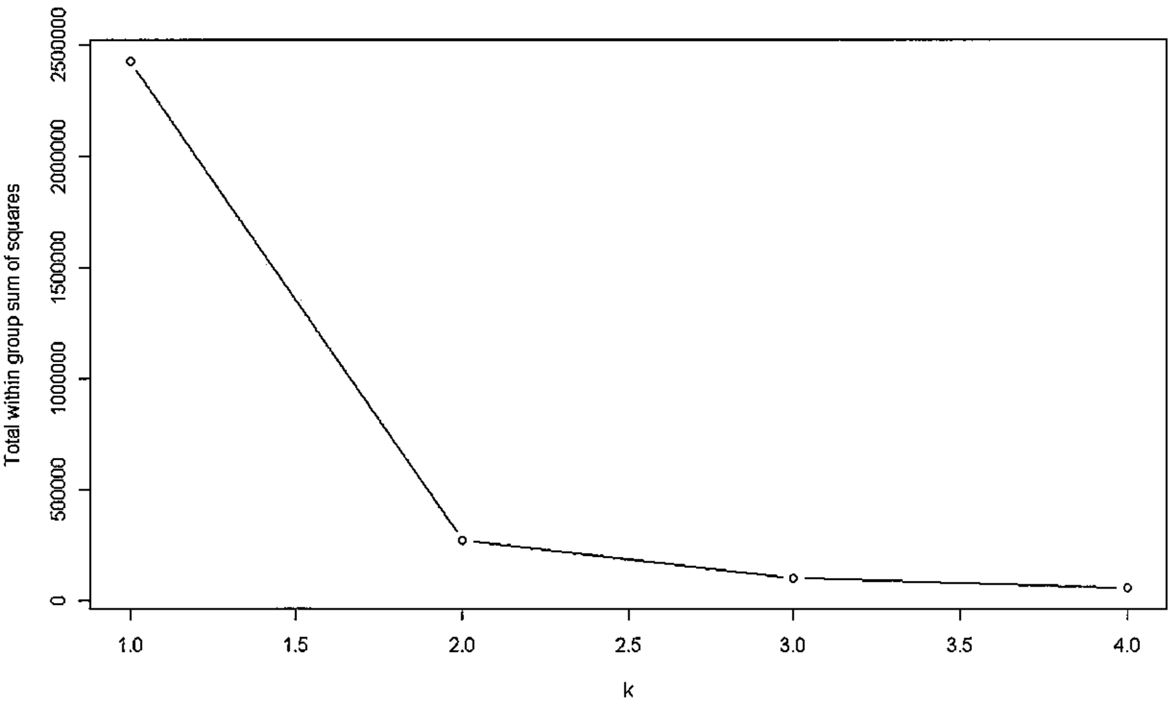
Number of clusters: 3

```
            Number of Obs.        WSS               Avg Dist to Centroid
Cluster 1        11               35539                   48.11
Cluster 2        19               21744                   28.31
Cluster 3        20               44361                   37.99
Sum              50               101644

Cluster Centroids:

            Cluster 1  Cluster 2  Cluster 3  Total Data
lat         -22.73     -21.70     -20.55     -21.47
long        181.31     183.07     181.14     181.91
depth       244.81     76.05      558.60     306.2
mag         4.59       4.68       4.49       4.58

Distance Between Cluster Centroids:

            Cluster 1  Cluster 2  Cluster 3
Cluster 1   0.00       168.78     313.79
Cluster 2   168.78     0.00       482.55
Cluster 3   313.79     482.55     0.00
```

Continued on next page

```
Number of clusters:   4

              Number of Obs.          WSS                Avg Dist to Centroid
Cluster 1          6                  11196                     31.60
Cluster 2          10                 16588                     37.53
Cluster 3          15                 11505                     23.89
Cluster 4          19                 21744                     28.31
Sum                50                 61034

Cluster Centroids:

              Cluster 1   Cluster 2   Cluster 3   Cluster 4   Total Data
lat           -21.59      -23.15      -20.00      -21.70      -21.70
long          181.47      181.21      181.10      183.07      183.07
depth         470.12      231.70      581.80      76.05       76.06
mag           4.45        4.57        4.53        4.78        4.68

Distance Between Cluster Centroids:

              Cluster 1   Cluster 2   Cluster 3   Cluster 4
Cluster 1     0.00        238.47      111.65      394.12
Cluster 2     238.47      0.00        350.11      155.67
Cluster 3     111.65      76.05       0.00        505.75
Cluster 4     394.12      155.67      505.75      0.00
```

Total within group sum of squares vs number of clusters



Question 2

Data were recorded on measurements of 100 specimens for each of 2 species of Crab (with 50 for each sex of each species). Initial statistics for the covariance and mean of the data is provided after the question, as is the output from two principal components analyses applied on the numeric (body measurement) variables of the data.

Variable	**Description**
SP	Indicator variable (1 for Blue species and 0 for Orange species).
SEX	Indicator variable (1 for Male and 0 for Female).
FL	Frontal lobe size (mm).
RW	Rear width (mm).
CL	Carapace length (mm).
CW	Carapace width (mm).
BD	Body depth (mm).

a) Provide a brief description on the benefits of Dimension Reduction as a statistical technique.

[3 marks]

b) Describe the pros and cons of both types of principal components analysis that are given in the subsequent pages.

[4 marks]

c) Making reference to either of the principal components outputs provided, explain the interpretation of the standard deviation row and how the values are generated.

[6 marks]

d) Provide an explanation for the remaining output of the principal component analysis that was considered in part c) and explain any conclusions found with reference to the Crab data.

[6 marks]

Question 2 continues on next page.....

.....Question 2 continued from previous page.

e) Describe an alternative dimension reduction technique that could be applied for this data and explain its difference to principal components analysis.

[6 marks]

Material for this Question in subsequent pages.

Summary statistics

Overall Mean:	*FL*	*RW*	*CL*	*CW*	*BD*
	15.6	12.7	32.1	36.4	14.0

Overall Covariance Matrix:

	FL	*RW*	*CL*	*CW*	*BD*
FL	12.2	8.2	24.4	26.6	11.8
RW	8.2	6.6	16.4	18.2	7.8
CL	24.4	16.4	50.7	55.8	24.0
CW	26.6	18.2	55.8	62.0	26.1
BD	11.8	7.8	24.0	26.1	11.7

Mean by sex and species:

SP	*SEX*	*FL*	*RW*	*CL*	*CW*	*BD*
1 (Blue)	1 (M)	14.8	11.7	32.0	36.8	13.4
1 (Blue)	0 (F)	13.2	12.1	28.1	32.6	11.8
0 (Orange)	1 (M)	16.6	12.2	33.7	37.2	15.3
0 (Orange)	0 (F)	17.6	14.8	34.6	39.0	15.6

Covariance Matrix by species:

Blue Species:

	FL	*RW*	*CL*	*CW*	*BD*
FL	9.1	6.2	20.7	23.6	9.2
RW	6.2	5.2	14.1	16.2	6.3
CL	20.7	14.1	47.6	54.2	21.0
CW	23.6	16.2	54.2	61.9	24.0
BD	9.2	6.3	21.0	24.0	9.4

Orange Species:

	FL	*RW*	*CL*	*CW*	*BD*
FL	10.7	7.7	21.9	24.5	10.1
RW	7.7	6.8	15.4	17.7	7.1
CL	21.9	15.4	45.8	50.8	21.2
CW	24.5	17.7	50.8	56.9	23.5
BD	10.1	7.1	21.2	23.5	9.9

Continued on next page

Principal Components applied on the Covariance Matrix:

```
Rotation:
           PC1     PC2      PC3      PC4      PC5
     FL    0.29    0.32     -0.51    0.73     0.12
     RW    0.20    0.86     0.41     -0.15    -0.14
     CL    0.60    -0.20    -0.18    -0.14    -0.74
     CW    0.66    -0.29    0.49     0.13     0.47
     BD    0.28    0.16     -0.55    -0.63    0.44

Importance of components:
           PC1     PC2      PC3      PC4      PC5
St. Dev.   11.9    1.1      1.0      0.4      0.3
Prop. Var. 0.982   0.00906  0.00698  0.00094  0.0005
Cum. Prop. 0.982   0.99153  0.99851  0.99946  1
```

Principal Components applied on the Correlation Matrix:

```
Rotation:
           PC1     PC2      PC3      PC4      PC5
     FL    0.45    0.14     0.53     0.70     0.10
     RW    0.43    -0.90    -0.01    -0.08    -0.05
     CL    0.45    0.27     -0.31    -0.01    -0.79
     CW    0.45    0.18     -0.65    0.09     0.57
     BD    0.45    0.26     0.44     -0.71    0.18

Importance of components:
           PC1     PC2      PC3      PC4      PC5
St. Dev.   2.19    0.39     0.22     0.11     0.04
Prop. Var. 0.958   0.0303   0.00933  0.00223  0.0003
Cum. Prop. 0.958   0.9881   0.99743  0.99966  1
```

Question 3

Data were recorded on the survival status of 765 passengers of the Titanic.
Logistic regression was used to determine if *PClass*, *Age*, and *Sex* is a good predictor of survival status. Output from the logistic regression is given after the question.

Variable	**Description**
PClass	Binary: 1=1st Passenger Class, 0=2nd or 3rd Passenger Class.
Age	Age in years.
Sex	Binary: 0=Female, 1=Male
Survived	Binary: 1=Yes, 0=No.

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

[3 marks]

b) Explain what the output from the logistic regression tells us about the relationship in this model between *PClass*, *Age*, and *Sex*, and whether the passenger survived the Titanic.

[6 marks]

c) Give the formula for the probability of Survived=1 that results from the logistic model. Use this formula along with the output from the logistic regression to predict the probability that a female from 1st passenger class who was aged 20 years would survive.

[6 marks]

d) For data with only binary entries, explain the Simple Matching (Hamming) and Jaccard dissimilarity measures (you should provide and explain the relevant formulas). Give a motivation for using the Jaccard measure instead of the Hamming measure.

[4 marks]

Question 3 continues on next page..........

..........Question 3 continued from previous page.

e) For data $\mathbf{x}^T=(x_1, x_2, \dots , x_m)$, $\mathbf{y}^T=(y_1, y_2, \dots , y_m)$, and $\mathbf{z}^T=(z_1, z_2, \dots , z_m)$, give the definition of the Maximum dissimilarity measure and show how this measure satisfies the common dissimilarity properties of non-negativity, symmetry, and the triangle inequality.

[6 marks]

Output from logistic regression

```
Call:

glm(formula=Survived~., family = binomial(logit), data=Titanic)

Deviance Residuals:

Min             1Q    Median        3Q        Max
-2.6406   -0.6052   -0.4648       0.7792    2.2699

Coefficients:
               Estimate  Std. Error     z value    Pr(>|z|)
(Intercept)    1.628748  0.248278       6.560      5.37e-11   ***
Age           -0.034331  0.007433      -4.619      3.86e-06   ***
PClass1        1.895242  0.236271       8.021      1.04e-15   ***
Sex1          -2.580914  0.194033      -13.301     < 2e-16    ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance:            1025.57 on 755 degrees of freedom
Residual deviance:        723.15 on 752 degrees of freedom
AIC:                      731.15

Number of Fisher Scoring iterations: 4
```