Analysis and Diagnostics for Censored Regression and Multivariate data

by

Nazrina Aziz

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Doctor of Philosophy
in Statistics.

Victoria University of Wellington
2010

Abstract

This thesis investigates three research problems which arise in multivariate data and censored regression. The first is the identification of outliers in multivariate data. The second is a dissimilarity measure for clustering purposes. The third is the diagnostics analysis for the Buckley-James method in censored regression.

Outliers can be defined simply as an observation (or a subset of observations) that is isolated from the other observations in the data set. There are two main reasons that motivate people to find outliers; the first is the researcher's intention. The second is the effects of an outlier on analyses, i.e. the existence of outliers will affect means, variances and regression coefficients; they will also cause a bias or distortion of estimates; likewise, they will inflate the sums of squares and hence, false conclusions are likely to be created. Sometimes, the identification of outliers is the main objective of the analysis, and whether to remove the outliers or for them to be down-weighted prior to fitting a non-robust model.

This thesis does not differentiate between the various justifications for outlier detection. The aim is to advise the analyst of observations that are considerably different from the majority. Note that the techniques for identification of outliers introduce in this thesis is applicable to a wide variety of settings. Those techniques are performed on large and small data sets. In this thesis, observations that are located far away from the remaining data are considered to be outliers.

Additionally, it is noted that some techniques for the identification of outliers are available for finding clusters. There are two major challenges in clustering. The first is identifying clusters in high-dimensional data sets

is a difficult task because of the curse of dimensionality. The second is a new dissimilarity measure is needed as some traditional distance functions cannot capture the pattern dissimilarity among the objects. This thesis deals with the latter challenge. This thesis introduces Influence Angle Cluster Approach (*iaca*) that may be used as a dissimilarity matrix and the author has managed to show that *iaca* successfully develops a cluster when it is used in partitioning clustering, even if the data set has mixed variables, i.e. interval and categorical variables. The *iaca* is developed based on the influence eigenstructure.

The first two problems in this thesis deal with a complete data set. It is also interesting to study about the incomplete data set, i.e. censored data set. The term 'censored' is mostly used in biological science areas such as a survival analysis. Nowadays, researchers are interested in comparing the survival distribution of two samples. Even though this can be done by using the logrank test, this method cannot examine the effects of more than one variable at a time. This difficulty can easily be overcome by using the survival regression model. Examples of the survival regression model are the Cox model, Miller's model, the Buckely James model and the Koul-Susarla-Van Ryzin model.

The Buckley James model's performance is comparable with the Cox model and the former performs best when compared both to the Miller model and the Koul-Susarla-Van Ryzin model. Previous comparison studies proved that the Buckley-James estimator is more stable and easier to explain to non-statisticians than the Cox model. Today, researchers are interested in using the Cox model instead of the Buckley-James model. This is because of the lack of function of Buckley-James model in the computer software and choices of diagnostics analysis. Currently, there are only a few diagnostics analyses for Buckley James model that exist.

Therefore, this thesis proposes two new diagnostics analyses for the Buckley-James model. The first proposed diagnostics analysis is called renovated Cook's distance. This method produces comparable results with

the previous findings. Nevertheless, this method cannot identify influential observations from the censored group. It can only detect influential observations from the uncensored group. This issue needs further investigation because of the possibility of censored points becoming influential cases in censored regression.

Secondly, the local influence approach for the Buckley-James model is proposed. This thesis presents the local influence diagnostics of the Buckley-James model which consist of variance perturbation, response variable perturbation, censoring status perturbation, and independent variables perturbation. The proposed diagnostics improves and also challenge findings of the previous ones by taking into account both censored and uncensored data to have a possibility to become an influential observation.

Dedication

This thesis is dedicated to my wonderful mom and in loving memory of my dad and also to my siblings with much love, appreciation and affection.

Acknowledgements

While a thesis has a single author by definition, many people are responsible for its existence. Dr Dong Wang, my supervisor, is perhaps the most important of these people. I wish to thank him sincerely for being very supportive and friendly throughout the whole of my PhD. I also greatly appreciate Dr Ivy Liu in reviewing the preliminary draft of this thesis and also advising, helping and supporting me when my supervisor was on sabbatical leave.

Besides my supervisor, I specially would like to thank Dr. Ray Browning, for teaching me how to use R and Latex and provided me with valuable support and instructive suggestions all the way. A special thank you to all examiners for their helpful comments to improve the thesis.

I would like to thank Ministry of Higher Education in Malaysia and Universiti Utara Malaysia for providing the Academic Staff Training Scholarship. I also thank Victoria University and School of Mathematics, Statistics and Operations Research (SMSOR) for awarding me Kathleen Stewart Postgraduate Scholarship, PhD Submission Scholarship and Faculty Strategic Research Grants.

Thank you to all SMSOR staff, especially Prof Megan Clark for giving me the opportunity to be a tutor. I also wants to thank Rebecca and Doris for helping me improve my English writing.

I also wish to express my gratitude to my parents for providing me with the opportunity to be where I am. Without them, none of this would even be possible. To my dear mom, Hajjah Nashrah Hanoum bt Hj Ahmad

Tajuddin, who had done a wonderful job in raising me. Your parenting, your personality, your genes, and your spirituality have helped to shape me into the person I am today and will be tomorrow. Thank you for loving me so very much. I know I am precious to you. You are to me too. I love you.

Life has it ups and downs, during my first year in Victoria University, I had lost a very special person in my life, Allahyarham Hj Aziz Bin Khalid, my father. With unbelievable sorrow I have got to say that a piece of my heart is gone. Words cannot express my feelings of loss. He taught me so much, I could never say all. As long as I'm living he'll stay in my heart. He also deserve my highest gratitude, thank you so much dad! I miss you.

Special thanks go to my brothers Azmel Hafiz, Muhamad Azim, Abdul Manan, Muhamad Zaki, Ahmad Muhammad and my sister Nur Fithrah for their love and support. I owe everything to them. Finally, I also thank my friends for their selfless help throughout my time at Wellington and for providing me with some memorable experiences.

Contents

1 Introduction **1**

1.1 Identification of outliers 1
1.2 Clustering 2
1.3 The Buckley-James method 3
1.4 Thesis Aims 4
1.5 Structure of the Thesis 4

2 Identification of outliers **7**

2.1 Introduction 7
2.2 What is an outlier? 9
2.3 Significance of identification of outliers 10
2.4 Source of outliers 14
2.5 Various methods for identification of outliers 15
2.6 Univariate methods 16
2.6.1 Test of discordance 16
2.6.2 Outlier labeling methods 18
2.7 Multivariate methods 21
2.7.1 Statistical methods 22
2.7.2 Multivariate robust measures 25
2.7.3 Eigenstructure Approach 29
2.7.4 Angles 30
2.7.5 Data mining methods 31
2.8 Solution to outliers 33

2.9 Conclusion 33

3 Outliers Identification by eigenstructure 35

3.1 Introduction 35
3.2 Eigenvalues and eigenvectors 39
3.3 Influence eigenvalues and eigenvectors 40
3.4 Influence eigen for identification of outliers 43
3.4.1 Influence eigen 43
3.4.2 Normalized influence eigen 50
3.5 Influence angle based on eigenstructure 51
3.5.1 Influence angle 51
3.5.2 Modified influence angle 55
3.6 Simulation data set 56
3.6.1 Scenario 1 56
3.6.2 Scenario 2 57
3.6.3 Scenario 3 57
3.7 Real data set 57
3.8 Illustration by simulation data set 60
3.9 Low dimension and small sample size 60
3.9.1 Scenario 1 with $n = 105,\ p = 3,\ m = 5$ 61
3.9.2 Scenario 2 with $n = 105,\ p = 3,\ m = 5$ 66
3.9.3 Scenario 3 with $n = 105,\ p = 3,\ m = 5$ 67
3.10 Low dimension and large sample size 68
3.10.1 Scenario 1 with $n = 1005,\ p = 10,\ m = 5$ 68
3.10.2 Scenario 2 with $n = 1005,\ p = 10,\ m = 5$ 71
3.10.3 Scenario 3 with $n = 1005,\ p = 10,\ m = 5$ 72
3.11 High dimension and large sample size 73
3.11.1 Scenario 1 73
3.11.2 Scenario 2 77
3.11.3 Scenario 3 with $n = 3005,\ p = 100,\ m = 5$ 80
3.12 Illustration using real data set 81

3.12.1 Hawkins Bradu Kass data 81
3.12.2 Stackloss data 82
3.12.3 Salinity data 83
3.13 Conclusion 84

4 An Overview of Clustering 86
4.1 Introduction 86
4.2 What is a clustering problem? 88
4.3 Proximity Measures 89
4.4 Choices of Clustering Algorithm 92
4.4.1 Partitioning Method 93
4.4.2 Hierarchical Method 95
4.5 Conclusion 97

5 Influence Angle Cluster Approach 98
5.1 Introduction 98
5.2 Influence Angle Cluster Approach as a Dissimilarity Measure 100
5.3 Partitioning methods 102
5.4 Strength Measurement of Cluster 105
5.5 Data set 107
5.6 Clustering low dimensional data 108
5.7 Clustering high dimensional data 111
5.8 Clustering real data set 114
5.8.1 Mammal milk data 114
5.8.2 Mortality data 116
5.8.3 Flower data set 117
5.9 Conclusion 118

6 Buckley-James censored regression 119
6.1 Introduction 119
6.2 Survival analysis 120
6.3 Censoring 122

6.4 How do we handle censored data? . . . 125
6.4.1 Kaplan-Meier estimator . . . 126
6.4.2 Regression method for censored data . . . 127
6.5 Buckley-James censored regression . . . 129
6.5.1 Multivariate censored regression . . . 132
6.5.2 Properties of the Buckley-James censored regression 134
6.6 Renovated diagnostics for Buckley-James censored regression 136
6.6.1 Renovated Scatterplots and Residual Plots . . . 136
6.6.2 Renovated Added Variable Plots . . . 137
6.6.3 Renovated Partial Residual Plot . . . 138
6.6.4 Renovated Hat Matrix . . . 140
6.6.5 Measures of explained variation . . . 141
6.7 Conclusion . . . 142

7 New diagnostics analysis for BJ model 144
7.1 Introduction . . . 144
7.2 Renovated Cook's distance for the Buckley-James model . . 146
7.3 Local influence for the Buckley-James model . . . 152
7.3.1 Perturbing the variance . . . 155
7.3.2 Perturbing response variables . . . 159
7.3.3 Perturbing censoring status . . . 160
7.3.4 Perturbing independent variables . . . 164
7.4 Analysis . . . 167
7.4.1 Illustration of renovated Cook's distance . . . 167
7.4.2 Illustration of local influence . . . 173
7.5 Conclusion . . . 181

8 Contributions and Future Work 182
8.1 Contributions and conclusions . . . 182
8.1.1 Contributions . . . 182
8.1.2 Conclusions . . . 185
8.2 Future Work . . . 189

List of Tables

2.1 Three types of simulated data sets 11
2.2 Dixon tests for univariate normal samples 18

3.1 The simulation data set used for illustration 56
3.2 The real data set used for illustration 58
3.3 $\lambda_{1(i)}$ of $\Delta^*_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^*_{1(i)}$–Condition 1, $n = 105$ 62
3.4 $\Delta^*_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^*_{1(i)}$–Condition 1, $n = 105$ 63
3.5 $\Delta^*_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^*_{1(i)}$–Condition 2, $n = 3005$ 75
3.6 $\Delta^*_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^*_{1(i)}$–Condition 3, $n = 3005$ 77
3.7 $\Delta^*_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^*_{1(i)}$–Scenario 2, $n = 3005$ 78

5.1 Interpretation of the Silhouette Index (SI) 107
5.2 The real data set used for clustering illustration 108
5.3 Silhouette width for interval variables; $n = 300, p = 30$ 110
5.4 Silhouette width for mixed variables; $n = 300, p = 30$ 110
5.5 Silhouette width for interval variables; $n = 4500, p = 100$ . . 113
5.6 Silhouette width for mixed variables; $n = 4500, p = 100$ . . . 113
5.7 Silhouette width for mammal milk data 115
5.8 Silhouette width for mortality data 116
5.9 Silhouette width for flower data 118

6.1 Regression Line for Stanford Heart Transplantation 139

7.1 Information of the SHT data 169

7.2 BJ model for SHT data, $n = 69$ 173
7.3 BJ model for SHT data, $n = 152$ 177
7.4 Buckley-James model for Lung cancer data 179

List of Figures

2.1 Boxplot for three simulated data sets 13
2.2 Mahalanobis distance plot for Hawkins Bradu Kass data . . 23
2.3 Wilks plot for Hawkins Bradu Kass data 24

3.1 Plot of Mahalanobis distance 59
3.2 Plot for $n = 105$, $p = 3$, $m = 5$ – condition 1, scenario 1 . . . 61
3.3 Plot for $n = 105$, $p = 3$, $m = 5$ – condition 2, scenario 1 . . . 64
3.4 Plot for $n = 105$, $p = 3$, $m = 5$ – condition 3, scenario 1 . . . 65
3.5 Plot for $n = 105$, $p = 3$, $m = 5$ – scenario 2 66
3.6 Plot for $n = 105$, $p = 3$, $m = 5$ – scenario 3 67
3.7 Plot for $n = 1005$, $p = 10$, $m = 5$ – condition 1, scenario 1 . . 68
3.8 Plot for $n = 1005$, $p = 10$, $m = 5$ – condition 2, scenario 1 . . 69
3.9 Plot for $n = 1005$, $p = 10$, $m = 5$ – condition 3, scenario 1 . . 70
3.10 Plot for $n = 1005$, $p = 10$, $m = 5$ – scenario 2 71
3.11 Plot for $n = 1005$, $p = 10$, $m = 5$ – scenario 3 72
3.12 Plot for $n = 3005$, $p = 100$, $m = 5$ – condition 1, scenario 1 . . 73
3.13 Plot for $n = 3050$, $p = 100$, $m = 50$ – condition 1, scenario 1 . 74
3.14 Plot for $n = 3005$, $p = 100$, $m = 5$ – condition 2, scenario 1 . . 75
3.15 Plot for $n = 3005$, $p = 100$, $m = 5$ – condition 3, scenario 1 . . 76
3.16 Plot for $n = 3005$, $p = 100$, $m = 5$ – scenario 2 78
3.17 Plot for $n = 3050$, $p = 100$, $m = 50$ – scenario 2 79
3.18 Plot for $n = 3005$, $p = 100$, $m = 5$ – scenario 3 80
3.19 Plot for Hawkins Bradu Kass data 81

3.20 Plot for Stackloss data 82
3.21 Plot for Salinity data 83

5.1 Cluster plot for $n = 300,\ p = 30$ – *pam* 109
5.2 Cluster plot for $n = 300,\ p = 30$; mixed variable – *pam* 111
5.3 Cluster plot for $n = 4500,\ p = 100$ – *clara* 112
5.4 Cluster plot for $n = 4500,\ p = 100$; mixed variable – *clara* . . 114
5.5 Cluster plots for mammal milk using *pam* 115
5.6 Cluster plots for mortality using *pam* 116
5.7 Cluster plots for flower using *pam* 117

6.1 Plot of right censored data 123
6.2 Plot of interval censored data 124
6.3 Renovated partial residuals plot using SHT data 139

7.1 Renovated leverage plot for SHT data 171
7.2 Renovated Cook's distance plot for SHT data 172
7.3 Plots of $|h_{max}|$ (perturbing variance) for SHT data 174
7.4 Plots of $|h_{max}|$ (perturbing censoring status) for SHT data . . 175
7.5 Plots of $|h_{max}|$ (perturbing response variable) for SHT data . 176
7.6 Plots of $|h_{max}|$ (perturbing x_1) for SHT data 176
7.7 Plots of $|h_{max}|$ (perturbing x_2) for SHT data, $n = 152$ 178
7.8 Plot of $|h_{max}|$ (perturbing variance) for Lung data 178
7.9 Plot of $|h_{max}|$ (perturbing response variable) for Lung data . 180
7.10 Plot of $|h_{max}|$ (perturbing censoring status) for Lung data . . 180

List of Abbreviations

(I)	a subscript used to indicate the omission a set of m observations
(i)	a subscript used to indicate the omission of the ith observation
$\mathbf{\Lambda}$	diagonal matrix of eigenvalues
$\mathbf{Q}$	The weights matrix
$\Delta^{*}_{j(i)}$	The influence eigen
$\Delta^{**}_{j(i)}$	The normalized influence eigen
δ_i	The censoring indicator for the ith observation
λ_j	The j eigenvalue of sample covariance matrix
$\mathbf{I}$	The identity matrix
$\mathbf{S}$	sample covariance matrix
$\mathbf{V}$	matrix of eigenvectors
$\theta^{*}_{j(i)}$	The modified influence angle
$\theta_{j(i)}$	The influence angle
v_j	The eigenvector corresponding to λ_j
$h(y)$	The hazard function at time y

H^* The renovated hat matrix

l_{ij} The principal component scores of the omitted observation in the principal component decomposition of the complete data **X**

m number of deleted obeservations

n number of observations

p number of explanatory variables

RD_i^* The renovated Cook's distance for the *i*th observation

$S(y)$ The survival function at time y

t_i The censoring time for the *i*th observation

Y A survival random variable

Y^* The renovated response for the *i*th observation

Z_i The observed survival time for the *i*th observation $Z_i = \min(Y_i, t_i)$

Chapter 1

Introduction

This thesis contributes to three areas of statistics. The first is identification of outliers. The second is dissimilarity measures for clustering purposes. The third is the diagnostics analysis for the Buckley-James method.

In this introductory chapter, a brief description of the subjects under studied are provided. In addition, the motivations and aims of the current study are provided. This chapter concludes with the organization of the structure for the remaining chapters of the thesis.

1.1 Identification of outliers

The identification of outliers is a part of the field of statistics. It is very important and deserves more attention because outliers are one of the possible reasons for the failure of analysis in explaining finding.

Outlier detection methods have been suggested for numerous applications (Hawkins, 1980; Barnett and Lewis, 1994; Penny and Jolliffe, 2001; Acuna and Rodriguez, 2004). Outlier detection methods can be classified into univariate methods and multivariate methods.

Or, one can classify them based on parametric methods (Hawkins, 1980; Rousseeuw and Leroy, 1987; Barnett and Lewis, 1994) and non-parametric methods (Williams, Baxter, He, Hawkins and Gu, 2002). The other ex-

ample of the outlier detection method categorization are clustering techniques (Kaufman and Rousseeuw, 1990; Ramaswamy, Rastogi and Shim, 2000; Acuna and Rodriguez, 2004)

However, some methods suffer from computational complexity, i.e. the efficiency of algorithms. Therefore, the current study aims to use the techniques which are based on the influence of eigenstructure for the identification of outliers. Chapter 3 of this thesis will illustrate the simple and exploratory nature of the techniques.

Additionally, the techniques are well suited in the identification of the outliers in a high dimensional data, in which the outliers appear to form a cluster from a separate sample. Thus study on the next subject is motivated by the desire to use the technique in chapter 3 to create a dissimilarity measure for clustering purpose.

1.2 Clustering

Clustering allows one to handle a large data set effectively. It is a technique for solving classification problems (Everitt, 1993). The basic idea of clustering is that it arranges objects, i.e. people, animals, plants and so forth, into groups where those objects in the same group will have a high degree of association, while the objects of different groups will have a low degree of association.

One of the main objectives of the current study is to propose a tool that can separate the objects in the data set to build a strong association between the objects in similar groups, yet a poor association associated between the objects in the other groups.

It is important to note that a data measurement is a very important step in clustering. Instead of using dissimilarity measures such as Euclidean distance or Manhattan distance, one may use the influence eigenstucture to measure the dissimilarity between observations.

Gnanadesikan and Kettenring (1972) mentioned that the total influence eigen can also be considered as the influence interpretation of the Euclidean distance.

1.3 The Buckley-James method

The next subject concerns a diagnostics analysis for the Buckley-James method, i.e. a method that handles censored data. A censored data set is a data set that contains observations with incomplete information; these observations occur when the event of interest is not fully observed.

Survival studies will normally have this type of data set because this type of study is related to looking at the life time of the subjects under studieds. At the end of this type of study, there could be patients who survived and also did not survive.

Therefore, the analyst would not be able to record the patients' exact time of survival or length of survival. Surviving patients normally reflect the success of a new treatment method (if this is the purpose of the study); therefore, the analyst would not want to label them as a missing data. The patient with incomplete information is called a censored observation.

Many methods exist for survival analysis. The current study, particularly described in chapters 6 and 7, is interested in the Buckley-James method because this method has a great performance (Miller and Halpern, 1982; Heller and Simonoff, 1990; Heller and Simonoff, 1992; Stare, Heinzl and Harrell, 2000)

Nevertheless, the Buckley-James method is rarely used as compared to some other methods for example Cox method. The reason for this is the deficiency of diagnostic analysis tools for this Buckley-James method. In line with this gap, chapter 7 of this thesis will outline a proposal for a few diagnostic analysis tools for the Buckley-James method.

1.4 Thesis Aims

The aims of this thesis are to:

- apply the influence eigenstructure as a tool for identifying outliers;
- construct clusters using the influence eigenstucture as a dissimilarity measure;
- create a diagnostics analysis for the Buckley-James method.

1.5 Structure of the Thesis

The thesis is organized as follows. This chapter, Chapter 1 is the introductory chapter. Next, Chapters 2 and 3 focus on the identification of outliers. The following chapters, Chapters 4 and 5 illustrate the dissimilarity measures for clustering purposes. Finally, Chapters 6 and 7 describe diagnostics analyses for the Buckley-James method.

Chapter 2: First, the most referred-to definitions of outliers in the literature are provided in this chapter. Second, the significance of identifying outliers are justified. This chapter also explains how outliers could easily influence the two important estimators (mean and variance) in most statistical analyses. Third, this chapter briefly discusses the various categories in classifying the techniques for the outliers identification. Next, some of the techniques which are normally used to detect outliers are presented. This chapter concludes with a brief suggestion on how to handle outliers should they exist in the data set.

Chapter 3: This chapter uses the influence eigenstructure for identifying outliers in a high-dimensional data. First, the definition and the computation of eigenstructure are briefly explained. The following section discusses the influence eigenstructure, and it will be used for identifying out-

liers. Finally, the performance of the techniques discussed in this chapter are evaluated by using the simulated data sets and the real data sets that have been used to evaluate the existing outlier detection methods.

Chapter 4: The first part of Chapter 4 explains the clustering problem and provides some examples of clustering applications. The subsequent part offers a brief discussion on the dissimilarity measures. A simple outline of clustering algorithms is given next and a few main clustering algorithm categories are discussed. The final part of this chapter concentrates on hierarchical clustering and partitioning clustering.

Chapter 5: This chapter starts by explaining the properties and algorithms of the new dissimilarity measure. Next, it describes the clustering algorithms that will be used in collaboration with the dissimilarity measures to identify clusters in the data set. A brief explanation regarding the simulated data set and the real data set used arethen provided in order to evaluate the performance of the new and existing dissimilarity measures. This chapter carries out a comparison between the new and existing dissimilarity measures in a few clustering algorithms for low dimensional and high dimensional data.

Chapter 6: Chapter 6 presents a discussion concerning various methods to resolve the problem of incomplete data sets, i.e. censored data sets. The term "censoring" was first used in 1949. One may find this term to be mostly used in the analyses in dealing with a life time data. This chapter focuses on survival analysis whereby the different types of censoring that may emerge in the survival analysis is explained. Besides, several methods that one may use to solve the problem involving censoring data sets are briefly explained.

The Buckley-James approach to survival regression method performs better than the other methods (see Miller and Halpern, 1982; Heller and

Simonoff, 1990; Heller and Simonoff, 1992; Stare et al., 2000). However, it is still rarely used by researchers as it is not well established in most computer software packages. In addition, there are only a few diagnostics analyses developed for this method thus far. As such, this chapter illustrates the application of the Buckley-James method as one of the survival regression methods.

Chapter 7: Two new diagnostics analyses for the Buckley-James method are proposed in this chapter. The first diagnostic analysis was based on Cook's idea, and the second one used Shi's approach. In censored regression, one will find that most diagnostic studies using local influence approach have only been applied to Cox method and Kaplan-Meier method (see, Reid, 1981; Pettitt and Daud, 1989; Weissfeld, 1990; Escobar and Meeker, 1992; Barlow, 1997).

In this chapter, Chapter 7, local influence diagnostics for the Buckley-James method which consist of variance perturbation, response variable perturbation, censoring status perturbation and independent variables perturbation, are presented. Shi's 1997 approach is easier to apply without considering a likelihood assumption. This method is able to assess the effect of perturbations to the data will have on inferences. It should be noted that it successfully discovers influential observations from the censored and uncensored data.

Chapter 8: This chapter concludes the thesis. It summarizes the contributions of the thesis, and points out several problems that should be explored further in order to ascertain that the proposed method works in a more efficient manner.

Chapter 2

Identification of outliers

2.1 Introduction

The word 'statistics' generally gives an imagery of numbers and figures to many people. In fact, statistics is a broad field of knowledge, and it is not simply associated with statisticians but is also relevant to many different fields of research, particularly to people who are occupied with quantitative research. Quantitative research in general is involved with the analysis of numerical data. Dealing with complicated numerical data usually results in the researcher failing to notice the nature of each observation in the data set.

Some might argue about the constraint of time and costs if they have to be particular with each observation in the data set before proceeding to the statistical analysis. It must be noted that this is one of the possible causes some analyses failing to explain a good finding. Many researchers are not aware of the significance of this issue. This issue leads us to a key word called 'outlier'.

In §2.2, four of the most commonly referred-to definitions of an outlier are presented. The concept of an outlier as defined in these four definitions are almost identical. Next, in §2.3, the significance of the identification of an outlier is explained. This section also explains how outliers could easily

influence two important estimators in most statistical analysis, i.e. mean and variance. Other than that, examples of outlier effects on the output of several statistical analyses are also given.

An outlier can appear in the data set for different reasons. In §2.4, three possible causes of outliers existing in the data set are listed. Nevertheless, in order to detect the source of outliers, first, one needs to examine whether outliers exist in the data set or not. Previous studies have proposed various methods for identifying outlier.

The method for identifying outliers can be divided into two main categories, namely the univariate and multivariate methods. This chapter is focuses on the multivariate methods category. Mahalanobis distance is the earliest and well known approach for the identification of outliers in the multivariate category. It was proposed by Mahalanobis (1930).

However, this approach suffers from masking and swamping effects, as the traditional mean and covariance could easily be inflated by outliers. What is meant by masking and swamping effects?

There are many definitions of masking and swamping effect (see, Hawkins, 1980; Iglewics and Martinez, 1982; Davies and Gather, 1993; Barnett and Lewis, 1994). Nevertheless, the definition given by Acuna and Rodriguez (2004) is simple and easy to understand. It is as follows:

Masking effect: It is said that if one outlier masks a second outlier, the second outlier can be considered as an outlier only by itself, but not in the presence of the first outlier. Thus, after the deletion of the first outlier, the second instance emerges as an outlier. Masking occurs when a cluster of outlying observations skews the mean and the covariance estimates toward it, and the resulting distance (the space between two objects or points) of the outlying point from the mean is small.

Swamping effect: It is said that one outlier swamps a second observation if the latter can be considered as an outlier only under the presence of the first one. In other words, after the deletion of the first outlier, the second observation becomes a non-outlying observation. Swamping occurs

when a group of outlying instances skews the mean and the covariance estimates toward it and away from other non-outlying instances, and the resulting distance from these non-outlying instances to the mean is large, making them look like outliers.

Many methods have emerged in order to find a solution to the masking and swamping effects, particularly methods that attempt to modify the traditional mean and variance, so that they become robust estimators and can be used for identification of outliers. In §2.6 and §2.7, various methods proposed by previous studies that can be used for the identification of outliers are briefly discussed.

Next, §2.8 briefly explains how to handle outliers if they exist in the data set. The story about the ozone hole above Antarctica, as an example of the outlier effect if they were not handled carefully, is also given in this section.

2.2 What is an outlier?

There are many definitions of an outlier(s) that are almost identical in meaning. The basic definition of an outlying observation is an observation(s) that does not fit the model of the rest of the data. However, given that there is currently no universally accepted definition for an outlier, the four most-commonly used definitions of outliers are provided. The first definition is given by Grubbs (1969); he says

> an outlier is one observation that appears to deviate markedly from other observations of the sample in which it occurs.

The next definition is presented by Hawkins, (1980, pg 1), where he defines

> an outlier as an observation which deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.

Following is the definition by Barnett and Lewis, (1994, pg 3), who state that

> an outlier in a set of data is an observation (or a subset of observations) which appears to be inconsistent with the remainder of that set of data.

Hair, Anderson, Tatham and Black, (2005, pg 64) explains that

> outliers are observations with a unique combination of characteristics identifiable as distinctly different from the other observations.

These definitions all refer to an observation(s) that is surprisingly different from the rest of the data. However, the words ”appears to deviate”, ”arouse suspicion”, ”inconsistent” and ”distinctly different” imply some kind of subjectivity or preconceived ideas about what the data should look like. This chapter uses the four definitions of outliers described above as the basis for discussion. In conclusion, an outlier can be simply defined a s as an observation (otherwise a subset of observations) that is isolated from the other observations in the data set.

2.3 Significance of identification of outliers

Barnett and Lewis (1994) mention two major reasons that could motivate the inspection of existent outliers before someone moves far above and away from the ground level of data analyses. The first reason is the intention of the study itself; see the example of Mr Haldum’s cases in Barnett and Lewis, (1994, pg 4). The second reason is related to the effect of outliers on the findings of analyses. A few examples of outlier effects on the findings of analyses are given in this section.

For a simple description of outlier influence, first look at the examples of simulated data sets given in Table 2.1. In Table 2.1, there are three

types of simulated data sets. The first column represents a normal data set, whereas the second and third columns are data sets with an outlier(s). This can be verified by constructing a box plot for each data set.

Observation	Data 1	Data 2	Data 3
1	-0.93	-0.93	-0.93
2	1.32	**3.97**	**13.24**
3	0.62	0.62	0.62
4	-0.04	-0.12	-0.40
5	-1.00	**-3.01**	-1.00
6	-0.82	-0.82	-0.82
7	-0.34	-1.02	-0.34
8	-1.53	-1.53	-1.53
9	-0.25	-0.25	-0.25
10	-1.14	-1.14	-1.14
Mean	-0.42	-0.43	0.73
Variance	0.76	3.31	19.65

Table 2.1: Three types of simulated data sets

The first data set is generated as normally distributed with mean zero and variance one, $N(0, 1)$. Based on Pickard, Kitchenham and Linkman (2001), outliers are generated for data 2 and data 3 in Table 2.1 by multiplying the chosen observations in data 1 by a constant value.

Moderate outliers can be generated by multiplying the chosen observation in data 1 by 3. The observations are selected by the following algorithm:

- Step 1: Generate a random value between 0 and 1;
- Step 2: Select the observation that has a random value less than 0.10. This indicates 10% of the observations are outliers;
- Step 3: Multiply the observation corresponding to the random value less than 0.10 by 3, otherwise preserve the observation value.

There are 4 observations in data 1 corresponding to the random value less than 0.10 and they were multiplied by 3. The new generated data set

is now labeled as data 2. However, note that out of the 4 chosen observations, only 2 observations in data 2 become the centre of attention, i.e. observations 2 and 5 (refer to their value in the Table 2.1 and Figure 2.1).

To create the severe outlier, multiply the chosen observations in data 1 by 10. The algorithm is given as follows:

- Step 1: Generate a random value between 0 and 1;
- Step 2: Select the observation in data 1 that corresponds to the random value less than 0.05. Therefore, 5% of the observations are outliers;
- Step 3: Multiply the chosen observation in data 1 that corresponds to the random value less than 0.05 by 10.

Consider the newly generated data set as data 3. Note that observation 2 in data 3 has a very large value compared to the others. Those peculiar observations from data 2 and 3 cause the sample estimators, i.e. mean, $\bar{x}$ and variance, s^2 for data 2 and 3 to be larger than the ones obtained from data 1. One can verify this by using the simple univariate test, i.e. the boxplot.

From Figure 2.1, boxplot of data 2 and 3 clearly show observations 2 and 5 in data 2 and observation 2 in data 3 as outliers. Even though the box and the whiskers of the boxplot for each data set look pretty normal, the existence of outliers may lead to the conclusion that observations in the data set are not normal even though all of them may be normal except for outlying observations.

Furthermore, these two sample estimators, i.e. mean, $\bar{x}$ and variance, s^2 play very important roles in the multivariate analyses. Both of these estimators are utilized in developing models for the data set. If outliers can change the values of these estimators severely, it may cause a big problem in more complex analyses, especially in multivariate analyses.

Note that every multivariate technique has underlying assumptions, both statistical and conceptual (Hair, Anderson, Tatham and Black, 2005).

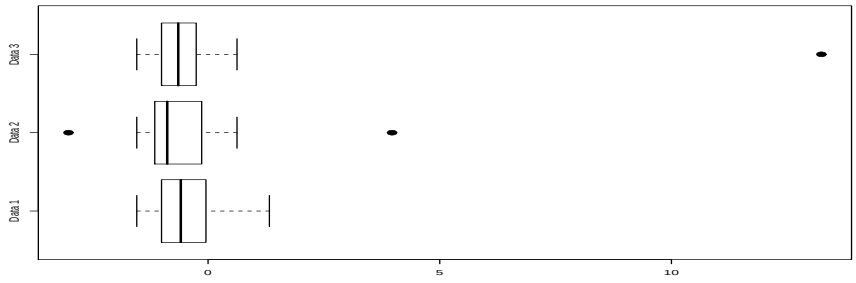


Figure 2.1: Boxplot for three simulated data sets

Examples of assumptions for techniques based on statistical inference are multivariate normality, linearity, independence of error terms and equality of covariance in a dependent relationship. The conceptual assumptions are related to such issues as model formulation and the types of relationships represented.

Both statistical and conceptual assumptions must all be met before any model estimation is attempted. However, there are situations when one cannot meet these assumptions. One of the reasons is probably due to the existence of outliers in the data set.

If the statistical models are simply applied to data sets containing outliers, one might get a misleading result. There are a few examples of outlier consequences in multivariate analyses.

For example, in regression analyses, one of the effects of the appearance of outliers is that they would control the regression line with the outliers pulling the regression line in their direction. In other words, outliers will influence the regression coefficient, which might result in all the predicted values calculated wrongly. Many authors have been critical in discussing these issues (Cook and Weisberg, 1982; Rousseeuw and Leroy, 1987; Chatterjee and Hadi, 1988)

In the case of principal component analyses or factor analyses, the existence of outliers will deflate the correlation coefficient and this will automatically influence the factor score. In the case of discriminant analyses,

outliers might cause problems when it comes to predicting the observations grouping. It might classify the observations into an incorrect group since the function value developed from eigenvectors might be affected by a variance value. A similar problem can also happen to analyses of variance; the appearance of outliers might prove to be a large influence on the estimate of variance, and this can cause a low probability of rejecting the hypothesis since it will affect the F statistics value.

Outliers are also a special target of interest in the real environment. Hodge (2004) listed a few applications which enable outlier detection. For example, in work that requires monitoring, one can detect mobile phone deception by monitoring phone activity or suspicious trades in the equity market, while in loan application processing, one can identify a potentially problematic customer. Outliers have also been utilized for detecting unauthorized access in computer networks and for monitoring medical conditions, such as heart-rate, etc.

2.4 Source of outliers

Outliers may arise coincidently without any anticipation by a researcher. Sometimes it cannot be explained. However, there are a few possible reasons for the existence of outliers in the data set. Barnett and Lewis (1994) classified outlier source into three types. The initial source is named as inherent variability, which implies a situation beyond one's control since it might arise from the natural characteristics of the individual variable. For example, if the data collection involves time duration, it may cause an occurrence of outliers since some of the observations might be influenced by any event that might occur unexpectedly throughout the period of the study. The next cause is measurement error such as reading, computing and typing errors during the data entry process. This possibly makes the observation peculiar compared to the other observations in the data set. The last reason is the execution error, related to the research design where

one might choose a biased sample or include individuals that are not true representatives of the population that is to be sampled. No matter what the causes of outliers are, the most important aspect of outlier issue is the technique to identify whether there are outliers in the data set or not. By identifying the existence of outliers, one may identify the source of the outliers.

2.5 Various methods for identification of outliers

There are many methods available for the identification of outliers. All of these methods can basically be grouped into two categories, namely the univariate method and the multivariate method (see Hawkins, 1980; Barnett and Lewis, 1994). The univariate method is performed independently on each variable, whereas the multivariate method investigates the relationship of several variables (Franklin, Thomas and Brodeur, 2000). One can also classify the methods in both categories into parametric and non-parametric approaches.

Other classifications of outlier detection methods can be found in Papadimitriou, Kitawaga, Gibbons and Faloutsos (2002), Hu and Sung (2003) and Acuna and Rodriguez (2004). This chapter will briefly explain outlier identification methods for high-dimensional data. Detailed explanations about those methods can be found in Hawkins (1980), Barnett and Lewis (1994), Papadimitriou et al. (2002), Hu and Sung (2003) and Acuna and Rodriguez (2004).

This chapter does not attempt to summarize literature covering the univariate method but some major concepts are reviewed before moving to the multivariate method.

2.6 Univariate methods

Many methods have been proposed for univariate outlier detection. The test of discordance, i.e. a formal test, and outlier labeling methods, i.e. informal test are the most popular approaches.

2.6.1 Test of discordance

The test of discordance needs test statistics for hypothesis testing and it is usually based on the assumption of well-behaved distribution. Normally the distribution is assumed to be identically and independently distributed. Additionally, the type of expected outlier and the distribution parameters are assumed to be known.

From Barnett and Lewis (1994), there are hundreds of discordance tests that have been developed for different conditions depending on

(i). the data distribution, i.e. whether the distribution parameters are known or not;

(ii). the number of expected outliers;

(iii). the types of expected outliers.

The test of discordance is quite powerful since it is based on distribution assumption. However, it is noted that most real world data may not follow a specific distribution or the distribution is unknown. The discordance test is thoroughly discussed in Barnett and Lewis (1994) and Iglewicz and Hoaglin (1993). Examples of discordance test are generalized extreme studentized deviate (ESD), kurtosis statistics and the Dixon test.

(i). **Extreme Studentized Deviate (ESD)**
The ESD test is suitable to use if we want to identify a single outlier in a normally distributed data. It is also known as the Grubb test.

The maximum deviation from the mean is given as

$$\tau = \frac{|x_i - \bar{x}|}{s} \tag{2.1}$$

where x_i is the observation, $\bar{x}$ and s are the mean and standard deviation of the data set, respectively. Equation 2.1 is calculated for each observation and the value is compared to the critical value, τ_α at the selected α. If τ is greater than the τ_α (see Iglewicz and Hoaglin (1993) for ESD test critical values), then the observation under consideration is an outlier.

(ii). **Dixon test**

The Dixon test is based on the ratio of the ranges and it is generally used for detecting a small number of outliers. There are six test statistics from Dixon for normal univariate samples. It is a very simple test. However, these tests are applicable to only sample sizes of up to 30. The algorithm is as follows:

- Step 1: Observations in the data set are sorted in ascending order, $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$ where $x_{(1)}$ is the lowest observation and $x_{(n)}$ is the highest one;
- Step 2: Compute the suitable test statistics and depending on the number of suspected outliers, different test statistics are used to identify potential outliers. The corresponding test statistics are given in Table 2.2.

  Tests r_{10}, r_{11}, r'_{11}, r_{12} and r'_{12} are the test statistics for an extreme outlier, $x_{(n)}$ or $x_{(1)}$ in a normal sample with population variance unknown, whereas tests r_{20}, r'_{20}, r_{21}, r'_{21}, r_{22} and r'_{22} are for two extreme observations either the upper-pair $x_{(n)}, x_{(n-1)}$ or the lower-pair $x_{(1)}, x_{(2)}$ in a similar normal sample;
- Step 3: Next the value of test statistics is compared to the critical value, r^* for a given number of observations n and at a

Table 2.2: Dixon tests for univariate normal samples

Applicability of test $n_{min} - n_{max}$	Value(s) tested	Test Statistics
3-30	Upper $x_{(n)}$	$r_{10} = (x_{(n)} - x_{(n-1)})/(x_{(n)} - x_{(1)})$
4-30	Upper $x_{(n)}$	$r_{11} = (x_{(n)} - x_{(n-1)})/(x_{(n)} - x_{(2)})$
4-30	Lower $x_{(1)}$	$r'_{11} = (x_{(2)} - x_{(1)})/(x_{(n-1)} - x_{(1)})$
5-30	Upper $x_{(n)}$	$r_{12} = (x_{(n)} - x_{(n-1)})/(x_{(n)} - x_{(3)})$
5-30	Lower $x_{(1)}$	$r'_{12} = (x_{(2)} - x_{(1)})/(x_{(n-2)} - x_{(1)})$
4-30	Upper pair $x_{(n)}, x_{(n-1)}$	$r_{20} = (x_{(n)} - x_{(n-2)})/(x_{(n)} - x_{(1)})$
4-30	Lower pair $x_{(1)}, x_{(2)}$	r'_{20}=$(x_{(3)} - x_{(1)})/(x_{(n)} - x_{(1)})$
5-30	Upper pair $x_{(n)}, x_{(n-1)}$	$r_{21} = (x_{(n)} - x_{(n-2)})/(x_{(n)} - x_{(2)})$
5-30	Lower pair $x_{(1)}, x_{(2)}$	r'_{21}=$(x_{(3)} - x_{(1)})/(x_{(n-1)} - x_{(1)})$
6-30	Upper pair $x_{(n)}, x_{(n-1)}$	$r_{22} = (x_{(n)} - x_{(n-2)})/(x_{(n)} - x_{(3)})$
6-30	Lower pair $x_{(1)}, x_{(2)}$	r'_{22}=$(x_{(3)} - x_{(1)})/(x_{(n-2)} - x_{(1)})$

given significance α. (The r^* critical value can be found in Kanji (1993));

- Step 4: If the test statistic is less than the critical value r^*, there are no outliers present. However, if the test statistic is greater than the critical value, the null hypothesis is rejected and the conclusion is that the most extreme value is an outlier. The test is applied consecutively for other extreme values until the null hypothesis is true.

2.6.2 Outlier labeling methods

Outlier labeling methods use the interval for identification of outliers. The interval will separate outliers into 'good region' and 'bad region'. Bad region refers to the area outside the interval. Any observations that fall in the bad region are considered as outliers. Normally, outlier labeling methods are appropriate to use if one is only interested in finding an observation that is extremely different from the majority data. This method is not suit-

able to be applied if one wants to identify the observation that violates the distribution assumption of statistical analyses, such as regression.

Another reason for using the outlier labeling method is when we have a large data set. Note that it is difficult to identify the distribution of a large data set. Therefore, in this condition, the labeling method is appropriate for outlier detection rather than discordance tests.

(i). **Standard Deviation (SD) method**

The simple classical approach of the outlier labeling method is Standard Deviation (SD) method. Given a data set of n observations of a variable x, let $\bar{x}$ be the mean and let s be standard deviation of the data distribution. One observation is declared as an outlier if it lies outside of the interval

$$\bar{x} - k's, \bar{x} + k's \tag{2.2}$$

where the value k' is usually taken as 2 or 3.

The problem with the given criteria is the mean and standard deviation are highly sensitive to outliers.

(ii). **Boxplot**

One of the well known and widely used labeling methods is the Boxplot. The Boxplot was introduced by Tukey in 1977. Tukey introduced the Boxplot as a graphical display on which outliers can be indicated.

The observation that falls between the inner fence and outer fence, or beyond the outer fence is labeled as an outlier. The inner fence is calculated as

$$[Q_1 - 1.5IQR, Q_3 + 1.5IQR], \tag{2.3}$$

where $IQR = Q_3 - Q_1$ is the inter quartile range of the data set with Q_3 and Q_1 are the upper quartile of the data set and the lower quartile of the data set, respectively. One can compute the outer fence as

$$[Q_1 - 3IQR, Q_3 + 3IQR]. \tag{2.4}$$

Notice that the upper and lower quartiles, Q_3 and Q_1 are used to obtain the robust measures for mean, $(Q_1 + Q_3)/2$ and the standard deviation, $Q_3 - Q_1$, which can replace $\bar{x}$ and s in equation 2.2.

The Boxplot is applicable to skewed data since it makes no distributional assumptions and it does not depend on a mean or standard deviation. However it is not suitable for a small sample size and it is noted that the more skewed the data are, the more observations may be detected as outliers.

(iii). **Adjusted Boxplot**

As a solution to the Boxplot, Vanderviere and Hubert (2008) presented an adjusted Boxplot. The difference between the former and latter Boxplot is the inner and outer fence. In the adjusted Boxplot, the medcouple (MC) is introduced. The MC value is between -1 and 1. If $MC = 0$, the data is symmetric and the adjusted Boxplot becomes Tukey's Boxplot. In addition, if $MC > 0$, the data is right skewed; if $MC < 0$, the data is left skewed.

Let $X = x_1, x_2, \ldots, x_n$ be the independent sample of a continuous univariate distribution. Sort each observation in X, from the smallest value to the largest value, $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Therefore, one can define the MC as

$$MC(x_1, x_2, \ldots, x_n) = med\frac{(x_j - med') - (med' - x_i)}{x_j - x_i}, \tag{2.5}$$

where med'= the median of X, i and j have to satisfy $x_i \leq med' \leq x_j$ and $x_i \neq x_j$. If $MC \geq 0$, one can develop the fence as below

$$[Q_1 - (1.5IQR \times e^{-3.5MC}), Q_3 + (1.5IQR \times e^{4MC})].$$

However, if $MC \leq 0$, the fence becomes

$$[Q_1 - (1.5IQR \times e^{-4MC}), Q_3 + (1.5IQR \times e^{3.5MC})].$$

Observations situated outside the fence are labeled as outliers.

2.7 Multivariate methods

Outliers become more difficult to detect in high dimensional data. One cannot claim multivariable observations as outliers if each variable is considered independently. Another scenario that could happen in multivariate cases is the masking and swamping problem.

Recall that the masking problem occurs when the appearance of one outlier covers the appearance of another outlier, whereas the swamping problem arises when the observation is identified as an outlier even if it is not. In other words, swamping is the opposite of masking. Instead of declaring too few outliers, the method declares more outliers than there actually are (Hawkins, Bradu and Kass, 1984).

Some of the multivariate outliers have been modified from the univariate method, so that it can take into account a multivariable. Examples are the generalized distance with studentized residual (Siotani, 1959), the ratio of generalized distance with all observations (Wilk, 1963) and the W statistics for normality (Shapiro and Wilk, 1965).

There are also examples of multivariate outlier detection method that are based on residuals. Cook (1977) recommended using plot of residuals or examining the standardized residuals or studentized residuals. Other suggestions of multivariate outlier detection method that are based on residuals can be found in David (1978) and Cook (1986).

2.7.1 Statistical methods

Observations that are situated far from the centre of the data distribution is labeled as outliers in the statistical method. One of the most widely used approaches for the detection of multivariate outlier in the statistical method is called the Mahalanobis distance. According to Stevens (1984), the Mahalanobis distance is a measure of the distance in factor space.
Let

$\mathbf{x}$ = consisting of n observations and p variables

$\mathbf{X}$ = matrix of the original data set with column centred by the mean

$\bar{\mathbf{x}}$ = p dimensional vector with the means of each variable

$\mathbf{S} = \frac{1}{n-1}(\mathbf{X^T X})$, covariances matrix of the p variables

Now, one can develop the Mahalanobis distance, D

$$D(\mathbf{x}, \bar{\mathbf{x}}) = \{(\mathbf{x} - \bar{\mathbf{x}})^{\mathbf{T}} \mathbf{S}^{-\mathbf{1}} (\mathbf{x} - \bar{\mathbf{x}})\}^{\mathbf{1/2}} \tag{2.6}$$

where D is the distance of $\mathbf{x}$ to the mean of the data set. For multivariate normally distributed data, the values of the Mahalanobis distance are approximately chi-square distributed with p degrees of freedom (χ^2_p). An observation with large Mahalanobis distance can be considered as an outlier.

The Mahalanobis distance works well when identifying scattered outliers (Rocke and Woodruff, 1996). However, it fails to perform when a data set contains clustered outliers. This is supported by Filzmoser (2004), who mentions that a single extreme observation or a group of observations far away from the main data structures can have a significant influence on the Mahalanobis distance.

They are subject both to the masking and swamping effect because both estimators, i.e. mean and covariance, are usually estimated in a non-robust manner. Robust estimators mean they are less affected by outliers. Penny

and Jolliffe (2001) explain the scenario of the masking and swamping effects if the Mahalanobis distance is used for identification of outliers. In the situation of masking effects, a value of Mahalanobis distance for outliers will decrease as the outliers will pull $\bar{\mathbf{x}}$ and S towards themselves. In contrast, in the swamping effect, Mahalanobis distance values for non-outliers might increase since outliers attract $\bar{\mathbf{x}}$ and blow S away from the majority of observations.

For an illustration of the masking effect, consider the Hawkins Bradu Kass data set (Hawkins et al., 1984). This data set contains 75 observations with 3 variables. Note that the first 14 observations are outliers. Nevertheless, the Mahalanobis distance only flagged cases 12, 13 and 14 as outlier, (refer to Figure 2.2). The other outliers emerge only after the deletion of cases 12, 13 and 14 (Hawkins et al., 1984).

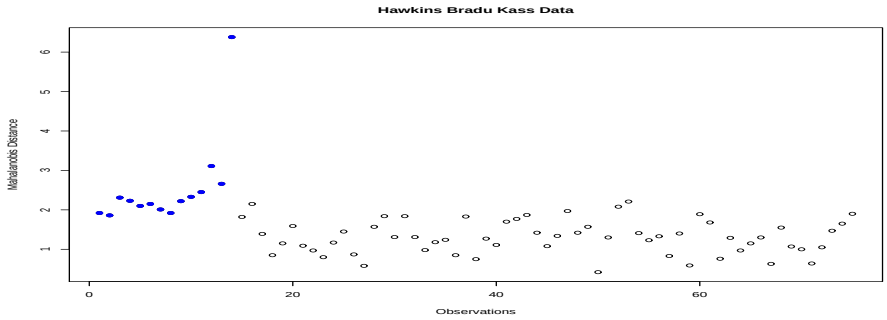


Figure 2.2: Mahalanobis distance plot for Hawkins Bradu Kass data

Wilk's statistics is also widely used for identification of outliers (Barnett and Lewis, 1994). It is equivalent to using the Mahalanobis distances of the n sample points, x_i from the sample mean, $\bar{x}$ (Caroni and Billor, 2007)

$$d_i(x_i, \bar{x}) = \{(x_i - \bar{x})^T S^{-1} (x_i - \bar{x})\}^{-1/2}.$$

Wilk's statistics is proposed by Wilk (1963) for identification of a single

outlier. Wilk's statistics is given by

$$R_i = \frac{|S_{(i)}|}{|S|} \sim B\Big(\frac{n-p-1}{2}, \frac{p}{2}\Big),$$

where $S_{(i)}$ is the covariance matrix of p variables when the ith row of matrix X is deleted, $X_{(i)}$. The subscript i in parentheses of $X_{(i)}$ is read as "with observation i is moved from X", i.e. the ith row of X is x_i^T then $X_{(i)}^T X_{(i)} = X^T X - x_i x_i^T$.

The sample points are ordered according to R_i. The outlier is the observation corresponding to minimum value of R_i. Since

$$\frac{|\hat{S}_{(i)}|}{\hat{S}} = 1 - \frac{n}{n-1}(x_i - \bar{x})^T \hat{S}^{-1}(x_i - \bar{x}),$$

minimization of R_i becomes equivalent to maximization of

$$(x_i - \bar{x})^T \hat{S}^{-1}(x_i - \bar{x}).$$

Wilk showed that the R_i are identically distributed with $B((n-p-1)/2, p/2)$ (but not independent). Figure 2.3 shows the index plot of $\{i, R_i\}$; notice that only observation 14 is located at the bottom of the plot.

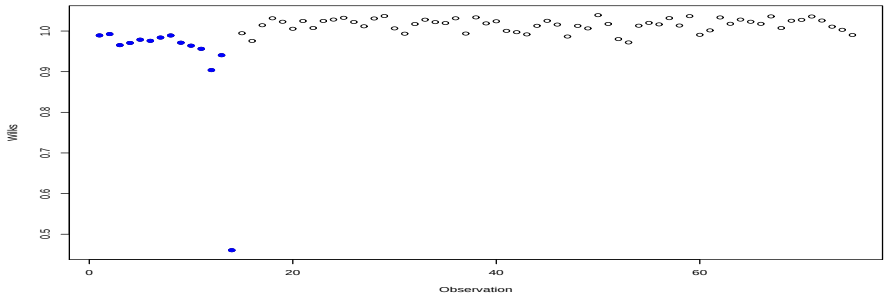


Figure 2.3: Wilks plot for Hawkins Bradu Kass data

Again, an obvious problem is that of masking. Suppose there are actually more outliers than the number being tested for, then the covariance

matrix will be inflated by these extra outliers. Hence, minimizing the distance, d_i makes it less likely that outliers will be declared. This consideration makes it desirable to consider a robust method of identifying outliers. This will be further discussed in the next section.

Now, consider Wilk's criterion for detecting I outlier is

$$R_I = \frac{|S_I|}{|S|},$$

where $S_{(I)}$ is the covariance matrix of p variables with set of observations I are removed from the data matrix" and $I = \{i_1, i_2, \ldots, i_m\}$ is the subset indices $1 \leq i_j \leq n$, $j = 1, 2, \ldots, m$. The detection of m joint outlier is based upon $\min_{I_m \subset n} R_{I_m}$. It is noted that the calculation of minimum R_I is more unfeasible if n and m are larger.

2.7.2 Multivariate robust measures

As a consequence of the Mahalanobis distance and Wilk's statistics problem in the statistical methods, many robust means and covariances have been introduced in previous studies. Examples are minimum volume ellipsoid (MVE) estimators (Rousseeuw and von Zomeren, 1990) and minimum covariance determinant (MCD) estimators by Rousseeuw and Driessen (1999). These estimators have the desirable properties of high breakdown point and affine equivariance.

Originally, the breakdown point definition was given by Hodges (1967), where the definition is limited to a one-dimensional estimation of location. Nevertheless, Hampel (1971) proposed a much more general formulation. The breakdown point is a percentage of outliers which will make the estimator take on the large values. Therefore, estimators with a large breakdown point are more robust. It is noted that the highest breakdown point value can possibly reach 50%. If the value goes beyond 50%, one cannot decide which data are outliers and which are from the main distribution.

Another desirable property of an estimator is affine equivariance. A location estimator $T_n \in \Re^p$ is affine equivariant if and only if for any vector $b \in \Re^p$ and any nonsingular $p \times p$ matrix A,

$$T_n(AX + b) = AT_n(X) + b.$$

A scale estimator $C_n \in PDS(p)$ (the set of positive-definite symmetric $p \times p$ matrices) is affine equivariant if and only if for any vector $b \in \Re^p$ and any nonsingular $p \times p$ matrix A,

$$C_n(AX + b) = AC_n(X)A^T.$$

If an estimator is affine equivariant, stretching or rotating the data won't affect the estimator.

Nevertheless, it is noted that the multivariate robust measures suffer from computational complexity, i.e. the efficiency of algorithms as run time and memory requirement permit.

(i). **M-estimator**

M-estimator is an early version of robust estimators, which are developed by a simple adjustment of the classical estimators. Maronna (1976) studied affinely equivariant M-estimators for covariance matrices and Campbell (1980) proposed using the Mahalanobis distance computed using the M-estimators for the mean and covariance matrix. To compute these estimators, each observation is given a weight. The given weight depends on the $d_i(x_i, \bar{x})$ values of each observation. Observation with a high value of $d_i(x_i, \bar{x})$ will be down weighted. Full weight is given to the observations with normal $d_i(x_i, \bar{x})$ value.

Note that the observations with the large value of $d_i(x_i, \bar{x})$ could be considered as outliers. Therefore by giving a reduced weight to the outlying observation in the data set, it hardly influences the estimator. However, the M-estimator has a low breakdown point, which is

$\frac{1}{p+1}$. It means the performance of these estimators is not consistent. Considering the M-estimator has a low breakdown point, a different approach have been proposed to overcome the difficulty.

(ii). **Minimum Volume Ellipsoide (MVE) estimator**

Minimum volume ellipsoid estimators are the mean and covariance matrix of subsample size h, where $h \leq n$. It minimizes the volume of the covariance matrix associated with the subsample. The basic idea of the MVE is to search among all such ellipsoids for the one having the smallest value. Therefore, the main problem of MVE is to find h that produces the smallest ellipse because the number of all subsamples containing half of the data is so large that determining the subsample with the minimum volume is impractical. It is noted that h is taken to be $h = (n + p + 1)/2$ which is the integer function. The h value can be assumed as the minimum number of instances that must not be outlying. Otherwise, one can state this approach has a breakdown point of approximately 50%.

(iii). **Minimum Covariance Determinant (MCD) estimator**

The minimum covariance determinant (MCD) estimator also has a breakdown point of approximately 50%. The MCD estimator is the mean and covariance of a subsample of size h $(h \leq n)$ that minimizes the determinant of the covariance matrix that corresponds to the subsample. As with MVE, it is impractical to consider all subsets of half of the data since it is computationally expensive.

(iv). **Application of multivariate robust measures**

Rousseeuw and von Zomeren (1990) used the MVE estimators to develop a method for outlier detection. The method was based on the basic resampling algorithm and they named it the Robust Distance method. However, Hadi (1992) pointed out three weaknesses of this

method, particularly a problem related to the situation when the covariance matrix has a zero determinant. Therefore, he solved this weakness by presenting an idea that makes outliers appear in one subset, with the other subset highly unlikely to contain outliers. The new approach still applies the MVE estimator, but it is easier to compute and the method is not dependent on the basic resampling algorithm. Later, Hadi (1994) modified his idea by giving an alternative step to the existing algorithm. The findings of this approach were almost similar to the findings of the previous solution in 1992.

The minimum covariance determinant (MCD) estimator had been used by Hawkins in 1994 to develop a feasible solution algorithm (FSA) to discover outliers. This approach still uses the subset to divide a data set from outliers. The disadvantage of this method is that large number of subsets need to be constructed from a data set, especially when one has a data set with a large sample size and variables. Therefore, in order to solve this problem, Rousseeuw and Driessen (1999) suggested the fast algorithm using the MCD estimator called FAST-MCD.

They introduced two techniques, which are, the selective iteration and the nested extension. They also presented *C-Step* where the *'C'* means concentration. The word concentration could be interpreted as their focus on h observations with least distances. It also could be described as the most recent chosen subset that provides a minimum determinant. The *C-Step* has four steps which are repeated until the last process fulfils the latter definition of *'C'*.

Hawkins and Olive (1999) also tried to improve the *FSA* by adding a condition called *C-Condition*. However, their approach still retained the similar computational complexity as *FSA* since it is only reduces the computation time for studies that use the fixed sample size, i.e. a subset with the same sample size.

2.7.3 Eigenstructure Approach

The eigenstructure of a data matrix plays an important role in some methods of statistical analyses. For example, in regression analyses, if the covariance matrix provides small eigenvalues, this is an indication of the presence of multicollinearity. Wang and Nyquist (1991) state that individual observations can highly influence the eigenstructure of a data matrix. They investigated the influence of each observation by comparing the eigenstructure of a completed data set and the eigenstructure of a data set without the observation under consideration. Their study summarized the properties of exact principal component eigenvalues influence measures based on a numerical point of view.

Next, Wang and Liski (1993) extended the study to the influence of a set of observations on the eigenstructure of a data matrix if they are removed from a data set. In this study, comparisons are made with results of single deleted observation cases from Critchley (1985).

Later, Mertens (1998) used the rice data to present the exact principal component eigenvalues influence measures from an applied statistician's perspective. Mertens (1998) developed a principal component from two types of eigenstructure and showed that the normalization of eigenvalues could create a distance of an observation. According to Mertens (1998), the analyses of eigenvalues is very important for high dimensional data sets.

In 2005, Gao, Li and Wang proposed a method for identification of outlier based on eigenstructure. Their method is called the Max-Eigen difference (MED).

$$\text{MED}_i = \frac{d_i'}{\sum_{j=1}^{n} d_j'} \tag{2.7}$$

where

$$d_i' = \| \lambda_1^{(i)} \upsilon_1^{(i)} - \lambda_1 \upsilon_1 \| (1 - \Pi_{j=1}^{p} I'_{(y_{ij}^2 < \lambda_j)})$$

and $\| \cdot \|$ represent the euclidean norm, $I'_{\{.\}}$ is an indicator function, $y_{ij} = (x_i - \bar{x})^T \upsilon_j$. $\lambda^{(i)}$ and $\upsilon^{(i)}$ are eigenvalues and eigenvectors, respectively,

calculated from covariance matrix of the data set, X with p dimension where the ith observation has been removed from it.

The function of $1 - \Pi_{j=1}^{p} I'_{(y_{ij}^2 < \lambda_j)}$ is to let the MED$_i$ become zero if all y_{ik}^2 is less than the corresponding λ_k where $k = 1, 2, \ldots, p$. This is because if x_is are close to the mean, $\bar{x}$, they should not be identified as outliers and their proportion with $y_{ik}^2 < \lambda_k$ for all k is not large if all observations x_i are identically and independently distributed with normal distribution (Gao, Li and Wang, 2005).

2.7.4 Angles

Instead of measuring two population using distances, one can also obtain an angle between them subtended at the origin. Note that the angle and the distance are using similar concepts, therefore the angle can be called the Mahalanobis angle. Fisher (1936) seems to be the first to use the concept of the Mahalanobis angle.

Let x_1 and x_2 be two independent random vectors with $E(x_i) = \mu_i$ and $\text{cov}(x_i) = \Sigma$ where $i = 1, 2$. If Σ is positive definite matrix, the Mahalanobis distance between the two populations with the random vectors x_1 and x_2 is defined as $D(\mu_1, \mu_2) = \{(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)\}^{1/2}$. In the same spirit, Mardia, 1977, defined the Mahalanobis angle as an angle between μ_1 and μ_2 subtended at the origin.

Angles also can be used for the identification of outliers. Juan and Prieto (2001) proposed a technique for the identification of outlier based on the analyses of certain angular properties of the observations. This technique is able to manage the data set with concentrated contamination that cannot be handled by methods developed from robust estimators. Angles for each observation are developed between the reference direction (refer to the equation (5) in Juan and Prieto (2001)) and the normalized distance. The Q-Q plot, i.e. a plot of quartiles and ordered angles is used to exhibit the outlier.

Later, Kriegel, Schubert and Zimek (2008) proposed a method called ABOD, i.e. Angle-Based Outlier Detection and some variants assessing the variance in the angles between the different vectors of an observation. The main advantage of this approach is it does not depend on any parameter selection influencing the quality of the achieved ranking. According to them, the relative contrast of the farthest observation and the nearest observation converges to 0 for increasing dimensionality, p as following:

$$\lim_{d\to\infty} \frac{dist_{max} - dist_{min}}{dist_{min}} \to 0$$

2.7.5 Data mining methods

Data mining methods can be considered as a non-parametric approach. There are a few examples of outlier detection methods in this category, i.e. clustering based-methods and distance-based methods.

(i). **Clustering based-methods**

Clustering based-methods will classify observations into clusters. Outlier is discovered as observation in a small cluster. There are many types of clustering methods. One of them is called partitioning around medoids (PAM). This approach was introduced by Kaufman and Rousseeuw (1990). Normally PAM is performed on small data sets. Therefore, Kaufman and Rousseeuw (1990) introduced a new approach called clustering large applications (CLARA).

Note that CLARA and PAM use the same algorithm, with CLARA merely utilizing the algorithm on the multisamples, i.e. the large data set dividing the observations into multisamples. The clustering based-methods are not always optimal for identification of outlier since their main purpose is clustering.

(ii). **Distance based-methods**

The idea of distance-based outlier was originally proposed by Knorr and Ng (1998). They defined a distance-based outlier as

Definition 2.7.1 *An observation O in a data set T is a $DB(p', D)$ outlier if at least fraction p' of the observations in T lie at a greater distance than D from O,*

where $DB(p', D)$ is a notation for a 'distance-based outlier', and outliers are detected using parameter p' and D. According to Knorr and Ng (1998), there are no criteria to choose p' and D; it is left to a human expert.

This approach is free from any distribution assumptions and observations with a large distance are possibly identified as outliers. Following the definition-based outlier, many methods have been proposed to detect distance-based outliers (see, Knorr and Ng, 1998; Ramaswamy et al., 2000; Angiulli and Pizzuti, 2002; Bay and Schwabacher, 2003). However, all of these methods can only identify the global outliers. It is noted that outliers can be classified into two groups: global outliers and local outliers.

A global outlier is an extreme observation with respect to every other observation in the data set, whereas a local outlier refers to an observation that is isolated from its surrounding neighbourhood rather than the whole data set. Therefore, as a solution to this problem, Papadimitriou, Kitawaga, Gibbons and Faloutsos (2003) and Breunig, Kriegel, Ng and Sander (2000) introduced the local correction integral (LOCI) and local outlier factor (LOF), respectively, to find the local outlier.

2.8 Solution to outliers

Normally when one finds outliers, there are a few ways to handle them. If the existence of outliers are only due to measurement error, this can be corrected. However, if the appearance is caused by an implementation error, the outliers probably should be removed. If the outlier emerges due to inherent variability, then it should remain. This is because one might use it as a new outcome that can lead to a new research.

A story about the 'ozone hole' above Antarctica gives a very good lesson about why outliers need to be considered rather than deleted from a dataset, `http://exploringdata.cqu.edu.au/ozone.htm`. It is based on three researchers who queried a data set provided by the British Antarctic Survey in 1985. The data set showed ozone levels not being normal; however, the Nimbus 7 satellite presented contrast findings. After doing an inspection, they found the solution to this problem. The different findings occurred because the computer programme that was used to record the ozone levels from the satellite had assumed that the low concentration levels were outliers and they were removed. The effect of the deleted outliers not being examined resulted in no one realizing the world atmosphere has been damaged for nine years (since 1976).

2.9 Conclusion

This chapter briefly presents a simple introduction about outliers and some methods for identifying outliers. As a conclusion, outliers can be defined simply as an observation (or a subset of observations) that is isolated from the other observations in the data set. There are two major reasons that motivate people to find outlier; first is the researcher intentions. The second reason is their effect on the analyses, i.e. the existence of outliers will affect means, variances and regression coefficients; they will also cause a bias or distortion of estimates; likewise, they will inflate the sums of

squares and certainly create false conclusions.

Outliers may exist because of inherent variability, measurement error and execution error. There have been many methods developed for the identification of outliers. They can be classified into the univariate method and the multivariate method (see Hawkins, 1980; Barnett and Lewis, 1994). The univariate method is performed independently on each variable, whereas the multivariate method investigates the relationship of several variables (Franklin et al., 2000). One can also classify them into a statistics approach, i.e. parametric and non-parametric approach.

§2.7 briefly explained the role of eigenstructure in some methods of statistical analyses and for identification of outliers. The approach looks easy to apply as a tool for identification of outliers. Therefore, Chapter 3 will use a few techniques based on eigenstructure as a tool for the identification of outliers.

Chapter 3

Outliers Identification by eigenstructure

3.1 Introduction

As discussed in the previous chapter, the covariance matrix is a very important tool in multivariate statistics. Suppose $\mathbf{X}$ is an $n \times p$ data matrix consisting of n observations on p variables and , therefore $\mathbf{X}$ can be written as follows:

$$\mathbf{X}_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}.$$

In practice, the sample covariance matrix is written as

$$\begin{aligned} \mathbf{S} &= \frac{1}{n-1}(\mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{X})^T(\mathbf{X} - \frac{1}{n}\,\mathbf{1}\mathbf{1}^T\mathbf{X}) \\ &= \frac{1}{n-1}\mathbf{X}^T(\mathbf{I} - \frac{1}{n}\,\mathbf{1}\mathbf{1}^T)\mathbf{X}, \end{aligned} \tag{3.1}$$

where

$$\frac{1}{n}\,\mathbf{1}\mathbf{1}^T\mathbf{X},$$

is the $n \times p$ matrix of means of the matrix $\mathbf{X}$ with $\mathbf{1}$ is n-vector of ones (Johnson and Wichern, 2007) and $\mathbf{I}$ be the $n \times n$ identity matrix. The purpose of the means is to allow the entries to be centred. If the means are subtracted out, the sample covariance matrix can be computed as

$$\mathbf{S} = n^{-1}(\mathbf{X}^T\mathbf{X}).$$

However, if one does not bother about dividing $\mathbf{S}$ by n, then

$$\mathbf{S} = \mathbf{X}^T\mathbf{X},$$

can be called a covariance matrix.

Normally, the covariance matrix will have p variances and $\frac{1}{2}p(p-1)$ different covariances (Johnson and Wichern, 2007).

$$\mathbf{S} \;=\; \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \dots & s_{pp} \end{bmatrix}, \tag{3.2}$$

where $s_{jk} = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ and $j, k = 1, 2, \dots, p$. Note that the denominator can also be replaced as n or 1 (Johnson and Wichern, 2007).

This chapter focuses on the identification of outliers using the eigenstructure of $\mathbf{S}$ and $\mathbf{S}_{(i)}$ in terms of eigenvalues, eigenvectors and principal components. Note that $\mathbf{S}_{(i)}$ is the sample covariance matrix of data matrix $\mathbf{X}_{(i)}$, where the subscript i in parentheses is read as "with observation i removed from $\mathbf{X}$".

The idea of using the eigenstructure as a tool for identification of outliers is motivated by Maximum Eigen Difference (MED). This method uti-

lizes the maximum eigenvalue and the corresponding eigenvector. It is noted that examination of the observations effect on the maximum eigenvalue is very significant. The reason is that outliers that lie in the direction close to the maximum eigenvalue or vice versa, will change the maximum eigenvalue (Gao et al., 2005). The maximum eigenvalue contains maximum variance, therefore, the outliers detected by the maximum eigenvalue have a greater effect on variance, and they need extra attention.

The definition and computation of eigenstructure are briefly explained in §3.2. Then in §3.3, the influence eigenstructure of covariance matrix will be discussed and it will be used to develop techniques for identifying outliers.

The main part of this chapter considers techniques as well as the index plot, mainly as a graphical tool for diagnostics, i.e. identification of outliers. These areas will be discussed in §3.4 and §3.5. This chapter discusses the problem of outliers without assuming any model distribution. Note that in practice, mean and variance or covariance are unknown and the data will often not have a multivariate normal distribution. Therefore, any distributional result derived under the restrictive assumption can only be approximations (Jackson, 1991; Jolliffe, 2002).

In previous studies, finding outliers is about identifying observations that are obviously different from others (Penny and Jolliffe, 2001), and there is no motivation to compute significance levels very accurately since an observation that is barely significant at 5%, typically is of no interest.

The techniques discussed in this chapter use the maximum eigenvalue with the corresponding eigenvector. It is noted that the first few eigenvalues are the most interesting for multivariate data because the first ones are sensitive to the outliers, as they could inflate the variance and covariance (Jolliffe, 2002).

An example is that eigenvalues can be referred to when examining the multicollinearity, i.e. correlation between independent variables in the model. The square root of the maximum eigenvalue, λ_{max} divided by the

smallest eigenvalues, λ_{min} represents the condition index (κ),

$$\kappa = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}.$$

The condition indices are used to identify whether there is a multicollinearity problem or not. If $\kappa = 1$, hence collinearity problems will not appear. However, as collinearity increases, eigenvalues will either be greater or smaller than 1. Eigenvalues close to zero indicate a multicollinearity problem and the condition indices will increase.

The behaviour of the maximum eigenvalue of a sample covariance matrix as a random object has been studied by Bejan (2005). The choice of the maximum eigenvalue as the object of interest is motivated by its importance in many techniques of multivariate statistics, for example, principal component analysis and the possibility of its use in statistics as test statistics (Bejan, 2005).

In particular, this chapter defines four techniques: $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$. The techniques discussed in this chapter are not entirely novel; they have been discussed earlier, i.e. influence eigen has been considered by Wang and Nyquist (1991) from a numerical point of view. Studies cited in this chapter are largely empirical and this chapter shows through examples how the suggested techniques perform in data sets containing outliers.

To investigate how the techniques perform, two types of data will be considered in this chapter. First is the simulated data set and second is the real data set. The simulated and real data sets used in this chapter are described in §3.6 and §3.7 respectively.

Next, the performance of the suggested techniques can be seen in §3.9, §3.10, §3.11 and §3.12. §3.9 considers low dimension and small sample size data. §3.10 studies those techniques on the low dimension but with a large sample size data, whereas §3.11 examines the performance of the techniques on high dimension with a large sample size data. The performance

of those techniques on a real data set can be found in §3.12.

3.2 Eigenvalues and eigenvectors

Definition 3.2.1 *(Johnson and Wichern 2007) Consider* $\boldsymbol{S}$ *is a* $p \times p$ *square matrix and* $\boldsymbol{I}$ *be the* $p \times p$ *identity matrix. Therefore, the scalars* $\lambda_1, \lambda_2, \ldots, \lambda_p$ *that satisfy the polynomial equation* $|\boldsymbol{S} - \lambda \boldsymbol{I}| = 0$ *are called the eigenvalues of matrix* $\boldsymbol{S}$ *and* $|\boldsymbol{S} - \lambda \boldsymbol{I}| = 0$ *is called characteristic equation.*

Definition 3.2.2 *(Johnson and Wichern 2007) Let* $\boldsymbol{S}$ *be a square matrix of dimension* $p \times p$ *and let* λ *be an eigenvalue of* $\boldsymbol{S}$*. If* υ *is a nonzero vector* $(\upsilon \neq 0)$ *such that*

$$\boldsymbol{S}\upsilon = \lambda\upsilon,$$

then υ *is said to be an eigenvector of the matrix* $\boldsymbol{S}$ *associated with the eigenvalue,*λ*.*

Let $\mathbf{S} = \mathbf{X}^T\mathbf{X}$, where $\mathbf{S}$ is a square $p \times p$ matrix. Let λ_j and υ_j, $j = 1, 2, \ldots, p$ be the eigenvalues and the corresponding normalized eigenvectors of $\mathbf{S}$ respectively.

According to Graybill (1976), there is an orthogonal matrix

$$\mathbf{V} = [\upsilon_1, \upsilon_2, \ldots, \upsilon_p], \tag{3.3}$$

such that

$$\begin{aligned} \mathbf{V}^T\mathbf{S}\mathbf{V} &= \mathbf{V}^T(\mathbf{X}^T\mathbf{X})\mathbf{V} \\ &= \text{diag}[\lambda_1, \lambda_2, \ldots, \lambda_p] \\ &= \mathbf{\Lambda}. \end{aligned} \tag{3.4}$$

3.3 Influence eigenvalues and eigenvectors

Some statistical methods are concerned with eigenstructure problems and a few statistics are the functions of eigenvalues in multivariate analysis. A test statistic is considered as a function of eigenvalues of a transition matrix to test a Markov chain for independence (Wang and Scott, 1989) and eigenstructure methods are applied to study the co-linear problem in multivariate linear regression (Wang and Nyquist, 1991).

Now, consider the influence of eigenvalues λ_j and eigenvectors v_j for matrix $\mathbf{X}^T\mathbf{X}$ where $\mathbf{X}$ is an $n \times p$ observation matrix consisting of n observations for p variables.

If ith row of matrix $\mathbf{X}$ is deleted, one can write it as $\mathbf{X}_{(i)}$ where the subscript i in parentheses is read as "with observation i is removed from $\mathbf{X}$", i.e. the ith row of $\mathbf{X}$ is x_i^T then $\mathbf{X}_{(i)}^T\mathbf{X}_{(i)} = \mathbf{X}^T\mathbf{X} - x_i x_i^T$. Let $\mathbf{X}^T\mathbf{X}$ have the eigenvalues-eigenvectors pairs

$$(\lambda_1, v_1), (\lambda_2, v_2), ..., (\lambda_p, v_p),$$

and the eigenvalues are in descending order

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p, \tag{3.5}$$

and let $\mathbf{X}_{(i)}^T\mathbf{X}_{(i)}$ have the eigenvalues and eigenvectors pairs

$$(\lambda_{1(i)}, v_{1(i)}), (\lambda_{2(i)}, v_{2(i)}), ..., (\lambda_{p(i)}, v_{p(i)}),$$

and the eigenvalues are also in descending order

$$\lambda_{1(i)} \geq \lambda_{2(i)} \geq ... \geq \lambda_{p(i)}. \tag{3.6}$$

Define,

$$\mathbf{V}_{(i)} = [v_{1(i)}, v_{2(i)}, \ldots, v_{p(i)}], \tag{3.7}$$

and

$$\begin{aligned}\mathbf{V}_{(i)}^T\mathbf{S}_{(i)}\mathbf{V}_{(i)} &= \mathbf{V}_{(i)}^T(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})\mathbf{V}_{(i)} \\ &= \text{diag}[\lambda_{1(i)}, \lambda_{2(i)}, \ldots, \lambda_{p(i)}] \\ &= \mathbf{\Lambda_{(i)}}. \end{aligned} \tag{3.8}$$

Then influence functions of eigenvalues λ_j and eigenvectors v_j are given respectively by Radhakrishnan and Kshirsagar (1981) as follows:

$$IF(x; \lambda_j) = (x^T v_j)^2 - \lambda_j \tag{3.9}$$

and

$$IF(x; v_j) = -x^T v_j \sum_{k \neq j} x^T v_k (\lambda_k - \lambda_j)^{-1} v_k. \tag{3.10}$$

If one wishes to examine the ith observation's influence on the eigenvalues and eigenvectors of $\mathbf{X}^T\mathbf{X}$, it is easy to remove the ith observation from the full data set and then compare the eigenvalues and eigenvectors of the remaining data with that of the complete data.

Lemma 3.3.1 *The properties of eigenvalues and eigenvectors are given as follows:*

(i). $\lambda_j \geq \lambda_{j(i)}$;

(ii). The relationship of eigenvalues λ_j *and* $\lambda_{j(i)}$ *is given by Gao et al. (2005):*

$$\begin{aligned}\lambda_{j(i)} = \lambda_j - \frac{1}{n-1}(l_{ij}^2 - \lambda_j) - \\ \frac{1}{2(n-1)^2} l_{ij}^2 \Big[1 + \sum_{k \neq j} \frac{l_{ij}^2}{\lambda_k - \lambda_j}\Big] + O(\frac{1}{n^3}),\end{aligned} \tag{3.11}$$

where $l_{ij} = (x_i - \bar{x})^T v_j$;

*(iii). The relationship between eigenvectors of v_j and $v_{j(i)}$ is obtained based on the observation matrix **X** given by Gao et al. (2005) as follows:*

$$v_{j(i)} = v_j + \frac{l_{ij}}{n-1}\sum_{k\neq j}\frac{l_{ik}v_k}{\lambda_k-\lambda_j} - \frac{1}{2(n-1)^2}\sum_{k\neq j}\Big[\frac{l_{ij}^2 l_{ik}^2 v_j}{(\lambda_k-\lambda_j)^2} - \frac{2l_{ik}^2 l_{ij}}{(\lambda_k-\lambda_j)}\sum_{k\neq j}\frac{l_{ik}v_k}{\lambda_k-\lambda_j} + \frac{2l_{ij}^3 l_{ik}v_k}{(\lambda_k-\lambda_j)^2}\Big] + O(\frac{1}{n^3}). \quad (3.12)$$

Proof:

(i) $\lambda_j \geq \lambda_{j(i)}$ is obtained from the following matrix operations: It is noted that

$$\mathbf{X}^T\mathbf{X} = \mathbf{X}_{(i)}^T\mathbf{X}_{(i)} + x_i x_i^T,$$

where $\mathbf{X}^T\mathbf{X}$, $\mathbf{X}_{(i)}^T\mathbf{X}_{(i)}$ and $x_i x_i^T$ are symmetric matrices and $x_i x_i^T$ is of rank unity, there exists on an orthogonal matrix **Q** such that

$$\mathbf{Q}^T(x_i x_i^T)\mathbf{Q} = \begin{pmatrix} s & 0 \\ 0 & 0 \end{pmatrix},$$

where s is the unique non-zero eigenvalues of $x_i x_i^T$, and consider

$$\mathbf{Q}^T(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})\mathbf{Q} = \begin{pmatrix} t & c^T \\ c & \mathbf{X}_*^T\mathbf{X}_* \end{pmatrix},$$

then there is an orthogonal matrix $\mathbf{P}_{(k-1)(k-1)}$ so that

$$\mathbf{P}^T(\mathbf{X}_*^T\mathbf{X}_*)\mathbf{P} = \Lambda_* = diag\{\lambda_1, \lambda_2, ...\lambda_{k-1}\}$$

and one can define an orthogonal matrix

$$\mathbf{G} = \mathbf{Q}\begin{pmatrix} 1 & 0 \\ 0 & \mathbf{P} \end{pmatrix},$$

then

$$\begin{aligned}\mathbf{G}^T(\mathbf{X}^T\mathbf{X})\mathbf{G} &= \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{P}^T \end{pmatrix} \mathbf{Q}^T(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})\mathbf{Q} \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{P} \end{pmatrix} \\ &+ \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{P}^T \end{pmatrix} \mathbf{Q}^T(x_i x_i^T)\mathbf{Q} \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{P} \end{pmatrix} \\ &= \begin{pmatrix} t+s & c^T\mathbf{P} \\ \mathbf{P}^T c & \Lambda_* \end{pmatrix},\end{aligned}$$

where

$$\begin{aligned}\sum_{j=1}^{k}\lambda_j &= t+s+\sum_{i=1}^{k-1}\lambda_i \\ &= t+\sum_{i=1}^{k-1}\lambda_i + s \\ &= \sum_{j=1}\lambda_{j(i)} + s. \end{aligned} \tag{3.13}$$

Note that $s \geq o$, and $\lambda_j \geq \lambda_{j(i)}$ is obtained for any $i = 1, 2, ..., n$. □

3.4 Influence eigen for identification of outliers

3.4.1 Influence eigen

Let the sample covariance matrix be

$$\mathbf{S} = \frac{1}{n}\mathbf{X}^T(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{X}, \tag{3.14}$$

where $\mathbf{1}$ is the n-vector of ones and $\mathbf{I}_n$ is the identity matrix of $n \times n$. If $\mathbf{X}_{(I)}$ and $\mathbf{S}_{(I)}$ are the data matrix and sample covariance matrix, respectively, when the m observations are deleted and the subscript I in parentheses is read as "with a set of m observations I removed from $\mathbf{X}$", note that

$I = \{i_1, i_2, \ldots, i_m\}$ where $1 \leq i_j \leq n$ and $j = 1, 2, \ldots, m$. Therefore, one has

$$\mathbf{S}_{(I)} = \frac{1}{n-m}\mathbf{X}_{(I)}^T(\mathbf{I}_{n-m} - \frac{1}{n-m}\mathbf{1}_{n-m}\mathbf{1}_{n-m}^T)\mathbf{X}_{(I)} \tag{3.15}$$

and

$$\mathbf{S}_I = \frac{1}{m}\mathbf{X}_I^T(\mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^T)\mathbf{X}_I. \tag{3.16}$$

Lemma 3.4.1 *It is noted that*

(i). The relationship among $\boldsymbol{S}$, $\boldsymbol{S}_I$ *and* $\boldsymbol{S}_{(I)}$ *is given as follows:*

$$\boldsymbol{S}_{(I)} = \tfrac{n}{n-m}\boldsymbol{S} - \tfrac{nm}{(n-m)^2}[\tfrac{n-m}{n}\boldsymbol{S}_I + (\bar{x}_I - \bar{x})(\bar{x}_I - \bar{x})^T];$$

(ii). If let $I = \{i\}$ *with a single observation, then*

$$\boldsymbol{S}_{(i)} = \tfrac{n}{n-1}\boldsymbol{S} - \tfrac{n}{(n-1)^2}(x_i - \bar{x})(x_i - \bar{x})^T.$$

Proof:

(i) Supposing that equations 3.14-3.16 are biased estimates, they can be used to developed unbiased estimates as in lemma 3.4.1.

$$\begin{aligned}
(n-m)\mathbf{S}_{(I)} &= X_{(I)}^T\Big(I_{n-m} - \frac{\mathbf{1}}{n-m}\mathbf{1}_{n-m}\mathbf{1}_{n-m}^T\Big)\mathbf{X}_{(I)} \\
&= \mathbf{X}_{(I)}^T\mathbf{X}_{(I)} - \frac{\mathbf{1}}{n-m}\mathbf{X}_{(I)}^T\mathbf{1}_{n-m}^T\mathbf{1}_{n-m}\mathbf{X}_{(I)} \\
&= \mathbf{X}^T\mathbf{X} - \mathbf{X}_I^T\mathbf{X}_I - \frac{\mathbf{1}}{n-m}\Big(\mathbf{X}^T\mathbf{1}_n - \mathbf{X}_I^T\mathbf{1}_m\Big)\Big(\mathbf{X}^T\mathbf{1}_n - \mathbf{X}_I^T\mathbf{1}_m\Big)^T \\
&= \mathbf{X}^T\mathbf{X} - \frac{\mathbf{X}^T\mathbf{1}_n\mathbf{1}_n^T\mathbf{X}}{n} - \frac{m\mathbf{X}^T\mathbf{1}_n\mathbf{1}_n^T\mathbf{X}}{n(n-m)} + \\
&\quad \frac{\mathbf{1}}{n-m}\Big(\mathbf{X}^T\mathbf{1}_n\mathbf{1}_m^T\mathbf{X}_I + \mathbf{X}_I^T\mathbf{1}_m\mathbf{1}_n^T\mathbf{X} - \mathbf{X}_I^T\mathbf{1}_m\mathbf{1}_m^T\mathbf{X}_I\Big) - \mathbf{X}_I^T\mathbf{X}_I \\
&= n\mathbf{S} - \frac{nm}{n-m}\Big(\bar{x}\bar{x}^T - \bar{x}_I\bar{x}^T - \bar{x}\bar{x}_I^T + \bar{x}_I\bar{x}_I^T\Big) + m\bar{x}_I\bar{x}_I^T - \mathbf{X}_I^T\mathbf{X}_I \\
&= n\mathbf{S} - \frac{nm}{n-m}(\bar{x} - \bar{x}_I)(\bar{x} - \bar{x}_I)^T + m\bar{x}_I\bar{x}_I^T - \mathbf{X}_I^T\mathbf{X}_I
\end{aligned} \tag{3.17}$$

Now simplify equation 3.17 as follows:

$$\begin{aligned}(n-m)\mathbf{S}_{(I)} &= n\mathbf{S} - \frac{nm}{n-m}(\bar{x}-\bar{x}_I)(\bar{x}-\bar{x}_I)^T - m\Big(\frac{\mathbf{X}_I^T\mathbf{X}_I}{m} - \bar{x}_I\bar{x}_I^T\Big)\\ &= n\mathbf{S} - \frac{nm}{n-m}(\bar{x}-\bar{x}_I)(\bar{x}-\bar{x}_I)^T - m\Big(\frac{\mathbf{X}_I^T\mathbf{X}_I}{m} - \frac{\mathbf{X}_I^T\mathbf{1}_m\mathbf{1}_m^T\mathbf{X}_I}{m^2}\Big)\\ &= n\mathbf{S} - \frac{nm}{n-m}(\bar{x}-\bar{x}_I)(\bar{x}-\bar{x}_I)^T - m\Big(\frac{1}{m}\mathbf{X}_I^T(\mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^T)\mathbf{X}_I\Big)\\ &= n\mathbf{S} - \frac{nm}{n-m}(\bar{x}-\bar{x}_I)(\bar{x}-\bar{x}_I)^T - m\mathbf{S}_I \qquad (3.18)\end{aligned}$$

(ii) By using equation 3.18, one can get the relationship between $\mathbf{S}$, $\mathbf{S}_I$ and $\mathbf{S}_{(I)}$ in (i) where $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ and $\bar{x}_I = \frac{\sum_{i\subset I} x_i}{m}$ represent the mean vector of all observations and the mean vector of the observations indexed by I respectively. Next, replace $m = 1$ in the following equation

$$(n-1)\mathbf{S}_{(I)} = n\mathbf{S} - \frac{nm}{n-m}(\bar{x}-\bar{x}_I)(\bar{x}-\bar{x}_I)^T - m\Big(\frac{\mathbf{X}_I^T\mathbf{X}_I}{m} - \frac{\mathbf{X}_I^T\mathbf{1}_m\mathbf{1}_m^T\mathbf{X}_I}{m^2}\Big)$$

hence one can find equation (ii) in lemma 3.4.1 as following

$$\begin{aligned}(n-1)\mathbf{S}_{(i)} &= n\mathbf{S} - \frac{n}{n-1}(\bar{x}-\bar{x}_i)(\bar{x}-\bar{x}_i)^T - 1\Big(\frac{\mathbf{X}_I^T\mathbf{X}_I}{1} - \frac{\mathbf{X}_I^T\mathbf{X}_I}{1^2}\Big)\\ &= n\mathbf{S} - \frac{n}{n-1}(\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})^T \qquad (3.19)\end{aligned}$$

This completes the proof of lemma 3.4.1. □

Lemma 3.4.2 *Let $\{(\lambda_j, v_j), j = 1, 2, ..., p\}$ be the pair of eigenvalues and eigenvectors of sample covariance matrix $\mathbf{S}$. $\{(\lambda_{j(i)}, v_{j(i)}), i = 1, 2, ..., n\}$ be the pair of eigenvalues and eigenvectors of covariance matrix $\mathbf{S}_{(i)}$. One now has*

(i). $\lambda_{j(i)} = \frac{n}{n-1}\lambda_j - \frac{n}{(n-1)^2}||x_i - \bar{x}_i||^2 G_i$

where the weights G_i satisfy $0 \le G_i \le 1$ and $\sum_i G_i = 1$;

(ii). $\frac{n}{n-1}\lambda_{j+1} \le \lambda_{j(i)} \le \frac{n}{n-1}\lambda_j$, $j = 1, 2, ..., p$.

Proof:

It follows immediately from Theorem 1 in Wang and Liski (1993, p. 222-223)

(i). Denote $\alpha_i = (x_i - \bar{x})/\|x_i - \bar{x}\|$ and from lemma 3.4.1, one has

$$\mathbf{S}_{(i)} = \frac{n}{n-1}\mathbf{S} - \frac{n}{(n-1)^2}(x_i - \bar{x})(x_i - \bar{x})^T. \tag{3.20}$$

Replace α_i in equation 3.20 which implies

$$\mathbf{S}_{(i)} = \frac{n}{n-1}\mathbf{S} - \frac{n}{(n-1)^2}\|x_i - \bar{x}\|^2\alpha_i\alpha_i^T. \tag{3.21}$$

Given that $\frac{n}{n-1}\lambda_j - \frac{n}{(n-1)^2}\|x_i - \bar{x}\|^2 \leq \lambda_{j(i)} \leq \frac{n}{n-1}\lambda_j, j = 1, 2, \ldots, p.$ Thus, the weights G_i satisfies $0 \leq G_i \leq 1$ such that

$$\lambda_{j(i)} = \frac{n}{n-1}\lambda_j - \frac{n}{(n-1)^2}\|x_i - \bar{x}\|^2 G_j. \tag{3.22}$$

Now, the preceding equation can be written as

$$\text{trace } S_{(i)} = \frac{n}{n-1}\text{trace } S - \frac{n}{(n-1)^2}\|x_i - \bar{x}\|^2. \tag{3.23}$$

From equation 3.22, one has

$$\begin{aligned}\text{trace } S_{(i)} &= \sum_{j=1}^{p} \lambda_{j(i)} \\ &= \frac{n}{n-1}\text{trace } S - \frac{n}{(n-1)^2}\|x_i - \bar{x}\|^2 \sum_{i=1}^{p} G_i. \end{aligned} \tag{3.24}$$

As a consequence of equations 3.23 and 3.24, one has $\sum_{j=1}^{p} G_j = 1.$

(ii). The proof is given in Corollary 1 and 2 in Wang and Liski, (1993, p.224). □

Theorem 3.4.3 *The influence eigen j for each observation i can be denoted by*

$$\Delta^*_{j(i)} = (x_i^T v_j)^2 + \sum_{\substack{k=1 \\ k\neq i}}^{n} \Big\{ (v_j + v_{j(i)})^T x_k x_k^T (v_j + v_{j(i)}) \Big\}, \tag{3.25}$$

where $j = 1, 2, \ldots, p$.

Proof:
According to Gnanadesikan and Kettenring (1972) an influence interpretation of the Euclidean distance can be considered as the total of influence eigen:

$$\frac{n}{(n-1)}(x_i - \bar{x})^T(x_i - \bar{x}) = \sum_{j=1}^{p} \Big\{ \frac{1}{n-1}\Big(l_{ij}^2 - \lambda_j \Big) + \frac{1}{2(n-1)^2} l_{ij}^2 \Big(1 + \sum_{k\neq j} \frac{l_{ij}^2}{\lambda_k - \lambda_j} \Big) \Big\}. \tag{3.26}$$

By using the relationship of influence eigenstructure in lemma 3.3.1, equation 3.26 can be re-written as follows:

$$\begin{aligned}
&\frac{n}{(n-1)}(x_i - \bar{x})^T(x_i - \bar{x}) \\
&= \sum_{j=1}^{p} \Big[(x_i^T v_j)^2 + \sum_{\substack{k=1 \\ k\neq i}}^{n} \Big[(x_k^T v_j)^2 - \Big\{ x_k^T \Big(v_j + \frac{l_{ij}}{n-1} \sum_{k\neq j} \frac{l_{ik} v_k}{\lambda_k - \lambda_j} - \frac{1}{2(n-1)^2} \times \\
&\quad \sum_{k\neq j} \Big[\frac{l_{ij}^2 l_{ik}^2 v_j}{(\lambda_k - \lambda_j)^2} - \frac{2 l_{ik}^2 l_{ij}}{(\lambda_k - \lambda_j)} \sum_{k\neq j} \frac{l_{ik} v_k}{\lambda_k - \lambda_j} + \frac{2 l_{ij}^3 l_{ik} v_k}{(\lambda_k - \lambda_j)^2} \Big] \Big) \Big\}^2 \Big] \Big] \\
&= \sum_{j=1}^{p} \Big[(x_i^T v_j)^2 + \sum_{\substack{k=1 \\ k\neq i}}^{n} \Big\{ (v_j + v_{j(i)})^T x_k x_k^T (v_j + v_{j(i)}) \Big\} \Big].
\end{aligned} \tag{3.27}$$

From equation 3.27, the influence eigen j for each observation i can be

denoted by

$$\Delta^*_{j(i)} = (x_i^T v_j)^2 + \sum_{\substack{k=1 \\ k\neq i}}^{n} \Big\{ (v_j + v_{j(i)})^T x_k x_k^T (v_j + v_{j(i)}) \Big\}, \tag{3.28}$$

where $j = 1, 2, \ldots, p$. $\square$

However, if one considers the influence eigen j on I, thus Theorem 3.4.3 now becomes

$$\begin{aligned} \Delta^*_{j(I)} = \frac{-m}{n-m} \sum_{k=1}^{n} (x_k^T v_j)^2 - \frac{nm}{(n-m)^2} v_j^T \times \\ \Big[\frac{n-m}{n} \mathbf{S}_I + (\bar{x}_I - \bar{x})(\bar{x}_I - \bar{x})^T \Big] v_j \end{aligned} \tag{3.29}$$

Suppose that the influence of an observation, i.e. an outlier on statistics such as jth eigenvalues, λ_j or eigenvectors, v_j of a sample covariance matrix is simply the change in λ_j or v_j when the ith observation is deleted from the sample.

Recall that this chapter considers the maximum eigenvalue and the corresponding eigenvector as the object of interest. From equation 3.5, it is given that

$$\begin{aligned} max\{\lambda_1, \lambda_2, \ldots, \lambda_p\} &= \lambda_{max} \\ &= \lambda_1, \end{aligned} \tag{3.30}$$

where λ_1 corresponds to v_1. Now, let $j = 1$, and equation 3.25 becomes

$$\Delta^*_{1(i)} = (x_i^T v_1)^2 + \sum_{\substack{k=1 \\ k\neq i}}^{n} \Big\{ (v_1 + v_{1(i)})^T x_k x_k^T (v_1 + v_{1(i)}) \Big\}. \tag{3.31}$$

Therefore, one can consider the influence eigen, $\Delta^*_{1(i)}$ as a tool to identify a potential influence observation, i.e. outlier in data matrix $\mathbf{X}$. Note that the test of significance for outlier was discussed briefly in §3.1 and cur-

rently, it is noted that the test of significance for outliers using the eigenstructure, such as principal component analysis, has not been widely used (Jolliffe, 2002). Perhaps the best advice is that the observation that is obviously more extreme than most of the remaining observations in the data set should be examined.

As a consequence, by using $\Delta^*_{1(i)}$, potential outliers in $\mathbf{X}$ can be identified by plotting the index plot of $\{i, \Delta^*_{1(i)}\}$. Note that ith observation can be considered as a potential outlier if it is located further away than the remaining observations in the data set. By using lemma 3.4.1, 3.4.2 and equation 3.31 the algorithm for influence eigen, $\Delta^*_{1(i)}$ is given as follows:

- Step 1 : Generate the sample covariance matrix $\mathbf{S}$ and $\mathbf{S}_{(i)}$;
- Step 2 : Compute the eigenstructure of $\mathbf{S}$ and $\mathbf{S}_{(i)}$. Denote the eigenstructure of $\mathbf{S}$ and $\mathbf{S}_{(i)}$ as $\{\mathbf{\Lambda}, \mathbf{V}\}$ and $\{\mathbf{\Lambda_{(i)}}, \mathbf{V_{(i)}}\}$ respectively. Note that $\mathbf{\Lambda}$, $\mathbf{V}$, $\mathbf{\Lambda_{(i)}}$ and $\mathbf{V}_{(i)}$ are from equations 3.4, 3.3, 3.8 and 3.7 respectively;
- Step 3 : Choose the maximum eigenvalue and the corresponding eigenvector pair, $max\{\lambda_j, v_j\}$ and $max\{\lambda_{j(i)}, v_{j(i)}\}$ of $\{\mathbf{\Lambda}, \mathbf{V}\}$ and $\{\mathbf{\Lambda_{(i)}}, \mathbf{V_{(i)}}\}$ respectively, i.e. $\{\lambda_1, v_1\}$ and $\{\lambda_{1(i)}, v_{1(i)}\}$;
- Step 4 : Compute $\Delta^*_{1(i)} = (x_i^T v_1)^2 + \sum_{\substack{k=1 \\ k \neq i}}^{n} \left\{ (v_1 + v_{1(i)})^T x_k x_k^T (v_1 + v_{1(i)}) \right\}$ for each observation;
- Step 5 : Develop the index plot of $\{i, \Delta^*_{1(i)}\}$, $i = 1, 2, \ldots, n$.

The outliers that are detectable from the index plot are those which inflate variance and covariance. If an outlier is the cause of a large increase in variances of the original variables, then it must be extreme on those variables (Gnanadesikan and Kettenring, 1972). Thus, one can identify it by looking at the index plot.

3.4.2 Normalized influence eigen

Using lemma 3.4.1, lemma 3.4.2 and considering the relationship between λ_j and $\lambda_{j(i)}$ in equation 3.11, one may compute the normalized influence eigen j for each ith observation as follows:

$$\Delta^{**}_{j(i)} = \Big\{\lambda_j - \lambda_{j(i)}\Big\}\Big[\sum_{j=1}^{p}\Big\{\frac{1}{n-1}\Big(l^2_{ij} - \lambda_j\Big) + \frac{1}{2(n-1)^2}l^2_{ij}\Big(1 + \sum_{k\neq j}\frac{l^2_{ij}}{\lambda_k - \lambda_j}\Big)\Big\}\Big]^{-1}. \tag{3.32}$$

The normalized influence eigen j for I is given by

$$\Delta^{**}_{j(I)} = \Big\{\lambda_j - \lambda_{j(I)}\Big\}\Big[\sum_{j=1}^{p}\Big\{\frac{-m}{n-m}\sum_{\substack{k=1\\k\neq i}}^{n}(x_k^T v_j)^2 - \frac{nm}{(n-m)^2}v_j^T \times \Big[\frac{n-m}{n}S_I + (\bar{x}_I - \bar{x})(\bar{x}_I - \bar{x})^T\Big]v_j\Big\}\Big]^{-1}, \tag{3.33}$$

where $\lambda_{j(I)} = \dfrac{n}{n-m}\lambda_j - \dfrac{nm}{(n-m)^2}v_j^T \times \Big[\dfrac{n-m}{n}S_I + (\bar{x}_I - \bar{x})(\bar{x}_I - \bar{x})^T\Big]v_j\Big\}$ is given in Wang and Liski (1993, p.219).

As this chapter considers the maximum eigenvalue, therefore, substitute $j = 1$ into equation 3.32 to give

$$\Delta^{**}_{1(i)} = \Big\{\lambda_1 - \lambda_{1(i)}\Big\}\Big[\sum\Big\{\frac{1}{n-1}\Big(l^2_{i1} - \lambda_1\Big) + \frac{1}{2(n-1)^2}l^2_{i1}\Big(1 + \sum_{k\neq j}\frac{l^2_{i1}}{\lambda_k - \lambda_1}\Big)\Big\}\Big]^{-1}. \tag{3.34}$$

From equation 3.13, $\lambda_j \geq \lambda_{j(i)}$, and if the difference of $\lambda_j - \lambda_{j(i)}$ is large, that means $\lambda_{j(i)}$ has a small value. Therefore ith observation affects λ_j and produces a large value of $\Delta^{**}_{j(i)}$ if the value of $\lambda_{j(i)}$ is very small.

Now, since $\Delta^{**}_{1(i)}$ is considering the maximum eigenvalue, it is noted that the ith observation influences λ_1 if the deletion of ith observation

causes the value of $\lambda_{1(i)}$ to become smaller and the difference of $\lambda_1 - \lambda_{1(i)}$ to become larger.

Therefore, the ith observation needs extra attention if it has large $\Delta^{**}_{1(i)}$ and is situated at the top of the index plot of $\{i, \Delta^{**}_{1(i)}\}$. This chapter will henceforth refer to the algorithm of normalized influence eigen, as $\Delta^{**}_{1(i)}$ and it is summarized as follows :

- Step 1 : Given $n \times p$ data matrix **X**, the sample covariance matrices **S** and $\mathbf{S}_{(i)}$ can be obtained by lemma 3.4.1;
- Step 2 : Compute the eigenvalues, $\mathbf{\Lambda}$ of **S** and the eigenvalues, $\mathbf{\Lambda_{(i)}}$ of $\mathbf{S}_{(i)}$;
- Step 3 : Choose the maximum eigenvalue, $max\{\lambda_j\}$ and $max\{\lambda_{j(i)}\}$ from $\mathbf{\Lambda}$ and $\mathbf{\Lambda_{(i)}}$ respectively, i.e. λ_1 and $\lambda_{1(i)}$;
- Step 4 : Calculate the normalized influence eigen, $\Delta^{**}_{1(i)}$ in equation 3.34 for each ith observations;
- Step 5 : Plot observations $\{i, \Delta^{**}_{1(i)}\}$, $i = 1, 2, ..., n$.

3.5 Influence angle based on eigenstructure for identification of outliers

3.5.1 Influence angle

Considering the relationship between eigenstructure in lemma 3.3.1 one can also develop the angle between v_j and $v_{j(i)}$ (Mertens, 1998). If ith is an outlier, therefore v_j will change when ith observation is deleted from the sample data matrix, **X**.

Let $\theta_{j(i)}$ be the angle between the jth eigenvectors of **S** for the given data **X**, and the $j(i)th$ eigenvectors when the ith observation is deleted in

X (i.e., $\mathbf{X}_{(i)}$), then one has the formula of $\theta_{j(i)}$ by Wang and Nyquist (1991) as

$$\cos(\theta_{j(i)}) = \frac{1}{2} \parallel v_j + v_{j(i)} \parallel^2 -1, \tag{3.35}$$

or it can be re-written as a function of eigenvalues and eigenvectors by

$$\theta_{j(i)} = \cos^{-1}\left\{ \frac{l_{ij}/\lambda^*_{j(i)}}{\sqrt{\sum_{k=1}^{p} l_{ik}^2/(\lambda^*_{j(i)} + (\lambda_k - \lambda_j))^2}} \right\}, \tag{3.36}$$

where $j = 1, 2, \ldots, p$; $i = 1, 2, \ldots, n$.

l_{ij} is the principal component scores of the omitted observation in the principal component decomposition of the complete data **X** and

$$\lambda^*_{j(i)} = \lambda_j - \frac{1}{n-1}(l_{ij}^2 - \lambda_j) - \frac{1}{2(n-1)^2} l_{ij}^2 [1 + \sum_{k \neq j}^{p} \frac{l_{ij}^2}{\lambda_k - \lambda_j}] + O(\frac{1}{n^3}).$$

The vector angle is defined as the angle between 0 and 180 degrees that satisfies the relationship $v_j^T v_{j(i)} = \|v_j\| \, \|v_{j(i)}\| \cos \theta_{j(i)}$ where $\|.\|$ refers to the vector length. If m observations are deleted from **X**, therefore

$$\begin{aligned}
\theta_{j(I)} &= \cos^{-1}\Big\{ \frac{v_j^T v_{j(I)}}{\|v_j\| \|v_{j(I)}\|} \Big\} \\
&= \cos^{-1}\Big\{ v_j^T \Big[v_j + \frac{m}{n-m} l_{jI} \sum_{k \neq j} l_{kI} (\lambda_k - \lambda_j)^{-1} v_k + \\
&\quad \frac{m}{n} \sum_{k \neq j} v_k^T S_I v_j (\lambda_k - \lambda_j)^{-1} v_k \Big] \Big\} \\
&= \cos^{-1}\Big\{ 1 + v_j^T \Big[\frac{m}{n-m} l_{jI} \sum_{k \neq j} l_{kI} (\lambda_k - \lambda_j)^{-1} v_k \\
&\quad + \frac{m}{n} \sum_{k \neq j} v_k^T S_I v_j (\lambda_k - \lambda_j)^{-1} v_k \Big] \Big\},
\end{aligned} \tag{3.37}$$

where $v_{j(I)} = v_j + \frac{m}{n-m} l_{jI} \sum_{k \neq j} l_{kI}(\lambda_k - \lambda_j)^{-1} v_k + \frac{m}{n} \sum_{k \neq j} v_k^T S_I v_j (\lambda_k - \lambda_j)^{-1} v_k$ and $l_{jI} = v_j^T(\bar{x}_I - \bar{x})$ is the mean of principal component score $l_{ji_m},\ i_m \in I$. Note that $v_{j(I)}$, l_{jI} and $l_{ji_m} = v_j^T(x_{i_m} - \bar{x})$ are given by Wang and Liski (1993).

Supposing that one only delete ith observation and considers the maximum eigenvalue, replacing $j = 1$ in equation 3.36 leads to

$$\theta_{1(i)} = \cos^{-1}\left\{ \frac{l_{i1}/\lambda^*_{1(i)}}{\sqrt{\sum_{k=1}^{p} l_{ik}^2/(\lambda^*_{1(i)} + (\lambda_k - \lambda_1))^2}} \right\}, \tag{3.38}$$

Next, one can apply the influence angle, $\theta_{1(i)}$ to identify the outlier in the data set; note that there are a few criteria that will control $\theta_{j(i)}$ value as following:

(i). First, consider $\lambda_j \geq \lambda_{j(i)}$ and $\lambda_{j(i)} \geq \lambda_{k+1}$ as given in lemma 3.4.2, where $j, k = 1, 2, \ldots, p$. One finds that the $\theta_{j(i)}$ value is dominated by the first component of the denominator, i.e. $l_{i1}^2/\{\lambda^*_{1(i)}\}^2$. If one substitutes $k = 1,$ into

$$l_{ik}^2/(\lambda^*_{1(i)} + (\lambda_k - \lambda_1))^2,$$

it becomes

$$l_{i1}^2/\{\lambda^*_{1(i)}\}^2. \tag{3.39}$$

Notice that

$$\frac{l_{ik}^2}{\{\lambda^*_{1(i)} + (\lambda_k - \lambda_1)\}} > \frac{l_{i(k+1)}^2}{\{\lambda^*_{1(i)} + (\lambda_{k+1} - \lambda_1)\}},$$

and the $\frac{l_{i(k+1)}^2}{\{\lambda^*_{1(i)} + (\lambda_{k+1} - \lambda_1)\}}$ value is always small because the denominator is $\{\lambda^*_{1(i)} + (\lambda_{k+1} - \lambda_1)\}$ of $\theta_{1(i)}$ usually large following $\lambda_{j(i)} \geq \lambda_{k+1}$, $j, k = 1, 2, \ldots, p$.

As a consequence, if the numerator value of the equation 3.38 is close to one, the denominator value will also be almost the same, noting that the numerator value is always less than the denominator value. This follows that the $\theta_{j(i)}$ yields almost a zero degree angle. Another point is that the value of cos $(\theta_{1(i)})$ is always between -1 and 1.

(ii). Next, if the principal component score is negative, $\theta_{1(i)}$ will be large. This corresponds to a negative cosine yielding a large angle.

Therefore, the supposed potential outlier will be situated further away than the remaining observations in the data set if:

(i). $\theta_{1(i)}$ for ith observation is larger than other observations following that $\{\lambda^*_{1(i)}\}$ (equation 3.39) in the first component of $\theta_{1(i)}$ is large; or $\theta_{1(i)}$ for ith observation is smaller than other observations corresponding to $\{\lambda^*_{1(i)}\}$ in the first component of $\theta_{1(i)}$ observation is small;

(ii). the principal component score for ith observation is negative while others are positive. Note that the negative principal component score produces larger $\theta_{1(i)}$ than the positive principal component score and vice versa.

The outlier can be exhibited by the index plot $\{i, \theta_{1(i)}\}$. Based on the influence angle $\theta_{1(i)}$, the following algorithm is proposed to find an outlier:

- Step 1 : Using lemma 3.4.1 and 3.4.2, find $\mathbf{S}$ and $\mathbf{S}_{(i)}$;
- Step 2 : Next find the eigenstructure of $\mathbf{S}$ and $\mathbf{S}_{(i)}$, and choose the maximum eigenpair (v_1, λ_1) and $(v_{1(i)}, \lambda_{1(i)})$ respectively;
- Step 3 : Find the principal component score, l_{ik} for each k or compute $l_{ik} = (x_i^T v_k)$;
- Step 4 : Compute $\theta_{1(i)}$;
- Step 5 : Identify the outlier from the index plot of $\{i, \theta_{1(i)}\}$.

3.5.2 Modified influence angle

From §3.5.1, note that $\lambda^*_{1(i)}$ in equation 3.39 plays an important role in determining $\theta_{j(i)}$ value in equation 3.38. Consequently, this section will ignore the principal component score by letting $l_{ij} = 1$, therefore equation 3.36 can be rewritten as

$$\theta^*_{j(i)} = \cos^{-1}\left\{ \frac{1/\lambda^*_{j(i)}}{\sqrt{\sum_{k=1}^{p} 1/(\lambda^*_{j(i)} + (\lambda_k - \lambda_j))^2}} \right\}, \tag{3.40}$$

where $j = 1, 2, \ldots, p$; $i = 1, 2, \ldots, n$. Let $j = 1$ as this chapter considers $max\{\lambda_j\}$. Therefore equation 3.40 is now given as

$$\theta^*_{1(i)} = \cos^{-1}\left\{ \frac{1/\lambda^*_{1(i)}}{\sqrt{\sum_{k=1}^{p} 1/(\lambda^*_{1(i)} + (\lambda_k - \lambda_1))^2}} \right\}, \tag{3.41}$$

One can use $\theta^*_{1(i)}$ to identify the outlier, and the algorithm of modified influence angle is similar to the original influence angle, except one does not have to compute the principal component scores of the omitted observation in the principal component decomposition of the complete data $\mathbf{X}$.

Note that, if ith observation is an outlier, $\theta^*_{1(i)}$ is larger than other observations in the data set following that $\lambda^*_{1(i)}$ is large. This is because deletion of ith observation, i.e. an outlier causes $\lambda_{j(i)}$ to become much smaller and the difference of $\lambda^*_{1(i)} + (\lambda_k - \lambda_j)$ for $j, k = 1$ to become larger.

The ith observation is considered as a potential outlier by $\theta_{1(i)}$ if it is located at the top of the index plot $\{i, \theta^*_{1(i)}\}$.

3.6 Simulation data set

The techniques in §3.4 and §3.5 are tested on the simulation data set given in the Table 3.1.

Table 3.1: The simulation data set used for illustration

Data	Sample size, n	Number of variables, p	Number of outlier, m
1	105	3	5
2	1005	10	5
3	3005	100	5
4	3050	100	50

Three different scenarios are considered to generate the data set from the multivariate distributions in Table 3.1 and this is further described in the following section.

3.6.1 Scenario 1: outliers with the same shapes but different locations

There are 3 conditions considered in the first scenario:

- Condition 1 : A random vector of $x_1, x_2, \ldots, x_n$ is drawn from a $p-$ variate normal distribution with mean vector μ and positive definite covariance matrix $\mathbf{\Sigma}$, i.e. $N(\mu, \mathbf{\Sigma})$. Next $x_1^*, x_2^*, \ldots, x_m^*$ is another random sample drawn from a $p-$ variate normal distribution with mean vector $\mu_{\mathbf{c1}}$ and a similar covariance matrix $\mathbf{\Sigma}$, i.e. $N(\mu_{\mathbf{c1}}, \mathbf{\Sigma})$. Note that m is the number of outliers. Later these two sets of data vector are merged;

- Condition 2 : The $x_1, x_2, \ldots, x_n$ random vector is developed as in condition 1. However, $x_1^*, x_2^*, \ldots, x_m^*$ is constructed by using $N(\mu_{\mathbf{c2}}, \mathbf{\Sigma})$,

which is closer to the majority of data parental distribution in condition 1, i.e. $\mu_{\mathbf{c2}} < \mu_{\mathbf{c1}}$;

- Condition 3 : In this condition, the random vector $x_1, x_2, \ldots, x_n$ is developed as in condition 1 and condition 2. Nevertheless, other m observations, which are $x_1^*, x_2^*, \ldots, x_m^*$, are developed by using $N(\mu_{\mathbf{c3}}, \mathbf{\Sigma})$, which is much closer to the majority of data parental distribution in condition 1 and condition 2, i.e. $\mu_{\mathbf{c3}} < \mu_{\mathbf{c2}} < \mu_{\mathbf{c1}}$.

3.6.2 Scenario 2: outliers with different shapes and different locations

In scenario 2, $x_1, x_2, \ldots, x_n$ is a random vector drawn for $p-$ variate normal distribution with mean vector μ and positive definite matrix $\mathbf{\Sigma}$ and $x_1^*, x_2^*, \ldots, x_m^*$ is another set of random vector from $p-$ variate distribution with mean vector $\mu_{\mathbf{s2}}$ and covariance matrix $\mathbf{\Sigma_{s2}}$. Note that $\mu \neq \mu_{\mathbf{s2}}$ and $\mathbf{\Sigma} \neq \mathbf{\Sigma_{s2}}$.

3.6.3 Scenario 3: outliers from a different probability law

Let $x_1, x_2, \ldots, x_n$ be a random sample drawn from $p-$ variate normal distribution with mean vector μ and positive definite covariance matrix $\mathbf{\Sigma}$. Now generate $x_1^*, x_2^*, \ldots, x_m^*$ drawn from $p-$ variate student t distribution with z degrees of freedom and correlation matrix $\mathbf{\Sigma_{s3}}$. Note that $\mathbf{\Sigma} \neq \mathbf{\Sigma_{s3}}$.

3.7 Real data set

This chapter will use data sets taken from Rousseeuw and Leroy (1987). Details of the data sets are given in Table 3.2. These data sets are chosen as they were often used to evaluate the performance of the outlier detection method (see, Rousseeuw and Leroy, 1987; Rousseeuw and von

Zomeren, 1990; Hadi, 1992; Maronna and Yohai, 1995; Rousseeuw and Driessen, 1999; Gao et al., 2005).

Table 3.2: The real data set used for illustration

Data set	Number of observations, n	Number of variables, p
Hawkins Bradu Kass	75	3
Stackloss	21	3
Salinity	28	3

By using these data sets, the influence eigen ($\Delta^{*}_{1(i)}$), normalized influence eigen ($\Delta^{**}_{1(i)}$), influence angle ($\theta_{1(i)}$) and modified influence angle ($\theta^{*}_{1(i)}$) are calculated and the index plot for each technique is developed for the identification of outliers. The index plots in this chapter will denote the potential outlier within the black circles.

Figure 3.1 contains the index plot of Mahalanobis distance approach for these three data sets. As one can see, the Mahalanobis distance approach does not perform consistently when it is tested on these three data sets. Note that the black solid circle in Figure 3.1 denotes the outlier.

In the first data set, which is the Hawkins Bradu Kass data, there are 14 observations ($i = 1, 2, \ldots, 14$) that were flagged as outliers by previous studies (see, Rousseeuw and von Zomeren, 1990; Hadi, 1992; Maronna and Yohai, 1995; Rousseeuw and Driessen, 1999; Gao et al., 2005). This artificial data set was generated by Hawkins et al. in 1984.

The second data set is the Stackloss data. There are 4 outliers in this data set, which are observations $1, 2, 3$ and 21. The Stackloss data set is a real data set that has been used by many statisticians and it is about the operation of a plant for the oxidation of ammonia to nitric acid (Rousseeuw and Leroy, 1987).

The third data set is the Salinity data. It is a set of measurements of water salinity, i.e. salt concentration and river discharge taken in North

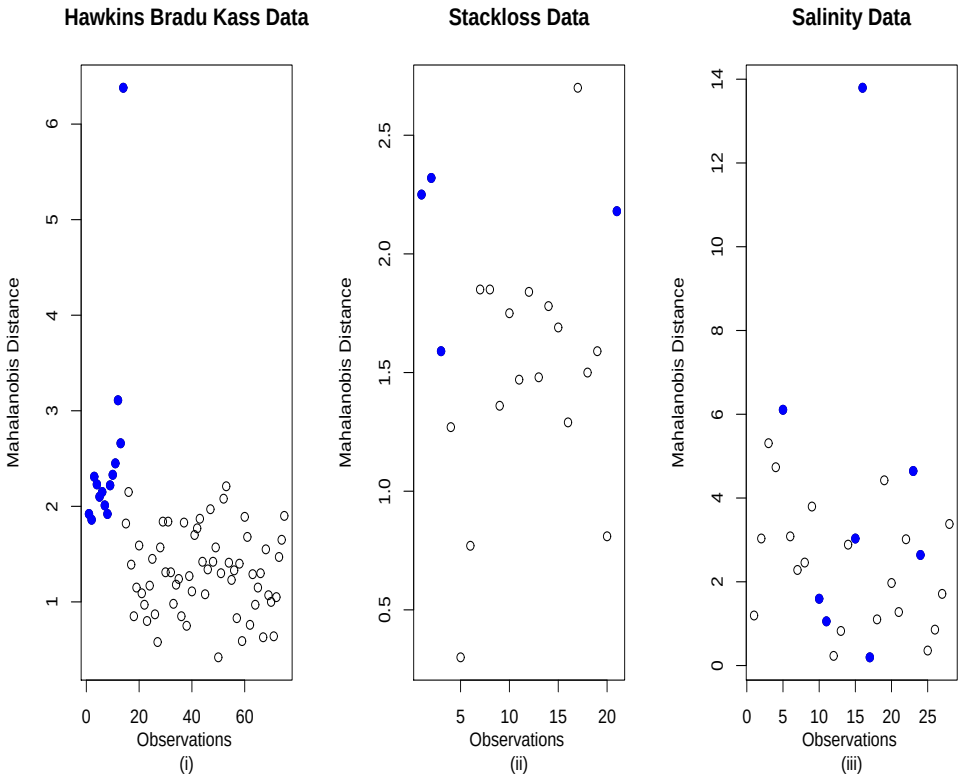


Figure 3.1: Mahalanobis distance plot for (i) Hawkins Bradu Kass data (ii) Stackloss data (iii) Salinity data

Carolina's Pamlico Sound (Rousseeuw and Leroy, 1987). This real data set contains 8 outliers($i = 5, 10, 11, 15, 16, 17, 23$ and 24).

Supposed outliers will be located at the top of the plot. Recall that the Mahalanobis approach will flag observations that are greater than $\chi^2_{p,\alpha}$ as outliers, where p is the number of variables. Therefore, an observation that is labeled as an outlier should have the Mahalanobis Distance (MD_i) value larger than other normal observations in the data matrix $\mathbf{X}$.

However, from Figure 3.1, it is noted that the Mahalanobis approach cannot identify all possible outliers in some of the data sets, i.e. the Stackloss data.

3.8 Illustration by simulation data set

Each technique in §3.4 and §3.5 is used, in turn, on each data set in Table 3.1 that is generated following the scenario described in §3.6. First, the performance of those techniques is evaluated on low dimension and a small sample size data. This is illustrated in §3.9.

Next, in §3.10, those techniques are considered with the data set with a large sample size yet still corresponding to the low dimension. Finally in §3.11, they are applied to high dimension data set with a large sample size data. Instead of using $m = 5$, §3.11 also examines those techniques for a data set containing $m = 50$. The purpose of choosing $m = 5$ to examine the performance of algorithms in §3.4 and §3.5 is to ensure that outliers can clearly be observed from the index plot by the reader.

The index plot will denote the potential outliers within the black circles. The algorithms for the techniques in §3.4 and §3.5 clearly mention that the index plot can be drawn by using two-dimensional scatterplot in which a comparison of 2 measures is presented, one measure along each axis.

However, note that the index plots in §3.9-§3.11 are represented by three-dimensional scatterplots to accommodate a better and clearer illustration. The x-axis and the y-axis denote the index while the z-axis denotes the influence value, i.e. $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ or $\theta^{*}_{1(i)}$. Examples in §3.9.1 corresponding to condition 1 will be described in detail, although a number of other examples will be discussed briefly.

3.9 Low dimension and small sample size

First, the techniques in §3.4 and §3.5 are used on the low dimension with a small sample size data. The data set is generated following the scenarios described in §3.6. The data set in this section contains 105 observations with 3 dimensions. Note that the last five observations in this data set can

be regarded as outliers.

3.9.1 Scenario 1 with $n = 105,\ p = 3,\ m = 5$

Condition 1

The $\Delta^*_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^*_{1(i)}$ in §3.4 and §3.5 are applied for identification of outliers. In the first condition, one can observe from Figure 3.2 that all techniques are able to identify the outlier. It is noted that the outliers are

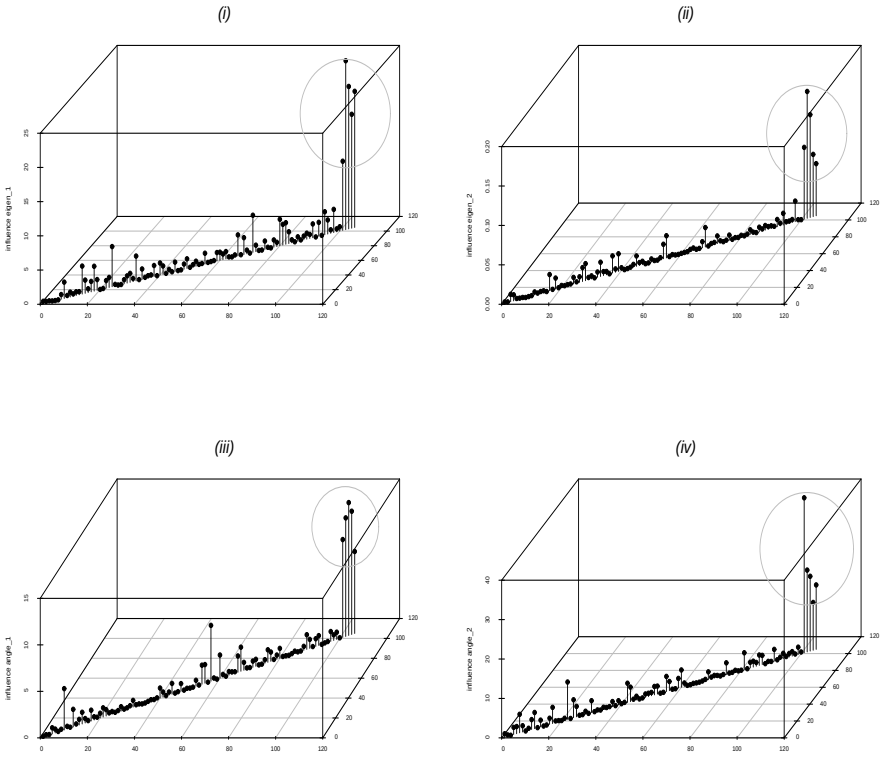


Figure 3.2: 3D scatterplot for $n = 105,\ p = 3,\ m = 5$ – visualization of outlier for condition 1, scenario 1 by using (i) $\Delta^*_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^*_{1(i)}$.

within the black circle located at the top of each index plot, which means

they all have large values from other observations in the data set. First, consider the influence eigen, i.e. $\Delta^{*}_{1(i)}$ and $\Delta^{**}_{1(i)}$. These two techniques are attached to each other since $\Delta^{**}_{1(i)}$ is the normalized value of $\Delta^{*}_{1(i)}$. Therefore, if $\Delta^{*}_{1(i)}$ is large, then $\Delta^{**}_{1(i)}$ is also large. The value of influence eigen depends on the $\lambda_1 - \lambda_{1(i)}$, supposing that $\lambda_1 - \lambda_{1(i)}$ is large, this is followed by $\Delta^{**}_{1(i)}$.

Provided that ith observation is an outlier, then the $\lambda_{1(i)}$ for i is smaller than the remaining observations in the data set, (Table 3.3 gives $\lambda_{1(i)}$ of $\Delta^{**}_{1(i)}$ for $i = 101, \ldots, 105$). The λ_1 value for the data set generated for condition 1 corresponding to $\Delta^{**}_{1(i)}$ is 196.28, whereas the maximum and minimum values of $\lambda_{1(i)}$ for observations 1 until 100 are 191.42 and 186.21, respectively. Information about $\lambda_{1(i)}$ for $\Delta^{*}_{1(i)}$ and $\Delta^{**}_{1(i)}$ can be found in Table 3.3.

Following that the value of $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ for all observations in Figure 3.2 are between 0 to 90, this indicates the principal score for all observations in the data set generated by condition 1 are positive values. Therefore, if ith are outliers, the value for $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ are larger than the remaining observations in the data set. Recall in §3.5.1 and §3.5.2, the criterion that causes $\theta_{1(i)}$ or $\theta^{*}_{1(i)}$ to become larger is when $\lambda^{*}_{1(i)}$ is large. This is because deletion of ith observation, i.e. an outlier, causes the $\lambda_{j(i)}$ to become much smaller and the difference of $\lambda^{*}_{1(i)} + (\lambda_k - \lambda_j)$ for $j, k = 1$ gets larger. The value of $\lambda_{1(i)}$ for $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ are presented in Table 3.3.

Table 3.3: $\lambda_{1(i)}$ of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ for observations 101-105– Condition 1, scenario 1 with $n = 105, p = 3, m = 5$

Technique	$\lambda_{1(101)}$	$\lambda_{1(102)}$	$\lambda_{1(103)}$	$\lambda_{1(104)}$	$\lambda_{1(105)}$
$\Delta^{*}_{1(i)}$	186.16	171.58	175.44	179.62	176.37
$\Delta^{**}_{1(i)}$	174.54	161.58	167.23	176.79	179.16
$\theta_{j(i)}$	184.37	181.43	179.49	180.79	186.76
$\theta^{*}_{j(i)}$	150.54	161.07	162.17	166.59	163.81

Next, the values of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ for observations 101 until 105 are given in Table 3.4. These five observations are the largest values of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ and Figure 3.2 shows they are located further away than other observations. The maximum and minimum values of $\Delta^{*}_{1(i)}$ among observations 1 until 100, i.e. the good data, are 5.86 (observation 24) and 7.1×10^{-6} (observation 70) respectively.

Observation 55 and observation 85 denote the maximum and the minimum value of $\Delta^{**}_{1(i)}$ among observations 1 until 100, where the values are given as 0.028 and 1.54×10^{-5}. Next, for $\theta_{1(i)}$, note that the maximum and the minimum value are present in observation 57 and observation 45. Among observations 1 until 100, the maximum and the minimum value of $\theta^{**}_{1(i)}$ are 9.46 and 1.99×10^{-5} by observation 22 and observation 54 respectively. Note that there is a gap between the maximum values of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ with the last five observations, i.e. outliers.

Table 3.4: $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ for observations $3001 - 3005$–Condition 1, scenario 1 with $n = 105, p = 3, m = 5$

Technique	101	102	103	104	105
$\Delta^{*}_{1(i)}$	10.10	24.70	20.82	16.64	19.90
$\Delta^{**}_{1(i)}$	0.09	0.16	0.13	0.08	0.06
$\theta_{j(i)}$	10.52	12.74	14.28	13.25	8.81
$\theta^{*}_{j(i)}$	39.30	20.75	19.10	12.11	16.33

Condition 2

The data set generated for condition 2 is almost similar to condition 1, except the mean is closer to the majority of data observations. It is noted that the most difficult situation for identifying outlier is when the good and bad data are drawn from the same multivariate normal distribution with a small difference in the location vector (Rocke and Woodruff, 1996).

However, Figure 3.3 verifies that those techniques in §3.4 and §3.5 successfully identify observations $101 - 105$ as outliers in the data set generated using condition 2.

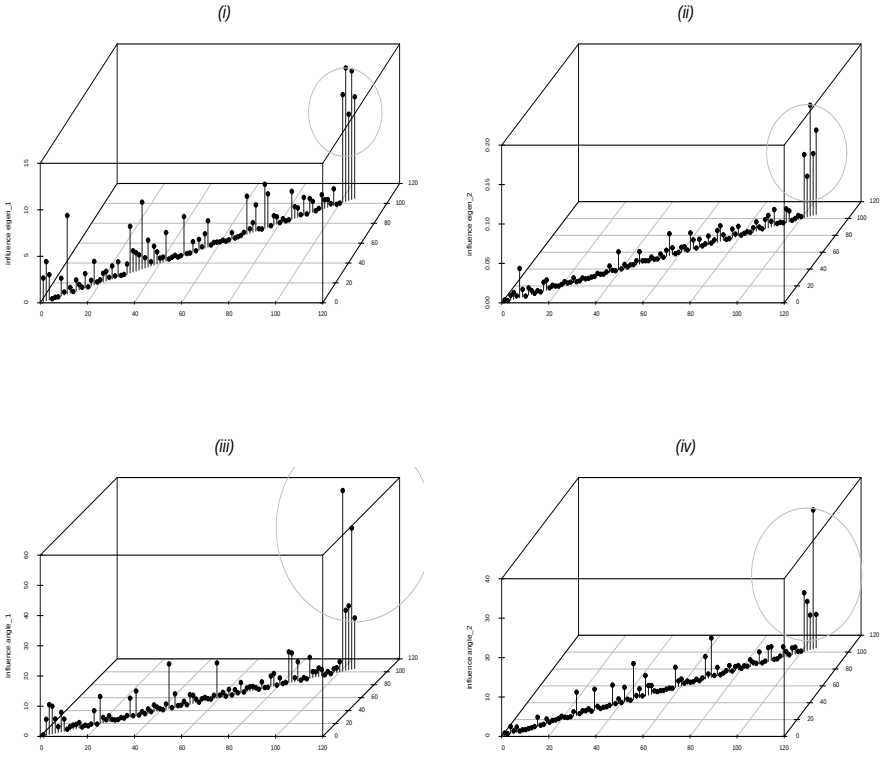


Figure 3.3: 3D scatterplot for $n = 105$, $p = 3$, $m = 5$ – visualization of outlier for condition 2, scenario 1 by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

Condition 3

Figure 3.4 displays the index plot for $\Delta^*_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^*_{1(i)}$, which is developed using the data set generated from the same population as conditions 1 and 2 but with a different mean. The mean in this condition is chosen to let the outliers become much closer to the majority of the data set compared to the one in condition 2. It is noted that an outlier is difficult to identify if the locations, i.e. mean, of the two populations are very close to each other. Instead of detecting 5 outliers, the index plot in Figure 3.4 only shows 4 extreme observations for Δ^*_i, 3 extreme observations for Δ^{**}_i and 2 extreme observations for $\theta_{j(i)}$ and $\theta^*_{j(i)}$ respectively.

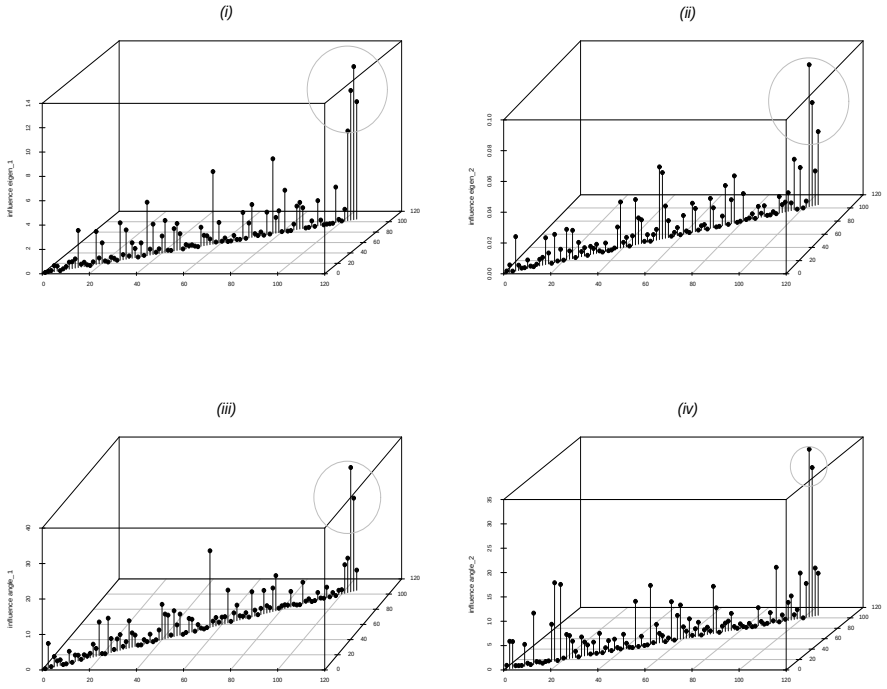


Figure 3.4: 3D scatterplot for $n = 105$, $p = 3$, $m = 5$ – visualization of outlier for condition 3, scenario 1 by using (i) $\Delta^*_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^*_{1(i)}$.

3.9.2 Scenario 2 with $n = 105,\ p = 3,\ m = 5$

Scenario 2 considers outliers with different shapes ($\Sigma \neq \Sigma_c$) and different locations ($\mu \neq \mu_{s2}$). Recall $(\mu,\ \Sigma)$ and $(\mu_{s2},\ \Sigma_c)$ are mean vector and covariance matrices for good and bad data respectively.

Figure 3.5 indicates all techniques in §3.4 and §3.5 are able to identify five outliers contained in the data set generated from scenario 2. The index plot for each technique marks the outlier within the black circles.

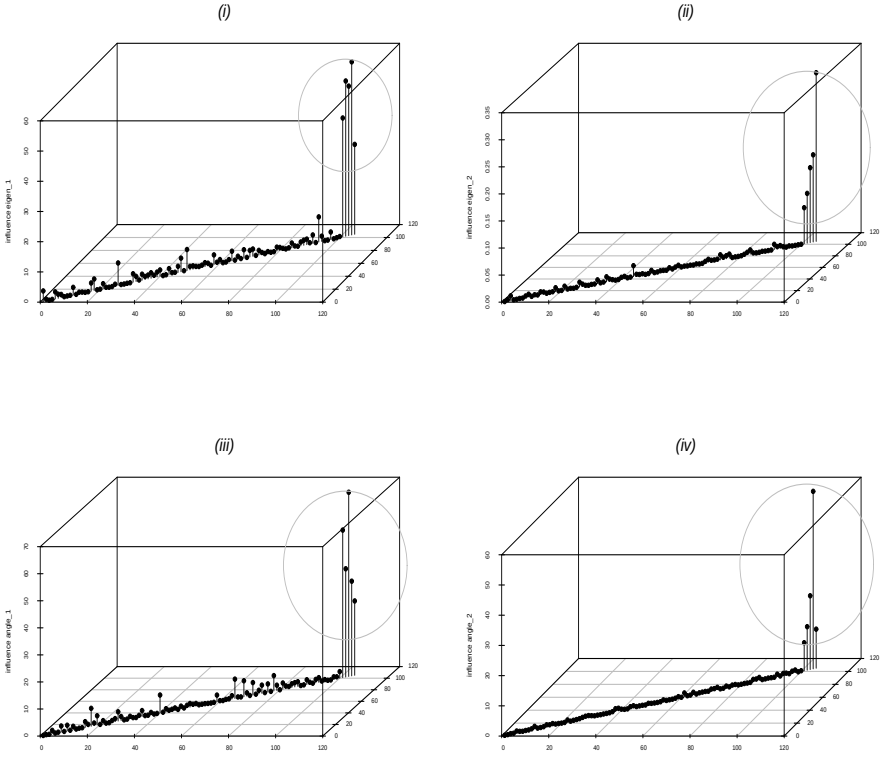


Figure 3.5: 3D scatterplot for $n = 105,\ p = 3,\ m = 5$ – visualization of outliers for scenario 2 by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

3.9.3 Scenario 3 with $n = 105,\ p = 3,\ m = 5$

The data set generated in scenario 3 considers outliers coming from a different probability law. The good data is generated using normal distribution, whereas the bad data uses the student-t distribution with z degrees of freedom.

Index plot of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ are shown in Figure 3.6. Overall, these four techniques do perform in identifying outliers, where the largest values of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ belong to observations $101 - 105$.

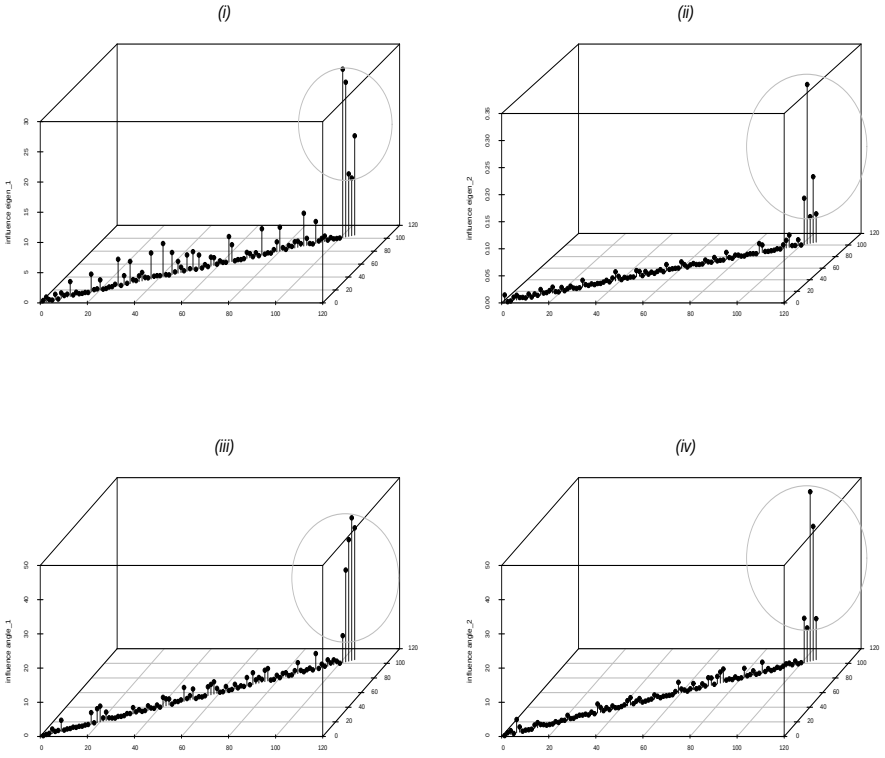


Figure 3.6: 3D scatterplot for $n = 105,\ p = 3,\ m = 5$ – visualization of outliers for scenario 3 by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

3.10 Low dimension and large sample size

Since the $\Delta^*_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^*_{1(i)}$ perform on low dimension with a small sample data size, i.e. examples in §3.9, this section will use them for identifying outliers in the large sample size yet still utilizing the low dimension. This section considers 1005 observations with 10 variables. The data set contains 5 outliers.

3.10.1 Scenario 1 with $n = 1005,\ p = 10,\ m = 5$

Condition 1

By generating a data set containing 1000 good data and 5 bad data on 10 variables using $N(\mu, \Sigma)$ and $N(\mu_{c1}, \Sigma)$ respectively, it is noted techniques in §3.4 and §3.5 can identify the outliers in this data set. This is shown in Figure 3.7.

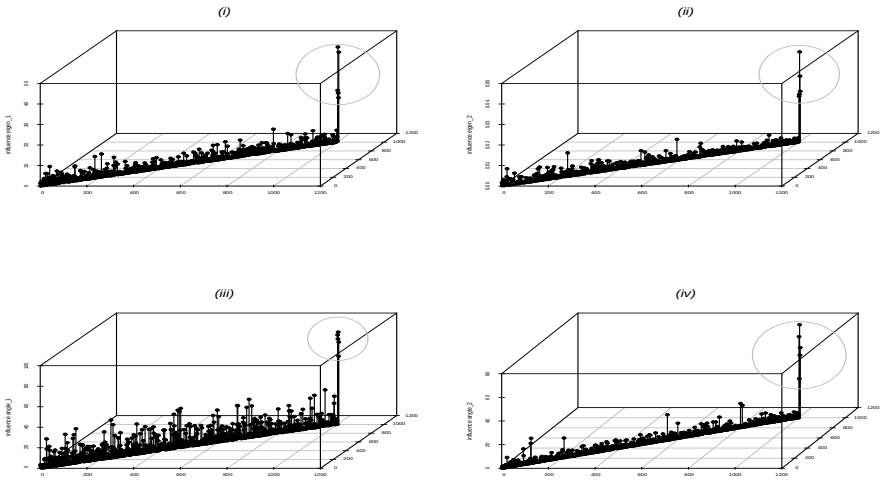


Figure 3.7: 3D scatterplot for $n = 1005,\ p = 10,\ m = 5$ – visualization of outlier for condition 1, scenario 1 by using (i) $\Delta^*_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^*_{1(i)}$.

Condition 2

Figure 3.8 corresponds to the condition of two observations of normal distribution respectively, using 1000 observations from $N(\mu, \Sigma)$ and 5 observations from $N(\mu_{c2}, \Sigma)$. Recall that μ_{c2} in condition 2 is closer to the majority of the data set than μ_{c1} in condition 1, i.e. $\mu_{c2} < \mu_{c1}$. However, it is noted that $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ are still able to identify the outliers in this data set. Cases $1001 - 1005$ are still located at the top of each index plot even though the gap between them and the majority of other observations is less than the gap shown by observations $1001 - 1005$ for each index plot in Figure 3.7.

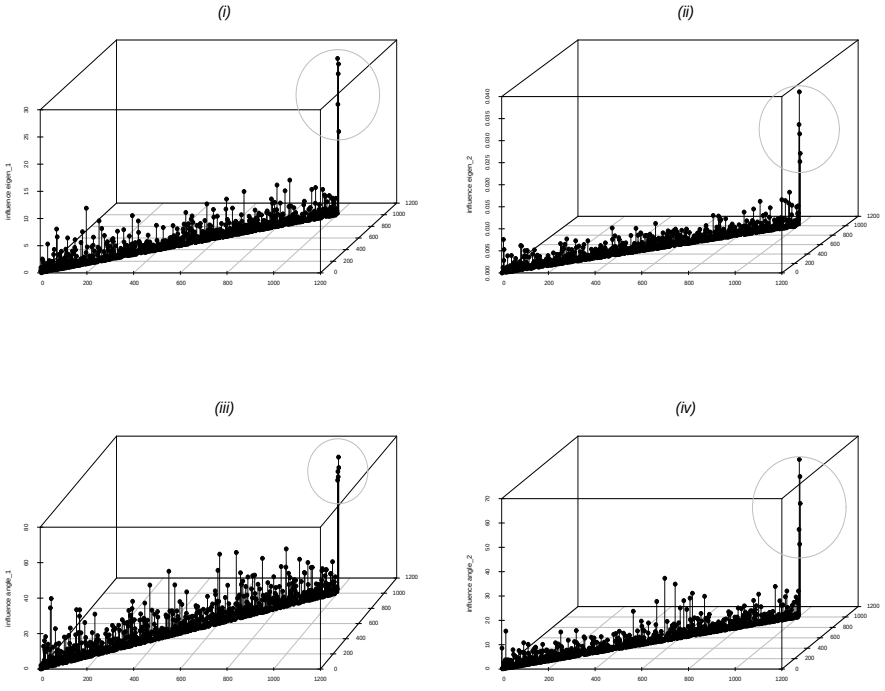


Figure 3.8: 3D scatterplot for $n = 1005$, $p = 10$, $m = 5$ – visualization of outlier for condition 2, scenario 1 by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

Condition 3

Figure 3.9 is an example of a large sample data size containing outliers that are generated much closer to the majority of the data set if compared to conditions 1 and 2, i.e. $\mu_{c3} < \mu_{c2} < \mu_{c1}$. Techniques in §3.4 and §3.5 cannot discover the whole 5 observations (1001-1005) as outliers. Nevertheless, note that the largest value of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ are among these 5 observations. Figure 3.9 flags them within the black circles.

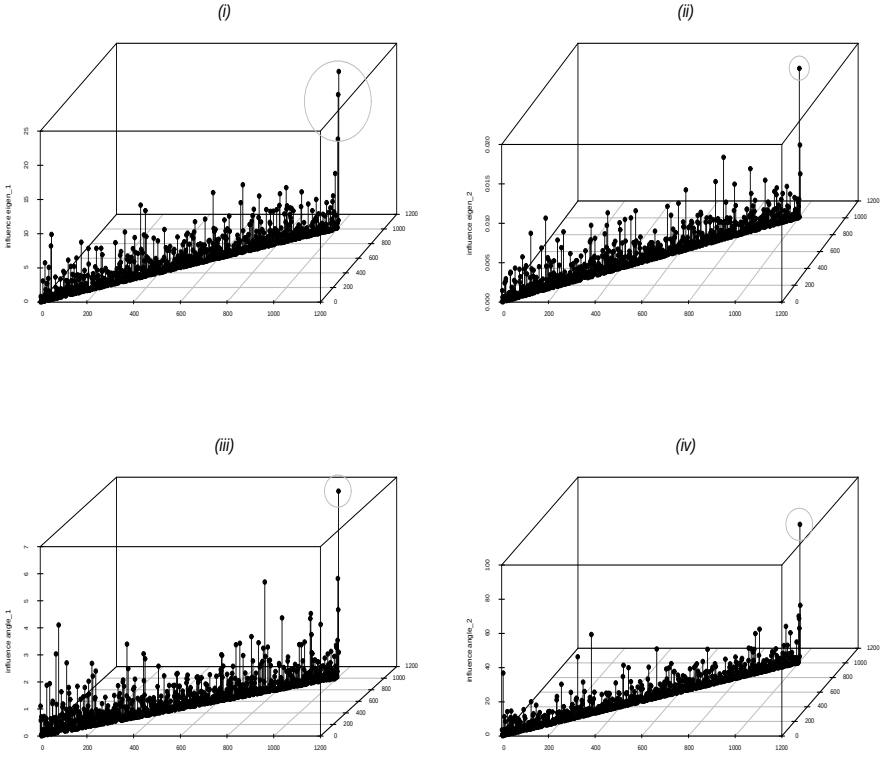


Figure 3.9: 3D scatterplot for $n = 1005$, $p = 10$, $m = 5$ – visualization of outlier for condition 3, scenario 1 by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

3.10.2 Scenario 2 with $n = 1005,\ p = 10,\ m = 5$

Scenario 2 denotes the data set generated from different locations ($\mu \neq \mu_c$) and different shapes ($\Sigma \neq \Sigma_c$). Referring to Figure 3.10 one can observe the index plot of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ is able to detect 5 observations that are generated as bad data. Notice there is a gap between these 5 observations and the majority of the data set for each index plot.

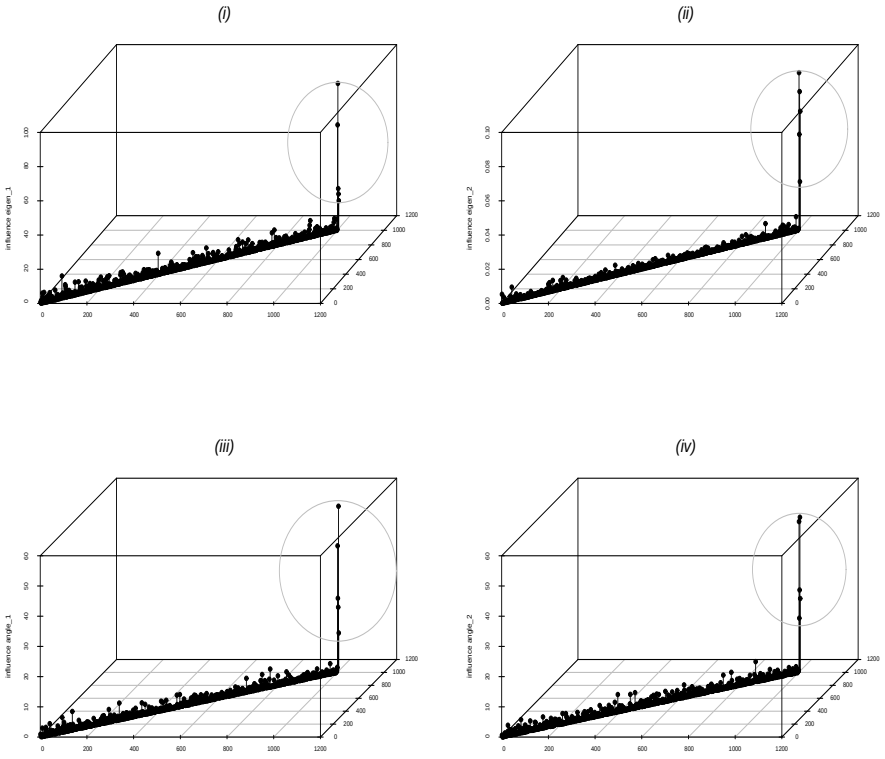


Figure 3.10: 3D scatterplot for $n = 1005,\ p = 10,\ m = 5$ – visualization of outlier for scenario 2 by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

3.10.3 Scenario 3 with $n = 1005,\ p = 10,\ m = 5$

In Figure 3.11, $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ are plotted against ith observations. It is obvious these four techniques succeed in identifying outliers generated from a different probability law.

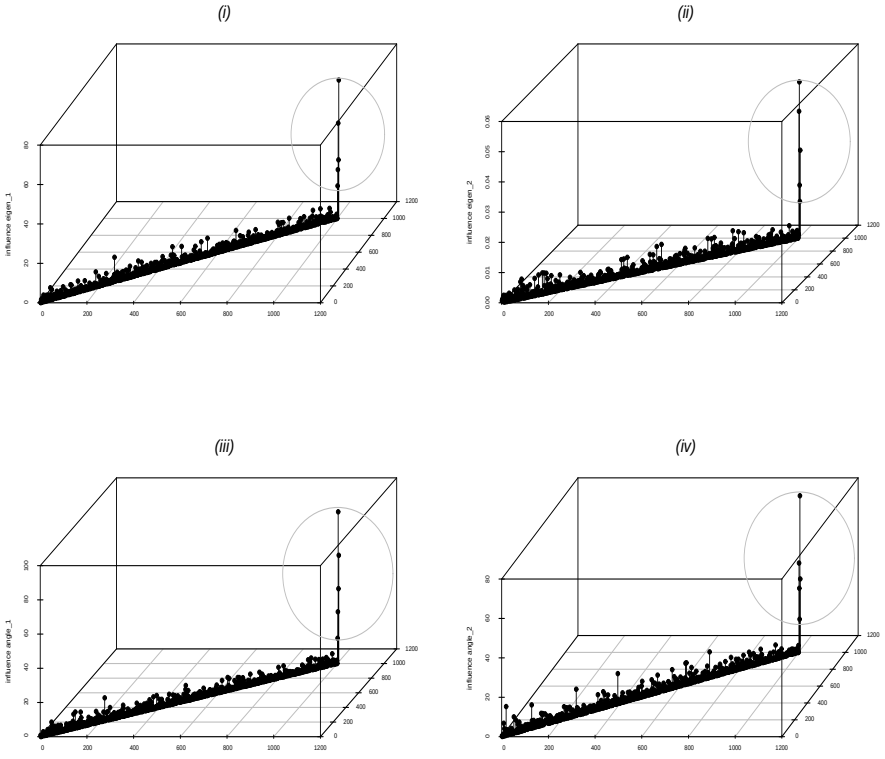


Figure 3.11: 3D scatterplot for $n = 1005,\ p = 10,\ m = 5$ – visualization of outlier for scenario 3 by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

3.11 High dimension and large sample size

Following the steps $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ performed in §3.9 and §3.11, they now will be examined on a high dimension and large sample data size. Note that this section considers $n = 3005$ and $n = 3050$ where each sample size contains $m = 5$ and $m = 50$ respectively.

3.11.1 Scenario 1

Condition 1 with $n = 3005,\ p = 100,\ m = 5$

Figure 3.12 presents the index plots of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$. These four index plots distinctly display 5 outliers at the top of the index plots. It is noted in each of the index plots that there is a very large gap between the outliers and the remaining observations, i.e. good data. The values of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ for each observation from the good data is almost zero.

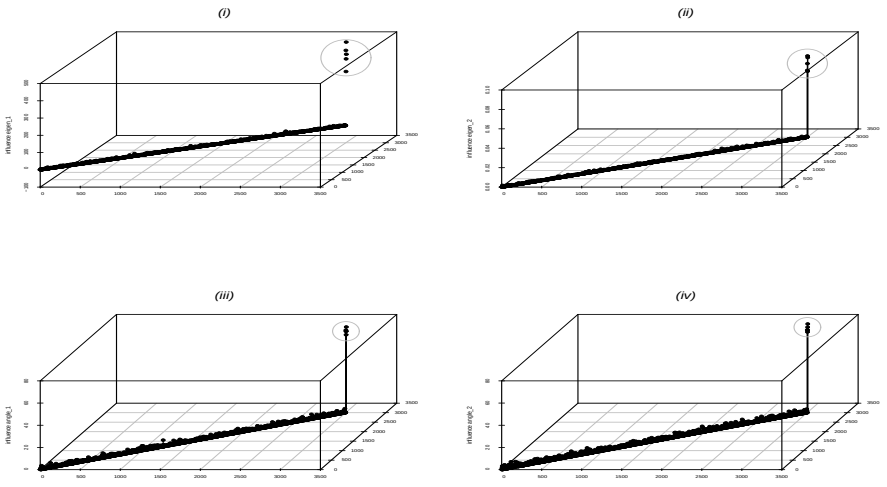


Figure 3.12: 3D scatterplot for $n = 3005,\ p = 100,\ m = 5$ – visualization of outliers for condition 1, scenario 1 by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

Condition 1 with $n = 3050,\ p = 100,\ m = 50$

Next, Figure 3.13 indicates the index plots for the generated data set the same as Figure 3.12, except the number of outliers, m has now increased to 50 observations. The results are consistent with those in Figure 3.12, where all 50 outliers are identified in the data set. The index plots denote the outliers within the black circles.

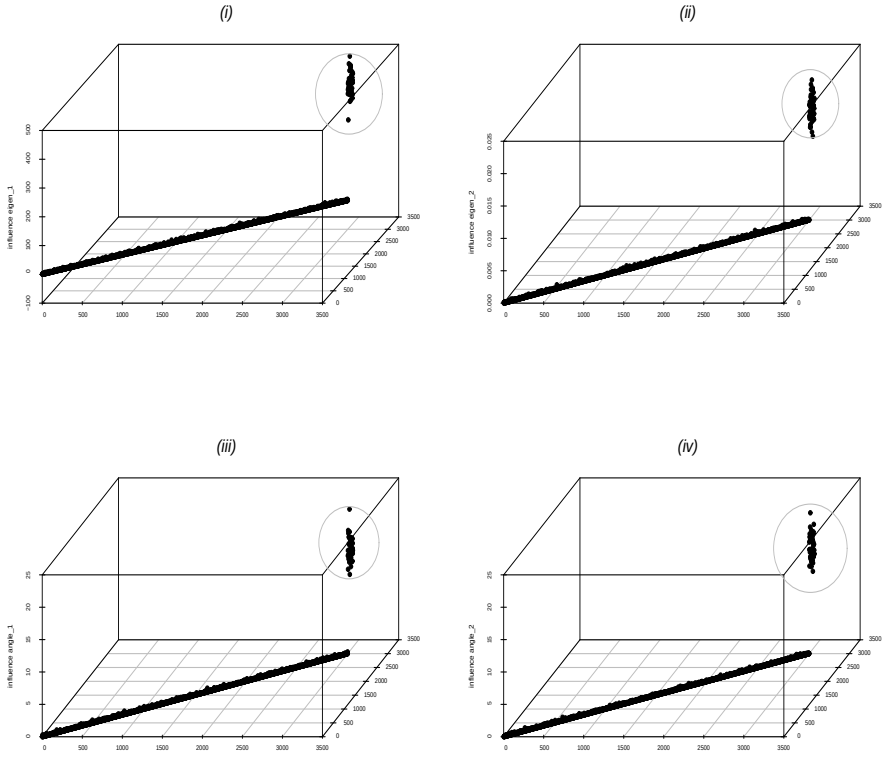


Figure 3.13: 3D scatterplot for $n = 3050,\ p = 100,\ m = 50$ – visualization of outliers for condition 1, scenario 1 by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

Condition 2 with $n = 3005,\ p = 100,\ m = 5$

Generation of a high-dimensional data set from a population where the good and bad data are closer to each other probably causes the suggested techniques in §3.4 and §3.5 not to perform. Nonetheless, Figure 3.14 indicates all 5 observations that are supposed to be outliers in the data set are considered for condition 2. The value of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ for observations $3001 - 3005$ are given in Table 3.5.

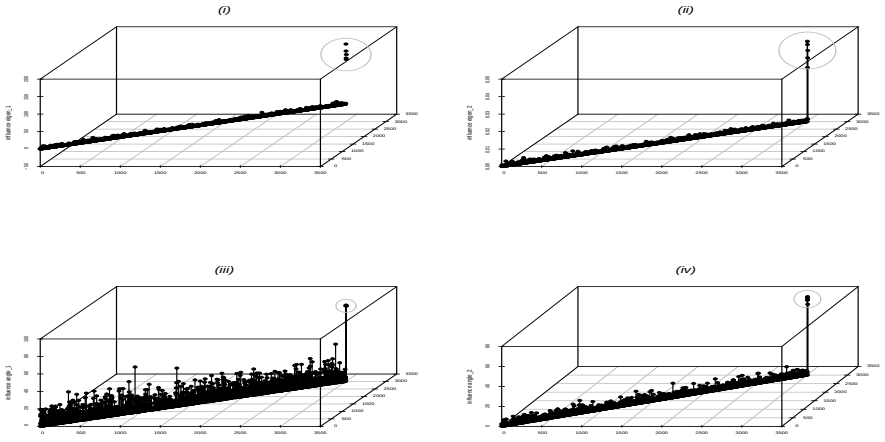


Figure 3.14: 3D scatterplot for $n = 3005,\ p = 100,\ m = 5$ – visualization of outlier for condition 2, scenario 1 by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

Table 3.5: $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ for observations 3001-3005–Condition 2, scenario 1 with $n = 3005, p = 100, m = 5$

Technique	3001	3002	3003	3004	3005
$\Delta^{*}_{1(i)}$	172.01	151.90	131.78	127.84	141.69
$\Delta^{**}_{1(i)}$	0.05	0.04	0.03	0.04	0.04
$\theta_{j(i)}$	86.51	87.38	86.77	86.42	86.42
$\theta^{*}_{j(i)}$	74.94	77.56	76.02	78.46	70.78

Condition 3 with $n = 3005,\ p = 100,\ m = 5$

Now, consider the high-dimensional data set containing outliers that are generated very close to the remaining good data. Figure 3.15 displays the index plot for each technique, where the outliers are denoted within the black circles. Surprisingly, almost all techniques are able to identify these 5 observations, i.e. the outliers, though the gap for some outliers with the good data is not large. Table 3.6 indicates the values of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ for observations $3001 - 3005$.

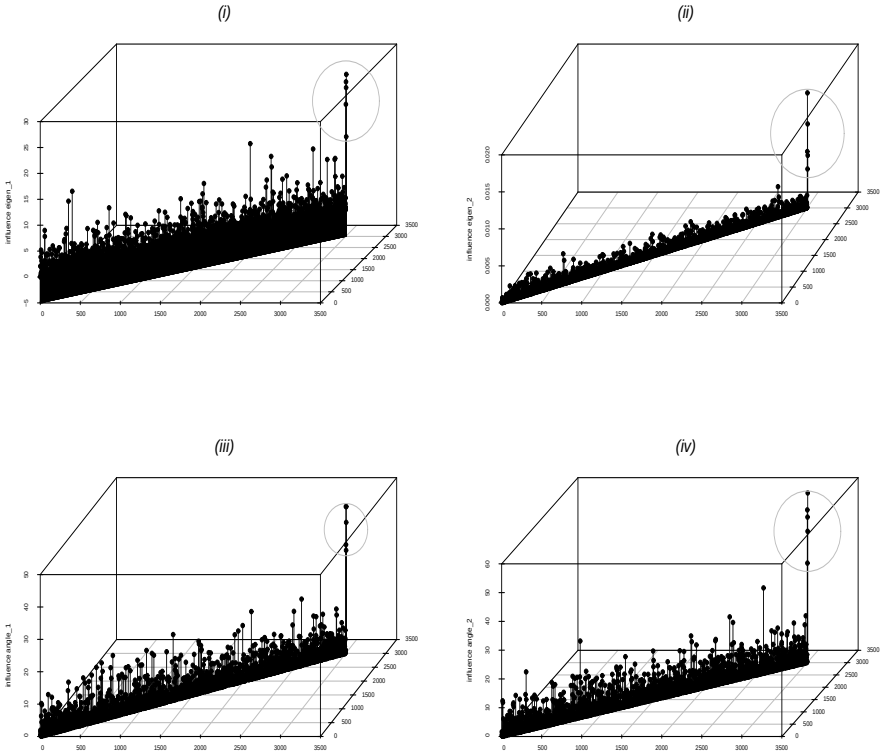


Figure 3.15: 3D scatterplot for $n = 3005,\ p = 100,\ m = 5$ – visualization of outliers for condition 3, scenario 1 by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

Note that Δ_i^* for observation 3004 is less than observation 2085 (see index plot (i) in Figure 3.15); only the values of observations 3001, 3002, 3003 and 3005 are a little bit larger than the remaining observations in the data set. The index plot of (iv) for $\theta^*_{1(i)}$ in Figure 3.15 also shows the value of $\theta^*_{j(i)}$ for observation 3002 to be very close to observation 2612.

Table 3.6: $\Delta^*_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^*_{1(i)}$ for observations 3001-3005–Condition 3, scenario 1 $n = 3005, p = 100, m = 5$

Technique	3001	3002	3003	3004	3005
$\Delta^*_{1(i)}$	20.50	24.84	23.71	14.24	26.27
$\Delta^{**}_{1(i)}$	0.01	0.02	0.01	0.01	0.01
$\theta_{j(i)}$	45.34	33.56	31.82	40.43	45.30
$\theta^*_{j(i)}$	53.06	34.57	50.55	45.61	58.92

3.11.2 Scenario 2

(i). $n = 3005,\ p = 100,\ m = 5$

Recall that scenario 2 generated a data set with different shapes and different locations. It is known the last 5 observations in this data set are the outliers. Figure 3.16 clearly displays these 5 outliers at the top of each index plot of $\Delta^*_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^*_{1(i)}$. It is noted there is a large gap between the outliers and the remaining observations.

The minimum value of $\Delta^*_{1(i)}$ among observations $3001 - 3005$ is 84 by observation 3003, whereas the maximum value is 472 by observation 3001. For $\Delta^{**}_{1(i)}$, the minimum value within observations $3001 - 3005$ is denoted by observation 3004 and the maximum value refers to observation 3001.

The value $\theta_{1(i)}$ for observations 3001, 3002, 3003, 3004 and 3005 are 40.96, 66.90, 55.50, 14.93 and 51.47 respectively, where observation 3002 exhibits the maximum value. The maximum value of $\theta^*_{1(i)}$ is

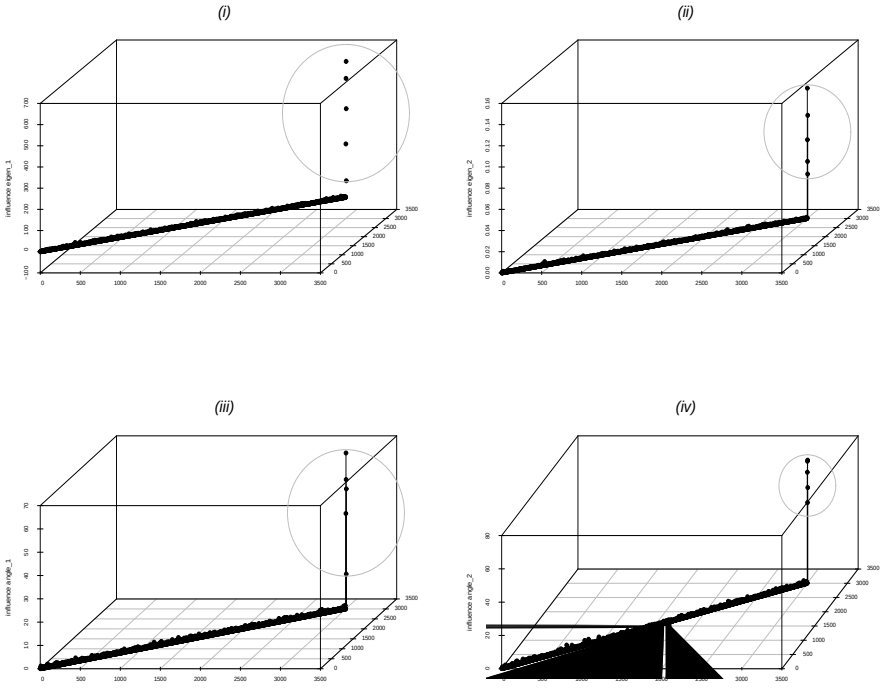


Figure 3.16: 3D scatterplot for $n = 3005$, $p = 100$, $m = 5$ – visualization of outlier for scenario 2 by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

represented by observation 3004; see index plot (iv) in Figure 3.16. Details of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ values for observations $3001-3005$ are given in Table 3.7.

Table 3.7: $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ for observations 3001-3005–Scenario 2 $n = 3005, p = 100, m = 5$

Technique	3001	3002	3003	3004	3005
$\Delta^{*}_{1(i)}$	252.07	561.41	641.47	418.15	77.84
$\Delta^{**}_{1(i)}$	0.12	0.07	0.05	0.04	0.10
$\theta_{j(i)}$	40.96	66.90	55.50	14.93	51.47
$\theta^{*}_{j(i)}$	66.70	73.31	48.45	73.87	57.41

(ii). $n = 3050,\ p = 100,\ m = 50$

Next, let the number of outliers, m become 50, now generate the new data set with sample size 3050, where 50 observations represent the outliers and the remaining observations represent good data. Figure 3.17 indicates all 50 outliers are located at the top of each index plot. It is noted the values of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ for the good data are almost zero.

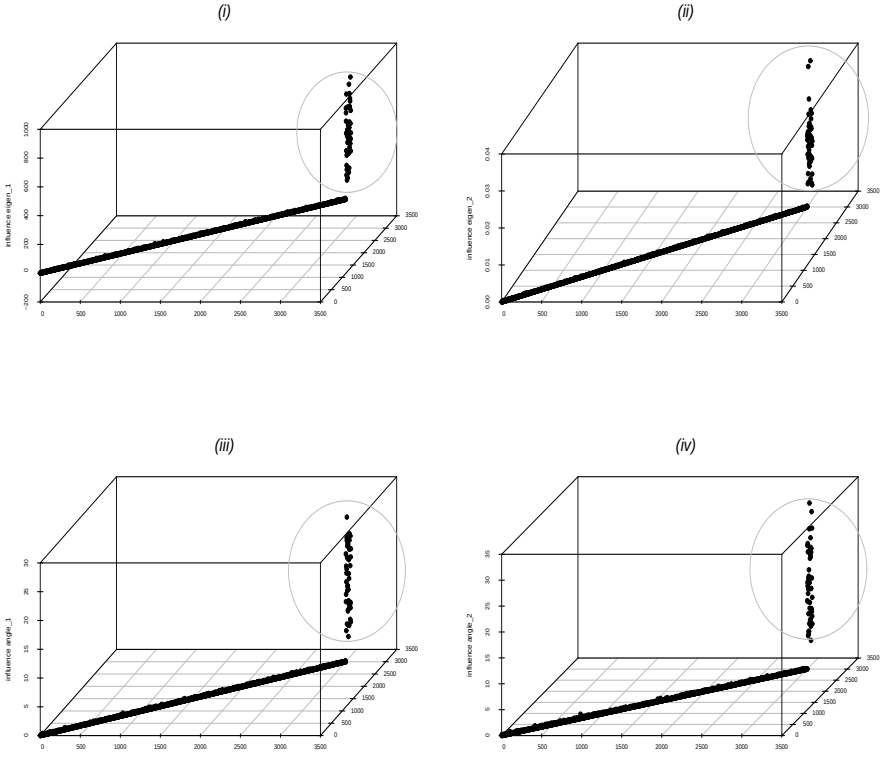


Figure 3.17: 3D scatterplot for $n = 3050,\ p = 100,\ m = 50$ – visualization of outliers for scenario 2 by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

3.11.3 Scenario 3 with $n = 3005,\ p = 100,\ m = 5$

Figure 3.18 also shows that $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ are capable of identifying outliers in a high-dimensional data set that contains outliers coming from different probability of laws. Note that outliers are denoted within the black circles.

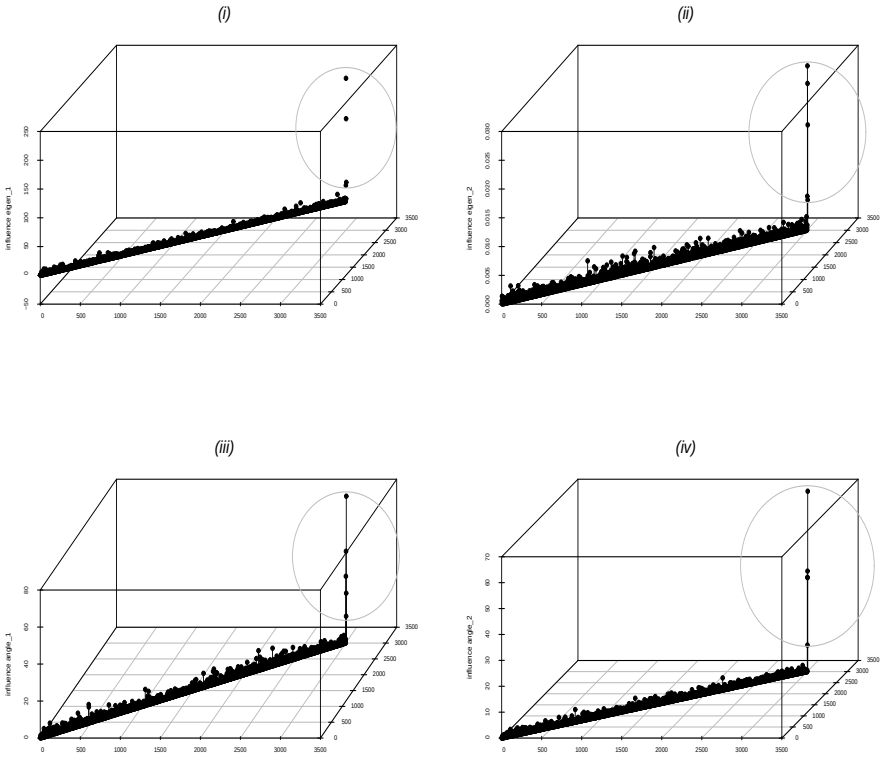


Figure 3.18: 3D scatterplot for $n = 3005,\ p = 100,\ m = 5$ – visualization of outliers for scenario 3 by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

3.12 Illustration using real data set

3.12.1 Hawkins Bradu Kass data

The first data set corresponds to a sample of 75 observations in 3 dimensions. Applying the influence angle $(\Delta^{*}_{1(i)})$, normalized influence angle $(\Delta^{**}_{1(i)})$, influence angle $(\theta_{1(i)})$ and modified influence angle $(\theta^{*}_{1(i)})$ results in the identification of 14 outliers among 75 observations. It is noted from Figure 3.19, that all 14 observations are located at the top of the index plot for each technique. The results agree well with Atkinson (1994), Rocke and Woodruff (1996) and Pena and Prieto (2001).

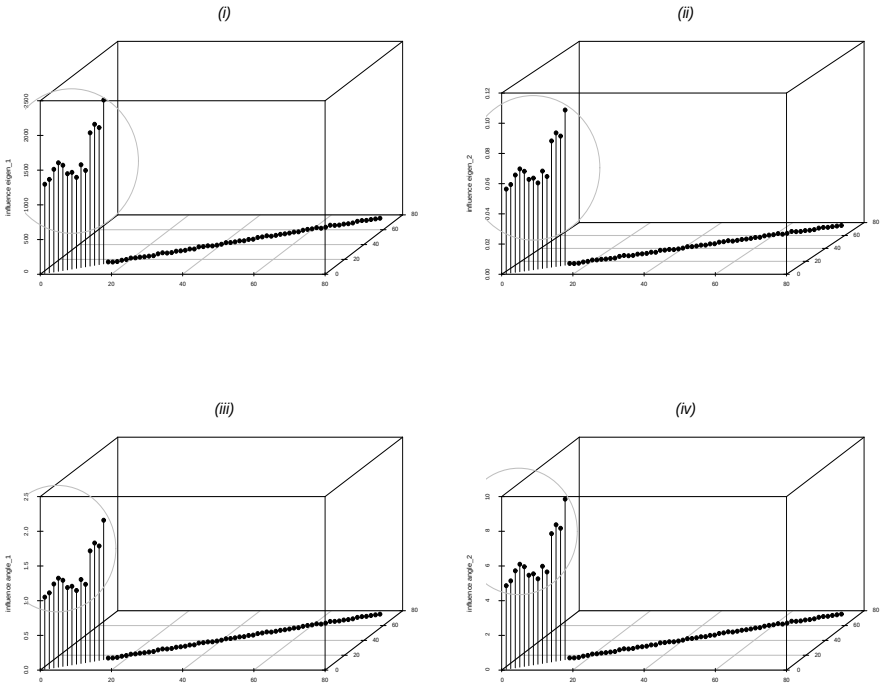


Figure 3.19: 3D scatterplot – visualization of outliers for Hawkins Bradu Kass data by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

3.12.2 Stackloss data

The Stackloss data set contains 21 observations in 3 dimensions. According to Rousseeuw and von Zomeren (1990), Hadi (1992) and Atkinson (1994) observations 1, 2, 3 and 21 are the outliers. However, Hawkins (1994) mentioned 9 observations as outliers in this data set. They are observations 1, 2, 3, 10, 15, 16, 18, 19 and 21. Pena and Prieto (2001) also declared more than 4 observations as outliers in the Stackloss data set (observations 1, 2, 3, 4, 13, 14, 20 and 21). Using techniques in §3.4 and §3.5, this reveals observations $1, 2, 3$and21 as the observations with the highest value of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$. This suggests that observations $1, 2, 3$ and 21 are outliers in this data set (see Figure 3.20).

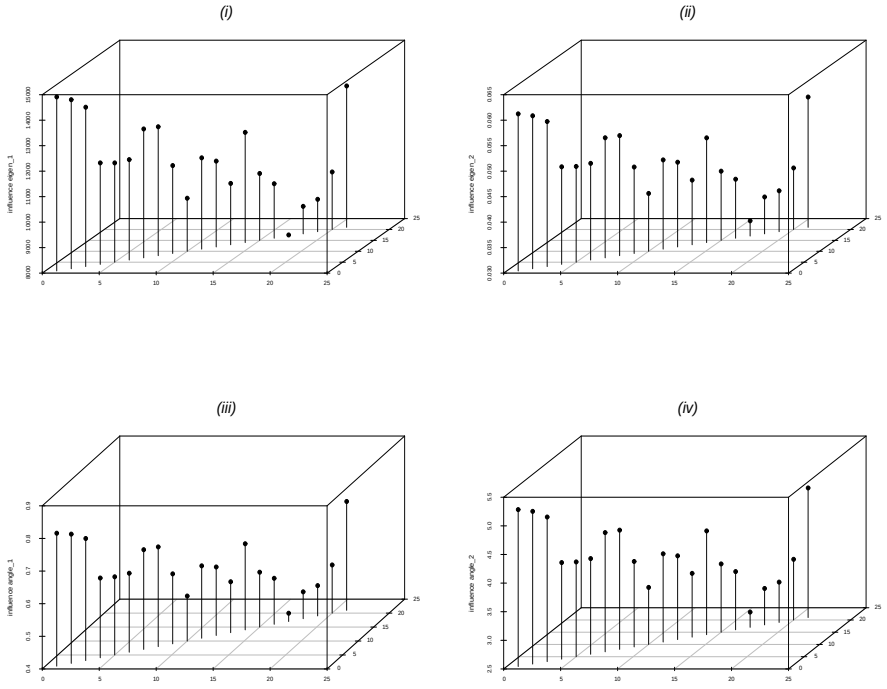


Figure 3.20: 3D scatterplot – visualization of outlier for Stackloss data by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

3.12.3 Salinity data

This data set comprises 28 measurements of water salinity and river discharge taken in North Carolina's Pamlico Sound. Rousseeuw and Leroy (1987) mentioned observations $3, 5$ and 16 as the outliers of the data set, whereas Hawkins (1994) and Pena and Prieto (2001) referred to observations 4, 5, 9, 10, 11, 16, 17, 19, 23 and 24, and observations 5, 10, 11, 15, 16, 17, 23 and 24 as outliers, respectively. However, Fung (1993) carried out the confirmatory analysis and concluded only observation 16 as the outlier in this data set. Plots in Figure 3.21 agree well with Fung (1993), where observation 16 has a large value of $\Delta^{*}_{1(i)}$, $\Delta^{**}_{1(i)}$, $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ among all observations.

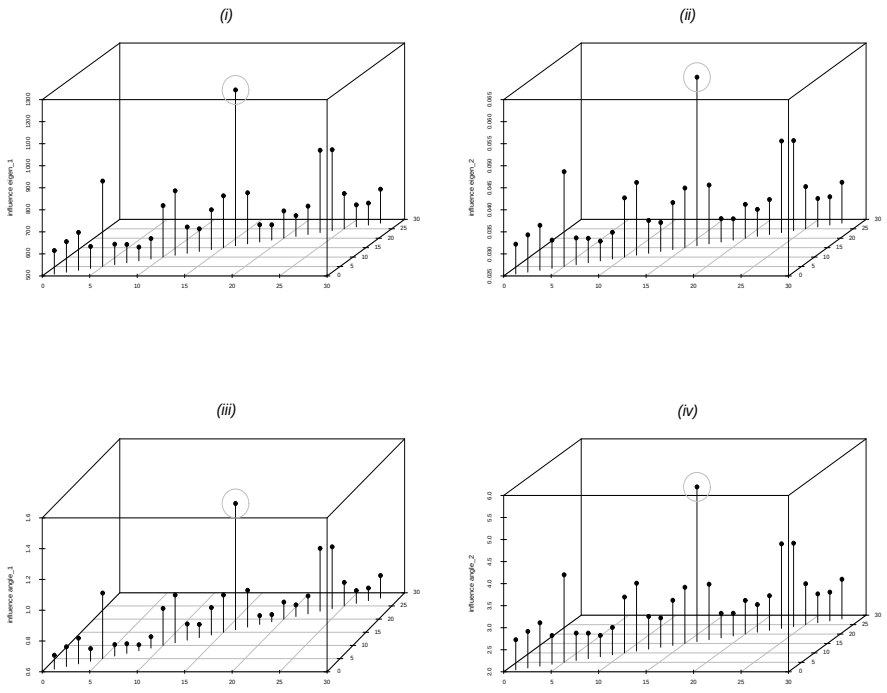


Figure 3.21: 3D scatterplot – visualization of outliers for Salinity data by using (i) $\Delta^{*}_{1(i)}$ (ii) $\Delta^{**}_{1(i)}$ (iii) $\theta_{1(i)}$ (iv) $\theta^{*}_{1(i)}$.

3.13 Conclusion

Sometimes, the identification of outliers is the main objective of the analysis, and whether to remove the outliers or for them to be down-weighted prior to fitting a non-robust model. This chapter does not differentiate between the various justifications for outlier detection. The aim is to advise the analyst of observations that are considerably different from the majority. Note that the techniques in §3.4 and §3.5 are, therefore, exploratory. It is applicable to a wide variety of settings. Techniques used in this chapter are performed on large and small data sets. They are used as a measurement, i.e. distance between observations. In this chapter, observations that are far away from the remaining data are considered to be outliers.

If the ith observation is a potential outlier, their values for $\Delta^{*}_{1(i)}$ and $\Delta^{**}_{1(i)}$ are all situated at the top of the index plot; see illustration of index plots in §3.8 until §3.12. This is because an outlier causes $\lambda_1 - \lambda_{1(i)}$ values to be larger than other observations. Note that $\lambda_{1(i)}$ value is smaller for an outlier. This follows that $\Delta^{*}_{1(i)}$ and $\Delta^{**}_{1(i)}$ become larger. Another thing is $\Delta^{**}_{1(i)}$ is the normalized value of $\Delta^{*}_{1(i)}$, thus if $\Delta^{*}_{1(i)}$ values for observation i is large, the $\Delta^{**}_{1(i)}$ value for observation i is also large than other observations.

Notice that the angles of $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ for all examples are between 0 and 90, which correspond to the principal component score, are all positive. The deletion of ith observation causes the $\lambda_{1(i)}$ to become smaller, and $\lambda^{*}_{1(i)}$ as well as the angle, i.e. $\theta_{1(i)}$ and $\theta^{*}_{1(i)}$ to get larger. As a consequence, the ith observation with a large angle, i.e. an outlier, is located at the top of the index plot.

Instead of identifying the outlier, another issue that one should consider is to determine whether the outlier is sufficiently extreme or influential to warrant further action. Chapter 2 gives some suggestions of what action should be taken if outliers exist, but this is not further discussed for each example in §3.8 until §3.12.

Additionally, it is noted that some techniques for the identification of outliers are also available for finding clusters. Clustering analysis also uses the distance as a measurement between observations to develop clusters among them. Jolliffe (2002) shows that principal component analysis is also capable of finding clusters in the data set. Note that the influence angle in §3.5.1 is partially developed by the principal component score and the outliers appear to form a cluster, separated from the other observations in the data set. In the next two chapters, $\theta_{1(i)}$ will be used as a tool to classify observations in the data set. First, Chapter 4 will briefly discuss existing measurement tools for clustering and the common clustering techniques. Next, Chapter 5 will use $\theta_{1(i)}$ to calculate the distance between observations for clustering purposes.

Chapter 4

An Overview of Proximity Measures and Clustering Algorithms

4.1 Introduction

Cluster analysis has been widely used in several disciplines, such as statistics, software engineering, biology, psychology and other social sciences, in order to identify natural groups in large amounts of data. These data sets are constantly becoming larger, and their dimensionality prevents easy analysis and validation of the results.

There are two major challenges in clustering. The first is identifying clusters in high-dimensional data sets is a difficult task because of the curse of dimensionality. According to Domeniconi, Papadopoulos, Gunopulos and Ma (2004), in high dimensional spaces, it is highly likely that, for any given pair of points within the same cluster, there exist at least a few dimensions on which the points are far apart from each other. As a consequence, distance functions that equally use all input features may not be effective. Furthermore, several clusters may exist in different subspaces, comprised of different combinations of features.

The second is a new dissimilarity measure is needed as some traditional distance functions cannot capture the pattern dissimilarity among the objects. For instance, some objects are not close to each other if they are measured by distance functions such as Euclidean, Manhattan, or Cosine. Chapter 4 and 5 deal with the latter challenge. Chapter 4 provides some choices of dissimilarity measures.

There are many techniques one can use to construct dissimilarity or similarity measures for continuous and binary data. Even though definitions of dissimilarity and similarity vary from one clustering approach to another, in most of these approaches the concept of dissimilarity is based on distances, i.e, Euclidean distance or Cosine distance. The other demand is that the dissimilarity measurement should have the ability to deal with a variety of data types, i.e. binary, ordinal and categorical values.

The clustering analysis covers multiple numbers of different algorithms and methods for grouping observations of similar kinds into respective clusters. Clustering algorithms can be classified into five main groups. They are partitioning methods, hierarchical methods, model-based methods, density-based methods and grid-based methods.

Partitioning methods : Given n objects, these methods construct k partitions of the data, by assigning objects to groups, with each partition representing a cluster. Generally, each cluster must contain at least one object, and each object may belong to one and only one cluster. However, as it is unfeasible to test all partitions for even moderate n and k, partitioning algorithms do not consider all partitions and can only find local optima.

Hierarchical methods : These methods create a hierarchical decomposition of the objects in the data set by either merging or splitting clusters sequentially. These are referred to as agglomerative and divisive hierarchical methods, respectively.

Model-based methods: These methods formulate a model and fit it to the data by estimating suitable parameters. The models are typically statistical mixture models or neural networks. Methods involving statistical

models are sometimes said to perform a "conceptual clustering", as clusters are given distributions that govern their behavior and may suggest some real-world meaning. Neural networks were originally motivated by an abstract attempt to model the way that brain clusters objects.

Density-based methods : In these methods clusters are defined as dense regions in the data space, i.e. a larger than expected number of points in a given subspace. Care must be taken (by statisticians) to remember that here "density" refers to the physical concept of density rather than a particular statistical distribution.

Grid-based methods : These methods are characterized by the practice of dividing the data space into a finite number of cells to form a grid. All clustering operations are then performed on the cells of this grid. Of course it is perhaps dangerous to pigeon-hole new algorithms as having to belong to one and only one of the above families. Furthermore, they serve to define some of the distinct styles in clustering , each style having its own set of qualities and problems.

4.2 What is a clustering problem?

The terms cluster, group and class have been used without any definite formal definition (Everitt, 1993). Cormack (1971) and Gordon (1999) however, reached an agreement for a definition of clusters, defining them as internal cohesion-homogeneity and external isolation-separation. On the other hand, Everitt, Landau and Leese (2001) emphasized that it is not appropriate to identify clusters through the plane but the feature of the recognition process appears to involve the assessment of a relative distance between points.

Clustering may simply represent a convenient method for organizing a large data set so that it can easily be understood and information can efficiently be retrieved . If the data can validly be summarized by clustered sets of that data, then it would probably give a more precise definition

about the large data set. Also, it would easily help in clarifying a product for a particular type of consumers. It is important to summarize a data set since a growing number of large databases can now easily be accessed in many areas of sciences.

To produce clustering, one has to find the distance measurement between observations in the data set. There are many types of distances that deal with continuous, categorical or mixed variables.

4.3 Proximity Measures

Objects (events) are usually represented as points (vectors) in a multi-dimensional space, where each dimension represents a distinct attribute (variable) describing the object. For simplicity, it is usually assumed that values are present for all attributes. Techniques for dealing with missing values are described in Jain and Dubes,1988; Kaufman and Rousseeuw,1990. Thus, a set of objects is represented as an n by p matrix, where there are n rows, one for each object, and p columns, one for each attribute.

The matrix has different names, e.g., pattern matrix or data matrix, depending on the particular field. The data is sometimes transformed before being used. One reason for this is that different attributes may be measured on different scales, e.g., centimeters and kilograms. In cases where the range of values differs widely from attribute to attribute, these differing attribute scales can dominate the results of the cluster analysis, and it is common to standardize the data so that all attributes are on the same scale.

A simple approach to such standardization is, for each attribute subtract off the mean of the attribute values and divide by the standard deviation of the values. While this is often sufficient, more statistically robust approaches are available, as described in Kaufman and Rousseeuw (1990). Another reason for initially transforming the data is to reduce the number of dimensions, particularly if the initial number of dimensions is large.

While cluster analysis sometimes uses the original data matrix, many clustering algorithms use a similarity matrix, say, S, or a dissimilarity matrix, say, D. For convenience, both matrices are commonly referred to as a proximity matrix. A proximity matrix, say, P, is an n by n matrix containing all the pairwise dissimilarities or similarities between the objects being considered.

If x_i and x_j are the ith and jth objects, respectively, then the entry at the ith row and jth column of the proximity matrix is the similarity, s_{ij}, or the dissimilarity, d_{ij}, between x_i and x_j. For simplicity, p_{ij} is represent either s_{ij} or d_{ij}.

For completeness, objects are sometimes represented by more complicated data structures than vectors of attributes, e.g., character strings or graphs. Determining the similarity (or dissimilarity) of two objects in such a situation is more complicated, but if a reasonable similarity (or dissimilarity) measure exists, then a clustering analysis can still be performed. In particular, clustering techniques that use a proximity matrix are unaffected by the lack of a data matrix.

The notion of similarity and dissimilarity (distance) seems fairly intuitive. However, the quality of a cluster analysis depends critically on the similarity measure that is used and, as a consequence, many different similarity measures have been developed for various situations.

The proximity measure (and the type of clustering used) depends on the attribute type and scale of the data. The three typical types of attributes are binary (two values, e.g., true and false), discrete (a finite number of values, or integers, e.g., counts.) and continuous (an effectively infinite number of real values, e.g., weight). The common data scales too are divided into qualitative and quantitative categories. The qualitative category is further divided into nominal (the values are just different names, e.g., colors or zip codes) and ordinal (the values reflect an ordering, nothing more, e.g., good, better, best). On the other hand one can classify the quantitative category into interval (the difference between values is mean-

ingful, i.e., a unit of measurement for example, temperature on the Celsius or Fahrenheit scales) and ratio (the scale has an absolute zero so that ratios are meaningful; examples are physical quantities such as electrical current, pressure, or temperature on the Kelvin scale)

The most commonly used proximity measure, at least for ratio scales (scales with an absolute 0) is the Minkowski metric, which is a generalization of the distance between points in Euclidean space.

$$p_{ij} = \Big(\sum_{k=1}^{d} |x_{ik} - x_{jk}|^r \Big)^{\frac{1}{r}}$$

where, r is a parameter, d is the dimensionality of the data object, and x_{ik} and x_{jk} are, respectively, the kth components of the ith and jth objects, x_i and x_j.

For $r = 1$, this distance is commonly known as the L_1 norm or city block distance. If $r = 2$, the most common situation, then one has the familiar L_2 norm or Euclidean distance. Occasionally one might encounter the $L_{\max}$ norm (L_∞ norm), which represents the case $r \to \infty$.

The r parameter should not be confused with the dimension, d. For example, Euclidean, Manhattan and supremum distances are defined for all values of $d = 1, 2, 3, \ldots, p$ and specify different ways of combining the differences in each dimension (attribute) into an overall distance. Finally, note that various Minkowski distances are metric distances. In other words, given a distance function, *dist*, and three points $a, b,$ and c these distances satisfy the following three mathematical properties:

- reflexivity ($dist(\mathbf{a}, \mathbf{a}) = 0$),
- symmetry ($dist(\mathbf{a}, \mathbf{b}) = dist(\mathbf{b}, \mathbf{a})$), and
- the triangle inequality ($dist(\mathbf{a}, \mathbf{c}) \leq dist(\mathbf{a}, \mathbf{b}) + dist(\mathbf{b}, \mathbf{a})$).

Not all distances or similarities are metric, i.e, the Jaccard measure. This introduces potential complications in the clustering process since in

such cases, **a** similar (close) to **b** and **b** similar to **c**, does not necessarily imply **a** similar to **c**.

The next chapter discusses how an alternative measurement to distance can provide useful tool to develop clusters even when the data set has mixed variables. By using this measurement, clusters can be obtained easily and it seems to work well in practice. Before going further with the discussion of the new measurement in the next chapter, this chapter will briefly explain the choices of clustering algorithms.

4.4 Choices of Clustering Algorithm

At the moment, there is no clustering technique that is universally applicable in uncovering the variety of structures present in multidimensional data sets (Jain, Murty and Flynn, 1999). Halkidi, Batistakis and Vazirgiannis (2001) and Jain et al. (1999), summarized and carried out comparisons of groups of algorithms; these are partitional clustering, hierarchical clustering, density-based clustering and grid-based clustering.

Clustering algorithms are increasing as there are many fields with different problems, patterns and types of data. For example, Rose (1998) presented the deterministic annealing approach to clustering and its extension, via introduction of appropriate constraints on the clustering solution, to attack a large and important set of optimization problems. Later, Xing and Karp (2001) proposed a new algorithm that iterated between two computational processes, namely feature filtering and clustering. In 2002, Wang, Wang, Yang and Yu tried to cluster objects that exhibited similar pattern on subset of dimension; they introduced *pCluster*. Recently, Fang, Liu, Yang, Luo and Li (2006) introduced the clustering algorithm that is based on graph structure.

This chapter discusses two types of clustering methods - the partitioning and hierarchical methods. The partitioning method divides the data into k clusters, so that the objects of the same cluster are close to each other

and objects of different cluster are well separated. The findings of the partition method can be easily viewed through graphical display. The hierarchical methods develop a dendogram, which is a tree of which the leaves are the data objects and each branch represent a cluster.

4.4.1 Partitioning Method

The partitioning methods try to improve the partitioning accuracy by moving observations from one cluster to another by iterative relocation to produce original partitions. The algorithm for this method divides the data set into k clusters, where the integer k needs to be specified by the user. Typically, the user runs the algorithm for a range of k-values and chooses the best. For each k, the algorithm carries out the clustering and also yields a "quality index", which allows the user to select a value of k later. The most familiar techniques in partitioning methods are K-means and K-medoids.

The K-means clustering algorithm is described in detail by Hartigan (1975). The objective of the K-means algorithm is to divide n points in p dimensions into k clusters so that the within-cluster sum of squares is minimized. The advantage of the k-means algorithm is that its time complexity is $O(n)$, making it slightly more scalable and it can work with any L_p norm.

Criticizing the method, one would complain about having to provide the number of clusters k. A common practice is to use a hierarchical clustering method to suggest a suitable k. The second criticism is, the method being sensitive to outliers, also tends to find spherical clusters of equal size and it has to find initial centroids to start the algorithm. Hartigan and Wong (1979) suggest using actual objects as initial cluster centres. These could be selected randomly. Hence the final set of clusters is dependent on the initial set of clusters having to convert to the distance space every time one needs to know how to cluster an object. However, this raises the

computational cost.

The K-medoids method is appealing because it is more robust and it allows a good characterization of all clusters that are not too elongated and makes it possible to isolate outliers in most situations (Kaufman and Rousseeuw, 1990). A few examples of algorithms in the k-medoids technique are *pam*, *clara* and *fanny*. See Struyf, Hubert and Rousseeuw (1996).

Partitioning Around Medoids (*pam*) is based on the search for k representative objects, called medoids, among the objects of the data set. In *pam*, the variables do not have to be continuous; they can be discrete variables. In this method, first , one needs to decide the number of clusters k required for the data set. Later, an observation will be selected as representative object (medoids). If one decides to build two groups, then two observations are needed as the representative objects (medoids).

The role of the representative object is being a medoid for the group. *pam* will select the representative objects that give a minimal total dissimilarity of all objects to their nearest medoid. Then dissimilarities between each observation to these two medoids are calculated. Later, each observation is assigned to the cluster where it has the smallest dissimilarity to the medoid; that is, observation i is put into cluster v_i when medoid m_{vi} is nearer to i than the other medoid m_{wi}. *pam* is more suitable to be used by small sample size data.

The k-medoids algorithm has a number of disadvantages, namely that the algorithm needs the number of clusters k to be entered as input. This requires a good guess from the user, which might not be available. Next , the time complexity per iteration of the algorithm is $O(n^2)$. For each iteration there are $k(n-k)i \leftrightarrow h$ swaps to consider; calculating each swap involves accessing $(n-k)$ distances, making one iteration $O(k(n-k)^2)$.

Clustering Large Application (*clara*) shares the same algorithm with *pam*. The user must provide the number of groups, k needed and determine the observations for representative objects (medoids). The advantage of *clara* is ,it can handle large sample size data. For a data set with

more than 250 observations, it is appropriate to use *clara* (Kaufman and Rousseeuw, 1990). This is because it does not store all possible dissimilarity matrixes but only the actual measurement.

The other method that is similar to the *pam* and *clara* is Fuzzy Analysis (*fanny*). The speciality of this method is that it can give a detailed explanation of which cluster the object should be assigned to. This method can exhibit the percentage or probability of the observations to be in particular groups. For example, there are 3 clusters to be built; this method can produce a result that says observation 1 belongs for 2% to cluster 1, for 95% to cluster 2 and for 3% to cluster 3. In *fanny* method, the observation is assigned to the group with the highest percentage or probability. So, in the example given above, observation 1 will be assigned to cluster 2.

This algorithm shares the usual problem of having to specify the number of clusters k. However, its attractive features are that the time complexity of the algorithm is only O(n) and the measure of confidence in each assignment of an object to a cluster is readily available through the membership coefficient and summaries of the clusters are available (i.e. the cluster centres) at the end of the algorithm.

4.4.2 Hierarchical Method

The hierarchical method works by grouping the observations in the data set into a hierarchy of clusters which is based on a dissimilarity measure. Hierarchical algorithms do not build a single partition with k cluster but they deal with all possible values of k in the same run. One might think the partitioning method as outdated as all possible values of k have to be found in a single run. However, this is not true because a clustering formed "along the way" is not necessarily very good (Kaufman and Rousseeuw, 1990). In fact, a partitioning method tries to select the best clustering with k clusters, and this is not the purpose of the hierarchical method. Another weakness of the hierarchical method is it can never redo

what was done in previous steps. For example, once agglomerative algorithm has joined two observations, they cannot be split.

There are two kinds of hierarchical methods: the agglomerative method and the divisive method. The agglomerative method is a bottom-up approach. That means it begins by placing each observation in its own cluster, and combines the clusters into larger ones step by step. On the other hand, the divisive method is a top-down approach. It starts by grouping all objects into one cluster, and separates the cluster into smaller ones step by step. An example of divisive method is Divisive Analysis (DIANA).

A brief analysis of these methods would show them to be of $O(n^3)$ time complexity, $O(n^2)$ on the algorithm for each iteration. This is even before the potentially expensive calculations that may take place albeit on a subset of the n original objects. The algorithm also needs the group averages between an object and the new and existing clusters to be recalculated after an object is moved. This will be costly in terms of the number of calculations and the amount of storage required.

The agglomerative method has some limitations, which includes the inability to handle outliers like the K-medoids method. Therefore the clustering results are easily affected by the outliers. For instance, by having outliers between clusters, two clusters may be grouped into one. It is also sensitive to the cluster size. For example, if small clusters are situated close to a large cluster, the agglomerative method cannot notice the small clusters.

The advantages of hierarchical method are that it is easy to implement computationally, and it is able to tackle large sample size data than the k-medoids method and it is unsupervised, in the sense that one can run the algorithm without having to provide the number of clusters. One disadvantage of the hierarchical method is, it has $O(n^3)$ time complexity. Even though the order of the distance matrix decreases with each iteration, the cost on iteration k is $O((n-k)^2)$, and there are $(n-k)$ iterations before getting to k. Secondly, the clusters produced are heavily dependent

on the metric $D_{i,j}$. Different metrics can produce very different clusters. For instance, the complete-link metric tends to produce spherical clusters, whereas the single-link metric produces elongated clusters. Therefore, one still has to decide which clusters, if any, one is going to choose.

4.5 Conclusion

This chapter briefly reviewed the proximity measures and clustering methods. Note that clustering may simply represent a convenient method for organizing a large data set so that it can easily be understood and information can efficiently be retrieved. If the data can validly be summarized by clustered sets of that data, then it would probably give a more precise definition about the large data set.

This chapter also described some hierarchical and partitioning clustering algorithms. The partitioning method divide the data set into k clusters, where the integer k needs to be specified by the user. The hierarchical approach do not build a single partition with k cluster but they deal with all values of k in the same run.

Even though one might label the partitioning method as outdated, as all possible values of k are searched in a single run, but the partitioning method tries to select the best clustering with k cluster, and this is not the purpose of the hierarchical method.

The next chapter will consider the partitioning method to examine the new dissimilarity matrix and a new approach called Influence Angle Cluster Approach (*iaca*).

Chapter 5

Influence Angle Cluster Approach

5.1 Introduction

In attempting to identify clusters of observations which may be present in a data is what is important is in knowing how close individuals are to each other. Two individuals are close either when their dissimilarity or distance is small, or when their similarity is large. Data set for clustering can either be from an $n \times p$ objects by attribute matrix, where rows represent objects and columns represent variables, or a $n \times n$ dissimilarity matrix where $d(i, j) = d(j, i)$ measures the "difference" or dissimilarity between the objects i and j.

To produce clustering, one has to find the measurement between observations in the data set. This can normally be obtained either by similarity or dissimilarity matrix. Similarity matrix is represented by correlation, where high correlation indicates similarity. The correlation focuses on the pattern rather than proximity. Therefore, to gain proximity, one needs to use the dissimilarity or distance. There are many types of distances, such as Euclidean, Maximum, Manhattan, Canberra and Binary. These distances deal with continuous variables (Everitt et al., 2001).

There are also situations where the data set has mixed variables, i.e. some variables are continuous and some are categorical. There are a few

approaches to handle this situation. One possibility would be to dichotomize all variables and use a similarity measure for binary data; another would be to construct a dissimilarity measure for each type of variables and combine these, either with or without differential weighting, into a single coefficient.

Let us consider the similarity measure proposed by Gower (1971) as an example to handle data with mixed types of variable. Gower (1971) defines a similarity measure by

$$s_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} s_{ijk}}{\sum_{k=1}^{p} w_{ijk}}.$$

Given that s_{ijk} is the similarity between the *i*th and *j*th individual measured using the *kth* variable and w_{ijk} is typically one or zero depending on whether or not the comparison is considered valid. The value of w_{ijk} is set to zero if the outcome of the *kth* variable is missing for either or both of individuals *i* and *j*. In addition, w_{ijk} can be set to zero if the *kth* variable is binary and it is thought appropriate to exclude negative matches. For binary variables and categorical variables with more than two categories, the component similarities, s_{ijk}, take the value one when the two individuals have the same value and zero otherwise. For continuous variables, Gower suggests using the similarity measure

$$s_{ijk} = 1 - |x_{ik} - x_{jk}|/R_k,$$

where R_k is the range of observations for the *kth* variable.

Kaufman and Rousseeuw (1990) also provide an algorithm called *daisy* which computes a dissimilarity matrix from objects-by-attributes. The main feature of *daisy* is the ability to handle nominal, ordinal, asymmetric binary and ratio-scaled variables, even if different types of variables occur in the same data set.

In §3.5.1, Chapter 3, reference was made to 'influence angle'. It is suggested here that 'influence angle' may be used as dissimilarity matrix and

it is defined as Influence Angle Cluster Approach (*iaca*) instead of influence angle to avoid confusion with §3.5.1. The influence angle is developed by principal component score and Jolliffe (2002) proved that principal component analysis is also capable of finding clusters in the data set. This chapter introduces *iaca* as a dissimilarity matrix in §5.2.

There have been many clustering techniques suggested. This chapter will also consider partitioning methods which divide data into several subsets. Unlike traditional hierarchical methods in which clusters are not revisited after being constructed, partitioning methods are otherwise. Partitioning methods are further divided to k-medoids and k-means. K-medoids have two advantages , namely the ability to cover any attribute types and having the embedded resistance against outliers. Two early versions of k-medoids methods are described in this chapter, i.e. Partitioning Around Medoids (*pam*) and Clustering Large Application (*clara*). §5.3 will briefly explain these two algorithms.

It is also very important to examine whether the combination of *iaca* and various clustering methods are likely to lead to interesting and informative classifications. Therefore, §5.4 will discuss in general about cluster validation. This chapter then continues by showing some examples using simulation and real data sets in §5.6 and §5.7 before it concludes.

5.2 Influence Angle Cluster Approach as a Dissimilarity Measure

Recall the definition of influence angle given by equation 3.35 in chapter 3 as shown below

$$\theta_{j(i)} = \cos^{-1}\left\{ \frac{l_{ij}/\lambda^*_{j(i)}}{\sqrt{\sum_{k=1}^{p} l_{ik}^2/(\lambda^*_{j(i)} + (\lambda_k - \lambda_j))^2}} \right\}, \tag{5.1}$$

where $j = 1, 2, \ldots, p$; $i = 1, 2, \ldots, n$, and l_{ij} is the principal component score of the omitted observation in the principal component decomposition of the complete data **X** and

$$\lambda^*_{j(i)} = \lambda_j - \frac{1}{n-1}(l^2_{ij} - \lambda_j) - \frac{1}{2(n-1)^2} l^2_{ij}[1 + \sum_{k \neq j} \frac{l^2_{ij}}{\lambda_k - \lambda_j}] + O(\frac{1}{n^3}).$$

The $\theta_{j(i)}$ is defined as the angle between 0 and 180 degrees that satisfies the relationship $v_j^T v_{j(i)} = \|v_j\| \, \|v_{j(i)}\| \cos \theta_{j(i)}$ where $\|.\|$ refers to the vector length.

Now consider the Influence Angle Cluster Approach (*iaca*) as a dissimilarity matrix. The *iaca* is developed in an algorithm as given below:

- **Step 1**: Assuming **X** is the complete data set, one can compute the vector of eigenvalues of $\mathbf{X}^T\mathbf{X}$. Supposed $\mathbf{\Lambda} = \{\lambda_\mathbf{1}, \lambda_\mathbf{2}, \ldots, \lambda_\mathbf{p}\}$ be the vector of eigenvalues of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{V}$ represents eigenvectors corresponding to the Λ. Next, estimate the eigenvalues for $\mathbf{X}^T_{(i)}\mathbf{X}_{(i)}$ after deleting *i*th observation. The vector of eigenvalues of $\mathbf{X}^T_{(i)}\mathbf{X}_{(i)}$ is $\mathbf{\Lambda_{(i)}} = \left\{\lambda_{\mathbf{1(i)}}, \lambda_{\mathbf{2(i)}}, \ldots, \lambda_{\mathbf{p(i)}}\right\}$ and $\mathbf{V}_{(i)}$ represents eigenvectors corresponding to the $\Lambda_{(i)}$ and $i = 1, 2, \ldots, n$.

- **Step 2**: Compute the principal component scores of the omitted observation in the principal component decomposition of the complete data set, l_{ij}. Next, determine the $\cos(\theta_{j(i)})$, where one can fix $j = 1$ as this chapter is considering the principal eigenvalues. $\theta_{j(i)}$ represents the influence angle between the *j*th eigenvector for the complete data set and the *j*th eigenvector when the *i*th observation is deleted for all $i = 1, 2, \ldots, n$.

Next, one can develop *iaca* as follows: one needs to find the difference between the influence angle of each observation as $|\theta_{j(i)} - \theta_{j(i^*)}|$ where $1 \leq i, i^* \geq n$ and the absolute value signs are used since one can consider the difference between influence angles to be an unsigned scalar value and it can be called *iaca*.

$$\Psi_j = \begin{pmatrix} 0 & |\theta_{j(1)} - \theta_{j(2)}| & \cdots & |\theta_{j(1)} - \theta_{j(n)}| \\ |\theta_{j(2)} - \theta_{j(1)}| & 0 & \cdots & |\theta_{j(2)} - \theta_{j(n)}| \\ \vdots & \vdots & \ddots & \vdots \\ |\theta_{j(n)} - \theta_{j(1)}| & |\theta_{j(n)} - \theta_{j(2)}| & \cdots & 0 \end{pmatrix}$$

Definition 1. An $n \times n$ matrix $\Psi_j = \psi^j_{(i)(i^*)}$ is called as the dissimilarity matrix of *iaca* if and only if there exist $\theta_{j(1)}, \ldots, \theta_{j(n)} \in R^d, n \geq 2$ points in some *d*-dimensional spaces, such that $\psi^j_{(i)(i^*)} = |\theta_{j(i)} - \theta_{j(i^*)}|$. The smallest *d* for which this is possible is the dimensionality of Ψ_j.

From Definition 1, if $\Psi_j = \psi^j_{(i)(i^*)}$ is the dissimilarity matrix of *iaca* then

- $\psi^j_{(i)(i^*)} = |\theta_{j(i)} - \theta_{j(i^*)}| \geq 0$ (Ψ has nonegative entries)
- Ψ_j is a hollow matrix where all elements on the diagonal of Ψ_j are equal to zero. The elements of Ψ_j are given by $\Psi_{n\times n} = \left\{\psi^j_{(i)(i^*)}\right\}$; $\psi^j_{(i)(i^*)} = 0$ if $i = i^*, 1 \leq i, i^* \leq n$.
- The trace of Ψ is zero (by the above property).
  $\text{tr}(\psi_j) = \sum_{i=i^*} |\theta_{j(i)} - \theta_{j(i^*)}| = \psi^j_{(1)(1)} + \psi^j_{(2)(2)} + \ldots + \psi^j_{(n)(n)} = \sum_{i=i^*} \psi^j_{(i)(i^*)} = 0$
- Ψ_j is symmetric ($\Psi_j = \Psi_j^T$), where $\psi^j_{(i)(i^*)} = |\theta^j_{(i)} - \theta^j_{(i^*)}| = |\theta^j_{(i^*)} - \theta^j_{(i)}| = \psi^j_{(i^*)(i)}$

Next, one can construct a partition of n objects into a set of k clusters by using *iaca* that is applied in the clustering algorithms which will be explained in the next section.

5.3 Partitioning methods

Given n objects, these methods construct k partitions of the data, with each partition representing a cluster. The general method works as follows: given the number of clusters, an initial partition is made; objects are then moved between partitions in an attempt to improve some objective function.

To find a global optimum for the objective function one needs to consider all $N(n,k)$ possible partitions, where

$$N(n,k) = \frac{1}{k!}\sum_{i=1} k(-1)^{(k-i)}\binom{k}{i}i^n$$

$N(n;k)$ is one of Stirling's numbers of the second kind. See Jensen (1969). With increasing n this soon becomes unfeasible, so inevitably, partitioning algorithms do not consider all partitions and may thus only find the local optima.

k-medoids

The k-medoids method partitions a distance-space into k clusters. A medoid is an object that is selected from the dataset representing a cluster. The algorithm selects k medoids to represent the k clusters. Clusters are then created by assigning each of the remaining objects to the nearest medoid. The most common k-medoids algorithm is the Partitioning Around Medoids (*pam*) algorithm of Kaufman and Rousseeuw (1990). The k-medoids algorithm is as follows:

- **STEP 1**: Arbitrarily select k objects from the data as medoids.
- **STEP 2**: Consider swapping the pair of objects (i,h), where $i \in$ selected objects and $h \in$ non- selected objects. Denote the swap as $i \leftrightarrow h$. Let $d(x_i, x_h)$ be the distance-measure between two objects i and h.

  Now consider another non-selected object j. Calculate T_{ih}, the "total swap contribution" for $i \leftrightarrow h$, as

$$T_{ih} = \sum_j C_{jih}$$

where C_{jih} is the contribution to $i \leftrightarrow h$ from object j defined below. There are four possibilities to consider when calculating C_{jih}.

- If j currently belongs to the cluster defined by medoid i (denote cluster i), consider the distance $d(x_j, x_h)$ between object j and object h. If h is further from j than the second best medoid i' is from j then the contribution from object j to the swap is:

$$C_{jih} = d(x_j, x_i') - d(x_j, x_i)$$

The result of $i \leftrightarrow h$ would be that object j now belongs to cluster i'
Else, if h is closer to j than i' is to j, the contribution from j to the swap is:

$$C_{jih} = d(x_j, x_h) - d(x_j, x_i)$$

The result of $i \leftrightarrow h$ would be that object j now belongs to cluster h.

- If j currently belongs to cluster k, where $k \neq i$, check the distance between object j and object h.
If h is further from j than the medoid k is from j, then the contribution from object j to the swap is:

$$C_{jih} = 0$$

The result of $i \leftrightarrow h$ would be that object j still belongs to cluster k.
Else, if h is closer to j than k is to j, the contribution from j to the swap is:

$$C_{jih} = d(x_j, x_h) - d(x_j, x_k)$$

The result of $i \leftrightarrow h$ would be that object j now belongs to cluster h.

- **STEP 3**: Let $(i^*, h^*) = \arg\min_{i,h} T_{ih}$. If $T_{i^*h^*} < 0$ then swap $i^* \leftrightarrow h^*$.

Now object $h \in$ selected objects and $i \in$ non-selected objects. Go to Step 2.

- **STEP 4**: Allocate each non-selected object to the cluster defined by the nearest medoid.

The most attractive property of this method is its robustness. The use of medoids to define clusters makes this method very resistant against outliers in the data. It does not have to store a vast amount of information in addition to the original data in memory; all that is required is the label of the selected object.

pam algorithm is intended to handle outliers efficiently. Instead of cluster centers, it chooses to represent each cluster by its medoid. The computational complexity of *pam* is $O(I''k(n-k)^2)$, with I'' being the number of iterations, making it very costly for large n and k values.

A solution to this is the *clara* algorithm by Kaufman and Rousseeuw (1990). This approach works on several samples of size s, of the n tuples in the database, applying *pam* on each one of them. The output depends on the s samples.

5.4 Strength Measurement of Cluster

Apart from developing a cluster for a data set, there is a method to measure whether one is constructing a strong or reasonable clustering structure. Levine and Domany (2001) had listed various methods and indicators that come under the name "cluster validation". They also proposed a method to check which clustering method would be more reliable. Details of the various cluster validity approaches also can be found in Halkidi et al. (2001).

This chapter is not proposing a new technique to choose an appropriate number of clusters in partitions method. Instead, this chapter would apply the existing method used to find a suitable group for the data set. This is to check whether the cluster constructed for high dimension data using *iaca* as a dissimilarity matrix is useful or not. Milligan and Cooper

(1985) compared 30 measures of the strengths of clusters for determining the number of clusters. The Calinski and Harabasz index (CH index) had the best performance. However, the Silhouette index introduced by Rousseeuw (1987) is used much more often and easy to interpret. From Rousseeuw (1987), the silhouette index is defined as

$$\overline{\mathrm{s}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{s}_i$$

where

$$\mathrm{s}_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

and a_i is the average distance of the data point y_i to other points in cluster A where y_i belongs to

$$\mathrm{a}_i = \frac{1}{n_A - 1} \sum_{j \in A, j \neq i} \mathrm{d}(y_i, y_j)$$

and b_i is the average distance to points in the nearest neighbor cluster besides its own. Define $\mathrm{d}(i, C)$ =average distance of the data point y_i to all data points in Cluster C. Then

$$\mathrm{b}_i = \min_{C \neq A} d(i, C)$$

The index s_i can take values from -1 to 1. If the index shows zero value, that means the data point has equal distance to its cluster and its nearest neighbor cluster. If the index is positive, then it shows the observation is assigned to a correct cluster, whereas if the index shows negative values, it means the reverse. Overall, if the data points are correctly assigned, the average for the index should be close to one. Experience has led to the subjective interpretation of the silhouette index (SI) as listed in Table 5.1 below. This interpretation does not depend on the number of observations.

Table 5.1: Interpretation of the Silhouette Index (SI)

SI	Proposed Interpretation
0.71-1.00	A strong structure has been found.
0.51-0.70	A reasonable structure has been found.
0.26-0.50	The structure is weak and could be artificial; try additional methods.
≤ 0.25	No substantial structure has been found.

5.5 Data set

This chapter considers a simulated data sets to examine the performance of the *iaca* as a dissimilarity matrix.

First a data set containing 3 groups of 30 dimensional observations is generated following normal distribution. Six columns i.e., variables are generated from this data set with mean, μ_i and variance, Σ_1, where $i = 1, 2, \ldots, 6$. Another 24 columns are generated with the first 100 rows, i.e., observations correspond to mean, μ_{1j} and variance, Σ_2 and $j = 1, 2, \ldots, 24$. The second 100 rows are generated with mean, μ_{2j} and variance, Σ_2 whereas the last 100 rows has the mean, μ_{3j} and the variance, Σ_2. Note that there are 300 observations in the first data set.

As we are also interested in high-dimensional data, a normal data set with 100 dimensions are generated to correspond to sample size of 4500. 18 columns are generated with μ_i and variance, Σ_1, where $i = 1, 2, \ldots, 18$. Later, the remaining 82 columns are generated so that the first 1500 rows are normal data with mean, μ_{1j} and variance, Σ_2, and $j = 1, \ldots, 82$. The second 1500 rows are normal data with mean μ_{2j} and variance, Σ_2 and the last 1500 rows are normal data with mean μ_{3j} and variance, Σ_2. It is noted that these two data sets are generated following the normal distribution and only contain interval variables.

As the data set might also contain nominal or ordinal variables, this chapter will also study the influence angle performance in handling a

data set with mixed variables. Once again the data sets containing 30 and 100 variables with 300 and 4500 observations respectively are generated in turn, but this time with mixed variables.

For a data set containing 30 dimensions, let 21 of the columns contain categorical values, i.e. nominal variables and the other nine columns containing interval variables. Assign some of the observations an identical categorical value and to the rest other categorical values. As an example, 300 observations are given X for the first 100 rows, Y to the second 100 rows and Z to the last 100 rows.

The real data sets are also utilized to evaluate the Influence Angle Cluster Approach. There are three real data sets considered and one of them contains mixed variables. They are mammal milk data, mortality data and flower data. The data sets details are given in Table 5.2.

Table 5.2: The real data set used for clustering illustration

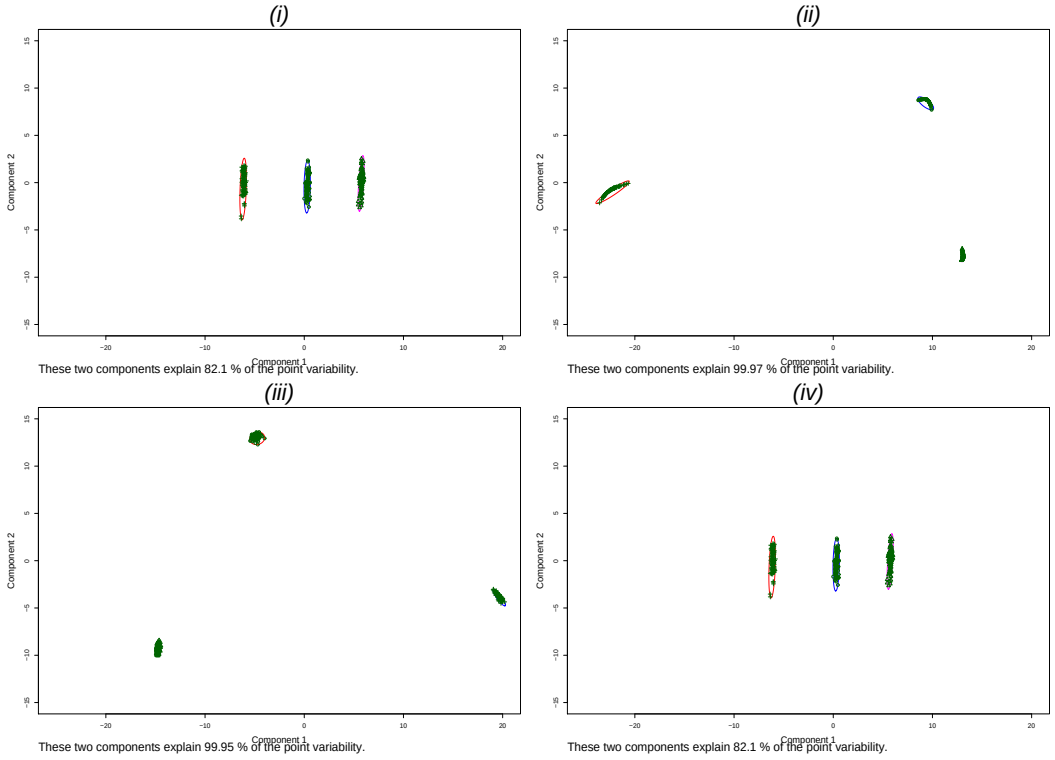
Data	Types of data	Number of variables	Sample size
Mammal milk	Interval variables	5	25
Mortality	Interval variables	4	48
Flower	Mixed variables	8	18

5.6 Clustering low dimensional data

In this section, first the low dimensional interval data set is considered. The *pam* algorithm was used to cluster the low dimensional data with Euclidean distance, while *iaca*, *daisy* and Manhattan distance in turn was used to determine the pairwise difference between objects.

This data set is generated following normal distribution containing 3 groups of 30 dimensional observations and 300 observations.

Figure 5.1 shows cluster plot developed using *pam* algorithm with Euclidean distance, and *iaca*, *daisy* and Manhattan distance in turn to calculate the dissimilarity between objects. Even though these four cluster plots can classify observations as to which group they are from, it is noted that *pam* algorithm with *iaca* and *daisy* are able to separate all three groups completely on a very large scale.



(a)

Figure 5.1: Cluster plot for $n = 300,\ p = 30$ – *pam* using (i) Euclidean distance, (ii) *iaca*, (iii) *daisy* and (iv) Manhattan distance.

Table 5.3 shows the silhouette width of the partition obtained with *pam*. The silhouette widths, $s(i)$ of all observations are visible at least above 0.95 when combination of *pam* and *iaca* is used to developed clusters. The silhouette index of *iaca* is the highest and the closest to 1 compared to the

other three dissimilarities.

Table 5.3: Silhouette width for interval variables with $n = 300, p = 30$

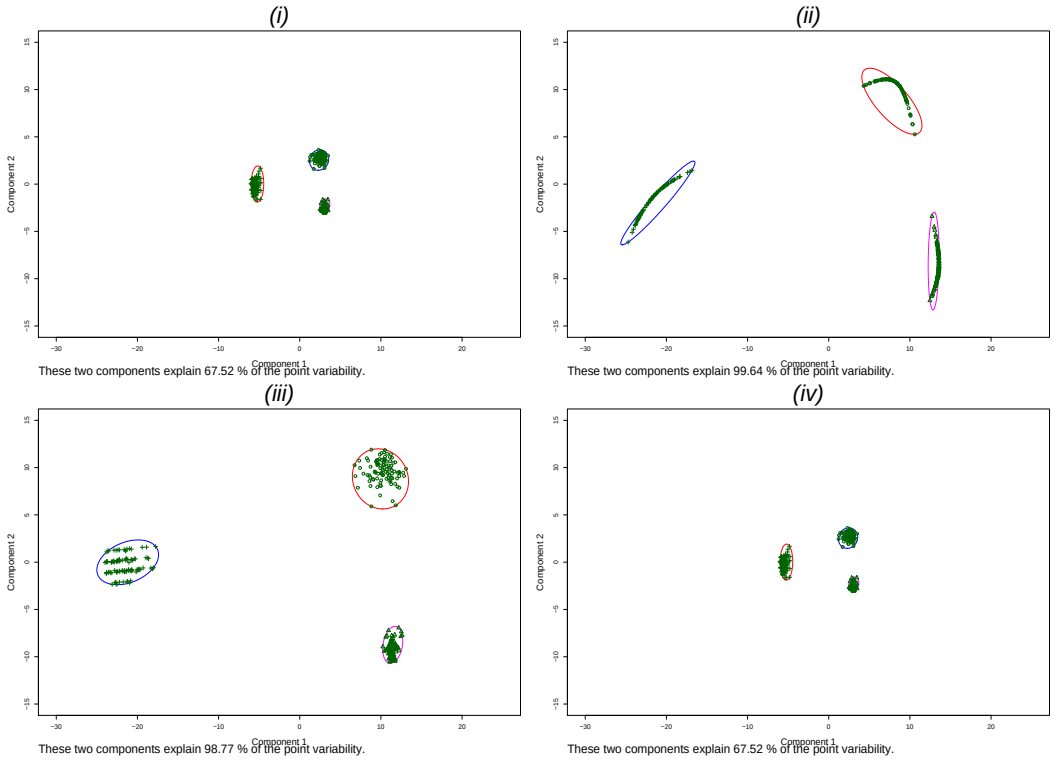
Data	Method	Dissimilarity	Silhouette width, $s(i)$	Silhouette index
$n = 300, p = 30$	*pam*	Euclidean	> 0.80	0.87
		iaca	> 0.95	0.97
		daisy	> 0.85	0.87
		Manhattan	> 0.80	0.89

The performance of *iaca* is now tested on low dimensional data with mixed variables. *pam* was used to partition the data into three clusters. The four resulting dissimilarities are given in Figure 5.2.

It can be verified that the combination of *iaca* and *pam* yields a strong clustering structure since the silhouette index is very close to 1 compared to the other three dissimilarities. See Table 5.4. Even though the silhouette widths of all objects for *iaca* are at least above 0.4, it is noted that only 5 objects have silhouette widths, $s(i)$ near to 0.4 and the silhouette widths for the remaining objects are above 0.80.

Table 5.4: Silhouette width for mixed variables with $n = 300, p = 30$

Data	Method	Dissimilarity	Silhouette width, $s(i)$	Silhouette index
$n = 300, p = 30$	*pam*	Euclidean	> 0.50	0.66
		iaca	> 0.40	0.89
		daisy	> 0.40	0.70
		Manhattan	> 0.60	0.73



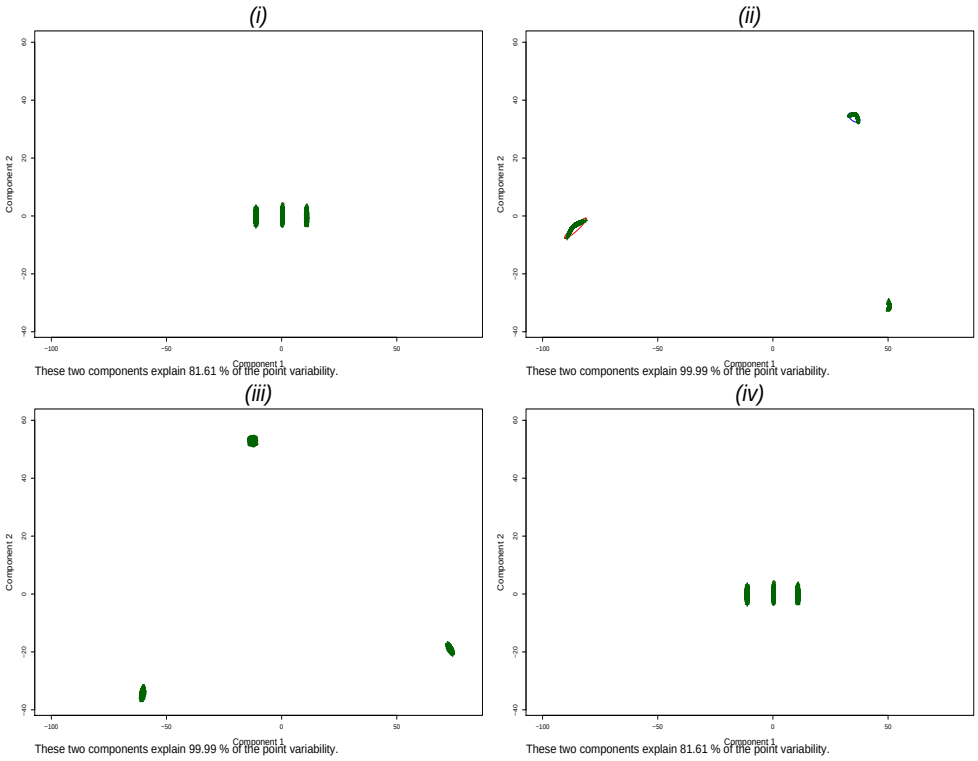
(a)

Figure 5.2: Cluster plot for $n = 300,\ p = 30$; mixed variable – *pam* using (i) Euclidean distance, (ii) *iaca*, (iii) *daisy* and (iv) Manhattan distance.

5.7 Clustering high dimensional data

Next, this chapter examines the performance of *iaca* for high-dimensional interval data set. This data set contains 4500 observations with 100 variables. *clara* algorithm was used to partition the high-dimensional data set.

Figure 5.3 shows the cluster plot and the ellipses represent the cluster boundaries as computed by *clara*. Combination of *clara* and Euclidean distance, *iaca*, *daisy* and Manhattan distance in turn managed to assign the 4500 observations into the correct cluster. Table 5.5 displays the silhouette widths, $s(i)$ of all observations and note that these 4 cluster plots have



(a) *clara*

Figure 5.3: Cluster plot for $n = 4500,\ p = 100$ – *clara* using (i) Euclidean distance, (ii) *iaca*, (iii) *daisy* and (iv) Manhattan distance

very high values of silhouette index, hence this indicates that combination of Euclidean distance, *iaca*, *daisy* and Manhattan distance with *clara* construct a very strong clustering structure.

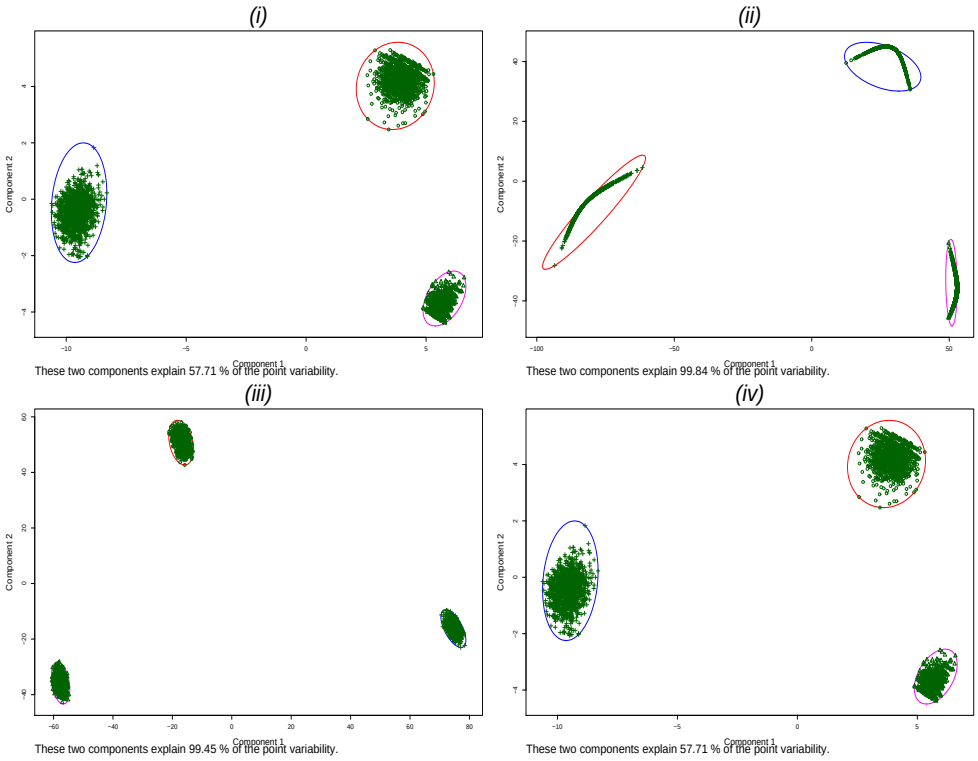
Table 5.5: Silhouette width for interval variables with $n = 4500, p = 100$

Data	Method	Dissimilarity measure	Silhouette width, $s(i)$	Silhouette index
$n = 4500, p = 100$	*clara*	Euclidean	> 0.80	0.88
		iaca	> 0.95	0.99
		daisy	> 0.95	0.99
		Manhattan	> 0.80	0.90

Next, we examined high-dimensional data set containing mixed variables. The finding shows that the combination of *clara* with *iaca* and *daisy* respectively managed to find a strong clustering structure where both dissimilarities obtained the value of silhouette index equal to 0.94. Refer to Table 5.6.

Table 5.6: Silhouette width for mixed variables with $n = 4500, p = 100$

Data	Method	Dissimilarity measure	Silhouette width, $s(i)$	Silhouette index
$n = 4500, p = 100$	*clara*	Euclidean	> 0.55	0.65
		iaca	> 0.80	0.94
		daisy	> 0.85	0.94
		Manhattan	> 0.55	0.67



(a) *clara*

Figure 5.4: Cluster plot for $n = 4500,\ p = 100$; mixed variable – *clara* using (i) Euclidean distance, (ii) *iaca*, (iii) *daisy* and (iv) Manhattan distance

5.8 Clustering real data set

5.8.1 Mammal milk data

The original Mammal Milk data set contains the ingredients of mammal's milk of 25 animals. It was taken from `http://www.uni-koeln.de/themen/Statistik/data/cluster/`. There are five variables in this data set and all variables, i.e. water, protein, fat, lactose and ash values are in percentage. Combination of *pam* with Euclidean distance, *iaca*, *daisy* and Manhattan distance are used in turn to

partition the data set into four clusters. Figures 5.5 shows the cluster plots developed by *pam*. The silhouette widths, $s(i)$ of all observations are at

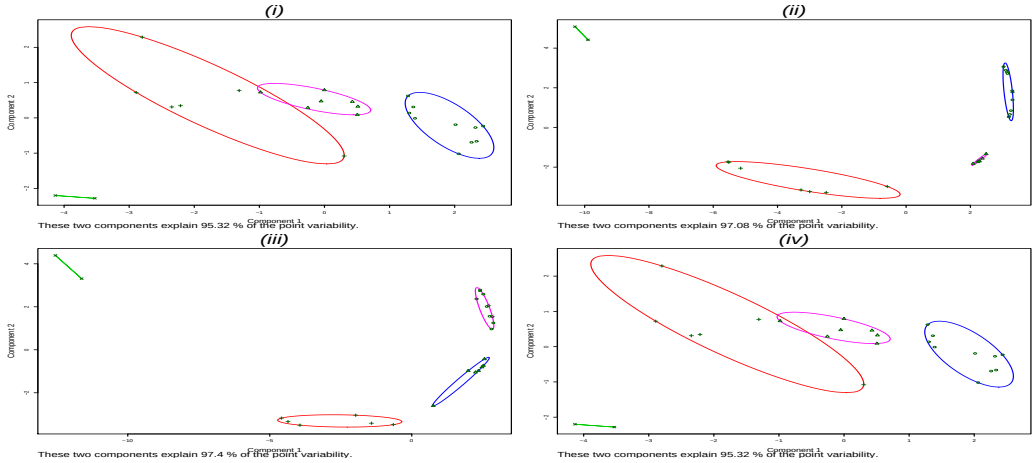


Figure 5.5: Cluster plots for mammals' milk using *pam* with (i) Euclidean distance (ii) *iaca* (iii) *daisy* (iv) Manhattan distance

least above 0.3 when the combination of *pam* and *iaca* are used to partition the data set. The value of silhouette index is 0.71, which indicates a strong clustering structure. Table 5.7 shows partition of objects i in mammal milk data when using the combination of *pam* and the other three dissimilarities only construct reasonable clustering structures since their silhouette index values are 0.60.

Table 5.7: Silhouette width for mammal milk data

Data	Method	Dissimilarity	Silhouette width, $s(i)$	Silhouette index
Mammal milk	*pam*	Euclidean	> 0.10	0.60
		iaca	> 0.30	0.71
		daisy	> 0.10	0.60
		Manhattan	> 0.10	0.60

5.8.2 Mortality data

The mortality data set is taken from Everitt et al. (2001). This data set contains 48 observations with 4 continuous variables. Cluster plots in Figure 5.6 exhibits a good clustering structure when *iaca* and *daisy* are used as dissimilarity matrix in *pam*. Table 5.8 displays the silhouette widths, $s(i)$ of

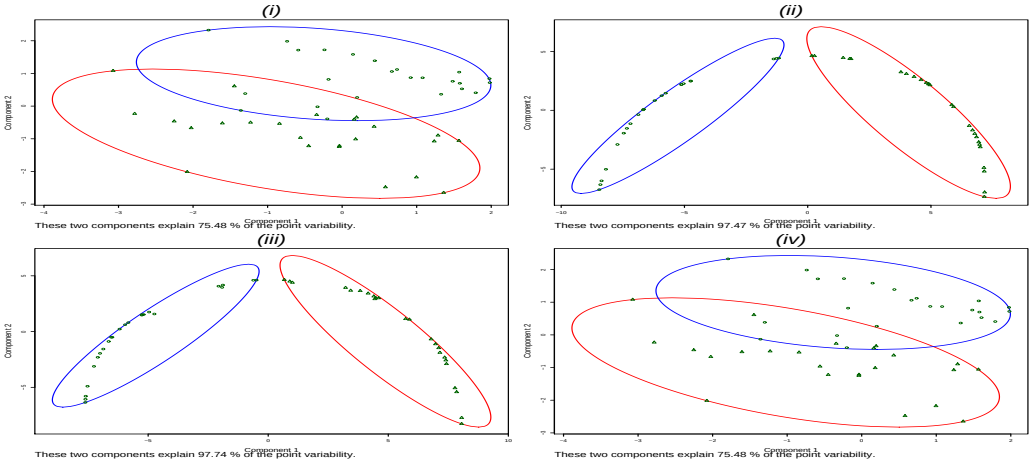


Figure 5.6: Cluster plots for mortality using *pam* (i) Euclidean distance (ii) *iaca* (iii) *daisy* (iv) Manhattan distance

all observations and the values of silhouette index yield from combination of 4 dissimilarities with *pam* in turn to partition the mortality data.

Table 5.8: Silhouette width for mortality data

Data	Method	Dissimilarity measure	Silhouette width, $s(i)$	Silhouette index
Mortality	*pam*	Euclidean	> 0.1	0.62
		iaca	> 0.1	0.64
		daisy	> 0.1	0.62
		Manhattan	> 0.1	0.58

5.8.3 Flower data set

Flower data set is taken from the R Library Cluster. There are 18 observations with eight variables in this data set. This data set contains mixed variables,. Six of them are categorical variables and the remaining are continuous variables. Cluster plots for the flower data exhibit very well the separation when combination of *pam* and *iaca* as dissimilarity matrix are used. See Figure 5.7.

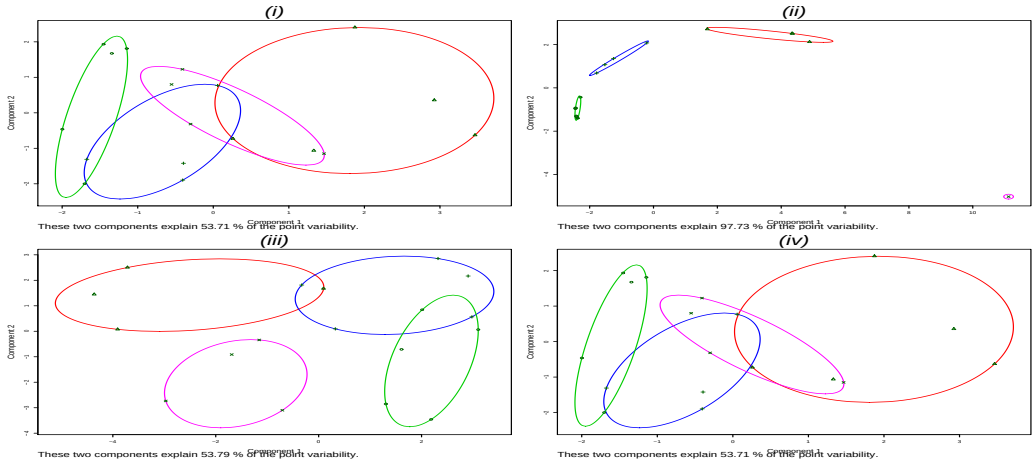


Figure 5.7: Cluster plots for flower using *pam* with (i) Euclidean distance (ii) *iaca* (iii) *daisy* (iv) Manhattan distance

The silhouette index values of 0.70 obtained shows a good clustering structure. Note that the combination of *pam* and the remaining three dissimilarities indicate weak clustering structure as the average silhouette width are below than 0.50. Refer to Table 5.9.

Table 5.9: Silhouette width for flower data

Data	Method	Dissimilarity	Silhouette width, $s(i)$	Silhouette index
Flower	*pam*	Euclidean	> 0.05	0.47
		iaca	> 0.10	0.70
		daisy	> 0.05	0.33
		Manhattan	> 0.10	0.45

5.9 Conclusion

In this chapter the author has managed to show that *iaca* successfully develops a cluster when it is used in partitioning clustering, even if the data set has mixed variables, i.e. interval and categorical variables. *iaca* is developed based on the influence eigenstructure. It can obtain clusters easily and hence, avoid the curse of dimensionality. It is also flexible to implement, and seems to work well in practice.

Chapter 6

The Buckley-James regression model for censored data

6.1 Introduction

There have been various methods created and modified to resolve the problem of censoring data sets. The term "censoring" was first used in 1949 and one can find this term mostly used in biological science areas such as survival, epidemiological and duration analysis. These analyses deal with the life time data. In addition to biological applications, censoring data sets can also be seen in educational testing and econometrics analysis (see, Greene, 2000).

This chapter will provide illustrations based on survival analysis, with §6.2 presenting the idea and concepts of survival analysis. In §6.3, the different types of censoring that can emerge in survival analysis are explained; and in §6.4, methods that can be used to solve the problem involving censoring data sets are listed. In survival analysis, the life table is the earliest and a well known method to handle the issue of censoring data sets.

Other than that, one also can choose a method which was introduced in 1958; this method is called Kaplan-Meier estimators. Details on Kaplan-

Meier estimators can be found in §6.4.1. Since researchers normally are interested in comparing Kaplan-Meier curves, various methods have been created based upon regression ideas, where these survival regression models have the ability to examine several effects of variables at a time. §6.4.2 discusses several survival regression models such as the Cox model, Miller's model, the Kaplan-Meier model and the Buckley-James model.

This chapter looks at the Buckley-James censored regression model, as this model performs well compared to other survival regression models (see Miller and Halpern, 1982; Heller and Simonoff, 1990; Heller and Simonoff, 1992; Stare et al., 2000).

Miller and Halpern (1982) made an effort to examine the potency of this model by comparing it with the Cox model, Miller's approach and Koul, Susarla and Van Ryzin's estimators by using the Stanford heart transplant data. They stated that Miller's approach and Koul, Susarla and Van Ryzin's estimators had problems with their methodology in contrast to the Cox method and the Buckley-James method. Details about the Buckley-James censored regression model are presented in §6.5.

Even though the Buckley-James approach performs better than other methods, it is still rarely practised by researchers as it is not established in many computer software programmes. The other reason is that there are few diagnostics analyses developed for the Buckley-James model. §6.6 discusses several diagnostics analyses for the Buckley-James censored regression approach. These include renovated scatterplot, plots of Hillis residuals, renovated leverage, renovated added variable plot, measures of explained variation and renovated partial residual plot.

6.2 Survival analysis

In a study, survival analysis is related to the life time distribution. In other words, it is studying the time between the subject's entry to a study and a subsequent event. Therefore, the main interest of the study is the relation

of the time to the event. Some examples of events that one can find in biological research are the time from diagnosis to death and the time it takes for a patient to respond to a new treatment. For industrial research, an example of special interest can be the life time of machine components.

However, one should note that the event may not happen for all subjects in the study. In this situation, one would have a censored data set, as some of the subjects do not have complete information. With censored data sets, one cannot use the standard analysis tools to analyse the data. Rather, the censored data set would need the use of survival analysis.

In survival analysis, there are two important functions that need to be understood. First is the survival function. Let Y be a random variable with probability density function f and Y become a survival random variable if an observed outcome, y of Y always lies in the interval $[0, \infty)$. Cumulative density function, F for Y is

$$F(y) = P(Y \leq y) = \int_0^y f(u)du. \tag{6.1}$$

The survival function is given as $S(y) = P(Y > y) = 1 - F(y)$ and expressing the survival function in terms of the probability density function, $f(u)$ can be written as below

$$S(y) = P(Y > y) = \int_y^\infty f(u)du. \tag{6.2}$$

By replacing $y = 0$ into equation 6.2, one can have $S(0) = 1$ which shows all observed subjects are alive as opposed to $S(\infty) = 0$ where all observed subjects are dead. The survival function is a decreasing monotony from $S(0) = 1$ through to $S(\infty) = 0$.

Another important function in survival analysis is the hazard function. The hazard function is the probability of failure in the time interval $[y, y +$

$\delta y]$. One can define the hazard function by the following equation

$$h(y) = \lim_{\delta y \to 0} \frac{P[y < Y < y + \delta y | Y \geq y]}{\delta y}. \tag{6.3}$$

Equation 6.3 is equal to $\frac{f(y)}{S(y)}$ for $y > 0$. $S(y)$ also can be written in the hazard function as $e^{-\int_0^y h(u)du}$. Sometimes the survival distribution is described by the cumulative hazard, $H(y) = \int_0^y h(u)d(u) = ln\, S(y)$. It should be noted that all these different functions $f(y)$, $F(y)$, $S(y)$, $h(y)$ and $H(y)$ are related and only one of the functions is needed to be able to calculate the other four.

6.3 Censoring

The word "censoring" was first recommended by Mr Kerrich to be used by Hald in 1949 and it is used mostly in studies that are involved with life time data, such as survival studies. Normally at the end of this type of study, there would be patients who survived till the final stage of the study. As having surviving patients reflects the success of the new treatment method (if this is the purpose of the study), the analyst would not want to label them as missing data. There would also be other patients with whom the analyst may lose contact. Those observations that contain incomplete information are called censored observations.

A well known censoring case is called right censored data. Right censored means the time to failure is to the right of the time line (the time line is the end time of the study), where those observations keep on running and the failure would happen after (or to the right of) the time line, refer to Figure 6.1.

The second type of censoring is interval censored data. Normally, one would use this type of censoring case when the data set is not continuously supervised. Therefore, the only information one would have is a certain interval of failure time. As an example, if one is testing a drug on eight

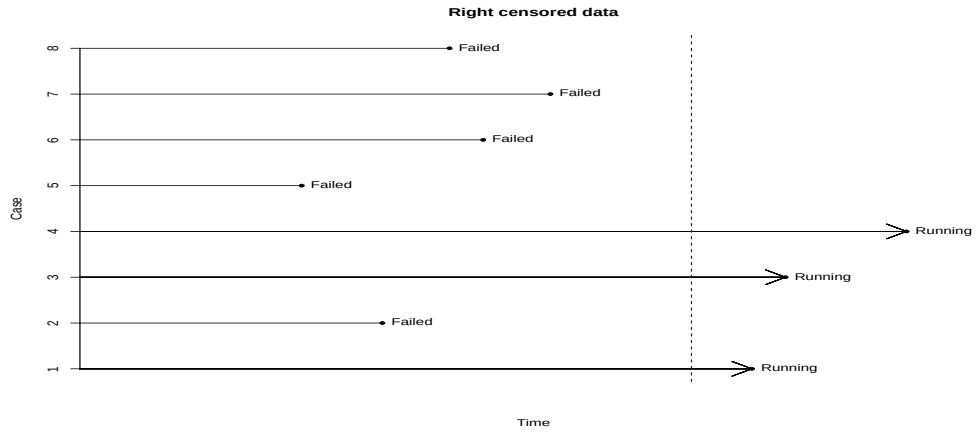


Figure 6.1: Plot of right censored data with the dashed lines representing the time line.

tissue-cells and inspecting them every 48 hours, and they are still alive at 48 hours, then the inspection is continued to 96 hours. If five of them died after 96 hours, one only knows the failure has occurred in the interval between 48 and 96 hours. Figure 6.2 reveals an example of the interval censored data and the dashed arrow means the failure time could have occurred at any time period up to the time line, however exact times for each failure are not available.

Interval censored data can also be seen as left censored data, which is another type of censoring. It is called left censored data and similar to interval censored data because the failure time occurs before the time line, and one will not know when the failure occurs. The interval starts from zero until the time line (for example, see Finkelstein, 1986).

Censoring case also can be classified into censoring type I data, censoring type II data and random censoring. Censoring type I data is also called right censored data since the times of failure to the right (i.e., larger than time line) are missing and the exact time of failure for each case occurred at any time period up to the time line is recorded. Let's say from T hours of test with n subjects, one can observe the total number of failed subjects,

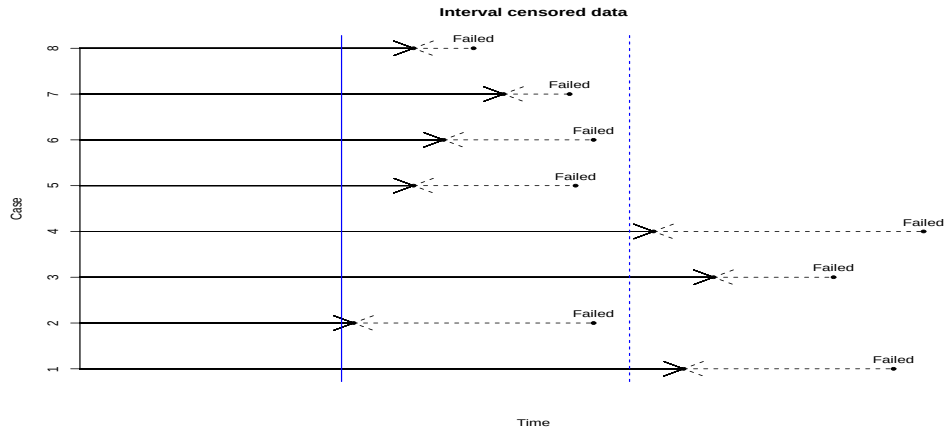


Figure 6.2: Plot of interval censored data with the vertical solid lines representing the first inspection and the vertical dashed lines representing the second inspection whereas the dashed arrows indicate the failure time could have occurred at any time period up to the time line, thus exact times for each failure are not available.

n_f before T. One will have the number of surviving cases in the T hours test as $n - n_f$.

In censoring type I data, T is fixed before the study begins and n_f is random as opposed to censoring type II data. For censoring type II data, one normally decides the total number of failures, n_f before starting the study, therefore T is unknown until the n_f failed cases emerge.

For random censoring, let each observation have a potential censoring time, t_i and potential survival time, Y_i which are independent variables. One can observe the life time of each observation, Z_i as the minimum of the censoring and survival time.

$$Z_i = \min(Y_i, t_i). \tag{6.4}$$

If t_i occurs earlier than the target event, then an ith observation will be randomly censored. The indicator variable for each observation is often

represented by δ_i where

$$\delta_i = \begin{cases} 0 & \text{(censored)} \quad \text{if} \quad Y_i > t_i, \\ 1 & \text{(uncensored)} \quad \text{if} \quad Y_i \leq t_i. \end{cases} \tag{6.5}$$

The situation exists with the random and type I censoring where there is no relationship between the censoring times and the survival variable. These two groups of censoring can be classified as non-informative censoring, which is often an assumption in survival analysis.

This chapter will consider the right censored data (or the censoring type I data). It should be noted that most techniques for right censored data can also be used with interval censored data (see, Glasson, 2007; Smith, 1996).

6.4 How do we handle censored data?

All methods in survival analysis can handle censored data. Examples of descriptive methods that one can use to estimate the distribution of survival time from a sample are life table and Kaplan-Meier survival function estimation. The life table method is the traditional approach in survival studies (see Berkson and Gage, 1950; Cutler and Ederer, 1958; Gehan, 1969), and it is mostly used in the presentation of large amounts of the right censored data.

Lawless (1982) proposed a simple algorithm of the life table analysis. According to his algorithm, the first step in the life table approach is assigning the life time into a certain number of intervals. Then for each interval, compute the conditional probability of survival. Later, one can estimate the survival function at the interval endpoints. In addition to survival function, one can obtain other information, such as the number of cases at risk, the proportion failing, the proportion surviving, the cumulative proportion, the probability density, hazard rate, median survival

time and required sample sizes.

6.4.1 Kaplan-Meier estimator

Another approach that one can use to estimate the survival function is called the Kaplan-Meier estimators or product limit estimators. It was introduced by Kaplan and Meier (1958). This approach is similar to the life table method.

The Kaplan-Meier estimators for n subjects with the ordered observations $(Z_{(1)} < Z_{(2)} < \ldots < Z_{(n)})$ corresponding to censor indicators $(\delta_{(1)}, \delta_{(2)}, \ldots, \delta_{(n)})$ can be computed as below if the tied data is absent:

$$\hat{S}(u) = \prod_{j:Z_{(j)} \leq u} \Big(\frac{n-j}{n-j+1} \Big)^{\delta_{(j)}}, \tag{6.6}$$

where j refers to number of observations surviving up to time u. The values of j are sequential integers $1, 2, \ldots, n$ if there are no censored observations.

However, if ties exist within the data, then equation 6.6 needs to be altered. There are three situations one needs to consider (Smith, 2002). First, ties among censored observations: let us say before time u, one found m individuals alive and at u there are d uncensored ties, this corresponds to a factor $(1 - \frac{d}{m})$ in the Kaplan-Meier estimators.

Second, ties among censored and uncensored observations, where in this situation, the priority should be given to uncensored observations. The last situation is where the largest ordered observation, $Z_{(n)}$ is censored; one has to change it to uncensored so as to let the $\hat{S}(u)$ reach zero for large values of u because $\hat{S}$ only jumps at uncensored observations. That is,

$$\lim_{u \to \infty} \hat{S}(u) > 0,$$

if the observations is censored.

Normally the survival studies intend to compare the differences between Kaplan-Meier curves (survival distribution of two samples). This can be done by using the logrank test. The logrank test is also called Mantel-Cox test. It is also appropriate to be used with the right censored data. This test was introduced by Mantel (1966) and the name was given to the test by Richard and Julian (1972). However, the logrank test cannot be used to examine the effects of several variables at a time. As a solution to the logrank test problem, the modeling of survival based on regression concept emerged.

In survival studies, the researcher is normally interested in investigating whether independent variables are correlated with the response variable (survival time). Nevertheless, this issue cannot be solved by using the typical multiple regression method since the response variable (survival time) follows exponential or weibull distribution (not normally distributed); the other reason is because of the censored data set.

6.4.2 Regression method for censored data

One of the survival models usually used as censored regression is called the Cox model (Cox, 1972). It is also known as the proportional hazards model and it is the regression model which is distribution-free. A good reason why the Cox model is widely utilised in survival analysis is because it has been included in most of the statistical software packages.

Distinct from the Cox model, there have been various methods created based upon standard regression ideas to resolve the problem of data sets containing censored observations, i.e. Miller's method, the Buckley-James estimators and the Koul-Susarla-Van Ryzin estimators.

Miller's method was proposed by Miller in 1976. This method is related to Kaplan and Meier's (1958) approach since it uses the weights of the Kaplan-Meier estimators to minimize the weighted sum of squares of the residuals. The estimators from this method are named Kaplan-Meier least

square estimators and are developed using iteration approaches. Later in 1979, Buckley and James presented the Buckley-James estimators, which are also known as BJ estimators. This approach is also developed using iteration methods. Details about the BJ estimators are explained in §6.5.

Subsequently Koul, Susarla and Ryzin (1981) suggested estimators for censored regression which would be developed without using the iteration method used by Miller (1976) and Buckley and James (1979). It is called the Koul-Susarla-Van Ryzin estimators. However, when Miller and Halpern (1982) compared the performance of these three methods, they found that only the Buckley-James regression method produced reliable estimators for use with censored observations.

In another study, Heller and Simonoff (1990) compared several methods of developing estimators in linear regression for a data set with censored observations. The finding was in agreement with Miller and Halpern (1982), whereby the Buckley-James method was selected over the other methods. Later, Heller and Simonoff (1992) re-examined the Buckley-James and the Cox (the proportional hazards model) methods. They determined that the choice of a method relied on the censoring proportion, the form of the failure distribution, the strength of the regression and the form of the censoring distribution.

Nevertheless, Stare et al. (2000) described three reasons to support the Buckley-James regression method over the Cox method: (i) Most researchers always failed to notice the basic assumptions of the Cox method, which is the proportionality; normally the assumption is not fulfilled (it might be due to no alternative method in the software resulting in the researcher omitting it); (ii) the Buckley-James method can provide prediction directly from estimators as opposed to the Cox method; (iii) The results of the fitting line with the Buckley-James method are easier to explain to non-statisticians.

6.5 Buckley-James censored regression

Buckley and James (1979) introduced the Buckley-James model as the regression model for censored data. The model was developed by modifying least square standard equations to make it suitable for a data set exposed to censored observations.

Before further discussions about the Buckley-James censored regression, first let us review the standard linear regression with a complete data set which can be written as

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \tag{6.7}$$

where

- Y_i is a response variable correponding to independent variable, x_i and $i = 1, 2, \ldots, n$;
- α and β are parameters to be estimated;
- ε_i is assumed to be independent and identical random variables with mean zero and variance σ^2 and the distribution $\varepsilon_i \sim F$.

By minimising the residual sum of squares, $RSS = \sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ one can obtained the β as below

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{6.8}$$

and get α by finding the difference of $\bar{Y} - \hat{\beta}\bar{x}$.

Nevertheless, if right censored observations exist in the data set, which causes the data set to be incomplete, then α and β will be biased by those censored observations. This problem can be solved by using the Buckley-James method, where one can replace censored observations with their

expected values, $E(Y_i|Y_i > t_i)$. Many studies have proved the efficiency of this method (see, Miller and Halpern, 1982; Weissfeld and Schneider, 1987; Heller and Simonoff, 1990; Hillis, 1993; Wu and Zubovic, 1995; Stare et al., 2000).

Let the response variable be subjected to right censoring, then ith subject will have a related censoring time, t_i. Now observed Z_i, δ_i and x_i for $i = 1, 2, \ldots, n$ where $Z_i = \min(Y_i, t_i)$ and δ_i is from equation 6.5.

Choose the survival time, Z_i as t_i; if the observation is censored, $\delta_i = 0$ whereas if the observation is uncensored, $\delta_i = 1$, then let the survival time, Z_i be as Y_i. Now, renovate each of the old response variable, Y_i based on their censored status, δ_i.

If $\delta_i = 1$, i.e. uncensored observation, then preserve the Y_i value therefore the new response variable, $Y_i^* = Y_i$. However if $\delta_i = 0$, compute the new response variable, Y_i^*.

From Smith (2002), one can find $E(Y_i^*(b)) = E(Y_i)$ which assures the linear regression model is not biased by $E(Y_i|Y_i > t_i)$ where b is an arbitrary slope to be estimated by Buckley-James algorithm.

$$Y_i^*(b) = \begin{cases} bx_i + \hat{E}_b(\epsilon_i(b)|\epsilon_i(b) > c_i(b)) & \text{if} \quad \delta_i = 0, \\ Y_i & \text{if} \quad \delta_i = 1, \end{cases} \tag{6.9}$$

where

$$\begin{aligned} \hat{E}_b(\epsilon_i(b)|\epsilon_i(b) > c_i(b)) &= \frac{\int_{e_i}^{\infty} \epsilon d\hat{F}_b(\epsilon)}{\int_{e_i}^{\infty} d\hat{F}_b(\epsilon)} \\ &= \sum_{k=1}^{n} q_{ik}(b) e_k(b) \end{aligned} \tag{6.10}$$

and $q_{ik}(b)$ are the weights developed from the probability mass assigned by the Kaplan-Meier estimator to $e_k(b)$.

Note that the residuals, $e_i(b)$ were obtained based on the selected survival time which corresponds to the δ_i. The different residual notations

are

$$c_i(b) = t_i - bx_i$$
$$\epsilon_i(b) = Y_i - bx_i$$
$$e_i(b) = Z_i - bx_i = \min\{\epsilon_i(b), c_i(b)\}.$$

The residuals, $e_i(b)$ play an important role in developing the weights. First, one has to sort the residuals from the smallest to the largest value as $e_1(b) < e_2(b) < \ldots < e_n(b)$. Later, one can find $\hat{E}_b(\epsilon_i(b)|\epsilon_i(b) > c_i(b))$ as (6.10) which q_{ik} can be computed as below

$$q_{ik} = \begin{cases} \dfrac{d\hat{F}(e_k(b))\delta_k(1-\delta_i)}{\hat{S}(e_i(b))} & \text{if} \quad k > i, \\ 0 & \text{if} \quad \text{otherwise.} \end{cases} \tag{6.11}$$

$d\hat{F}(e_k(b))$ is the probability mass assigned by the Kaplan-Meier estimator to e_k and $\hat{S}(e_i(b))$ is the Kaplan-Meier estimate for $e_k(b)$. After finding the renovated response variable, Y^*, which is given by 6.9, one can develop the Buckley-James estimator of β as follows

$$\sum_{i=1}^{n}(x_i - \bar{x})(Y_i^* - x_i\hat{\beta}) = 0. \tag{6.12}$$

By using the iteration, first get the initial estimate of the slope, $\hat{\beta}^{(0)}$, then, further, the Buckley-James estimator of β can be obtained as below

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i^*(\hat{\beta}_m)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \hat{\beta}_{m+1}, \tag{6.13}$$

where Y^* is given by equation 6.9 and $\hat{\beta}_m$ is the estimate of β for the mth iteration, $m = 1, 2, \ldots$.

The iteration is stopped when $|\hat{\beta}_{m+1} - \hat{\beta}_m|$ is small and reaches convergence. However, note that sometimes the convergence may not be ob-

tained, even after several iterations, especially when there are many censored observations in the data set compared to uncensored observations; but there is always at least one consistent solution (James and Smith, 1984).

Later one can estimate $\hat{\alpha}$ as follows

$$\hat{\alpha} = \frac{Y^*(\hat{\beta}) - \hat{\beta}x_i}{n}. \tag{6.14}$$

6.5.1 Multivariate censored regression

Considering that researchers normally deal with data sets that contain more than one covariate, the multivariate censored regression emerges and can be defined as below

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim F$$

where

- $\mathbf{Y}$ is a $n \times 1$ vector of response variable, which is right censored;
- $\mathbf{X}$ is a known $n \times (p+1)$ matrix as the first column of 1's to provide an intercept;
- β is a $(p+1) \times 1$ vector of parameters where it is estimated by $\mathbf{b^T} = (b_0, b_1, \ldots, b_p)$;
- ε is $n \times 1$ vector of errors and the distribution has an unknown survival function, $S = 1 - F$.

If the matrix, $\mathbf{X}$ contains only uncensored observations, then the regression parameters can be estimated as

$$\mathbf{b} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y}. \tag{6.15}$$

However, if $\mathbf{X}$ contains censored observations, then the regression parameters cannot be estimated directly as equation 6.15. Firstly, one needs to

renovate the response variable for multivariate censored regression based on the censor indicator, $\delta^{\mathbf{T}} = (\delta_1, \delta_2, \ldots, \delta_n)$ as one did for linear censored regression.

This can be done by the following equation

$$\mathbf{Y}^*(\mathbf{b}) = \mathbf{X}\mathbf{b} + \mathbf{Q}(\mathbf{b})(\mathbf{Z} - \mathbf{X}\mathbf{b}), \tag{6.16}$$

where

$$\begin{aligned} Q(\mathbf{b}) &= \text{diag}(\delta) + \{q_{ik}(\mathbf{b})\} \\ &= \begin{pmatrix} \delta_1 & q_{12}(\mathbf{b}) & q_{13}(\mathbf{b}) & \ldots & q_{1n}(\mathbf{b}) \\ 0 & \delta_2 & q_{23}(\mathbf{b}) & \ldots & q_{2n}(\mathbf{b}) \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & q_{(n-1)n}(\mathbf{b}) \\ 0 & 0 & 0 & \ldots & \delta_n \end{pmatrix} \end{aligned} \tag{6.17}$$

is the upper triangle Renovation Weight Matrix containing censored status on the main diagonal (Smith, 2002) and $\mathbf{Z}^{\mathbf{T}} = (Z_1, Z_2, \ldots, Z_n)$ are the observed responses subject to censoring indicator, δ and q_{ik} is

$$q_{ik}(\mathbf{b}) = \begin{cases} \dfrac{d\hat{F}(e_k(\mathbf{b}))\delta_k(1-\delta_i)}{\hat{S}(e_i(\mathbf{b}))} & \text{if} \quad k > i, \\ 0 & \text{if} \quad \text{otherwise}, \end{cases} \tag{6.18}$$

where $d\hat{F}(e_k(\mathbf{b}))$ is the probability mass assigned by the Kaplan-Meier estimator to e_k and $\hat{S}(e_i(\mathbf{b}))$ is the Kaplan-Meier estimate for $e_k(\mathbf{b})$. The weight matrix $\mathbf{Q}$ satisfies (Smith, 2002):

- $\mathbf{Q}^2 = \mathbf{Q}$ and $(\mathbf{I}\text{-}\mathbf{Q})^2 = \mathbf{I}\text{-}\mathbf{Q}$ (idempotence);
- $\mathbf{Q}\mathbf{1} = \mathbf{1}$ where $\mathbf{1}$ is an $n \times 1$ vector of 1's (row sums);
- $\mathbf{1}^T\mathbf{Q} = n\mathbf{v}^T$ where $\mathbf{v} = (\upsilon_1, \upsilon_2, \ldots, \upsilon_n)^T$ (column sums);

- $\mathbf{W} = \mathbf{I}$, the $n \times n$ identity matrix, in the absence of any censoring.

In multivariate censored regression, the iteration concept is still applied to develop the Buckley-James estimators:

$$b_{m+1} = (X^T X)^{-1} X^T (X b_m + Q(b_m)(Z - X b_m)). \tag{6.19}$$

Recall that $m = 1, 2, \ldots$ refers to the number of iterations as in the linear censored regression. The solution of (6.19) can be obtained as the norm of $b_{m+1} - b_m$ is small (James and Smith, 1984) and (Lin and Wei, 1992). Nevertheless if the iteration fails to converge, one can solve this problem by taking the average of all possible solutions of β (Wu and Zubovic, 1995). Note that where there is an exact solution, the Buckley-James estimators are given, as below

$$\hat{\beta} = (\mathbf{X^T Q X})^{-1} \mathbf{X^T Q Z}. \tag{6.20}$$

Since $QY^* = QZ$, equation (6.20) can be rewritten as the following equation

$$\hat{\beta} = (\mathbf{X^T Q X})^{-1} \mathbf{X^T Q Y}^*. \tag{6.21}$$

6.5.2 Properties of the Buckley-James censored regression

Many comparisons and simulation studies on the Buckley-James model have been done to evaluate its performance (see, Buckley and James, 1979; Miller and Halpern, 1982; Moon, 1989; Heller and Simonoff, 1990).

In addition, studies on the Buckley-James censored regression asymptotic properties also have been given a great deal of attention, as in James and Smith (1984), James (1986), Smith (1988) and Ritov (1990). In 1991, Lai and Ying tried to modify and stabilize the Buckley-James estimators so as to make them consistent as well as asymptotically normal under regularity condition.

Studies on the Buckley-James covariance matrix can be found in Buckley and James (1979), Smith (1986), Weissfeld and Schneider (1987), Ritov

(1990), Hillis (1993) and Hillis (1994). The variance estimator proposed by Buckley and James in 1979 was lacking in theoretical justification even though previous studies showed it performed well in most situations (see, Weissfeld and Schneider, 1987; Lin and Wei, 1992; Hillis, 1993). The covariance estimator for $\hat{\beta}$ given by Buckley and James (1979) can be written as following

$$\begin{aligned}\hat{\Sigma}_{\hat{\beta}} &= \hat{\sigma}^2_{BJ}\Big[X_u^T X_u\Big]^{-1} \\ &= \Big[\sum_{i=1}^{n}\delta_i\Big\{e_i(b) - n_u^{-1}\sum_{i=1}^{n}\delta_i e_i(b)\Big\}^2\Big/\Big(n_u - p\Big)\Big]\Big[X_u^T X_u\Big]^{-1}, \qquad (6.22)\end{aligned}$$

where n_u is the number of uncensored observations and X_u represents the design matrix corresponding to the uncensored observations. Later, Smith (1986) proposed a variance estimator based on the asymptotic variance.

Weissfeld and Schneider (1987) also proposed an alternate covariance matrix for $\hat{\beta}$ as follows

$$\begin{aligned}\hat{\Sigma}_{\hat{\beta}} &= \hat{\sigma}^2_{WS}\Big[X_u^T X_u\Big]^{-1} \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big[\delta_i\{e_i(b)\}^2 + (1-\delta_i)\sum_{k=1}^{n}q_{ik}(b)\{e_k(b)\}^2\Big]\Big[X_u^T X_u\Big]^{-1}, \qquad (6.23)\end{aligned}$$

However, the Hillis (1993) and Hillis (1994) simulation studies showed that the Smith (1986) variance estimator performed best.

Ritov (1990) suggested that variance estimation for Buckley-James model can be done by following Tsiatis's (1990) method. However, this idea was disapproved by Wei, Ying and Lin (1995) because it was not stable for uncensored data.

6.6 Renovated diagnostics for Buckley-James censored regression

There are various techniques used to examine a model so as to discover the outlying and influential observations in regression with a common data set (details can be found in Belsley, Kuh and Welsch, 1980; Cook and Weisberg, 1982; Chatterjee and Hadi, 1988).

Thus in censored regression, particularly the estimators estimated using the Buckley-James method, one can find a few diagnostic tools. For example, Smith and Zhang (1995) proposed renovated leverage value and renovated scatterplot. One can also find scatterplot proposed by Hillis (1995). In 1999, Smith and Peiris suggested using renovated added variable plot. And the latest studies used measures of explained variation (MVE) and renovated partial residual plot, proposed by Glasson (2007) and Wang, Zhang, Ahmed and Aziz (2009) respectively.

6.6.1 Renovated Scatterplots and Residual Plots

Once the solution of the Buckley-James estimator is obtained, one will also have the new response variable (Y^*), particularly for censored observations, noting that the response variable for uncensored observations would remain the same: recall equation 6.9. By using Y^*, one can now develop a scatterplot of X vs Y^*. This means the plot contains renovated points and uncensored points and it tends to display less scatter in the censored values since the censored residuals are renovated to their expected positions, which tends to be close to the final regression line.

Renovated scatterplots and associated residual plots are not completely the same as those used in simple linear regression. However, renovated scatterplots are useful in visualising the upwards movement of the censored points caused by the Buckley-James algorithm, and their effect on the final regression line (Glasson, 2007).

Next, Hillis (1995) made an effort to develop a residual plot similar to standard residual plot for standard regression. This plot was developed by using modified residuals to examine heteroscedacity and the violation of other distributional assumptions. The modified residuals are

$$e_i^* = \delta_i(Y_i - x_i^T\hat{\beta}) + (1 - \delta_i)D_i, \tag{6.24}$$

where D_i is randomly generated from the conditional distribution estimated from the fitted model (Glasson, 2007).

6.6.2 Renovated Added Variable Plots

The added variable plots are diagnostic tools that permit an evaluation of the role of individual variables within the multiple regression model. They are used to assess visually (i) whether a variable should be included or not in the model and (ii) the presence of outliers and influential cases. An added variable plot is a way to look at the marginal role of variable X_k in the model, given that other independents are already in the model.

Smith and Peiris (1999) proposed the renovated added variable plot for censored regression. Assume the censored regression model,

$$Y^* = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

then the renovated added variable plot for censored regression can be defined in terms of residuals as the plot $e^*(Y^*|X_1)$ against $e^*(X_2|X_1)$ where $e^*(Y^*|X_1)$ is the renovated residual (Y^* regress on X_1) and $e^*(X_2|X_1)$ is the renovated residual (X_2 regress on X_1).

It can be shown that the slope of the added variable plot of $e^*(Y^*|X_1)$ on $e^*(X_2|X_1)$ is equal to the estimated coefficient β_2 of X_2 in the censored regression model $Y^* = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ (see Smith and Peiris, 1999).

The stronger the linear relationship in the added variable plot, the more important the additional contribution of X_2 to the regression equation al-

ready containing other predictors. If the scatter of the points shows no marked slope, the variable is unlikely to be useful in the model.

6.6.3 Renovated Partial Residual Plot

The partial residual plot is also called the residual plus component plot. This plot examines whether the linearity assumption in a multiple regression model appears to be satisfied. The plot can therefore suggest possible transformations for linearizing the data. The indication of normality is however not present in the added variable plot because the horizontal scale in the plot is not the variable itself.

The partial residual plot is a scatter plot of $(e + \hat{\beta}_j X_j)$ versus X_j where e is the ordinary least square residual when Y is regressed on all predictor variables and $\hat{\beta}_j$ is the cefficient of X_j in this regression. As in the added variable plot, the slope of the points in this plot is $\hat{\beta}_j$, the regression coefficient of X_j.

In the case of censored regression, let us say one has the censored regression model $Y^* = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, Wang et al. (2009) defined the renovated partial residual vector for X_2 as $R^*_{X_2} = (I - H^*)Y^* + X_2\beta_2$, where H^* is the renovated hat matrix. From the plot, if the point lies very close to the straight lines, it suggests the X_2 affects the Y^* strength linearity. Wang et al. (2009) also proved that the slope of the renovated partial residual plot is equal to the β_2 in Y^*.

By using the Stanford heart transplantation model from Miller and Halpern (1982), Wang et al. (2009) illustrated the renovated partial residual plots and their properties in the censored regression model. The coefficients of the fitted regression line are given in the Table 6.1 and it presents strong evidence that the partial slope of X_i for renovate partial plots and the corresponding regression coefficient for the full model, $\hat{\beta}_i$ display almost the same value.

Table 6.1: Regression Line for Stanford Heart Transplantation

Using Variable	Estimate of Coefficient $\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
X_1 and X_2	182.750	1.136	62.197
X_1	236.974	1.501	—
X_2	228.950	—	64.550

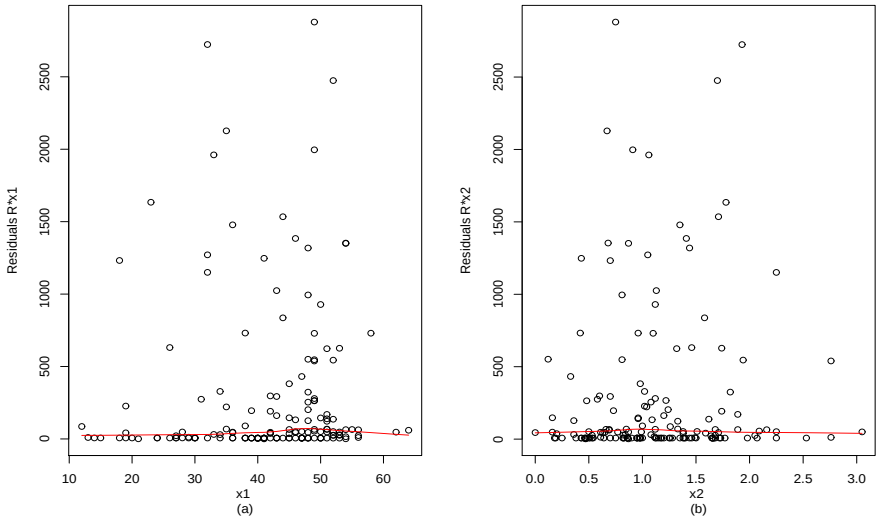


Figure 6.3: Renovated partial residuals plot using Stanford heart transplant data for (a) X_1 (b) X_2

Both plots are useful, but the partial renovated residual plot is more sensitive than the added variable plot in detecting nonlinearities in the variable being considered for introduction in the model. The added variable plot is, however easier to interpret and points out the influential observations.

6.6.4 Renovated Hat Matrix

The hat matrix is used to identify the outlying observations (Belsley et al., 1980). In 1995, Smith and Zhang proposed the renovated hat matrix, H^* for censored regression. H^* is developed from Lemma 2.1 in Chatterjee and Hadi (1988). The renovated hat matrix for censored regression is given as

$$H^* = X(X^T Q X)^{-1} X^T Q$$

Next, one may define the vector of renovate residual,

$$e^* = Y^* - \hat{Y}^* = Y^* - H^* Y^*,$$

so that $e^* = (I - H^*)Y^*$. The H^* is not symmetric, however it fulfills $(H^*)^2 = H^*$, $(I - H^*)^2 = I - H^*$, $\text{tr}(H^*) = p$ and $H^*(Y^* - X\beta) = 0$. It follows that the variance of the renovate residual estimate is $\sigma^2(e^*) = \sigma^2(I - H^*)$.

Thus, the variance of an individual renovate residual, e_i^*, is

$$\sigma^2(e_i^*) = \sigma^2(1 - h_{ii}^*),$$

where h_{ii}^* is from

$$\text{diag}(h_{11}^*, h_{22}^*, \ldots, h_{nn}^*) = \text{diag}(H^*)$$

and h_{ii}^* can be calculated without calculating the whole H^*,

$$h_{ii}^* = x_i^T (X^T Q X)^{-1} X^T q_i,$$

where q_i can be calculated as equation 6.11. h_{ii}^* measures the leverage of an observation.

For standard regression cases where all observations are uncensored, one can identify the high-leverage observation by comparing the h_{ii} value with $2p/n$ (see, Belsley et al., 1980; Myers, 1990) where the h_{ii} value is given

as below

$$h_{ii} = x_i^T (X^T X)^{-1} x_i. \tag{6.25}$$

In censored regression, it is noted the h_{ii}^* is equal to zero for $\delta_i = 0$, i.e. censored observation. In the case of $\delta_i = 1,$ i.e uncensored cases, if the $h_{ii}^* > 2(p+1)/n$, then the observation could be flagged as uncommonly large (Smith, 2004).

6.6.5 Measures of explained variation

Measures of explained variance (MEV) is a summary of the fit of a linear regression. In common regression, it can be measured using coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})}. \tag{6.26}$$

The regression line perfectly fits the data if the $R^2 = 1$. However, this measurement cannot be used in survival analysis due to the existence of the censored point. There are a few suggested MEV for survival analysis in previous studies by Harrell (1986) and Kent and O'Quigley (1988) that are based on a likelihood assumption; nevertheless, this proposed method is not suitable to be applied to the Buckley-James model.

Therefore, in 2007, Glasson proposed a few versions of MEV for the Buckley-James model; the first measurement of MEV was developed following the ideas of Smith and Zhang (1995) by using the renovated response variable, Y^* (see 6.9) where

$$R_{G1}^2 = 1 - \frac{\sum_{i=1}^n (\hat{e}_i^*)^2}{\sum_{i=1}^n (y_i^* - \bar{y^*})}. \tag{6.27}$$

However, the first measurement did not work well under high censoring rates where it produced a large value of R_{G1}^2 as the censored observations tended to give more information in the Buckley-James model (Glasson,

2007). Hence, he tried to solve this problem by suggesting MEV that computed using uncensored observations only. The second measurement is given as

$$R^2_{G2} = 1 - \frac{\sum_{i=1}^{n} \delta_i \hat{e}_i^2}{\sum_{i=1}^{n} \delta_i (y_i - (\sum_{i=1}^{n} \delta_i y_i \Big/ \sum_{i=1}^{n} \delta_i))}. \tag{6.28}$$

However, the second measurement produced negative values of R^2_{G2} when he applied it to Stanford heart transplant data. Recall that R^2 must be between 0 and 1.

Consequently, he proposed a third measurement that was based on pearson correlation coefficient between the uncensored response and the uncensored predicted response; this approach follows Hocking (2003) and can be defined as

$$r^2_{G3} = \frac{\sum_{i=1}^{n} \delta_i \{y_i - (\sum_{i=1}^{n} \delta_i y_i \Big/ \sum_{i=1}^{n} \delta_i)\}\{\hat{y}_i - (\sum_{i=1}^{n} \delta_i \hat{y}_i \Big/ \sum_{i=1}^{n} \delta_i)\}}{\{(\sum_{i=1}^{n} \delta_i) - 1\} s_1 s_2}, \tag{6.29}$$

where s_1 and s_2 correspond to standard deviation of the uncensored response and the uncensored predicted response. The third measurement yielded a value of r^2_{G3} between 0 and 1 and the value was not inflated by censored observations. Nonetheless, the diagnostic of r^2_{G3} with Stanford heart transplant data showed very small values. From these three measurements, only r^2_{G3} is reliable in practice; however, measures of explained variation in the Buckley-James model need more attention and work in the future.

6.7 Conclusion

This chapter mainly looked at the survival regression model, particularly the Buckley-James model and the corresponding diagnostics analysis. In the first instance, the idea of survival analysis and the relation to the censoring data were briefly discussed .

Note that in censoring, one can find three types of censoring, which are censoring type I, censoring type II and random censoring. Censoring data are related to the time and they cannot be analysed using the standard analysis tools. Therefore, the use of survival analysis appears to solve this problem. In survival analysis, there are two important functions: the survival function and hazard function.

There are many methods in survival analysis that can be used to analyse censoring data. Nowadays, researchers are interested in comparing the survival distribution of two samples. Even though this can be done by using the logrank test, this method cannot examine the effects of more than one variable at a time. This difficulty can easily be overcome by using the survival regression model. Examples of the survival regression model are the Cox model, Miller's model, the Buckely-James model and the Koul-Susarla-Van Ryzin model.

The Buckley-James model's performance is comparable with the Cox model and the former performs best when compared both to the Miller model and the Koul-Susarla-Van Ryzin model. Previous comparison studies prove that the Buckley-James estimator is more stable and easier to explain to non-statisticians than the Cox model. Today, researchers are interested in using the Cox model instead of the Buckley-James model. This occurred because of the lack of function of Buckley-James model in the computer software and choices of diagnostics analysis.

Currently, there are only a few diagnostics analyses for Buckley-James model that exist. Therefore, two new diagnostics analyses for the Buckley-James model are proposed in Chapter 7.

Chapter 7

New diagnostics analysis for the Buckley-James model

7.1 Introduction

In this chapter, two new diagnostics analyses of the Buckley-James model will be discussed. The first diagnostic analysis is based on Cook's idea, and the second one is using Shi's approach. It is acknowledged that Cook's statistics (Cook, 1977) are perhaps the best summary of influence due to its tendency to amplify the influence of a case.

Therefore, Cook's statistics are chosen to be modified in an attempt to produce the quickest way to detect the influential case in censored regression, particularly the Buckley-James model. In this chapter, the first proposed diagnostic is called renovated Cook's distance, RD_i^* which can be found in §7.2. This approach seems to have advantages (depending on the analyst's demands) over

(i). DFIT$_i^* = x_i^T\hat{\beta} - x_i^T\hat{\beta}_{(i)}$ as it measures the influence of case i on all n fitted values $\hat{Y}_i^*$ (not just the fitted value for case i as DFIT$_i^*$);

(ii). DBETA$_i^* = \hat{\beta} - \hat{\beta}_{(i)}$ since DBETA$_i^*$ corresponds to the number of variables, p so it is usually easier to look at a diagnostic measure such as RD_i^* since information in p can be considered simultaneously.

DFIT$_i^*$ measure effect of change in fit and DBETA$_i^*$ evaluate change in the estimated regression coefficients for censored regression if the ith row of $X_{n\times(p+1)}$ is deleted. $\hat{\beta}$ represents the coefficients estimated for censored regression of all cases and $\hat{\beta_{(i)}}$ are the coefficients estimated for censored regression when the ith row is deleted. The subscript i in parentheses is read as "with case i is removed from $X_{n\times(p+1)}$". Recall

$$\hat{\beta} = (X^T Q X)^{-1} X^T Q Y^*$$

where Y^* and Q is from equation 6.9 and 6.17 respectively.

Secondly, the local influence approach for the Buckley-James model is proposed. In censored regression, one finds that most diagnostic studies using the local influence approach have only been applied to the Cox model and the Kaplan-Meier model (see, Reid, 1981; Pettitt and Daud, 1989; Weissfeld, 1990; Escobar and Meeker, 1992; Barlow, 1997). §7.3 presents the local influence diagnostics of the Buckley-James model, which consist of

(i). variance perturbation;

(ii). response variable perturbation;

(iii). censoring status perturbation;

(iv). independent variables perturbation.

The advantage of local influence analysis for the Buckley-James model is that this approach can discover influential censored observation. Recall that the renovated leverage used by Smith and Zhang (1995) can only identify influence observations from the uncensored group.

Next, these two suggested diagnostic methods in §7.4 are illustrated using the Stanford heart transplant data set and the lung cancer data set. The latter data set is considered so as to examine whether the proposed method can be utilised by a large covariates data set. In this chapter, plotting of survival data will denote the censored observations as solid circles and uncensored observations as hollow triangles.

7.2 Renovated Cook's distance for the Buckley-James model

Cook's distance was originally proposed by Cook in 1977. A large value of Cook's distance possibly indicates that a case is influential, that when it is excluded from the regression, it will cause a substantial change in the estimated regression function (Cook and Weisberg, 1980). As opposed to DFIT_i^*, Cook's distance measures the influence of ith cases on all n fitted values $\hat{Y}_i^*$.

DFIT_i^* for the Buckley-James model measures the influence of case i on its own fitted value, $\hat{Y}_i^*$. DFIT_i^* in §7.1 is given by Smith (2002) as

$$\text{DFIT}_i^* = x_i^T\hat{\beta} - x_i^T\hat{\beta}_{(i)} = \frac{h_{ii}^*\epsilon_i^*}{(1-h_{ii}^*)}, \tag{7.1}$$

where $\epsilon_i^* = Y_i^* - x_i^T\hat{\beta}$ and $h_{ii}^* = x_i^T(X^TQX)^{-1}X^Tq_i$. Y_i^*, q_i and Q is from equation 6.9, 6.11 and 6.17 respectively.

DFIT_i^* represents the number of estimated standard deviations of $\hat{Y}_i^*$ where the fitted value $\hat{Y}_i^*$ increases or decreases with the inclusion case i in regression.

In a general version of Cook's distance for least square regression (LSR), one can have

$$D_i = \frac{(\hat{Y} - \hat{Y}_{(i)})^T(\hat{Y} - \hat{Y}_{(i)})}{p\sigma^2}, \tag{7.2}$$

where $\hat{Y}_{(i)}$ is the deleted fitted value when the ith point is deleted.

Theorem 7.2.1 *The renovated Cook's distance is given as*

$$RD_i^* = \frac{(\hat{e}_i^*)^2}{ps^2}\left\{\frac{h_{ii}^{**}}{(1-h_{ii}^*)^2}\right\}$$

where $h_{ii}^{**} = q_i^T X(X^TQX)^{-1}X^Tq_i$ *and* $h_{ii}^* = x_i^T(X^TQX)^{-1}X^Tq_i$.

Proof:
First, let the Buckley-James estimators be

$$\hat{\beta} = (X^TQX)^{-1}(X^TQY^*). \tag{7.3}$$

Therefore, the Buckley-James estimators without ith observation are given as

$$\begin{aligned}
\hat{\beta}_{(i)} &= (X_{(i)}^TQ_{(i,i)}X_{(i)})^{-1}(X_{(i)}^TQ_{(i,i)}Y_{(i)}^*) \\
&= \left[X^TQX - x_iq_i^TX\right]^{-1}\left[X^TQY^* - x_iq_i^TY^*\right] \\
&= \left[(X^TQX)^{-1}X^TQY^* + \frac{(X^TQX)^{-1}X^Tq_ix_i^T(X^TQX)^{-1}X^TQY^*}{1-x_i^T(X^TQX)^{-1}X^Tq_i}\right] - \\
&\quad \left[\left\{(X^TQX)^{-1} + \frac{(X^TQX)^{-1}X^Tq_ix_i^T(X^TQX)^{-1}}{1-x_i^T(X^TQX)^{-1}X^Tq_i}\right\}X^Tq_iy_i^*\right] \\
&= \left[\hat{\beta} + \frac{(X^TQX)^{-1}X^Tq_ix_i^T\hat{\beta}}{1-x_i^T(X^TQX)^{-1}X^Tq_i}\right] - \\
&\quad \left[\left\{1 + \frac{x_i^T(X^TQX)^{-1}X^Tq_i}{1-x_i^T(X^TQX)^{-1}X^Tq_i}\right\}(X^TQX)^{-1}X^Tq_iy_i^*\right] \\
&= \left[\hat{\beta} + \frac{(X^TQX)^{-1}X^Tq_ix_i^T\hat{\beta}}{1-x_i^T(X^TQX)^{-1}X^Tq_i}\right] - \left[\frac{(X^TQX)^{-1}X^Tq_iy_i^*}{1-x_i^T(X^TQX)^{-1}X^Tq_i}\right],
\end{aligned} \tag{7.4}$$

where

$$Q_{(i,i)} = \begin{pmatrix} \delta_1 & q_{12} & q_{13} & \cdots & q_{1(i-1)} & q_{1(i+1)} & \cdots & \cdots & q_{1n} \\ 0 & \delta_2 & q_{23} & \cdots & q_{2(i-1)} & q_{2(i+1)} & \cdots & \cdots & q_{2n} \\ \vdots & 0 & \ddots & \ddots & & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & & & & \\ 0 & \vdots & & \ddots & \delta_{(i-1)} & q_{(i-1)(i+1)} & \cdots & \cdots & q_{(i-1)n} \\ \vdots & & & & 0 & \delta_{(i+1)} & \cdots & \cdots & q_{(i+1)n} \\ & & & & \vdots & 0 & \ddots & \ddots & \vdots \\ & & & & \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & \ldots & \ldots & \ldots & 0 & \delta_n \end{pmatrix}_{(n-1)\times(n-1)} \tag{7.5}$$

is the upper triangle Renovation Weight Matrix when ith row and column are deleted from the matrix.

So $h^*_{ii} = x_i^T(X^TQX)^{-1}X^Tq_i$ becomes a renovated leverage for censored regression, thus replacing h^*_{ii} in 7.4, resulting in the Buckley-James estimators without ith observation as

$$\begin{aligned} \hat{\beta}_{(i)} &= \left\{\hat{\beta} + \frac{(X^TQX)^{-1}X^Tq_ix_i^T\hat{\beta}}{1-h^*_{ii}}\right\} - \left\{\frac{(X^TQX)^{-1}X^Tq_iy^*_i}{1-h^*_{ii}}\right\} \\ &= \hat{\beta} + \left\{(X^TQX)^{-1}X^Tq_i\right\}\left\{\frac{(x_i^T\hat{\beta}) - y^*_i}{1-h^*_{ii}}\right\} \\ &= \hat{\beta} - \left\{(X^TQX)^{-1}X^Tq_i\right\}\left\{\frac{y^*_i - \hat{y}_i}{1-h^*_{ii}}\right\} \\ &= \hat{\beta} - \left\{(X^TQX)^{-1}X^Tq_i\right\}\left\{\frac{\hat{e}^*_i}{1-h^*_{ii}}\right\}. \end{aligned} \tag{7.6}$$

By using equation 7.6, the renovated Cook's distance for censored regression can be written as

$$\begin{aligned} RD_i^* &= \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T S(\hat{\beta}_{(i)} - \hat{\beta})}{ps^2} \\ &= \frac{1}{ps^2}\left[-\frac{\left\{(X^TQX)^{-1}X^Tq_i\right\}\hat{e}_i^*}{1-h_{ii}^*}\right]^T X^TQX\left[-\frac{\left\{(X^TQX)^{-1}X^Tq_i\right\}\hat{e}_i^*}{1-h_{ii}^*}\right] \\ &= \frac{(\hat{e}_i^*)^2}{ps^2}\left\{\frac{q_i^TX(X^TQX)^{-1}X^Tq_i}{(1-h_{ii}^*)^2}\right\} \\ &= \frac{(\hat{e}_i^*)^2}{ps^2}\left\{\frac{h_{ii}^{**}}{(1-h_{ii}^*)^2}\right\}, \end{aligned} \tag{7.7}$$

where $S = X^TQX$, s^2 is estimate variance and $h_{ii}^{**} = q_i^TX(X^TQX)^{-1}X^Tq_i$ and $\hat{e}_i^* = y_i^* - \hat{y}_i$.□

Theorem 7.2.2 *The renovated leverage of an observation in censored regression, h_{ii}^*, can be presented in the following form, $q_i^TX(X^TQX)^{-1}X^Tq_i$, which is defined as h_{ii}^{**}. Therefore, $h_{ii}^{**} = h_{ii}^*$.*

Proof:
Let the renovated leverage become

$$H^* = X(X^TQX)^{-1}X^TQ$$

and

$$X = (X_1 \;\; X_2),$$

where X_1 is an $(n \times r)$ matrix of rank r and X_2 is an $n \times (k-r)$ matrix of rank $k-r$. From lemma 1 in Smith and Peiris (1999, page 1990), one can find

$$H^* = H_1^* + (I - H_1^*)(X_2MX_2^TQ)(I - H_1^*),$$

where $H_1^* = X_1(X_1^TQX_1)^{-1}X_1^TQ$ and $M = [X_2^TQ(I - H_1^*)X_2]^{-1}$.

By using Lemma 2.1 in Chatterjee and Hadi (1988), H^{**} can be developed as below

$$\begin{aligned}
H^{**} &= QX(X^TQX)^{-1}X^TQ \\
&= \begin{pmatrix} QX_1 : & QX_2 \end{pmatrix} \begin{pmatrix} X_1^TQX_1 & X_1^TQX_2 \\ X_2^TQX_1 & X_2^TQX_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1^TQ \\ X_2^TQ \end{pmatrix} \\
&= \begin{pmatrix} QX_1 : & QX_2 \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & M \end{pmatrix} \begin{pmatrix} X_1^TQ \\ X_2^TQ \end{pmatrix} \\
&= X_1(X_1^TQX_1)^{-1}X_1^TQQ + (I - H_1^*)(X_2MX_2^TQQ)(I - H_1^*) \\
&= X_1(X_1^TQX_1)^{-1}X_1^TQ^2 + (I - H_1^*)(X_2MX_2^TQ^2)(I - H_1^*),
\end{aligned}$$

where

$v_{11} = (X_1^TQX_1)^{-1} + (X_1^TQX_1)^{-1}(X_1^TQX_2)M(X_2^TQX_1)(X_1^TQX_1)^{-1}$;
$v_{12} = -(X_1^TQX_1)^{-1}(X_1^TQX_2)M$;
$v_{21} = -M(X_2^TQX_1)(X_1^TQX_1)^{-1}$.

From the properties of the weight matrix, it is known that $Q^2 = Q$, idempotence, see the proof of $Q^2 = Q$ in Smith (2004, page 167).
Hence,

$$\begin{aligned}
H^{**} &= X_1(X_1^TQX_1)^{-1}X_1^TQ^2 + (I - H_1^*)(X_2MX_2^TQ^2)(I - H_1^*) \\
&= X_1(X_1^TQX_1)^{-1}X_1^TQ + (I - H_1^*)(X_2MX_2^TQ)(I - H_1^*) \\
&= H_1^* + (I - H_1^*)(X_2MX_2^TQ)(I - H_1^*) \\
&= H^*.\square
\end{aligned}$$

Since the renovated leverage, h_{ii}^*, comprises the diagonal entries of H^*, therefore $h_{ii}^{**} = h_{ii}^*$. Based on Theorem 7.2.2 above, Theorem 7.2.3 is given as follows:

Theorem 7.2.3 *The renovated Cook's distance is given as*

$$RD_i^* = \frac{(\hat{e}_i^*)^2}{ps^2}\left\{\frac{h_{ii}^*}{(1-h_{ii}^*)^2}\right\}$$

where $h_{ii}^* = x_i^T(X^TQX)^{-1}X^Tq_i$.

The formulae shows that RD_i^* is large when either renovated residual, e_i^*, or the renovated leverage, h_{ii}^*, is large, or both. It should be noted that due to censoring estimates of the residual variance, s^2 could easily inflate the RD_i^*. This problem is solved by calculating s^2 using the variance estimator proposed by Smith in 1986. Simulation studies by Hillis (1993) and Hillis (1994) showed Smith's estimator performed best. Smith's (1986) variance estimator is given as

$$\hat{\sigma}^2_{SMITH} = \frac{n_u}{n_u-2}g^{-2}\Big[\sum_{i=1}^{n}(x_i-\bar{x})^2\tilde{\sigma}_i^2\Big], \tag{7.8}$$

where n_u is the number of uncensored observations, $\tilde{\sigma}_i^2$ and g are defined by

$$\tilde{\sigma}_i^2 = \int \epsilon^2 d\hat{F}_{\hat{\beta}}(\epsilon) - (1-\delta_i)\left[\frac{\int_{e_i}^{\infty}\epsilon^2 d\hat{F}_{\hat{\beta}}(\epsilon)}{\int_{e_i}^{\infty} d\hat{F}_{\hat{\beta}}(\epsilon)} - \left\{\frac{\int_{e_i}^{\infty}\epsilon d\hat{F}_{\hat{\beta}}(\epsilon)}{\int_{e_i}^{\infty} d\hat{F}_{\hat{\beta}}(\epsilon)}\right\}^2\right],$$

where

$$\int \epsilon^2 d\hat{F}_{\hat{\beta}}(\epsilon) = \frac{1}{n}\sum_{i=1}^{n}\left[\delta_i[e_i(b)]^2 + (1-\delta_i)\sum_{k=1}^{n}q_{ik}(b)[e_k(b)]^2\right]$$

and

$$g = \sum_{i=1}^{n}(x_i-\bar{x})^2\Big[1-(1-\delta_i)\hat{p}_i(b)\Big],$$

where

$$\hat{p}_i(b) = 1 + \hat{\lambda}(e_i)\left[e_i - \frac{\int_{e_i}^{\infty}\epsilon d\hat{F}_{\hat{\beta}}(\epsilon)}{\int_{e_i}^{\infty} d\hat{F}_{\hat{\beta}}(\epsilon)}\right].$$

$\hat{\lambda}(e_i)$ is the estimated hazard function for e_i and it is calculated using the life table method as in Lee (1980). Influence cases can easily be detected by using the index plot $\{i, RD_i^*\}$ where i is the case number, particularly for influential observations that belong to the uncensored group.

Recall that the original Cook's distance will flag observations in standard regression from normal data i.e. uncensored data that is greater than 1 or 2 as influential (Velleman and Welsch, 1981). The principle occurs as there is an argument that Cook's distance does not have F-distribution (Chatterjee and Hadi, 1988). Following this reference point, the uncensored data using the renovated Cook's distance also will be given extra attention if their RD_i^* value is greater than 1 or 2.

In censored regression, it is noted that the RD_i^* is equal to zero for observation with $\delta_i = 0,$ i.e. censored observation. This follows from h_{ii}^*, recall that $h_{ii}^* = 0$ for censored observations. Even though the circumstances agree well with Weissfeld and Schneider's (1990) analysis, censored observations have a high tendency to be less influential than uncensored observations; nevertheless, one still has to be aware of the potency of censored observations to influence the censored regression. This issue is futher discussed and a new diagnostic tool based on local influence is proposed in §7.3 to overcome this issue.

7.3 Local influence for the Buckley-James model

Another method that one can use to discover influential observations in a data set is called local influence. It was also proposed by Cook in 1986 and was based on likelihood displacement. It is an alternative method to the global influence, i.e. deletion case, which suffers from a form of the masking effect. Details regarding diagnostics based on case deletion can be found in Andrews and Pregibon (1978), Atkinson (1981) and Johnson and Geisser (1983).

Even though the local influence method has been applied mostly to

regression models, it also works well in other statistical areas. As an example, Shi (1997) studied local influence in a multivariate model. He presented the idea of combining a general influence function and generalised Cook statistic as a new concept of local influence. This concept is easier to apply without considering a likelihood assumption.

In a censored regression, most diagnostic studies based on local influence have been done for the Cox model and the Kaplan-Meier model (see, Reid, 1981; Pettitt and Daud, 1989; Weissfeld, 1990; Escobar and Meeker, 1992; Barlow, 1997).

Studies on influence observations for the Cox model using the local influence method can be found in Pettitt and Daud (1989) and Weissfeld (1990). Pettitt and Daud (1989) proposed an overall measure of influence that uses the asymptotic covariance matrix, where this measure approximates the change in likelihood displacement if the individual observation is deleted.

The local influence method proposed by Weissfeld (1990) was different from Pettitt and Daud (1989) since it was based on perturbation of the likelihood function and perturbation of covariates included in the model.

Barlow (1997) also suggested a different local influence approach from Pettitt and Daud (1989) by measuring the estimated influence of each individual on the maximum likelihood estimate of the regression parameters. Another study that used the local influence approach was done by Escobar and Meeker (1992). They used the local influence approach to detect data perturbations that have an important effect on the maximum likelihood estimates of regression model parameters based on the censored data.

This section presents the local influence diagnostics in the Buckley-James model, which consist of

(i). variance perturbation;

(ii). response variable perturbation;

(iii). censoring status perturbation;

(iv). independent variables perturbation.

To evaluate the local change of small perturbation on some issues, we first define the general influence function and generalised Cook statistics proposed by Shi (1997). The general influence function of $T \in R^{p+1}$, can be displayed as

$$GIF(T,h) = \lim_{\varepsilon \to 0} \frac{T(w_o + \epsilon h) - T(w_o)}{\epsilon}$$

where $w = w_o + \epsilon h \in R^n$ describes a perturbation with the null perturbation, w_o fulfils $T(w_o) = T$ and $h \in R^n$ refers to a unit-length vector. Next, one can specify generalised Cook statistics to measure the influence of the perturbations on T as

$$GC(T,h) = \frac{\{GIF(T,h)\}^T M \{GIF(T,h)\}}{c},$$

where M is a $p \times p$ positive-definite matrix and c is a scalar. One may find a direction of $h_{max}(T)$ to perturb a datum and maximize local change in T. The direction of $h_{max}(T)$ can be derived by maximizing the absolute value of $GC(T,h)$ with respect to h. The serious local influence appears if maximum value $GC_{max}(T) = GC(T, h_{max}(T))$.

7.3.1 Perturbing the variance for censored regression

By using the Buckley-James estimators as follows

$$b = (X^T QX)^{-1} X^T QY^* \tag{7.9}$$

perturb the variance of the error in equation 7.9, by replacing ϵ as

$$\epsilon_w \sim N(0, \sigma^2 W^{-1}).$$

Let W be diagonal matrix

$$W = \begin{pmatrix} w_1 & 0 & \ddots & 0 \\ 0 & w_2 & \ddots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix} \tag{7.10}$$

and vector $w^T = (w_1, w_2, \ldots, w_n)$ and w is given by,

$$w = w_\circ + \epsilon h,$$

where

$$w_\circ^T = (1, 1, \ldots, 1),$$

the n-vector of ones and

$$h^T = (h_1, h_2, \ldots, h_n),$$

refers to a unit-length vector.

Lemma 7.3.1 *It is noted that the general influence function of b under the perturbation is obtained as $GIF(b,h) = (X^TQX)^{-1}X^TQD(e^*)h$.*

Proof:
Hence, W in equation 7.10 can be written as

$$\begin{aligned} W &= D(w) \\ &= \begin{pmatrix} w_1 & 0 & \ddots & 0 \\ 0 & w_2 & \ddots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix} \\ &= \begin{pmatrix} 1+\epsilon h_1 & 0 & \ddots & 0 \\ 0 & 1+\epsilon h_2 & \ddots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1+\epsilon h_n \end{pmatrix} \\ &= \mathbf{I}_n + \epsilon D(h), \end{aligned} \tag{7.11}$$

where $I_n = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$ and $D(h) = \begin{pmatrix} h_1 & 0 & 0 & 0 \\ 0 & h_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_n \end{pmatrix}$.
Now equation 7.9 becomes

$$b(w) = (X^TWQX)^{-1}X^TWQY^*. \tag{7.12}$$

By replacing $W = \text{diag}(w_1, w_2, \ldots, w_n)$ in equation 7.12, $b(w)$ can be rewritten as below

$$\begin{aligned} b(w) &= (X^T\{\mathbf{I}_n + \epsilon D(h)\}QX)^{-1}X^TWQY^* \\ &= (X^TQX + \epsilon X^TD(h)QX)^{-1}X^TWQY^* \\ &= \left\{(X^TQX)^{-1} - \epsilon[(X^TQX)^{-1}X^TQD(h)X(X^TQX)^{-1}] + O(\epsilon^2)\right\} \times X^TWQY^* \\ &= [(X^TQX)^{-1} - \epsilon\left\{(X^TQX)^{-1}X^TQD(h)X(X^TQX)^{-1}\right\}] \times X^TWQY^*, \end{aligned}$$

where $X^TWQY^* = X^T\{\mathbf{I}_n + \epsilon D(h)\}QY^* = X^TQY^* + \epsilon X^TQD(h)Y^*$.
Therefore, $b(w)$ is given by

$$\begin{aligned} b(w) &= [(X^TQX)^{-1} - \epsilon\left\{(X^TQX)^{-1}X^TQD(h)X(X^TQX)^{-1}\right\}] \times X^TWQY^* \\ &= [(X^TQX)^{-1} - \epsilon\left\{(X^TQX)^{-1}X^TQD(h)X(X^TQX)^{-1}\right\}] \\ &\quad \times (X^TQY^* + \epsilon X^TQD(h)Y^*) \\ &= b + \epsilon\left\{(X^TQX)^{-1}(X^TQD(h)e^*)\right\} + O(\epsilon^2). \end{aligned} \tag{7.13}$$

From equation 7.13, the general influence function of b under the perturbation is obtained as

$$\begin{aligned} GIF(b,h) &= (X^TQX)^{-1}X^TQD(h)e^* \\ &= (X^TQX)^{-1}X^TQD(e^*)h.\square \end{aligned} \tag{7.14}$$

Theorem 7.3.2 *To assess the influence of the variance perturbations, the generalized Cook statistics are defined as*

(i). $GC_1(b,h) = h^TD(e^*)(H^*)^2 \triangle D(e^*)h/ps^2$ *and*

(ii). $GC_2(b,h) = h^TD(e^*)(H^*)^2D(e^*)h/ps^2$

Proof:
By using lemma 7.3.1, the generalised Cook statistic of b is developed. It is

scaled by $M = X^T \triangle X$ in censored regression following that

$$cov(b) = (X^T \triangle X)^{-1}\sigma^2_{BJ},$$

where $\triangle = \text{diag}(\delta_1, \delta_2, \ldots, \delta_n)$.

$$\begin{aligned} GC_1(b,h) &= \frac{GIF(b,h)X^T \triangle XGIF(b,h)}{ps^2} \\ &= \frac{h^T D(e^*)QX(X^TQX)^{-1}X^T \triangle X(X^TQX)^{-1}X^TQD(e^*)h}{ps^2} \\ &= \frac{h^T D(e^*)H^* \times \triangle \times H^*D(e^*)h}{ps^2}. \end{aligned}$$

Therefore

$$GC_1(b,h) = \frac{h^T D(e^*)(H^*)^2 \triangle D(e^*)h}{ps^2}, \tag{7.15}$$

where $H^* = X(X^TQX)^{-1}X^TQ$ is renovated leverage for censored regression.

By applying $M = X^TX$ to the scaled generalised Cook statistic, which is based on least square regression framework ' $cov(b) = (X^TX)^{-1}\sigma^2$, one can find $GC_2(b,h)$ as follows

$$\begin{aligned} GC_2(b,h) &= \frac{GIF(b,h)X^TXGIF(b,h)}{ps^2} \\ &= \frac{h^T D(e^*)QX(X^TQX)^{-1}X^TX(X^TQX)^{-1}X^TQD(e^*)h}{ps^2} \\ &= \frac{h^T D(e^*)H^* \times H^*D(e^*)h}{ps^2} \\ &= \frac{h^T D(e^*)(H^*)^2D(e^*)h}{ps^2}.\square \end{aligned} \tag{7.16}$$

The diagnostic direction h_{max} can be obtained by calculating the eigenvector correpsonding to the largest eigen value of matrices $D(e^*)(H^*)^2 \triangle D(e^*)$ and $D(e^*)(H^*)^2D(e^*)$ from equation (7.15) and (7.16) respectively.

7.3.2 Perturbing response variables for censored regression

Theorem 7.3.3 *To assess the influence of the response variables perturbations, two generalised Cook statistics can be developed by using the scale* $M = X^T \triangle X$ *and* $M = X^T X$ *based on censored regression and the least square regression framework (LSR), which are* $h^T(H^*)^2 \triangle h/ps^2$ *and* $h^T(H^*)^2 h/ps^2$ *respectively.*

Proof:
The response variable can be perturbed as follows

$$Y_w^* = Y^* + \varepsilon h,$$

where $h \in R^n$ refers to a unit-length vector. Let equation $(X^TQX)^{-1}X^TQY^*$ become

$$\begin{aligned}(X^TQX)^{-1}X^TQY_w^* &= (X^TQX)^{-1}X^TQ(Y^* + \varepsilon h)\\ &= (X^TQX)^{-1}X^TQY^* + \varepsilon(X^TQX)^{-1}X^TQh\\ &= b + \varepsilon(X^TQX)^{-1}X^TQh. \end{aligned} \tag{7.17}$$

Therefore, the general influence function of b under the perturbation can be shown as

$$GIF(b,h) = (X^TQX)^{-1}X^TQh. \tag{7.18}$$

Now two generalised Cook statistics can be developed by using the scale $M = X^T \triangle X$ and $M = X^T X$ based on censored regression and the least square regression framework (LSR), which are

$$cov(b) = \begin{cases} (X^T \triangle X)^{-1}\sigma_{BJ}^2 & \text{if} \quad \text{(censored regression)},\\ (X^TX)^{-1}\sigma^2 & \text{if} \quad \text{(LSR)}. \end{cases} \tag{7.19}$$

respectively, where $\triangle = \text{diag}(\delta_1, \delta_2, \ldots, \delta_n)$. Hence,

$$\begin{aligned} GC_1(b,h) &= \frac{GIF(b,h)(X^T \triangle X)GIF(b,h)}{ps^2} \\ &= \frac{h^T QX(X^TQX)^{-1}X^T \triangle X(X^TQX)^{-1}X^TQh}{ps^2} \\ &= \frac{h^T(H^*)^2 \triangle h}{ps^2} \end{aligned} \tag{7.20}$$

and

$$\begin{aligned} GC_2(b,h) &= \frac{GIF(b,h)(X^TX)GIF(b,h)}{ps^2} \\ &= \frac{h^T QX(X^TQX)^{-1}X^TX(X^TQX)^{-1}X^TQh}{ps^2} \\ &= \frac{h^T(H^*)^2h}{ps^2}.\square \end{aligned} \tag{7.21}$$

7.3.3 Perturbing censoring status for censored regression

Theorem 7.3.4 *To assess the influence of the censoring status perturbations, the generalized Cook statistics are defined as*

(i). $GC_1(b,h) = \dfrac{\{\mathbf{1}+\varepsilon h\}^T D(e^*)X(X^TQX)^{-1}(X^T \triangle X)(X^TQX)^{-1}X^TD(e^*)\{\mathbf{1}+\varepsilon h\}}{ps^2}$ *and*

(ii). $GC_2(b,h) = \dfrac{\{\mathbf{1}+\varepsilon h\}^T D(e^*)X(X^TQX)^{-1}(X^TX)(X^TQX)^{-1}X^TD(e^*)\{\mathbf{1}+\varepsilon h\}}{ps^2}$

Proof:
Note that the weight matrix is given as

$$\begin{aligned} Q(\mathbf{b}) &= \text{diag}(\delta) + \{\mathbf{q_{ik}}(\mathbf{b})\} \\ &= \begin{pmatrix} \delta_1 & q_{12}(\mathbf{b}) & q_{13}(\mathbf{b}) & \dots & q_{1n}(\mathbf{b}) \\ 0 & \delta_2 & q_{23}(\mathbf{b}) & \dots & q_{2n}(\mathbf{b}) \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & q_{(n-1)n}(\mathbf{b}) \\ 0 & 0 & 0 & \dots & \delta_n \end{pmatrix} \end{aligned} \tag{7.22}$$

If the censored status is perturbed as

$$Q_w(\mathbf{b}) = Q(\mathbf{b}) + W.$$

Recall

$$\begin{aligned} W &= \begin{pmatrix} w_1 & 0 & 0 & 0 \\ 0 & w_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix} \\ &= \begin{pmatrix} 1+\epsilon h_1 & 0 & 0 & 0 \\ 0 & 1+\epsilon h_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1+\epsilon h_n \end{pmatrix} \\ &= I_n + \epsilon D(h). \end{aligned}$$

Therefore

$$
\begin{aligned}
Q_w(\mathbf{b}) &= Q(\mathbf{b}) + \mathbf{I}_n + \epsilon D(h) \\
&= \text{diag}(\delta) + \{q_{ik}(\mathbf{b})\} + \text{diag}(\mathbf{1} + \varepsilon\mathbf{h}) \\
&= \begin{pmatrix} \delta_1 & q_{12}(\mathbf{b}) & q_{13}(\mathbf{b}) & \dots & q_{1n}(\mathbf{b}) \\ 0 & \delta_2 & q_{23}(\mathbf{b}) & \dots & q_{2n}(\mathbf{b}) \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & q_{(n-1)n}(\mathbf{b}) \\ 0 & 0 & 0 & \dots & \delta_n \end{pmatrix} + \begin{pmatrix} 1+\varepsilon h_1 & 0 & 0 & \dots & 0 \\ 0 & 1+\varepsilon h_2 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \dots & 1+\varepsilon h_n \end{pmatrix} \\
&= \text{diag}(\delta + \mathbf{1} + \varepsilon\mathbf{h}) + \{q_{ik}(\mathbf{b})\} \\
&= \begin{pmatrix} \delta_1 + 1 + \varepsilon h_1 & q_{12}(\mathbf{b}) & q_{13}(\mathbf{b}) & \dots & q_{1n}(\mathbf{b}) \\ 0 & \delta_2 + 1 + \varepsilon h_2 & q_{23}(\mathbf{b}) & \dots & q_{2n}(\mathbf{b}) \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & q_{(n-1)n}(\mathbf{b}) \\ 0 & 0 & 0 & \dots & \delta_n + 1 + \varepsilon h_n \end{pmatrix}. \qquad (7.23)
\end{aligned}
$$

Hence,
$(X^T Q_w X)^{-1} X^T Q_w Y^*$

$$
\begin{aligned}
&= (X^T(Q+W)X)^{-1}X^T(Q+W)Y^* \\
&= (X^TQX + X^TWX)^{-1}(X^TQY^* + X^TWY^*) \\
&= \left\{(X^TQX)^{-1}X^TQY^* - \frac{(X^TQX)^{-1}X^TWX(X^TQX)^{-1}X^TQY^*}{1 + x_i^T(X^TQX)^{-1}X^Tw_i}\right\} + \\
&\qquad \left[\left\{(X^TQX)^{-1} - \frac{(X^TQX)^{-1}X^TWX(X^TQX)^{-1}}{1 + x_i^T(X^TQX)^{-1}X^Tw_i}\right\} X^TWY^*\right].
\end{aligned}
$$

Let $W = \mathbf{I}_n + \varepsilon D(h)$, therefore

$(X^TQ_wX)^{-1}X^TQ_wY^*$

$$= \left\{ b - \frac{(X^TQX)^{-1}X^T\{\mathbf{I}_n + \varepsilon D(h)\}Xb}{1 + x_i^T(X^TQX)^{-1}X^Tw_i} \right\} + \left\{ \frac{(X^TQX)^{-1}X^T\{\mathbf{I}_n + \varepsilon D(h)\}Y^*}{1 + x_i^T(X^TQX)^{-1}X^Tw_i} \right\}$$
$$= b + (X^TQX)^{-1}X^Te^* + \varepsilon(X^TQX)^{-1}X^TD(h)e^* + O(\epsilon^2)$$
$$= b + (X^TQX)^{-1}X^T\{e^* + \varepsilon D(h)e^*\} + O(\epsilon^2).$$

Now one can describe the general influence function as the follows

$$\begin{aligned} GIF(b,h) &= (X^TQX)^{-1}X^T\{e^* + \varepsilon D(h)e^*\} \\ &= (X^TQX)^{-1}X^TD(e^*)\{\mathbf{1} + \varepsilon h\}, \end{aligned}$$

where $\mathbf{1^T} = (1, 1, \ldots, 1)$ and $h^T = (h_1, h_2, \ldots, h_n)$. Therefore generalised Cooks using $M = X^T \triangle X$ becomes
$GC_1(b,h)$

$$= \frac{GIF(b,h)(X^T \triangle X)GIF(b,h)}{ps^2}$$
$$= \frac{\{\mathbf{1} + \varepsilon h\}^TD(e^*)X(X^TQX)^{-1}(X^T \triangle X)(X^TQX)^{-1}X^TD(e^*)\{\mathbf{1} + \varepsilon h\}}{ps^2}. \tag{7.24}$$

Now, let $M = X^TX$, hence
$GC_2(b,h)$

$$= \frac{GIF(b,h)(X^T \triangle X)GIF(b,h)}{ps^2}$$
$$= \frac{\{\mathbf{1} + \varepsilon h\}^TD(e^*)X(X^TQX)^{-1}(X^TX)(X^TQX)^{-1}X^TD(e^*)\{\mathbf{1} + \varepsilon h\}}{ps^2}.\square \tag{7.25}$$

7.3.4 Perturbing independent variables for censored regression

In global influence, the *ith* case can be considered as influential on independent variables if deleting it from the data set will change the estimated regression function. This crisis can be seen in local influence by introducing small perturbations to independent variables. If one perturbs the ith column of X as

$$X_w = X + \epsilon l_i h d_i^T,$$

where

- l_i represents the scale factor, this accounts for the different measurement units associated with the columns of X. Normally l_i is the standard deviation of the ith coefficient (Weissfeld, 1990);
- $i = 1, 2, \ldots, p$ and
- d_i is a $p \times 1$ vector with one in the ith position and zeroes elsewhere.

Lemma 7.3.5 *It is noted that the general influence function of b under the perturbation is obtained as $GIF(b, h) = l_i(X^TQX)^{-1}[d_i(e^*)^T - b_iX^T]Qh$.*

Proof:
Therefore,

$$\begin{aligned}(X_w^TQX_w)^{-1} &= \left\{(X + \epsilon l_i h d_i^T)^T Q (X + \epsilon l_i h d_i^T)\right\}^{-1} \\ &= \left\{X^TQX + \epsilon l_i (X^TQhd_i^T + d_ih^TQX + d_ih^{'}hd_i^T)\right\}^{-1} \\ &= (X^TQX)^{-1} - \epsilon l_i (X^TQX)^{-1} \times \\ &\quad (X^TQhd_i^T + d_ih^TQX + d_ih^Thd_i^T)(X^TQX)^{-1} + O(\epsilon^2)\end{aligned}$$

and $X_w^TQY^* = (X + \epsilon l_i h d_i^T)^TQY^* = X^TQY^* + \epsilon l_i d_i h^TQY^*$.

Later, one can find $(X_w^T Q X_w)^{-1}(X_w^T Q Y^*)$ as

$$
\begin{aligned}
&(X_w^T Q X_w)^{-1}(X_w^T Q Y^*) \\
&\quad = (X^T Q X)^{-1} X^T Q Y^* + \epsilon l_i (X^T Q X)^{-1} \\
&\quad\quad \{d_i h^T Q Y^* - (X^T Q h d_i^T + d_i h^T Q X)(X^T Q X)^{-1}(X^T Q Y^*) \\
&\quad\quad - \epsilon l_i d_i h^T h d_i (X^T Q X)^{-1}(X^T Q Y^*)\} + O(\epsilon^2) \\
&\quad = b + \epsilon l_i (X^T Q X)^{-1}\{d_i h^T Q Y^* - (X^T Q h d_i^T + d_i h^T Q X) \\
&\quad\quad (X^T Q X)^{-1}(X^T Q Y^*)\} + O(\epsilon^2) \\
&\quad = b + \epsilon l_i (X^T Q X)^{-1}\left\{d_i h^T Q (e^*) - X^T Q h d_i^T b\right\} + O(\epsilon^2). \square
\end{aligned}
$$

Thus the general influence function of b under the perturbation can be shown as

$$GIF(b,h) = l_i (X^T Q X)^{-1}[d_i h^T Q (e^*) - X^T Q h d_i^T b].$$

One can replace the *ith* element of b, therefore $d_i^T b = b_i$ and now one has

$$GIF(b,h) = l_i (X^T Q X)^{-1}[d_i (e^*)^T - b_i X^T] Q h. \tag{7.26}$$

Theorem 7.3.6 *To assess the influence of the independent variables perturbations, the generalized Cook statistics are defined as*

(i). $GC_1(b,h) = \dfrac{l_i^2 h^T H^* \triangle \left\{e^* d_i^T - b_i X\right\}(X^T Q X)^{-1}\left\{d_i (e^*)^T - b_i X^T\right\} Q h}{p s^2}$

and

(ii). $GC_2(b,h) = \dfrac{l_i^2 h^T H^* \left\{e^* d_i^T - b_i X\right\}(X^T Q X)^{-1}\left\{d_i (e^*)^T - b_i X^T\right\} Q h}{p s^2}$

Proof:
By using lemma 7.3.5, two generalised Cook statistics for b are constructed as:

$GC_1(b,h)$

$$= \frac{GIF(b,h)(X^T \triangle X)GIF(b,h)}{ps^2}$$
$$= \frac{l_i^2 h^T Q \left\{e^* d_i^T - b_i X\right\} (X^T QX)^{-1}(X^T \triangle X)(X^T QX)^{-1} \left\{d_i (e^*)^T - b_i X^T\right\} Qh}{ps^2}$$
$$= \frac{l_i^2 h^T H^* \triangle \left\{e^* d_i^T - b_i X\right\} (X^T QX)^{-1} \left\{d_i (e^*)^T - b_i X^T\right\} Qh}{ps^2}, \qquad (7.27)$$

whereas
$GC_2(b,h)$

$$= \frac{GIF(b,h)(X^T X)GIF(b,h)}{ps^2}$$
$$= \frac{l_i^2 h^T Q \left\{e^* d_i^T - b_i X\right\} (X^T QX)^{-1}(X^T X)(X^T QX)^{-1} \left\{d_i (e^*)^T - b_i X^T\right\} Qh}{ps^2}$$
$$= \frac{l_i^2 h^T H^* \left\{e^* d_i^T - b_i X\right\} (X^T QX)^{-1} \left\{d_i (e^*)^T - b_i X^T\right\} Qh}{ps^2}.\square \qquad (7.28)$$

Equations (7.27) and (7.28) were developed using similar scales as §7.3.3 and $\triangle = \text{diag}(\delta_1, \delta_2, \ldots, \delta_n)$. One can obtain the diagnostic direction h_{max} by computing the eigenvector corresponding to the largest eigenvalue of the following matrice

$$H^* \triangle \left\{e^* d_i^T - b_i X\right\} (X^T QX)^{-1} \left\{d_i (e^*)^T - b_i X^T\right\} Q,$$

or

$$H^* \left\{e^* d_i^T - b_i X\right\} (X^T QX)^{-1} \left\{d_i (e^*)^T - b_i X^T\right\} Q$$

from equation (7.27) and (7.28) respectively.

7.4 Analysis

The Stanford heart transplant data set will be analysed to illustrate the proposed diagnostics analysis of the Buckley-James model. This is a standard data set for censored regression. It is taken from a Stanford heart transplant programme which ran from October 1967 until April 1974. It has had a number of versions since then. In this chapter, three different versions of the Stanford heart transplant data set, corresponding to sample sizes of 69, 152 and 184, are used. Details about this data set are explained in Crowley and Hu (1977).

The second data set is lung cancer data, which are taken from Lawless (1982). This data set is considered in order to examine whether the diagnostics analysis suggested in this chapter is able to handle large covariates, due to the concern that it is more difficult to identify influential and peculiar observations in high dimensions.

7.4.1 Illustration of renovated Cook's distance

To illustrate the renovated Cook's distance in §7.2, the Stanford heart transplant data, with a sample size of 69 patients is considered. The involved variables are:

(i). date at acceptance into the programme, t_1;

(ii). date last seen, t_2;

(iii). survival time in days (y), where $y = t_2 - t_1$;

(iv). date of transplantation, t_3 where $t_1 \leq t_3 \leq t_2$;

(v). date of birth, t_4;

(vi). transplant status, s_1

$$s_1 = \begin{cases} 0 & \text{if} \quad \text{non transplant}, \\ 1 & \text{if} \quad \text{transplant}; \end{cases}$$

(vii). the censored status (δ_i)

$$\delta_i = \begin{cases} 0 & \text{if} \quad \text{censored, i.e patient alive until 1st April 1974}, \\ 1 & \text{if} \quad \text{uncensored, i.e patient deceased}; \end{cases}$$

(viii). age last seen in days (x), where $x = t_2 - t_4$.

In this section, the data are taken from R library. The data on patients who were admitted to the programme but did not receive the transplant $(s_1 = 0)$ have been omitted. That is why, for the illustration of renovated Cook's distance, only 69 patients were used and of these 69 patients, 45 were deceased, i.e. were uncensored and 24 were alive, i.e. were censored.

The explanatory variables are censored status and age in years. Since the data for the age are given in days, it is divided by 365. A patient who died on the same day as his/her transplant is given a survival time of one day. The response variable is survival time, and it is transformed to log base 10, as the linear model is often appropriate when the response variable is measured on the logarithm scale (Buckley and James, 1979).

Table 7.1 provides detailed information about each observation for residual $(\hat{e}_i^*)$, censored status (δ_i), leverage (h_{ii}), renovated leverage (h_{ii}^*) or (h_{ii}^{**}) and renovated Cook's distance (RD_i^*). One can clearly see that h_{ii}^* is similar to h_{ii}^{**} from Table 7.1. These findings agree well with theorem 7.2.2 in §7.2.

Table 7.1: Detailed information of the Stanford heart transplant data based on age, $\hat{e}_i^*, \delta_i, h_{ii}, h_{ii}^*, h_{ii}^{**}, RD_i^*$

Cases	Age	$\hat{e}_i^*$	δ_i	h_{ii}	$h_{ii}^* = h_{ii}^{**}$	RD_i^*
1	35.1	-2.620	0	0.027	0.000	0.000
2	41.5	-2.440	1	0.021	0.030	1.881
3	54.1	-2.086	1	0.019	0.029	1.334
4	40.3	-1.996	1	0.017	0.037	1.565
5	29.2	-1.706	1	0.015	0.155	6.239
6	28.6	-1.688	0	0.028	0.000	0.000
7	40.3	-1.327	1	0.026	0.038	0.718
8	55.3	-1.052	1	0.017	0.035	0.417
9	36.2	-0.945	1	0.036	0.071	0.722
10	54.3	-0.904	1	0.030	0.030	0.257
11	23.6	-0.901	0	0.016	0.000	0.000
12	45.0	-0.864	0	0.016	0.000	0.000
13	42.8	-0.812	1	0.043	0.027	0.184
14	42.5	-0.749	1	0.021	0.028	0.165
15	52.1	0.556	1	0.045	0.021	0.066
16	53.0	-0.719	1	0.027	0.024	0.128
17	19.6	-0.697	1	0.015	0.347	3.903
18	56.9	-0.645	1	0.078	0.045	0.204
19	26.7	-0.083	0	0.016	0.000	0.000
20	53.8	-0.632	1	0.020	0.027	0.112
21	46.3	-0.606	1	0.016	0.017	0.064
22	47.1	-0.575	1	0.059	0.016	0.054
23	45.3	0.769	1	0.018	0.019	0.117
24	49.0	-0.497	1	0.018	0.015	0.039
25	50.6	-0.476	1	0.016	0.017	0.040
26	53.3	-0.427	1	0.015	0.025	0.047
27	52.5	-0.423	1	0.033	0.022	0.041
28	49.1	-0.413	1	0.016	0.015	0.027
29	51.3	-0.345	1	0.015	0.019	0.023
30	51.1	-0.337	1	0.031	0.018	0.021
31	54.6	-0.265	1	0.014	0.031	0.023
32	56.4	-0.222	1	0.016	0.041	0.022
33	61.5	-0.206	1	0.015	0.084	0.042
34	43.9	-0.166	1	0.022	0.024	0.007

table 7.1 continued

Cases	Age	$\hat{e}_i^*$	δ_i	h_{ii}	$h_{ii}^* = h_{ii}^{**}$	RD_i^*
35	48.0	-0.153	1	0.024	0.015	0.004
36	47.4	-0.107	1	0.015	0.016	0.002
37	26.7	-0.633	0	0.019	0.000	0.000
38	51.8	-0.017	1	0.016	0.020	0.000
39	64.5	-0.015	1	0.048	0.116	0.000
40	42.7	0.067	1	0.016	0.031	0.001
41	47.8	0.112	0	0.020	0.000	0.000
42	48.8	0.169	1	0.145	0.017	0.005
43	32.7	0.224	0	0.015	0.000	0.000
44	49.5	0.232	1	0.015	0.017	0.010
45	48.7	0.247	0	0.023	0.000	0.000
46	48.0	0.251	1	0.015	0.019	0.012
47	46.5	0.360	0	0.084	0.000	0.000
48	49.0	0.361	0	0.034	0.000	0.000
49	38.8	0.426	0	0.067	0.000	0.000
50	54.4	0.453	0	0.021	0.000	0.000
51	36.7	0.469	0	0.016	0.000	0.000
52	41.4	0.481	0	0.026	0.000	0.000
53	47.4	0.496	0	0.015	0.000	0.000
54	48.8	0.507	1	0.023	0.027	0.071
55	52.9	0.523	0	0.024	0.000	0.000
56	52.1	-0.727	0	0.017	0.000	0.000
57	48.0	0.562	0	0.016	0.000	0.000
58	33.2	0.576	0	0.015	0.000	0.000
59	44.9	0.578	1	0.027	0.048	0.175
60	50.9	0.620	1	0.019	0.035	0.144
61	43.4	0.624	1	0.021	0.056	0.241
62	45.9	0.637	1	0.015	0.044	0.193
63	40.6	0.725	0	0.015	0.000	0.000
64	48.6	0.757	1	0.015	0.043	0.269
65	45.3	-0.520	0	0.021	0.000	0.000
66	48.5	0.893	0	0.084	0.000	0.000
67	58.4	0.899	1	0.109	0.072	0.671
68	48.9	0.955	0	0.071	0.000	0.000
69	54.0	1.042	1	0.037	0.117	1.607

First, the result of the renovated leverage observation is examined before proceeding to the influence observation in censored data using Cook's, renovated method ' RD_i^*.

The values of leverage, h_{ii} and renovated leverage, h_{ii}^* obtained in this research are similar to those shown by Smith and Zhang (1995). From Table 7.1, one can find that the highest renovated leverage value is from case 17 (patient aged 19.6 years) followed by case 5 (patient aged 29.2 years). Other patients with high renovated leverage values are case 69 (patient aged 54.0 years) and case 39 (patient aged 64.5 years). The high renovated leverage values indicate those cases have the potential to affect the parameter estimates and one needs to be aware of them.

The plot of renovated leverage in Figure 7.1 clearly represents the youngest patient (age 19.6 years), (case 7), and the patient aged 29.2 years (case 5) as the two cases with the largest h_{ii}^*.

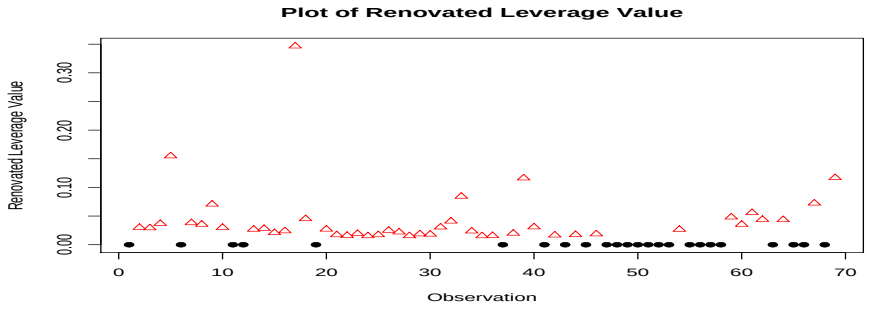


Figure 7.1: Renovated leverage plot for Stanford heart transplant data $(n = 69)$.

The result of h_{ii}^* value is zero for all patients corresponding to censored data $(\delta_i = 0)$ (refer to Table 7.1). The findings agree well with Weissfeld and Schneider's (1990) analysis, as censored observations have a high tendency to be less influential than uncensored observations.

Next, the RD_i^* value for each case in Table 7.1 is scrutinised to measure how much the cases affect the parameter estimates. If the h_{ii}^* is large,

this indicates the existence of leverage observation, which can cause the renovated residual, $\hat{e}_i^*$ of corresponding observation to be small. This situation shows the fitted value $(\hat{y}_i)$ relationship will likely to be close to the corresponding y_i^* since h_{ii}^* is close to 1. If y_i^* is such as to pull the fitted relationship from where it would be placed, the potential for influence is clear and will become real.

One can find that case 17, who is the youngest uncensored patient, with an age of 19.6 years, does not give the largest value of RD_i^* even though this observation shows the highest value of h_{ii}^*. Case 5, which is the uncensored patient with an age of 29.2 years, gives the highest value of RD_i^*. This patient has a higher residual value than the youngest patient. Refer to Figure 7.2, which is the plot of the renovated Cook's distance and one can see similar cases showing the two largest values of RD_i^*, with the patient aged 29.2 years (case 5) leading.

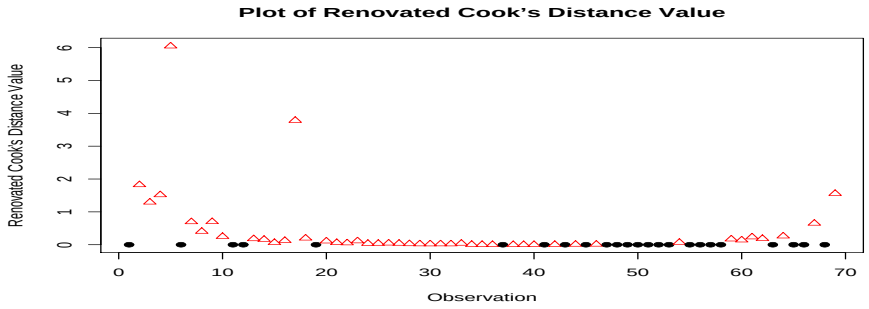


Figure 7.2: Renovated Cook's distance plot for Stanford heart transplant data $(n = 69)$ with Smith (1986)'s variance estimator.

The result of modified Cook's statistics, without doubt, clearly shows influence cases for the censored regression. However, note that the censored points cannot be influential cases as the points have no renovated leverage $(h_{ii}^* = 0)$; it follows that RD_i^* is also equal to zero. This issue needs further investigation because of the possibility of censored points becoming influential cases in censored regression.

Therefore, this chapter proposes the solution for this issue in §7.3, and the illustrations of the diagnostics analysis can be found in §7.4.2. Next, look at Table 7.2 which displays the Buckley-James model of cases, without cases 5 and 17. Table 7.2 shows the estimator for age only decreases

Table 7.2: Buckley-James model for Stanford heart transplant data, $n = 69$

Case deleted	Age	δ_i	β_1	SE	p-value	$\hat{\beta} - \hat{\beta}_{(i)}$
None			-0.028	0.015	0.060	
17	19.6	1	-0.035	0.019	0.060	0.007
5	29.2	1	-0.038	0.016	0.018	0.010

by 0.007 and 0.010 when cases 17 and 5 are excluded from the data set one at a time. Excluding these two cases from the data set do not significantly affect to the age estimator values. Nevertheless, when the p-value is scrutinised, one can find deleting case 5 causes the age estimator, β_1 to be significant at $\alpha = 5\%$.

7.4.2 Illustration of local influence

Two data sets have been considered for the illustration of local influence in the Buckley-James model. The first data set is the Stanford heart transplant data, which was taken from Miller and Halpern (1982) with sample size, $(n = 152)$. The next data set is lung cancer data, which was obtained from Lawless (1982).

Stanford heart transplant data

This data set contains 184 patients with variables such as survival time(days), censored status, age at time of first transplant (in years) and T5 mismatch score. The mismatch score refers to the continuous score derived from antibody responses of pregnant women by Charles Bieber of Stanford University (Crowley and Hu, 1977).

In this section, only 152 patients are considered, corresponding to a survival time equal to at least 10 days and with complete records. From 152 patients, 55 were deceased, i.e. were uncensored and 97 were alive, i.e. were censored. The Buckley-James model for this data set was developed as

$$Y = \beta_0 + \beta_1 AGE + \beta_2 AGE^2 + \beta_3 T5.$$

First, consider the variance perturbation. The index plot of $|h_{max}(b)|$ in Figure 7.3 shows patients aged below 20 years as the most influential cases. This finding agrees well with Reid and Crepeau (1985), and Pettitt and Daud (1989) where patients aged 13, 15 and 12 years in order have the greatest influence on variance. Note that the patient aged 15 years old is a censored observation.

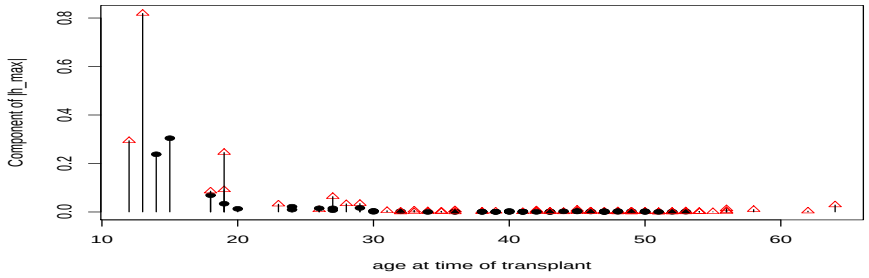


Figure 7.3: Index plots of $|h_{max}|$ for perturbing variance for Stanford heart transplant data $(n = 152)$.

Second, consider the perturbation of response variable and individual independent variables. It is obvious that the most influential patients are aged below 20 years and two patients aged above 60 years. Removal of the patients aged 12 and 13 decreases $\hat{\beta}_1$ by 0.010 and 0.030 respectively, while removal of the patient aged 15 increases $\hat{\beta}_1$ by 0.015. There is no impact on the estimator values in the Buckley-James model when deleting those observations (one at a time) since the maximum eigenvalues for the perturbation of the variance, response variable, x_1 and x_2 are small at 0.142,

0.021, -0.002 and 1.000 respectively.

However, when the p-value is scrutinised, one can find the p-value for age is roughly five times larger when deleting case 1, and triple when deleting case 4, whereas deleting case 2 has a large effect on the p-value of age^2 where the value becomes fourteen times larger.

When the observations flagged by the diagnostics based on perturbation of the censoring vector are considered, patients aged 29, 33 and 36 are flagged as influential. These three cases account for almost half of the variability of the elements of h_{max} since their sum of squares of elements of h_{max} is 0.424.

Most patients aged below 20 years are not influential to case censoring perturbations because they have small residuals. It is noted most cases with large residuals only exhibit as influential cases if their covariate variable (age) are large while cases with similar ages are less. Thus, the patient aged 13 does not appear as an influential observation even though their data have a large residual, see Figure 7.4.

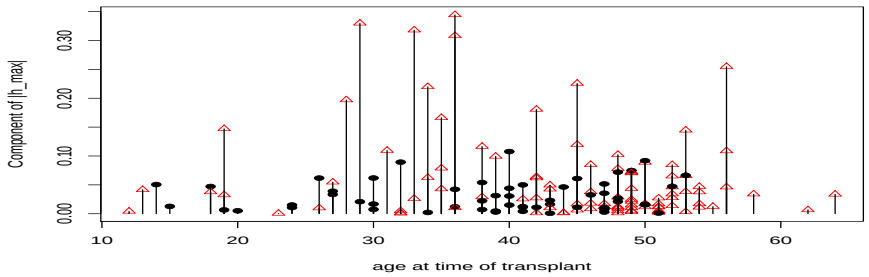


Figure 7.4: Index plots of $|h_{max}|$ for perturbing censoring status for Stanford heart transplant data

Of interest here is the plot of elements $|h_{max}(b)|$ against observation based on perturbation response variable, x_1, x_2 and censoring status given in Figure 7.5, 7.6 and 7.7 respectively. No attention is given to x_3 since this variable is not strongly associated with survival time (refer to the p-value

in Table 7.3).

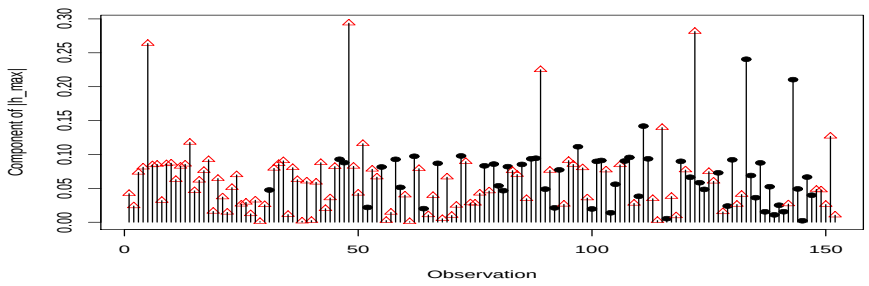


Figure 7.5: Index plots of $|h_{max}|$ for perturbing response variable for Stanford heart transplant data

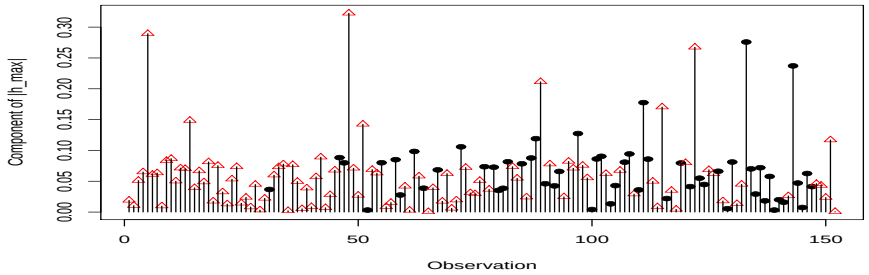


Figure 7.6: Index plots of $|h_{max}|$ for perturbing x_1 for Stanford heart transplant data $(n = 152)$.

Lung cancer data

In this data set, there are 40 observations with covariates such as survival times, censored status, performance status (x_1), age (x_2), months from diagnosis (x_3), tumor type $(x_4 - x_6)$ and treatment/standards (x_7). Three patients were deceased, i.e. were uncensored and 37 patients were alive, i.e. were censored.

Table 7.3: Buckley-James model for Stanford heart transplant data $n = 152$

Case deleted	age	δ_i	$(\beta_1, \beta_2, \beta_3)$	SE	p-value	$\hat{\beta} - \hat{\beta}_{(i)}$
None			(0.105,-0.002,-0.032)	(0.038,0.000,0.117)	(0.006,0.001,0.784)	
2	13	1	(0.075,-0.001,-0.015)	(0.042,0.001,0.116)	(0.007,0.014,0.898)	(0.030,-0.001,-0.017)
4	15	0	(0.120,-0.002,-0.035)	(0.038,0.000,0.117)	(0.002,0.000,0.767)	(-0.015,0.000,0.003)
1	12	1	(0.095,-0.002,-0.030)	(0.043,0.001,0.118)	(0.028,0.005,0.801)	(0.010,0.000,-0.002)
151	62	1	(0.107,-0.002,-0.030)	(0.040,0.001,0.118)	(0.007,0.001,0.800)	(-0.002,0.000,-0.002)
152	64	1	(0.115,-0.002,-0.023)	(0.041,0.001,0.120)	(0.005,0.001,0.848)	(-0.010,0.000,-0.009)

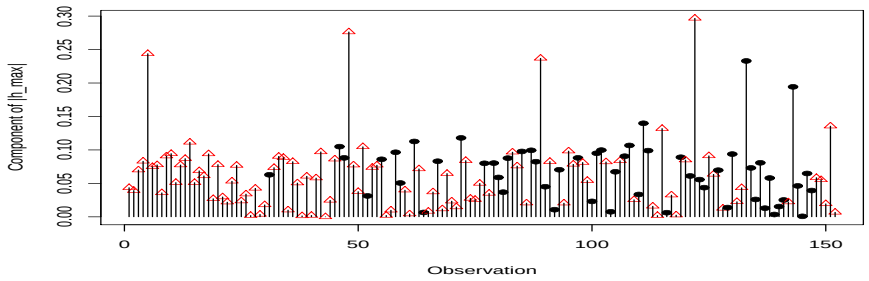


Figure 7.7: Index plots of $|h_{max}|$ for perturbing x_2 for Stanford heart transplant data $(n = 152)$.

Diagnostics based on the perturbation of variance show the patient from case 37 as the most influential case (see Figure 7.8). Weissfeld in 1990 also mentioned case 37 as one of the influential observation based on their covariates perturbation. Deletion of this observation results in little change in the all parameter estimates except for x_3. From Table 7.4, one can find that β_3 decreases almost 0.011 when case 37 is deleted from the data set. This is followed by the p-value of x_3 where it is also reduced from 0.973 to 0.166.

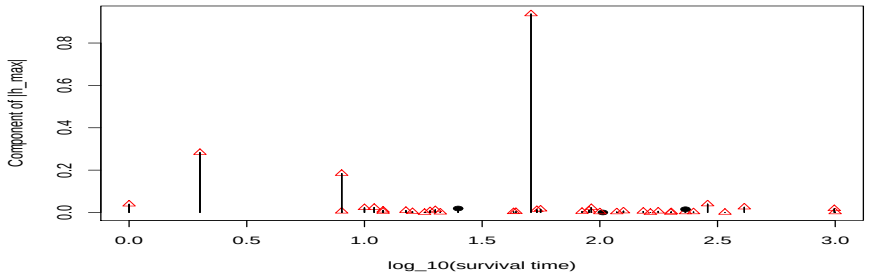


Figure 7.8: Index plots of $|h_{max}|$ for perturbing variance for Lung cancer data

Next, the perturbation of response and independent variables are examined, and one can find case 24 gives the largest value of $|h_{max}|$ for re-

Table 7.4: Buckley-James model for Lung cancer data

		Deleted Case			
		None	9	24	37
Variable	x_1	0.026	0.025	0.027	0.026
	x_2	0.007	0.005	0.008	0.006
	x_3	-0.0002	-0.001	0.0002	-0.011
	x_4	-0.106	-0.063	-0.087	-0.063
	x_5	-0.150	-0.153	-0.128	-0.190
	x_6	-0.346	-0.342	-0.321	-0.349
	x_7	-0.090	-0.105	-0.078	-0.047
SE	x_1	0.005	0.005	0.005	0.005
	x_2	0.010	0.010	0.010	0.009
	x_3	0.006	0.006	0.006	0.008
	x_4	0.220	0.232	0.228	0.212
	x_5	0.241	0.245	0.252	0.231
	x_6	0.288	0.293	0.301	0.275
	x_7	0.177	0.182	0.183	0.171
p-value	x_1	<0.001	<0.001	<0.001	<0.001
	x_2	0.441	0.622	0.419	0.461
	x_3	0.973	0.925	0.970	0.166
	x_4	0.631	0.786	0.703	0.766
	x_5	0.532	0.532	0.612	0.410
	x_6	0.230	0.243	0.286	0.204
	x_7	0.614	0.566	0.670	0.782

sponse variable and $x_1 - x_3$, whereas for x_4, x_5, x_6 and x_7, the largest values of $|h_{max}|$ are by case 9, 30, 9 and 25, respectively.

For censoring perturbations, case 37 stands out as the most influential, but with influence in the opposite direction ($h_{max} = -0.652$), with case 2 and 1 in second and third places (see Figure 7.10).

Next, cases 9, 24 and 37 are removed to examine whether they would affect the parameter estimates. Case 9 was selected as it is the most influential case in 2 variables x_4 and x_6 and additionally, an analysis by Lawless (1982) showed cell type was marginally significant, in particular, cell type

adeno, x_6 which is associated with an increase risk of failure.

Case 24 was chosen since it is the most influential to response perturbations as well as to $x_1 - x_3$. Likewise, the study by Lawless (1982) presented x_1 as highly significant. Later, case 37 was picked out from the data set as it is most influential on variance and censoring perturbations.

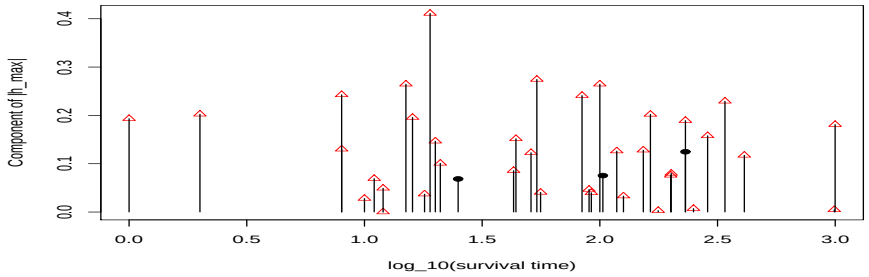


Figure 7.9: Index plots of $|h_{max}|$ for perturbing response variable for Lung cancer data

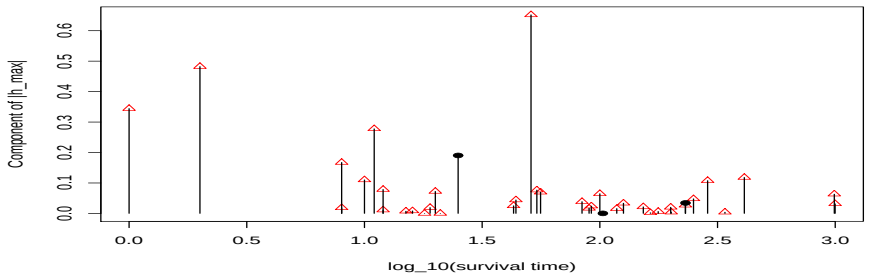


Figure 7.10: Index plots of $|h_{max}|$ for perturbing censoring status for Lung cancer data

However, deletion of any of these observations from the data set results in minor changes in the parameter estimates except for x_3. From Table 7.4, one finds the deletion of case 24 causes the change in β_3 sign, nevertheless, the p-value remains the same. Findings of other parameter estimates agree

well with the results of analyses done by Reid and Crepeau (1985) and Weissfeld (1990).

7.5 Conclusion

In this chapter, two new diagnostics analyses are proposed for the Buckley-James model. The first analysis, which can be called renovated Cook's distance, produces comparable results with previous findings. Nevertheless, this method cannot identify influential observations from the censored group as renovated leverage (Smith and Zhang, 1995). It can only detect influential observations from the uncensored group. This issue needs further investigation because of the possibility of censored points becoming influential cases in censored regression.

The second approach uses the local influence method. This approach follows Shi (1997), where they combined a general influence function and generalised Cook statistic as a new concept of local influence. This concept is easier to apply without considering a likelihood assumption. This method is able to assess the effect of perturbations to the data will have on inferences. It successfully discovers influential observations from both groups, i.e the censored and uncensored groups.

In this chapter, the Stanford heart transplant data and the lung cancer data are used for illustrations. These two methods are computationally simple and the results are easy to display through plotting. The diagnostics computation and graphics in this chapter's illustrations are done using the R software package.

Chapter 8

Contributions and Future Work

8.1 Contributions and conclusions

This section lists the contributions that this thesis has made to the outliers issue, cluster analysis and diagnostics analysis for the Buckley-James method, and outlines its main conclusions.

8.1.1 Contributions

This thesis mainly studies three problems: first is the identification of outliers in multivariate data set, second is a design of dissimilarity measure for clustering purpose and third is about diagnostics analysis for the Buckley-James method. This study has been pursued through the following efforts:

(i). Influence eigenstructure for identification of outliers.

In Chapter 3 we explored techniques based on influence eigenstructure for identifying outliers and identified four such techniques for identifying outliers .

They are:

- Influence eigen, Δ_i^*;
- Normalized influence eigen, Δ_i^{**};
- Influence angle, $\theta_{j(i)}$;
- Modified influence angle, $\theta^*_{j(i)}$.

These four techniques use the maximum eigenvalue and the corresponding eigenvector. The choice of the largest eigenvalue as the object of interest was motivated by its importance for many techniques. Examples are principle component analysis and the possibility of its statistical use as test statistics (Bejan, 2005). Gao et al. (2005) mentioned that examination of the observations' effect on the maximum eigenvalue is very significant because outliers that lie in the direction close to the maximum eigenvalue or vice versa, will change the maximum eigenvalue.

Chapter 3 does not distinguish between the various reasons for identifying outliers. The aim is to inform the analyst of observations that are considerably different from the majority. The techniques are therefore exploratory and applicable to a wide variety of settings. They are also well suited for identifying outliers in high dimensional data.

(ii). Influence Angle Cluster Approach (*iaca*).

In Chapter 4, a new dissimilarity measure is proposed for clustering purposes. The dissimilarity measure is also one of the techniques for identifying outliers, i.e. influence angle $\theta^*_{j(i)}$. It can be called the Influence Angle Cluster Approach (*iaca*), in order to differentiate it from influence angle for outliers detection, which was referred to in Chapter 3.

iaca successfully produces a cluster when it is used in partitioning clustering, even if the data set has mixed variables, i.e. interval and

categorical variables. *iaca* is developed based on the influence eigenstructure. It can obtain clusters easily and hence, avoid the curse of dimensionality. It is also flexible to implement, and seems to work well in practice. *iaca* can deal with continuous, categorical or mixed variables. Additionally, when *iaca* is used as a dissimilarity measure in partitioning clustering, those algorithms produce a good clustering structure compared to ones using the Euclidean distance, *daisy* or Manhattan distance as dissimilarity measures.

(iii). New diagnostics analysis for the Buckley James method.

In Chapter 7, two new diagnostics analyses for censored data that use the Buckley-James method are introduced. The first diagnostic analysis is called the renovated Cook's distance, RD_i^*. It produces comparable results to existing findings. The renovated Cook's distance also seems to have advantages (depending on the analyst's requirements) because it measures the influence of observation i on all fitted values, $\hat{y}$ instead of a single fitted value, $\hat{y}_i$ for case i such as DFIT_i^* as described by Smith (2002)

$$\text{DFIT}_i^* = x_i^T \hat{\beta}^* - x_i^T \hat{\beta}^*_{(i)}.$$

Moreover, this approach is also appropriate if one wants to measure the influence of observation i in whole variables rather than using $\text{DBETA}_i^* = \hat{\beta}^* - \hat{\beta}^*_{(i)}$ from Smith (2002). By using RD_i^* information for all variables, p can be considered simultaneously.

Chapter 7 also proposes the local influence approach for the Buckley-James method. The idea of local influence is based on general influence function and generalized Cook's statistics as used by Shi (1997). This idea is easier to apply without considering a likelihood assumption. This method is able to assess the effect of perturbations to the observations that will then influence inferences. It successfully discovers influential observations from both groups, i.e censored and

uncensored groups, as opposed to the current diagnostics used with the Buckley-James method, i.e. renovated leverage value Smith and Zhang (1995). The chapter presents the local influence diagnostics of the Buckley-James model, which consist of

- variance perturbation;
- response variable perturbation;
- censoring status perturbation;
- independent variables perturbation.

8.1.2 Conclusions

This study focused on three aspects related to diagnostics analysis for multivariate data and censored regression. The main conclusions are explained by chapter, with respect to the aims of the thesis as follows.

Recall that the first problem studied in this thesis is about identifying outliers using the influence eigenstructure. First, in Chapter 2, a simple introduction about outliers and the effects of outliers on the analysis are given in order to justify the significance of studying the outliers problem. Chapter 2 also explains about existing outliers detection methods as well as the motivations for using techniques based on influence eigenstructure. Chapter 3 discusses and presents some results about identifying outliers based on the influence eigenstructure. Chapter 3 also introduces and describes four techniques for identifying outliers. The objective of the techniques reported in this chapter is to advise the analyst of observations that are considerably different from the majority. The techniques in this chapter are, therefore, exploratory and they are applicable to a wide variety of settings. Techniques explored in this chapter can be performed on large and small data sets. They are used as to evaluate the deviation between observations. Observations that are further away from the remaining data are considered as outliers. Note that Chapter 3 formulates the problem

of outliers without assuming any model distribution. As discussed in the introduction of Chapter 3, the test of significance for outliers using the eigenstructure, such as principal component analysis, has not been widely used (Jolliffe, 2002). Consequently, the best advice is that the observation that is obviously more extreme than most of the remaining observations in the data set should be examined. This can be done simply through graphical illustration.

Hence, apart from considering the techniques for identifying outliers, Chapter 3 also concentrates on the graphical tool for diagnostics, i.e. identification of outliers. The illustrative results in Chapter 3 show that if the ith observations are potential outliers, their values for those techniques used in this chapter are all situated at the top of the index plot. This is because an outlier causes the difference of the eigenvalue of the full data set and the data set without ith observation to be larger than other observations. Recall that $\lambda_{1(i)}$ value is smaller for an outlier. Note that Chapter 3 only shows numerical examples for the techniques of influence eigenstructure for a single observation, even though it also describes techniques to handle multiple bad observations. Recall that if the data set has more than one outlier, the cases may mask each other, making finding the outliers difficult and an influence measure for the multiple cases is needed.

Only the techniques of influence eigenstructure for a single observation are shown in Chapter 3 because they are able to handle a masking effect in all data sets used for the numerical examples in Chapter 3. Hence, this problem is not further investigated using the influence eigenstructure for multiple observations. Recall that Chapter 3 shows that the examples of Mahalanobis distance cannot identify all fourteen outliers in Hawkin Bradu Kass data because of the masking effect.

It is also noted that some techniques for identifying outliers are also available for finding clusters. In Chapter 5, one of the techniques for identifying outliers is the dissimilarity measure for clustering purpose, i.e. influence angle. In Chapter 5, it is called the Influence Angle Cluster Ap-

proach (*iaca*), since it now measures the dissimilarity between every observation for clustering purposes, as well as to avoid any reader confusion.

The reason for choosing only the influence angle as a dissimilarity measure is Jolliffe (2002) shows that principal component analysis is also capable of finding clusters in a data set. Note that the influence angle is partially developed by the principal component score and the outliers appear to form a cluster, separated from the other observations in the data set. Chapter 4 briefly discusses previous measurement tools for clustering and the common clustering techniques. *iaca* is evaluated by examining the performance on continuous, categorical or mixed variables. A comparative study is also undertaken in this chapter. This is to identify whether *iaca* as a dissimilarity measure on partitioning clustering performs at a similar or better rate of efficiency when it is compared to Euclidean distance, *daisy* and Manhattan distance dissimilarity measures. It is noted that *iaca* produces a good clustering structure compared to Euclidean distance, *daisy* and Manhattan distance when it is used on those clustering algorithms.

Recall that the first two problems in this thesis are dealing with a complete data set. It is noted that using the incomplete data set, i.e. censoring data set, is also very important. This type of data set is widely used in biological science, educational testing and econometrics analysis.

Chapter 6 mainly looks at censoring data sets and the analysis that can handle this type of data set, particularly survival analysis data. Initially the idea of survival analysis and its relation to the censoring data is briefly discussed. Note that in censoring, there are three types of censoring, namely censoring type I, censoring type II and random censoring. This chapter discusses the censoring type I, i.e. right censored data.

There are many methods in survival analysis that can be used to analyse censoring data. One of them is the survival regression method. Examples of the survival regression method are the Cox method, Miller's method, the Buckley-James method and the Koul-Susarla-Van Ryzin model. Chapter 6 deals with the Buckley-James method as this method's perfor-

mance is comparable with the Cox method and performs better than both Miller's method and the Koul-Susarla-Van Ryzin model.

Chapter 6 explained about previous comparative studies showing that the Buckley-James estimator is more stable and it can be more easily explained to non-statisticians than the Cox model. However, now days, researchers are interested in using the Cox method instead of the Buckley-James method. This is so because of the relative lack of availability of Buckley-James method in the statistical software and limited choices of diagnostics analysis (Glasson, 2007). Hence, two new diagnostics analyses for the Buckley-James method are proposed in Chapter 7.

Chapter 7 introduces a diagnostics analysis called renovated Cook's distance. This method produces comparable results with existing findings. Nevertheless, this method cannot identify influential observations as renovated leverage from the censored group (Smith and Zhang, 1995). It can only detect influential observations from the uncensored group. This issue needs further investigation because of the possibility of censored points becoming influential cases in censored regression.

Another diagnostic analysis method introduced in Chapter 7 is the local influence method. This approach is similar to that of Shi (1997), where they combined a general influence function and generalized Cook statistics to create a new concept of local influence. This concept is easier to apply without considering a likelihood assumption. This method is able to assess the effect of perturbations to the data will have on inferences. Chapter 7 shows that the second approach successfully discovers influential observations from both groups, i.e the censored and uncensored groups. The local influence diagnostics of the Buckley-James model in Chapter 7 consists of variance perturbation, response variable perturbation, censoring status perturbation and independent variables perturbation.

8.2 Future Work

The following paragraphs outline the number of areas for possible future work on each problem. These areas either describe aspects of this thesis's work that might be worthy of further investigation, or that arise as a consequence of the findings of this thesis.

Chapters 2 and 3 discuss the issue of outliers identification. In particular, Chapter 3 discusses two issues -first, which observations could be candidates for influential observations, i.e. an outlier, and second, how can one evaluate the influence of more than one observation, i.e. multiple outliers. Chapter 3 presents the influence eigenstructure for a single observation, i and the influence eigenstructure for a multiple of observations, I.

However, note that the illustrative examples in Chapter 3 only consider the case of a single observation i. Thus, in future research it would also be useful to examine the influence eigenstructure for multiple observations by illustrative examples. This can be done by generating a data set which has severe masking and swamping problems. Thereby the performance of influence eigenstructure for I can be evaluated. Next, instead of identifying the outliers, another issue that one should consider is to determine whether the outliers are sufficiently extreme or influential to warrant further action.

Chapter 5 shows the Influence Angle Cluster Approach (*iaca*) performing as a dissimilarity measure in hierarchical clustering and partitioning clustering. In future research, *iaca* can be applied to other clustering algorithms, especially the high-dimensional algorithms.

Note that Chapter 7 only derives an influence measure for the single case, not for multiple cases. As mentioned before, an influence measure for multiple cases is needed to solve the masking effect. Thus future research could extend investigations of the influence measures for multiple cases for the Buckley-James method.

Bibliography

Acuna, E. and Rodriguez, C. (2004). A meta analysis study of outlier detection methods in classification, *Technical report*, Department of Mathematics, University of Puerto Rico, In proceedings IPSI 2004, Venice.

Andrews, D. F. and Pregibon, D. (1978). Finding the outliers that matter, *Journal of Royal Statist. Soc. B* **40**: 85–93.

Angiulli, F. and Pizzuti, C. (2002). Fast outlier detection in high dimensional spaces, *Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*.

Atkinson, A. C. (1981). Robustness, transformations and two graphical displays for outlying and influential observations in regression, *Biometrika* **68**: 13–20.

Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers, *Journal of The American Statistical Association* **89**(428): 1329–1339.

Barlow, W. E. (1997). Global measures of local influence for proportional hazards regression models, *Biometrics* **53**(3): 1157–1162.

Barnett, V. and Lewis, T. (1994). *Outliers in statistical data*, Wiley and Son, New York.

Bay, S. D. and Schwabacher, M. (2003). Mining distance-based outliers in near linear time with randomization and a simple pruning rule, *Pro-*

ceedings of Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C. USA, pp. 29–38.

Bejan, A. I. (2005). Largest eigenvalues and sample covariance matrices. tracy-widom and painlevè ii: Computational aspects and realization in s-plus with applications. Preprint http://www.vitrum.md/andrew/MScWrwck/TWinSplus.pdf.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression diagnostics identifying influential data and sources of collinearity*, John Wiley and Sons, New York.

Berkson, J. and Gage, R. R. (1950). Calculation of survival rate for cancer, *Staff Meetings, 25*, Mayo Clinic, pp. 270–286.

Breunig, M. M., Kriegel, H., Ng, R. T. and Sander, J. (2000). Lof: Identifying density-based local outliers, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, Texas, USA, pp. 93–104.

Buckley, J. and James, I. (1979). Linear regression with censored data, *Biometrika* **66**(3): 429–436.

Campbell, N. A. (1980). Robust procedures in multivariate analysis i: Robust covariance estimation, *Appl. Statist.* **29**(3): 231–237.

Caroni, C. and Billor, N. (2007). Robust detection of multiple outliers in grouped multivariate data, *Journal of Applied Statistics* **34**(10): 1241–1250.

Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity analysis in linear regression*, John Wiley, United States.

Cook, R. D. (1977). Detection of influential observation in linear regression, *Technometrics* **19**(1): 15–18.

Cook, R. D. (1986). Assessment of local influence, *Journal of Royal Statist. Soc. B* **48**(2): 133–169.

Cook, R. D. and Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression, *Technometrics* **22**(4): 495–508.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in regression*, Chapman and Hall, New York.

Cormack, R. M. (1971). A review of classification, *Journal of the Royal Statistic Society A* **134**: 321–367.

Cox, D. R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society B* **34**: 187–202.

Critchley, F. (1985). Influence in principal components analysis, *Biometrika* **72**: 627–636.

Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data, *Journal of the American Statistical Association* **72**(357): 27–36.

Cutler, S. J. and Ederer, F. (1958). Maximize utilization of the life table method in analysis survival, *Journal of chronic disease* **8**: 699–712.

David, H. A. (1978). *Contributions to survey sampling and applied statistics*, Academic Press, Inc, New York.

Davies, L. and Gather, U. (1993). The identification of multiple outliers, *Journal of the American Statistical Association* **88**(423): 782–792.

Domeniconi, C., Papadopoulos, D., Gunopulos, D. and Ma, S. (2004). Subspace clustering of high dimensional data, *Proceedings Fourth SIAM International Conference Data Mining*, pp. 517–521.

Escobar, L. A. and Meeker, W. Q. (1992). Assessing influence in regression analysis with censored data, *Biometrics* **48**(2): 507–528.

Everitt, B. S. (1993). *Cluster Analysis*, 3rd edn, Edward Arnold, London.

Everitt, B. S., Landau, S. and Leese, M. (2001). *Cluster Analysis*, 4th edn, Edward Arnold, London.

Fang, Z., Liu, L., Yang, J., Luo, Q.-M. and Li, Y.-X. (2006). Comparisons of graph-structure clustering methods for gene expression data, *Acta Biochimica et Biophysica Sinica* **38**(6): 379–384.

Filzmoser, P. (2004). A multivariate outlier detection method, *in* S. Aivazian, P. Filzmoser and Y. Kharin (eds), *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling*, Vol. 1, Belarusian State University, Minsk, pp. 18–22.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure data, *Biometrics* **42**: 845–854.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**: 179–188.

Franklin, S., Thomas, S. and Brodeur, M. (2000). Robust multivariate outlier detetection using mahalanobis distance and modified stahel-donoho estimators, *Proceeding International Conference on Establishment Surveys*, New York, pp. 697–706.

Fung, W. K. (1993). Unmasking outliers and leverage points: a confirmation, *Journal of the American Statistical Association* **88**: 515–519.

Gao, S., Li, G. and Wang, D. Q. (2005). A new approach for detecting multivariate outliers, *Communication in Statistics-Theory and Method* **34**: 1857–1865.

Gehan, E. A. (1969). Estimating survival function from the lifetable, *Journal of chronic disease* **21**: 629–644.

Glasson, S. (2007). *Censored Regression Techniques for Credit Scoring*, PhD thesis, RMIT University.

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data, *Biometrics* **28**: 81–124.

Gordon, A. D. (1999). *Classification*, 2nd edn, Chapman and Hall, FL.

Gower, J. (1971). A general coefficient of similarity and some of its properties, *Biometrics* **27**: 857–872.

Graybill, F. A. (1976). *Theory and application of the linear model*, Duxbury Press, Boston.

Greene, W. H. (2000). *Econometric Analysis*, Prentice Hall International, London.

Grubbs, F. (1969). Procedures for deleting outlying observation in samples, *Technometrics* **11**(1): 1–21.

Hadi, A. S. (1992). Identyfying multiple outliers in multivariate data, *Journal Royal Statistics Soc B* **54**(3): 761–777.

Hadi, A. S. (1994). A modification of a method for the detection of outliers in multivariate sample, *Journal Royal Statistics Soc B* **56**(2): 393–396.

Hair, J. H., Anderson, R. E., Tatham, R. L. and Black, W. C. (2005). *Multivariate data analysis*, Prentice Hall, United States.

Hald, A. (1949). Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point, *Skandinavisk Aktuarietid- skrift* **32**: 119–134.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). On clustering validation techniques, *Journal of Intelligent Information Systems* **17**: 107–145.

Hampel, F. R. (1971). A general qualitative definition of robustness, *Annals of Mathematics Statistic* **42**(6): 1887–1896.

Harrell, F. E. (1986). *SUGI Supple-mental Library User's Guide*, R. P. Hastings ed, SAS Institute Inc, North Carolina, chapter The PHGLM procedure, pp. 437–466.

Hartigan, J. A. (1975). *Cluster Analysis*, John Wiley, New York.

Hawkins, D. M. (1980). *Identification of outliers*, Chapman and Hall, London.

Hawkins, D. M. (1994). The feasible solution algorithm for the minimum covariance deteminant estimator in multivariate data, *Computational Statistics and Data Analysis* **17**(2): 197–210.

Hawkins, D. M., Bradu, D. and Kass, G. V. (1984). Location of several outliers in multiple regression data using elemental sets, *Technometrics* **26**(3): 197–208.

Hawkins, D. M. and Olive, D. J. (1999). Improved feasible solution algorithms for high breakdown estimation, *Computational Statistics and Data Analysis* **30**: 1–11.

Heller, G. and Simonoff, J. S. (1990). A comparison of estimators for regression with a censored response variable, *Biometrika* **77**(3): 515–520.

Heller, G. and Simonoff, J. S. (1992). Prediction in censored survival data: A comparison of the proportional hazards and linear regression models, *Biometrika* **48**(1): 101–115.

Hillis, S. L. (1993). A comparison of three buckley-james variance estimators, *Communication in Statistics B* **22**(4): 955–973.

Hillis, S. L. (1994). A heuristic generalisation of smith's buckley-james variance estimator, *Communications in statistics. Simulation and computation* **23**: 713–831.

Hillis, S. L. (1995). Residual plots for the censored data linear regression model, *Statistics in Medicine* **14**: 2023–2036.

Hocking, R. R. (2003). *Methods and applications of linear models: regression and the analysis of variance*, John Wiley and Sons, Hoboken, NJ.

Hodge, V. J. (2004). A survey of outlier detection methodologies, *Artificial Intelligence Review* **22**(2): 85–126.

Hodges, J. L. J. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location, *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability Date:*, Vol. 1, Univ. California Press, Berkeley, Calif, pp. 163–186.

Hu, T. and Sung, S. Y. (2003). Detecting pattern-based outliers, *Pattern Recognition Letters* **24**: 3059–3068.

Iglewics, B. and Martinez, J. (1982). Outlier detection using robust measures of scale, *Journal of Sattistical Computation and Simulation* **15**: 285–293.

Iglewicz, B. and Hoaglin, D. (1993). *How to detect and handle outliers*, ASQC Quality Press.

Jackson, J. E. (1991). *A user's Guide to Principle Component*, Wiley, NEw York.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*, Prentice Hall.

Jain, A., Murty, M. and Flynn, P. (1999). Data clustering: A review, *ACM Computing Surveys* **31**(3): 254–323.

James, I. R. (1986). On estimating equations with censored data, *Biometrika* **73**: 35–42.

James, I. R. and Smith, P. J. (1984). Consistency results for linear regression with censored data, *The Annals of Statistics* **12**(2): 590–600.

Jensen, R. (1969). A dynamic programming algorithm for cluster analysis, *Operations research* **17**: 1034–1057.

Johnson, R. A. and Wichern, D. W. (2007). *Applied multivariate statistical analysis*, Pearson Prentice Hall, New Jersey.

Johnson, W. and Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis, *Journal Amer. Statist. Ass* **78**: 137–144.

Jolliffe, I. T. (2002). *Principal Component Analysis*, Springer-Verlag, New York.

Juan, J. and Prieto, F. J. (2001). Using angles to identify concentrated multivariate outliers, *American Statistical Association and the American Society for Quality* **43**(3): 311–322.

Kanji, G. K. (1993). *100 Statistical Tests*, London : SAGE Publication Ltd.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal America Stat. Assoc* **53**: 457–481.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley, New York.

Kent, J. T. and O'Quigley, J. (1988). Measures of dependence for censored survival data, *Biometrika* **75**(3): 525–534.

Knorr, E. M. and Ng, R. T. (1998). Algorithms for mining distance-based outliers in large datasets, *Proceedings of the 24rd International Conference on Very Large Data Bases*, pp. 392–403.

Koul, H., Susarla, V. and Ryzin, J. V. (1981). Regression analysis with randomly righ censored data, *The annals of statistics* **9**(6): 1276–1288.

Kriegel, H.-P., Schubert, M. and Zimek, A. (2008). Angle-based outlier detection in high-dimensional data, *Proceedings of the Knowledge Discovery and Data Mining*, pp. 269–276.

Lai, T. L. and Ying, Z. (1991). Large sample theory of a modified buckley-james estimator for regression analysis with censored data, *The Annals of Statistics* **19**(3): 1370–1402.

Lawless, J. F. (1982). *Statistical models and methods for lifetime data*, John Wiley and Sons, New York.

Lee, E. T. (1980). *Statistical methods for survival data analysis*, Lifetime Learning, California.

Levine, E. and Domany, E. (2001). Resampling methods for unsupervised estimation of cluster validity, *Neural Computation* **13**: 2573–2593.

Lin, J. S. and Wei, L. J. (1992). Linear regression analysis based on buckley-james estimating equation, *Biometrics* **48**(3): 679–681.

Mahalanobis, P. C. (1930). On tests and measures of group divergence, *Journal of the asiatic society of bengal* **26**: 541–588.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Reports* **50**(3): 163–170.

Mardia, K. V. (1977). Mahalanobis distances and angles, *Multivariate Analysis IV*, North-Holland Pbulishing Company, 4th International Symposium on Multivariate Analysis, pp. 495–511.

Maronna, R. A. (1976). Robust m-estimators of multivariate location and scatter, *The Annals of Statistics* **4**(1): 51–67.

Maronna, R. A. and Yohai, V. J. (1995). The behavior of the stahel-donoho robust multivariate estimator, *Journal of the American Statistical Association* **90**(429): 330–341.

Mertens, B. J. A. (1998). Exact principle component influence measure applied to the analysis of spectroscopic data on rice, *Applied Statistics* **47**(4): 527–542.

Miller, R. G. (1976). Least squares regression with censored data, *Biometrika* **63**(3): 449–464.

Miller, R. and Halpern, J. (1982). Regression with censored data, *Biometrika* **69**(3): 521–531.

Milligan, G. W. and Cooper, M. C. (1985). An examination procedures for determining the number of clusters in a data set, *Psychometrika* **50**: 159–179.

Moon, C. G. (1989). A monte carlo comparison of semiparametric tobit estimators, *Journal of Applied Econometrics* **4**(4): 361–382.

Myers, R. H. (1990). *Classical and modern regression with applications*, PWS-KENT, USA.

Papadimitriou, S., Kitawaga, H., Gibbons, P. G. and Faloutsos, C. (2002). Loci: Fast outlier detection using the local correlation integral, *Technical report*, Intel research Laboratory. Technical report no. IRP-TR-02-09.

Papadimitriou, S., Kitawaga, H., Gibbons, P. G. and Faloutsos, C. (2003). Loci: Fast outlier detection using the local correlation integral, *Proceedings of the 19th International Conference on Data Engineering*, pp. 315–328.

Pena, D. and Prieto, F. J. (2001). Multivariate outlier detection and robust covariance matrix estimation, *Technometrics* **43**(3): 286–299.

Penny, K. I. and Jolliffe, I. T. (2001). A comparison of multivariate outlier detection methods for clinical laboratory safety data, *The Statistician* **50**(3): 295–308.

Pettitt, A. N. and Daud, I. B. (1989). Case-weighted measures of influence for proportional hazards regression, *Applied Statistics* **38**(1): 51–67.

Pickard, L., Kitchenham, B. and Linkman, S. J. (2001). Using simulated data sets to compare data analysis techniques used for software cost modelling, *IEE Proc-Softw* **148**(6): 165–174.

Radhakrishnan, R. and Kshirsagar, A. M. (1981). Influence functions for certain parameters in multivariate analysis, *Communication in Statistics-Theory and Method* **10**(6): 515–529.

Ramaswamy, S., Rastogi, R. and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets, *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 427–438.

Reid, N. (1981). Influence functions for censored data, *The Annals of Statistics* **9**(1): 78–92.

Reid, N. and Crepeau, H. (1985). Influence functions for proportional hazards regression, *Biometrika* **72**(1): 1–9.

Richard, P. and Julian, P. (1972). Asymptotically efficient rank invariant test procedures, *Journal of the Royal Statistical Society A* **135**(2): 185–207.

Ritov, Y. (1990). Estimation in a linear regression model with censored data, *Annals of Statistics* **18**: 303–328.

Rocke, D. M. and Woodruff, D. L. (1996). Identification of outliers in multivariate data, *Journal of The American Statistical Association* **91**(435): 1047–1061.

Rose, K. (1998). Deterministic annealing for clustering, compression, classification, regression and related optimization problems, *Proceeding IEEE*, Vol. 86, pp. 2210–2239.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics* **20**: 53–65.

Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator, *American Statistical Association and the American Society for Quality* **41**(3): 212–223.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust regression and outlier detection*, John Wiley, New York.

Rousseeuw, P. J. and von Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association* **85**(411): 633–639.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples), *Biometrika* **52**: 591–611.

Shi, L. (1997). Local influence in principle component analysis, *Biometrika* **84**(1): 175–186.

Siotani, M. (1959). The extreme value of the generalized distance of the individual points in the multivariate normal sample, *Annals of the Institute of Statistical Mathematics* **10**: 183–208.

Smith, P. J. (1986). Estimation in linear regression with censored response, *Pacific Statistical Congress*, Amsterdam, Holland, pp. 261–265.

Smith, P. J. (1988). Asymptotic properties of linear regression estimators under a fixed censorship model, *Australian Journal of Statistics* **30**: 52–66.

Smith, P. J. (1996). Renovating interval-censored responses, *Lifetime Data Analysis* **2**: 1–11.

Smith, P. J. (2002). *Analysis of failure and survival data*, Chapman and Hall, United States.

Smith, P. J. (2004). Using linear regression techniques with censored data, *International Journal of Realibility, Quality and Safety Engineering* **11**(2): 163–173.

Smith, P. J. and Peiris, L. W. (1999). Added variable plots for linear regression with censored data, *Communication in Statistics-Theory and Method* **28**(8): 1987–2000.

Smith, P. J. and Zhang, J. (1995). Renovated scatterplots for censored data, *Biometrika* **82**(2): 447–452.

Stare, J., Heinzl, H. and Harrell, F. (2000). On the use of buckley and james least squares regression for survival data, *New Approach in Applied Statistics* **12**: 125–134.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis, *Psychological Bulletin* **95**: 334–344.

Struyf, A., Hubert, M. and Rousseeuw, P. J. (1996). Clustering in an object-oriented environment, *Journal of Statistical Software* . 1. http://www.stat.ucla.edu/journals/jss/.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data, *Annals of Statistics* **18**: 354–372.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison Wesley, Reading, MA.

Vanderviere, E. and Hubert, M. (2008). An adjusted boxplot for skewed distributions, *Computational Statistics and Data Analysis* **52**(12): 5186–5201.

Velleman, P. F. and Welsch, R. E. (1981). Efficient computing of regression diagnostics, *The American Statistician* **35**(4): 234–242.

Wang, D. Q. and Scott, D. J. (1989). Testing a markov chain for independence, *Communication in Statistics-Theory and Method* **18**(11): 4085–4103.

Wang, D. Q., Zhang, L., Ahmed, S. E. and Aziz, N. (2009). Renovated partial plots and hat matrix for censored regression model, *Pakistan Journal of Statistics* **25**(4): 631–645.

Wang, H., Wang, W., Yang, J. and Yu, P. S. (2002). Clustering by pattern similarity in large data sets, *Proceeding ACM SIGMOD International Conference on Management of Data*, pp. 394–405.

Wang, S.-G. and Liski, E. P. (1993). Effects of observations on the eigensystem of a sample covariance matrix, *Journal of Statistical Planning and Inference* **36**: 215–226.

Wang, S. G. and Nyquist, H. (1991). Effects on the eigenstructure of a data matrix when deleting an observation, *Computational Statistics and Data Analysis* **11**(2): 179–188.

Wei, L. J., Ying, Z. and Lin, D. Y. (1995). Linear regression analysis of censored survival data based on rank tests, *Journal of Statistical Computation and Simulation* **77**: 845–851.

Weissfeld, L. A. (1990). Influence diagnostics for the proportional hazards model, *Statistics and Probability Letters* **10**(5): 411–417.

Weissfeld, L. A. and Schneider, H. (1987). Inferences based on the buckley-james procedure, *Communication in Statistics A* **16**: 177–187.

Weissfeld, L. A. and Schneider, H. (1990). Influence diagnostics for the normal linear model with censored data, *Australian Journal Statistics* **32**(1): 11–20.

Wilk, S. S. (1963). Multivariate statistical outliers, *Sankhya* **25**: 407–426.

Williams, G. J., Baxter, R. A., He, H. X., Hawkins, S. and Gu, L. (2002). A comparitive study of rnn for outlier detection in data mining, *Technical report*, Japan. Technical report CMIS-02/102.

Wu, C. P. and Zubovic, Y. (1995). A large-scale monte carlo study of the buckley-james estimator with censored data, *Journal of Statistical Computation and Simulation* **51**: 97–119.

Xing, E. P. and Karp, R. M. (2001). Cliff: Clustering of high dimensional microarray data via iterative feature filtering using normalized cuts, *Bioinformatics* **17**(1): 306–315.