LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Department of Statistics and Mathematical Sciences

COMPUTATIONAL METHODS FOR TRANSFORMATIONS TO MULTIVARIATE NORMALITY

by

Ham–Mukasa Mulira

Thesis submitted for the degree of

Doctor of Philosophy in the University of London

April, 1992

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

The classical multivariate theory has been largely based on the multivariate normal distribution (MVN): the scarcity of alternative models for the meaningful and consistent analysis of multiresponse data is a well recognised problem. Further, the complexity of generalising many non–normal univariate distributions makes it undesirable or impossible to use their multivariate versions. Hence, it seems reasonable to inquire about ways of transforming the data so as to enable the use of more familiar statistical techniques that are based implicitly or explicitly on the normal distribution.

Techniques for developing data–based transformations of univariate observations have been proposed by several authors. However, there is only one major technique in the multivariate (p–variable) case by Andrews et. al. [1971]. Their approach extended the power transformations proposed by Box & Cox [1964] to the problem of estimating power transformations of multiresponse data so as to enhance joint normality. The approach estimates the vector of transformation parameters λ by numerically maximising the log–likelihood function. However, since there are several parameters to be estimated, $p(p+5)/2$ for multivariate data without regression, the resulting maximisation is of high dimension, even with modest values of p and sample size n. The purpose of the thesis is to develop computationally simpler and more informative statistical procedures which are incorporated in a package. The thesis is in three main parts:

– A proposed complementary procedure to the log–likelihood approach which attempts to reduce the size of the computational requirements for obtaining the estimates of λ. Though computational simplicity is the main factor, the statistical qualities of the estimates are not compromised, indeed the estimated values are numerically identical to those of the log–likelihood. Further, the procedure implicitly produces diagnostic statistics and some useful statistical quantities describing the structure of the data. The technique is a generalisation of the constructed variables method of obtaining quick estimates for transformation parameters [Atkinson 1985]. To take into account the multiresponse nature of the data and, hence, joint estimates for λ, a seemingly unrelated regression is carried out. The algorithm is iterative. However, there is considerable savings in the number of iterations required to converge to the maximum likelihood (MLE) estimates compared to those using the log–likelihood function. The technique is refered to as the *Seemingly Unrelated Regressions/Constructed Variable* (*SURCON*) analysis, and the estimates obtained are the *Surcon estimates.*

– The influence of individual observations on the need for transformations is quite crucial and, hence, it is necessary to investigate the data for any spurious or suspicious observations, outliers. The thesis also proposes an iterative technique for detecting and identifying outliers based on Mahalanobis distances computed from sub–samples of the observations. The results of the analysis are displayed in a graphical summary called the *Stalactite Chart*, hence, the analysis is refered to as the *Stalactite Analysis.*

– The development of a userfriendly microcomputer–based statistical package which incorporates the above techniques. The package is written in the C programming language.

To Rebecca Nakabito and Henry–Medadi Ndaula

COMPUTATIONAL METHODS FOR TRANSFORMATIONS
TO MULTIVARIATE NORMALITY

TABLE OF CONTENTS

	Page
Acknowledgements	7
List of Figures and Tables	8
CHAPTER ONE - INTRODUCTION	
1.1 Background	12
1.2 The role of the Multivariate Normal (MVN) distribution	15
1.3 The Stalactite Analysis	16
1.4 The SURCON Analysis	17
1.5 The tSTAT Package	17
1.6 Notation	18
CHAPTER TWO - MULTIVARIATE OUTLIER DIAGNOSTICS	
2.1 Introduction	20
2.1.1 The Outlier Model	23
2.1.1.1 Single Observation Formulation	23
2.1.1.2 General Formulation	24
2.2 Identification of Outliers	32
2.2.1 Discordancy Tests	32
2.2.2 Mahalanobis Distance	36
2.2.2.1 Case Deletion	38
2.2.2.2 Distribution of Mahalanobis Distance	40
2.3 Graphical Techniques	41
2.3.1 Univariate and Bivariate Plots	42
2.3.1 Multivariate Plots	42
2.3.2.1 Scatter Plot Matrix	43
2.3.2.2 Index Plot	43
2.3.2.3 Probability Plots	44
2.3.3 Parallel Coordinates Plots (Z-curves)	45
2.4 Classical Mahalanobis Distance Approach	46

2.5 Minimum Volume Ellipsoid (MVE) Approach 46
2.6 Hat Matrix Approach 49
2.7 Proposed Stalactite Analysis Approach 50
2.8 Examples 59
CHAPTER THREE - TRANSFORMATIONS TO MULTIVARIATE NORMALITY
3.1 Introduction 119
3.2 Marginal Symmetry 124
3.2.1 Graphical Assessment of Symmetry 124
3.2.2 Transformations to Symmetry 127
3.3 Likelihood Approach 128
3.4 The Proposed SURCON Approach 132
3.4.1 Seemingly Unrelated Regressions (SUR) 132
3.4.1.1 Estimation with unknown covariance matrix 136
3.4.1.2 Hypothesis testing 136
a) Linear Restrictions on the coefficient vector β 136
b) Testing for a diagonal covariance matrix σ 139
3.4.2 Constructed Variables 139
3.4.2.1 Structured sample case 139
3.4.2.2 Unstructured sample case 143
3.4.3 The proposed SURCON Analysis algorithm 143
3.4.3.1 Hypothesis testing 145
a) Significance of γ 145
b) Testing for independence of the variables 147
3.4.3.2 Convergence of SURCON estimates to ML estimate 147
3.5 Assessing Normality 154
3.5.1 Probability Plots 154
3.5.2 Rao's Score Test 154
3.6 Examples 159
CHAPTER FOUR - tSTAT PACKAGE
4.1 Introduction 192
4.2 System Design 194
4.3 Main Algorithms 196

4.3.1 Summary Statistics 196
4.3.2 Stalactite Analysis 198
4.3.2.1 Initial Sub-sample Selection 198
4.3.2.2 Matrix Inversion 200
4.3.2.3 Mahalanobis Distance 201
4.3.2.4 Stalactite Analysis - The Complete Algorithm 202
4.3.3 SURCON Transformations 202
4.3.4 Probability Distributions 202
a) Normal N[0,1] 205
b) Students-t t(κ) 206
c) Chi-square $\chi^2(p)$ 209
d) Fisher $F_{\kappa 1 \kappa 2}$ 211
4.4 User Manual 214
4.4.1 Using the tSTAT package 214
4.4.2 tSTAT Functions 223
4.4.2.1 Main menu 223
4.4.2.2 Data Option 224
4.4.2.3 Statistics Option 226
4.4.2.4 Transformations Option 229
4.2.4.5 Plots Option 231
4.2.4.6 Outliers Option 233
4.2.4.7 Utilities Option 234
4.2.4.8 Ending a tSTAT Session 236
4.5 Example 237
CHAPTER FIVE - CONCLUSIONS AND RECOMMENDATIONS 238
APPENDIX
Appendix A Expected Maximum χ^2_p i) n 50; ii) n 60 242
Appendix B Example of tSTAT Output 244
Appendix C Data Sets of Chapter Four 254
BIBLIOGRAPHY 256

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank all those who made it possible for me to carry out this research. Firstly, I wish to express my sincere gratitute to Professor Anthony Atkinson, my supervisor, whose invaluable guidance, patience and careful supervision has made it all possible. I would also like to thank Professor Sam Tulya–Muhika and Dr.James Ntozi, the consecutive Directors of the Institute of Statistics and Applied Economics, Makerere University, Kampala, Uganda, who have been very supportive throughout the entire period of the research, Dr.Chris Kershaw of Rothamsted Experimental Station who so kindly gave me some initial ideas and data. I also thank the United Nations Economic Commission for Africa (UNECA) for having provided me with the initial funding. It is not possible to mention all those who have been of assistance in so many different ways but my gratitude to them is immense. I would also like to thank my parents for the initial inspiration and continual motivation. Lastly, I wish to thank my wife, Nora, for her patience and understanding through it all together with my children, Rebecca and Henry, for having endured it without not knowing what it was all about.

List of Figures (Main Text)

1. Figure 2.1 Types of outliers in 2–dimensions.

2. Figure 2.2 Parallel Coordinates Plot (Z–Curves) for Four Variables.

3. Figure 2.3 Direction of Pull in Two Dimensional Space.

4. Figure 2.4 Direction of Pull in p–Dimensional Space.

5. Figure 3.1 Bivariate Densities.

6. Figure 3.2 Test for Symmetry Plots.

7. Figure 3.3 Likelihood Ratio Test, Wald Test and Score Test.

8. Figure 3.4 Loglikelihood Contours for BVN data.

9. Figure 3.5 Score Statistic vs Lambda. (Bivariate Normal Data, ρ=0, y_1).

10. Figure 3.6 Loglikelihood Surface Plot for BVN data.

11. Figure 4.1 System Flowchart for the tSTAT Package.

12. Figure 4.2 Flowchart for the Initial Sub–sample Selection for the Stalactite Analysis Algorithm (Selection Sampling).

13. Figure 4.3 Flowchart for the Stalactite Analysis Algorithm.

14. Figure 4.4 Flowchart for the SURCON Analysis Algorithm.

15. Figure 4.5 Main Screen.

16. Figure 4.6 The tSTAT Menu Structure.

17. Figure 4.7 Data Editor.

18. Figure 4.8 Data Editor Menu Structure.

List of Figures (Examples)

1. Figure E.1 Simulated Bivariate Normal Data (with no outliers).
(a) Box Plots; (b) Scatter Plot; (c) Normal Plot of Z(r–i); (d) Stalactite Chart;
(e) Index Plot of the Mahalanobis Distances (90% Sample); (f) Index Plot of the Mahalanobis Distances (Full Sample); (g) Means Plot.

2. Figure E.2 Simulated Bivariate Normal Data (with 4 outliers).
(a) Box Plots; (b) Scatter Plot; (c) Normal Plot of Z(r–i); (d) Stalactite Chart;
(e) Index Plot of the Mahalanobis Distances (90% Sample); (f) Index Plot of the Mahalanobis Distances (Full Sample); (g) Means Plot.

3. Figure E.3 Belgian Phone Calls Data.
(a) Box Plot; (b) Scatter Plot; (c) Normal Plot of Z(r–i); (d) Stalactite Chart;

4. Figure E.4 Hertzsprung–Russell Star Data.
(a) Box Plots; (b) Scatter Plot; (c) Normal Plot of Z(r–i); (d) Stalactite Chart;
(e) Means Plot.

5. Figure E.5 Hawkins, Bradu and Kass Artificial Data.
(a) Box Plots; (b) Scatter Plot; (c) Normal Plot of $Z(r_{-i})$. i/ y1 & y2; ii/ y1 & y3; iii/ y2 & y3.5; (d) Stalactite Chart (e) 50% sub–sample squared Mahalanobis distance Index Plot; (f) 90% sub–sample squared Mahalanobis distance Index Plot (MIP); (g) 100% full–sample squared Mahalanobis distance Index Plot (MIP); (h) 50% sub–sample Normal Plot of cube–root (MD^2); (i) 90% sub–sample Normal Plot of cube–root(MD^2); (j) full–sample Normal Plot of cube–root(MD^2); (k) Means Plot; (l) Stalactite Analysis (with initial subsample of 14 observations); (m) (Non–random start) Means Plot.

6. Figure E.6 Simulated Bivariate Normal Data (w/o outliers).
(a) Test for Symmetry Plots i/ Type I; ii/ Type II; iii/ Type III; (b) Surcon Analysis;

7. Figure E.7 Simulated Bivariate Normal Data (with 4 outliers).
(a) Surcon Analysis; (b) Transformed Bivariate Normal Data (with 4 outliers) Stalactite Chart.

8. Figure E.8 Peruvian Data (with observation 39 omitted).
(a) Normal Plots. i/ Weights; ii/ Heights; (b) Chi–square Probability Plot i/ Full sample; ii/ Without observation 39; (c) Surcon Analysis; (d) Surcon Analysis (w/o observation 39).

9. Figure E.9 Surcon Analysis for the Minitab Tree Data (Response vs X_2).

10. Figure E.10 Fisher's Iris Setosa Data (w/o petal width).
(a) Surcon Analysis; (b) Stalactite Chart; (c) Index Plot of Estimated Lambda with Case Deletion. i/ Sepal length; ii/ Sepal width; iii/ Petal length; iv/ Petal width; (d) Surcon Analysis for Fisher's Iris Versicolor; (e) Surcon Analysis for Fisher's Iris Virginica; (f) Surcon Analysis for Fisher's Iris Data (Setosa+Versicolor+Viginica).

11. Figure E.11 Repeat Soil Sampling Survey Data (RSSS). (a) Surcon Analysis;
(b) Histograms.

List of Tables (Main Text)

1. Table 2.1 Discordancy Tests for a Single Outlier.

2. Table 2.2 Lower Bounds of Sample Size n for E[Max χ_p^2] > $\chi_p^2(1-a)$.

2. Table 3.1 Helpful Transformations to Near Normality.

3. Table 3.2 Type of elements in $x_{(3)}$ and $x_{(4)}$.

4. Table 3.3 Hinkley's Quick Transformations to Marginal Symmetry.

5. Table 3.4 Summary of the Box–Cox Transformations to Joint Normality.

6. Table 4.1 Executable File to Module Structure.

7. Table 4.2 Deletion Summary Statistics.

List of Tables (Examples)

1. Table E.1 Simulated Bivariate Normal Data (with no outliers).
(a) Data; (b) Summary Statistics; (c) M–Estimators;

2. Table E.2 Simulated Bivariate Normal Data (with 4 outliers).
(a) Data; (b) Summary Statistics; (c) M–Estimators; (d) Stalactite Analysis.

3. Table E.3 Belgian Phone Calls Data.
(a) Data; (b) Summary Statistics (c) M–Estimators; (d) Stalactite Analysis; (e) Case Deletion Correlation Coeffiecient, Diagonal Elements of the Hat Matrix, Mahalanobis Distances and Stalactite Scores.

4. Table E.4 Hertzsprung–Russell Diagram of the Star Cluster CYG OB1 Data.
(a) Data; (b) Summary Statistics; (c) M–Estimators; (d) Stalactite Analysis; (e) Case Deletion Correlation Coefficient, Diagonal Elements of the Hat Matrix, Mahalanobis Distances and Stalactite Scores.

5. Table E.5 Hawkins–Bradu–Kass Artificial Data.
(a) Data; (b) Summary Statistics; (c) M–Estimators; (d) Stalactite Analysis;
(e) Stalactite Analysis (with initial subsample of 14 observations); (f) Diagonal Elements of the Hat Matrix, Squared Mahalanobis Distances, MVE Robust Distances and Stalactite Scores.

6. Table E.6 Effect of Correlation on Transformations to Joint Normality.

CHAPTER ONE

1.0 INTRODUCTION

1.1 Background

This thesis deals with the problem of transforming multivariate data to the multivariate normal (MVN) distribution. In general, the transformations can be based on theoretical considerations or be estimated from the data that are being analysed. The thesis concentrates on the latter which is sometimes refered to as "data–based" transformation. Techniques for data–based transformations of univariate data have been proposed by several authors e.g. Moore & Tukey [1954], Box & Cox [1964] and Andrews [1971]. However, there is only one major technique in the general multivariate (p–variable) case. The technique is by Andrews et.al. [1971] and is an extension of the power transformations to normality, proposed by Box & Cox [1964], to multivariate data so as to enhance joint normality. The approach estimates the vector of transformation parameters λ by numerically maximising the log–likelihood function. Since there are several parameters to be estimated, $p(p+5)/2$ for p dimensional multivariate data without regression, the resulting maximisation problem is of high dimension even with modest values of p and sample size n.

The main aim of the thesis, therefore, is to propose a complementary procedure to the log–likelihood approach which attempts to reduce the size of the computational requirements for obtaining the estimates λ. Although computational simplicity is the focus of the technique, the statistical qualities of the estimates are not compromised; indeed the estimates derived are numerically identical to those from the log–likelihood. The procedure also implicitly produces diagnostic statistics and some useful quantities which describe the structure of the data. The technique is a combination of two regression analysis methods, namely, that of obtaining "quick" estimates for transforming the response in a regression model using constructed variables [Atkinson 1985] and that of seemingly unrelated regressions [Zellner 1962] and is thus referred to as the *Seemingly Unrelated*

Regressions/Constructed Variable (*SURCON*) analysis. The estimates obtained are called the *Surcon estimates.* The SURCON method is an iterative generalisation of the constructed variables method, the seemingly unrelated regressions being adopted to take into account the multiresponse nature of the data through the covariance structure. There is considerable savings in the number of iterations required to converge to the maximum likelihood (MLE) estimates compared to those using the log–likelihood.

The influence of individual observations on the need for transformations is quite crucial to a proper understanding of data and, hence, it is necessary to investigate the data for any spurious or suspicious observations or outliers. Even normal data may fail to exhibit normality due to such observations. These spurious observations may be valid (outliers) or may not be from the same population (contaminants) or may be genuine errors introduced into the sample during the different stages of data collection (e.g. enumeration errors, data capture errors or even in the sampling design). It is, therefore, necessary to perform a thorough statistical check on the data for the existence of outliers. The thesis discusses the problem of detecting and identifying such observations. It proposes an iterative technique for the task based on Mahalanobis distances computed from sub–samples of the observations. The results of the analysis are summarised in a graphical display called the *Stalactite Chart* (or *Plot*) and the analysis is refered to as the *Stalactite Analysis.* The technique is compared to a number of other outlier detection methods.

The transformations process should be carried out in a number of stages. The multivariate data should first be tested for any departures from multivariate normality. If there is evidence of departures, the next stage would be to check for any outlying observations which may be causing them. A decision can be made on any such observations to either discard them and proceed with the transformations or accommodate them and make note of the fact. In an ideal case, both alternatives should be used and comparisons of the results made. The whole process is exploratory and so a dedicated software tool for the task is required. Such a tool is also presented in the thesis. It implements the proposed

techniques together with the related existing methods into a microcomputer–based user–friendly statistical package called *tSTAT* (short for transformation STATistics). The package is not designed to be a comprehensive statistical analysis package but to complement the well–known statistical packages with the specific tasks of carrying out transformations to normality and of outlier detection. The package is written in C.

The structure of the thesis follows the order of outlier detection and then transformations. In Chapter two, techniques for detecting outliers in multivariate data are investigated and compared. These include univariate screening techniques which are used to study the marginal characteristics of the data. The main result of the chapter is the proposed Stalactite analysis. A number of examples are analysed using both simulated data sets and some well known data used for outlier detection techniques in the multivariate literature.

In Chapter three, discussion and presentation of some techniques for assessing the violation of the normality assumption is made. The chapter then investigates computational methods for transformations to multivariate normality. The main result of the chapter is the presentation of the proposed Surcon analysis. Some well known data sets are used as examples to demonstrate the theory. Simulated data sets are also used to study the expected behaviour of the techniques under known predetermined conditions.

The tSTAT software package is discussed and presented in Chapter four. The chapter begins by describing the overall design and structure of the package. It displays some of the main algorithms used together with the source code in the C programming language. The general usage of the package is outlined in the form of a reference manual.

Finally, Chapter five is a summary of conclusions and recommendations.

The rest of this chapter outlines the the main results obtained and described in each of the chapters. The following section is an overview of the role of the multivariate normal distribution in the theory of multivariate analysis and thus provides the justification for the need for transformations.

1.2 The Role of the MVN

Since the beginnings of the theory of multivariate analysis the multivariate normal distribution (MVN), defined in Definition 1.1, has played a central role in the subject. However, the computations were usually very time–consuming, even with a good desktop calculator, and the distributions of the various test statistics were not tabulated. For example, likelihood ratio test statistics were developed for a wide range of hypothesis tests, but they all involve finding $p \times p$ determinants, where p is the dimension of the data, and only asymptotic distributions were known. A great deal of effort was then put in to find good approximations for the distributions of these test statistics and their power functions.

The advent of powerful computers, however, helped the subject to free itself from the multivariate normal straitjacket and multivariate problems could be tackled in a more general manner without being cramped by lack of computational power. Automated procedures have allowed the graphical exploration of data, which is so necessary for good data analysis, to become a practical possibility and have facilitated the calculation of tables of exact percentage points of a number of test statistics. However, good approximations are still important for the automatic presentation of significance levels in a computer printout.

The multivariate normal distribution, however, still has a central role in multivariate analysis since the classical multivariate theory has been largely based on it. Among the reasons for its ascendancy in the multivariate context are the following:

- The MVN is an easy generalisation of its univariate counterpart, and the mutlivariate analysis runs almost parallel to the corresponding analysis based on univariate normality. Generally, the same cannot be said of other multivariate generalisations of univariate distributions.

- The MVN distribution is entirely defined by its first and second order moments – a total of only $\frac{1}{2}p(p+3)$ parameters in all. This compares with 2^p-1 for the multivariate binary or logit distributions [Mardia et.al., 1979; p59]. This economy of parameters simplifies the problems of estimation.

• In the case of normal variables zero correlation implies independence.

• Linear functions of a multivariate normal vector are themselves univariate normal.

• Even when the original data is not MVN one can often appeal to the central limit theorem which proves that certain functions such as the sample mean are normal for large samples.

• The equiprobability contours of the MVN are simple ellipses, which by a suitable change of coordinates can be made into circles (or in general, hyperspheres). Amongst other things, this geometric simplicity allows us to graphically assess departures from the distribution with ease especially in the bivariate case.

DEFINITION 1.1 Multivariate Normal Distribution (MVN)

Let $y = (y_1, y_2,..., y_p)$ be a p–dimensional vector of random variables. Then y is said to have a nonsingular MVN distribution with mean vector μ and covariance matrix Σ (denoted by $N[\mu, \Sigma]$) if its density function is

$$f(y) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp[-\frac{1}{2} (y-\mu)^T \Sigma^{-1} (y-\mu)] \qquad (1.1)$$

where $(-\infty < y_j < \infty, j=1, 2,..., p)$ and $\Sigma > 0$.

□

1.3 The Stalactite Analysis

As mentioned previously, the proposed Stalactite analysis algorithm is presented in Chapter two. The procedure is used for the detection of multiple outliers in multivariate data using Mahalanobis distances and iterative resampling of the data. It involves a sequential construction of an outlier free subset of the data, starting from a small random subset. The Stalactite plot provides a summary of suspected outliers as the subset size increases. A combination of the diagnostic quantities derived and probability plots leads to the identification of multivariate outliers. These outliers are identified even in the presence of appreciable masking where the presence of a clustering of outliers may obscure their outlyingness due to the influence they would have on the estimates of the means and

covariance matrix used in calculating the distances. Usually, the classical approach of using the Mahalanobis distances computed from the full sample cannot detect such outliers. A recent method to overcome this problem by Rousseeuw and van Zomeren [1990] uses distances based on robust estimates of location and covariance. The method is theoretically appealing but the computational requirements are immense and so make it prohibitive as a quick outlier detection procedure on small computers. In comparison, the Stalactite analysis algorithm has the advantage of computational modesty while yielding a simple graphical summary.

1.4 The SURCON Analysis

The proposed SURCON analysis algorithm is presented in Chapter three. The algorithm involves initial normalised Box–Cox transformations on each variable based on some hypothesised values for the transformation parameters, λ_0, which are usually 1 (no transformation for all λ's). A seemingly unrelated regression analysis is then performed on the fit of the constructed variables centered about the means on the transformed variables also centered about their means. The derived transformation parameters, $\hat{\lambda}$ are functions of λ_0 and the regression coefficient estimates of the constructed variables. When λ_0 is taken as the maximum likelihood estimate (MLE) λ_{mle} of λ the fit should not be significant; if it is the process is repeated using an interpolated value between λ_0 and the new $\hat{\lambda}$. The process is, therefore, repeated until all the regression in the model is removed.

The method is compared to the likelihood approach and although the transformation parameter estimates are identical there are significant savings in the number of iterations required to converge to the MLE's. The output produced also includes confidence intervals for the estimates and other useful statistics.

1.5 The tSTAT Package

The tSTAT package described in Chapter five aims at providing a quick tool for implementing the algorithms and theory discussed in the thesis. The package is designed to serve as a quick tool and so it was designed to facilitate maximum ease of use. As Cooper

1984] remarked about a user–friendly statistical package "...its prime aim was not only to 'keep the user away from the operating system' but to meet the claim that if the user knew what he wanted in data analysis terms the system would provide all the help and guidance at the computing level necessary to do it". The tSTAT package was designed in line with this concept of user–friendliness. The user base for statistical systems has changed with the advent of microcomputers and so systems have to be designed to cater for both the 'naive' user and also to remove the burden of preparation of program routines for the expert user such that effort is shifted towards the interpretation of the results.

The main disadvantage of a statistical package is that the user is locked into the system he is using. This means that he is restricted to the options available in the particular package. Nelder [1984] suggests that the user needs the ability to move easily between packages using each for that step of the analysis for which it is best suited e.g. package A output can easily be accepted as input to package B. The tSTAT package does not, in general, have the ability of producing output in a suitable format which can be used as input in another package apart from the transformed data. However, for the purposes it was designed for, exploratory analysis, its output is in "final form" format. In addition, since it produces a logfile, which is in plain ASCII format, the results can easily be imported into many modern wordprocessing packages or edited using any text editor.

1.6 Notation

In general, the notation adopted in the thesis conforms to familiar conventions. Thus, for instance, x, y, x, y,... denote column vectors and $\mathbf{x}^T$,$\mathbf{y}^T$,x^T,y^T,..., row vectors; and X, Y, X, Y,..., matrices. The data matrix is denoted by Y (and X in certain contexts). It is of order n by p where n is the sample size and p the number of variables. A distinction is made between parameters and random variables by using the familiar convention that the former are denoted by Greek letters and the latter by letters of the English alphabet. Most of the concepts and methods discussed are based on observed or sample statistics. A hat ($\hat{}$) is placed over the parameter symbol for its estimate. The rest of the notation is as used in

multivariate statistics literature eg. Σ for the population covariance matrix and S for its estimate.

Equations, figures and tables in the main text are numbered sequentially within a chapter and have a numerical prefix indicating the chapter. However, figures and tables appearing in the examples have a prefix E followed by the example number.

CHAPTER TWO

2.0 MULTIVARIATE OUTLIER DIAGNOSTICS

2.1 Introduction

This chapter is devoted to investigating and comparing techniques for detecting outliers in multivariate data sets. It includes univariate screening techniques which are used to study the marginal characteristics of the data. These provide useful insights in the interpretation of the results obtained from joint screening (multivariate) techniques.

This initial section discusses the general outlier problem by first defining an outlier together with the associated outlier models. The models are presented in two contexts, namely, the single observation formulation and the general formulation. The remainder of the chapter is as follows. In section 2 the techniques and tests for identifying outliers are discussed. The multivariate outlier diagnostics theory depends strongly on the Mahalanobis distances and this section places specific emphasis on them. The main result in the section is the single case deletion formula for Mahalanobis distances. Section 3 contains the graphical procedures that are used. The main multivariate techniques are discussed in the next four sections. The Classical approach to multivariate outlier detection is section 4, the Hat–Matrix approach in section 5 and the Minimum Volume Ellipsoid approach in section 6. Section 7 presents the main result of the chapter which is the proposed iterative technique sampling technique, the Stalactite Analysis approach.

To begin, then, with the general problem of outliers. A recurring difficulty in the creation and maintenance of a large computerised data base is the accuracy of the information entering the base. If high volumes of data are involved, then the data capture tends to be carried out by personnel with varied levels of efficiency and accuracy, and verification may be unsatisfactory in part. Thus, action is required to maintain the base's integrity. The fact that large volumes of machine–readable material are involved suggests that, as far as possible, this screening should be computerised.

The screening process (at the data processing stage) can involve several stages,

including:

A/ Verification – the cross–validation of data being transcribed from the "source documents" to computer magnetic media. This could be carried out by the data being entered twice by two different operators and comparing the two versions.

B/ Validation – testing the data to ensure that all the responses lie within certain pre–specified or known (a priori) limits; i.e. range checks. Examples are $0 <= \text{age} < 120$, sex = 1 or 2, etc.

C/ Editing – this tests for consistency between responses ie. two or more responses should jointly be valid. Examples are if sex = male and number of births = 2 then error.

Stages B and C require some prior knowledge of the data (responses) to be able to set the "rules". However, it is possible to use the data itself to provide the "rules". Here exploratory or initial studies are done to understand the nature of the data, to detect measurement errors, recording errors and "outliers".

The classical work in the field of computerised data screening related to census data is based on the redundancy built into the census return. This enables checks of internal consistency to be carried out and inconsistent records flagged for appropriate action.

On the other hand, data bases consisting of vectors of data following some (known or assumed) distribution can be screened using the distributional properties. In general in this type of data base, no deliberate redundancy is built into the records but rather the extent that the components are statistically related can be used for statistical checking of mutual consistency.

The data in these databases are multivariate in nature and so the problem of detecting and identifying "unusual" observations should strictly be a multivariate one. The term "outliers" can have several different interpretations and one choice of interpretation may render an observation as an outlier (or alternatively as a "good" observation) whereas another interpretation may yield different conclusions. As Kruskal [1960] and Gnandesikan [1977: p.272] have noted, an observation may be an outlier for one purpose but not for

another.

In detecting outliers, therefore, it is necessary to have both an operational definition of an outlier and operational procedures for the identification of such points. Anscombe and Tukey [1963] considered outliers to be "observations that have large residuals, in comparison with most of the others, as to suggest that they ought to be treated specially." To propose and compare outlier procedures, one must know what information is sought for. Two possible aims were mentioned by David [1981: p.218]:

(a) to determine whether outliers are present in the data

(b) to determine those observations that are aberrant

Clearly, if either or both of these are the objectives, the outliers themselves are the primary concern of the analysis. On the other hand, if fitting a model, estimating a set of parameters, or testing a hypothesis is the main interest, outliers are a complication and need to be handled in an appropriate fashion. The aim there is:

(c) to modify a statistical analysis by using information regarding the presence and identity of outliers.

Methods suitable for one of these tasks may or may not be suitable for the others.The focus in this thesis is primarily on aims (a) and (b), that is examination for the presence of outliers and their identification.

In defining and discussing the general outlier problem the theory developed does not depend on any particular distribution, the data are assumed to be a random sample from a multivariate normal distribution. So any observation whose distribution departs from this model is regarded as an outlier. The multivariate normal error structure has been adopted for several reasons, including mathematical tractability, and even more importantly, the fact that many standard multivariate methods are derived under the assumption of normality. This makes it crucial to check for outliers, as well as other types of nonnormality, as their presence will strongly affect inferences made from normal–based procedures. For example, Layard [1974] showed that the normal theory likelihood ratio test

for equality of covariance matrices is highly nonrobust against departures from normality, including contamination.

The definition of outliers adopted in this thesis is as follows:

DEFINITION 2.1: Outliers

Outliers are observations that deviate from the model suggested by the majority of the point cloud, where the central model is the multivariate normal (or at least a unimodal elliptical distribution). #

This definition, therefore, considers a mixture of clean data and arbitrary contaminants. In practice, a further refinement to the definition is required in order to decide what the majority of the point cloud is and, hence, which observations can be considered as outlying.

2.1.1 The Outlier Model

2.1.1.1 Single Observation Formulation

Let the vector $x = (x_1,...,x_p)^T$ represent an arbitrary observation vector distributed as $N(\mu,\Sigma)$. Assume without loss of generality that the x_i have been scaled to zero mean and unit variance. Also, let the input record be $y = (y_1,...,y_p)^T$ where

$$y = x + e \tag{2.1}$$

e being a vector of data capture errors. The screening of y consists, therefore, of a test of the null hypothesis

$$H_0: e = 0 \tag{2.2}$$

One possible alternative to H_0 is H_1: $e \neq 0$, that is, an arbitrary vector is present. In practice, H_1 can be specialised considerably. Suppose that with low probability q a given variable is entered incorrectly. In this case the majority of the data vectors will either be correct or contain a single error located randomly, and primary concern is with the alternative hypothesis

$$H_2: e_j \neq 0 \text{ for some unknown } j$$
$$e_i = 0 \text{ for all } i \neq j$$

and test statistics which are more powerful against H_2 are required. Hawkins [1974] proposes five screening procedures – a "one–at–a–time" test, the standard χ^2 test, and three statistics derived from principal component analysis.

2.1.1.2 General Formulation

Consider a random sample from a multivariate normal distribution. The model for these data can be specified by the matrix

$$Y = e\mu + U \tag{2.3}$$

where the $n \times p$ observation matrix Y has i.i.d rows $Y_1,\ldots,Y_n$, e is an $n \times 1$ vector of 1's, μ is the unknown $1 \times p$ mean vector, and the rows of the $n \times p$ matrix U are i.i.d. $N(0,\Sigma)$ with covariance matrix Σ unknown. It will be assumed that $n \geq p + 1$ to ensure that μ and Σ are estimable.

To reflect the possibility of outliers, the model can be embedded in a *multivariate mean model with mean slippage* [Schwager & Margolin 1982]:

$$Y = e\mu + \Delta^* A^* + U \tag{2.4}$$

Here e, μ, and U are as above, and $n \geq p + 1$. Furthermore, Δ^* is a nonnegative scalar, and A^* is an arbitrary $n \times p$ matrix such that:

(C1) $\|A^*\| = \sqrt{(\sum_{ij} a_{ij}^2)} = 1$, unless $\Delta^* = 0$, in which case $A^* = 0$

(C2) more than half the rows of A^* are zero

In this model, the observation Y_i is an *outlier* if the i–th row of A^* is nonzero. No outliers are present if (iff) $\Delta^* = 0$. Condition (C2) requires that more than half of the observations are drawn from the $N(\mu,\Sigma)$ population. The general outlier problem, therefore consists of:

- Model $Y = e\mu + \Delta^* A^* + U$ (all terms as above)
- Hypothesis $H_0:\Delta^* = 0$ vs $H_1:\Delta^* > 0$
- Action space $\mathcal{A} = \{D_0,D_1\}$, where D_i denotes the decision to act as if hypothesis H_i is true, i=0,1

- State space $\Omega = \{(\Delta^*, A^*, \mu, \Sigma)$: $\Sigma > 0$, $\Delta^* \geq 0$, (C1),(C2) hold$\}$
- Loss function L with $L(\theta, D_i) = i$, if $\Delta^* = 0$ and $L(\theta, D_i) = 1-i$, if $\Delta^* > 0$

Schwager and Margolin [1982] discuss the problem of detecting multivariate normal outliers using mean slippage and demonstrate that the locally best invariant test for outliers is based on Mardia's [1970] multivariate sample kurtosis $b_{2,p}$.

If H_1 is true then D_1 needs to be defined. However, the key question on outliers is *Should one or more observations from a data set be discarded simply because they appear to be "inconsistent" with the rest of the set?* The current thinking on the closely related topic , robust estimation, supports some form of truncation or modification of the data by minimising the influence of such outliers on the fitted model. However, a closer look at an extreme observation is often warranted, as it may shed light on underlying structures or reveal something about the recording of the data and procedures directed specifically at deleting outliers can be useful [Dixon, 1953; Grubbs, 1969]. For example, an observation may deviate sharply from a fitted hypothesised model, because the model breaks down at that particular point and not because the observation is spurious. It also happens not infrequently that only part of the data obeys a different model. A single outlier which is sufficiently far away can ruin, for example, a least squares analysis completely; some sources for gross errors such as keypunch errors or wrong decimal points do indeed easily change values of order of magnitude; and with the modern trend of entering masses of data unscreened into the computer, outliers can easily escape attention if no precautions are taken. So wrong measurements or wrongly recorded data in either the experimentation or computational stage are much more common than is generally recognised. Hampel, et al. [1986:, pp.25–28] cite numerous examples for the frequency of gross errors and other outliers in real data.

Some sources of gross errors are:

- copying errors
- interchange of two values or groups of values in a structured design

- inadvertent observation of a member of a different population
- equipment failure
- transient effects

With sufficient care, gross errors can often be prevented but the necessary care cannot always be afforded. Moreover, with fully automatic data recording and properly working equipment, there may be large transient effects. An example [Hampel, et al. 1986: p.26] is about a quarter of a million electroencephalographic data, fully automatically recorded by properly working equipment; the histogram looked normal except for some jittering of the plotter way out in the tails, but the third and fourth moments were far too large. A search revealed that there was a huge spike of about two dozen data points when the equipment was switched on; these few transient points caused the high moments and jitter in the plot. Distant gross errors are one of the most dangerous deviations from the usual statistical assumptions; but they are also the ones which can most easily be treated. The number of distant gross errors which statisticians get to see is frequently decreased considerably below the original one because the subject matter specialists often "clean" their data in some informal way before consulting the statistician. Even so, the frequency of gross errors varies considerably. Crudely speaking, one has to distinguish between high–quality data with no (or virtually no) gross errors, and routine data, with about 1–10% or more gross errors. Whenever a distant outlier occurs, some robust method is required, and be it just a subjective look at the data with subsequent special treatment of the outlier.

Another approach to dealing with outliers is to construct outlier diagnostics. These are quantities computed from the data with the purpose of pinpointing influential observations, which can then be studied and corrected or deleted, followed by an analysis on the remaining observations. When there is only a single outlier it may be possible to use methods effectively by looking at the effect of deleting one point at a time. For example, denote by $\hat{\theta}(i)$ the estimate for a parameter θ computed from the sample without the i–th

case. Then the difference between $\hat{\theta}$ (the full sample estimate) and $\hat{\theta}$(i) gives the extent to which the presence of the k–th case affects the corresponding estimates. These are so–called *single case diagnostics*, which are computed for each case i.

Unfortunately, it is much more difficult to diagnose outliers when there are several of them. It is, of course, possible to generalise the single–case diagnostics for highlighting the simultaneous influence of several cases. However, it may not be obvious at all which cases should be deleted. It may happen that some points are jointly influential but the individual points are not! Moreover, the computations involved are often infeasible because of the large number of subsets that would be considered. For example if consideration is given to deletion of 4 out of 30 cases, there are 27405 possibilities. In some examples the sequential employment of single deletion methods leads to the detection of important sets of observations. In others, the importance of the observations is not evident unless several observations are deleted at once. In the case of multiple regression analysis, a two stage method for the detection of outliers and influential observations when masking is present can be adopted [Atkinson 1986]. The first, exploratory, stage uses least median of squares regression, a method which resists nearly 50% of contamination in the data [Rousseeuw, 1984; Hampel et al., 1986: p.330]. In the second, confirmatory, stage the diagnostic methods of least–squares regression are used to confirm the findings of the robust method.

Robust methods and diagnostic methods have the same goal but proceed in an opposite order: the robust approach first fits a model that does justice to the majority of the data and then discovers the outliers as those points which have large residuals from the robust solution, whereas in the diagnostic setting, one first wants to identify the outliers and then fit the good data in the classical way.

To aid the search for outliers, graphical techniques can be employed to provide possible candidates for further investigation, followed by suitable tests of "discordancy" of these observations with the rest of the data. With univariate data the observations are readily ranked so that the largest and/or smallest observations come up for scrutiny. As

mentioned earlier, when there are several possible outliers, we can test the extreme observations one at a time or as a group. Testing one at a time, however, may suffer from a "masking effect" in which the less extreme outliers mask the discordancy of the most extreme observations under investigation.

When dealing with multivariate data, the situation is even more complicated. First, for more than two dimensions, there is the problem of representing the data graphically so as to highlight the presence of any outliers. Second, there is the problem of ordering the data so that we can isolate the extremes of observations that separate themselves from the bulk of the data [See Barnet 1976]. Third, a multivariate outlier can distort the measures of orientation (correlation) as well as the measures of location and scale.

Consider the bivariate data in Figure 2.1: observation *A* will inflate both variances, but will have little effect on the correlation; observation *B* will reduce the correlation and inflate the variance of Y_1 but will have little effect on the variance of Y_2; and observation *C* has little effect on the variances but reduces the correlation. From another viewpoint, *B* and *C* add what could be considered as an insignificant second dimension to data that are essentially one dimensional, lying on a straight line. This could have serious implications in especially dimensional reduction techniques such as principal components. Suppose C was to appear very far away from the majority point cloud in the Y_1 space; it would lead to large variability in that direction and thus "flip" the principal components resulting into that direction becoming erroneously the first principal component.

To illustrate the last point we shall briefly state the theory of principal components and discuss a simple bivariate example to show the effect of outliers on principal components transformations.

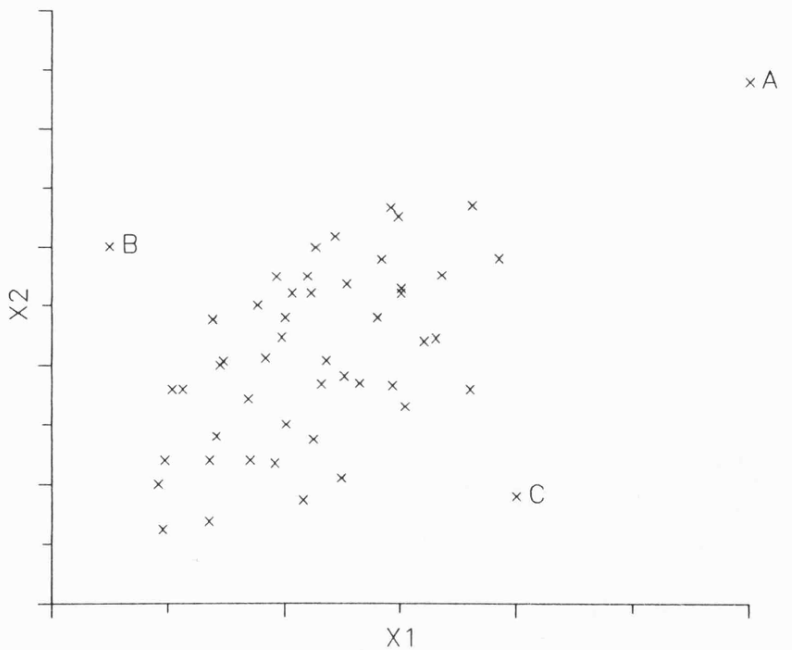
Let $Y^T = [y_1, y_2, \ldots, y_p]$ be a *p*–variate sample of size *n* with covariance matrix Σ. If Σ has eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ and considering the linear combinations

$$u_j = l_j^T Y = l_{1j}^T y_1 + l_{2j}^T y_2 + \ldots + l_{pj}^T y_p \qquad (2.5)$$

$j=1,2,\ldots,p$ with

$$\mathrm{Var}(u_j) = l_j^T \Sigma l_j \qquad (2.6)$$

Figure 2.1 Types of Outliers in 2--Dimensions



$$\text{Cov}(u_j,u_k) = l_j^T \Sigma l_k \tag{2.7}$$

j,k = 1,2,...,p, then the principal components are those "uncorrelated" (orthogonal) linear combinations whose variances are as large as possible. So the j–th principal component is that linear combination which provides a solution to the following constrained optimisation problem

$$\begin{aligned} &\text{Maximise } \text{var}(u_j) = l_j^T \Sigma l_j \\ &\text{Subject to } l_j^T l_j = 1 \\ &\text{Cov}(l_j^T Y, l_k^T Y) = 0 \qquad \text{for } k < j. \end{aligned} \tag{2.8}$$

In particular, if Σ is the covariance matrix associated with the random vector $Y^T = [y_1, y_2,..., y_p]$ having the eigenvalue–eigenvector pairs (λ_1,e_1), (λ_2,e_2),...,(λ_p,e_p) where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ the j–th principal component is given by

$$u_j = e_j^T Y = e_{1j}^T y_1 + e_{2j}^T y_2 + ... + e_{pj}^T y_p \tag{2.9}$$

and

$$\begin{aligned} \text{Var}(u_j) &= e_j^T \Sigma e_j = \lambda_j \quad ,j = 1,2,...,p \\ \text{Cov}(u_j,u_k) &= e_j^T \Sigma e_k = 0 \quad ,j \neq k \end{aligned}$$

It is, therefore, evident that principal components depend entirely on the covariance matrix Σ and so are very sensitive to any observations which may affect variances and correlations.

EXAMPLE 2.1 (Effect of outlier on principal components)

Let $Y^T = [y_1, y_2]$ be from a bivariate distribution with covariance matrix

$$\Sigma = \begin{bmatrix} 100 & 4 \\ 4 & 1 \end{bmatrix}.$$

Assume a scatter plot of the observations with an extreme observation P, say, which is greatly displaced from the majority of the data in the y_1 space but well within range in the y_2 space. Its consequence is to inflate the y_1 variance and also reduce the correlation between the two variables.

In carrying out a principal components transformation the eigenvalue–eigenvector pairs from Σ are

$$\lambda_1 = 100.16 \qquad e_1^T = [0.999, -0.040]$$

$$\lambda_2 = \ 0.84 \qquad e_2^T = [0.040, \ 0.999]$$

So the respective principal components become

$$u_1 = 0.999y_1 - 0.040y_2$$

$$u_2 = 0.040y_1 + 0.999y_2$$

We note that because of its large variance, y_1 completely dominates the first principal component determined from Σ. Moreover, this first principal component explains a proportion

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{100.16}{101.00} = 0.992 \qquad (2.10)$$

of the total population. This would erroneously suggest that the principal components are as displayed in Figure 2.2(a). Figure 2.2(b) displays a scatter plot of the data without observation P. A reverse solution for the principal components transformation is indicated. □

A fourth problem with multivariate outliers is that an outlier can arise because of either a gross error in one of its components or small systematic errors in several components or even from a few observations coming from a completely different distribution from the rest of the data. This is the swamping phenomenon or "Masking". The situation is complex and, as emphasised by Gnanadesikan and Kettenring [1972: p.109], there is no point in looking for an omnibus outlier protection procedure: Rather an arsenal of methods designed for specific purposes is required.

This thesis presents a proposed technique for use in detecting these outliers based on iterative resampling of the data. The technique is compared with some classical identification methods, and the Minimum volume ellipsoid method [Rousseeuw and van Zomeren 1990] in terms of both the capacity for detecting the outliers and for computational simplicity. Some diagnostic quantities are also proposed together with a graphical display which summarises the results. The technique is refered to as the *Stalactite Analysis* a term based on the nature of the graphical summary display which resembles

geological stalactites.

2.2 Identification of Outliers

2.2.1 Discordancy Tests

As a first step in detecting outliers, the p univariate marginal distributions can be inspected and the univariate techniques applied to them [See Barnett and Lewis 1978]. The coefficient of kurtosis, b_2, is a useful statistic for detecting outliers among normally distributed data. If b_2 is significant, the most extreme observation is removed and b_2 retested on the remaining observations. A major weakness of this one–dimensional approach is that outliers like C in Figure 2.1, which mainly affect correlation, may not be detected.

In the case of bivariate data, the sample correlation r is very sensitive to outliers and can therefore be used for their detection. For example, when the data are bivariate normal, Gnanadesikan and Kettenring [1972] suggest a normal probability plot of the

$$Z(r_{-i}) = \frac{1}{2}\log\left[\frac{1 + r_{-i}}{1 - r_{-i}}\right] \qquad (2.11)$$

where r_{-i} is the sample correlation between two variables based on the data with X_i omitted.

Devlin et al. [1975] use the concept of the *Influence Curve* [Hampel 1974] and present two graphical methods based on the sample *Influence Function* of r. One of the methods leads to the function $(n-1)[Z(r) - Z(r_{-i})]$ which, for a reasonably large sample of normal bivariate data, is approximately distributed as a product of two independent N(0,1) variables. they propose a further normalising transformation prior to probability plotting.

For higher–dimensional data we can examine the $\frac{1}{2}p(p-1)$ scatter plots for all the bivariate marginals using scatter plot matrices, say, provided that p is not too large, and

then apply the bivariate methods for detecting outliers. However, although marginal distributions will detect a gross error in one of the variables, they may not show up an outlier with small systematic errors in all its variables. Also, there is the problem of outliers like B and C in Figure 2.1, which add spurious dimensions.

Apart from B and C the data lie almost in a one–dimensional space, the regression line. Going up a dimension we envisage a set of points that all lie elose to a plane except for one or two outliers at some distance from the plane. Generally, such outliers can be uncovered by working with robust principal components of the observations instead of the observations. One procedure is given by Campbell, [1980] and is based on using the robust estimation of the covariance matrix. Gnanandesikan and Kettenring [1972: p.111] note that the first few principal components are sensitive to outliers that inflate variances and covariances (if working with $\hat{\Sigma}$ or S) or correlations (if working with the sample correlation matrix R),and the last few are sensitive to outliers which add spurious dimensions.

One advantage of using principal components is that they are likely to be more normal than the original data. Particularly when p is not so large, approximate normality being achieved by the central limit theorem argument being applied to linear combinations. The multivariate kurtosis $b_{2,p}$ can also be used as an outlier test [Schwager and Margolin 1982].

Plotting techniques used to assess multivariate normality can be used for detecting outliers. Gamma values can be "normalised" using Fisher's transformation $y = \sqrt{x}$ or Wilson and Hilferty's transformation $y = \sqrt[3]{x}$. In particular, for a given scale parameter λ and shape parameter η (not too small)

$$y = x^{1/2} \rightarrow N_1\left[\left[\lambda(\eta - \tfrac{1}{4})\right]^{1/2}, \tfrac{1}{4}\lambda\right] \tag{2.12}$$

$$y = x^{1/3} \rightarrow N_1\left[\left[\lambda\eta\right]^{1/3}\left[1 - (9\eta)^{-1}\right], \tfrac{1}{9}\left[\frac{\lambda^2}{\eta}\right]^{1/3}\right] \tag{2.13}$$

With either transformation we can apply to y a discordancy test for a sample of size

n from a normal distribution with unknown mean and variance. Barnett and Lewis [1978: pp.91–92] list 17 such tests.

The quantity

$$d^2 = (x-\mu)^T \Sigma (x-\mu) \rightarrow \chi^2(p) \tag{2.14}$$

if $x \rightarrow N_p[\mu, \Sigma]$. Hence, we can carry out one of the gamma normalising transformations on d^2 and apply these discordancy tests. Three such tests are:

1/ $$T = \max\left\{ \frac{\bar{y} - y[1]}{s}, \frac{y[n] - \bar{y}}{s} \right\} \tag{2.15}$$

where $s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)$ and y[i] is the i–th ordered observation.

2/ $$b_2 = n \sum_{i=1}^{n} (y_i - \bar{y})^4 / \{ \sum_{i=1}^{n} (y_i - \bar{y})^2 \}^2 \tag{2.16}$$

and by Kimber [1979]

3/ $$Z = \max_{1 \leq i \leq n} \left\{ \gamma_i / \sum_{g=1}^{n} \gamma_g \right\} \tag{2.17}$$

where $\gamma_i = -\log u_i - (n-1)\log\left\{ \frac{n - u_i}{n - 1} \right\}$, $u_i = y_i/\bar{y}$ and the y_i have a gamma distribution.

Large values of T, b_2 and Z signify the presence of a very large or a very small observation in the sample. Significance points for b_2 are given in D'Agostino and Tietjen [1971: Table 1] and D'Agostino and Pearson [1973:Figures 1 and 2]. The critical values for Z can be found in the table for a 5% and 1% discordancy test for a single gamma outlier, Kimber [1979: Table 1, n=5(1)20] and Barnett and Lewis [1978: Table I, r=0.5, n>20].

Also, a test for a single outlier in a sample of size n from $N_p[\mu, \Sigma]$, (μ, Σ unknown) can be constructed using the following test statistic

$$\begin{aligned} d^2(n) &= \max_{1 \leq i \leq n} (y_i - \bar{y})^T S^{-1} (y_i - \bar{y}) \\ &= \max_{1 \leq i \leq n} d_i^2 \end{aligned} \tag{2.18}$$

where $S = \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^T/(n-1)$. The critical values can be obtained from Barnett

TABLE 2.1 Discordancy Tests for a Single Outlier

DATA	n	p	Discordancy Test† T	b_2	Z	d^2_{max}
1. Bivariate Normal data (without outliers)	50	2	2.601 (1.960)	3.091 (3.990)	0.131 (0.206)	4.912 (12.230)
2. Bivariate Normal data (with 4 outliers)	50	2	2.946 (1.960)	4.124 (3.990)	0.187 (0.206)	16.542 (12.230)
3. Belgian Phone Calls data	24	2	1.819 (1.960)	1.785 (4.168)	0.136 (0.343)	6.497 (9.780)
4. Hertzsprung-Russell Star data	47	2	2.196 (1.960)	2.791 (3.996)	0.091 (0.215)	10.970 (12.020)
5. Hawkins-Bradu-Kass Artificial data	75	3	4.343 (1.960)	6.387 (3.860)	0.348 (0.156)	41.280 (15.315)
6. Repeat Soil Sample Survey data						
- X1, X2	57	2	2.088 (1.960)	2.610 (3.945)	0.081 (0.184)	8.839 (12.510)
- X1, X2, X3	57	2	3.586 (1.960)	5.272 (3.945)	0.243 (0.184)	38.736 (12.510)
- X1, X2, X3, X4, X5	57	5	3.142 (1.960)	4.022 (3.945)	0.201 (0.184)	38.802 (14.495)

† The terms in brackets indicate the critical values of the discordancy tests at the 5% level of significance.

Notes:

1/ T is tested against the standard normal, $N[0,1]$.

2/ b_2 is tested against the critical values from D'Agostino and Tietjen [1971: Table 1] for the sample coefficient of kurtosis from a random sample of size n from a normal distribution.

3/ Z is tested against the critical values from Kimber [1979: Table 1, n=5(1)20] and Barnett and Lewis [1978: Table I, r=0.5, n>20] for 5% discordancy tests of a single outlier in a sample of size n from a gamma distribution.

4/ d^2max is tested against the critical values from Barnett and Lewis [1978: Table XXVIII] for 5% discordancy tests of a single outlier in a sample of size n from $Np[\mu, \Sigma]$; μ and Σ are unknown.

and Lewis [1978: Table XXVIII] for 5% and 1% tests.

An outlier tends to inflate $\bar{x}$ and S, and possible reduce d_i^2, so that it may strictly warrant the use of the robust version of d_i^2, namely

$$d_i^{2*} = (y - \mu^*)^T (S^*)^{-1} \Sigma (y - \mu^*) \tag{2.19}$$

where μ^* and S^* are the robust estimators of μ and Σ, [Campbell 1980].

2.2.2 Mahalanobis Distance

The approaches to outlier detection depend on determining the point cloud and then identifying those observations which seem to lie to remotely from the majority of the data. It is for this reason that the quantity D_i^2 is adopted in many of the tests since it looks at the remoteness of the observations from the center of the point cloud together with taking into account its shape. This positive square root of this quantity is called the Mahalanobis distance.

DEFINITION: Sample Mahalanobis Distance

Consider a data set with n observations measured on p variables with data matrix Y_{pxn}, then the Sample Mahalanobis distance for the i–th observation is given by

$$d_i = \left[(y_i - \bar{y})^T S^{-1} (y_i - \bar{y})\right]^{1/2} \tag{2.20}$$

where i = 1,2,...,n and

y_i - i–th observation vector ie. y_i is the i–th column of Y

$\bar{y}$ – sample mean vector

S – sample var–covariance matrix #

If we define d as an n×1 vector with elements d_i, i=1,2,...,n then

$$D = d^T d = (Y - \overline{Y})^T S^{-1} (Y - \overline{Y}) \tag{2.21}$$

where Y is the data matrix, $\overline{Y} = 1.y^T/n$ and 1 is an n×1 vector with all the elements equal to unity

Also, let $V = Y - \overline{Y}$

then $D = VS^{-1}V^T$

but $$S = \frac{1}{n-1} V^T V$$

thus, $$D = (n-1)V(V^TV)^{-1}V^T \tag{2.22}$$

The i–th diagonal element of D is given by

$$d_{ii} = d_i^2 = (n-1)v_i^T(V^TV)^{-1}v_i \tag{2.23}$$

and is the square of the Mahalanobis distance from the i–th observation which could be regarded as a "standardised" measure of remoteness of the i–th observation from the centre of gravity of the data set. An observation with a large d_i^2 value is atypical and should be examined further.

The minimum value of d_i^2 is 0 and occurs when $y_i = \bar{y}$. Since d_i^2 behaves like a χ_p^2 if the parent population is normal the maximum value of d_i^2 can be compared to the expected maximum χ_p^2. This value could be used as a test value to indicate atypical or influential observations. Thus, values of d_i^2 such that

$$d_i^2 > E\left[\text{Max } \chi_p^2 \right] \tag{2.24}$$

can be considered atypical and, hence,require further examination.

A Chi–plot can be constructed in which the pairs $\{d_{[i]}^2, \chi_p^2([i - \frac{1}{2}]/n)\}$, i=1,2,...,n are graphed where $d_{[i]}^2$ is the i–th ordered squared distance and $\chi_p^2([i - 1/2]/n)$ is the $100(i - \frac{1}{2})/n$ percentile of the Chi–square distribution with p degrees of freedom. It therefore, follows that the expected maximum χ_p^2 is given when i=n. Thus,

$$E\left[\text{Max } \chi_p^2\right] = \chi_p^2\left[\frac{n-1/2}{n}\right] \tag{2.25}$$

Another value of interest is the "total" distance within the data set together with the average distance.

Now,
$$\begin{aligned} tr(D) &= (n-1)tr\{ V(V^TV)^{-1}V^T \} \\ &= (n-1)tr\{ (V^TV)^{-1}VV^T \} \\ &= (n-1)tr(I_p) \\ &= (n-1)p \end{aligned} \tag{2.26}$$

it follows that $$\Sigma d_i^2 = tr(D) = (n-1)p$$

and $$\bar{d}^2 = \frac{(n-1)}{n}p \qquad (2.27)$$

2.2.2.1 Case Deletion

As mentioned earlier, it may be of interest to perform single case diagnostics, that is to study the behaviour of the of the quantities computed from the data when one or more observations (cases) are deleted. The following is a derivation of a computational procedure for the Mahalanobis distances when one observation is deleted from the sample.

Let the i–th squared Mahalanobis distance with the k–th observation deleted be $d_i^2(k)$, then

$$d_i^2(k) = (y_i - \bar{y}(k))^T S^{-1}(k)(y_i - \bar{y}(k)) \qquad (2.28)$$

but

$$\begin{aligned} y_i - \bar{y}(k) &= x_i - \frac{1}{n-1}(n\bar{y} - y_k) \\ &= \frac{1}{n-1}\{n(y_i - \bar{y}) - (y_i - y_k)\} \end{aligned} \qquad (2.29)$$

Further, excessive computation for the inverse of S(k) can be avoided by using the matrix identity [Bartlett 1951]

$$S^{-1}(k) = aS^{-1} + \frac{abS^{-1}(y_k - \bar{y})(y_k - \bar{y})^T S^{-1}}{\{\, 1 - b(y_k - \bar{y})^T S^{-1}(y_k - \bar{y}) \,\}} \qquad (2.30)$$

where $a = (n-2)/(n-1)$ and $b = n/(n-1)^2$.

So each $d_i^2(k)$ can be obtained by using just $\bar{x}$, S^{-1} and y_k.

LEMMA:

Let $d^2(k)$ denote the squared Mahalanobis distance of the k–th observation with observation k deleted from a sample of size n when computing the sample mean and covariance matrix, then

$$d^2(k) = \frac{(n-2)n^2}{(n-1)^3}\left[\frac{d^2_k}{1 - \frac{nd^2_k}{(n-1)^2}}\right] \qquad (2.31)$$

#

PROOF:

Let $y = (y_1,\ldots,y_p)$ be a sample of size n then,

$$s_{ii} = \sum_k (y_{ik} - \bar{y}_{i\cdot})^2/(n-1) \tag{2.32}$$

$$s_{ij} = \sum_k (y_{ik} - \bar{y}_{i\cdot})(y_{jk} - \bar{y}_{j\cdot})/(n-1) \tag{2.33}$$

and

$$(n-2)s_{ij}(k) = (n-1)s_{ij} - r_{ik}r_{ik}/(1-h_k) \tag{2.34}$$

where $h_k = 1/n$

ie.
$$t_{ij}(k) = t_{ij}(k) - r_{ik}r_{ik}/(1-h_k) \tag{2.35}$$

or
$$T(k) = T - r_k r_k^T/a_k$$
$$= T - \tilde{r}_k \tilde{r}_k^T \tag{2.36}$$

So, the squared Mahalanobis distance for the k–th observation is

$$d_k^2 = (n-1)r_k^T T^{-1} r_k \tag{2.37}$$

and
$$d^2(k) = (n-2)r^T(k)T^{-1}(k)r(k) \tag{2.38}$$

The corresponding residuals are

$$r_k = y_k - \hat{\beta} \tag{2.39}$$

and
$$r(k) = y_k - \hat{\beta}(k)$$
$$= y_k - \hat{\beta} + \hat{\beta} - \hat{\beta}(k)$$
$$= r_k + \hat{\beta} - \hat{\beta}(k) \tag{2.40}$$

But
$$\hat{\beta} - \hat{\beta}(k) = (X^T X)^{-1} x r_k/(1-h_k)$$
$$= \frac{1}{n} r_k \frac{1}{1 - 1/n} = \frac{r_k}{n-1} \tag{2.41}$$

So
$$r(k) = r_k/(1-h_k) = r_k/a_k \tag{2.42}$$

Also
$$T^{-1}(k) = T^{-1} + T^{-1}\tilde{r}_k(I - \tilde{r}_k T^{-1}\tilde{r}_k^T)^{-1}\tilde{r}_k T^{-1} \tag{2.43}$$

where $\tilde{r}_k = r_k/\sqrt{a_k}$.

So
$$\tilde{r}_k T^{-1}\tilde{r}_k^T = \tilde{r}_k T^{-1}\tilde{r}_k^T/a_k = q/a \tag{2.44}$$

From (2.43)

$$T^{-1}(k) = T^{-1} + T^{-1}\tilde{r}_k\tilde{r}_k T^{-1}/c \tag{2.45}$$

where $c = 1-q/a$.

So
$$d^2(k)/(n-2) = r^T(k)T^{-1}(k)r(k)$$
$$= \frac{1}{a_k^2} r_k{}^T T^{-1}(k) r_k$$

and
$$a_k^2 d^2(k)/(n-2) = r^T{}_k T^{-1}(k) r_k \qquad (2.46)$$

Then using (2.45)
$$= r_k^T T^{-1} r_k + r_k^T T^{-1} r_k r_k^T T^{-1} r_k / a_k c \qquad (2.47)$$
$$= r_k^T T^{-1} r_k r_k^T + \frac{r_k^T T^{-1} r_k \; r_k^T T^{-1} r_k}{a(1-q/a)}$$
$$= q + q^2/a(1-q/a)2$$
$$= q/(1-q/a) \qquad (2.48)$$

but
$$q = r_k^T T^{-1} r_k = d_k^2/(n-1)$$

so
$$q/(1-q/a) = \frac{d_k^2/(n-1)}{1 - \frac{d_k^2}{(n-1)} \cdot \frac{n}{(n-1)}}$$
$$= \frac{1}{(n-1)} \left[\frac{d_k^2}{1 - \frac{n d_k^2}{(n-1)^2}} \right]$$

and hence,
$$d^2(k) = \frac{(n-2)n^2}{(n-1)^3} \left[\frac{d^2{}_k}{1 - \frac{n d^2{}_k}{(n-1)^2}} \right] \qquad (2.49)$$

□

2.2.2.2 Distribution of Malanobis Distance

If $y \longrightarrow N_p[\mu, \Sigma]$ then it can easily be shown that

$$d^2 = (y-\mu)^T \Sigma (y-\mu) \xrightarrow{\text{indept}} \chi^2(p) \qquad (2.50)$$

However, if μ and Σ are unknown and have to be estimated by $\bar{x}$ and S respectively the following complications arise:

1/ The d_i^2 are no longer independent since each d_i^2 involves values of $\bar{y}$ and S.

2/ Since $\bar{y}$ and S are only estimates it implies that d_i^2 are no longer χ^2 under the null hypothesis of normality.

Fortunately, if n is large ($\geq$ 10p) the difference between using the true distribution and the χ^2 approximation is negligible, and for samples of that size the departure from

independence of the d_i^2's is also insignificant.

For small samples, the following modifications are required. If $\bar{y}$ and S are from a sample of size n from a $N_p[\mu, \Sigma]$ population and y is a further independent observation from this population then

$$(y - \bar{y})^T S^{-1} (y - \bar{y}) \longrightarrow \frac{p(n^2 - 1)}{n(n - p)} F_{p,n-p} \qquad (2.51)$$

Hence, these F quantities must be used in place of χ^2. Additionally, independence of each y_i from $\bar{y}$ and S must be ensured. To do this "Jack–knifed" means and covariance matrices can be used instead of the single mean vector $\bar{y}$ and covariance matrix S. The "Jack–knifed" mean $\bar{y}(i)$ and covariance matrix S(i) for use with observation y_i are simply the mean vector and covariance matrix, respectively, of the (n–1) observations excluding observation y_i. (See Section 2.2.2.1 for computational formulae).

2.3 Graphical Techniques

Plots are important aids in all aspects of data analysis because they provide a visual perception of the data from which its structure (or non–structure) can be quickly assessed. Although it is not possible to plot simultaneously all the measurements made on several variables and study the configurations, plots of individual variables and plots of pairs of variables can still be informative. The advancement in computer hardware technology has made it possible to have sophisticated graphics hardware at very reasonable costs. This has led to a parallel development of complex and agile computer graphics software capable of examining data in one, two or three dimensions with relative ease. There are, therefore, numerous elegant and effective methods for displaying data [See Tukey 1977]. In the quest for detecting outliers some of the more common methods have been adopted in the thesis. A new plot which summarises the results from the proposed outlier detection technique, Stalactite Analysis, is also presented.

2.3.1 Univariate and Bivariate Plots

– Scatter Plots

It is often important to plot pairs of variables and to visually inspect the pattern of association. The Scatter Plot is used for this purpose and is a plot of n points in two dimensions with each axis representing a variable ie. the coordinates being determined by the paired measurements $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$.

Any unusual points from the majority point cloud can, therefore, be easily seen from the scatter plot although their identity may not be ascertained. This technique is useful in detecting all the three categories of outliers (those that affect variation and not correlation, variation and correlation, and correlation but not variation) in two dimensions

– Box Plots

Apart from assessing association between variables, it is useful to visually study the locality, spread and skewness of a data set. A Box Plot provides such a plot. It is composed of a box with lines protruding from either side (whiskers). The box indicates the median as well as the first quartile (q_1) and the third quartile (q_2).

For a completely symmetrical distribution the box should be divided into equal halves and the whiskers should be of equal length. Departures from this norm indicate levels of skewness and spread. An intrinsic feature of the box plot is the ability to indicate outliers (which can either be exceptionally large or small observations). Different criteria according to which an outlier can be identified are available. A useful criterion is that of Tukey [1977] where observations larger than $q_3 + t$ or smaller than $q_1 - t$, with $t = 1.5(q_3 - q_1)$, are regarded as outliers.

The box plot furthermore offers a useful way of comparing two or more data sets observed in the same units with each other, with regard to locality, spread and symmetry.

2.3.2 Multivariate Plots

For more than two dimensions, there is the problem of representing the data graphically so as to highlight the presence of any outliers. However, the data can be

"collapsed" into a one–dimensional statistic which still retains as much of the multivariate information. The Mahalanobis distance provides such a statistic, it can, therefore, be used in place of the original observations. It is, thus, possible to exhibit the multivariate data in terms of univariate (or bivariate) plots using these distances.

There are several possible plots which could be used, but for the purposes of this thesis only three are considered in addition to the proposed displays. These have been selected to display the association between variables, to provide a form of identification to the suspicious observations and to also portray any swamping phenomena, masking, if any. As mentioned in Section 2.3.1 the scatter plot in two dimensions can show any discrepant observations but cannot identify them, the Index Plot provides a method of identifying the particular observations under suspicion.

2.3.2.1 Scatter Plot Matrix

In a multiresponse data set where p variables are recorded on n observations, scatter plots can be made for all possible pairs of variables, provided p is not too large. The Scatter Plot Matrix is a generalisation of the scatter plot described above (Section 2.3.1). This is a plot of n points in p dimensions.

Consider the extension of the scatter plot where the p measurements $[y_{1i},y_{2i},\cdots,y_{pi}]^T$ on the i–th observation represent the coordinates of a point in p–dimensional space. The coordinate axes are taken to correspond to the variables, so that the i–th point is y_{1i} units along the first axis, y_{2i} units along the second,..., y_{pi} units along the p–th axis. The resulting plot with n points will in fact not only exhibit the overall pattern of the variability, but will show similarities (and differences) among the n observations. The plot appears in a form of grid and, hence, the name Scatter Plot matrix. (It is also possible to plot the p–points in n–dimensional space).

2.3.2.2. Index Plot

DEFINITION 2.3: Index Plot

An "Index Plot" is a plot of some measurement on the observations against their

observation number. #

In this section, the Mahalanobis Index plot (MIP) is considered ie. the Mahalanobis distance $d_i = \sqrt{d_i^2}$ for each observation plotted against the observation number. This plot provides a quick visual display of the general pattern of the d_i's, so extreme values of d_i^2 are easily spotted and identified in the data set.

2.3.2.3 Probability Plots

The probability plots are empirical cumulative distribution functions (ecdf) which may be defined as plots of the i–th ordered observation against $(i - 1/2)/n$ [Gnanadesikan 1977: p.198]. Wilk and Gnanadesikan [1968] describe two basic types of probability plots, called the *P–P* and *Q–Q* plots, respectively. A plot of points whose coordinates are the cumulative probabilities $\{p_x(q), p_y(q)\}$ for different values of q is a *P–P* plot, while a plot of the points whose coordinates are the quantities $\{q_x(p), q_y(p)\}$ for different values of p is a *Q–Q* plot. A usual form of comparison is one in which an ecdf for a body of univariate data, for x say, is compared to a specified (or theoretical) distribution function, for y say.

In particular, if an ecdf of an unstructured sample, $y_1, y_2, \ldots, y_n$, of size n is to be compared with a hypothesised standardised distribution $F(y;\theta)$ (where the parameters θ have specified values): if $y(1) \leq y(2) \leq \ldots \leq y(n)$ are the ordered observations, then a plot of the n points $\{y(i), \tilde{y}_i\}$, $i = 1,2,\ldots,n$, where $\tilde{y}_i$ is the quantile of the distribution F corresponding to a cumulative probability $p_i = (i - a)/(n - 2a + 1)$ with $a = 1/2, 1/3$, or 0 as some of the choices. So $\tilde{y}_i$ is defined by $F(\tilde{y}_i;\theta) = p_i$.

For the purposes of applying the outlier tests using the Mahalanobis distances two probability plots are used based on the distributional theory of the d_i^2 ,namely, the Chi–square plot and the normal plot with the d_i^2 transformed accordingly using the Wilson and Hilferty or Fisher's transformations. In the normal plot F is taken as the distribution function Φ the standard normal distribution.

These plots should exhibit linearity if the two distributions under comparison are not different. It, therefore, follows that the plots can be used to indicate the presence of

points (if any) which depart from this otherwise linear plot.

– Chi–square Probability plot (Chi–plot)

If $d^2_{[1]},d^2_{[2]},\ldots,d^2_{[n]}$ are the ordered squared Mahalanobis distances the Chi–plot is obtained by plotting $d^2_{[i]}$ vs $\chi^2_p[(i-1/2)/n]$ where χ^2_p $[(i-1/2)/n]$ is the $100(i - \frac{1}{2})/n$ percentile of the Chi–square distribution with p degrees of freedom.

– Normal Probability plot (normal plot)

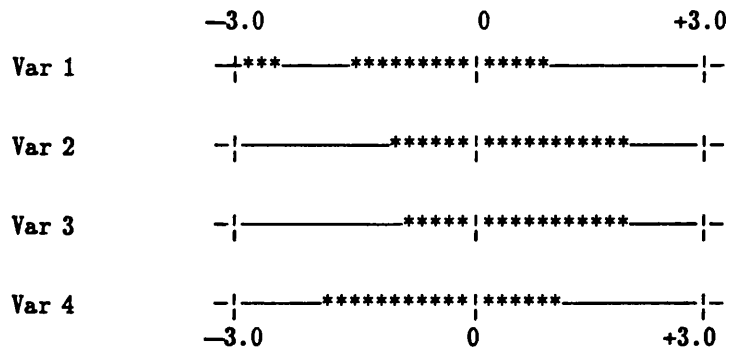
If $d_{[1]},d_{[2]},\ldots,d_{[n]}$ are the ordered Mahalanobis distances the normal plot is obtained by plotting $d_{[i]}$ vs $\Phi[(i-1/2)/n]$ where $\Phi[(i-1/2)/n]$ is the $100(i - \frac{1}{2})/n$ percentile of the standard normal distribution.

2.3.3 Parallel Coordinate Plots (Z–Curves)

If the data are considered as p marginal (univariate) variables then "dot diagrams" could be constructed for each of the variables. Links between observations across the p dot diagrams could be made and these would present some insight into the general spread of the data across variables. An extreme variable would tend to cause the rest of the data to appear to cluster.

Due to different units of measurement across variables, it is necessary to first standardise observations before constructing the plots. The standardising method should not be sensitive to extreme observations.

Figure 2.2 Parallel Coordinates Plot (Z–Curves) for Four Variables



2.4 The Classical Approach

The classical approach to outlier detection is to compute the squared Mahalanobis distance for each observation based on the arithmetic mean $T(X) = (1/n)\Sigma_{i=1}^{n} x_i$ and the unbiased covariance estimator $C(X) = (1/(n-1))\Sigma_{i=1}^{n}(x_i - T(X))^T(x_i - T(X))$. Observations with large d_i^2 (possibly compared to some χ_p^2 quantile) are then considered as outliers.

However, this approach suffers from the fact that it is based on exactly those statistics that are most sensitive to outliers, namely, $T(X)$ and $C(X)$. This is particularly acute when there are several outliers forming a small cluster, because they will pull the arithmetic mean towards them and possibly even inflate the tolerance ellipsoid in their direction. It, thus, follows that they would not necessarily have large d_i^2. This is known as the *masking effect*. A natural consideration is to replace $T(X)$ and $C(X)$ by robust estimators.

A technique proposed by Campbell [1980] is to use M–estimators, but the low breakdown point of M–estimators (ie. the fraction of outliers they can tolerate) which is at most $1/(p+1)$ limits the applicability of the technique since the breakdown point goes down with higher dimensionality. Unfortunately it is exactly when there are more coordinates that there are more dimensions in which outliers can occur. Indeed in the modified wood gravity data ($p=5$) with four outliers in a sample of size 20, the limit of the breakdown, $1/(p+1) = 16.7\%$, has already been passed.

To proceed it is necessary to consider estimators of multivariate location and covariance with a high breakdown point. The first such estimator was proposed by Stahel [1981] and Donoho [1982]. Rousseeuw [1985] introduced the Minimum volume estimator (MVE).

2.5 The Minimum Volume Estimator (MVE) approach

In the Minimum Volume Estimator approach [Rouseeuw & van Zomeren 1990]

$T(Y)$ is taken as the center of the minimum volume ellipsoid covering half of the observations, and C(Y) is determined by the same ellipsoid.

DEFINITION 2.4: Minimum Volume Ellipsoid estimator

The Minimum Volume Ellipsoid estimator (MVE) is defined as the pair (T,C), where T(Y) is a p–vector and C(Y) is a positive–semidefinite p–by–p matrix such that the determinant of C is minimised subject to

$$\#\{i;\ (y_i - T)C^{-1}(y_i - T)^T \leq a^2\} \geq h \tag{2.52}$$

where $h = [(n + p + 1)/2]$ in which [q] is the integer part of q. The number a^2 is a fixed constant, which can be chosen as $\chi_p(0.50)$ when we expect the majority of the data to come from a normal distribution. #

For small sample size n a correction factor $c^2(n,p)$, which depends on n and p, is required. The MVE has a breakdown point of nearly 50%, which means that $T(Y)$ will remain bounded and the eigenvalues of C(Y) will stay away from zero and infinity when half the data is replaced by arbitrary values. The robust distances are defined relative to the MVE:

$$RD_i = \left[(y_i - T(Y))C^{-1}(Y)(y_i - T(Y))^T \right]^{1/2} \tag{2.53}$$

One can then compute a weighted mean,

$$T_1(Y) = (\sum_{i=1}^{n} w_i)^{-1} \sum_{i=1}^{n} w_i y_i \tag{2.54}$$

and weighted covariance matrix,

$$C_1(Y) = (\sum_{i=1}^{n} w_i - 1)^{-1} \sum_{i=1}^{n} (y_i - T(Y))C^T(Y)(y_i - T(Y)) \tag{2.55}$$

where the weights $w_i = w(RD_i)$ depend on the robust distances.

Rousseeuw and van Zomeren [1990] consider two approximate algorithms for the MVE. The first is the resampling algorithm described in Rouseeuw and Leroy [1987: pp.258–261]. It is based on the idea of looking for a small number of good points, rather than of k bad points, where k=1,2,3,..... .Subsamples of size p+1 different observations are drawn, indexed by $J = \{i_1,...,i_J\}$, say. The mean and covariance matrix of such a

subsample are

$$T_J = \Sigma_J y_i/(p+1) \tag{2.56}$$

and

$$C_J = \Sigma_J(y_i - T_J)(y_i - T_J)^T/p \tag{2.57}$$

Then the corresponding ellipsoid should be inflated or deflated to contain exactly h points, which amounts to computing

$$m_J^2 = \{(y_i - T_J)C_J^{-1}(y_i - T_J)^T\}_{h:n} \tag{2.58}$$

because m_J is the right magnification factor. The squared volume of the resulting ellipsoid is proportional to $m_J^{2p}\det(C_J)$, of which the smallest value is kept. For this "best" subset J we compute

$$T(Y) = T_J \tag{2.59}$$

and

$$C(Y) = [\chi^2(p,0.5)]^{-1}c^2(n,p)m_J^2 C_J \tag{2.60}$$

as an approximation to the MVE estimator, followed by a reweighting step as above. The number of subsamples J depends on a probabilistic argument, because it is required that enough subsamples consisting of p+1 good points are encountered. Rouseeuw and van Zomeren recommend $c^2(n,p) = (1+15/(n-p))^2$ for the small sample correction factor.

The second algorithm is the projection algorithm which is a variant of one by Gasko and Donoho [1982]. For each point y_i consider

$$u_i = \max_v \frac{|y_i v^T - L(y_1 v^T,\ldots,y_n v^T)|}{S(y_1 v^T,\ldots,y_n v^T)} \tag{2.61}$$

where L and S are the MVE estimates in one dimension, which are computed as follows: For any subset of numbers $z_1 \le z_2 \le \ldots \le z_n$ one can determine its shortest half by taking the smallest differences

$$z_h - z_1,\ z_{h+1} - z_2,\ldots,z_n - z_{n-h+1}.$$

If the smallest difference is $z_j - z_{j-h+1}$ we put L equal to the midpoint of the corresponding half,

$$L(z_1,...,z_n) = (z_j + z_{j-h+1})/2 \tag{2.62}$$

and S as its length,

$$S(z_1,...,z_n) = c(n)(z_j - z_{j-h+1}) \tag{2.63}$$

up to a correction factor $c(n)$, which depends on the sample size.It is noted that u is exactly a one dimensional version of RD but applied to projections y_iV^T of the y_i on the direction v. As not all possible directions v can be tried, a selection has to be made. We take all v of the form $y_l - M$ where $l = 1,...,n$ and M is the coordinatewise median:

$$M = (\underset{j=1,...,n}{\text{median}}\ y_{j1},..., \underset{j=1,...,n}{\text{median}}\ y_{jp}) \tag{2.64}$$

In the algorithm the array u_i i=1,...,n is updated while l loops over 1,...n. The final u_i are approximations of RD_i which can be plotted or used for reweighting.

2.6 The Hat Matrix

Some quantities that occur frequently in classical diagnostics are the diagonal elements of the least squares (LS) projection matrix *H*. This matrix is known under the name of *Hat Matrix*, because it puts a hat on the column vector $y = (y_1,...,y_n)^T$. This means that $\hat{y} = Hy$, where $\hat{y}$ is the LS prediction for y. The diagonal elements of the hat matrix are often used as diagnostic tools and in particular in linear regression they are used to detect *leverage points* ie. outliers in the carrier space. [See Atkinson 1985; Rousseeuw & Leroy 1987].

DEFINITION 2.5: Hat Matrix

Consider the $n \times 1$ vector of responses denoted by $y = (y_1,...,y_n)^T$;the linear model states that

$$y = X\theta + e \tag{2.65}$$

where X is the n×p matrix of the explanatory variables, θ is the vector of unknown parameters, and e is the error vector. The Hat Matrix is defined by

$$H = X(X^TX)^{-1}X^T \tag{2.66}$$

(it is assumed X^TX is invertible). #

The set of p–dimensional points x that satisfy

$$h_x = x(X^TX)^{-1}x^T \leq \max_i h_{ii} \qquad (2.67)$$

determine an ellipsoid [See eg. Montgomery & Peck 1982: p.143] which contains the smallest convex set enclosing the n observations. One can, thus, say that the point x lies close to the bulk of the space formed by the variables in X if h_x is small. The h_{ii} can be compared to some "cut–off" point. Most authors use a "cut–off" point of $2p/n$ and so determine potentially influential points as those having $h_{ii} > 2p/n$.

2.7 Proposed Iterative Sampling Technique – Stalactite Analysis

The proposed technique, *Stalactite Analysis*, involves the iterative computation of the Mahalanobis distances based on means and covariances computed from suitably selected subsamples of size m($<$ n). Initially, a subsample of size m=p+1 observations (sometimes refered to as an "elemental set") is chosen at random from which the mean vector and covariance matrix are computed. The Mahalanobis distances for all the n observations are computed and a new subsample of size m = m+k ($1 \leq k < n-p-1$) is selected based on the observations with the smallest Mahalanobis distances. The process is repeated until m=n. A new graphical display called the *Stalactite Chart* is proposed and used in the analysis together with some corresponding diagnostic quantities.

2.7.1 Stalactite Analysis Algorithm

Consider a random sample $y = (y_1, y_2, \ldots, y_p)$ of size n ($p < n$), we require to compute the squared Mahalanobis distances based on subsamples of size m from the main sample.

If we let d^2_{iJ} be the squared Mahalanobis distance of the i–th observation on the

J–th (J=0,2,...,n–p–2) iteration based on a subsample of size m_J($\leq$ n) then

$$d_{iJ}^2 = (y_i - \bar{y}_J)^T S_J^{-1} (y_i - \bar{y}_J) \qquad (2.68)$$

i = 1,2, ..., n

and $\bar{y}_J$ – the sample mean based on m_J observations

S_J – the sample var–covariance matrix based on m_J observations

The procedure is iterative and, hence, an initial value for m_0 needs to be determined. There are several possible choices for the initial value but a good starting point is when $m_0 = p + 1$, since this is the minimum number of points required to define an ellipsoid for the data. A subsample of size m_0 is selected at random and the d_{i0}^2 (for the full sample) are computed together with the mean vector, covariance matrix and other diagnostic quantitites. The subsample size is incremented by some quantity k (< n–p–1) ie. $m_J = m_{J-1} + k$ (J=1,2, ..., n–p–1) and the new subsample is selected to include those observations with the smallest m_J Mahalanobis distances d_{i0}^2. The required computations are performed and the process is repeated until $m_J = n$.

ALGORITHM I STALACTITE ANALYSIS ALGORITHM

The following is a summary of the algorithm:

Step 1: [Initialise.] Set $m_0 = p + 1$, $J = 0$, $k\ (=1)$.

Step 2: [Select.] Initial subsample $y_{\{i\}}$ $(i=1,2,...,m_0)$ randomly.

Step 3: [Compute.] The mean vector, covariance matrix and Mahalanobis distances according to the following relationships:

$$\bar{y}_J = \sum_{i=1}^{m_J} y_{\{i\}}/m_J\,, \tag{2.69}$$

$$S_J = \sum_{i=1}^{m_J} (y_{\{i\}} - \bar{y}_J)^2/(m_J - 1) \tag{2.70}$$

and

$$d_{iJ}^2 = (y_i - \bar{y}_J)^T S_J^{-1} (y_i - \bar{y}_J) \tag{2.71}$$

$(i = 1,2,...,n)$

Step 4: [Test.] If $m_J = n$ algorithm terminates goto Step 7

Step 5: [Increment.] The iteration number and the subsample size ie. $J = J+1$ and $m_J = m_{J-1} + k$.

Step 6: [Select.] New subsample $y_{\{i\}}$ $(i=1,2,...,m_J)$ based on the observations with the smallest $d_{i(J-1)}^2$ and go to Step 3.

Step 7: [Terminate.] Algorithm. #

The proposed display for a summary of the results of the procedure is termed as the *Stalactite Chart* (the term being borrowed from Geology).

DEFINITION 2.6: Stalactite Chart

If $y = (y_1, y_2, \ldots, y_p)^T$ is a sample of size n and $\bar{y}_J$ is the sample mean vector and S_J the sample covariance matrix on the J–th iteration based on a subsample of size m_J with

$$d_{iJ}^2 = (y_i - \bar{y}_J)^T S_J^{-1} (y_i - \bar{y}_J) \tag{2.72}$$

being the squared Mahalanobis distance of the i–th observation and also let

$$I_{iJ}^* = \begin{cases} 1, & d_{iJ}^2 > a^2 \\ 0, & d_{iJ}^2 \leq a^2 \end{cases} \tag{2.73}$$

where a^2 is some constant, then the Stalactite Chart is defined as the plot of I_{iJ}^* on i and J axes i=1,2,...,n and J=0,1,...,n–p–1. #

The quantity a^2 is some "cut–off" point suitably chosen, for instance, using the distributional properties of d_i^2 ie. if y comes from a multivariate normal population then

$$d_i^2 \xrightarrow{\text{approx}} \chi_p^2 \tag{2.74}$$

so a^2 could be taken as $\chi_p^2(1-\alpha)$ for some significance level α. In this study an alternative choice for a^2 is adopted, namely, E[Max χ_p^2] for the following reasons:

– E[Max χ_p^2] { = f(n,p) } takes into account information from both the variable space,p, and the the observation space n and so provides a more data specific statistic, ie. the data suggest their own "cut–off" point, whereas χ_p^2 { = f(p) } could be the same for data sets with very diverse magnitudes of sample sizes.

– E[Max χ_p^2] by definition can be easily computed as $\chi_p^2((n-1/2)/n)$ and is independent of the individual observations ,hence, robust.

– As mentioned earlier, the choice of a "cut–off" point (hence, rendering an

observation as an outlier) is quite subjective and ,thus, as much tolerance of an observation is required before it is classified as an outlier. In addition to the above reasons E[Max χ^2_p] provides more tolerance than $\chi^2_p(1-\alpha)$ provided n is not too small. In particular, the following table gives limits for n in comparison to selected α levels.

TABLE 2.2 Lower Bounds of Sample Size n for E[Max χ^2_p] > $\chi^2_p(1-\alpha)$

$1-\alpha$	n
0.90	$\geq$ 5
0.95	$\geq$ 10
0.99	$\geq$ 50

Using Wilson and Hilferty's result (Cf. Section 2.1) we can define

$$I^{\varphi}_{iJ} = \begin{cases} 1, & \sqrt[3]{d^2_J} > a^{\varphi} \\ 0, & \sqrt[3]{d^2_J} \leq a^{\varphi} \end{cases} \qquad (2.75)$$

where a^{φ} is the $\Phi^{-1}(1-\alpha)$ or $\Phi^{-1}((n-1/2)/n)$.

As mentioned earlier, the selection of a "cut–off" point is quite subjective and, thus, tolerance of an observation should be fairly high before it is classified as an outlier. Since, the proposed method is iterative it is possible to retain the "history" of an observation across iterations and summarise its presence (or absence) in the "bad" class into a single statistic. The proposed statistic here is termed as the Stalactite Score (since it is derived directly from the Stalactite Chart).

DEFINITION 2.7: Stalactite Score

If $y = (y_1, y_2, \ldots, y_p)$ is a sample of size n with d^2_{iJ} being the squared Mahalanobis distance for the i–th observation on the J–th iteration with

$$I^*_{iJ} = \begin{cases} 1, & d^2_{iJ} > a^2 \\ 0, & d^2_{iJ} \leq a^2 \end{cases} \tag{2.76}$$

i=1,2,..,n and a^2 some "cut–off" point let

$$r_i = \sum_{J=0}^{n-p-2} I^*_{iJ}/(n-p-1) \tag{2.77}$$

then the Stalactite Score for the i–th observation, SS_i, is defined as

$$SS_i = \begin{cases} 0, & r_i = 0 \\ c, & (c-1)/w < r_i \leq c/w \end{cases} \tag{2.78}$$

c = 1,2,...,w where w is an arbitrary constant representing the highest score. #

An observation can be deemed for further scrutiny if it has a high Stalactite Score, ideally w. The choice of w is subjective. However, a low value tends to retain the influence of an extreme observation on the Compound Mean described below, hence, it shows the general direction in the p space where there may be an extreme observation (or cluster of observations). A high value for w reduces the influence of an extreme observation and, thus, makes the Compound Mean a more robust measure of location. The recommended values for w are 4 when interest is on the general direction of influence and 9 or 10 when interest is on providing a robust estimate for location.

The usual mean vector $\bar{y}$ is constructed using the marginal arithmetic means of the variables. It is then used to measure the centre of the point cloud of the data. However, since its construction does not take into consideration the interrelationships between the variables, the centre thus obtained may not be the true one. This is especially so if a point cluster existed which only affected correlation and with little or no effect on both the location and scale on any of the dimensions.

It is, therefore, possible to exploit this fact by constructing a measure of the centre

taking into account the interrelationships between the variables with a specific view for detecting these point clouds.

A proposed statistic termed as the *Compound Mean* is such a statistic.

DEFINITION 2.8: Compound Mean

If $y = (y_1, y_2, \ldots, y_p)$ is a sample of size n with Stalactite Score vector $SS = (SS_1, SS_2, \ldots, SS_n)$ then the Compound Mean, $\bar{y}_c$, is defined as

$$\bar{y}_c = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{2.79}$$

where the weights $w_i(SS_i) = 1/(SS_i + 1)$. #

A weight of 1/w suggests that the observation is probably atypical and a weight of 1/(w+1) certainly indicates an atypical observation.

If $\bar{y}_j$ is the arithmetic mean computed marginally for the j–th variable and $\bar{y}_{cj}$ its compound mean then we have the following

$|\bar{y}_j - \bar{y}_{cj}| < \delta$ (for some small δ) – no presence of a "pull" of the centre in the j–th direction

$|\bar{y}_j - \bar{y}_{cj}| >> \delta$ (for some small δ) – presence of a "pull" of the centre in the j–th direction.

A "pull" refers to the tendency of an observation (or set of observations) to attract the center of the point cloud towards them.

If we let

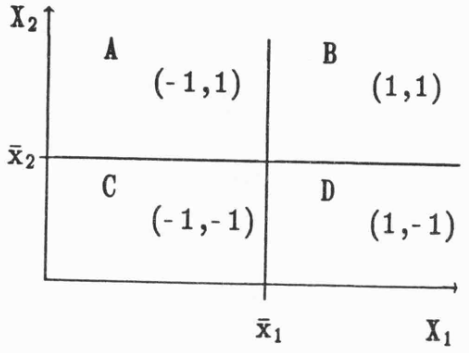
$$\tilde{p}_j = \begin{cases} 1, & \underline{y}_j < \underline{y}_{cj} \\ 0, & \bar{\underline{y}}_j = \bar{\underline{y}}_{cj} \\ -1, & y_j > y_{cj} \end{cases} \tag{2.80}$$

y_j, y_{cj} as above, then $\tilde{p}_j$ is an indicator of a "pull" in the j–th direction eg. a negative value

of $\tilde{p}_j$ implies that the center of the point cloud is lower than the sample mean in the j–th direction. A vector of all the values of $\tilde{p}_j$ provides a quick summary of the directions of pull, $\tilde{P} = (\tilde{p}_1,\tilde{p}_2,\ldots,\tilde{p}_p)$.

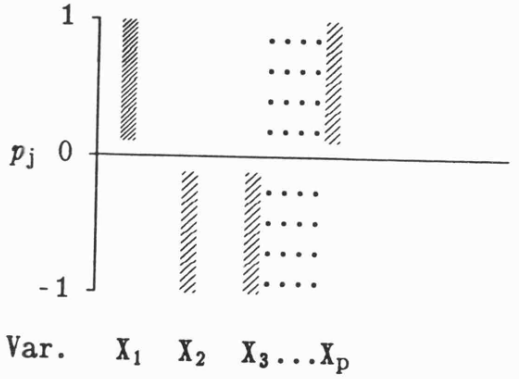
The pull can also be used as a crude measure of detection for the direction (if any) in which clustering appears. Figure 2.3 demonstrates the this in two dimensions.

FIGURE 2.3 Direction of Pull in Two Dimensional Space



Each of the quadrants A,B,C and D displays the possible composition of the P vector and hence the direction of the "pull". In more than two dimensions a possible representation is displayed in Figure 2.4.

FIGURE 2.4 Direction of Pull in p–Dimensional Space



A further refinement to the displays would be to replace the $\tilde{p}_j$ values with the actual magnitudes and signs of the differences between the arithmetic mean and the compound mean. These provide a measure of the relative "pull" in each of the variables

and so making it easier to identify the most outlying one. The drawback here is be that all the variables have to be suitably scaled and standardised for direct comparison.

It is useful to study the behaviour of the arithmetic means during the iterations. This behaviour is displayed in what is termed as the *Means Plot.*

DEFINITION 2.9: Means Plot

A Means Plot is the plot of the arithmetic mean at each iteration of the Stalactite Analysis against the iteration number. #

Another diagnostic statistic of interest is the proportion of "bad" observations to "good" ones. In this study, this proportion has been referred to as the *Contamination Index.* For completeness this index is computed at each iteration.

DEFINITION 2.10: Contamination Index

If $x = (x_1,x_2,...,x_p)$ is a sample of size n with d^2_{iJ} being the squared Mahalanobis distance for the i–th observation on the J–th iteration then let G_J and B_J be the number of "good" observations and "bad" observations, respectively, on the J–th iteration ie.

$$\#\{G_J;\ d^2_{iJ} \leq a^2\} \tag{2.81}$$

and

$$\#\{B_J;\ d^2_{iJ} > a^2\} \tag{2.82}$$

i=1,2,...,n , then the Contamination Index for the J–th iteration, C^*_J, is defined as

$$C^*_J = B_J/G_J \tag{2.83}$$

#

The Contamination Index C^* can be compared to some tolerance level, τ, say and a subsample of relatively "good" observations can be selected. The size of this subsample n_J can be determined as

$$\{n_J;\ C^*_J \leq \tau\} \tag{2.84}$$

where J is the iteration number. Also, the particular observations that constitute the subsample can be determined by selecting those observations with the smallest Stalactite Scores first ie. with SS=0 then SS=1, etc until the desired size n_J is reached. Conversely,

the observations with the largest Stalactite Score can be removed from the sample until the desired size n_J, so that the remaining observations would constitute the required subsample.

2.8 EXAMPLES

The approaches and tests for the detection of outliers exemplified in this chapter are just a few of the numerous ones. Each approach has specific circumstances within which it is suitable, for example the $Z(r_{-i})$ (see Section 2.2.1) can only be applied to bivariate data. Furthermore, each test is designed to cope with and detect a specific number of outliers. The discordancy tests in Table 2.1 are all designed to test for the presence of a single outlier whereas the Classical approach (Section 2.4), the Minimum Volume Ellipsoid (MVE) approach (Section 2.5), the Hat–Matrix approach (Section 2.6) and the proposed Stalactite Analysis (Section 2.7) can detect multiple outliers.

This section presents examples of the application of the approaches discussed. Each data set has been chosen to portray specific characteristics which may arise in typical data. The first two examples are based on simulated bivariate normal data with 50 observations. These data are used as control data to demonstrate the expected behaviour of the techniques under known and predetermined conditions. The first of these data, Example E.1, is generated so as not to contain any obvious outliers and this is used as the null data set. The second, Example E.2, is a contamination of the first data set with the introduction of four outlying observations. These outlying observations are introduced to depict the three types of effects which outliers may have, as portrayed in Figure 2.1; inflating variances and no effect on correlations, reducing correlation and inflating the variance in one variable, and no effect on variances but with reduction of the correlation. The remaining data are well known and have been used extensively in the literature on robust regression and outlier diagnostics.

Example E.3 is the Belgian Telephone calls data [Rousseeuw and Leroy 1987] and is the number of international telephone calls (in tens of millions) in the years 1950–1973.

There is heavy contamination from 1964 to 1969 and it turns out that another recording system was used, giving the total number of minutes of these calls. This data set demonstrates the masking effect but with one variable, the year (although a sequential variable) has no outlyingness and the observations causing the masking lie centrally in this variable.

The fourth data set, Example E.4, is the Hertzsprung–Russell diagram of the star cluster CYG OB1, which contains 47 stars in the direction of Cygnus. This contains the logarithm of the effective temperature at the surface of the star and the logarithm of its light intensity [Rousseeuw and Leroy 1987: p.27]. These data also exhibit masking. Although one of the variables does not contain obvious outliers the observations causing the masking lie at the extreme.

The fifth data set, Example E.5, is the Hawkins–Bradu–Kass artificial data [Hawkins, Bradu and Kass 1984]. These data being artificially generated offer the advantage that the position of the bad points is known exactly. The data set consists of 75 observations in four dimensions (one response and three explanatory variables). For the purposes of this thesis, following Rousseeuw and van Zomeren [1990], only the explanatory variables are used. In the complete data set (with all variables) the first 10 observations are bad leverage points (outlyingness in the explanatory variables), and the next four are good leverage points (outlyingness in the explanatory variables but corresponding responses fit the model quite well). In all, therefore, there are 14 outliers and these form a cluster far away from the rest of the data and hence exhibit masking.

The initial step in the analyses is to obtain the summary statistics so as to study the marginal behaviour of the variables. These statistics include the M–estimators of the means. This is followed by graphical studies with Box plots and scatter plots for a visual display of the behaviour of the variables marginally and jointly, respectively. In the case of bivariate data, the analysis based on the correlation coefficient with case deletion is carried out. The results are then summarised and tested using a normal plot of the $Z(r_{-i})$. The

multivariate outlier diagnostics are then carried out. These include the Classical approach, the Hat–Matrix approach and the proposed Stalactite Analysis approach. In the Hawkins–Bradu–Kass example the results of results from the Minimum Volume Ellipsoid approach are included.

EXAMPLE E.1 Simulated Bivariate Normal Data: The data consist of two sets p=2 of 50 (=n) computer generated standard normal deviates. These deviates are transformed pairwise to construct bivariate normal deviates with a given correlation between them, ρ. The following is the relationship used for the transformations.

$$y_{1i} = x_{1i}$$

$$y_{2i} = \rho x_{1i} + \sqrt{(1-\rho^2)}x_{2i}$$

($i = 1,2,...,50$), to obtain 50 samples (y_{1i}, y_{2i}) from

$$N\left[\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right].$$

To avoid negative values, the mean vector is shifted sufficiently away from the origin by adding a constant vector (c_1, c_2) to each of the observations. The observation vector is also rescaled by a matrix A such that

$$\begin{bmatrix}z_1\\z_2\end{bmatrix} = A\begin{bmatrix}y_1\\y_2\end{bmatrix} = \begin{bmatrix}\sigma_1 y_1\\ \sigma_2 y_2\end{bmatrix}$$

where $A = \begin{bmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{bmatrix}$ and so the final sample is

$$\begin{bmatrix}z_1\\z_2\end{bmatrix} = \left[\begin{bmatrix}c_1\\c_2\end{bmatrix}, \begin{bmatrix}\sigma^2 & \rho\\ \rho & \sigma^2\end{bmatrix}\right].$$

The constants used in the analysis are $c_1 = 10$, $c_2 = 20$, $\sigma_1 = 5$ $\sigma_2 = 10$ and $\rho = 0.6$.

The three measures of location (Table E.1(b)), the mean, median and 5% trimmed mean for both variables are in close agreement which suggests that both variables are fairly marginally symmetrical. This is further confirmed by the almost zero coefficient of skewness values for both and the interquartile range is approximately twice the standard deviation. The negative values for kurtosis suggest that both variables are concentrated close to the respective means. There is also negligible differences between the M–estimators. Graphically, the box plots show close symmetry for both variables although with slightly longer stems to the right.

The scatter plot in Figure E.1(b) is evenly distributed around an elliptical point cloud with a correlation of 0.59. There are no obvious unusual points, outliers. Figure E.1(c) is the normal plot of the case deletion correlation coefficient function $Z(r_{-i})$. Apart from a few straggling points in both extremes, the plot is fairly linear which means that no individual observation affects the correlation between the variables and hence, no outlying observations.

From Table 2.1 all the discordancy tests applied to these data are not significant at the 5% level and so this further confirms that there does not appear to be any single outlying observations.

The Stalactite Chart and Stalactite diagnostics are displayed in Figure E.1(d) and Table E.1(d) respectively. According to Figure E.1(d) the "deepest" stalactite has a "depth" of 80 i.e. it ceases after 80% of the closest observations to the center of the point cloud have been selected for the computation of the means and covariance matrix (Note: A depth of 100 is equivalent to the Classical approach). It further implies that there is only one observation at that point which is outlying (a fact which can also be observed from the Contamination Index CIX (ratio of bad to good observations) of 0.02 from Table E.1(d). The identity of this observation can be determined from the Stalactite Scores and it is observation 2. It lies on the edge of the bottom of the ellipsoid. Two other observations lying on the edge are observations 24 and 37 which again are identified by the Stalactite

Scores. If a tolerance level $\tau = 0.05$ for the CIX is used it leads to only one outlier in the data and a selection of at least 40 observations and at most 49 without observation 2 will ensure a "clean" data set. Although three observations are detected, by quick inspection of the scatter plot (Figure E.1(b)) they all lie well within range of the marginal distributions and hence do not affect the marginal measures of location and spread. It is worth noting that this is the reason that none of the discordancy tests could detect them. However, the effect of these observations, albeit slight, is in affecting the correlation between the variables.

Figures 1(e) and 1(f) display the Mahalanobis index plot (MIP) of the data at 90% sub–sample size and full sample size, respectively. At 90% sub–sample size the $\chi^2(0.95)$ cut–off detects two possible outliers whereas the E[Max χ^2] does not. The labeling of an observation as an outlier should be after it has quite clearly failed the tests and this example shows how using a less tolerant cut–off point may condemn an observation when it is in fact a viable point. Using the full sample both cut–off points do not detect any outliers as is also visible from the Stalactite Chart.

Finally, Figure E.1(g) is the Means Plot for the data. The variation of the mean for Y_1 is very slight and is concentrated around the full sample mean and the compound mean. This is verified by noting that the three observations which were identified as outliers all lie well within the range of the Y_1 space and are fairly close to its mean. On the other hand, the mean for Y_2 starts off large and is "pulled" down as the sub–sample size increases. Further, the compound mean is lower than the full sample mean. This suggests that the inclusion of some of the observations tends to pull the mean down and these are the observations with high Stalactite Scores. On inspection on the scatter plot it is noted that the three outlier observations are far below the central tendency in the Y_2 space. Referring to Figure 2.3 the "pull" falls within quadrant D i.e. with a pull vector $\tilde{P} = (1, -1)$.

EXAMPLE E.1 Simulated Bivariate Normal Data (with no outliers). Table E.1(a) displays the data from Example E1 where the constants used in the analysis are $c_1 = 10$, $c_2 = 20$, $\sigma_1 = 5$ $\sigma_2 = 10$ and $\rho = 0.6$.

TABLE E.1(a) Simulated Bivariate Normal Data (with no outliers)

Obs	Y_1	Y_2	Obs	Y_1	Y_2	Obs	Y_1	Y_2
1	11.1	26.03	18	12.13	30.83	35	16.75	27.54
2	12.4	10.52	19	16.50	22.25	36	5.32	18.00
3	4.6	10.00	20	9.18	20.57	37	10.79	8.67
4	5.65	18.00	21	16.00	22.00	38	6.90	23.80
5	7.08	14.00	22	4.78	6.22	39	7.00	20.00
6	6.79	12.00	23	10.00	24.00	40	12.65	26.81
7	18.07	33.46	24	15.19	16.50	41	14.15	28.91
8	11.76	20.37	25	14.55	33.29	42	15.00	26.00
9	4.87	12.00	26	8.53	12.00	43	14.00	24.00
10	14.64	18.30	27	8.45	17.15	44	10.31	26.00
11	11.24	13.74	28	8.00	15.00	45	8.83	25.00
12	7.25	20.00	29	14.87	32.52	46	11.56	18.41
13	10.06	15.00	30	15.21	28.97	47	7.38	20.32
14	9.64	27.40	31	12.54	19.10	48	12.00	22.00
15	11.32	29.89	32	6.77	6.90	49	9.86	22.30
16	10.00	12.00	33	13.21	18.49	50	10.98	27.44
17	15.00	726.42	34	9.57	11.74			

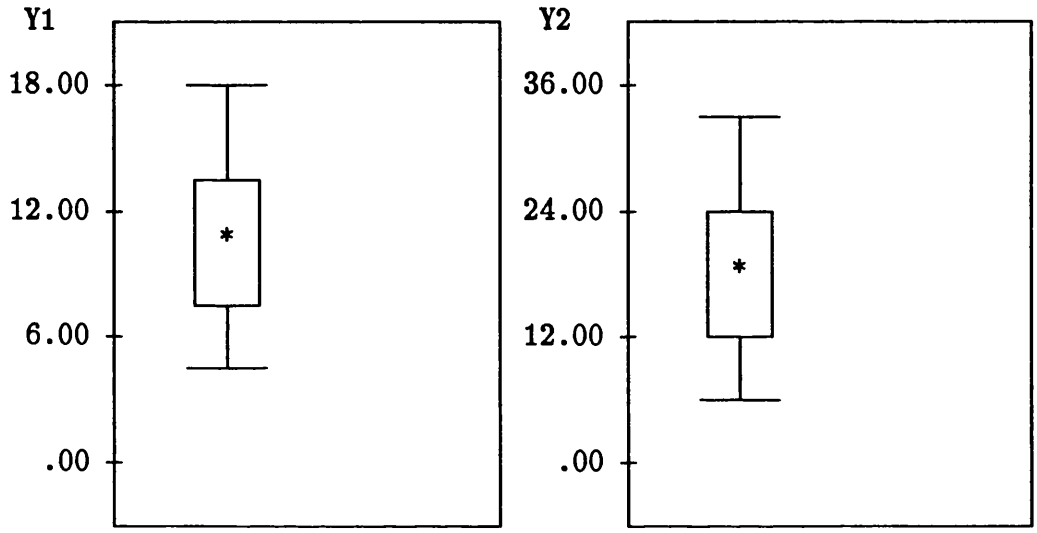
TABLE E.1(b) Summary Statistics for Simulated Bivariate Normal Data (with no outliers)

Statistic	Y_1	Y_2
Location		
Mean	10.81	20.44
Median	10.89	20.35
5% Trim	10.79	20.47
Std Err	0.50	1.01
Dispersion		
Variance	12.37	51.02
Std Dev	3.52	7.14
Min	4.59	6.23
Max	18.07	33.47
Range	13.48	27.24
IQR	6.19	11.38
Skewness & Kurtosis		
Skewness	0.05	-0.10
S E Skew	0.34	0.34
Kurtosis	-0.87	-0.81
S E Kurt	0.66	0.66

TABLE E.1(c) M-Estimators

Statistic	Y_1	Y_2
Huber (1.34)	10.81	20.57
Hampel (1.70,3.40,8.50)	10.81	20.48
Tukey (4.69)	10.80	20.55
Andrew (1.3 * pi)	10.80	20.55

FIGURE E.1(a) Box Plots for Simulated Bivariate Normal Data (with no outliers)



Symbol Key: * - Median (...) - Outliers

Figure E.1(b) Simulated Bivariate Normal Data (with no outliers)
Scatter Plot

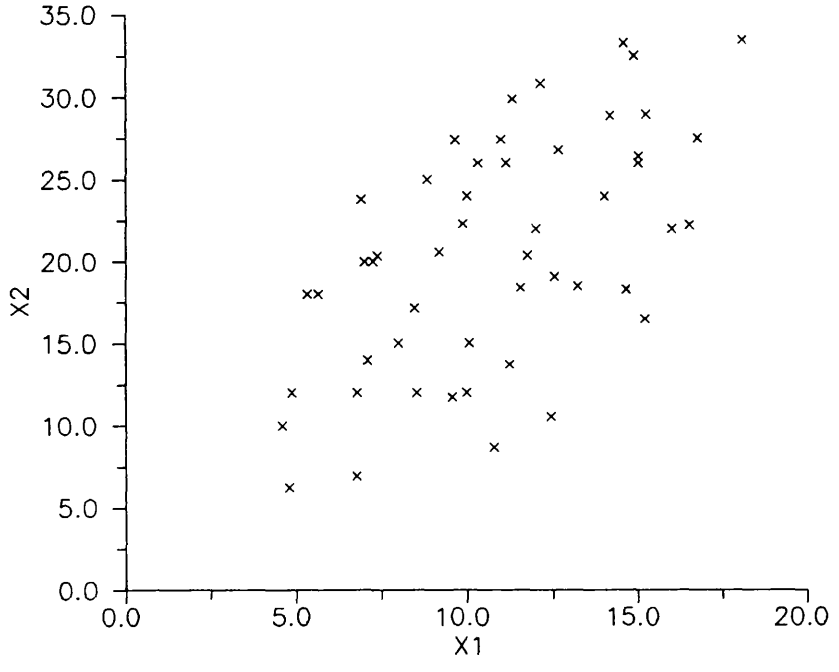


Figure E.1(c) Simulated Bivariate Normal Data (with no outliers)
Normal Plot of $Z(r_{-i})$

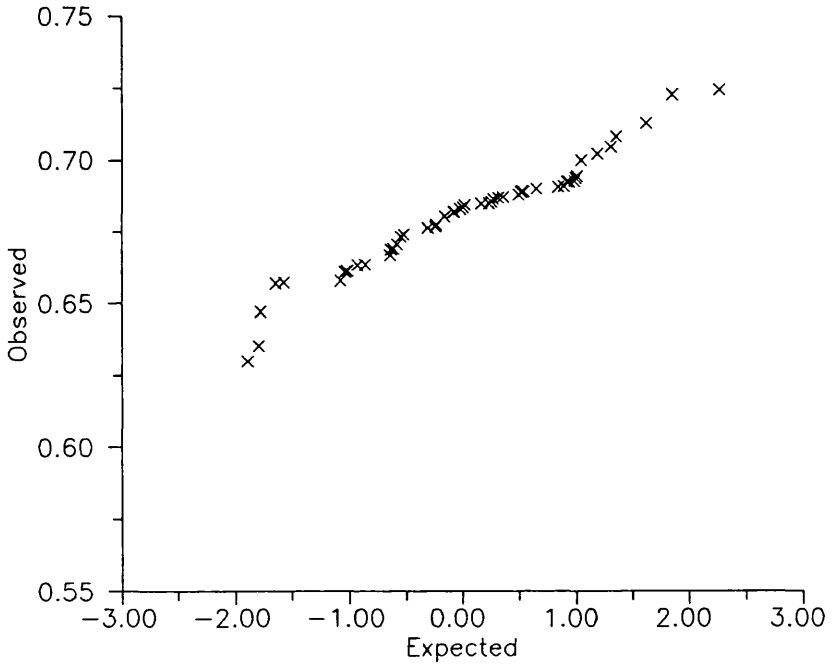


FIGURE E.1(d) Simulated Bivariate Normal Data (w/o outliers) Stalactite Chart

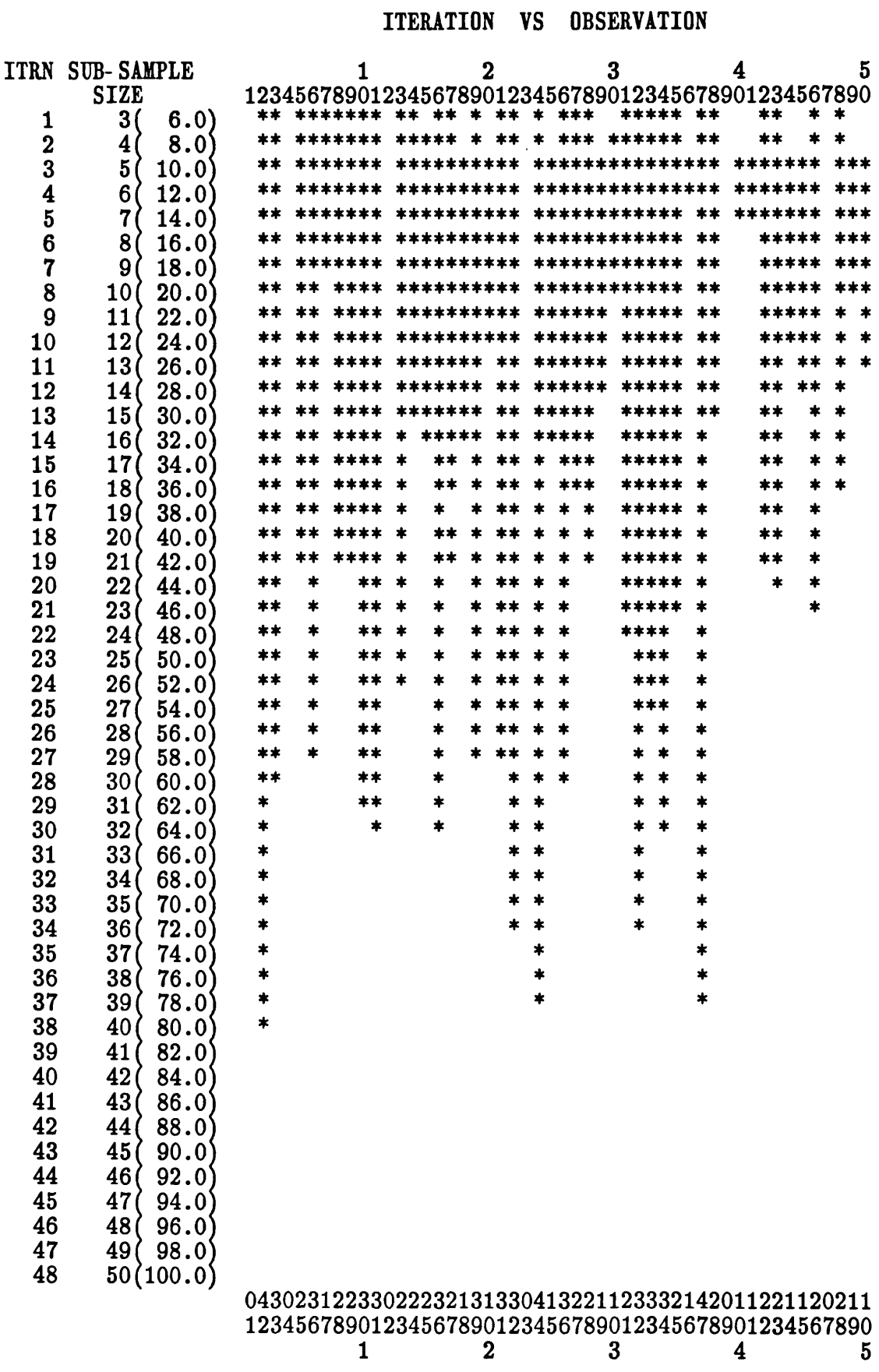


TABLE E.1(d) Simulated Bivariate Normal Data (w/o outliers) Stalactite Analysis

ITRN	SUB-SAMPLE SIZE	OBSERVATION GOOD #(%)	OBSERVATION BAD #(%)	BAD:GOOD RATIO	TOTAL SQ. DISTANCE OBS.	TOTAL SQ. DISTANCE EXP.
1	3(6.0)	19(38.0)	31(62.0)	1.63	4.00	4.00
2	4(8.0)	17(34.0)	33(66.0)	1.94	0.78	6.00
3	5(10.0)	6(12.0)	44(88.0)	7.33	8.00	8.00
4	6(12.0)	6(12.0)	44(88.0)	7.33	10.00	10.00
5	7(14.0)	7(14.0)	43(86.0)	6.14	12.00	12.00
6	8(16.0)	9(18.0)	41(82.0)	4.56	14.00	14.00
7	9(18.0)	9(18.0)	41(82.0)	4.56	16.00	16.00
8	10(20.0)	10(20.0)	40(80.0)	4.00	18.00	18.00
9	11(22.0)	12(24.0)	38(76.0)	3.17	20.00	20.00
10	12(24.0)	12(24.0)	38(76.0)	3.17	22.00	22.00
11	13(26.0)	14(28.0)	36(72.0)	2.57	24.00	24.00
12	14(28.0)	15(30.0)	35(70.0)	2.33	26.00	26.00
13	15(30.0)	17(34.0)	33(66.0)	1.94	28.00	28.00
14	16(32.0)	19(38.0)	31(62.0)	1.63	30.00	30.00
15	17(34.0)	22(44.0)	28(56.0)	1.27	32.00	32.00
16	18(36.0)	22(44.0)	28(56.0)	1.27	34.00	34.00
17	19(38.0)	25(50.0)	25(50.0)	1.00	35.24	36.00
18	20(40.0)	24(48.0)	26(52.0)	1.08	36.96	38.00
19	21(42.0)	24(48.0)	26(52.0)	1.08	40.00	40.00
20	22(44.0)	30(60.0)	20(40.0)	0.67	42.00	42.00
21	23(46.0)	31(62.0)	19(38.0)	0.61	44.00	44.00
22	24(48.0)	33(66.0)	17(34.0)	0.52	46.00	46.00
23	25(50.0)	34(68.0)	16(32.0)	0.47	48.00	48.00
24	26(52.0)	34(68.0)	16(32.0)	0.47	50.00	50.00
25	27(54.0)	35(70.0)	15(30.0)	0.43	52.00	52.00
26	28(56.0)	36(72.0)	14(28.0)	0.39	54.00	54.00
27	29(58.0)	36(72.0)	14(28.0)	0.39	56.00	56.00
28	30(60.0)	39(78.0)	11(22.0)	0.28	58.00	58.00
29	31(62.0)	41(82.0)	9(18.0)	0.22	60.00	60.00
30	32(64.0)	42(84.0)	8(16.0)	0.19	61.80	62.00
31	33(66.0)	45(90.0)	5(10.0)	0.11	64.00	64.00
32	34(68.0)	45(90.0)	5(10.0)	0.11	65.90	66.00
33	35(70.0)	45(90.0)	5(10.0)	0.11	67.81	68.00
34	36(72.0)	45(90.0)	5(10.0)	0.11	69.57	70.00
35	37(74.0)	47(94.0)	3(6.0)	0.06	72.00	72.00
36	38(76.0)	47(94.0)	3(6.0)	0.06	74.00	74.00
37	39(78.0)	47(94.0)	3(6.0)	0.06	76.00	76.00
38	40(80.0)	49(98.0)	1(2.0)	0.02	77.47	78.00
39	41(82.0)	50(100.0)	0(0.0)	0.00	80.00	80.00
40	42(84.0)	50(100.0)	0(0.0)	0.00	82.00	82.00
41	43(86.0)	50(100.0)	0(0.0)	0.00	84.00	84.00
42	44(88.0)	50(100.0)	0(0.0)	0.00	86.00	86.00
43	45(90.0)	50(100.0)	0(0.0)	0.00	88.00	88.00
44	46(92.0)	50(100.0)	0(0.0)	0.00	90.00	90.00
45	47(94.0)	50(100.0)	0(0.0)	0.00	92.00	92.00
46	48(96.0)	50(100.0)	0(0.0)	0.00	94.00	94.00
47	49(98.0)	50(100.0)	0(0.0)	0.00	96.00	96.00
48	50(100.0)	50(100.0)	0(0.0)	0.00	98.00	98.00

Figure E.1(e) Simulated Bivariate Normal Data (with no outliers)
Index Plot of the Mahalanobis Distances (90% Sample)

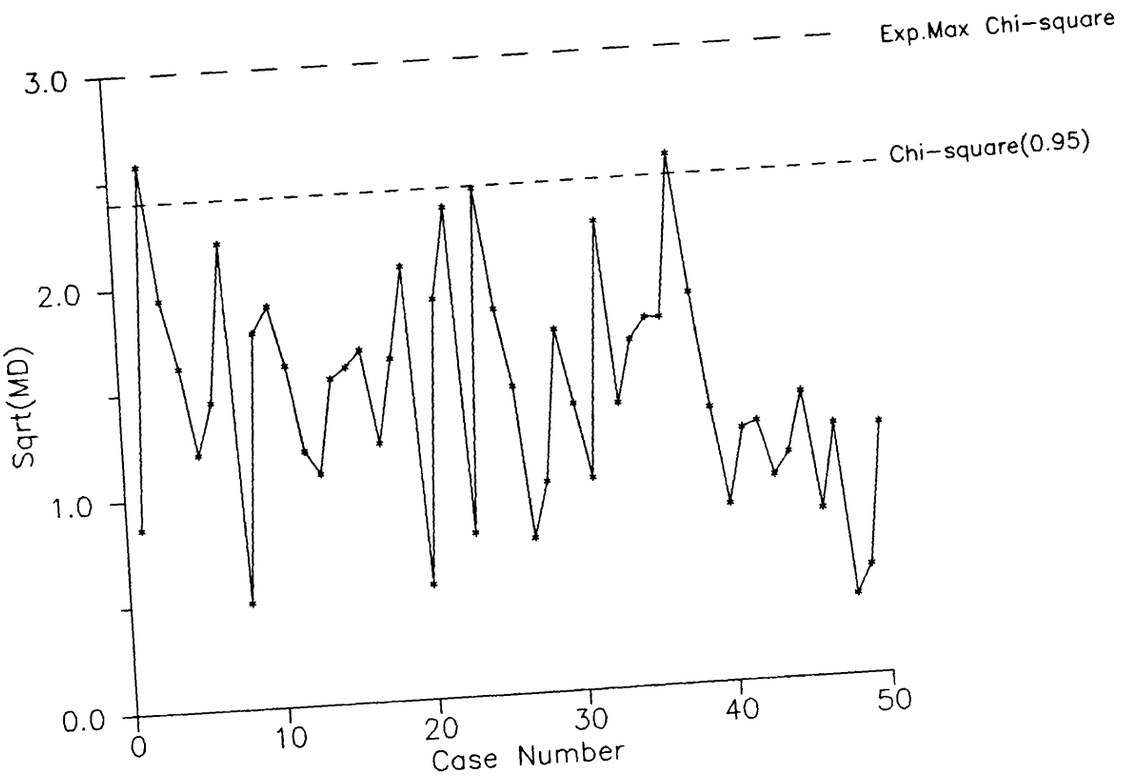


Figure E.1(f) Simulated Bivariate Normal Data (with no outliers)
Index Plot of the Mahalanobis Distances (Full Sample)

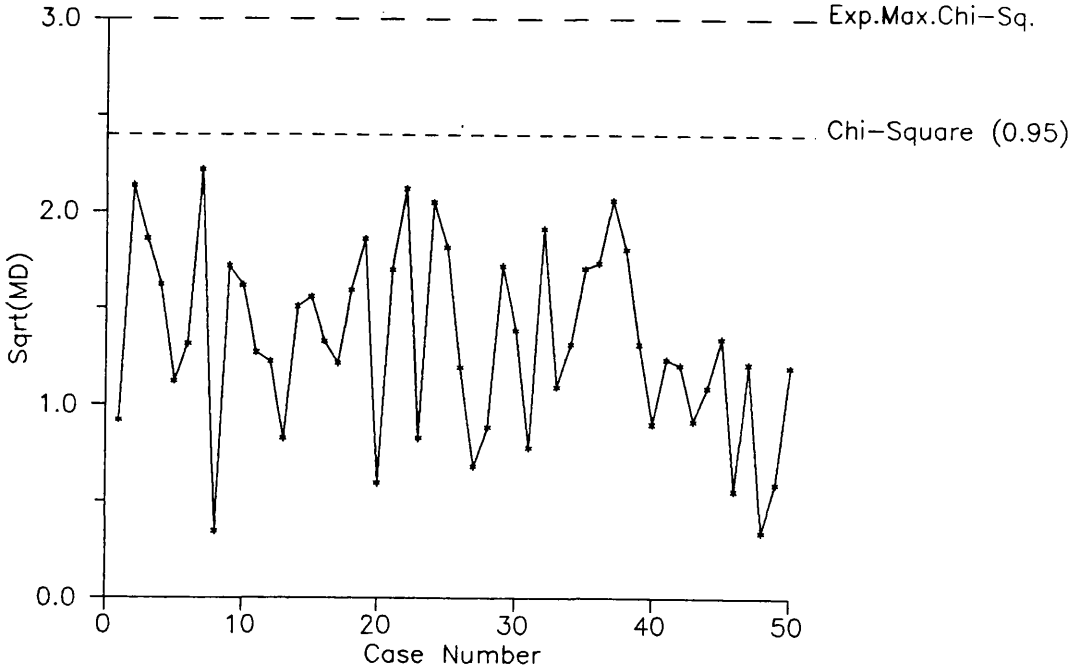
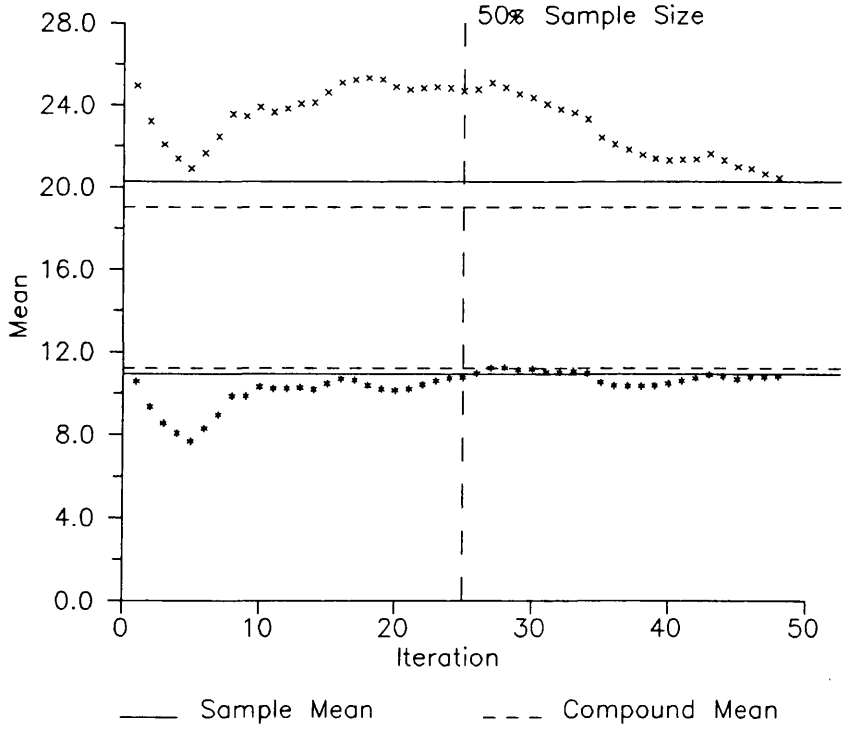


Figure E.1(g) Simulated Bivariate Normal Data (with no outliers)
Means Plot



EXAMPLE E.2 Simulated Bivariate Normal Data (with 4 outliers). Table E.2(a) displays the data from Example E.1 with four observations replaced by outlying ones. The four observations are 16, 28, 39 and 48.

TABLE E.2(a) Simulated Bivariate Normal Data (with 4 outliers)

Obs	Y_1	Y_2	Obs	Y_1	Y_2	Obs	Y_1	Y_2
1	11.12	26.03	18	12.13	30.83	35	16.75	27.54
2	12.43	10.52	19	16.50	22.25	36	5.32	18.00
3	4.59	10.00	20	9.18	20.57	37	10.79	8.67
4	5.65	18.00	21	16.00	22.00	38	6.90	23.80
5	7.08	14.00	22	4.78	6.22	39	2.50	30.00
6	6.79	12.00	23	10.00	24.00	40	12.65	26.81
7	18.07	33.46	24	15.19	16.50	41	14.15	28.91
8	11.76	20.37	25	14.55	33.29	42	15.00	26.00
9	4.87	12.00	26	8.53	12.00	43	14.00	24.00
10	14.64	18.30	27	8.45	17.15	44	10.31	26.00
11	11.24	13.74	28	25.00	40.00	45	8.83	25.00
12	7.25	20.00	29	14.87	32.52	46	11.56	18.41
13	10.06	15.00	30	15.31	28.97	47	7.38	20.32
14	9.64	27.40	31	12.54	19.10	48	30.00	20.00
15	11.32	29.89	32	6.77	6.90	49	9.86	22.30
16	30.00	44.00	33	13.21	18.49	50	10.98	27.44
17	15.00	26.42	34	9.57	11.74			

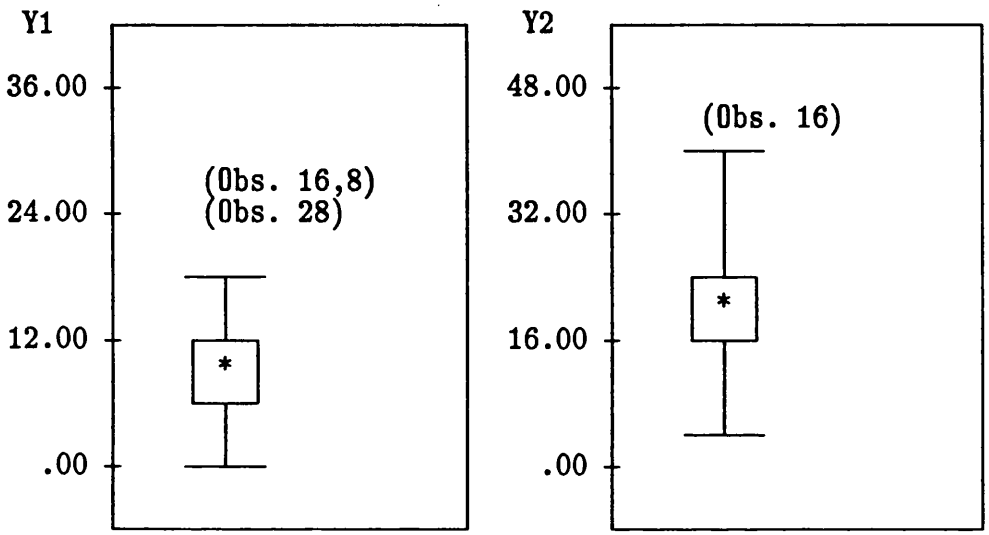
TABLE E.2(b) Summary Statistics for Simulated Bivariate Normal Data (with 4 outliers)

Statistic	Y_1	Y_2
Location		
Mean	11.90	21.74
Median	11.18	21.29
5% Trim	11.40	21.53
Std Err	.80	1.17
Dispersion		
Variance	32.29	68.18
Std Dev	5.68	8.26
Min	2.50	6.23
Max	30.00	44.00
Range	27.5	
IQR	6.52	11.29
Skewness & Kurtosis		
Skewness	1.37	0.28
S E Skew	0.34	0.34
Kurtosis	2.86	0.05
S E Kurt	0.66	0.66

TABLE E.2(c) M-Estimators

Statistic	Y_1	Y_2
Huber (1.34)	11.24	21.64
Hampel (1.70,3.40,8.50)	11.14	21.51
Tukey (4.69)	10.88	21.36
Andrew (1.34 * pi)	10.86	21.35

FIGURE E.2(a) Box Plots for Simulated Bivariate Normal Data (with 4 outliers)



Symbol Key: * - Median (...) - Outliers

EXAMPLE E.2 Simulated Bivariate Normal Data (with 4 outliers) Analysis

The three measures of location (Table E.2(b)), the mean, median and 5% trimmed mean for both variables are in close agreement but with a larger standard error than the uncontaminated data. The coefficient of skewness value for Y_1 is also high and positive whereas that of Y_2 is small. This means that Y_1 has a long tail to the right and some unusual observations may exist but Y_2 is almost symmetric. The interquartile range for Y_1 is also not much larger than the standard deviation and that of Y_2 is not too different from the original data. The coefficient of kurtosis for Y_1 is large compared to the change in that of Y_2. There also differences between the M–estimators. Graphically, the box plots show close symmetry for both variables although with slightly longer stems to the right. In particular, three observations (16, 8 and 28) in the Y_1 space are detected as outliers and one (observation 16) in the Y_2 space.

The scatter plot in Figure E.2(b) is evenly distributed around an elliptical point cloud with a correlation of 0.56. There are four obvious unusual points, outliers, one on either side of the point cloud and two within the direction of the point cloud although they are outlying in both the dimensions. Figure E.2(c) is the normal plot of the case deletion correlation coefficient function $Z(r_{-i})$. There are four points which are far away from the otherwise linear plot. Two of these have low observed values for $Z(r_{-i})$ this implies that deleting either of them reduces the correlation. The other two have high values and so deleting them increases correlation. On inspecting the Scatter Plot (Figure E.2(b)) it is visible that the lefttmost two outlying observations correspond to the former category and the rightmost two to the latter.

From Table 2.1 all the discordancy tests apart from the $Z(r_{-i})$ are significant at the 5% level and so this further confirms that there does appear to be at least one outlying observation.

The fact that all tests so far indicate that there is at least one outlier makes it useful to apply the multivariate tests for further investigation of the data together with

obtaining the identities of these outlying observations.

The Stalactite Chart and Stalactite diagnostics are displayed in Figure E.2(d) and Table E.2(d) respectively. According to Figure E.2(d) there are two stalactites which have a depth of 100, there are also two more with a depth of 98 and 96. This implies that at full depth there are two observations that are outlying, the Contamination Index CIX being 0.04 from Table E.2(d). From the Stalactite Scores these observations are identified as observations 16 and 48, the other two are observations 28 and 39. The pairing of these observations is identical to the pairing obtained in the case deletion correlation test. Similarly, observation 16 is detected by the box plots in both variables but in Y_1 even observation 48 is detected. These are the rightmost observations in the scatter plot and so their exclusion from the data would increase the correlation. If a tolerance level $\tau = 0.05$ for the CIX is used it leads to two outliers in the data and a selection of 48 observations without observations 16 and 48 will ensure a "clean" data set.

Figures 2(e) and 2(f) display the Mahalanobis index plot (MIP) of the data at 90% sub–sample size and full sample size, respectively. At 90% sub–sample size both the $\chi^2(0.95)$ cut–off and the E[Max χ^2] detect the four outliers. Using the full sample $\chi^2(0.95)$ cut–off detects upto eleven outliers whereas the E[Max χ^2] cut–off points detect the four known outliers correctly.

Figure E.2(g) displays the Means Plot for the data. The mean for Y_1 is starts off relatively small in magnitude but is "pulled" up as the sample size increases. This is verified by noting that three of the outlying observations clearly lie far from the majority of the data in the Y_1 space and so they do affect the central tendency by pulling the mean towards them. On the other hand, the mean for Y_2 has very slight variation about the full sample mean. This is expected since the outlying observations fall close to the majority of the data in this dimension and so do not affect the mean. Further, the compound mean is higher than the full sample mean in both variables indicating a positive pull for the mean in both dimensions. This suggests that the inclusion of some of the observations tends to

pull the mean up and these are the observations with high Stalactite Scores. On inspection of the scatter plot it is noted that the three outlier observations are far above the central tendency in both variables. Refering to Figure 2.3 the "pull" falls within quadrant B i.e. with a pull vector $\tilde{P} = (1, 1)$.

Figure E.2(b) Simulated Bivariate Normal Data (with 4 outliers)
Scatter Plot

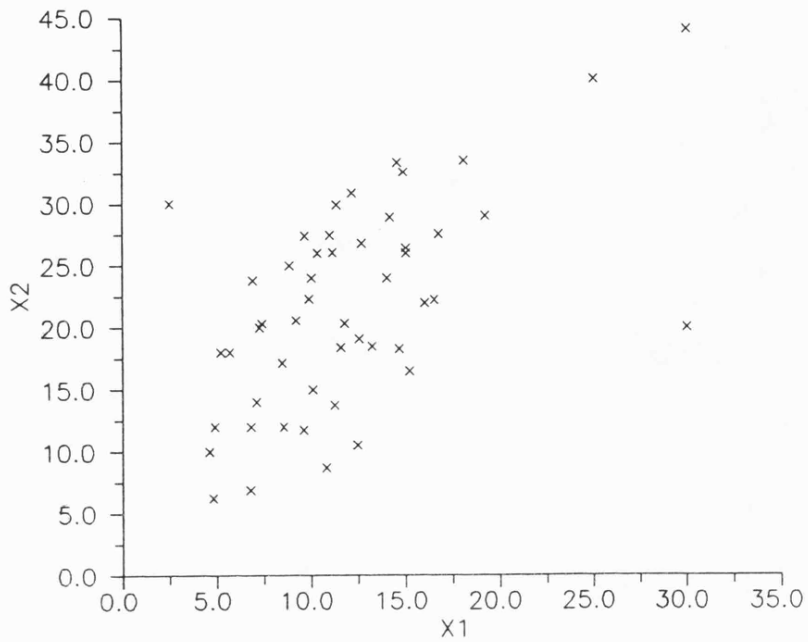


Figure E.2(c) Simulated Bivariate Normal Data (with 4 outliers)
Normal Plot of $Z(r_{-i})$

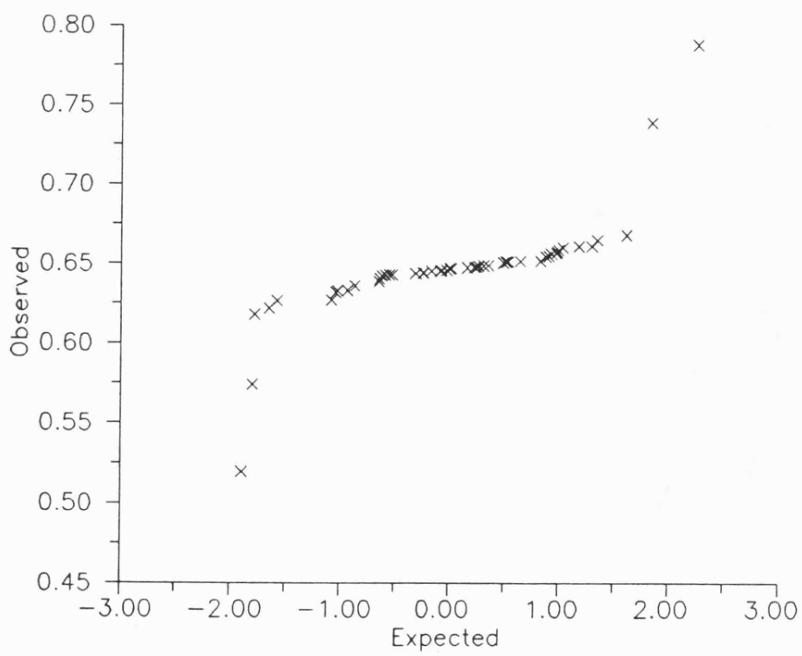


FIGURE E.2(d) Simulated Bivariate Normal Data (w/4 outliers) Stalactite Chart

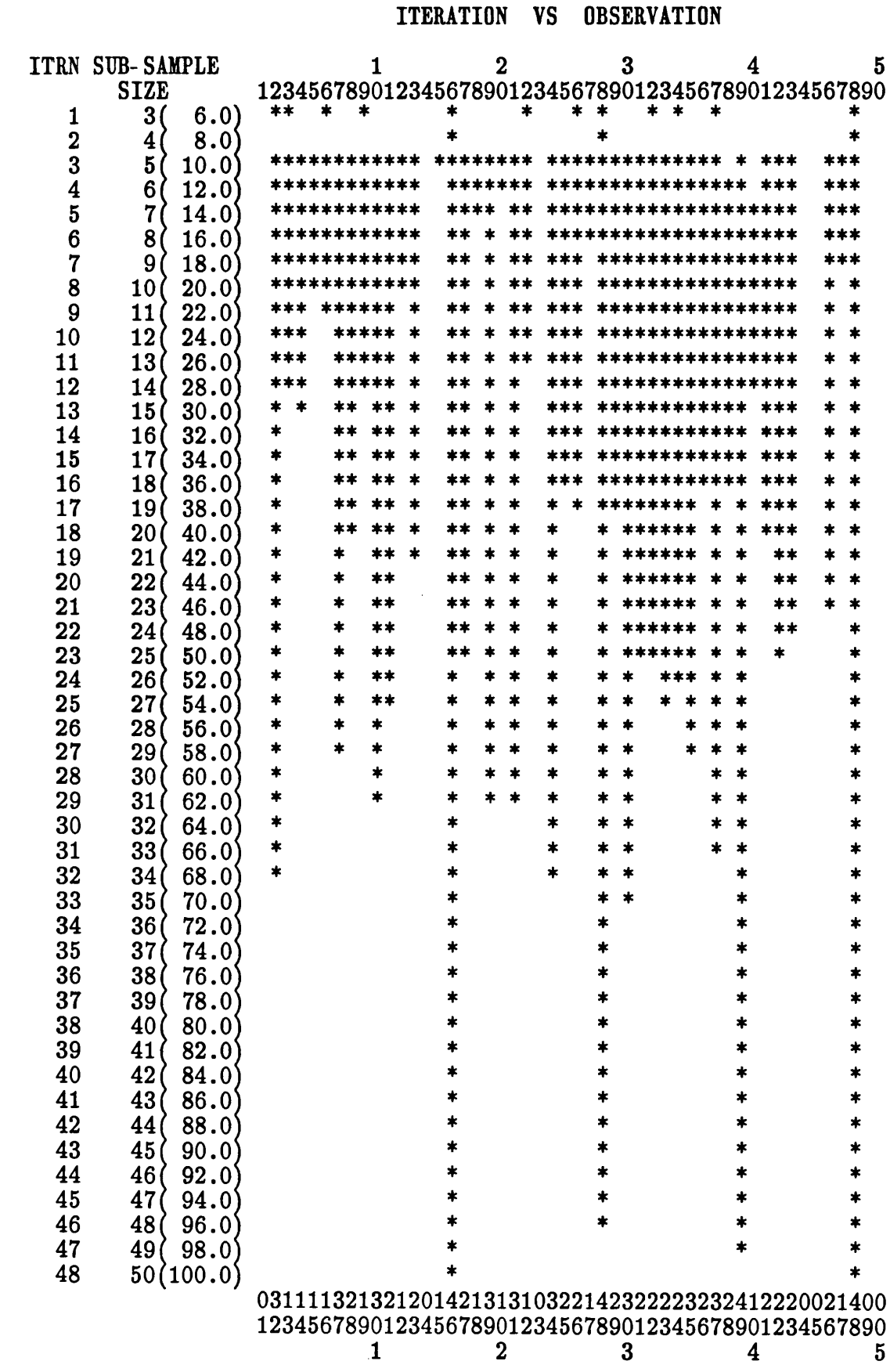


TABLE E.2(d) Simulated Bivariate Normal Data (w/4 outliers) Stalactite Analysis

ITRN	SUB-SAMPLE SIZE	OBSERVATION GOOD #(%)	OBSERVATION BAD #(%)	BAD:GOOD RATIO	TOTAL SQ. DISTANCE OBS.	TOTAL SQ. DISTANCE EXP.
1	3(6.0)	38(76.0)	12(24.0)	0.32	4.00	4.00
2	4(8.0)	47(94.0)	3(6.0)	0.06	0.40	6.00
3	5(10.0)	9(18.0)	41(82.0)	4.56	7.48	8.00
4	6(12.0)	9(18.0)	41(82.0)	4.56	10.00	10.00
5	7(14.0)	9(18.0)	41(82.0)	4.56	12.00	12.00
6	8(16.0)	10(20.0)	40(80.0)	4.00	14.00	14.00
7	9(18.0)	11(22.0)	39(78.0)	3.55	16.00	22.00
11	13(26.0)	15(30.0)	35(70.0)	2.33	24.00	24.00
12	14(28.0)	16(32.0)	34(68.0)	2.13	26.00	26.00
13	15(30.0)	19(38.0)	31(62.0)	1.63	28.00	28.00
14	16(32.0)	20(40.0)	30(60.0)	1.50	30.00	30.00
15	17(34.0)	20(40.0)	30(60.0)	1.50	32.00	32.00
16	18(36.0)	20(40.0)	30(60.0)	1.50	34.00	34.00
17	19(38.0)	23(46.0)	27(54.0)	1.17	36.00	36.00
18	20(40.0)	25(50.0)	25(50.0)	1.00	38.00	38.00
19	21(42.0)	27(54.0)	23(46.0)	0.85	40.00	40.00
20	22(44.0)	28(56.0)	22(44.0)	0.79	42.00	42.00
21	23(46.0)	28(56.0)	22(44.0)	0.79	44.00	44.00
22	24(48.0)	29(58.0)	21(42.0)	0.72	44.95	46.00
23	25(50.0)	30(60.0)	20(40.0)	0.67	47.36	48.00
24	26(52.0)	34(68.0)	16(32.0)	0.47	50.00	50.00
25	27(54.0)	35(70.0)	15(30.0)	0.43	53.59	54.00
27	29(58.0)	37(74.0)	13(26.0)	0.35	56.00	56.00
28	30(60.0)	39(78.0)	11(22.0)	0.28	57.68	58.00
29	31(62.0)	39(78.0)	11(22.0)	0.28	59.43	60.00
30	32(64.0)	42(84.0)	8(16.0)	0.19	62.00	62.00
31	33(66.0)	42(84.0)	8(16.0)	0.19	64.00	64.00
32	34(68.0)	43(86.0)	7(14.0)	0.16	66.00	66.00
33	35(70.0)	45(90.0)	5(10.0)	0.11	68.00	68.00
34	36(72.0)	46(92.0)	4(8.0)	0.09	70.00	70.00
35	37(74.0)	46(92.0)	4(8.0)	0.09	72.00	72.00
36	38(76.0)	46(92.0)	4(8.0)	0.09	74.00	74.00
37	39(78.0)	46(92.0)	4(8.0)	0.09	76.00	76.00
38	40(80.0)	46(92.0)	4(8.0)	0.09	78.00	78.00
39	41(82.0)	46(92.0)	4(8.0)	0.09	80.00	80.00
40	42(84.0)	46(92.0)	4(8.0)	0.09	82.00	82.00
41	43(86.0)	46(92.0)	4(8.0)	0.09	84.00	84.00
42	44(88.0)	46(92.0)	4(8.0)	0.09	86.00	86.00
43	45(90.0)	46(92.0)	4(8.0)	0.09	87.97	88.00
44	46(92.0)	46(92.0)	4(8.0)	0.09	90.00	90.00
45	47(94.0)	46(92.0)	4(8.0)	0.09	92.00	92.00
46	48(96.0)	46(92.0)	4(8.0)	0.09	94.00	94.00
47	49(98.0)	47(94.0)	3(6.0)	0.06	96.00	96.00
48	50(100.0)	48(96.0)	2(4.0)	0.04	98.00	98.00

Figure E.2(e) Simulated Bivariate Normal Data (with 4 outliers)
Index Plot of the Mahalanobis Distances (90% Sample)

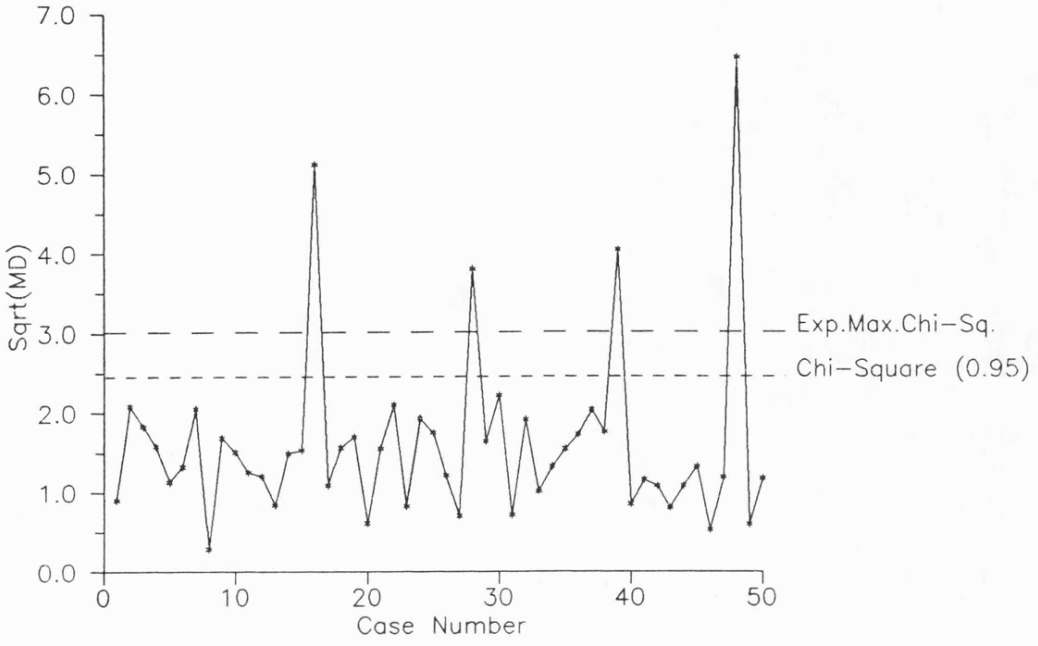


Figure E.2(f) Simulated Bivariate Normal Data (with 4 outliers)
Index Plot of the Mahalanobis Distances (Full Sample)

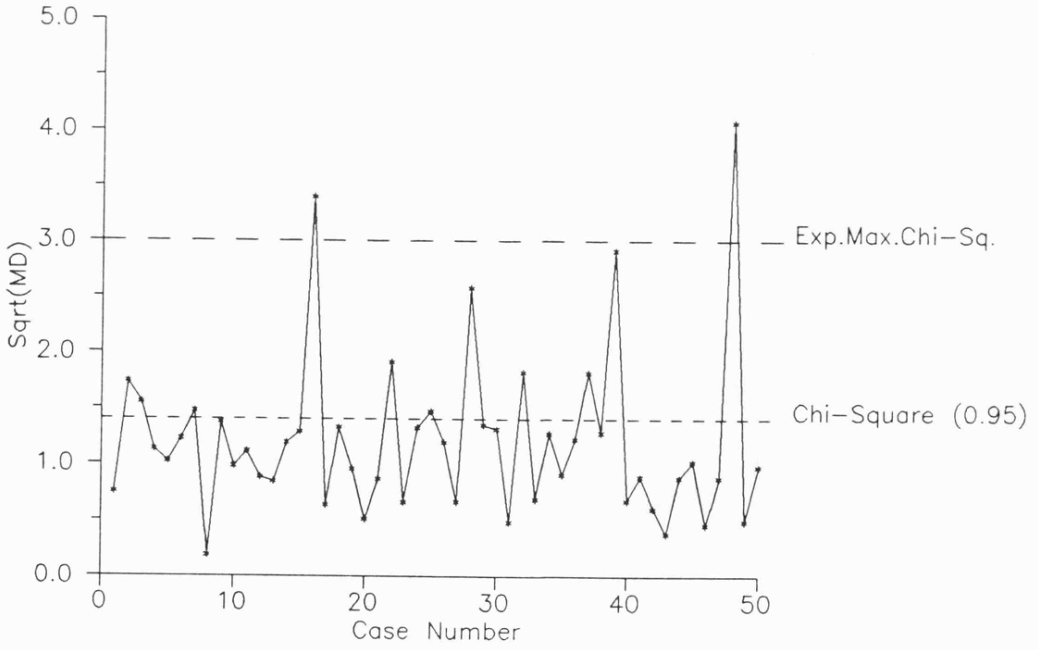
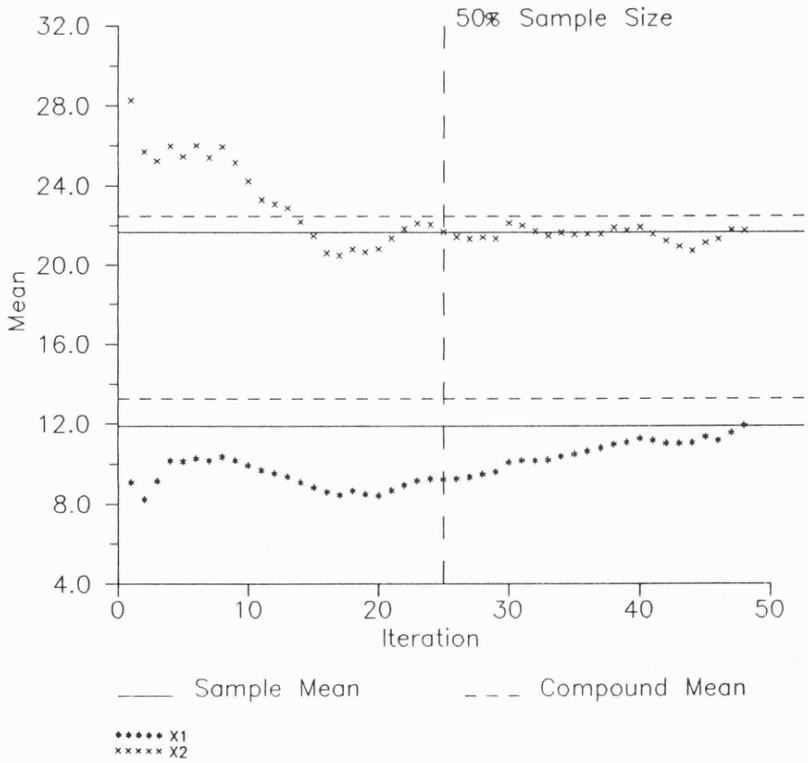


Figure E.2(g) Simulated Bivariate Normal Data (with 4 outliers)
Means Plot



EXAMPLE E.3 Belgian Phone Calls. The data consist of the total number (in tens of millions) of international phone calls from Belgium in the years 1950-1973 (Y_1 - Year, Y_2 - Number of calls).

TABLE E.3(a) Belgian Phone Calls

Obs	Y_1	Y_2	Obs	Y_1	Y_2
1	50	0.44	13	62	1.61
2	51	0.47	14	63	2.12
3	52	0.47	15	64	11.90
4	53	0.59	16	65	12.40
5	54	0.66	17	66	14.20
6	55	0.73	18	67	15.90
7	56	0.81	19	68	18.20
8	57	0.88	20	69	21.20
9	58	1.06	21	70	4.30
10	59	1.20	22	71	2.40
11	60	1.35	23	72	2.70
12	61	1.49	24	73	2.90

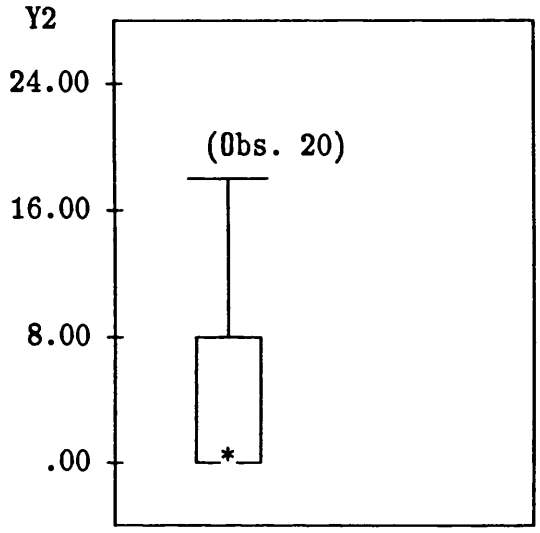
TABLE E.3(b) Summary Statistics for Belgian Phone Calls

Statistic	Y_2
Location	
Mean	5.09
Median	1.49
5% Trim	4.48
Std Err	1.39
Dispersion	
Variance	44.69
Std Dev	6.69
Min	0.44
Max	11.17
Range	10.73
Skewness & Kurtosis	
Skewness	1.35
S E Skew	0.48
Kurtosis	0.34
S E Kurt	0.94

TABLE E.3(c) M-Estimators

Statistic	Y_2
Huber (1.33)	1.73
Hampel (1.70,3.40,8.50)	1.28
Tukey (6.69)	1.20
Andrew (1.34 * pi)	1.20

FIGURE E.3(a) Box Plot for Belgian Phone Calls



Symbol Key: * - Median (...) - Outliers

EXAMPLE E.3 Belgian Phone Calls Analysis

In this analysis the summary statistics of Y_2 only are considered (the number of phone calls) since Y_1 is just a sequential variable (year). There are marked differences in the three measures of location (Table E.3(b)). The mean is significantly larger than the median and the 5% trimmed mean is no different. This suggests that Y_2 has a very long tail to the right. This is further confirmed by the large positive value of the coefficient of skewness. There is, however, negligible differences between the M–estimators and these compare well to the median. Graphically, the box plot shows strong asymmetry and indeed it does not have a left stem.

The scatter plot in Figure E.3(b) displays a linear trend apart from six observations which clearly stand out from the rest of the data. Figure E.3(c) is the normal plot of the case deletion correlation coefficient function $Z(r_{-i})$. There are a few points which are far away from the otherwise linear plot in both extremes.

From Table 2.1 all the discordancy tests are not significant at the 5% level. This is explained by the fact that these tests are for detecting a single outlier but in these data the effect is masked by the fact that there are several observations which are jointly influential but have little influence if looked at individually.

The discordancy tests, therefore, show no evidence of outliers and yet the summary statistics and the graphical techniques clearly indicate their presence. This makes it necessary to conduct multivariate tests.

The Stalactite Chart and Stalactite diagnostics are displayed in Figure E.3(d) and Table E.3(d) respectively. According to Figure E.3(d) there is no stalactite with a depth of 100, there is one with a depth of 95.8, two with 91.7. All the six outlying observations are detected up to a depth of 83.3. The fact that at full depth there are no observations that are outlying implies that the Classical approach does not detect any outliers. From the Stalactite Scores all the six outlying observations are identified and are observations 15, 16, 17, 18, 19 and 20. If a tolerance level $\tau = 0.05$ for the CIX is used it leads to one outlier in

the data and a selection of 23 observations without observations 20 will ensure a "clean" data set at the 5% CIX tolerance level.

Table E.3(e) is a summary of the different multivariate results from the different approches. These are the case deletion correlation coefficient, the $Z(r_{-i})$, the diagonal elements of the Hat matrix, the Mahalanobis matrix and the Stalactite Scores. Apart from the Stalactite Scores, which identify all the six outlying observations, none of the other values detect these outliers

Figure E.3(b) Belgian Phone Calls Data
Scatter Plot

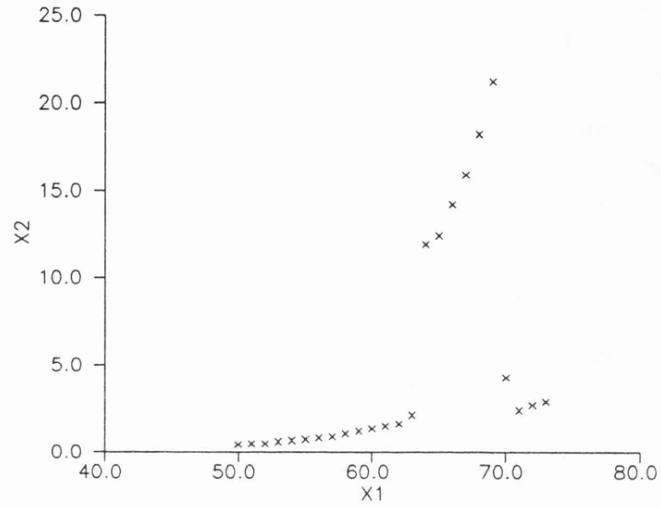


Figure E.3(c) Belgian Phone Calls Data
Normal Plot of $Z(r_{-i})$

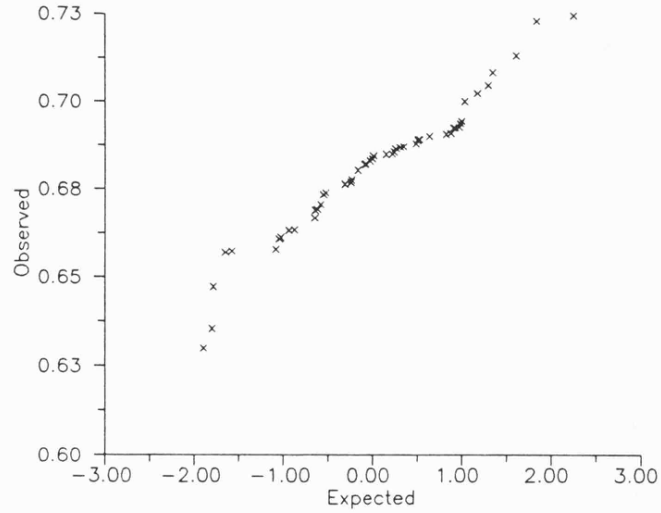


FIGURE E.3(d) Belgian Data Stalactite Chart

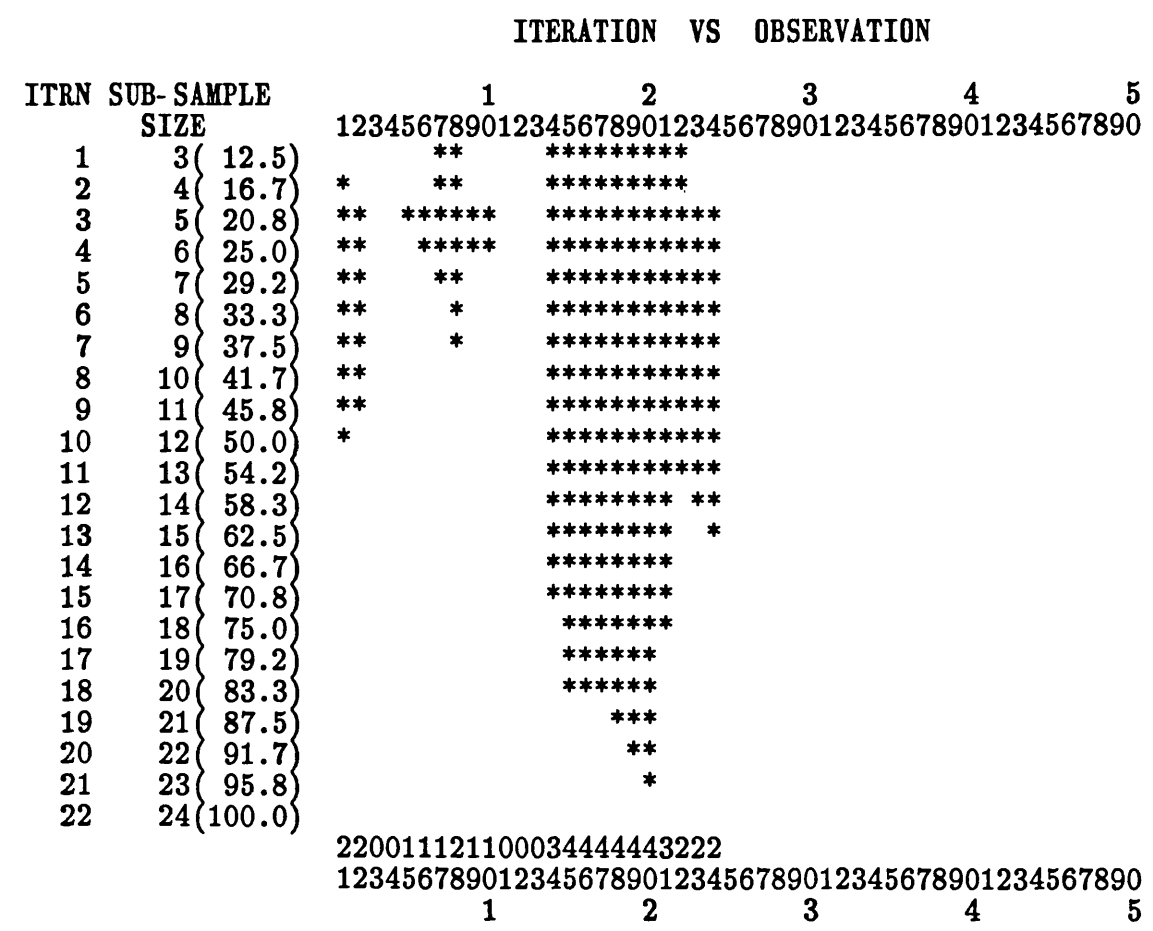


TABLE E.3(d) Belgian Data Stalactite Analysis

ITRN	SUB-SAMPLE SIZE	OBSERVATION GOOD #(%)	OBSERVATION BAD #(%)	BAD:GOOD RATIO	TOTAL SQ. DISTANCE OBS.	TOTAL SQ. DISTANCE EXP.
1	3(12.5)	13(54.2)	11(45.8)	0.85	4.00	4.00
2	4(16.7)	12(50.0)	12(50.0)	1.00	1.87	6.00
3	5(20.8)	5(20.8)	19(79.2)	3.80	8.00	8.00
4	6(25.0)	6(25.0)	18(75.0)	3.00	10.00	10.00
5	7(29.2)	9(37.5)	15(62.5)	1.67	14.00	14.00
7	9(37.5)	10(41.7)	14(58.3)	1.40	16.00	16.00
8	10(41.7)	11(45.8)	13(54.2)	1.18	18.00	18.00
9	11(45.8)	11(45.8)	13(54.2)	1.18	20.00	20.00
10	12(50.0)	12(50.0)	12(50.0)	1.00	22.00	22.00
11	13(54.2)	13(54.2)	11(45.8)	0.85	24.00	24.00
12	14(58.3)	14(58.3)	10(41.7)	0.71	26.00	26.00
13	15(62.5)	15(62.5)	9(37.5)	0.60	28.00	28.00
14	16(66.7)	16(66.7)	8(33.3)	0.50	30.00	30.00
15	17(70.8)	16(66.7)	8(33.3)	0.50	32.00	32.00
16	18(75.0)	17(70.8)	7(29.2)	0.41	34.00	34.00
17	19(79.2)	18(75.0)	6(25.0)	0.33	36.00	36.00
18	20(83.3)	18(75.0)	6(25.0)	0.33	38.00	38.00
19	21(87.5)	21(87.5)	3(12.5)	0.14	40.00	40.00
20	22(91.7)	22(91.7)	2(8.3)	0.09	42.00	42.00
21	23(95.8)	23(95.8)	1(4.2)	0.04	44.00	44.00
22	24(100.0)	24(100.0)	0(0.0)	0.00	46.00	46.00

TABLE E.3(e) Case Deletion Correlation Coefficient, Diagonal Elements of the Hat Matrix, Mahalanobis Distances, and Stalactite Scores for the Belgian Phone Calls Data

Obs. i	r(-i)	z[r(-i)]†	h_{ii} (0.167)	d_{ii} (2.71)	SS_i (4)
1	0.531	0.592	0.157	1.67	2
2	0.530	0.590	0.138	1.52	2
3	0.530	0.589	0.120	1.37	0
4	0.530	0.591	0.105	1.22	0
5	0.531	0.592	0.091	1.08	1
6	0.532	0.593	0.078	0.95	1
7	0.534	0.595	0.068	0.83	2
8	0.536	0.598	0.059	0.73	2
9	0.538	0.601	0.052	0.64	1
10	0.540	0.605	0.047	0.59	1
11	0.543	0.608	0.044	0.57	0
12	0.546	0.613	0.042	0.60	0
13	0.549	0.617	0.042	0.68	0
14	0.551	0.620	0.044	0.70	3
15	0.542	0.608	0.047	1.10	4
16	0.537	0.600	0.052	1.16	4
17	0.533	0.594	0.059	1.44	4
18	0.526	0.585	0.068	1.70	4
19	0.519	0.575	0.078	2.06	4
20	0.513	0.567	0.091	2.54	4
21	0.569	0.646	0.105	1.53	3
22	0.595	0.686	0.120	1.94	2
23	0.600	0.693	0.138	2.07	2
24	0.607	0.703	0.157	2.21	3

† $z[r(-i)] = (1/2)\log_e\{(1 + r_{-i})/(1 - r_{-i})\}$

Note: $h_{ii} > 0.167$ (=2p/n), distances d_{ii} exceeding "cutoff" value $\sqrt{\chi^2_2(0.975)} = 2.71$ and $SS_i = 4$ are underlined. Also, $\sqrt{\chi^2_2(0.950)} = 2.45$.

EXAMPLE E.4 Hertzsprung-Russell Diagram of the Star Cluster CYG OB1. The data in Table E.5 form the Hertzsprung-Russell diagram of the star cluster CYG OB1, which contains 47 stars in the direction of Cygnus. Here Y_1 is the logarithm of the effective temperature at the surface of the star (T_e), and Y_2 is the logarithm of its light intensity (L/L_o).

TABLE E.4(a) Hertzsprung-Russell Diagram of the Star Cluster CYG OB1.

Obs	Y_1	Y_2	Obs	Y_1	Y_2	Obs	Y_1	Y_2
1	4.37	4.23	17	4.23	3.94	33	4.45	5.22
2	4.56	5.74	18	4.42	4.18	34	3.49	6.29
3	4.26	4.93	19	4.23	3.94	35	4.23	4.34
4	4.56	5.74	20	3.49	5.89	36	4.62	5.62
5	4.30	5.19	21	4.29	4.38	37	4.53	5.10
6	4.26	5.57	24	4.49	4.85	40	4.43	5.57
9	4.57	5.27	25	4.38	5.02	41	4.38	4.62
10	4.37	5.12	26	4.42	4.66	42	4.45	5.06
11	3.49	5.73	27	4.29	4.66	43	4.50	5.34
12	4.43	5.45	28	4.38	4.90	44	4.45	5.34
13	4.48	5.42	29	4.22	4.39	45	4.55	5.54
14	4.01	4.05	30	3.48	6.05	46	4.45	4.98
15	4.29	4.26	31	4.38	4.42	47	4.42	4.50
16	4.42	4.58	32	4.56	5.10			

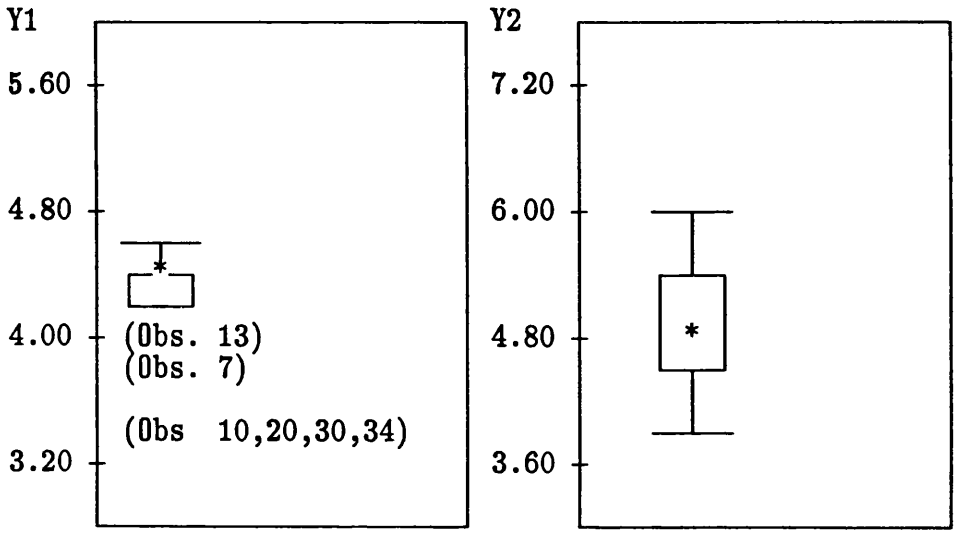
TABLE E.4(b) Summary Statistics for Hertzsprung-Russell Star Data

Statistic	Y_1	Y_2
Location		
Mean	4.31	4.99
Median	4.42	5.08
5% Trim	4.34	4.99
Std Err	0.04	0.09
Dispersion		
Variance	0.09	0.33
Std Dev	0.29	0.58
Min	3.48	3.94
Max	4.62	6.29
Range	1.14	2.35
IQR	0.18	0.89
Skewness & Kurtosis		
Skewness	-2.01	0.03
S E Skew	0.35	0.35
Kurtosis	3.38	-0.66
S E Kurt	0.69	0.69

TABLE E.4(c) M-Estimators

Statistic	Y_1	Y_2
Huber (1.34)	4.39	5.00
Hampel (1.70,3.40,8.50)	4.40	4.99
Tukey (4.69)	4.41	5.00
Andrew (1.34 * pi)	4.41	5.00

FIGURE E.4(a) Box Plots for Hertzsprung-Russell Star Data



Symbol Key: * - Median (...) - Outliers

EXAMPLE E.4 Hertzsprung–Russell Diagram of the Star Cluster CYG OB1 Analysis

In this data set the three measures of location (Table E.4(b)), the mean, median and 5% trimmed mean for both variables are in close agreement and with a small standard error. The coefficient of skewness value for Y_1 is also high and negative whereas that of Y_2 is almost zero. This means that Y_1 is has a long tail to the left and some unusual observations may exist but Y_2 is almost symmetric. The coefficient of kurtosis for Y_1 is close to the expected value but that of Y_2 is small and negative. There are negligible differences between the M–estimators. Graphically, the box plot for Y_1 confirms its asymmetry and shows a long left stem whereas Y_2 exhibits close symmetry. In particular, six observations (13, 7, 10, 20, 30 and 34) in the Y_1 space are detected as outliers and none in the Y_2 space.

The scatter plot in Figure E.4(b) is evenly distributed around an elliptical point cloud. There are four obvious unusual points which are distinctly far away from the point cloud in the Y_1 space with two more which are also not to near. Figure E.4(c) is the normal plot of the case deletion correlation coefficient function $Z(r_{-i})$. There are four points which jump and form an independent linear cluster from the one formed by the majority of the points. The influence of the four outlying observations is so strong that they make the correlation coefficient negative which can easily be refuted by inspecting the scatter plot. This explains why there is a jump in the observed $Z(r_{-i})$ when any of the four observations is deleted because the correlation then tends towards being positive.

From Table 2.1 all the discordancy tests apart from the T are not significant at the 5% level. This is explained by the fact that T is a form of range test and these outlying observations are distinctly far away from the rest of the data in the Y_1 direction.

The Stalactite Chart and Stalactite diagnostics are displayed in Figure E.4(d) and Table E.4(d) respectively. According to Figure E.4(c) there are two stalactites which have a depth of 100, there also a two more with a depth of 97.9. From the Stalactite Scores the most extreme observations are identified as observations 30 and 34, the other two are

observations 11 and 20. If a tolerance level $\tau = 0.05$ for the CIX is used it leads to two outliers in the data and a selection of 46 observations without observations 30 and 34 will ensure a "clean" data set at the 5% CIX tolerance level.

Figure E.4(e) displays the Means Plot for the data. The outlying observations clearly lie far from the majority of the data in the Y_1 space and so they do affect the central tendency by pulling the mean towards them so as the sub–sample size increases the outlying observations are include in the computations thus reducing the mean in this case. The mean for Y_2 starts off high but is "pulled" down as the sample size increases until a sub–sample size of 50% at which point it oscillates about the full sample mean. The reason for this is that the data is almost symmetric and so after 50% of it is used in the computations the observations are selected randomly since they all have similar distances from the centre of the point cloud. The compound means are lower than the full sample means in both variables indicating a negative pull for the mean in both dimensions. This suggests that the inclusion of some of the observations tends to pull the mean down and these are the observations with high Stalactite Scores. Referring to Figure 2.3 the "pull" falls within quadrant B i.e. with a pull vector $\tilde{P} = (-1, -1)$.

Table E.4(e) is a summary of the different multivariate results from the different approches. These are the case deletion correlation coefficient, the $Z(r_{-i})$, the diagonal elements of the Hat matrix, the Mahalanobis matrix and the Stalactite Scores. Apart from the two correlation related approaches, which fail to identify outlying observations, all the other approaches managed to detect and identify them.

Figure E.4(b) Hertzsprung–Russell Star Data
Scatter Plot

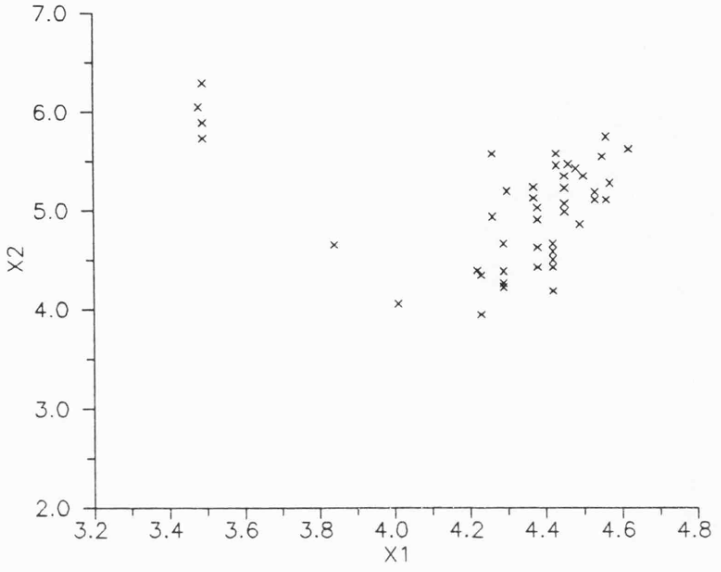


Figure E.4(c) Hertzsprung–Russell Star Data
Normal Plot of $Z(r_{-i})$

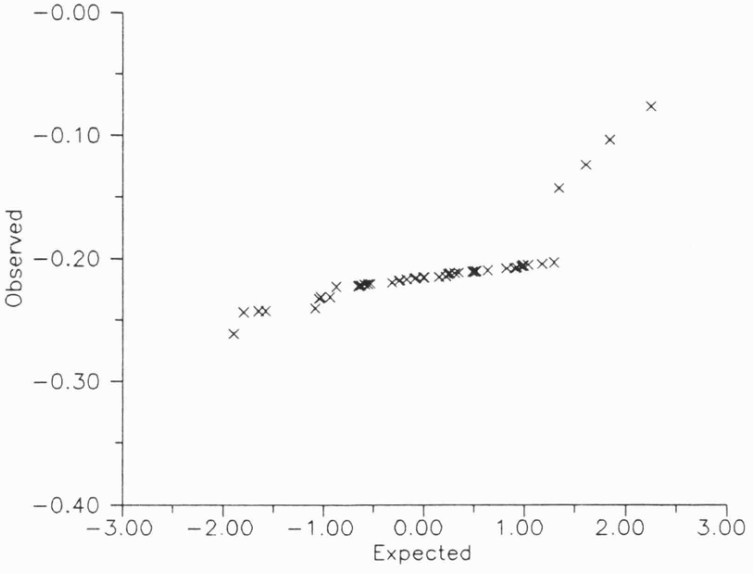


FIGURE E.4(d) Hertzsprung-Russell Star Data Stalactite Chart

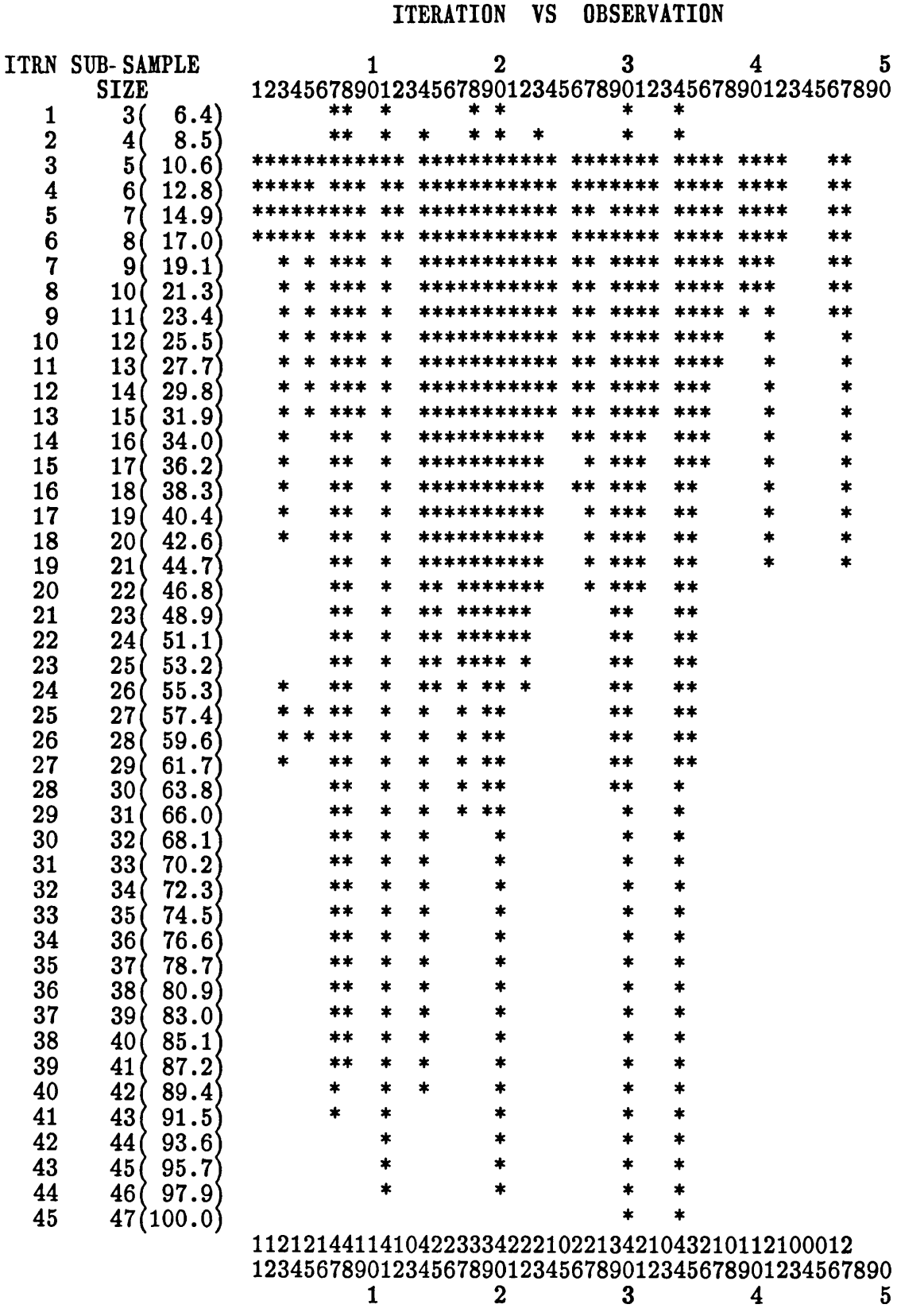


TABLE E.4(d) Hertzsprung-Russell Star Data Stalactite Analysis

ITRN	SUB-SAMPLE SIZE	OBSERVATION GOOD #(%)	OBSERVATION BAD #(%)	BAD:GOOD RATIO	TOTAL SQ. DISTANCE OBS.	TOTAL SQ. DISTANCE EXP.
1	3(6.4)	40(85.1)	7(14.9)	0.17	4.00	4.00
2	4(8.5)	38(80.9)	9(19.1)	0.24	0.38	6.00
3	5(10.6)	7(14.9)	40(85.1)	5.71	8.00	8.00
4	6(12.8)	9(19.1)	38(80.9)	4.22	10.00	10.00
5	7(14.9)	9(19.1)	38(80.9)	4.22	12.00	12.00
6	8(17.0)	9(19.1)	38(80.9)	4.22	14.00	14.00
7	9(19.1)	15(31.9)	32(68.1)	2.13	16.00	16.00
8	10(21.3)	15(31.9)	32(68.1)	2.13	18.00	18.00
9	11(23.4)	16(34.0)	31(66.0)	1.94	20.00	20.00
10	12(25.5)	18(38.3)	29(61.7)	1.61	22.00	22.00
11	13(27.7)	18(38.3)	29(61.7)	1.61	24.00	24.00
12	14(29.8)	19(40.4)	28(59.6)	1.47	26.00	26.00
13	15(31.9)	19(40.4)	28(59.6)	1.47	28.00	28.00
14	16(34.0)	23(48.9)	24(51.1)	1.04	30.00	30.00
15	17(36.2)	24(51.1)	23(48.9)	0.96	32.00	32.00
16	18(38.3)	24(51.1)	23(48.9)	0.96	34.00	34.00
17	19(40.4)	25(53.2)	22(46.8)	0.88	36.00	36.00
18	20(42.6)	25(53.2)	22(46.8)	0.88	38.00	38.00
19	21(44.7)	26(55.3)	21(44.7)	0.81	40.00	40.00
20	22(46.8)	29(61.7)	18(38.3)	0.62	42.00	42.00
21	23(48.9)	32(68.1)	15(31.9)	0.47	44.00	44.00
22	24(51.1)	32(68.1)	15(31.9)	0.47	45.93	46.00
23	25(53.2)	33(70.2)	14(29.8)	0.42	47.67	48.00
24	26(55.3)	33(70.2)	14(29.8)	0.42	49.98	50.00
25	27(57.4)	34(72.3)	13(27.7)	0.38	52.00	52.00
26	28(59.6)	34(72.3)	13(27.7)	0.38	54.00	54.00
27	29(61.7)	35(74.5)	12(25.5)	0.34	56.00	56.00
28	30(63.8)	37(78.7)	10(21.3)	0.27	58.00	58.00
29	31(66.0)	38(80.9)	9(19.1)	0.24	60.00	60.00
30	32(68.1)	40(85.1)	7(14.9)	0.17	62.00	62.00
31	33(70.2)	40(85.1)	7(14.9)	0.17	64.00	64.00
32	34(72.3)	40(85.1)	7(14.9)	0.17	66.00	66.00
33	35(74.5)	40(85.1)	7(14.9)	0.17	68.00	68.00
34	36(76.6)	40(85.1)	7(14.9)	0.17	70.00	70.00
35	37(78.7)	40(85.1)	7(14.9)	0.17	72.00	72.00
36	38(80.9)	40(85.1)	7(14.9)	0.17	74.00	74.00
37	39(83.0)	40(85.1)	7(14.9)	0.17	76.00	76.00
38	40(85.1)	40(85.1)	7(14.9)	0.17	78.00	78.00
39	41(87.2)	40(85.1)	7(14.9)	0.17	80.00	80.00
40	42(89.4)	41(87.2)	6(12.8)	0.15	82.00	82.00
41	43(91.5)	42(89.4)	5(10.6)	0.12	84.00	84.00
42	44(93.6)	43(91.5)	4(8.5)	0.09	86.00	86.00
43	45(95.7)	43(91.5)	4(8.5)	0.09	88.00	88.00
44	46(97.9)	43(91.5)	4(8.5)	0.09	90.00	90.00
45	47(100.0)	45(95.7)	2(4.3)	0.04	92.00	92.00

Figure E.4(e) Hertzsprung–Russell Star Data
Means Plot

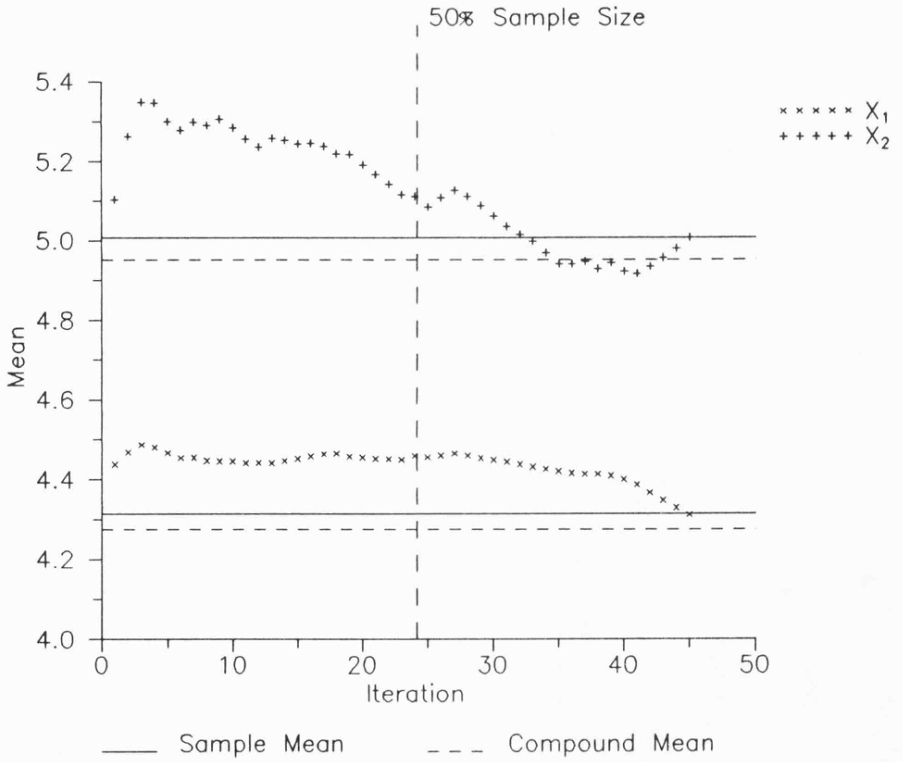


TABLE E.4(e) Case Deletion Correlation Coefficient, Diagonal Elements of the Hat Matrix, Mahalanobis Distances, and Stalactite Scores for the Hertzsprung Russell Star Data

Obs. i	r(-i)	z(-i)†	h_{ii} (0.085)	d_{ii} (2.71)	SS_i (4)
1	-0.210	-0.213	0.022	0.48	1
2	-0.239	-0.243	0.037	1.72	1
3	-0.208	-0.211	0.022	0.24	2
4	-0.239	-0.243	0.037	1.72	1
5	-0.208	-0.211	0.021	0.32	2
6	-0.219	-0.223	0.027	1.05	1
7	-0.237	-0.241	0.078	1.90	4
8	-0.220	-0.223	0.038	0.98	4
9	-0.209	-0.212	0.022	1.11	1
10	-0.142	-0.143	0.022	0.32	1
11	-0.217	-0.220	0.195	2.93	4
12	-0.219	-0.223	0.025	0.97	1
13	-0.256	-0.262	0.029	1.04	0
14	-0.214	-0.217	0.044	2.18	4
15	-0.203	-0.206	0.021	1.34	2
16	-0.228	-0.232	0.024	0.78	2
17	-0.203	-0.206	0.023	1.97	3
18	-0.228	-0.232	0.024	1.44	3
19	-0.201	-0.204	0.023	1.97	3
20	-0.124	-0.124	0.195	3.01	4
21	-0.212	-0.215	0.021	1.13	2
22	-0.214	-0.218	0.021	1.41	2
23	-0.202	-0.205	0.024	1.03	2
24	-0.205	-0.208	0.030	0.64	1
25	-0.208	-0.211	0.023	0.25	0
26	-0.204	-0.207	0.024	0.66	2
27	-0.209	-0.213	0.021	0.63	2
28	-0.207	-0.210	0.023	0.28	1
29	-0.218	-0.221	0.023	1.20	3
30	-0.104	-0.104	0.196	3.14	4
31	-0.205	-0.208	0.023	1.02	2
32	-0.213	-0.216	0.037	0.93	2
33	-0.213	-0.216	0.026	0.68	0
34	-0.077	-0.077	0.195	3.31	4
35	-0.218	-0.222	0.023	1.27	3
36	-0.239	-0.244	0.046	1.70	2
37	-0.212	-0.216	0.034	0.83	1
38	-0.213	-0.216	0.026	0.68	0
39	-0.215	-0.218	0.034	0.89	1
40	-0.220	-0.223	0.025	1.16	1
41	-0.205	-0.208	0.023	0.68	2
42	-0.210	-0.213	0.026	0.52	1
43	-0.218	-0.222	0.031	0.98	0
44	-0.215	-0.219	0.026	0.84	0
45	-0.229	-0.233	0.036	1.39	0
46	-0.208	-0.211	0.026	0.48	1
47	-0.203	-0.206	0.024	0.90	2

† $z[r(-i)] = (1/2)\log_e\{(1 + r_{-i})/(1 - r_{-i})\}$

Note: $h_{ii} > 0.085$ (=2p/n), distances d_{ii} exceeding "cutoff" value $\sqrt{\chi^2_2}(0.975) = 2.71$ and $SS_i = 4$ are underlined. Also, $\sqrt{\chi^2_2}(0.950) = 2.45$.

EXAMPLE E.5 Hawkins-Bradu-Kass Artificial Data. Table E.5(a) displays data generated by Hawkins, Bradu, and Kass [1984]. The artificial data consist of 75 observations in four dimensions (one response and three explanatory variables) but for the purposes of the analysis associated with the thesis consideration is restricted to the explanatory variables.

TABLE E.5(a) Hawkins-Bradu-Kass Artificial Data

Obs	Y_1	Y_2	Y_3	Obs	Y_1	Y_2	Y_3	Obs	Y_1	Y_2	Y_3
1	10.1	19.6	28.3	26	0.9	3.3	2.5	51	2.3	1.5	0.4
2	9.5	20.5	28.9	27	3.3	2.5	2.9	52	3.3	0.6	1.2
3	10.7	20.2	31.0	28	1.8	0.8	2.0	53	0.3	0.4	3.3
4	9.9	21.5	31.7	29	1.2	0.9	0.8	54	1.1	3.0	0.3
5	10.3	21.1	31.1	30	1.2	0.7	3.4	55	0.5	2.4	0.9
6	10.8	20.4	29.2	31	3.1	1.4	1.0	56	1.8	3.2	0.9
7	10.5	20.9	29.1	32	0.5	2.4	0.3	57	1.8	0.7	0.7
8	9.9	19.6	28.8	33	1.5	3.1	1.5	58	2.4	3.4	1.5
9	9.7	20.7	31.0	34	0.4	0.0	0.7	59	1.6	2.1	3.0
10	9.3	19.7	30.3	35	3.1	2.4	3.0	60	0.3	1.5	3.3
11	11.0	24.0	35.0	36	1.1	2.2	2.7	61	0.4	3.4	3.0
12	12.0	23.0	37.0	37	0.1	3.0	2.6	62	0.9	0.1	0.3
13	12.0	26.0	34.0	38	1.5	1.2	0.2	63	1.1	2.7	0.2
14	11.0	34.0	34.0	39	2.1	0.0	1.2	64	2.8	3.0	2.9
15	3.4	2.9	2.1	40	0.5	2.0	1.2	65	2.0	0.7	2.7
16	3.1	2.2	0.3	41	3.4	1.6	2.9	66	0.2	1.8	0.8
17	0.0	1.6	0.2	42	0.3	1.0	2.7	67	1.6	2.0	1.2
18	2.3	1.6	2.0	43	0.1	3.3	0.9	68	0.1	0.0	1.1
19	0.8	2.9	1.6	44	1.8	0.5	3.2	69	2.0	0.6	0.3
20	3.1	3.4	2.2	45	1.9	0.1	0.6	70	1.0	2.2	2.9
21	2.6	2.2	1.9	46	1.8	0.5	3.0	71	2.2	2.5	2.3
22	0.4	3.2	1.9	47	3.0	0.1	0.8	72	0.6	2.0	1.5
23	2.0	2.3	0.8	48	3.1	1.6	3.0	73	0.3	1.7	2.2
24	1.3	2.3	0.5	49	3.1	2.5	1.9	74	0.0	2.2	1.6
25	1.0	0.0	0.4	50	2.1	2.8	2.9	75	0.3	0.4	2.6

EXAMPLE E.5 Hawkins–Bradu–Kass Data Analysis

In all the three variables medians are significantly different from the means and the 5% trimmed means. The coefficient of skewness value for all three variables is also high and positive. This means that all the variables have long tails to the right and some unusual observations may exist. The interquartile range is far smaller than the standard deviation in all three cases. The coefficients of kurtosis for all the variables is small although that of Y_2 is slightly larger. There are negligible differences between the M–estimators. Graphically, all the box plots have no stems to the left since there are several outlying observations far removed from the median. In particular, there are fourteen observations (1,...,14) in all the variables space that are detected as outliers. It is worth noting that although the same 14 observations are detected in each of the variables their degree of outlyingness varies across variables.

The scatter plot matrix in Figure E.5(b) displays the three possible combinations of bivariate scatter plots for the variables. All the plots show the fourteen observations as outlying but as mentioned earlier the outlyingness varies in each. Figures E.5(c) i., E.5(c) ii. and E.5(c) iii. are the normal plots of the case deletion correlation coefficient function $Z(r_{-i})$ for the corresponding scatter plot matrix, namely, between Y_1 and Y_2, Y_1 and Y_3, Y_2 and Y_3 respectively. In the Y_1 and Y_2 space the normal plot is linear with a jump at a $Z(r_{-i})$ value of 1.77 from where it continues to be linear until a further jump by a single observation. In this space the the 14 outlying observations correspond to the low values for $Z(r_{-i})$. This means that deleting any of the 14 observations has the effect of reducing the correlation, a feature which is also visible from the scatter plot. The Y_1 and Y_3 space shows linearity with a jump at the $Z(r_{-i})$ value of 1.95. Again the 14 outlying observations correspond to the low values for $Z(r_{-i})$ and hence reduce correlation when any of them is deleted. The Y_2 and Y_3 space normal plot is very linear (with a small jump in the middle) apart from one extreme point. This time the 14 outlying observations do not have that strong an influence on the correlations when deleted as can also be verified from the scatter

plot.

From Table 2.1 all the discordancy tests are highly significant at the 5% level and so this suggests that there is strong evidence that there is at least one outlying observation with a strong possibility of several outliers.

The fact that all tests so far indicate that there is at least one outlier makes it useful to apply the multivariate tests for further investigation of the data together with obtaining the identities of these outlying observations.

The Stalactite Chart and Stalactite diagnostics are displayed in Figure E.5(d) and Table E.5(d) respectively. According to Figure E.5(d) there is one stalactite which has a depth of 100. The next deepest stalactite has a depth of 97.3. From the Stalactite Scores these observations are identified as observations 1 to 14. If a tolerance level $\tau = 0.05$ for the CIX is used it leads to three outliers in the data and a selection of 69 observations without the most outlying observations 12, 13 and 14 will ensure a "clean" data set at 5% CIX tolerance.

Figures E.5(e), E.5(f) and E.5(g) display the Mahalanobis index plot (MIP) of the data at the 50%, 90% and full sample sizes, respectively. At 50% and 90% sub–sample sizes both the $\chi^2(0.95)$ cut–off and the E[Max χ^2] detect all the fourteen outliers. Using the full sample $\chi^2(0.95)$ cut–off detects upto three outliers whereas the E[Max χ^2] cut–off points detects one.

Figures E.5(h), E.5(i) and E.5(j) show the 50%, 90% and full sub–sample sizes normal plots of the Mahalanobis distances transformed using the Wilson and Hilferty result (Section 2.2.1), respectively. The 50% sub–sample size plot is linear with a distinct jump. The observations in the cluster formed after the jump are actually the fourteen outlying observations. This exposes the masking in the data. This feature is not well detected at the 90% sub–sample. The full sample plot does not even suggest any possibility of masking, it only exposes the one most outlying observation (on identification it is observation 14). According to the data it appears that the observations detected as most outlying by the

Stalactite Analysis approach (observations 11,12,13,14) are the ones that were designed to be good leverage points. An explanation for this is that good leverage points are actually outlying observations in each of the variables simultaneously which could be regarded as joint outlyingness. Such outliers are like point A in Figure 2.1; they do not affect the correlations but inflate the variances of the variables simultaneaously.

Figure E.5(k) displays the Means Plot for the data. The means for all three variables are relatively stable until about iteration 57 when the outlying observations start becoming selected in the sub–sample at which point the means are "pulled" up sharply as the sample size increases. Also, the compound means are much higher than the full sample means in all the variables indicating a strong positive pull of the mean in each dimension. Refering to Figure 2.4 the "pull" would have all the bars above the zero line i.e. with a pull vector $\tilde{P} = (1, 1, 1)$.

The Stalactite Analysis algorithm selects the initial sub–sample size $m = p + 1$ randomly. If this initial sub–sample happens to come from an outlying cluster as in this data set then it would mean that some of the "good" observations (which lie within the majority point cloud) would have large distances and so be erroneously detected as outliers. Fortunately, if the sample size n is large compared to p then the sub–samples selected would always gravitate towards the majority point cloud. The worst possible case when gravitation does not take place, i.e. when the method breaks down, is as the proportion of contamination tends to 50% of the sample size. This is the *Breakdown Point* of the method.

Figure E.5(l) is an example of the behaviour of the method if the initial sub–sample is selected from an outlying cluster. In the first iteration all the fourteen outlying points are used in computing the means and covariance matrix to start the algorithm off. These fourteen observations are not detected and all the "good" observations are detected as outliers. As the iterations proceed the "bad" points start being detected until the 35th iteration (with a sub–sample size of 50.7%) when all the "bad" points are detected and the "good" ones cease. Figure E.5(m) shows that at about this iteration the variations of the

means are similar to those observed when the analysis was done with a random start. This example demonstrates the fact that there is no change in the Stalactite Chart after the 50% sub–sample whether the algorithm is initiated with a random or fixed start. There is, however, a change in the Stalactite Scores due to the fact that they contain the history of the presence/absence of the observations.

Table E.5(f) is a summary of the different multivariate results from the different approches. These are the the diagonal elements of the Hat matrix, the Mahalanobis distances, the Minimum Volume Ellipsoid (MVE) robust distances and the Stalactite Scores. The Hat matrix managed to detect three observations (12, 13 and 14), the Mahalanobis distance detected two (12 and 14) whereas both the Minimum Volume Estimator (MVE) robust distances and the Stalactite Scores detected all the fourteen "bad" observations.

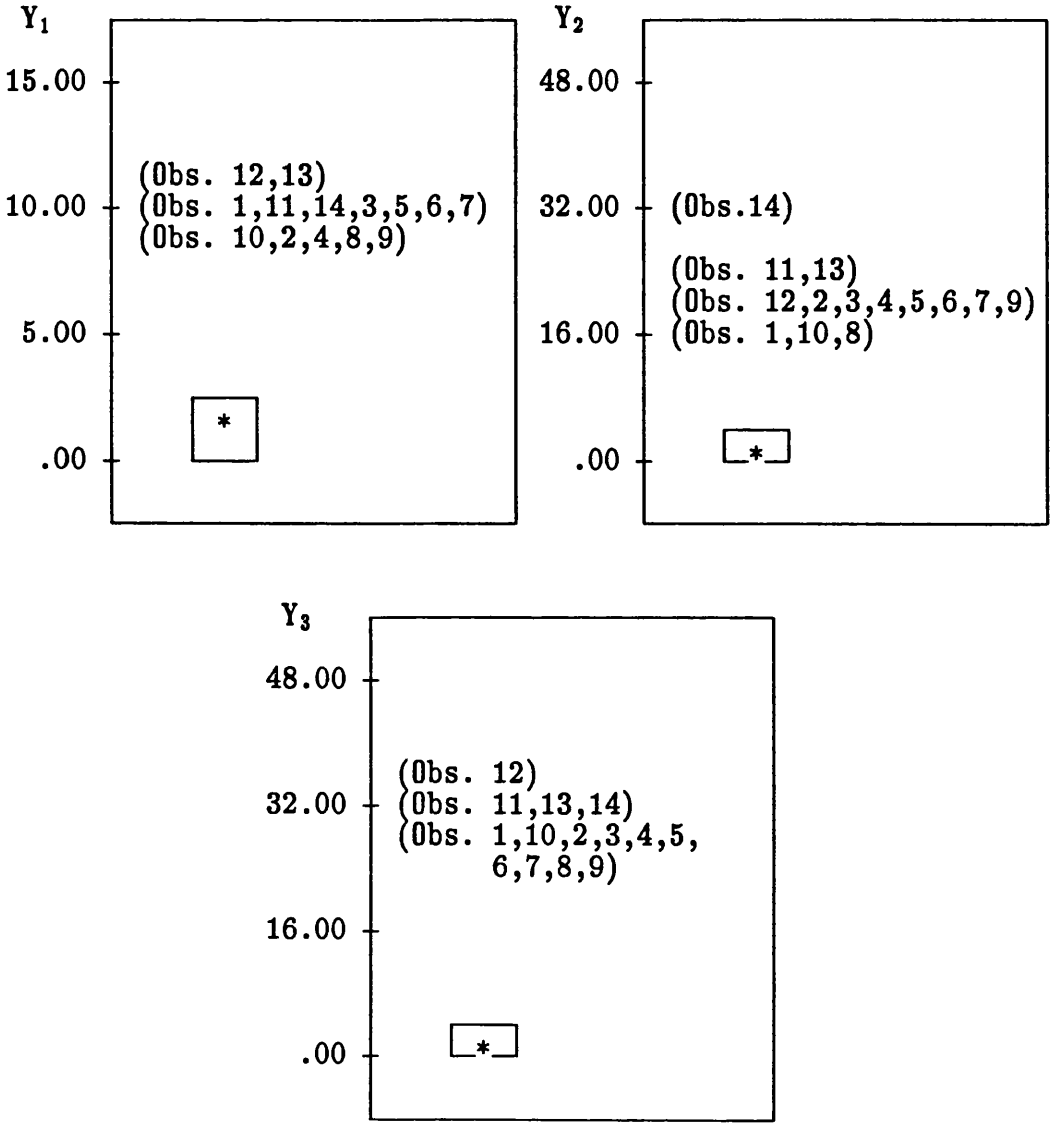
TABLE E.5(b) Summary Statistics for Hawkins, Bradu and Kass Artificial Data

Statistic	Y_1	Y_2	Y_3
Location			
Mean	3.21	5.60	7.23
Median	1.80	2.20	2.10
5% Trim	2.92	4.72	6.07
Std Err	0.42	0.95	1.36
Dispersion			
Variance	13.34	67.88	137.84
Std Dev	3.65	8.24	11.74
Min	0.00	0.00	0.20
Max	12.00	34.00	37.00
Range	12.00	34.00	36.80
IQR	2.30	2.30	2.10
Skewness & Kurtosis			
Skewness	1.42	1.77	1.65
S E Skew	0.28	0.28	0.28
Kurtosis	0.48	1.77	0.87
S E Kurt	0.55	0.55	0.55

TABLE E.5(c) M-Estimators

Statistic	Y1	Y2	Y3
Huber (1.34)	1.95	2.23	2.09
Hampel (1.70,3.40,8.50)	1.71	1.78	1.69
Tukey (4.69)	1.52	1.80	1.68
Andrew (1.34 * pi)	1.52	1.80	1.68

FIGURE E.5(a) Box Plots for Hawkins, Bradu and Kass Artificial Data



Symbol Key: * - Median (...) - Outliers

Figure E.5(b) Hawkins, Bradu and Kass Artificial Data
Scatter Plot Matrix

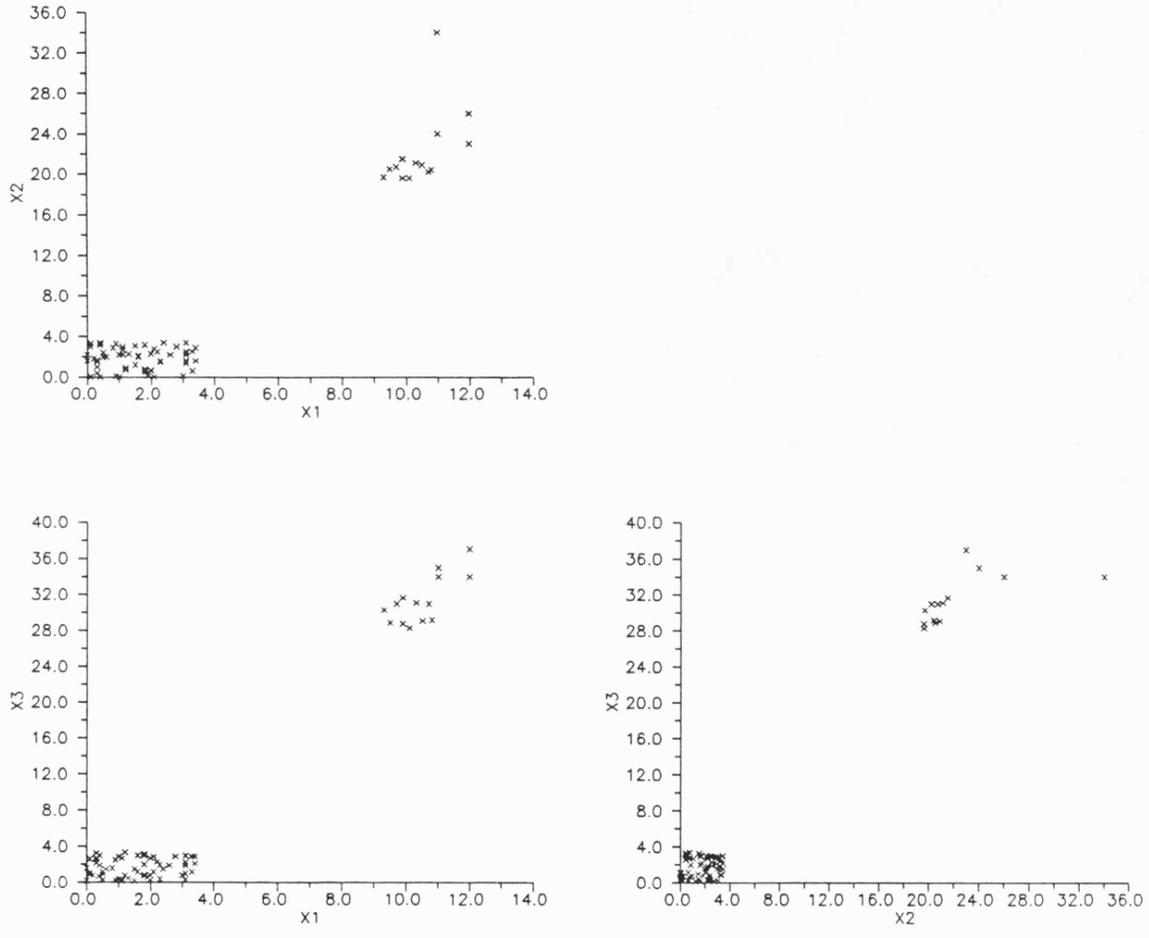


Figure E.5(c) Hawkins, Bradu and Kass Artificial Data
Normal Plot of $Z(r_{-i})$

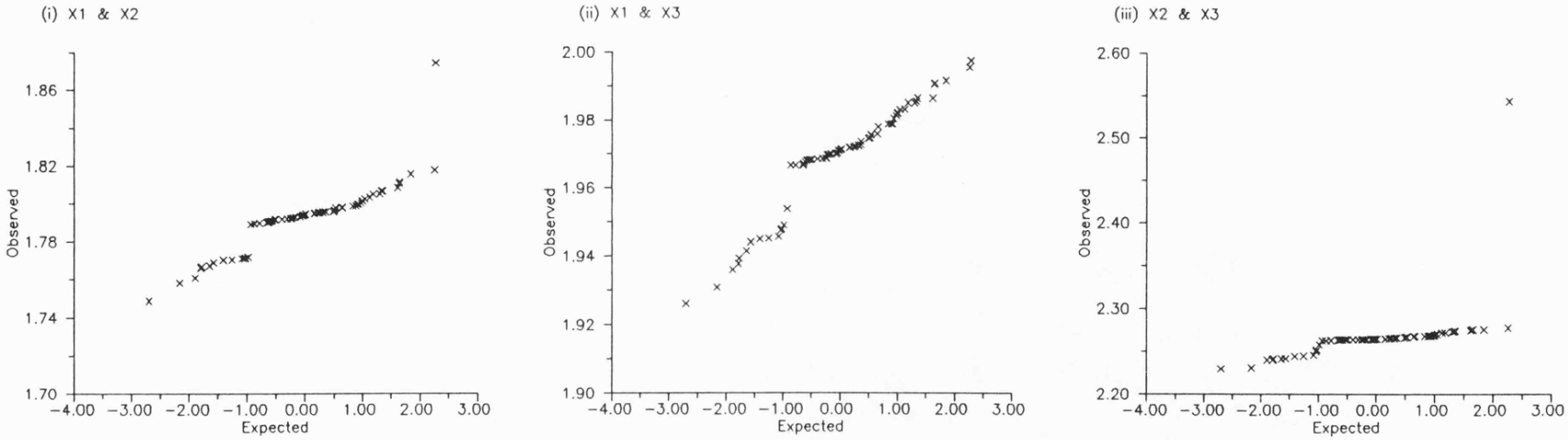


FIGURE E.5(d) Hawkins, Bradu and Kass Artificial Data Stalactite Chart

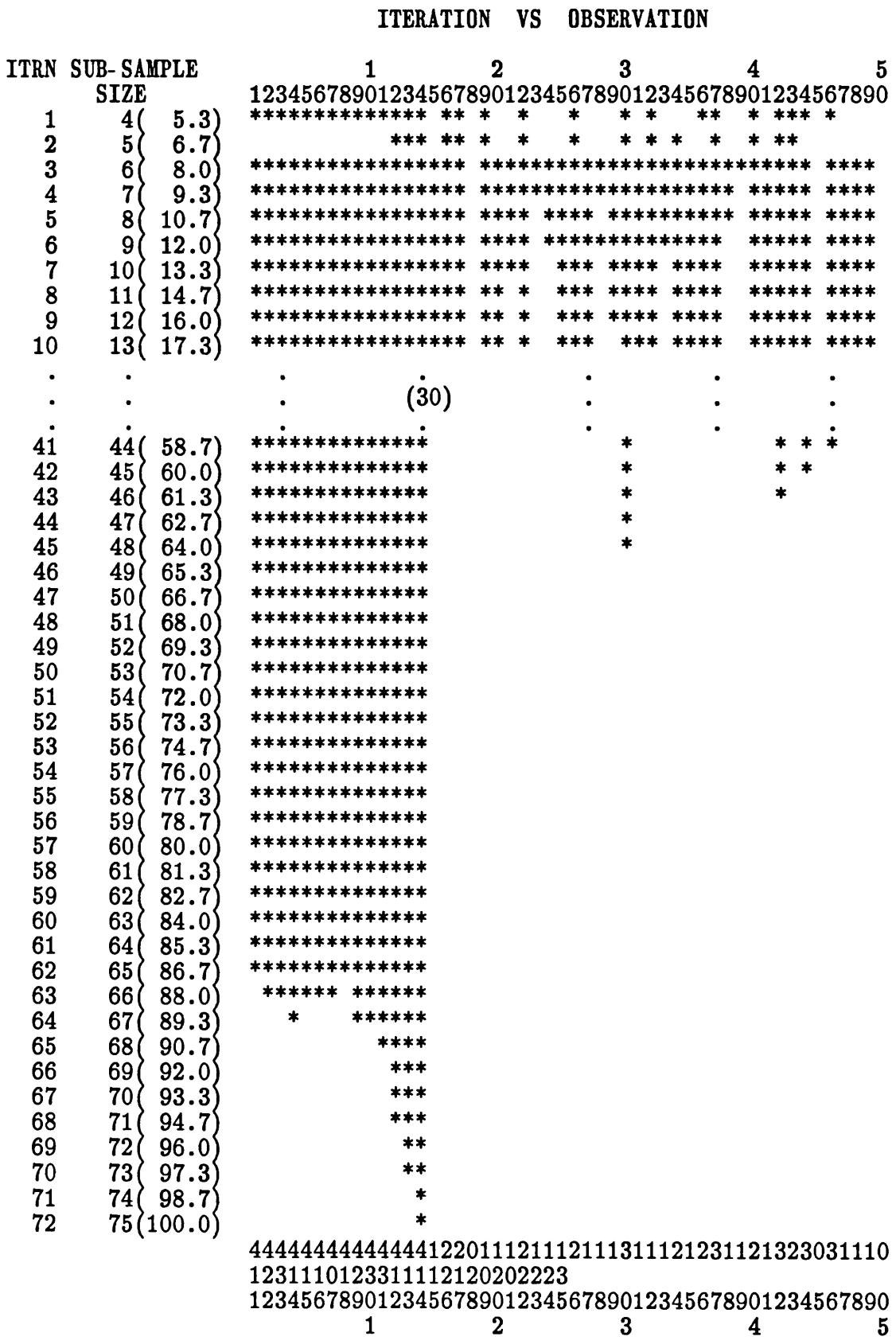


TABLE E.5(d) Hawkins, Bradu and Kass Artificial Data Stalactite Analysis

ITRN	SUB-SAMPLE SIZE	OBSERVATION GOOD #(%)	OBSERVATION BAD #(%)	BAD:GOOD RATIO	TOTAL SQ. DISTANCE OBS.	TOTAL SQ. DISTANCE EXP.
1	4(5.3)	33(44.0)	42(56.0)	1.27	9.00	9.00
2	5(6.7)	46(61.3)	29(38.7)	0.63	3.55	12.00
3	6(8.0)	6(8.0)	69(92.0)	11.50	15.00	15.00
4	7(9.3)	7(9.3)	68(90.7)	9.71	18.00	18.00
5	8(10.7)	11(14.7)	64(85.3)	5.82	21.00	21.00
6	9(12.0)	11(14.7)	64(85.3)	5.82	24.00	24.00
7	10(13.3)	15(20.0)	60(80.0)	4.00	27.00	27.00
8	11(14.7)	17(22.7)	58(77.3)	3.41	30.00	30.00
9	12(16.0)	19(25.3)	56(74.7)	2.95	33.00	33.00
10	13(17.3)	19(25.3)	56(74.7)	2.95	36.00	36.00
.	.	.	.	.	.	.
.	.	.	(30)	.	.	.
.	.	.	.	.	.	.
41	44(58.7)	54(72.0)	21(28.0)	0.39	129.00	129.00
42	45(60.0)	55(73.3)	20(26.7)	0.36	132.00	132.00
43	46(61.3)	56(74.7)	19(25.3)	0.34	135.00	135.00
44	47(62.7)	57(76.0)	18(24.0)	0.32	138.00	138.00
45	48(64.0)	57(76.0)	18(24.0)	0.32	141.00	141.00
46	49(65.3)	60(80.0)	15(20.0)	0.25	144.00	144.00
47	50(66.7)	60(80.0)	15(20.0)	0.25	147.00	147.00
48	51(68.0)	60(80.0)	15(20.0)	0.25	150.00	150.00
49	52(69.3)	60(80.0)	15(20.0)	0.25	153.00	153.00
50	53(70.7)	60(80.0)	15(20.0)	0.25	156.00	156.00
51	54(72.0)	60(80.0)	15(20.0)	0.25	159.00	159.00
52	55(73.3)	61(81.3)	14(18.7)	0.23	162.00	162.00
53	56(74.7)	61(81.3)	14(18.7)	0.23	165.00	165.00
54	57(76.0)	61(81.3)	14(18.7)	0.23	168.00	168.00
55	58(77.3)	61(81.3)	14(18.7)	0.23	171.00	171.00
56	59(78.7)	61(81.3)	14(18.7)	0.23	174.00	174.00
57	60(80.0)	61(81.3)	14(18.7)	0.23	177.00	177.00
58	61(81.3)	61(81.3)	14(18.7)	0.23	180.00	180.00
59	62(82.7)	61(81.3)	14(18.7)	0.23	183.00	183.00
60	63(84.0)	61(81.3)	14(18.7)	0.23	186.00	186.00
61	64(85.3)	61(81.3)	14(18.7)	0.23	189.00	189.00
62	65(86.7)	61(81.3)	14(18.7)	0.23	192.00	192.00
63	66(88.0)	63(84.0)	12(16.0)	0.19	195.00	195.00
64	67(89.3)	68(90.7)	7(9.3)	0.10	198.00	198.00
65	68(90.7)	71(94.7)	4(5.3)	0.06	201.00	201.00
66	69(92.0)	72(96.0)	3(4.0)	0.04	204.00	204.00
67	70(93.3)	72(96.0)	3(4.0)	0.04	207.00	207.00
68	71(94.7)	72(96.0)	3(4.0)	0.04	210.00	210.00
69	72(96.0)	73(97.3)	2(2.7)	0.03	213.00	213.00
70	73(97.3)	73(97.3)	2(2.7)	0.03	216.00	216.00
71	74(98.7)	74(98.7)	1(1.3)	0.01	219.00	219.00
72	75(100.0)	74(98.7)	1(1.3)	0.01	222.00	222.00

Figure E.5(e) Hawkins, Bradu and Kass Artificial Data (50% Sample)
Index Plot of the Squared Mahalanobis Distances (MIP)

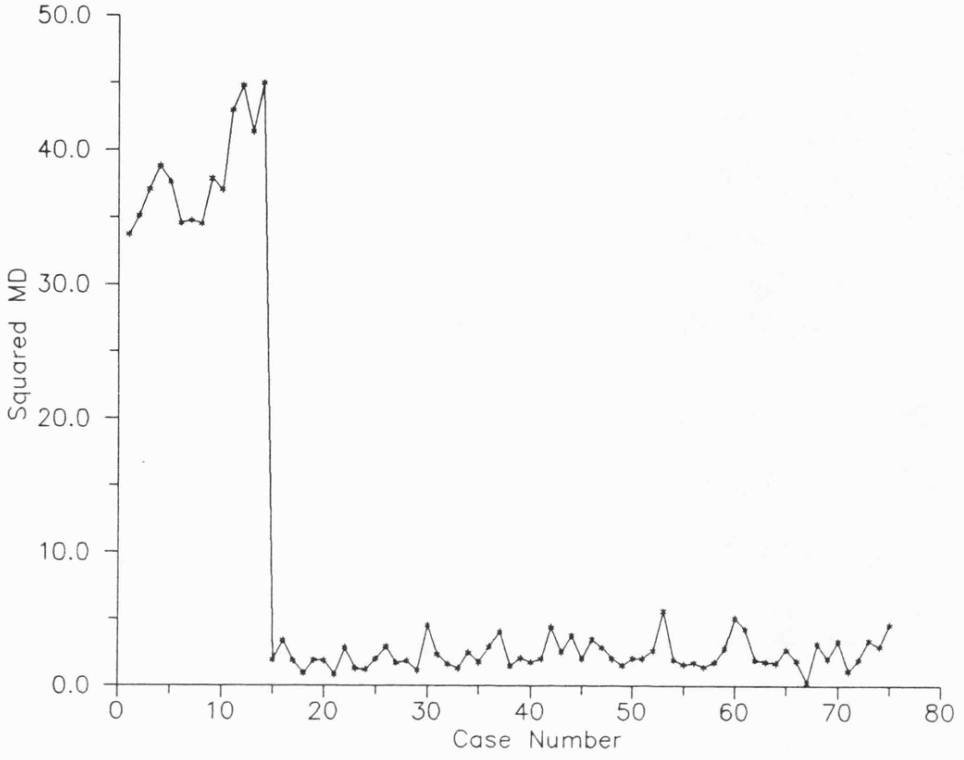


Figure E.5(f) Hawkins, Bradu and Kass Artificial Data (90% Sample)
Index Plot of the Squared Mahalanobis Distances (MIP)

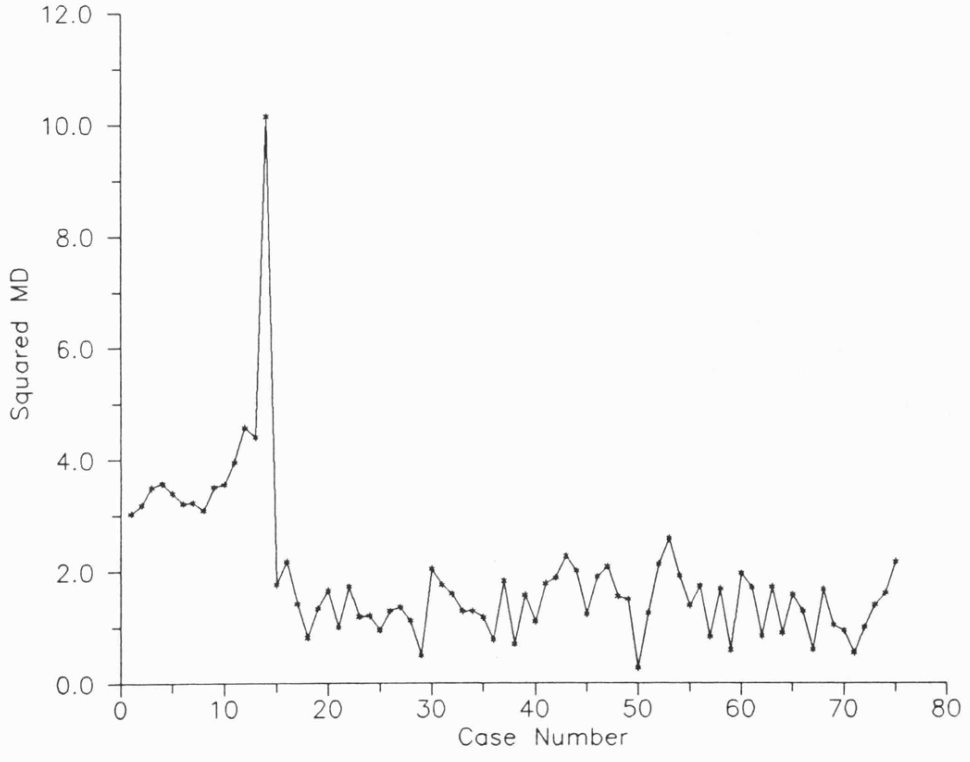


Figure E.5(g) Hawkins, Bradu and Kass Artificial Data (Full Sample)
Index Plot of the Squared Mahalanobis Distances (MIP)

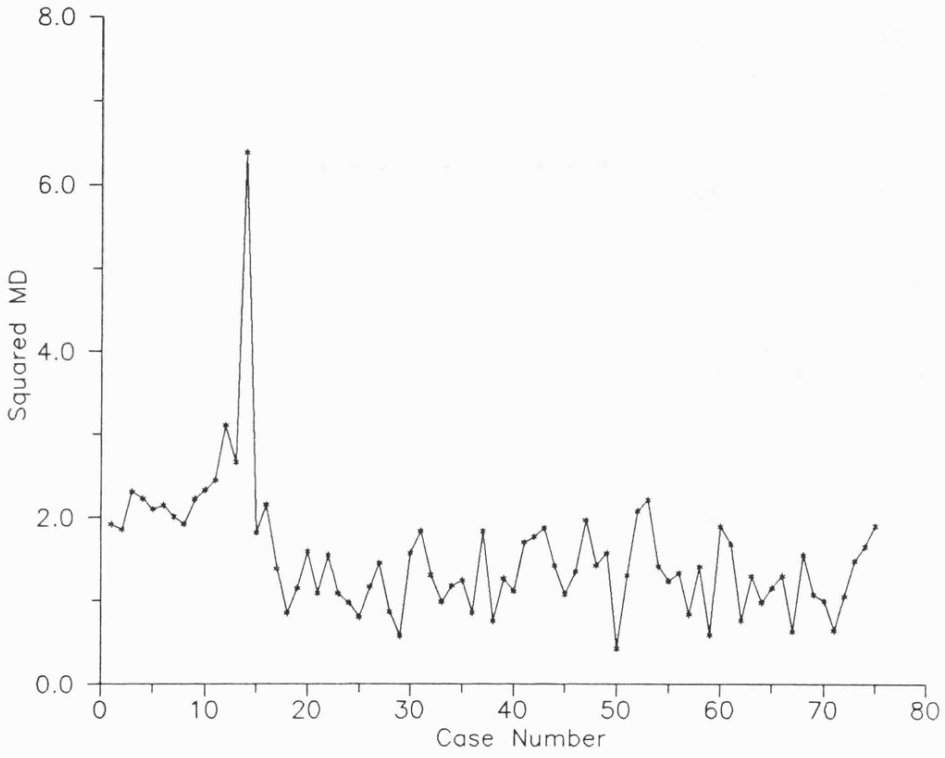


Figure E.5(h) Hawkins, Bradu and Kass Artificial Data (50% Sample)
Normal Plot of cube−root(MD^2)

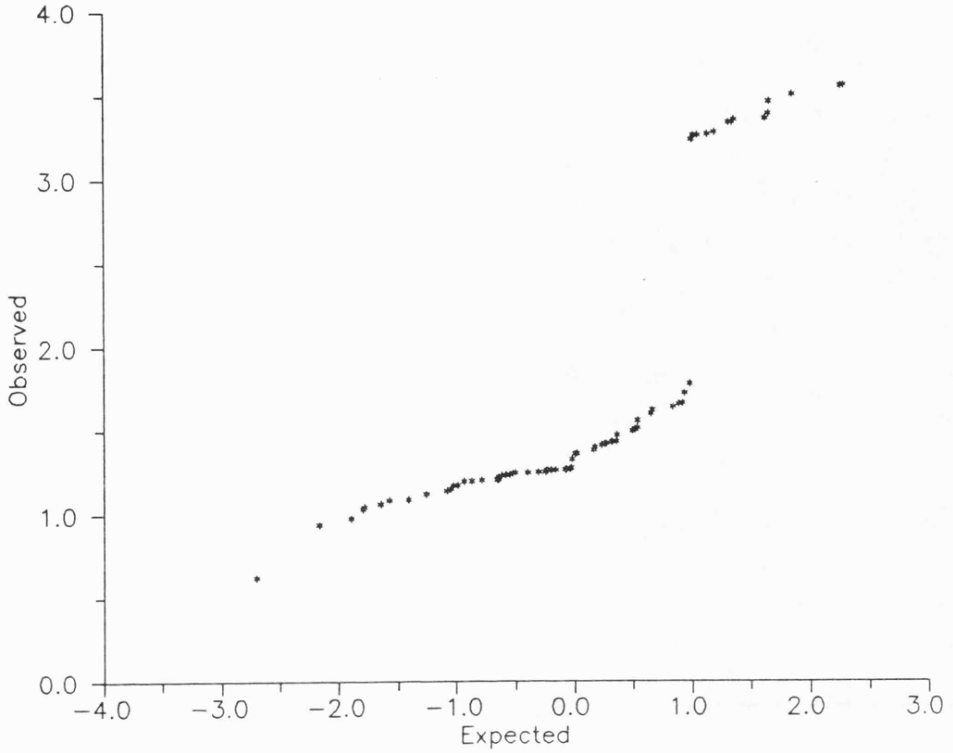


Figure E.5(i) Hawkins, Bradu and Kass Artificial Data (90% Sample)
Normal Plot of cube−root(MD^2)

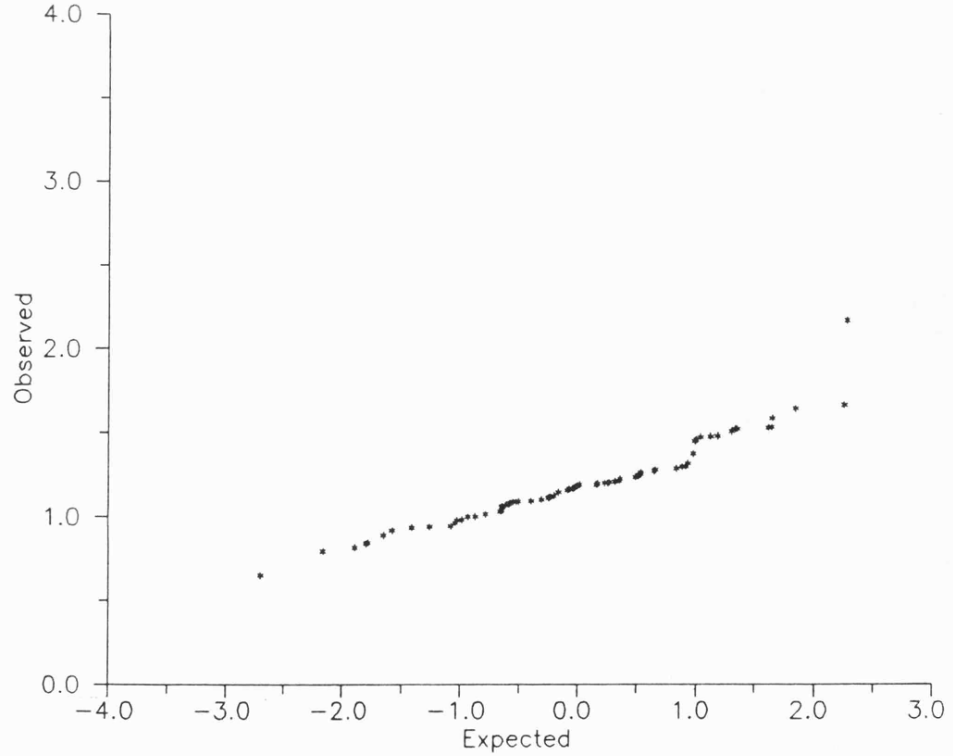


Figure E.5(j) Hawkins, Bradu and Kass Artificial Data (Full Sample)
Normal Plot of cube−root(MD^2)

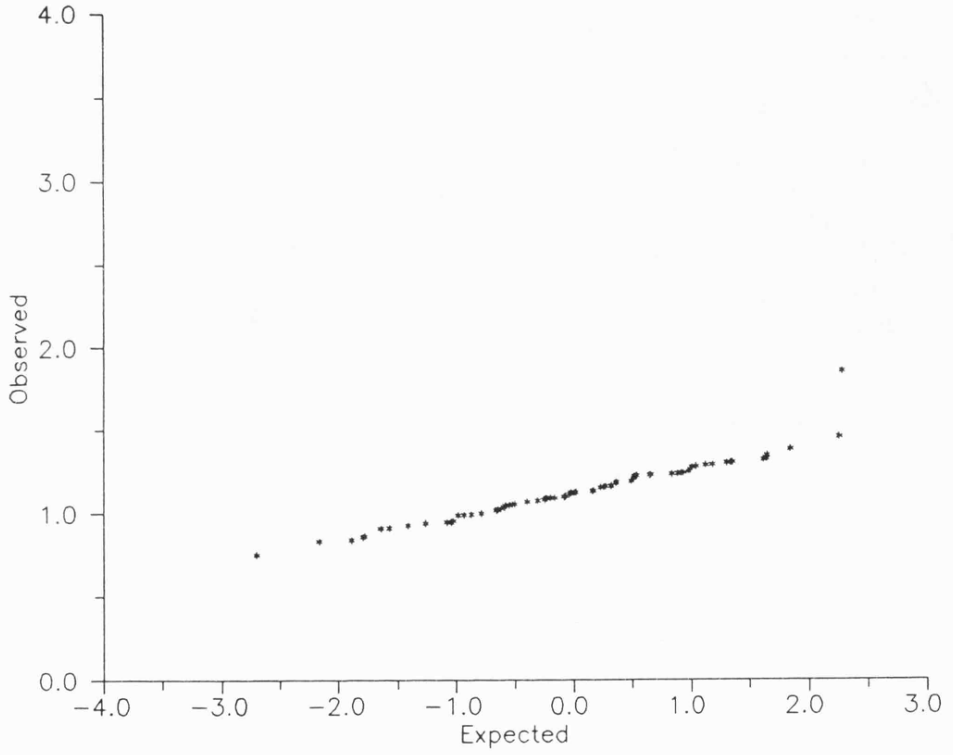


Figure E.5(k) Hawkins, Bradu and Kass Artificial Data
Means Plot

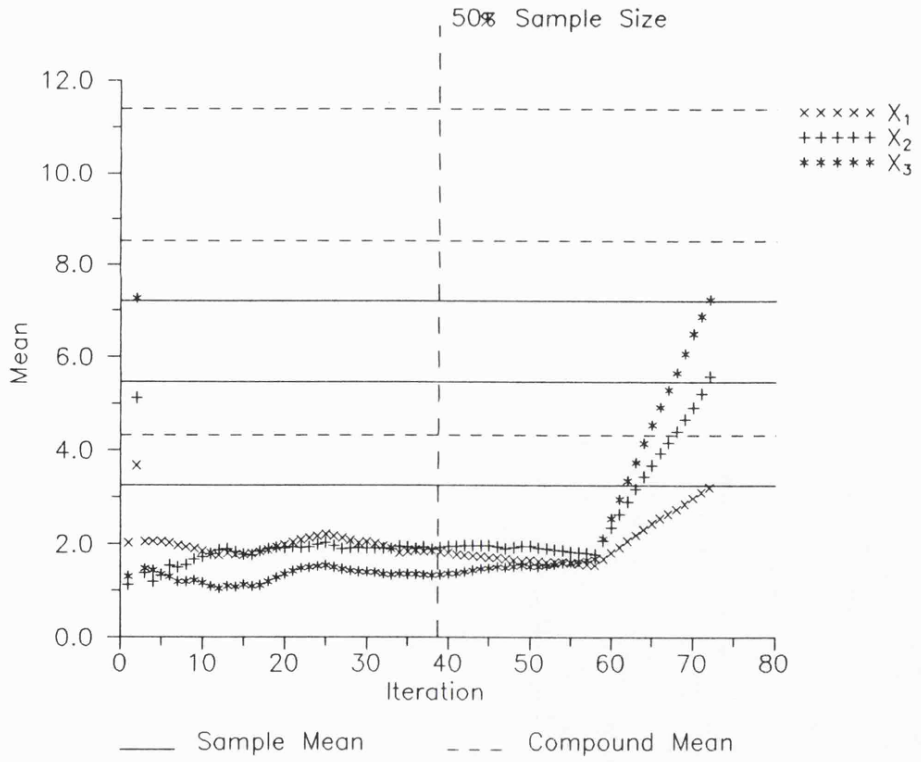


FIGURE E.5(1) Hawkins, Bradu and Kass Artificial Data Stalactite Chart
(with initial subsample of 14 observations)

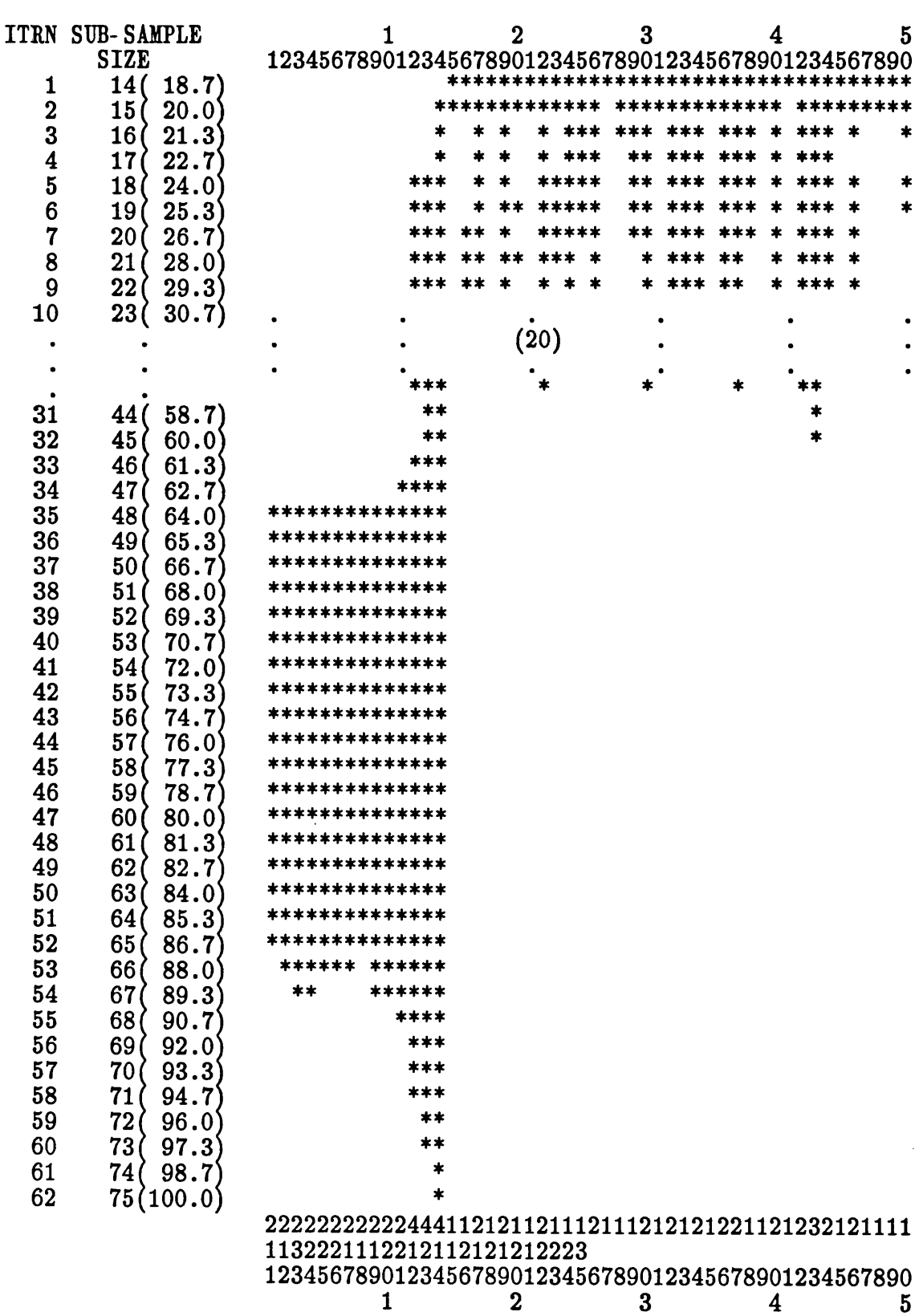


TABLE E.5(e) Hawkins, Bradu and Kass Artificial Data Stalactite Analysis

ITRN	SUB-SAMPLE SIZE	OBSERVATION GOOD #(%)	OBSERVATION BAD #(%)	BAD:GOOD RATIO	TOTAL SQ. DISTANCE OBS.	TOTAL SQ. DISTANCE EXP.
1	14(18.7)	14(18.7)	61(81.3)	4.36	42.00	39.00
2	15(20.0)	15(20.0)	60(80.0)	4.00	44.13	42.00
3	16(21.3)	35(46.7)	40(53.3)	1.14	41.49	45.00
4	17(22.7)	38(50.7)	37(49.3)	0.97	44.04	48.00
5	18(24.0)	33(44.0)	42(56.0)	1.27	51.96	51.00
6	19(25.3)	32(42.7)	43(57.3)	1.34	55.50	54.00
7	20(26.7)	33(44.0)	42(56.0)	1.27	57.94	57.00
8	21(28.0)	36(48.0)	39(52.0)	1.08	62.08	60.00
9	22(29.3)	38(50.7)	37(49.3)	0.97	65.33	63.00
10	23(30.7)	37(49.3)	38(50.7)	1.03	68.49	66.00
.	.	.	.	.	.	.
.	.	.	(20)	.	.	.
.	.	.	.	.	.	.
31	44(58.7)	69(92.0)	6(8.0)	0.09	132.00	129.00
32	45(60.0)	70(93.3)	5(6.7)	0.07	134.97	132.00
33	46(61.3)	70(93.3)	5(6.7)	0.07	134.78	135.00
34	47(62.7)	69(92.0)	6(8.0)	0.09	121.42	138.00
35	48(64.0)	60(80.0)	15(20.0)	0.25	142.92	141.00
36	49(65.3)	60(80.0)	15(20.0)	0.25	147.00	144.00
37	50(66.7)	60(80.0)	15(20.0)	0.25	150.00	147.00
38	51(68.0)	60(80.0)	15(20.0)	0.25	153.00	150.00
39	52(69.3)	60(80.0)	15(20.0)	0.25	156.00	153.00
40	53(70.7)	60(80.0)	15(20.0)	0.25	159.00	156.00
41	54(72.0)	60(80.0)	15(20.0)	0.25	162.00	159.00
42	55(73.3)	61(81.3)	14(18.7)	0.23	165.00	162.00
43	56(74.7)	61(81.3)	14(18.7)	0.23	168.00	165.00
44	57(76.0)	61(81.3)	14(18.7)	0.23	171.00	168.00
45	58(77.3)	61(81.3)	14(18.7)	0.23	174.00	171.00
46	59(78.7)	61(81.3)	14(18.7)	0.23	177.00	174.00
47	60(80.0)	61(81.3)	14(18.7)	0.23	180.00	177.00
48	61(81.3)	61(81.3)	14(18.7)	0.23	183.00	180.00
49	62(82.7)	61(81.3)	14(18.7)	0.23	186.00	183.00
50	63(84.0)	61(81.3)	14(18.7)	0.23	189.00	186.00
51	64(85.3)	61(81.3)	14(18.7)	0.23	192.00	189.00
52	65(86.7)	61(81.3)	14(18.7)	0.23	195.00	192.00
53	66(88.0)	63(84.0)	12(16.0)	0.19	198.00	195.00
54	67(89.3)	67(89.3)	8(10.7)	0.12	201.00	198.00
55	68(90.7)	71(94.7)	4(5.3)	0.06	204.00	201.00
56	69(92.0)	72(96.0)	3(4.0)	0.04	207.00	204.00
57	70(93.3)	72(96.0)	3(4.0)	0.04	210.00	207.00
58	71(94.7)	72(96.0)	3(4.0)	0.04	213.00	210.00
59	72(96.0)	73(97.3)	2(2.7)	0.03	216.00	213.00
60	73(97.3)	73(97.3)	2(2.7)	0.03	219.00	216.00
61	74(98.7)	74(98.7)	1(1.3)	0.01	222.00	219.00
62	75(100.0)	74(98.7)	1(1.3)	0.01	225.00	222.00

Figure E.5(m) Hawkins, Bradu and Kass Artificial Data
Means Plot (Non-random start)

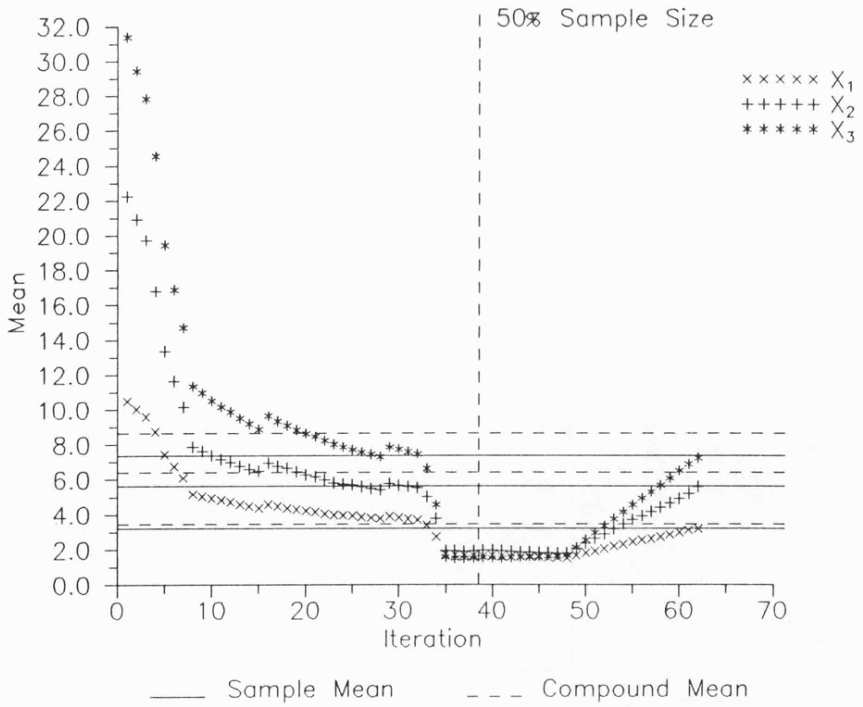


TABLE E.5(f) Diagonal Elements of the Hat Matrix, Squared Mahalanobis Distances, MVE Robust Distances and Stalactite Scores for the Hawkins-Bradu-Kass Artificial Data

Obs. i	h_{ii} (0.107)	d_{ii} (3.06)	RD_i (3.06)	SS_i (4)
1	0.063	1.92	16.20	4
2	0.060	1.86	16.62	4
3	0.086	2.31	17.65	4
4	0.081	2.23	18.18	4
5	0.073	2.10	17.82	4
6	0.076	2.15	16.80	4
7	0.068	2.01	16.82	4
8	0.063	1.92	16.44	4
9	0.080	2.22	17.71	4
10	0.087	2.33	17.21	4
11	0.094	2.45	20.23	4
12	0.144	3.11	21.14	4
13	0.109	2.66	20.16	4
14	0.564	6.38	22.38	4

Note: $h_{ii} > 0.107$ (=2p/n), distances d_{ii} and RD_i exceeding "cutoff" value $\sqrt{\chi^2_3}(0.975) = 3.06$ and $SS_i = 4$ are underlined. Also, $\sqrt{\chi^2_3}(0.950) = 2.80$. Only the first 14 observations listed.

CHAPTER THREE

3.0 TRANSFORMATIONS TO MULTIVARIATE NORMALITY

3.1 Introduction

In Chapter Two we dealt with the problem of detecting and identifying outliers in multivariate data sets. Having identified the outliers (if any) and appropriate corrective measures having been taken it is then possible to investigate the distributional properties of the data taking into account any deficiencies it may have. In particular, since interest is in the multivariate normal distribution, the conformity of the data to this distribution can now be tested and where necessary appropriate transformations can be carried out to normalise them (or at least obtain a unimodal, symmetrical distribution in the p–dimensional space). The aim of this chapter is to present some techniques for assessing the violation of the normality assumption together with providing computational methods for transformations to multivariate normality. The main result of the chapter is the presentation of the proposed computational procedure to perform transformations to multivariate normality. This procedure is based on "seemingly unrelated regressions" and "constructed variables". The technique is referred to as the *Seemingly Unrelated Regressions/Constructed Variable* (*SURCON*) analysis, and the estimates obtained are the *Surcon estimates.*

This section discusses the need for transformations to multivariate normality due to the central role of normality in the theory of multivariate analysis. It is often informative to study the symmetry of each variable as this provides some intuition as to how the data would behave jointly. In section 2, the assessment of marginal symmetry is discussed using graphical methods including a quick computational technique for transforming to marginal symmetry. Section 3 describes the likelihood approach to obtaining the joint transformations to multivariate normality. The main result of this chapter is the proposed SURCON method which is discussed in section 4 where the first two subsections describe the theory on which it is based, namely, that of seemingly unrelated regressions and constructed variables, and how these are combined to obtain it. The full computational

algorithm is also given. Section 5 discusses ways of assessing normality in both the univariate and multivariate cases. The main test for multivariate normality used is based on Rao's Score test [Mardia et al., 1991] so it is described in detail. Finally, section 6 demonstrates most of the theory discussed by using examples of some well known data sets. Simulated data sets are also used to study the expected behaviour of the techniques under known and predetermined conditions.

The classical multivariate theory has been largely based on the multivariate normal distribution (MVN): the scarcity of alternative models for the meaningful and consistent analysis of multiresponse data is a well recognised problem. Further, the complexity of generalising many non–normal univariate distributions makes it undesirable or impossible to use their multivariate versions. Hence, it seems reasonable to inquire about ways of transforming the data so as to enable the use of more familiar statistical techniques that are based implicitly or explicitly on the normal distribution. On the other hand, in situations where the sample size is large and the techniques depend solely on the behaviour of the mean vector, or distances involving it, the assumption of normality for the individual observations is less crucial. However, to some degree, the quality of inferences made by these methods depends on how closely the true parent population resembles the multivariate normal form. It is imperative, then, that procedures exist for detecting cases where the data exhibit moderate to extreme departures from what is expected under multivariate normality and also to adapt the data to conform to the normality model.

Some of the questions to be addressed in studying departures from the normality assumptions are [Johnson & Wichern, 1982: p.151]:

1/ Do the marginal distributions of the variables appear to be normal? What about a few linear combinations of these?

2/ Do the scatter plots of pairs of observations give the elliptical appearance expected from the normal population?

3/ Are there any "wild" observations (outliers) that should be checked for accuracy?

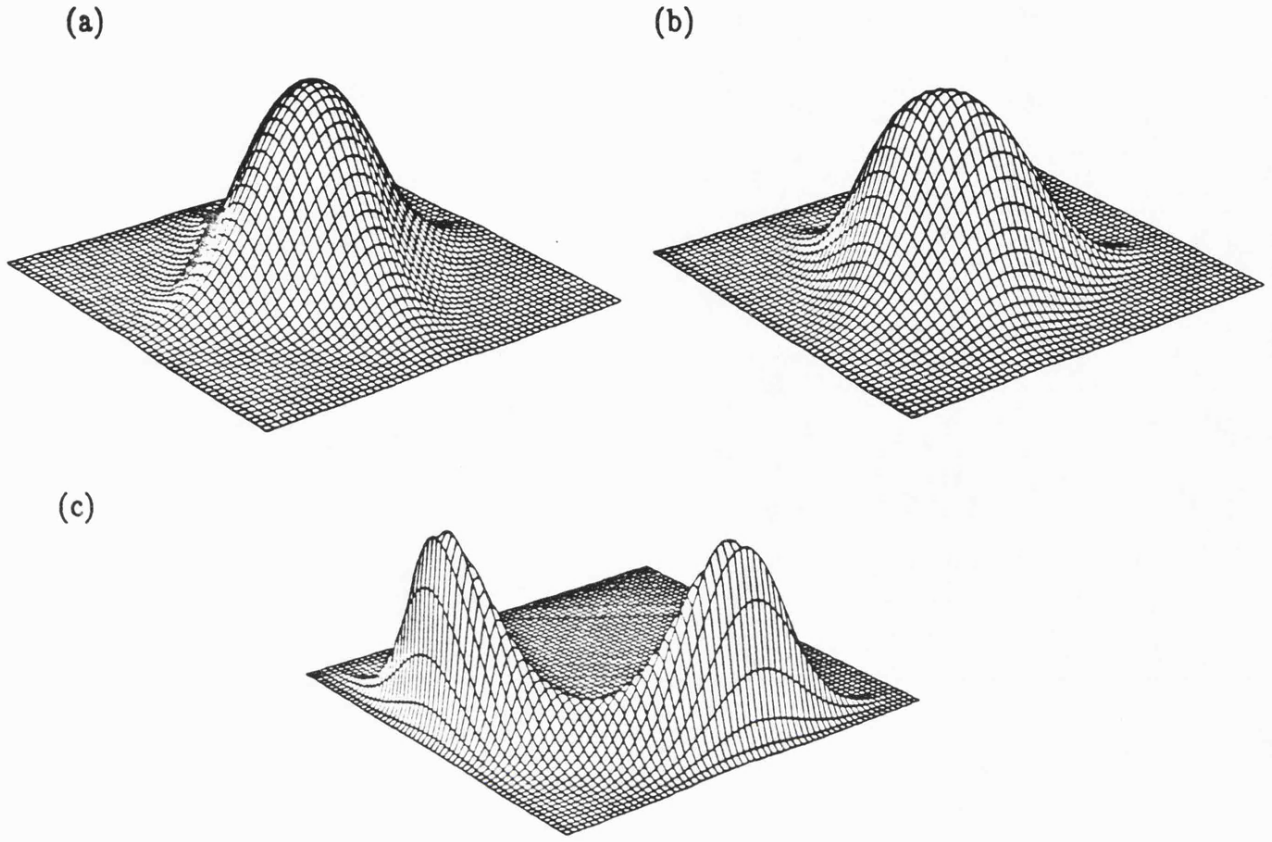
This chapter attempts to provide answers to questions one 1 and 2, whereas,

Chapter 2 answers question 3.

In practice investigations of normality concentrate on the behaviour of the observations in one or two dimensions (for example, marginal distributions and scatter plots) since it can prove difficult to have an overall test (especially graphically) of joint normality in more than three dimensions because of the large number of things that can go wrong. However, for a complete analysis all the information within the data relating to the interdependencies which exist between the variables should be used. This can be achieved by the inclusion of the covariance matrix within the analysis. It is also generally true that the marginal normality of variables does not necessarily imply their joint normality. The plots in Figure 3.1 are from Gelman & Meng [1991] and illustrate three cases where the variables are conditionally normal but jointly non–normal. In Figure 3.1(a) we have a joint density with zero conditional means that differ from the Gaussian by having non–constant conditional variances. Its joint density function is of the form $f(x_1, x_2) \propto \exp(-\frac{1}{2}[x_1^2x_2^2 + x_2^2])$. Figure 3.1(b) has a conditional distribution $(x_1|x_2) \longrightarrow N[1/(x_2^2 + 1), 1/(x_2^2 + 1)]$ and vice–versa, so the conditional mean equals the conditional variance at all points. Its joint density function is of the form $f(x_1, x_2) \propto \exp(-\frac{1}{2}[x_1^2x_2^2 + x_1^2 + x_2^2 - 2x_1 - 2x_2])$. Figure 3.1(c) is the most interesting in that the marginals are normal yet the joint density is clearly non–normal and is bimodal.

A transformation may be based on theoretical considerations or estimated from the data that are being analysed. Examples of the former are the logistic transformation of binary data proposed by Cox [1970, 1972] and the variance–stabilising transformations for the binomial, the Poisson, the correlation coefficient, etc. There are also several empirical indications as to whether a transformation may be useful. One indication is if the variable is non–negative e.g. the times until an event occurs and the measured diameters of particles are both non–negative and so cannot strictly follow a normal distribution. In these cases it is quite likely that the log of the variables will be more normally distributed

Figure 3.1 Bivariate Densities



than the variables themselves. If all the values of the variables are far from zero and the scatter in the observations is relatively small, the transformation will have little effect e.g. the heights of adult men in millimeters can be modelled by either a normal or by the lognormal distribution. If, however, the ratio of the largest observation to the smallest one is one or more powers of ten, so that the data covers several cycles, a transformation is often desirable [Atkinson, 1985]. If the data are counts then they can often be made more normal by taking their square roots. Table 3.1 shows some helpful transformations to near normality based on theoretical considerations. These are based on theoretical considerations on the variable under study.

TABLE 3.1 Helpful Transformations to Near Normality

Original Scale	Transformed Scale
1. Counts, y	$\sqrt{y}$
2. Proportions, $\hat{p}$	$\text{logit}(\hat{p}) = \frac{1}{2}\log\left\{\frac{\hat{p}}{1-\hat{p}}\right\}$
3. Correlations, r	Fisher's $z(r) = \frac{1}{2}\log\left\{\frac{1+r}{1-r}\right\}$

In instances where the choice of a transformation to improve the approximation to normality is not obvious it is convenient to let the data suggest a transformation. The family of transformations considered in this thesis is the Power Transformations. These are defined only for positive variables. However, a single constant can be added to each observation in the data set if some values are negative. A sequence of possible transformations for a variable y is

$$\underbrace{\ldots,\ y^{-1} = \frac{1}{y},\ y^{0} = \log_e y,\ y^{1/4} = \sqrt[4]{y},\ y^{1/2} = \sqrt{y},}_{\text{Shrinks large values of y}}\quad \underbrace{y^{1}}_{\text{No transformation required}},\quad \underbrace{y^{2},\ y^{3},\ \ldots}_{\text{Increases large values of y}}$$

Techniques for developing data–based transformations of univariate observations have been proposed by several authors. However, there is only one major technique in the multivariate (p–variable) case by Andrews et al [1971]. Their approach extended the power transformations proposed by Box & Cox [1964] to the problem of estimating power transformations of multiresponse data so as to enhance joint normality. The approach estimates the vector of transformation parameters λ by numerically maximising the log–likelihood function. However, since there are several parameters to be estimated, $p(p+5)/2$ for multivariate data without regression, the resulting maximisation is of high dimension, even with modest values of p and sample size n. The aim of the proposed method is to provide a complementary procedure to the log–likelihood approach which

attempts to reduce the size of the computational requirements for obtaining the estimates of λ. Though computational simplicity is the main factor, the statistical qualities of the estimates are not compromised, indeed the estimated values are numerically identical to those of the log–likelihood. Further, the procedure implicitly produces diagnostic statistics and some useful statistical quantities describing the structure of the data. The technique is a generalisation of the constructed variables method of obtaining quick estimates for transformation parameters [Atkinson, 1985]. To take into account the multiresponse nature of the data and, hence, joint estimates for λ, a seemingly unrelated regression is carried out. The algorithm is iterative. However, there is considerable savings in the number of iterations required to converge to the maximum likelihood (MLE) estimates compared to those using the log–likelihood function.

To begin with, then, the next section discusses the assessment and transformations to marginal symmetry.

3.2 Marginal Symmetry

Means and covariances (which provide the basic summary statistics for all multivariate procedures) may have little meaning unless the underlying distributions are symmetric and not too platykurtic. Moreover, the notion of an atypical or unduly influential observation (gross error or outlier) only makes sense when some form of reference distribution is assumed. Since many of the multivariate techniques assume multivariate normality or at least a symmetric distribution it is appropriate to investigate the validity of this assumption. This section deals with both the assessment of symmetry using graphical techniques and selecting transformations to obtain approximate symmetry by considering the variables marginally.

3.2.1 Graphical Assessment of symmetry

Graphical output is very useful in studying the structure of a data set. In particular, a histogram or frequency polygon is a useful visual tool in assessing the symmetry of a variable. However, for a rigourous assessment, numerical procedures for examining symmetry should be adopted. One such procedure can be based on the simple idea that for

symmetry the average of pairs of ordered observations y[i] and y[n–i+1] should remain constant for all i. That is

$$\frac{(y[i] + y[n-i+1])}{2} = K \qquad (3.1)$$

where y[i] is the i–th ordered observation and K is some constant.

Gnanadesikan [1977] considers some graphical displays, of ordered observations, based on the procedure. The suggested plots include:

(1) y[i] vs y[n–i+1], i=1,2,...,[n/2] which for a symmetrical distribution should be linear with slope –1 and intercept 2y[m], where y[m] denotes the median. [See Figure 3.2(a)]

(2) y[m] – y[i] vs y[n–i+1] – y[m], i=1,2,...,[n/2] which for a symmetric distribution should be linear with unit slope and zero intercept. [See Figure 3.2(b)]

(3) y[n–i+1] – y[i] vs y[i] + y[n–i+1], i=1,2,...,[n/2] which for a symmetric distribution should be horizontal with intercept 2y[m], y[m] as above. [See Figure 3.2(c)]

The three plots shall be referred to as Type I, Type II and Type III plots, respectively.

For each plot, the deviations from the expected behaviour are given by

$$D(i) = y[i] + y[n-i+1] - 2y[m] \qquad (3.2)$$

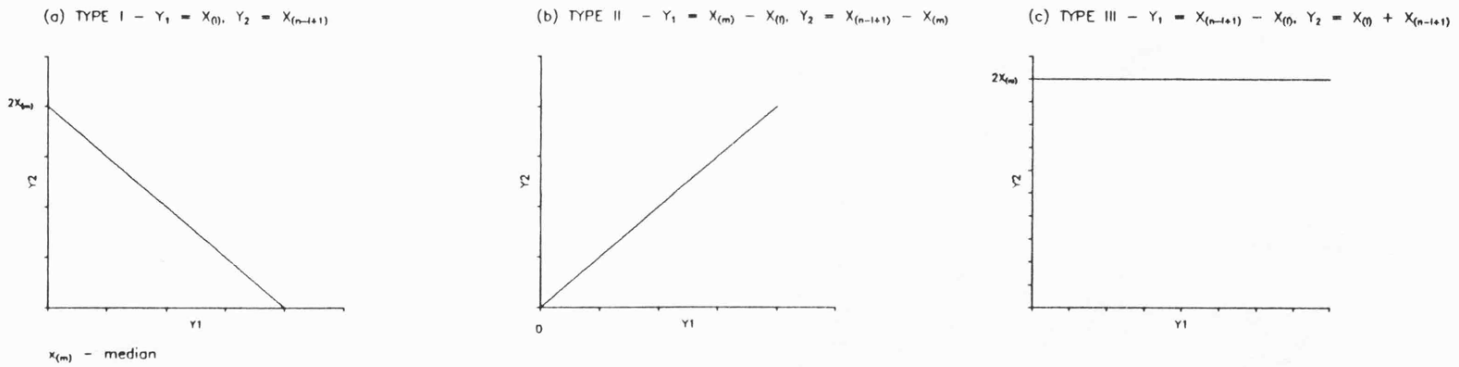
i=1,2,...,[n/2]. To achieve closer agreement with symmetry, a transformation can be selected from the family of power transformations. The chosen transformation is that which achieves closest agreement with the expected shape of the plot under the assumption of symmetry for the main body of the data, specifically that which minimises

$$\Sigma D_i^2 = \Sigma(y[i] + y[n-i+1] - 2y[m])^2 \qquad (3.3)$$

i=[0.1n],[0.1n]+1,...,[n/2].

Having assessed the marginal symmetry of the variables graphically it would be desirable to attach a numerical value to the results by obtaining a possible transformation to symmetry if the data required so. The next section discusses a "quick" computational method to provide such values.

Fig 3.2 Test for Symmetry Plots



3.2.2 Transformations to symmetry

Hinkley [1977] proposes a "quick" method of selecting a transformation for symmetrising based on the relationship between the mean and the median, where for a symmetric distribution the two values should be equal.

The general class of transformations considered is defined by

$$y_t = \frac{(y^t - 1)}{t} \tag{3.4}$$

where y is the original response. The desired value of t is that which gives approximate symmetry of the distribution of y_t.

If $y_1, y_2,\ldots,y_p$ is a homogenous random sample from a symmetric distribution with finite mean, then the sample will tend to reflect the identity

$$\text{mean} = \text{median} \tag{3.5}$$

which holds in the population. Thus, given an appropriate measure of scale, S, the degree of asymmetry in the sample may be measured by

$$d_t = \frac{\bar{y} - y[m]}{S} \tag{3.6}$$

where $\bar{y}$ is the sample mean and y[m] the sample median. For arbitrary positive data $y_1, y_2,\ldots,y_p$ a symmetrising transformation of type (3.4) is chosen so as to make d_t, the value of d obtained using y_t as data, as small as possible. The choices of t are normally restricted to -1, 0, 1/2, 1, 2 [Tukey, 1970] which are the reciprocal, log, square root, no transformation and square, respectively. A few values of d_t can be used to interpolate the solution to $d_t = 0$ if desired.

Reasonably efficient methods of estimating t so as to achieve normality or symmetry are heavily influenced by unduly influential observations, asymmetry is most apparent in the tails. The measure (3.6), while inefficient, is not so sensitive to these extreme observations, particularly if a robust scale is employed. Two suggestions are

$$s1 = \text{sample standard deviation} = \sqrt{\frac{1}{n}\Sigma(y - \bar{y})^2}$$

$$s2 = \text{sample interquartile range} = Q(3) - Q(1)$$

where Q(i) is the i–th quartile. The latter choice is more robust and provides a quicker

computational method. Additional robustness can be obtained by replacing the sample mean with the Winsorized mean [Tiao and Guttman, 1967; Guttman and Smith, 1969; Lorenzen, 1980].

A good property of the method is its moderate robustness relative to efficient methods. The method is useful for data sets which contain two or more variables where interest is in carrying out subsequent analysis assuming symmetric deviations from the mean (most simple statistical procedures make such an assumption).

The above technique does not make any probability distributional assumptions and cannot be easily extended to more than one variable at a time. The following sections provide techniques which are dependant on the normal distribution both univariate and multivariate.

3.3 Likelihood Approach

The Likelihood Approach is an extension by Andrews et al. [1971] of the Box & Cox [1964] approach to the problem of estimating a power transformation of multiresponse data so as to enhance normality.

If $Y^T = (y_1, y_2,...,y_p)$ denotes the set of p response variables of size n, the general problem may be formulated as follows:

.... to determine the vector of transformation parameters λ, such that the transformed variables $[g_1(Y^T;\lambda), g_2(Y^T;\lambda),..., g_p(Y^T;\lambda)]$ are more nearly p–variate normal, $N[\mu,\Sigma]$, than the original p variables. The elements of λ are unknown, as are those of μ and Σ. Provided that one can obtain an appropriate estimate $\hat{\lambda}$ of λ (as well as of μ and Σ) from the data, the original observations can be transformed one at a time to yield new observations, $[g_1(Y^T;\hat{\lambda}), g_2(Y^T;\hat{\lambda}),..., g_p(Y^T;\hat{\lambda})]$, which may then be considered as conforming more to a p–variate normal model than the original observations.

The work of Andrews et al. [1971] is concerned with transformation functions g_j, which are direct extensions of the power transformation of a single non–negative response X, to $X(\lambda)$, considered by Box & Cox [1964], where

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & , \lambda \neq 0 \\ \log_e \lambda & , \lambda = 0 \end{cases} \qquad (3.7)$$

Andrews et al. [1971] consider both marginal and joint transformations, where the latter is to achieve joint normality. Although marginal normality does not imply joint normality (See Figure 3.1 and Section 3.5.1), the choice of transformations to improve marginal normality may in many cases yield data more amenable to standard analyses.

The above approach leads to estimating λ by maximum likelihood (ML). In the multivariate case the ML estimate for λ is obtained by numerically maximising the joint log–likelihood function. Andrews et al. [1971] consider three possible cases depending on the objectives of the analysis requiring the need for transformations. The variables are transformed to achieve:

Case 1 – Marginal normality of the variables

Case 2 – Joint normality

Case 3 – Directional normality

For the purposes of the thesis only cases 1 and 2 are discussed.

Case 1 – Marginal Normality

If $Y^T = (y_1, y_2,...,y_p)$ denotes the set of p response variables of size n and $\lambda^T = (\lambda_1, \lambda_2,..., \lambda_p)$ the corresponding vector of transformation parameters then the following family of power transformations is defined as

$$g_j(Y^T;\hat{\lambda}) = y_j(\lambda_j) = \begin{cases} \dfrac{y_i^{\lambda_j} - 1}{\lambda_j} & , \lambda \neq 0 \\ \log_e \lambda_j & , \lambda = 0 \end{cases} \qquad (3.8)$$

where j=1,2,...,p.

A sensible starting point is to choose λ_j so as to improve the marginal normality of $y_j(\lambda_j)$. The logarithm of the profile likelihood function (which has been initially maximised with respect to the unknown mean and variance for a given λ_j), $l_{max}(\lambda_j)$, is maximised to provide the estimate $\hat{\lambda}_j$. If $Y^T = (y_1, y_2,...,y_p)$ denotes the n×p matrix of the original observations, and if the transformed observations obtained by using equation (3.8) are $Y^T(\lambda) = [y_1(\lambda_1), y_2(\lambda_2),...,y_p(\lambda_p)]$ where $y_j(\lambda_j)$ denotes the vector of n observations on the

j–th variable, each of which is obtained by transforming according to (3.8), then

$$l_{max}(\lambda_j) = -\frac{n}{2}\log_e\hat{\sigma}_{jj} + (\lambda_j - 1)\sum_{i=1}^{n}\log_e y_{ij} \qquad (3.9)$$

where y_{ij} denotes the i–th observation on the untransformed j=th response, and $\hat{\sigma}_{jj}$ is the maximum likelihood estimate of the variance of the presumed normal distribution of $y_j(\lambda_j)$, ie.

$$\hat{\sigma}_{jj} = \frac{1}{n}[y_j(\lambda_j) - \xi_j]^T[y_j(\lambda_j) - \xi_j] \qquad (3.10)$$

where ξ_j is the maximum likelihood estimate of $\xi_j = E[y_j(\lambda_j)]$. For an unstructured sample, ξ_j would be an n×1 vector all of whose elements are equal to the mean of the transformed observations on the j–th variable, while for the more general case of a linear model specification, $\xi_j = X\theta_j$, the appropriate estimate would be $X\hat{\theta}_j$. In addition to the second term on the right–hand–side of (3.9), $\hat{\sigma}_{jj}$ is also a function of λ_j, and the required maximum likelihood estimate, $\hat{\lambda}_j$, is the value of λ_j which maximises $l_{max}(\lambda_j)$. Since the maximisation is with respect to a single parameter λ_j, despite the complication of $\hat{\sigma}_{jj}$ being a function of λ_j, the computations involved are quite simple.

The value of $l_{max}(\lambda_j)$ for a sequence of values of $\hat{\lambda}_j$ can be computed to empirically determine the value, $\hat{\lambda}_j$, for which it is a maximum. Also, for a single parameter a graph of $l_{max}(\lambda_j)$ can be plotted so as to study its behaviour near $\hat{\lambda}_j$.

An approximate confidence interval for λ_j can be obtained by using asymptotic theory. So a 100(1–α)% confidence interval for λ_j is defined by

$$2\{l_{max}(\hat{\lambda}_j) - l_{max}(\lambda_j)\} \leq \chi^2_1(\alpha) \qquad (3.11)$$

where $\chi^2_\nu(\alpha)$ denotes the upper 100α% point of a chi–squared distribution with ν degrees of freedom.

Case 2 – Joint Normality

In Case 1, concern is with estimating power transformations of multiresponse data so as to improve marginal normality. Case 2 describes a method for choosing the transformations of (3.8) so as to enhance joint normality. Thus the n×p matrix $Y^T = [(y_{ij})]$ i=1,2,...,n; j=1,2,...,p, is the data matrix whose rows, Y_i^T, are the multivariate observations, and it is assumed that after a transformation of the form (3.7) the

transformed data $Y^T(\lambda)$ may be statistically described by a multivariate normal density function with mean μ^T and covariance Σ.

Let $\theta = E[Y^T(\lambda)] = 1\mu^T$. If $\lambda^T = (\lambda_1, \lambda_2,..., \lambda_p)$ is the set of transformation parameters yielding multivariate normality, the density function for the original data, Y, is

$$f(Y|\mu,\Sigma,\lambda) = |\Sigma|^{-n/2}(2\pi)^{-n}\exp\left[-\frac{1}{2}tr\Sigma^{-1}[Y(\lambda) - \theta]\,[Y(\lambda) - \theta]^T\right]J \qquad (3.12)$$

where , J the Jacobian of the transformation from Y to $Y(\lambda)$ is

$$\prod_{j=1}^{p}\prod_{i=1}^{n} y_{ij}^{\lambda_j - 1} \qquad (3.13)$$

Thus, the log–likelihood of μ, Σ and λ is given by (apart from an additive constant)

$$l(\mu,\Sigma,\lambda\,|\,Y) = -\frac{n}{2}\log_e|\Sigma| - \frac{1}{2}tr\Sigma^{-1}[Y(\lambda) - \theta]\,[Y(\lambda) - \theta]^T + \sum_{j=1}^{p}\left[(\lambda_j - 1)\sum_{i=1}^{n}\log_e y_{ij}\right] \qquad (3.14)$$

For specified λ_j, j=1,2,...,p, the maximum likelihood estimates of μ and Σ are given, respectively, by

$$\hat{\mu} = \frac{1}{n}Y(\lambda)1 \qquad (3.15)$$

and

$$\hat{\Sigma} = \frac{1}{n}\,[Y(\lambda) - \hat{\theta}]\,[Y(\lambda) - \hat{\theta}]^T \qquad (3.16)$$

where $\hat{\theta} = 1\hat{\mu}^T = \frac{1}{n}\,11^TY^T(\lambda)$.

If these estimates are substituted in the above log–likelihood function, the resulting maximised function (up to an additive constant) is

$$l_{max}(\lambda_1, \lambda_2,..., \lambda_p) = -\frac{n}{2}\log_e|\hat{\Sigma}| + \sum_{j=1}^{p}\left[(\lambda_j - 1)\sum_{i=1}^{n}\log_e y_{ij}\right] \qquad (3.17)$$

where y_{ij} is the i–th observation on the untransformed j–th reponse (i=1,2,...,n; j=1,2,...,p).

The maximum likelihood estimates $\hat{\lambda}_j$ (j=1,2,...,p) can be obtained by numerically maximising (3.17). However, in the general p–response case l_{max} is a function of p variables and thus the problem of studying and numerically maximising it is quite complex. In the bivariate case, p=2, (3.17) is a function of two variables so it can easily be computed and studied. This suggests the possibility of studying all possible pairs of responses, thus, considering them as bivariate data and so losing the joint relationships which may exist

across all the variables.

Using the normalised Box–Cox transformation (See Section 3.4.2.1) with covariance matrix Σ_λ (estimated by $\hat{\Sigma}_\lambda$) leads to the alternative log–likelihood

$$l_{max}(\lambda_1, \lambda_2,..., \lambda_p) = -\frac{n}{2}\log_e|\hat{\Sigma}_\lambda|. \tag{3.18}$$

It can be easily verified that (3.17) and (3.18) are numerically identical.

If $\hat{\lambda}_1, \hat{\lambda}_2,..., \hat{\lambda}_p$ are the values that maximise $l_{max}(\lambda)$, an approximate confidence region for $\lambda_1, \lambda_2,...,\lambda_p$ is defined by

$$2\{l_{max}(\hat{\lambda}_1, \hat{\lambda}_2,..., \hat{\lambda}_p) - l_{max}(\lambda_1, \lambda_2,..., \lambda_p)\} \leq \chi^2_p(a) \tag{3.19}$$

where $\chi^2_p(a)$ is the upper $100a\%$ point of the chi–squared distribution with p degrees of freedom.

3.4 The Proposed SURCON Approach

The proposed SURCON Analysis theory is based on two ideas from linear regression. The first of these is the simultaneous linear regression of several models which are related through their error terms. These are sometimes referred to as "disturbance related equations" or more commonly Zellner's "Seemingly Unrelated Regressions" [Zellner, 1962]. The second idea is based on "Constructed Variables" so called by Box and are commonly used in Regression Diagnostics. The first two subsections provide the background theory of these two techniques which is adapted into the SURCON (Seemingly Unrelated Regressions/Constructed Variables) analysis.

3.4.1 Seemingly Unrelated Regressions (SUR)

Consider a set of p regression equations

$$y_j = X_j\beta_j + e_j \tag{3.20}$$

j=1,2,...,p where $y_j(n\times1)$, $X_j(n\times k_j)$, β $(k_j\times1)$, $e_j(n\times1)$ and each equation obeys the classical regression assumptions. Assume, further, that there may be correlation between the random error e in different equations. In this case we have disturbance related sets of equations or seemingly unrelated regressions (SUR).

A convenient way to write these equations is

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \dots & 0 \\ 0 & X_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & X_p \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{bmatrix} \tag{3.21}$$

or

$$y_{np\times 1} = X_{np\times K}\beta_{K\times 1} + e_{np\times 1} \tag{3.22}$$

where $K = \sum_{j=1}^{p} k_j$. (NB: If $k_1 = k_2 = \dots = k_p = k$ then $K = k.p$)

Let e_{it} be the error for the t–th observation in the i–th equation then the assumption of disturbances being related between equations but not within equations implies that

$$E[e_{it}e_{js}] = \begin{cases} \sigma_{ij} & ,t=s \\ 0 & ,\text{otherwise} \end{cases} \tag{3.23}$$

or

$$E[e_i e_j^T] = \sigma_{ij} I_n \tag{3.24}$$

and thus the covariance matrix for the complete error vector can be written as

$$E[ee^T] = \begin{bmatrix} \sigma_{11}I_n & \sigma_{12}I_n & \dots & \sigma_{1p}I_n \\ \sigma_{21}I_n & \sigma_{22}I_n & \dots & \sigma_{2p}I_n \\ \vdots & \vdots & & \vdots \\ \sigma_{p1}I_n & \sigma_{p2}I_n & \dots & \sigma_{pp}I_n \end{bmatrix} \tag{3.25}$$

$$\underset{\sim}{\Phi} = \Sigma \otimes I \tag{3.26}$$

where

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix}$$

When the system (3.21) is viewed as a single equation (3.22) we can estimate β and, hence, all β_i by the generalised least squares (GLS) procedures ie. by minimising

$$e^T\underset{\sim}{\Phi}^{-1}e = (y - X\beta)^T\Phi^{-1}(y - X\beta) \tag{3.27}$$

obtaining the GLS estimator

$$\begin{aligned} \hat{\beta} &= (X^T\underset{\sim}{\Phi}^{-1}X)^{-1}X^T\underset{\sim}{\Phi}^{-1}y \\ &= [X^T(\Sigma^{-1}\otimes I)X]^{-1}X^T(\Sigma^{-1}\otimes I)y \end{aligned} \tag{3.28}$$

In detail

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \begin{bmatrix} \sigma^{11}X_1X_1 & \sigma^{12}X_1X_2 \dots & \sigma^{1p}X_1X_p \\ \sigma^{21}X_2X_1 & \sigma^{22}X_2X_2 \dots & \sigma^{2p}X_2X_p \\ \vdots & \vdots & \vdots \\ \sigma^{p1}X_pX_1 & \sigma^{p2}X_pX_2 \dots & \sigma^{pp}X_pX_p \end{bmatrix} \begin{bmatrix} \Sigma\sigma^{1i}X_1y_i \\ \Sigma\sigma^{2i}X_2y_i \\ \vdots \\ \Sigma\sigma^{pi}X_py_i \end{bmatrix} \tag{3.29}$$

where the σ^{ij} element is the (i,j)–th element of Σ^{-1}.

The covariance of $\hat{\beta}$ is

$$\text{Var–cov}(\beta) = (X^T\Phi^{-1}X)^{-1} = [X^T(\Sigma^{-1} \otimes I)X]^{-1} \tag{3.30}$$

This resulting estimator possesses the same properties as the estimator b, where

$$b = (X^TX)^{-1}X^Ty \tag{3.31}$$

obtained by considering the equations one at a time, i.e. it is unbiased and if y is normally distributed it is the maximum likelihood estimator and has minimum variance within the class of all unbiased estimators.

However, $\hat{\beta}_i$, the i–th vector $\hat{\beta}$ (estimator for the i–th equation using SUR) is better than b_i, the single equation estimator for the i–th equation since

a) it allows for correlation between e_i and error vectors from the equations

b) it uses information on explanatory variables that are included in the system but excluded from the i–th equation.

This gain in efficiency can be shown as follows. The estimator obtained by applying least squares (LS) to each separate equation is

$$\begin{aligned} b &= (X^TX)^{-1}X^Ty \\ &= [(X^TX)^{-1}X^T + (X^T\Phi^{-1}X)^{-1}X^T\Phi^{-1} - (X^T\Phi^{-1}X)^{-1}X^T\Phi^{-1}]y \\ &= (X^T\Phi^{-1}X)^{-1}X^T\Phi^{-1}y + Ay \\ &= (X^T\Phi^{-1}X)^{-1}X^T\Phi^{-1}(X\beta + e) + A(X\beta + e) \\ &= \beta + (X^T\Phi^{-1}X)^{-1}X^T\Phi^{-1}e + AX\beta + AXe \end{aligned} \tag{3.32}$$

where $A = (X^TX)^{-1}X^T - (X^T\Phi^{-1}X)^{-1}X^T\Phi^{-1}$

So,

$$E[b] = \beta + AX = \beta \tag{3.33}$$

since $AX = (X^TX)^{-1}X^TX - (X^T\Phi^{-1}X)^{-1}X^T\Phi^{-1}X$.

Hence, b is unbiased.

The covariance matrix is

$$\begin{aligned}\Sigma_b &= E[(b-\beta)(b-\beta)^T] \\ &= (X^T\Phi^{-1}X)^{-1}X^T\Phi^{-1}E(ee^T)\Phi^{-1}X(X^T\Phi^{-1}X)^{-1} + AE(ee^T)A \\ &= (X^T\Phi^{-1}X)^{-1} + A\Phi^{-1}A \qquad (3.34)\end{aligned}$$

(since $E(ee^T) = \Phi$) which is the variance–covariance of $\hat{\beta}$ plus some other positive semi–definite matrix.

Consequently,

$$\Sigma_b - \Sigma_{\hat{\beta}} = A\Phi A^T \qquad (3.35)$$

where $A\Phi^{-1}A^T$ is at least positive semi–definite and so $\hat{\beta}$ is strictly efficient unless $A = 0$ and thus $b = \hat{\beta}$

ie. $$\Sigma_b > \Sigma_{\hat{\beta}} \qquad A \neq 0 \qquad (3.36)$$

In general the efficiency gain tends to be higher when the errors among different equations are highly correlated.

There are, however, two cases when $b = \hat{\beta}$ and thus no gain in efficiency.

Case 1. If Σ is a diagonal matrix ie $\sigma_{ij} = 0$ for all $i \neq j$, in otherwords, no correlation between the random vectors of different equations.

Case 2. If $X_1 = X_2 = \ldots = X_p = X_0$

This follows because in this case

$$X = (I_p \otimes X_0)$$

and therefore

$$\begin{aligned}&[\,X^T(\Sigma^{-1}\otimes I)\,X\,]^{-1}\,X^T(\Sigma^{-1}\otimes I) \\ &= [(I\otimes X_0)^T(\Sigma^{-1}\otimes I)(I\otimes X_0)\,]^{-1}(I\otimes X_0)^T(\Sigma^{-1}\otimes I) \\ &= [(\Sigma^{-1}\otimes X_0)(I\otimes X_0)\,]^{-1}(I\otimes X_0)(\Sigma^{-1}\otimes I) \\ &= (\Sigma^{-1}\otimes X_0^TX_0)^{-1}(\Sigma^{-1}\otimes X_0)^T \\ &= (\Sigma\otimes(X_0^TX_0)^{-1})(\Sigma^{-1}\otimes X_0) \\ &= (I\otimes(X_0^TX_0)^{-1}X_0^T \\ &= (X^TX)^{-1}X^T \qquad (3.37)\end{aligned}$$

So on post–multiplying (3.37) by y the least squares estimator in (3.31) is obtained.

This result also demonstrates that in general the gain in efficiency tends to be higher when the explanatory variables in different equations are not highly correlated.

The above theory is all based on the covariance matrix Σ being known. However, in many instances this may not be the case so it has to be estimated. The following section discusses the estimator used when Σ is unknown.

3.4.1.1 Estimation with unknown covariance matrix

If Σ is unknown, it can be estimated by $\hat{\Sigma}$ (or S) based on the LS residuals

$$\hat{e}_i = y_i - X_i b_i \tag{3.38}$$

and has elements

$$\hat{\sigma}_{ij} = s_{ij} = \frac{1}{n} \hat{e}_i \hat{e}_j \tag{3.39}$$

i,j = 1,2,...,p and the resulting estimator is

$$\hat{\hat{\beta}} = [\, X^T(S^{-1} \otimes I)\, X\,]^{-1} X^T(S^{-1} \otimes I)y \tag{3.40}$$

Having obtained the parameter estimates it may be required to test hypothesis on them. A hypothesis of great interest is the significance of these estimates. The following section discusses hypothesis testing in the Seemingly Unrelated Regressions context in general and in particular the hypothesis for the equality of the coefficients.

3.4.1.2 Hypothesis Testing

In this section we consider tests for two types of hypotheses. The first test (Section 3.4.1.2a) deals with linear restrictions on the coefficients β. These constraints can be used to assert linear relationships between the β's e.g. in testing for their equality. The second test (Section 3.4.1.2b) deals with testing for a diagonal Σ since the least squares estimator b $= (X^TX)^{-1}X^Ty$ is fully efficient if this condition holds (See Case 1. Section 3.4.1) and, hence, nothing is achieved by carrying out a 2SLS. Judge et al. [1985] give a thorough discussion on these tests and others associated with Seemingly Unrelated Regressions.

3.4.1.2a Linear Restrictions on the Coefficient Vector β

Here we consider testing a set of linear restrictions represented by $R\beta = r$, for example a test for the equality of all the β's i.e. $\beta_1 = \beta_2 = \ldots = \beta_p$ (See Zellner [1962]). The relevant test statistics depends on Σ and because it is unknown is replaced with $\hat{\Sigma}$.

Therefore, the tests now have large sample justification and secondly it is possible to test restrictions that relate coefficients in one equation with the coefficients in other equations.

Under the assumption that e is normally distributed and that the null hypothesis $R\beta = r$ is true the statistic

$$\delta = \frac{(r - R\hat{\beta})^T(RCR^T)^{-1}(r - R\hat{\beta})}{(y - X\hat{\beta})^T(\Sigma^{-1} \otimes I)(y - X\hat{\beta})} \cdot \frac{np - k}{J} \sim F_{(J,\, np-k)} \tag{3.41}$$

where $C = [X^T(\Sigma^{-1} \otimes I)\, X]^{-1}$ and J is the number of restrictions (ie. the number of rows in R)

When Σ is replaced by $\hat{\Sigma}$ and $\hat{\beta}$ is replaced by $\hat{\hat{\beta}}$ the limiting distribution of

$$(r - R\hat{\hat{\beta}})^T(R\hat{C}R^T)^{-1}(r - R\hat{\hat{\beta}}) \tag{3.42}$$

with $C = [X^T(\hat{\Sigma}^{-1} \otimes I)\, X]^{-1}$ is $\chi^2_{(J)}$. The limiting distribution of

$$\hat{\delta} = \frac{(r - R\hat{\hat{\beta}})^T(R\hat{C}R^T)^{-1}(r - R\hat{\hat{\beta}})}{(y - X\hat{\hat{\beta}})^T(\hat{\Sigma}^{-1} \otimes I)(y - X\hat{\hat{\beta}})} \cdot \frac{np - k}{J} \tag{3.43}$$

is $(1/J)\, \chi^2_{(J)}$.

Also, $F_{(J,\ np\ -\ k)}$ converges in distribution to $(1/J)\chi^2_{(J)}$ as $n \to \infty$ and asymptotically it makes no difference whether $\hat{\lambda}$ is used in conjunction with the F or χ^2 distribution or whether we simply use the χ^2 distribution.

Lemma 3.1

The test statistic $\hat{\delta}$ can be written in terms of restricted and unrestricted sums of squares as follows.

$$\hat{\delta} = \frac{(y - X\hat{\beta}^*)^T(\hat{\Sigma}^{-1} \otimes I)(y - X\hat{\beta}) - (y - X\hat{\beta})^T(\hat{\Sigma}^{-1} \otimes I)(y - X\hat{\beta})}{(y - X\hat{\beta})^T(\hat{\Sigma}^{-1} \otimes I)(y - X\hat{\beta})} \cdot \frac{np - k}{J} \tag{3.44}$$

Proof:

In proving this Lemma it is sufficient to show that the numerators in (3.43) and (3.44) are identical.

Let

$$\hat{\beta}^* = \hat{\beta} + CR^T(RCR^T)^{-1}(r - R\hat{\beta}) \tag{3.45}$$

then the LS estimator obtained by minimising

$$(y - X\beta)^T(\Sigma^{-1}\otimes I)(y - X\beta) \tag{3.46}$$

Subject to
$$R\beta = r \tag{3.47}$$

then
$$\begin{aligned}(y - X\hat{\beta}^*)^T(\hat{\Sigma}^{-1}\otimes I)(y - X\hat{\beta}^*) &= (y - X\hat{\beta})^T(\hat{\Sigma}^{-1}\otimes I)(y - X\hat{\beta}) \\ &\quad - (y - X\hat{\beta})^T(\hat{\Sigma}^{-1}\otimes I)XCR^T(RCR^T)^{-1}(r - R\hat{\beta}) \\ &\quad + [XCR^T(RCR^T)^{-1}(r - R\hat{\beta})]^T(\hat{\Sigma}^{-1}\otimes I)[XCR^T(RCR^T)^{-1}(r - R\hat{\beta})] \\ &\quad - XCR^T(RCR^T)^{-1}(r - R\hat{\beta})(\hat{\Sigma}^{-1}\otimes I)(y - X\hat{\beta})\end{aligned} \tag{3.48}$$

So
$$(y - X\hat{\beta}^*)^T(\hat{\Sigma}^{-1}\otimes I)(y - X\hat{\beta}^*) - (y - X\hat{\beta})^T(\hat{\Sigma}^{-1}\otimes I)(y - X\hat{\beta}) = -A+B-A^T \tag{3.49}$$

Now,
$$\begin{aligned}B &= (r - R\hat{\beta})^T(RCR^T)^{-1}RCX^T(\hat{\Sigma}^{-1}\otimes I)XCR^T(RCR^T)^{-1}(r - R\hat{\beta}) \\ &= (r - R\hat{\beta})^T(RCR^T)^{-1}RCR^T(RCR^T)^{-1}(r - R\hat{\beta}) \\ &= (r - R\hat{\beta})^T(RCR^T)^{-1}(r - R\hat{\beta})\end{aligned} \tag{3.50}$$

also
$$\begin{aligned}A &= (r - R\hat{\beta})^T(RCR^T)^{-1}RCX^T(\hat{\Sigma}^{-1}\otimes I)(y - X\hat{\beta}) \\ &= (r - R\hat{\beta})^T(RCR^T)^{-1}RCX^T(\hat{\Sigma}^{-1}\otimes I)(I - XCX^T(\hat{\Sigma}^{-1}\otimes I)y \\ &= (r - R\hat{\beta})^T(RCR^T)^{-1}RCX^T(\hat{\Sigma}^{-1}\otimes I)[I - XCX^T(\hat{\Sigma}^{-1}\otimes I)](X\beta + e) \\ &= (r - R\hat{\beta})^T(RCR^T)^{-1}\{RCX^T(\hat{\Sigma}^{-1}\otimes I)X\beta - RCX^T(\hat{\Sigma}^{-1}\otimes I)XCX^T(\hat{\Sigma}^{-1}\otimes I)X\beta \\ &\quad + RCX^T(\hat{\Sigma}^{-1}\otimes I)e - RCX^T(\hat{\Sigma}^{-1}\otimes I)XCX^T(\hat{\Sigma}^{-1}\otimes I)Xe\} \\ &= (r - R\hat{\beta})^T(RCR^T)^{-1}\{R\beta - R\beta + RCX^T(\hat{\Sigma}^{-1}\otimes I)e - RCX^T(\hat{\Sigma}^{-1}\otimes I)e \\ &= 0 = A^T\end{aligned} \tag{3.51}$$

Therefore, by (3.48)
$$(y - X\hat{\beta}^*)^T(\hat{\Sigma}^{-1}\otimes I)(y - X\hat{\beta}^*) - (y - X\hat{\beta})^T(\hat{\Sigma}^{-1}\otimes I)(y - X\hat{\beta}) = (r - R\hat{\beta})^T(RCR^T)^{-1}(r - R\hat{\beta})$$

□

If the equation is written in this way, it shows that $\hat{\delta}$ is an extension of the F statistic commonly used to test the significance of the increase in residual sums of squares that result from the imposition of linear constraints. If Σ is replaced by $\hat{\Sigma}$ and $\hat{\beta}$, $\hat{\beta}^*$ by $\hat{\beta}$ and $\hat{\beta}^*$, respectively, the resulting expression is equivalent to $\hat{\delta}$.

– <u>Estimate of Σ used in $\hat{\delta}$</u>

In deriving $\hat{\delta}$ the question of whether to estimate Σ using LS residuals from the restricted or unrestricted model arises (ie from H_0 or H_A).

However, since we are interested in the probability distribution of $\hat{\delta}$ when

$$H_0: R\beta = r \tag{3.52}$$

is true, it could be argued that it is more logical to base $\hat{\delta}$ on an estimate of Σ that assumes that assumes that the null hypothesis is true, although asymptotically it makes no difference as to which one is adopted.

3.4.1.2b Testing for a Diagonal Covariance Matrix Σ

If the error terms across equations are uncorrelated (Σ is diagonal) then the least squares estimator $b = (X^TX)^{-1}X^Ty$ is fully efficient. It is, therefore, useful to have test statistics to test this hypothesis. Assuming normality, Breusch and Pagan [1980] have shown that the Lagrange Multiplier (LM) statistic for testing the null hypothesis of a diagonal Σ is given by

$$\eta = n \sum_{i=2}^{p} \sum_{j=1}^{i-1} r_{ij}^2 \tag{3.53}$$

where $r_{ij} = \hat{\sigma}_{ij}/\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}$ and $\hat{\sigma}_{ij} = (y_i - X_ib_i)^T(y_j - X_jb_j)$. Under H_0, η has an asymptotic $\chi^2_{[p(p-2)/2]}$ distribution.

Another test for a diagonal Σ can be based on the likelihood ratio test [Judge et al 1985] but for the purposes of this thesis only the LM test is used.

3.4.2 Constructed Variables

3.4.2.1 Structured sample case

If we take one equation from the system of equations defined in (3.20) then a simple score test can be derived which is the t–test for the significance of a regression coefficient. The model is first replaced by a model on the normalised Box–Cox transformation where

$$z_j(\lambda) = \begin{cases} \dfrac{y_i^{\lambda_j} - 1}{\lambda \dot{y}_j^{\lambda-1}} & , \lambda \neq 0 \\ \dot{y}_j \log_e y_j & , \lambda = 0 \end{cases} \tag{3.54}$$

and $\dot{y}_j = (\prod_{i=1}^{n} y_{ij})^{1/n}$ is the geometric mean of y.

It is the hope that for some λ

$$z(\lambda) = X\beta + \epsilon \tag{3.55}$$

The value of λ for which this applies is estimated. Expanding $z(\lambda)$ by a Taylor series about the known (hypothesised) value λ_0 yields

$$z(\lambda) \simeq z(\lambda_0) + (\lambda - \lambda_0)\omega(\lambda_0) \tag{3.56}$$

and the approximate linear model

$$z(\lambda_0) = X\beta - (\lambda - \lambda_0)\omega(\lambda_0) + \epsilon$$

$$z(\lambda_0) = X\beta + \gamma\,\omega(\lambda_0) + \epsilon \tag{3.57}$$

where $\omega(\lambda_0) = \partial z(\lambda)/\partial\lambda$ evaluated at λ_0 and was called the *Constructed Variable* by Box [1980] also $\gamma = -(\lambda - \lambda_0)$.

The least squares estimate for the regression coefficient of $\omega(\lambda_0)$ is

$$\hat{\gamma} = \omega^T(\lambda_0)Az(\lambda_0)/\omega^T(\lambda_0)A\omega(\lambda_0) \tag{3.58}$$

where $A = I - H$ and $H = X(X^TX)^{-1}X^T$ is the Hat matrix for the model in (3.57).

For brevity, we shall write z and ω in (3.58) and similar expressions, unless dependence on λ is important.

The variance of $\hat{\gamma}$ is given by $\sigma^2/(\omega^TA\omega)$. To form a t test for the significance of $\hat{\gamma}$ an estimate of σ^2 is required. We shall use the estimate under the null hypothesis, s^2. The purpose of the Score test is to avoid calculation of $\hat{\lambda}$, so that the maximum likelihood estimator of σ^2 is not available. An approximation to this estimate is given by

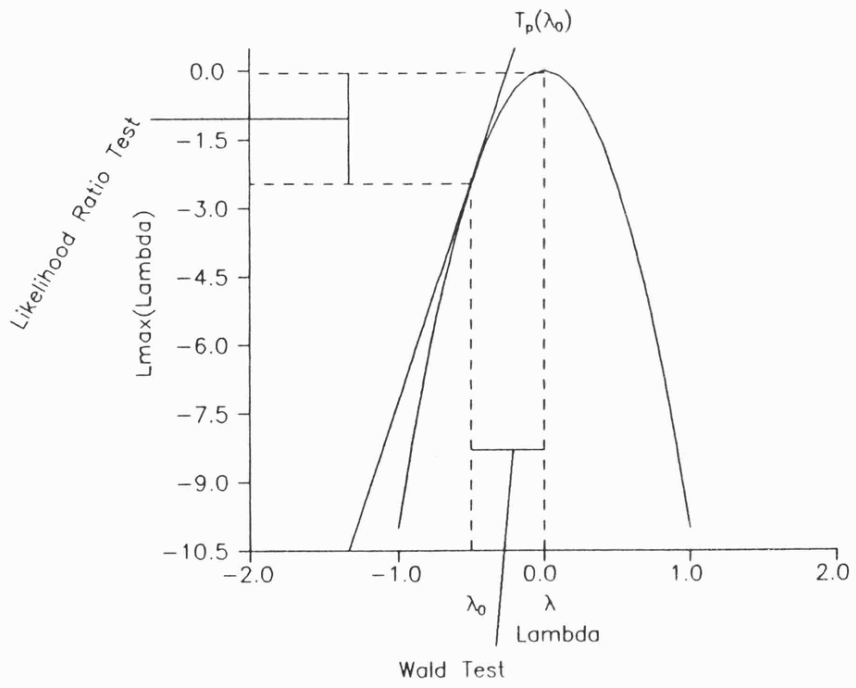
$$(n - p - 1)s_z^2 = z^TAz - (z^TA\omega)^2/(\omega\,A\omega) \tag{3.59}$$

The t test for the hypothesis $\gamma = 0$ is the approximate Score statistic [Atkinson 1973]

$$T_p(\lambda) = -\frac{z(\lambda_0)^TAz(\lambda_0)}{s_z\sqrt{\{\ \omega(\lambda_0)\ ^TA\omega(\lambda_0)\ \}}} \tag{3.60}$$

The negative sign arises because, in (3.57), $\gamma = -(\lambda - \lambda_0)$. Use of s_z^2 rather than s^2, to estimate σ^2 yields a test with higher power. To the extent that the linear approximation which leads to (3.57) is exact, $\hat{\gamma} = -(\lambda - \lambda_0)$. $T_p(\lambda)$, therefore, provides an approximate test of the hypothesis $\lambda = \lambda_0$. This score test is an approximation to the likelihood ratio test (See Figure 3.3). A third test based on the likelihood function is the Wald Test as displayed in Figure 3.3.

Fig 3.3 Likelihood Ratio Test, Wald Test and Score Test



The score test can usefully be interpreted in terms of the univariate regression through the origin of the residual values of $z(\lambda)$ on the residual constructed variables.

For the power transformation (3.53) the constructed variable is given by

$$\omega_p(\lambda) = \frac{\partial z(\lambda)}{\partial \lambda} = \frac{y^{\lambda}\log y}{\lambda \dot{y}^{\lambda-1}} - \frac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}}(1/\lambda + \log \dot{y}) \tag{3.61}$$

Usually interest is in testing hypotheses about a few special values of interest of λ. For λ_0 ie. H_0: No transformation

$$\omega_p(1) = y\{\log(y/\dot{y}) - 1\} + \log \dot{y} + 1 \tag{3.62}$$

Presence of regression on this constructed variable would, therefore, be evidence that a transformation is required. Similarly, for the null hypothesis of a log transformation, $\lambda_0 = 0$

$$\omega_p(0) = \dot{y} \log y(\log y/2 - \log \dot{y}) \tag{3.63}$$

Calculation of the score statistics does not require $\omega_p(\lambda)$ directly but rather the residual constructed variable $\overset{*}{\omega}_p(\lambda)$ which is formed by

$$\overset{*}{\omega}_p(\lambda) = (I - H)\omega_p(\lambda) = A\omega_p(\lambda) \tag{3.64}$$

where $H = X(X^TX)^{-1}X^T$ is the Hat matrix for the model in (3.57) and $A = I - H$.

At $\hat{\lambda}$ the value of $\hat{\gamma}$ is identically zero. An approximation to the variance of $\hat{\lambda}$ is $s^2/\omega^T(\hat{\lambda})A\omega(\hat{\lambda})$ and the corresponding $100\alpha\%$ confidence interval for λ accordingly has limits

$$\hat{\lambda} \pm t_{\alpha,n-p-1}s/\sqrt{\{\omega^T(\hat{\lambda})A\omega(\hat{\lambda})\}} \tag{3.65}$$

NB: All the quantities are calculated at $\hat{\lambda}$ and s^2 is the residual mean square estimate of σ^2.

As mentioned earlier the score test can usefully be interpreted in terms of the univariate regression through the origin of the residual values of $z(\lambda)$ on the residual constructed variable ie. if

$$\overset{*}{\omega} = (I - H)\omega = A\omega \quad \text{and} \quad \overset{*}{z} = (I - H)z = Az$$

and treating these as variables rather than residuals, we have the univariate regression

$$\overset{*}{z}_j = \gamma_j \overset{*}{\omega}_j + e_j \tag{3.66}$$

j=1,...,p which leads to the LS estimate $\hat{\gamma}$ for γ. This is the estimate of the slope of the

added variable plot of z_j^* against ω_j^*.

To find a quick estimate for λ the linearised model (3.57) can be used to give an approximation to $\hat{\lambda}$ .If $\tilde{\lambda}$ denotes the quick estimate it follows that

$$\tilde{\lambda} - \lambda_0 = -\hat{\gamma}$$

or

$$\tilde{\lambda} = \lambda_0 - \hat{\gamma} \tag{3.67}$$

3.4.2.2 Unstructured sample case

In the case of the unstructured sample the system of equations, model (3.20), becomes

$$y_j = 1\mu_j + \epsilon_j \tag{3.68}$$

j=1,2,...,p where μ_j would be interpreted as the population mean of y_j.

All the quantities as discussed previously continue to hold except the hat matrix now takes the form

$$H = \frac{1}{n}11^T \tag{3.69}$$

where 1 is an n×1 vector with all elements equal to unity.

This leads to

$$\omega_j^* = (I - H)\omega_j(\lambda_0) = \omega_j(\lambda_0) - \bar{\omega}_j(\lambda_0) \tag{3.70}$$

and

$$z_j^* = (I - H)z_j(\lambda_0) = z_j(\lambda_0) - \bar{z}_j(\lambda_0) \tag{3.71}$$

where $\bar{\omega}_j(\lambda_0)$, $\bar{z}_j(\lambda_0)$ are the sample means of the constructed variable and the normalised Box–Cox transformation for y_j, respectively, evaluated at λ_0.

We note that the linearised model (equation) (3.57) now takes the form

$$z_j = 1\mu_j + \gamma_j\omega_j + \epsilon_j \tag{3.72}$$

j=1,2,...,p.

The next section discusses the proposed SURCON algorithm based on the theory in the previous two sections (3.4.1) and (3.4.2).

3.4.3 The proposed SURCON analysis algorithm

In general, the system of equations (3.20) considering an unstructured sample and no transformation is as in (3.68) where μ_j is the population mean of the j–th equation

(variable in this case). On taking into account the constructed variable and using the linearised model in the form of (3.72) we can write the model as

$$y = X\beta + e \tag{3.73}$$

where y is an np×1 , X is an np×2p block–diagonal matrix of the form

$$X = \begin{bmatrix} X_1 & 0 & \dots & 0 \\ 0 & X_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & X_p \end{bmatrix} \tag{3.74}$$

with $X_i = \begin{bmatrix} 1 & \omega_j \\ 1 & \omega_j \\ \vdots & \vdots \\ 1 & \omega_j \end{bmatrix}$ of order n×2 also $\beta = (\mu_1, \gamma_1, \mu_2, \gamma_2, \dots, \mu_p, \gamma_p)^T$ and is of order 2p×1, e is an np×1 vector of residuals across all the variables.

If we let $A_{ij} = \begin{bmatrix} \sigma^{ij}n & \sigma^{ij}\Sigma\omega_j \\ \sigma^{ij}\Sigma\omega_i & \sigma^{ij}\Sigma\omega_i\omega_j \end{bmatrix}$, i,j = 1,2,...,p,

then

$$X^T(\Sigma^{-1} \otimes I)\, X = \begin{bmatrix} A_{11} & A_{12} \dots A_{1p} \\ A_{21} & A_{22} \dots A_{2p} \\ \vdots & \vdots \quad \dots \quad \vdots \\ A_{p1} & A_{p2} \dots A_{pp} \end{bmatrix} \tag{3.75}$$

Also,

$$X^T(\Sigma^{-1} \otimes I) = [\sum_{j=1}^{p}\sigma^{1j}y_j, \sum_{j=1}^{p}\sigma^{1j}\omega_1 y_j, \dots, \sum_{j=1}^{p}\sigma^{pj}y_j, \sum_{j=1}^{p}\sigma^{1j}\omega_p y_j]^T \tag{3.76}$$

where Σ is the error covariance matrix and σ^{ij} is the i,j–th element of its inverse Σ^{-1}.

From the two expressions (3.75) and (3.76) the GLS estimate for β can be obtained as

$$\hat{\beta} = [X^T(\Sigma^{-1} \otimes I)\, X]^{-1} X^T(\Sigma^{-1} \otimes I)y \tag{3.77}$$

where Σ is as above.

However, the expression could greatly be simplified for computational purposes by first centering both y and ω ie.replacing each with its deviation from its mean. This would make all terms of the form $\Sigma\omega_j$ be equal to zero. Further consequences are that the model (equation) can be reformulated as follows

$$y_j^* = \gamma_j \omega_j^* + e_j \tag{3.78}$$

j=1,2,...,p.

NB μ_j disappears since we are dealing with the unstructured sample and is estimated by $\bar{y}$.

The model as in (3.78) conforms to the one defined in (3.66) regarding the regression

of the residual values of $z(\lambda)$ on the residual constructed variable, and since we are mainly concerned with the estimate of γ we can adopt (3.78) as the single equation formulation of the problem.

Using (3.78) not only reduces the computational requirement but also provides the necessary values used for the diagnostic plots eg. the added variable plots.

The system of equations (3.20) can, therefore, be written as

$$\begin{bmatrix} z_1^* \\ z_2^* \\ \vdots \\ z_p^* \end{bmatrix} = \begin{bmatrix} \sigma^{11}\omega_1^{*T}\omega_1^* & \sigma^{11}\omega_1^{*T}\omega_2^* \ldots \sigma^{11}\omega_1^{*T}\omega_p^* \\ \sigma^{11}\omega_2^{*T}\omega_1^* & \sigma^{11}\omega_2^{*T}\omega_2^* \ldots \sigma^{11}\omega_2^{*T}\omega_p^* \\ \vdots & \vdots \\ \sigma^{11}\omega_p^{*T}\omega_1^* & \sigma^{11}\omega_p^{*T}\omega_2^* \ldots \sigma^{11}\omega_p^{*T}\omega_p^* \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{bmatrix} \qquad (3.79)$$

If we substitute σ^{ij} with the estimates $\hat{\sigma}^{ij}$ then we obtain the GLS estimator $\hat{\hat{\gamma}}$ for γ.

So if $\tilde{\lambda}$ is the vector of "quick estimates" for λ and λ_0 is the vector of the hypothesised values then

$$\tilde{\lambda} = \lambda_0 - \hat{\gamma}$$

i.e.
$$\begin{bmatrix} \tilde{\lambda}_1 \\ \tilde{\lambda}_2 \\ \vdots \\ \tilde{\lambda}_p \end{bmatrix} = \begin{bmatrix} \lambda_{01} - \hat{\gamma}_1 \\ \lambda_{02} - \hat{\gamma}_2 \\ \vdots \\ \lambda_{0p} - \hat{\gamma}_p \end{bmatrix} \qquad (3.80)$$

where λ_{0i} (i=1,2,...,p) is the hypothesised λ for the i–th variable. Usually $\lambda_{0i} = 1$, $\forall$ i is a reasonable starting point since it corresponds to no transformation required for all the variables.

The standard error of $\hat{\gamma}_i$ is given by the square root of the i–th diagonal element of $[X^T(\Sigma^{-1} \otimes I)\, X]^{-1}$.

Tests for the significance of γ (hence λ) are necessary since the decisions and conclusions on the parameter estimates are based on it. The following section discusses these tests.

3.4.3.1 Hypothesis Testing (Significance of γ)

3.4.3.1a Significance of γ

Section 3.4.1.2 discusses the testing of a set of linear constraints (restrictions) represented by $R\beta = r$. In the SURCON context, the hypothesis to be tested is whether

there is a presence of regression of z^* on the residual constructed variable ω^*.

This can be formulated as follows:

H_0: No regression of z^* on w^*

vs

H_A: Presence of regression of z^* on w^*

Alternatively, we wish to test

$$H_0: \gamma = 0 \text{ vs } H_A: \gamma \neq 0$$

where $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p)$ and $\gamma_1 = \gamma_2 = \gamma_p = 0$.

In the notation used previously we have

$$R\gamma = 0$$

i.e.
$$\begin{bmatrix} 1 & -1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & -1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & & & \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{3.81}$$

Under the assumption that e is normally distributed and that H_0 is true

$$\hat{\delta} = \frac{(R\hat{\gamma})^T(R\hat{C}R^T)^{-1}(R\hat{\gamma})}{(z^* - \hat{\gamma}\omega^*)^T(\hat{\Sigma}^{-1} \otimes I)(z^* - \hat{\gamma}\omega^*)} \cdot \frac{np - p}{p} \sim F_{(p,\, np-p)} \tag{3.82}$$

where $\hat{C} = [\omega^{*T}(\hat{\Sigma}^{-1} \otimes I)\omega^*]^{-1}$ which is the covariance matrix for $\hat{\gamma}$.

Considering the model under H_0 we have

$$z^* = e \tag{3.83}$$

and so by (3.44)

$$\hat{\delta} = \frac{z^{*T}(\hat{\Sigma}^{-1} \otimes I)\, z^* - (z^* - \hat{\gamma}\omega^*)^T(\hat{\Sigma}^{-1} \otimes I)(z^* - \hat{\gamma}\omega^*)}{(z^* - \hat{\gamma}\omega^*)^T(\hat{\Sigma}^{-1} \otimes I)(z^* - \hat{\gamma}\omega^*)} \cdot \frac{np - p}{p} \tag{3.84}$$

or

$$\hat{\delta} = \frac{\text{SS of scaled residual(under } H_0 - \text{under } H_A)}{\text{SS of scaled residuals under } H_A} \cdot \frac{np - p}{p} \tag{3.85}$$

The value $\hat{\delta}$ can be tested against either an $F_{(p,\, np-p)}$ or $(1/p)\chi^2_{(p)}$.

The choice of estimate for Σ could be either under H_0 or under H_A, however, asymptotically there is no difference in the results obtained from either choice.

Under H_0 the estimate of Σ would be the covariance matrix of z. It can be readily shown, therefore, that the first term in the numerator of (3.85), $z^{*T}(\hat{\Sigma}^{-1} \otimes I)\, z^*$, can be written as np^2.

3.4.3.1b Testing for the Independence of the variables

Section 3.4.2.1b discusses the test used for testing the diagonality of Σ. In the context of the SURCON algorithm this is equivalent to testing for the independence of the variables. If the variables are uncorrelated it implies that their transformations can be performed marginally i.e. by considering one variable at a time. The consequence of this is that the joint transformations would not necessarily enhance joint normality and it would , therefore, not be worthwhile to obtain them.

The actual computations are carried out by substituting the relevant values into (3.53). So $\hat{\sigma}_{ij} = (z_i^* - \gamma_i \omega_i^*)^T (z_j^* - \gamma_j^* \omega_j^*)$.

3.4.3.2 Convergence of SURCON estimates to Maximum Likelihood estimates

The theory in Section 3.4.3 so far is used to obtain "quick estimates" for the transformation parameters based on some hypothesised values λ_{0j},j=1,2,...,p. It is, however, desirable to obtain the maximum likelihood estimates, $\hat{\lambda}_j$, j=1,2,...p. This can be achieved by using an iterative scheme [Atkinson, 1985].

The following discussion is based on one variable being considered at a time.

The maximum likelihood estimate of λ for the j–th variable is the value satisfying

$$T_p(\hat{\lambda}_j) = 0 \qquad (3.86)$$

Solving this expression numerically would yield the required estimates. The *false position method* was selected for the SURCON algorithm because of its better numerical properties against Newton methods.

Given two values of λ_j at which the values of $T_p(\lambda_j)$ have opposite signs, a third value of λ_j for which $T_p(\lambda_j)$ should be zero is found by linear interpolation. If the magnitude of $T_p(\lambda_j)$ at this new value of λ_j is not sufficiently small, the process is repeated with the two values of $T_p(\lambda_j)$ which are smallest in magnitude and opposite in sign. The iteration continues until a sufficiently small absolute value of $T_p(\lambda_j)$ is obtained. To initiate the method a grid search has to be made to locate two value of λ_j bracketing the solution to (3.86). For the sake of clarity the notation is changed slightly so the $\lambda = \lambda_j$. A satisfactory starting point is with the values 1 and $\tilde{\lambda}$. The first iteration is then given by

$$\lambda_1 = \frac{T_p(\tilde{\lambda}) - \tilde{\lambda}T_p(1)}{T_p(\tilde{\lambda}) - T_p(1)} \tag{3.87}$$

and the general step in the false position method is given by

$$\lambda_k = \frac{\tilde{\lambda}_{k-1}T_p(\tilde{\lambda}) - \tilde{\lambda}T_p(\tilde{\lambda}_k)}{T_p(\tilde{\lambda}_k) - T_p(\tilde{\lambda}_{k-1})} \tag{3.88}$$

In (3.88) successive values of $\tilde{\lambda}_k$ and $\tilde{\lambda}_{k-1}$ are chosen to give values of $T_p(\lambda)$ with opposite signs and smallest absolute values among the three candidate points. The method can be considered to converge with the condition $|T_p(\lambda)| < 10^{-q}$, where q is some integer.

After the above iteration is performed on the j–th variable a test for the significance of γ (Section 3.4.3.1) can be made. If $\hat{\delta}$ is significant then a new variable is selected and the false position method applied to it. The process continues until $\hat{\delta}$ ceases to be significant.

Figure 3.4 is a graphical demonstration of the algorithm with two variables. It displays the contours of the log–likelihood function for the simulated bivariate normal sample of Example E1. The x and y axes are the transformation parameters λ_1 and λ_2, respectively.

The lines A and A′ represent the initial values of λ for Y_1 and Y_2 respectively, which are both 1 (no transformations) in this case. The algorithm keeps one line constant, A′, and searches along it for a value of λ_1 which has an opposite sign to the previous λ_1, line B. This would bracket the solution for λ_1. It then searches along B for a value of λ_2 which has an opposite sign to the previous λ_2, line B′. The process continues until the algorithm converges. The coordinates of the points of intersection (a,b and c) of each pair of lines are the estimated values for λ at that iteration and so the total number of iterations is six (3×2) in this example.

Figure 3.5 is a plot of the score statistic $T_p(\lambda)$ against λ for X1 from the same data as above. Provided the solution is properly bracketed this plot can easily be used to read off $\hat{\lambda}$ which corresponds to $T_p(\lambda) = 0$. A confidence interval can also be read off from the plot. The 95% confidence interval for this example is wider than the theoretical limits from a Student's–t distribution with n–1 degrees of freedom but it is quite possible for the

Fig 3.4 Loglikelihood Contours for BVN data

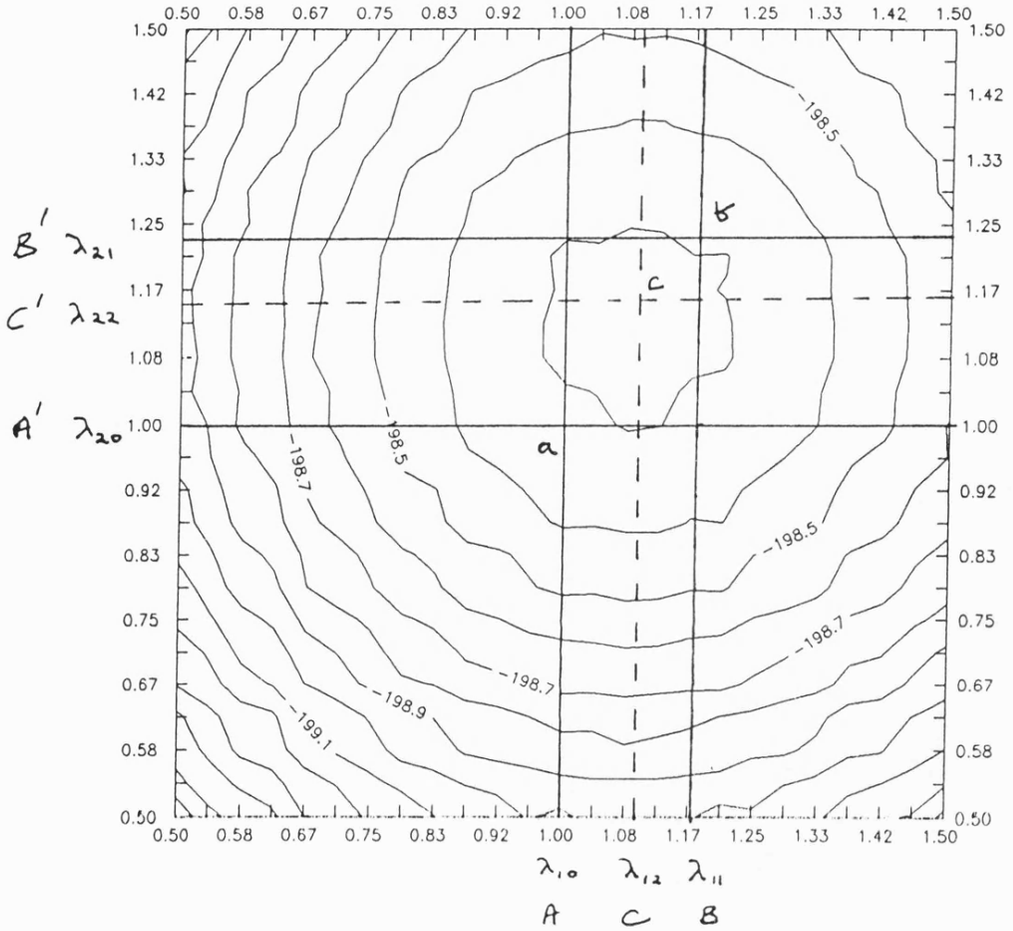


Fig 3.5 Score Statistic vs Lambda
(Bivariate Normal Data, $\rho=0$, X_1)

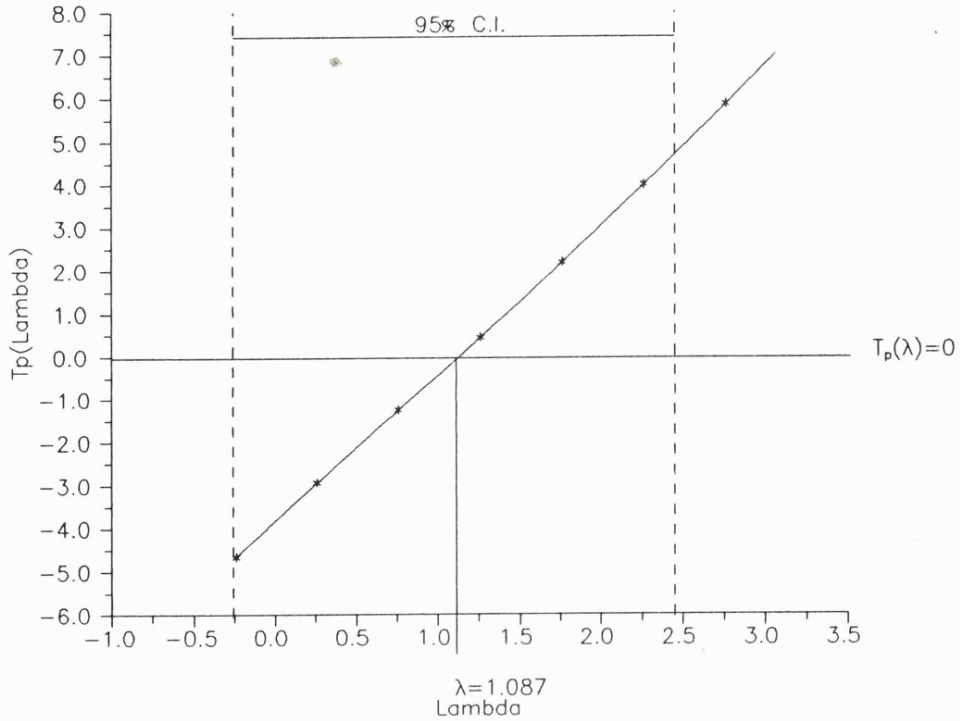
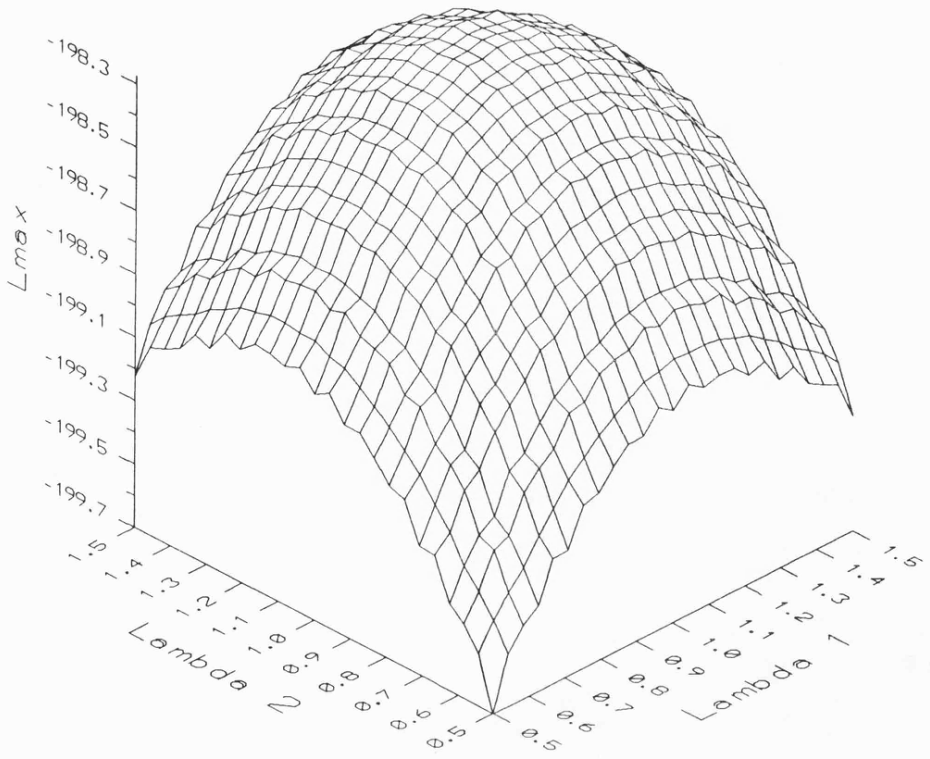


Fig 3.6 Loglikelihood Surface Plot for BVN data



ALGORITHM II. SURCON ANALYSIS ALGORITHM

The following is the summary of the SURCON analysis algorithm:

Step 1: [Initialise.] No.of iterations K=0, J=1, $\lambda_j = \lambda_{0j} = 1$, j=1,2,...,p

Step 2: [Compute.] The Box–Cox power transformations for all the variables

$$z_j(\lambda_j) = \begin{cases} \dfrac{y_i^{\lambda_j} - 1}{\lambda_j \dot{y}^{\lambda_j - 1}}, & \lambda \neq 0 \\ \dot{y} \log_e y, & \lambda = 0 \end{cases} \tag{3.89}$$

where $\dot{y}_j = (\prod_{i=1}^{n} y_{ij})^{1/n}$, j=1,2,...,p.

Step 3: [Compute.] The constructed variables

$$w_j(\lambda_j) = \frac{y_i^{\lambda_j} \log_e y_j}{\lambda_j y_j^{\lambda_j - 1}} - \frac{y_i^{\lambda_j} - 1}{\lambda_j \dot{y}^{\lambda_j - 1}}(1/\lambda_j + \log_e \dot{y}) \tag{3.90}$$

j=1,2,...,p.

Step 4: [Compute.] Deviations from the means for z and w

$$z_j^*(\lambda_j) = z_j(\lambda_j) - \bar{z}_j(\lambda_j) \tag{3.91}$$

$$w_j^*(\lambda_j) = w_j(\lambda_j) - \bar{w}_j(\lambda_j) \tag{3.92}$$

j=1,2,...,p.

Step 5: [Fit.] The single equation models

$$z_j^*(\lambda_j) = \gamma_1 w_j^*(\lambda_j) + e_{j1} \tag{3.93}$$

j=1,2,...,p.

Step 6: [Estimate.] The covariance matrix of the error terms, Σ, using the residuals from Step 5.

$$\hat{\Sigma} = \mathrm{Cov}(\hat{e}_{i1}, \hat{e}_{j1}) \tag{3.94}$$

i,j = 1,2,...,p and $e_{j1} = z_j^*(\lambda_j) - \gamma_{1j} w_j^*(\lambda_j)$.

Step 7: [Fit.] The SUR model

$$z^*(\lambda) = \gamma_2 w^*(\lambda) + e_2 \tag{3.95}$$

and obtain $\hat{\gamma}_2 = [w^{*T}(\hat{\Sigma}^{-1} \otimes I) w^*]^{-1} w^{*T}(\hat{\Sigma}^{-1} \otimes I) z^*$

Step 8: [Compute.] The Score statistics $T_p(\lambda_j)$, j=1,2,...,p.

Step 9: [Test.] For the significance of the score statistics

If $T_p(\lambda_J) > t_{n-1}(\alpha)$ then goto Step 11.

Step 10: [Test.] For overall significance of the 2SLS estimates.

$$\text{If } \hat{\delta} = \frac{SSHo - SSHA}{SSHA} < 10^{-q} \text{ then goto Step 14.}$$

Step 11: [Interpolate.] Compute new value for λ_J using the False Position Method.

If $|T_p(\lambda_J)| < 10^{-q}$ then J=J+1.

Step 12: [Increment.] K=K+1.

Step 13: [Goto.] Step 2.

Step 14: [End.] Terminate algorithm.

observed significance to be much larger than the nominal one. (See Atkinson and Lawrance, [1989]).

Figure 3.6 is the surface plot of the log–likelihood function.

Algorithm II is a summary of the SURCON algorithm.

3.5 Assessing Normality

3.5.1 Probability Plots

Probability plots as discussed in Section 2.3.2.3 are useful visual tools for assessing the conformance of a random variable to some theoretical distribution.

In assessing multivariate normality the multivariate nature of the observations can be initially ignored and a study of the marginal Q–Q plots carried out. These would be plots of the sample quantiles versus the expected quantiles from the normal distribution. If the points lie very nearly along a straight line, the normality assumption would be reasonable. Normality is suspect if the points deviate from a straight line. A further useful feature of these plots is that the patterns of deviations can provide clues about the nature of nonnormality. Once the reasons for nonnormality have been ascertained corrective action is possible e.g. by removing possible extreme observations (outliers) or by carrying out transformations to normality.

3.5.2 Rao's Score Test

Consider a regular exponential family $\mathcal{M} = \{ M(\theta): \theta \in \otimes \}$, where the distribution $M(\theta)$ has probability density function

$$f(x;\theta) = c(\theta)\exp\{ \theta^T u(x) \}, \quad x \in \mathcal{H} \tag{3.96}$$

with respect to some base measure $v(dx)$, $x \in \mathcal{H}$. Here the natural parameter θ and the sufficient statistic $u(x)$ are p×1 vectors and $\otimes$ denotes the natural parameter space. Let $X \sim M(\theta)$ be a random observation from this distribution.

Mardia and Kent [1991] consider the Rao Score test for departures from M based on another q×1 vector $v(x)$. For example, if $\mathcal{M}$ is the normal family $u(x) = (x, x^2)^T$, $x \in \mathbb{R}^1$ and if $v(x) = x^3$, then we are testing for skewness.

Under the model (3.96), define the mean vector and covariance matrix of

$$w(x) = \{ u(x)^T, v(x)^T \}^T \tag{3.97}$$

under $M(\theta)$ by

$$\mu(\theta) = \begin{bmatrix} \mu_u(\theta) \\ \mu_v(\theta) \end{bmatrix}, \quad \Omega(\theta) = \begin{bmatrix} \Omega_{uu}(\theta) & \Omega_{uv}(\theta) \\ \Omega_{vu}(\theta) & \Omega_{vv}(\theta) \end{bmatrix} \tag{3.98}$$

with inverse

$$\Omega^{-1}(\theta) = \begin{bmatrix} \Omega^{uu}(\theta) & \Omega^{uv}(\theta) \\ \Omega^{vu}(\theta) & \Omega^{vv}(\theta) \end{bmatrix} \tag{3.99}$$

Next, let $x_1,...,x_n$ denote the independent identically distributed observations from $M(\theta)$ and set $\bar{w} = (\bar{u}^T, \bar{v}^T)^T = n^{-1}\Sigma w(x_i)$. Note that u is sufficient for θ and the maximum likelihood esitimate $\hat{\theta}$ satisfies$_c$ $\hat{\mu}_u = \bar{u}$. For brevity, we shall write $\hat{\mu} = \mu(\hat{\theta})$, $\hat{\Omega} = \Omega(\hat{\theta})$ etc.

Since $\bar{u}$ is sufficient for θ it is natural to look for departures from $M(\theta)$ using the conditional distribution of v given u. Further, since $(\bar{u}^T, \bar{v}^T)^T$ is asymptotically normal, a natural test statistics is Rao's Score statistic [Rao, 1948] defined by

$$T = n(\bar{v} - \hat{\mu}_v)^T \hat{\Omega}^{vv} (\bar{v} - \hat{\mu}_v) \tag{3.100}$$

Up to first–order asymptotics, $\hat{\mu}_v$ is the conditional mean of v given u and

$$(n \hat{\Omega}^{vv})^{-1} = n^{-1} (\hat{\Omega}_{vv} - \hat{\Omega}_{vu}\hat{\Omega}_{uu}^{-1}\hat{\Omega}_{uv}) \tag{3.101}$$

is the conditional variance matrix of $\bar{v}$ given $\bar{u}$. Hence, $T \sim \chi^2_q$ asymptotically [Cox & Hinkley, 1974].

In the case of testing multivariate normality departures in a sample of p–dimensional vectors, the Rao Score statistic is based on the third and fourth moments.

Some notation is needed to express the results in a concise form.

Given a p–vector x and an integer $d \geq 1$, let

$$x_{(d)} = \{ x_{i1}, \cdots, x_{id} : 1 \leq i_1 \leq \cdots \leq i_d \leq p \} \tag{3.102}$$

$$x_{[d]} = \{ x_{i1}, \cdots, x_{id} : 1 \leq i_1, \cdots, i_d \leq p \} \tag{3.103}$$

each arranged as a column vector in lexicographic order, say. Thus, $x_{(d)}$ is a vector of the distinct monomials of degree d formed of components of x. The number of elements in $x_{(d)}$ is $\begin{bmatrix} d+p-1 \\ p-1 \end{bmatrix} = p(d)$, say.

On the other hand $x_{[d]}$ contains repeated copies of some of the elements of $x_{(d)}$. Also, $x_{[d]}$ can be thought of as the Kronecker product of x with itself d times and contains

p^d elements. For example if p=2, $x_{(1)} = x_{[1]} = x = (x_1, x_2)^T$, whereas $x_{(2)} = (x_1^2, x_1x_2, x_2^2)$ and $x_{[2]} = (x_1^2, x_1x_2, x_1x_2, x_2^2)^T$.

A partition of d is a is a collection of positive integers arranged in nonincreasing order, $L = (l_1, ..., l_k)$ say, such that $l_1 + ... + l_k = d$. Each element of $x_{(d)}$ can be written in the form $x_{i_1}^{l_1}...x_{i_k}^{l_k}$ for some partition L and some collection of distinct indices $i_1,...,i_k \in \{1,...,p\}$. Say that such an element of $x_{(d)}$ is of type L. For fixed p and d, let $m(L)$ denote the number of elements of $x_{(d)}$ of type l and let the multinomial coefficient

$$r(L) = \frac{d!}{l_1!...l_2!} \tag{3.104}$$

denote the number of times an element of type L in $x_{(d)}$ is repeated in $x_{[d]}$. Clearly, for fixed p and d, $\Sigma\, m(L) = p(d)$ and $\Sigma\, m(L)r(L) = p^d$, where the sum is over partition types L. Table 3.2 lists the values of L, $m(L)$ and $r(L)$ for d=3 and d=4.

Now let x denote an observation from the multivariate normal distribution (MVN) $N_p(\delta, A)$ with mean δ and covariance matrix A. The sufficient statistic

$$u(x) = (x_{(1)}^T, x_{(2)}^T)^T \tag{3.105}$$

To test for departures from multivariate normality we use $v(x) = (x_{(3)}^T, x_{(4)}^T)$. To construct Rao's Score statistic we require the mean of $v(x)$ and the residual covariance matrix of $v(x)$ after fitting a linear regression on $u(x)$. Further, without loss of generality we may calculate the moments under the assumption $\delta = 0$ and $A = I_p$, where I_p denotes the p×p identity matrix.

Using the symmetry of the normal distribution makes the mean of $v(x)$ easy to calculate. If the partition type of an element $x_{(d)}$ contains an odd power, then the mean of that element must be equal to 0. In particular all the elements of $x_{(3)}$ have mean 0. The remaining means can be deduced directly from the formula for even–power expectations for the standard normal distribution ie. if $Z \sim N(0,1)$,

$$E(Z^{2k}) = 1 \times 3 \times ... \times (2k-1) \quad (k \geq 1) \tag{3.106}$$

Thus, elements of $x_{(4)}$ of partition types $L = (4)$ and $L = (2,2)$ have means 3 and 1 respectively and all other elements of $x_{(4)}$ have mean 0. Let $\mu_{[4]}$ be p^4–vector arranged in the same order as $x_{[4]}$ containing these mean values.

Table 3.2 Types of elements in $x_{(3)}$ and $x_{(4)}$; d, degree; L partition type; $m(L)$, number of such elements in $x_{(d)}$; $r(L)$, number of times each element is repeated in $x_{(d)}$.

d	L	m(L)	$r(L)$
3	(3)	p	1
3	(2,1)	p(p–1)	3
3	(1,1,1)	p(p–1)(p–2)/6	6
4	(4)	p	1
4	(3,1)	p(p–1)	4
4	(2,2)	p(p–1)/2	6
4	(2,1,1)	p(p–1)(p–2)/2	12
4	(1,1,1,1)	p(p–1)(p–2)(p–3)/24	24

The residual covariance matrix of $v(x)$ after regressing on $u(x)$ is rather more tedious to evaluate but has a simple answer.

THEOREM 3.1 *Let $x \sim N_p(0,I_p)$. After regressing on $u(x)$, the elements of $v(x)$ are residually uncorrelated. Further, the residual variance of an element of $v(x)$ of degree d and partition type L is $d!/r(L)$.* □

Proof: From the symmetry of the normal distribution it follows immediately that all odd order moments are uncorrelated with all even order moments. In particular, when evaluating the residual variance of $x_{(3)}$, regressing on $x_{(1)}$ and $x_{(2)}$ is the same as on $x_{(1)}$. Similarly when evaluating the residual variance of $x_{(4)}$, regressing on $x_{(1)}$ and $x_{(2)}$ is the same as regressing on $x_{(2)}$. Further all the elements of $x_{(3)}$ are residually uncorrelated with all the elements of $x_{(4)}$, after regressing on $x_{(1)}$ and $x_{(2)}$.

For the remaining calculations it is necessary to use brute force. Although the answer is elegant, there does not seem to be a simple method of proof. As an example of the sort of calculation needed, we evaluate the residual covariance of $x_1^2x_2$ and $x_2x_3^2$, after regressing on $x_{(1)}$. This quantity is given by

$$\mathrm{E}(x_1^2x_2^2x_3^2) - \mathrm{E}[x_1^3x_2 \;\; x_1^2x_2^2, \; x_1^2x_2x_3]I_3^{-1}\mathrm{E}\begin{bmatrix} x_1x_2x_3^2 \\ x_2^2x_3^2 \\ x_2x_3^3 \end{bmatrix}$$

$$= 1 - [0, 1, 0]\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = 0 \tag{3.107}$$

It should be noted that this theorem does not generalise to d > 4, at least not in the form relevant for our purpose. If we let A_d denote the residual covariance matrix of $x_{(d)}$ after regressing on $x_{(1)}$ and $x_{(2)}$, for d=3, 4 then A_d is a p(d) × p(d) diagonal matrix with typical element $d!/R(L)$, where L is the partition type of the corresponding element of $x_{(d)}$.

□

All the information required to compute Rao's score statistic is now available. Let

$$\{ x_r; r = 1, ..., n \}$$

be a sample of p–vectors with elements x_{ri} (i=1,...,p) and d–th order powers $x_{(d),r}$ and $x_{[d],r}$. Transform the data to $z_r = \sqrt{S^{-1}}(x_r - \bar{x})$, where $\bar{x}$ and S are the sample mean vector and sample covariance matrix. Let $\bar{z}_{(d)} = n^{-1}\Sigma\, z_{(d),r}$ and $\bar{z}_{[d]} = n^{-1}\Sigma\, z_{[d],r}$. Then Rao's score statistic takes the form

$$\begin{aligned} s1 &= n\{ \bar{z}^T_{(3)}A^{-1}_{(3)}\bar{z}_{(3)} + (\bar{z}_{(4)} - \mu_{(4)})^T A^{-1}_{(4)}(\bar{z}_{(4)} - \mu_{(4)}) \} \\ &= n\left\{ \frac{1}{3!}\bar{z}^T_{[3]}\bar{z}_{[3]} + \frac{1}{4!}(\bar{z}_{[4]} - \mu_{[4]})^T(\bar{z}_{[4]} - \mu_{[4]}) \right\} \\ &= T_3 + T_4, \text{ say} \end{aligned} \tag{3.108}$$

Under H_0, $T_3 \sim \chi^2_{p(3)}$ and $T_4 \sim \chi^2_{p(4)}$, independtly of one another so $T \sim \chi^2_{p(3)+p(4)}$. The representation of T using square brackets is the simplest to write down. The second line follows from the first line in (P.4) because the diagonal elements of $d!A_d^{-1}$ just count the number of replications of each component of $x_{(d)}$ in $x_{[d]}$ so that

$$d!x^T_{(d)}A_d^{-1}x_{(d)} = x^T_{[d]}x_{[d]} \tag{3.109}$$

It can be shown that

$$T_3 = nb_{1,p}/6 \tag{3.110}$$

$$T_4 = n\{ b^*_{2,p} - 6b_{2,p} + 3p(p+2) \}/24 \tag{3.111}$$

where

$$b_{1,p} = \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} D^3_{rs}$$

$$b_{2,p} = \frac{1}{n}\sum_{r=1}^{n} D^2_{rr}$$

$$b^{*}_{2,p} = \frac{1}{n}\sum_{r=1}^{n}\sum_{s=1}^{n} D^4_{rs}$$

with $D_{rs} = (x_r - \bar{x})^T S^{-1}(x_s - \bar{x}) = z_r^T z_s^T$. The quantities $b_{1,p}$ and $b_{2,p}$ are the multivariate skewness and kurtosis introduced by Mardia [1970]. These expressions are useful in computing the valuse of T_3 and T_4. Also, $b_{2,p}$ is asymptotically normally distributed with mean $\gamma_1 = p(p + 2)$ and variance $\gamma_2 = 8p(p + 1)/n$.

Note that $b_{2,p}$ depends on the fourth–order moments only through the elements D^2_{rr}, the 'radial' part of the data. Thus we might expect $b_{2,p}$ to be more powerful than T_4 in picking up departures from multivariate normality (MVN) in elliptical families. Further, it is possible to partition the chi–squared statistic T_4 into two asymptotically independent pieces,

$$T_4 = Z^2 + (T_4 - Z^2) \qquad (3.112)$$

where $Z = (b_{2,p} - \gamma_1)/\sqrt{\gamma_2}$ has one degree of freedom and $T_4 - Z^2$ has $p(4) - 1$ degrees of freedom.

3.6 Examples

This section gives examples of the techniques discussed in this chapter. Each of the data sets has been specially selected to demonstrate the behaviour and consequences of the techniques under a variety of attributes which typical data may have.

The first two data sets, Examples E.6 and E.7, are taken from Section 2.8 (Examples E.1 and E.2, respectively) and are used as control data to show the expected behaviour of the techniques under known predetermined conditions. The first of these, Example E.6, is simulated bivariate normal data with 50 observations and does not contain any obvious outliers; it is used as the null data set. The second, Example E.7, is a contaminated version of the first data set with the outliers introduced as described in Section 2.8. This data set is chosen to see how the outliers affect the need for transformations in general and the effects of the individual types of outliers on these

transformations. The rest of the data sets, apart from Examples E.11 and E.12, are well known and have been widely used in multivariate analysis literature.

In Example E.8 we have bivariate observations on the weights and heights of 39 Peruvian Indians [Ryan et al., 1976]. From previous analysis it is known that one of the observations is an outlier. So these data are used to demonstrate the effect of outliers on the transformations using real data (a real data complement of Example E.7).

Example E.8 is data taken from Ryan et al. [1976] and is commonly referred to as the Minitab Tree Data. It contains the volumes in cubic feet and the heights in feet of 31 black cherry trees. This data set is used to compare the results obtained from the multivariate transformations techniques with those from a regression approach as carried out by Atkinson [1985].

The data in Example 10 are the widely used Fisher's Iris data [Anderson, 1935; Fisher, 1936] (See eg. Mardia et al., [1979]: pp.6–7). They consist of 50 quadrivariate observations of three species of iris (Iris setosa, Iris versicolor and Iris virginica). The variables are measurements in centimeters of the sepal length and width and of petal length and width. This data set is selected because of its well studied and known properties. It is known to be generally well behaved with no particular peculiarities although the Iris setosa has been found to be distinguishable from the other two species. For this reason the discussion on this data set is centered on this specie.

Example E.11 is based on data from the representative soil sampling survey of arable and grassland fields in England and Wales between 1969 and 1973 carried out by the Rothamsted Experimental Research Station to study the pH nutrient status of the soils. This data was kindly provided by the Rothamsted Experimental Research Station and the orgininal data consists of samples taken from all the regions in England and Wales with replicates over the years. For the purposes of the thesis data from only one replicate of one region is used and a subset of five variables analysed. These are the pH values of water (H_2O) and calcium chloride ($CaCl_2$), the available Phosphorus (P), Potassium (K) and

TABLE 3.3 Hinkley's Quick Transformations to Marginal Symmetry

Data	n		T = -1	0	0.5	1	2
1.Simulated bivariate normal	50	x1	-0.239	-0.116	-0.062	-0.012*	0.080
		x2	-0.296	-0.122	-0.053	0.008*	0.119
2.Simulated bivariate normal (with 4 outliers)	50	x1	-0.323	-0.082	0.018*	0.116	0.337
		x2	-0.340	-0.119	-0.034*	0.042	0.179
3.Peruvian	39	x1	-0.005*	0.026	0.043	0.060	0.096
		x2	0.056**	0.065	0.070	0.074	0.083
4.Peruvian (minus obs 39)	38	x1	-0.025	0.000*	0.013	0.026	0.051
		x2	0.034**	0.043	0.048	0.053	0.062
5.Minitab tree	31	x1	-0.072	0.000*	0.035	0.070	0.138
		x2	-0.068	-0.033	-0.016	0.000*	0.032
6.Fisher's iris (Iris Setosa)	50	x1	-0.046	-0.015	-0.000*	0.015	0.045
		x2	-0.030	0.014*	0.035	0.056	0.096
		x3	-0.305	-0.245	-0.217	-0.190	-0.138**
		x4	0.126**	0.308	0.385	0.460	0.628
7.Fisher's iris (All groups)	150	x1	-0.054	-0.011*	0.011*	0.033	0.079
		x2	-0.009*	0.054	0.084	0.115	0.176
		x3	-0.329	-0.255	-0.213	-0.169	-0.073**
		x4	-0.451	-0.243	-0.154	-0.067*	0.103
8.Repeat soil sample survey	57	x1	-0.095	-0.053	-0.031	-0.010*	0.033
		x2	-0.107	-0.062	-0.039	-0.015*	0.034
		x3	-0.250	0.073*	0.209	0.363	0.887
		x4	0.066**	0.203	0.289	0.393	0.706
		x5	-0.150	0.064*	0.183	0.321	0.722

* - selected value for transformation parameter.
** - actual transformation parameter is greater/less than selected value.

Note: T is the transformation parameter in the expression $x_T = (x^T - 1)/T$ that gives approximate symmetry and the displayed values are $d_T = (\bar{x}_T - x_T[m])/s_T$, where $\bar{x}_T$ is the sample mean, $x_T[m]$ the sample median and s_T the inter-quartile range all evaluated at T.

TABLE 3.4 Summary of the Box-Cox Transformations to Joint Normality

Data	n		Rao's	Method: MLE		SURCON	
			τ	$\hat{\lambda}$	$\tilde{\lambda}$	$T_p(1)$	$\hat{\lambda}$
1.Simulated bivariate normal	50	x1	6.0587	0.73	0.29	1.0755	0.73
		x2	(0.7340)	0.99	0.91	0.1519	0.99
2.Simulated bivariate normal (with 4 outliers)	50	x1	56.1905	0.35	-0.30	5.5704*	0.34
		x2	(0.0000)	0.78	0.15	1.8051	0.78
3.Peruvian	39	x1	73.9785	-3.21	-5.60	5.6675*	-3.08
		x2	(0.0000)	3.99	10.33	-1.0993	4.09
4.Peruvian (minus obs 39)	38	x1	6.4314	-1.36	-5.18	2.2197	-1.34
		x2	(0.6961)	5.11	12.44	-1.3770	5.13
5.Minitab tree	31	x1	11.5487	-0.16	-0.63	5.8372*	-0.16
		x2	(0.2400)	2.31	-0.82	3.6725*	2.33
6.Fisher's iris (Iris Setosa)	50	x1	25.9760	0.36	-0.22	0.5435	0.40
		x2	(0.4089)	1.26	1.41	-0.3676	1.25
		x3		0.66	0.15	0.6388	0.69
		x4		-	-	-	-
7.Fisher's iris (Iris Versicolour)	50	x1	48.3249	-0.80	-1.51	1.2005	-0.65
		x2	(0.7257)	2.51	4.10	-2.1307*	2.53
		x3		2.26	3.18	-2.1184*	2.36
		x4		0.80	0.90	0.1073	0.81
7.Fisher's iris (Iris Virginica)	50	x1	60.1830	1.15	-0.04	0.6777	1.08
		x2	(0.2937)	-0.01	-1.09	1.5402	0.01
		x3		-0.70	-2.66	2.5136*	-0.83
		x4		1.40	1.81	-0.5343	1.37
8.Repeat soil sample survey	57	x1	9.3435	-0.49	-2.84	2.9243*	-0.61
		x2	(0.4062)	-0.22	-2.32	2.8398*	-0.31
		x3	1558.0977	-0.03	0.03	13.2128*	-0.03
		x4	(0.0000)	-0.81	-0.43	12.9889*	-0.82
		x5		-0.26	-0.22	12.2840*	-0.26
		x1	1783.3890	-0.45	-3.11	3.2194*	-0.55
		x2	(0.0000)	-0.12	-2.54	3.1244*	-0.30
		x3		-0.17	0.08	12.9942*	0.02
		x4		-0.88	-0.46	13.7725*	-0.89
		x5		-0.23	-0.23	12.3944*	-0.23

Notes: 1/ The "quick estimate" $\tilde{\lambda}$ is obtained by taking $\lambda_0 = 1$, i.e. no transformation.

2/ A * indicates that $T_p(1)$ is significant at level $\alpha = 0.05$ with n-1 degrees of freedom.

3/ The terms in brackets under Rao's τ indicate the α' values from $P[T \leq \tau] = 1-\alpha'$, where $\tau \rightarrow \chi^2(p[3] + p[4])$, $p[d] = {}^{d+p-1}C_{p-1}$ and p is the number of variables.

Magnesium (Mg). The data set was selected to study the new techniques on a fresh real data set i.e. with unknown properties from a multivariate outliers and transformations perspective.

The final example, Example E.12, consists of 50 sets of computer generated bivariate normal deviates. Pairs of these random deviates were transformed to obtain 50 new samples with induced correlation as in Example E2.1. A range of values for ρ was used to provide a basis for comparing the different approaches discussed for transforming observations and seeing how correlation affects them.

Tables 3.2 and 3.3 are the summaries of the estimates obtained from the examples using Hinkley's quick transformations to marginal symmetry and the SURCON analysis, respectively.

Appendix C contains the listings of the data used in this chapter (excluding data already exhibited in Chapter Two).

EXAMPLE E.6 Simulated Bivariate Normal data.

This example is based on the data set from Example E.1 in Chapter Two and consists of 50 two dimensional computer generated normal samples. It is used as the "null" data set to study and demonstrate the analysis under known, predetermined conditions. The summary statistics and Stalactite analysis are discussed in Example E.1.

As a first step in carrying out the analysis, Hinkley's "quick" transformations to marginal symmetry are performed. The best choice for the transformation parameters in both cases is T=1 (See Table 3.3) indicating that no transformation is required as should be evident from the nature of the data. The next phase is to test for joint normality using Rao's Score test (referred to as Rao's τ in Table 3.4) which for joint normality is also not significant so the joint normality assumption for the two variables is viable. The log–likelihood method and the SURCON method suggest the same transformation parameters which for both variables are in the vicinity of unity. However, λ_1 is further

away from 1 and this is due to the fact that X_1 has less variability. The "quick" estimates for joint normality are not significant although they still provide some insight as to the possible transformations and can also be used as starting points to both the log–likelihood ans SURCON methods.

The three types of tests for symmetry plots for X_1 are shown in Figure E.6(a) and they all confirm its symmetry. Those of X_2 (not displayed) also yield the same conclusions.

Figure E.6(b) displays the full output from the SURCON analysis run. The lambda values shown are those obtained when the method converges to the maximum likelihood estimates (MLE). The left–hand side of the table shows the single equation (first stage least squares) estimates which are the marginal transformation parameter estimates. The right–hand part of the table shows the SURCON (two stage least squares) estimates i.e. the parameter estimates for joint normality. There is slight numerical difference between the two sets of parameters so the data are marginally and jointly normal. A test for the joint significance of the slopes, the gammas, is constructed using the $\hat{\delta}$ statistic (Section 3.4.3.1a) denoted here by the F–Statistic. The output contains the p–values from the F distribution and the χ^2 distribution. The 95% confidence intervals for the marginal and joint estimates for each variable are also displayed. In this example the value 1 (no transformation) is firmly included in both variables. Finally, the Lagrange Multiplier (LM) test for independence of the variables (Section 3.4.3.1b) is included since it helps in ascertaining whether or not it is worthwhile to consider searching for joint estimates. The LM of 17.5389 is highly significant and so although the variables are marginally normal the fact that they are correlated makes it desirable to consider their joint estimates.

The output also displays the total number of iterations taken to converge to the maximum likelihood estimates for a given tolerance factor, ϵ where $|\hat{\delta}| \leq \epsilon$ (in this and subsequent examples $\epsilon = 10^{-4}$). The number of iterations in this example is 6.

Figure E.6(a) Symmetry Plots for X1 in Example E.6

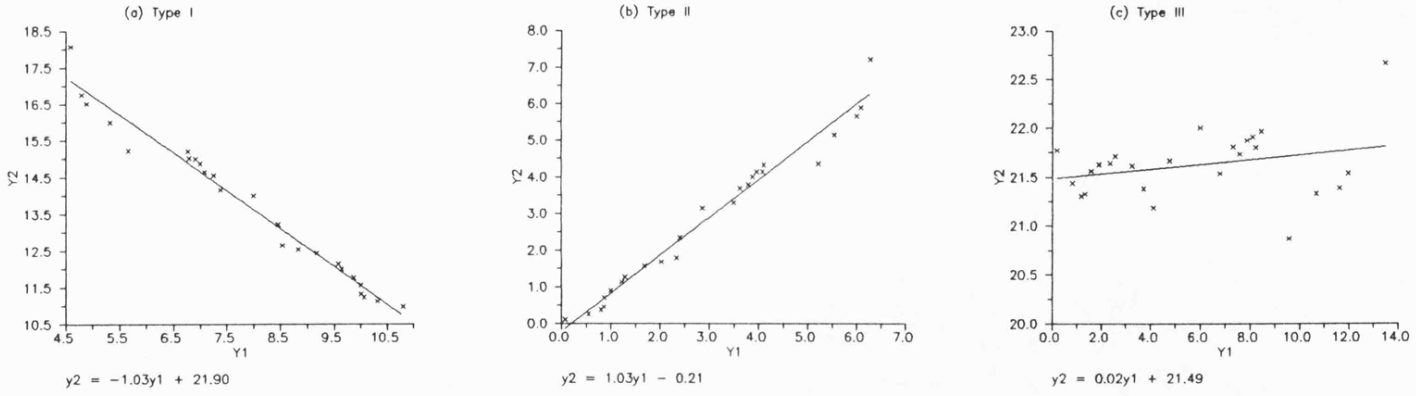


FIGURE E.6(b)

SURCON ANALYSIS

ESTIMATES OF TRANSFORMATION PARAMETERS FOR DATA: BIVARIATE NORMAL DATA

LAMBDA = 0.73 0.99

	SINGLE EQUATION ESTIMATES				SUR ESTIMATES			
VAR.	GAMMA	S.E.	T-VALUE	EST. LAMBDA	GAMMA	S.E.	T-VALUE	EST. LAMBDA
1	-0.1325	0.7910	-0.1676 (0.4338)	0.8666	-0.0024	0.6554	-0.0036 (0.4986)	0.7364
2	0.1175	0.7026	0.1672 (0.4339)	0.8755	0.0000	0.5821	0.0000 (0.5000)	0.9930

*** TERMS IN BRACKETS ARE p-VALUES FOR T(N-1) ***

*** JOINT TEST STATISTIC FOR H0: G1 = G2 = ...= GP = 0

F-STATISTIC = 0.0000 D.O.F. = 2, 98

p-VALUE: F-DISTRIBUTION = 1.0000 D.O.F. = 2, 98

CHI-DISTRIBUTION/P = 0.5000 D.O.F. = 2

*** CONFIDENCE INTERVALS

VAR	95.0% C.I. LS	95.0% C.I. 2SLS
1	[-1.4588, 1.1937] (-0.4596, 2.1928)	[-1.1011, 1.0964] (-0.3624, 1.8351)
2	[-1.0605, 1.2955] (-0.3025, 2.0535)	[-0.9760, 0.9760] (0.0171, 1.9690)

*** TERMS IN SQUARE BRACKETS ARE GAMMAS AND ROUND BRACKETS ARE LAMBDAS ***

*** TEST STATISTIC FOR H0: DIAGONAL COVARIANCE MATRIX
USING THE LAGRANGE MULTIPLIER TEST (LM)

LM-STATISTIC = 17.5389 D.O.F. = 1

p-VALUE FOR CHI-SQUARE WITH 1 D.O.F. = 0.0000

TOTAL NO OF ITERATIONS TO CONVERGE = 6 WITH TOLERANCE FACTOR 10.0**(-4)
****

EXAMPLE E.7 Simulated Bivariate Normal data (with 4 outliers).

This is the contaminated form of the data in the preceeding example (See Example E.2) where four outliers are introduced. Two of the four are outlying in both variables whereas the other two outlie in the extremes of X_1 only. The purpose of this example is to demonstrate the effects of the types of outliers on the transformations. The identities of the four observations are known, namely, 16, 28 (outlying in both variables), 39 and 48 (outlying in X_1 only).

The square root transformation is suggested by Hinkley's "quick" transformations for both variables which is reasonable since it would shrink the tails to the right due to observations 16 and 28. Rao's Score test is highly significant, hence, joint non–normality is suspected. The existance of only 4 outliers has distorted the otherwise jointly normal data from Example E.6. The "quick" estimate for λ_1 is significant whereas that of λ_2 is not. This suggests that the lack of fit to normality is due to the marginally outlying observations in X_1 (observations 39 and 48) and these are the very observations which affect the correlation between the two variables. The log–likelihood and SURCON estimates for λ_1 are significantly reduced from 0.73 to 0.34 but the change in λ_2 is not as dramatic. There is also a wider difference between the marginal estimates and the joint estimates for λ_1 which indicates that in this case marginal normality would not ensure joint normality. The 95% confidence interval for λ_1 excludes no transformation and this is further exhibited by the marginal one which firmly excludes unity whereas both confidence intervals for λ_2 include unity.

The LM statistic (Figure E.7(a)) is not as significant as before which again is influenced by the marginally outlying observations. It is also interesting to note that the number of iterations increased from 6 to 12 as a result of the outliers.

This example has shown the significant influence that outliers can have on the need for transformation. It also demonstrates the ability for jointly outlying observations being obscured when carrying out transformations but the influence of marginally outlying

FIGURE E.7(a)

SURCON ANALYSIS

ESTIMATES OF TRANSFORMATION PARAMETERS FOR DATA: SIMULATED BIVARIATE NORMAL DATA (WITH 4 OUTLIERS)

LAMBDA = 0.34 0.78

	SINGLE EQUATION ESTIMATES				SUR ESTIMATES			
VAR.	GAMMA	S.E.	T-VALUE	EST. LAMBDA	GAMMA	S.E.	T-VALUE	EST. LAMBDA
1	0.3234	0.3840	0.8421 (0.2019)	0.0166	-0.0014	0.3415	-0.0042 (0.4984)	0.3414
2	0.2702	0.5518	0.4897 (0.3133)	0.5065	0.0000	0.4907	0.0000 (0.5000)	0.7767

*** TERMS IN BRACKETS ARE P-VALUES FOR T(N-1) ***

*** JOINT TEST STATISTIC FOR H0: G1 = G2 = ...= GP = 0

F-STATISTIC = 0.0000 D.O.F. = 2, 98

p-VALUE: F-DISTRIBUTION = 1.0000 D.O.F. = 2, 98

CHI-DISTRIBUTION/P = 0.5000 D.O.F. = 2

*** CONFIDENCE INTERVALS

VAR	95.0% C.I. LS	95.0% C.I. 2SLS
1	[-0.3205, 0.9672]	[-0.5740, 0.5711]
	(-0.6273, 0.6604)	(-0.2312, 0.9139)
2	[-0.6549, 1.1953]	[-0.8227, 0.8227]
	(-0.4186, 1.4316)	(-0.0460, 1.5994)

*** TERMS IN SQUARE BRACKETS ARE GAMMAS AND ROUND BRACKETS ARE LAMBDAS ***

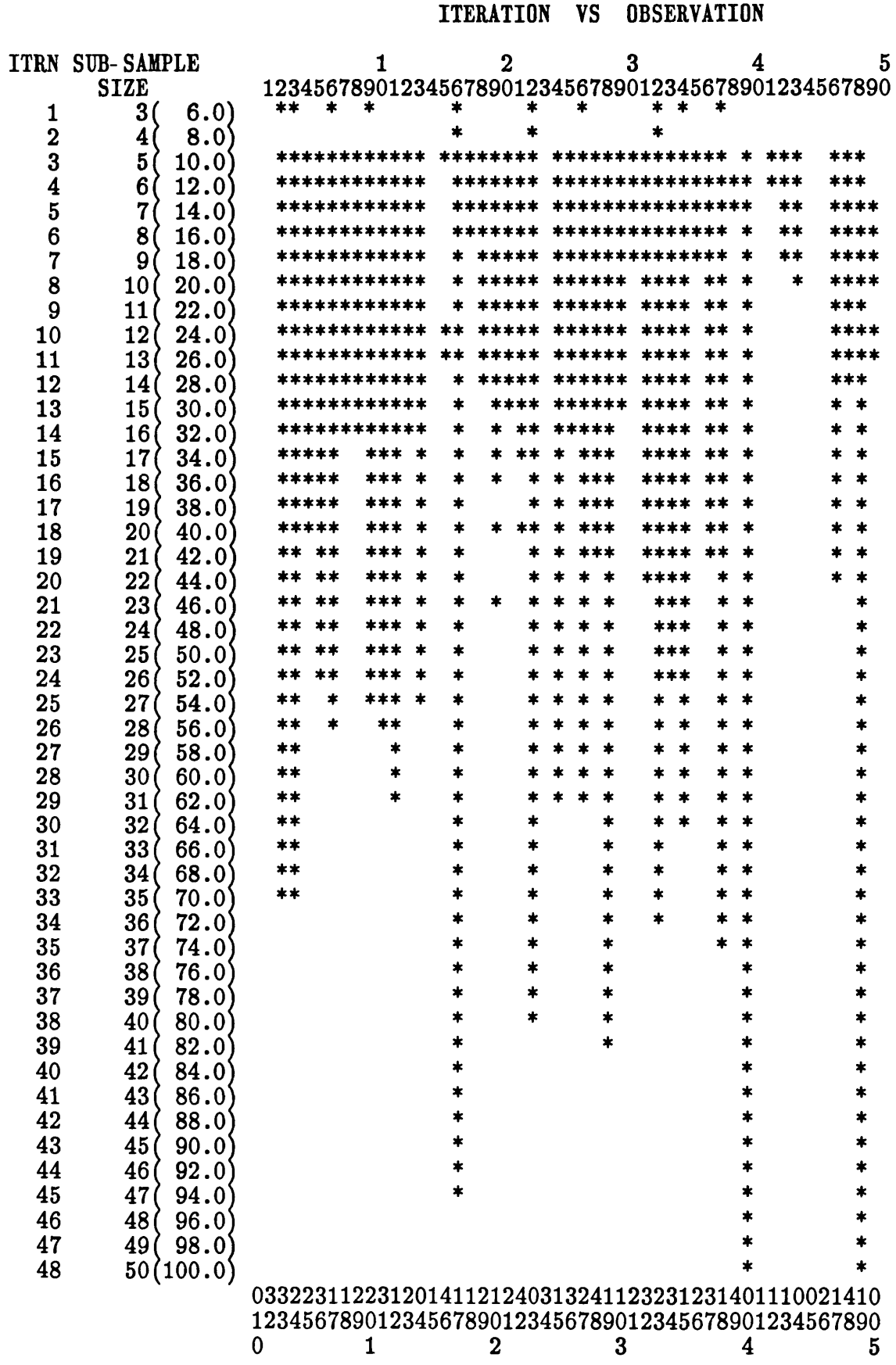
*** TEST STATISTIC FOR H0: DIAGONAL COVARIANCE MATRIX
USING THE LAGRANGE MULTIPLIER TEST (LM)

LM-STATISTIC = 13.1990 D.O.F. = 1

p-VALUE FOR CHI-SQUARE WITH 1 D.O.F. = 0.0003

TOTAL NO OF ITERATIONS TO CONVERGE = 12 WITH TOLERANCE FACTOR 10.0**(-4)
****

FIGURE E.7(b) Transformed Bivariate Normal Data (with 4 outliers) Stalactite Chart



observations can still exist. As a further display of this fact the data was transformed according to the SURCON estimates and a Stalactite analysis carried out. Figure E.7(b) is the Stalactite chart after transformation and it clearly shows observations 39 and 48 as still outlying but the outlyingness of observations 16 and 28 is greatly reduced. This example also demonstrates the ability of the Stalactite chart to highlight outliers since from Chapter Two the Classical approach to outlier detection (which corresponds to the final iteration of the Stalactite analysis) would only detect observations 39 and 48.

EXAMPLE E.8 Weights and Heights of 39 Peruvian Indians.

The data are X_1 the weights in kilograms and X_2 the heights in millimeters of 39 Peruvian Indians. [Ryan et al., 1976].

This data set is included to demonstrate the effect of outliers on transformations using real data. A scatter plot of the data (not displayed) suggests that observation 39 is an outlier.

As an initial test for marginal normality the normal probability plots were made for the weights (X_1) and heights (X_2) (See Figure E.8(a)). The probability plot for the weights looks S shaped suggesting a short–tailed marginal distribution for weight. The plot for heights, however, is reasonably linear up to a height of about 1600mm, and then there is discontinuity. On inspecting the scatter plot there are about seven taller Indians with lower than average weights in the range 61–64 kg approximately. This cluster of observations would have the effect of raising the probability plot for heights in the region of $0 < z < 0.5$.

The inspection of joint normality of the variables was also carried out using χ^2 probability plots of the Mahalanobis distances of the observations (with and without observation 39). These plots are given in Figure E.8(b). Figure E.8(b)i. is the full sample version and is linear for the smaller distances but has a distinct point far removed from the rest. In fact the effect of such a point is to compress the remaining distances which tends to make them appear to be linear. To obtain a clearer picture of the shape of the plot the

Figure E.8(a) Normal Probability Plots for the Peruvian Data
(with observation 39 omitted)

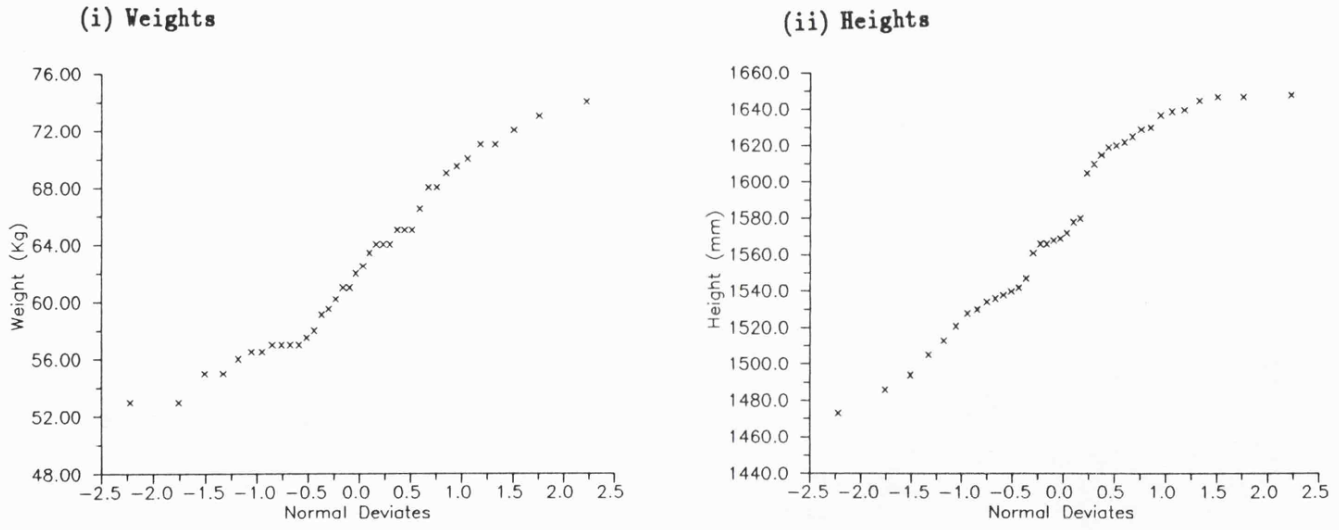


Figure E.8(b) Mahalanobis Distances χ^2 Probability Plot for the Peruvian Data

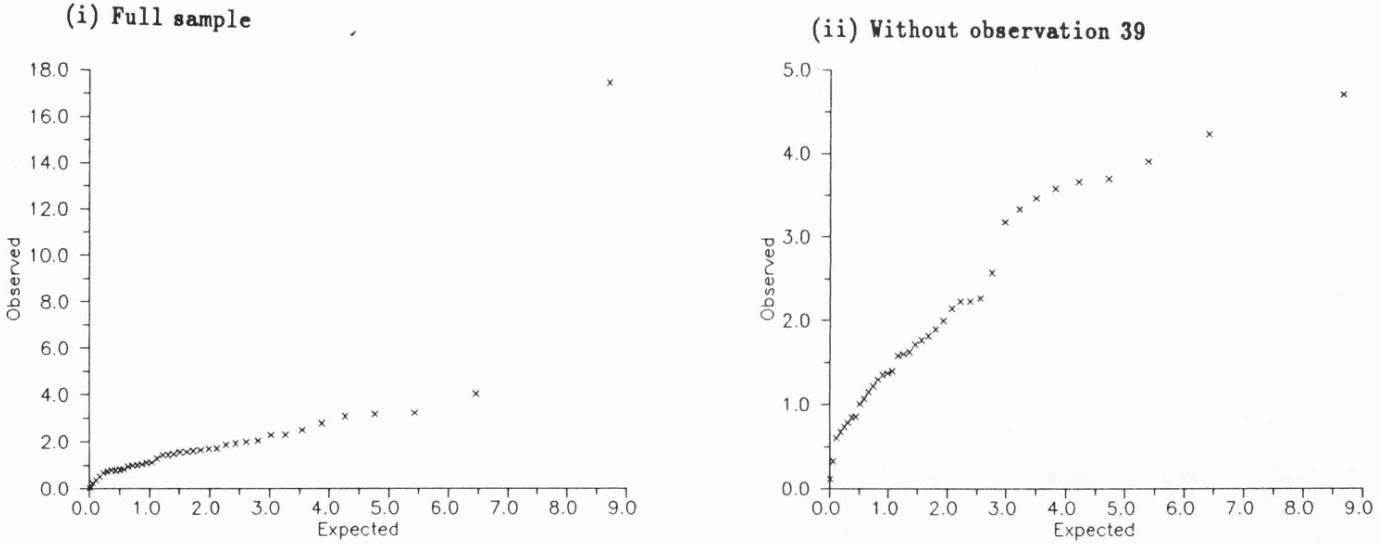


FIGURE E.8(c)

SURCON ANALYSIS

ESTIMATES OF TRANSFORMATION PARAMETERS FOR DATA: WEIGHT AND HEIGHT OF 39 PERUVIAN INDIANS

LAMBDA = -3.08 4.09

	SINGLE EQUATION ESTIMATES				SUR ESTIMATES			
VAR.	GAMMA	S.E.	T-VALUE	EST. LAMBDA	GAMMA	S.E.	T-VALUE	EST. LAMBDA
1	-2.9016	2.7450	-1.0571 (0.1486)	-0.1771	0.0058	2.4193	0.0024 (0.4990)	-3.0845
2	-3.2196	11.5581	-0.2786 (0.3910)	7.3065	-0.0002	10.1869	0.0000 (0.5000)	4.0871

*** TERMS IN BRACKETS ARE P-VALUES FOR T(N-1) ***

*** JOINT TEST STATISTIC FOR H0: G1 = G2 = ...= GP = 0

F-STATISTIC = 0.0000 D.O.F. = 2, 76

p-VALUE: F-DISTRIBUTION = 1.0000 D.O.F. = 2, 76

CHI-DISTRIBUTION/P = 0.5000 D.O.F. = 2

*** CONFIDENCE INTERVALS

VAR	95.0% C.I. LS	95.0% C.I. 2SLS
1	[-7.5295, 1.7263]	[-4.0731, 4.0847]
	(-4.8050, 4.4508)	(-7.1633, 0.9944)
2	[-22.7060, 16.2668]	[-17.1749, 17.1744]
	(-12.1798, 26.7929)	(-13.0875, 21.2618)

*** TERMS IN SQUARE BRACKETS ARE GAMMAS AND ROUND BRACKETS ARE LAMBDAS ***

*** TEST STATISTIC FOR H0: DIAGONAL COVARIANCE MATRIX
USING THE LAGRANGE MULTIPLIER TEST (LM)

LM-STATISTIC = 9.5571 D.O.F. = 1

p-VALUE FOR CHI-SQUARE WITH 1 D.O.F. = 0.0020

TOTAL NO OF ITERATIONS TO CONVERGE = 6 WITH TOLERANCE FACTOR 10.0**(-4)
****

FIGURE E.8(d)

SURCON ANALYSIS

ESTIMATES OF TRANSFORMATION PARAMETERS FOR DATA: WEIGHT AND HEIGHT OF 39 PERUVIAN INDIANS (W/O OBS 39)

LAMBDA = -1.33 5.13

	SINGLE EQUATION ESTIMATES				SUR ESTIMATES			
VAR.	GAMMA	S.E.	T-VALUE	EST. LAMBDA	GAMMA	S.E.	T-VALUE	EST. LAMBDA
1	-2.2143	3.7751	-0.5866 (0.2805)	0.8799	-0.0009	3.0093	-0.0003 (0.4999)	-1.3336
2	-2.8949	11.8638	-0.2440 (0.4043)	8.0233	-0.0002	9.4573	0.0000 (0.5000)	5.1287

*** TERMS IN BRACKETS ARE P-VALUES FOR T(N-1) ***

*** JOINT TEST STATISTIC FOR HO: G1 = G2 = ...= GP = 0

F-STATISTIC = 0.0000 D.O.F. = 2, 74

p-VALUE: F-DISTRIBUTION = 1.0000 D.O.F. = 2, 74

CHI-DISTRIBUTION/P = 0.5000 D.O.F. = 2

*** CONFIDENCE INTERVALS

VAR	95.0% C.I. LS	95.0% C.I. 2SLS
1	[-8.5833, 4.1546]	[-5.0779, 5.0761]
	(-5.4891, 7.2488)	(-6.4106, 3.7434)
2	[-22.9103, 17.1205]	[-15.9555, 15.9551]
	(-11.9921, 28.0387)	(-10.8266, 21.0840)

*** TERMS IN SQUARE BRACKETS ARE GAMMAS AND ROUND BRACKETS ARE LAMBDAS ***

*** TEST STATISTIC FOR HO: DIAGONAL COVARIANCE MATRIX
USING THE LAGRANGE MULTIPLIER TEST (LM)

LM-STATISTIC = 14.5341 D.O.F. = 1

p-VALUE FOR CHI-SQUARE WITH 1 D.O.F. = 0.0001

TOTAL NO OF ITERATIONS TO CONVERGE = 5 WITH TOLERANCE FACTOR 10.0**(-4)
****

extreme observation is deleted and the plot is repeated. The new plot is shown in Figure E.8(b)ii. and is reasonably linear apart from a curious hump near the upper end. This is no doubt due to those taller Indians with lower than average weight that caused the hump in the normal probability plot for the heights [Figure E.8(b)i.].

The probability plots provide a graphical assessment of the characteristics of the data. It is now possible to carry out a formal analysis so as to attach some numerical quantities to the findings. When the full data set is considered the reciprocal transformation is suggested for both variables to achieve marginal symmetry. The exclusion of observation 39 alters the transformations to log for X_1 but that of X_2 remains unchanged. This is verified by the fact that observation 39 was outlying in the X_1 direction. Rao's Score test is highly significant for the full sample and ceases to be significant with the exclusion of observation 39. The suggested SURCON estimates (Figures E.8(c) and E.8(d)) for λ_1 drop from -3.08 to -1.34 for the reduced sample and there is little change in that of λ_2. On comparing the single equation estimates with the SURCON estimates it can be seen that marginal normality would not imply joint normality in this case.

EXAMPLE E.9 Minitab Tree Data (volume and heights).

The data are X_1 the volumes in cubic feet and X_2 the heights in feet of 31 black cherry trees (the original data consisted of a third variable, the girth, but for the purposes of this analysis it is omitted as discussed below). It is referred to as the Minitab Tree Data because it originates from the *Minitab Student Handbook* [Ryan et al. 1976] which is an introductory statistics textbook to complement the Minitab statistical package.

Atkinson [1985] performs an analysis for transformation of the variables based on regression diagnostics where the volume is taken as the response with the other two variables as the carriers. There is very strong correlation between the volume and the girth and for that reason the latter is dropped from the present analysis. On the other hand the plot of the volume against the height produces is megaphone shaped; so to demonstrate the

FIGURE E.9

SURCON ANALYSIS

ESTIMATES OF TRANSFORMATION PARAMETERS FOR DATA: MINITAB TREE DATA (RESPONSE VS X2)

LAMBDA = -0.16 2.33

	SINGLE EQUATION ESTIMATES				SUR ESTIMATES			
VAR.	GAMMA	S.E.	T-VALUE	EST. LAMBDA	GAMMA	S.E.	T-VALUE	EST. LAMBDA
1	-0.2620	0.5938	-0.4412 (0.3311)	0.1059	0.0033	0.4902	0.0068 (0.4973)	-0.1594
2	-2.0196	3.8885	-0.5194 (0.3037)	4.3510	0.0001	3.2100	0.0000 (0.5000)	2.3313

*** TERMS IN BRACKETS ARE P-VALUES FOR T(N-1) ***

*** JOINT TEST STATISTIC FOR H0: G1 = G2 = ...= GP = 0

F-STATISTIC = 0.0000 D.O.F. = 2, 60

p-VALUE: F-DISTRIBUTION = 1.0000 D.O.F. = 2, 60

CHI-DISTRIBUTION/P = 0.5000 D.O.F. = 2

*** CONFIDENCE INTERVALS

VAR	95.0% C.I. LS	95.0% C.I. 2SLS
1	[-1.2697, 0.7458]	[-0.8286, 0.8352]
	(-0.9019, 1.1137)	(-0.9913, 0.6725)
2	[-8.6194, 4.5802]	[-5.4480, 5.4483]
	(-2.2488, 10.9508)	(-3.1168, 7.7794)

*** TERMS IN SQUARE BRACKETS ARE GAMMAS AND ROUND BRACKETS ARE LAMBDAS ***

*** TEST STATISTIC FOR H0: DIAGONAL COVARIANCE MATRIX
USING THE LAGRANGE MULTIPLIER TEST (LM)

LM-STATISTIC = 12.4285 D.O.F. = 1

p-VALUE FOR CHI-SQUARE WITH 1 D.O.F. = 0.0004

TOTAL NO OF ITERATIONS TO CONVERGE = 13 WITH TOLERANCE FACTOR 10.0**(-4)
****

SURCON technique and also compare the results thereof with those obtained from the regression approach, this combination of variables is more appealing.

Hinkley's "quick" estimates (Table 3.3) for transformations to marginal symmetry suggest a log transformation for X_1 and no transformation for X_2. This is reasonable considering that the volume is a cubic quantity and the heights would be expected to be symmetrical. Rao's Score test (Table 3.4) is not significant so joint normality can be assumed. However, from the scatter plot, the data do not exhibit a good elliptical shape so estimates to achieve joint normality would be desirable. Both the "quick" estimates $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ are highly significant although they do not provide reasonable values. The SURCON estimates (Figure E.9) suggest the log for X_1 and the square for X_2. These values compare favourably with those obtained by the regression approach especially for the volume. The regression approach provides a number of possible transformations for the volume depending on whether a first order, second order or log–log model is fitted. In our case although one carrier variable is missing the confidence intervals obtained include all the transformations derived i.e. from log to the cube–root for the volume and no transformation to log in the heights.

This example shows that joint transformations can be used to obtain guidelines for the actual transformations which would be used in fitting regression models.

EXAMPLE E.10 Fisher's Iris Data.

The data consist of 50 quadrivariate (p=4) observations of three species of iris (Iris setosa, Iris versicolor and Iris virginica). The variables are the sepal length and width and the petal length and width all in centimeters. This is a well known data set in multivariate literature and has been used by several authors as the basis for testing different classification and clustering algorithms (e.g. Friedman and Rubin, 1967). The data set is considered to be well behaved with no peculiarities and it has been found that Iris setosa is

easily distinguishable from the other two species [See Fisher, 1936; Friedman and Rubin, 1967; Gnanadesikan, 1977: pp.217–222]. It is for this reason that only the Iris setosa analysis will be dsicussed in detail for this example although the results for the other two species are displayed (See Table 3.4; Figures E.10(d) and E.10(e)).

For the Iris setosa data X_1 and X_2 are the sepal lengths and widths respectively and X_3 and X_4 the petal lengths and widths. The transformations suggested for marginal symmetry are square root, log, square and reciprocal for X_1 to X_4 respectively. On inspecting the data X_1 and X_2 have right hand tails and so these transformations would shrink these tails. X_3 has slight variation about above unity and so to stretch it out a square transformation would be in order. On the other hand X_4 has very slight variation above zero and so in order to stretch it out the reciprocal transformation would be required.

For subsequent analysis X_4 shall be dropped because its lack of variabilty causes singularity in the covariance matrix. Rao's test is not significant which means that the data exhibit joint normality on the reduced variable space. In particular, the joint estimates for multivariate normality are 0.4, 1.25 and 0.69 (Figure E.10(a)) which are the square root, no transformation and square–root/no–transformation for X_1, X_2 and X_3 respectively. The joint estimates are significantly different from the marginal estimates so marginal normality does not imply joint normality for the data. The confidence intervals for all the variables firmly includes unity. Considering the joint estimates is worthwhile since the LM statistic is highly significant. The total number of iterations required to converge to the maximum likelihood estimator (from λ_0=1 for all variables) is only 9 at a tolerance factor of 10^{-4}. The results for the other two species is displayed in Figures E.10(d) and E.10(e).

This data set is further used to test the effect of deletion of observations on the transformation paramater estimates. As a first step it is necessary to study the level of outlyingness in the data (if any) before examining the observation (case) deletions. A

FIGURE E.10(a)

SURCON ANALYSIS

ESTIMATES OF TRANSFORMATION PARAMETERS FOR DATA: FISHER'S IRIS DATA (SETOSA) (W/O PETAL WIDTH)

LAMBDA = 0.40 1.25 0.69

	SINGLE EQUATION ESTIMATES				SUR ESTIMATES			
VAR.	GAMMA	S.E.	T-VALUE	EST. LAMBDA	GAMMA	S.E.	T-VALUE	EST. LAMBDA
1	0.3238	3.2473	0.0997 (0.4605)	0.0798	0.0151	2.3053	0.0065 (0.4974)	0.3885
2	0.7601	1.5559	0.4885 (0.3137)	0.4856	-0.0031	1.1295	-0.0027 (0.4989)	1.2487
3	-0.4556	1.4266	-0.3194 (0.3754)	1.1418	0.0000	1.3999	0.0000 (0.5000)	0.6862

*** TERMS IN BRACKETS ARE P-VALUES FOR T(N-1) ***

*** JOINT TEST STATISTIC FOR HO: G1 = G2 = ...= GP = 0

F-STATISTIC = 0.0025 D.O.F. = 3, 147

p-VALUE: F-DISTRIBUTION = 0.9998 D.O.F. = 3, 147

CHI-DISTRIBUTION/P = 0.3333 D.O.F. = 3

*** CONFIDENCE INTERVALS

VAR	95.0% C.I. LS	95.0% C.I. 2SLS
1	[-5.1205, 5.7681] (-5.3645, 5.5241)	[-3.8499, 3.8800] (-3.4764, 4.2535)
2	[-1.8484, 3.3685] (-2.1229, 3.0940)	[-1.8968, 1.8906] (-0.6450, 3.1424)
3	[-2.8475, 1.9362] (-1.2500, 3.5336)	[-2.3470, 2.3470] (-1.6608, 3.0331)

*** TERMS IN SQUARE BRACKETS ARE GAMMAS AND ROUND BRACKETS ARE LAMBDAS ***

*** TEST STATISTIC FOR HO: DIAGONAL COVARIANCE MATRIX
USING THE LAGRANGE MULTIPLIER TEST (LM)

LM-STATISTIC = 32.7495 D.O.F. = 3

P-VALUE FOR CHI-SQUARE WITH 3 D.O.F. = 0.0000

TOTAL NO OF ITERATIONS TO CONVERGE = 9 WITH TOLERANCE FACTOR 10.0**(-4)
****

FIGURE E.10(b) Fisher's Iris Data (Setosa without X_4) Stalactite Chart

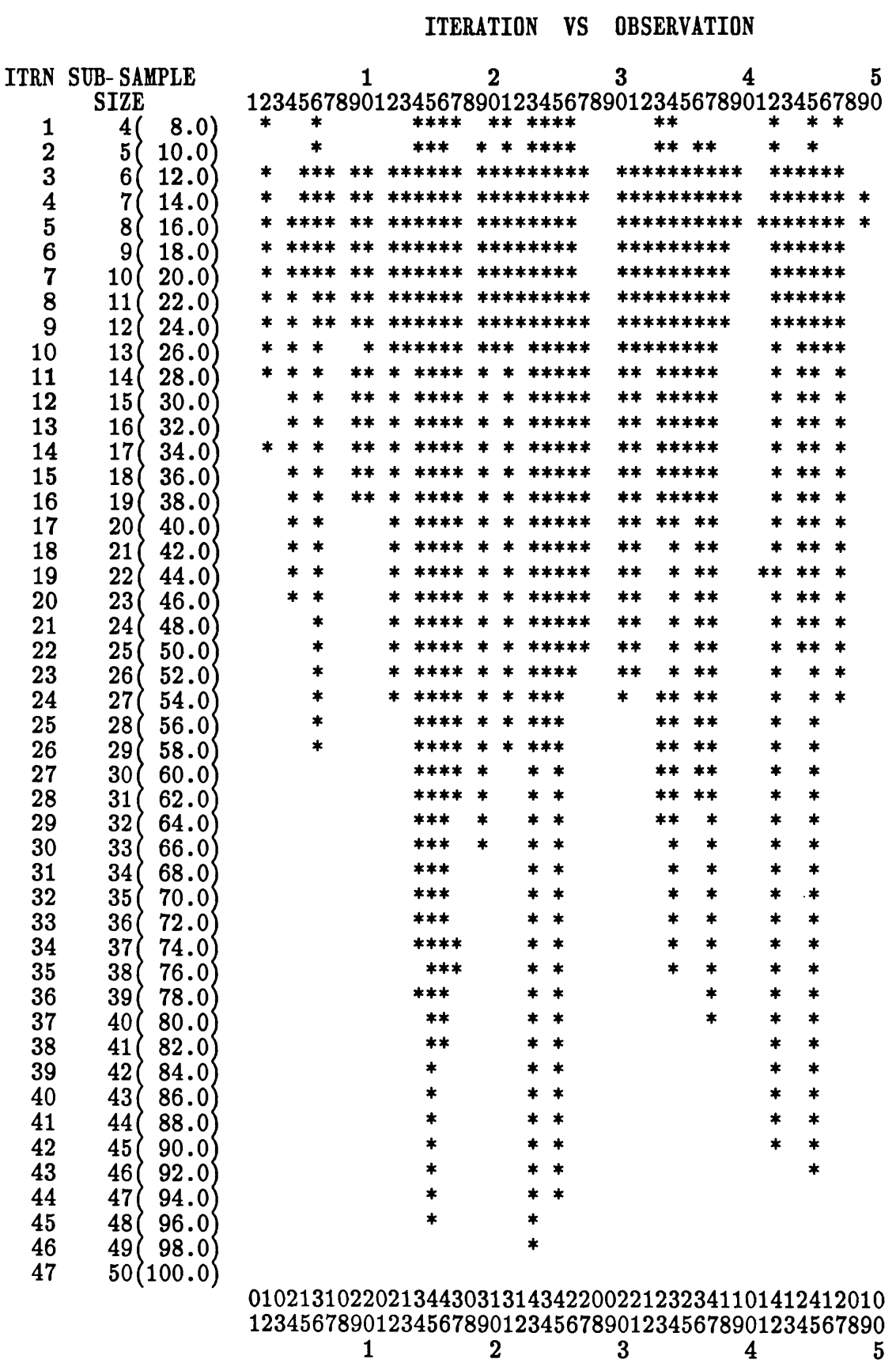
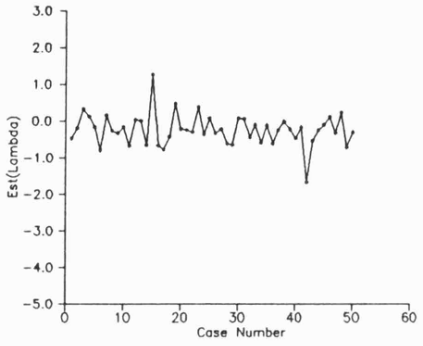
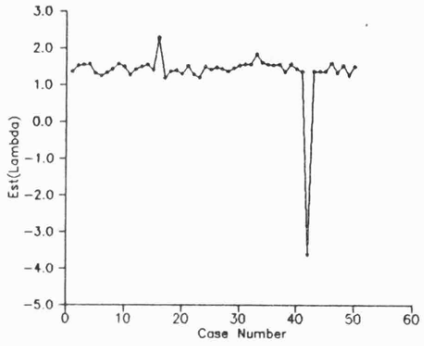


Figure E.10(c) Index Plot of Estimated Lambda with Case Deletion

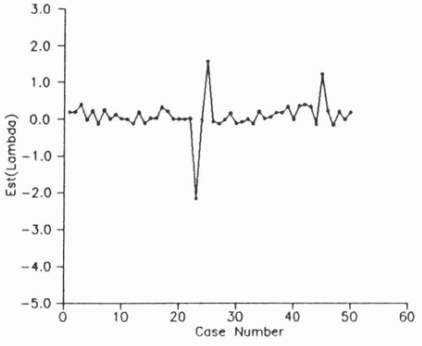
i. X_1 — Sepal Length



ii. X_2 — Sepal Width



iii. X_3 — Petal Length



iv. X_4 — Petal Width

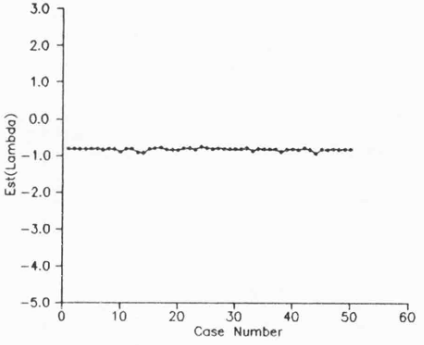


FIGURE E.10(d)

SURCON ANALYSIS

ESTIMATES OF TRANSFORMATION PARAMETERS FOR DATA: FISHER'S IRIS DATA (VERSICOLOUR)

LAMBDA = -0.65 2.53 2.36 0.81

	SINGLE EQUATION ESTIMATES				SUR ESTIMATES			
VAR.	GAMMA	S.E.	T-VALUE	EST. LAMBDA	GAMMA	S.E.	T-VALUE	EST. LAMBDA
1	-3.2785	2.6597	-1.2327 (0.1118)	2.6269	0.0004	1.9377	0.0002 (0.4999)	-0.6520
2	1.0092	2.1555	0.4682 (0.3209)	1.5173	0.0005	1.6295	0.0003 (0.4999)	2.5260
3	-2.1379	2.0774	-1.0291 (0.1542)	4.4959	0.0002	1.1968	0.0002 (0.4999)	2.3577
4	-0.6545	1.6104	-0.4064 (0.3431)	1.4606	0.0000	0.9529	0.0000 (0.5000)	0.8061

*** TERMS IN BRACKETS ARE P-VALUES FOR T(N-1) ***

*** JOINT TEST STATISTIC FOR HO: G1 = G2 = ...= GP = 0

F-STATISTIC = 0.0001 D.O.F. = 4, 196

p-VALUE: F-DISTRIBUTION = 1.0000 D.O.F. = 4, 196

CHI-DISTRIBUTION/P = 0.2500 D.O.F. = 4

*** CONFIDENCE INTERVALS

VAR	95.0% C.I. LS	95.0% C.I. 2SLS
1	[-7.7376, 1.1806]	[-3.2482, 3.2489]
	(-1.8322, 7.0860)	(-3.9006, 2.5966)
2	[-2.6047, 4.6230]	[-2.7314, 2.7324]
	(-2.0965, 5.1312)	(-0.2059, 5.2579)
3	[-5.6208, 1.3449]	[-2.0062, 2.0066]
	(1.0130, 7.9787)	(0.3513, 4.3641)
4	[-3.3545, 2.0455]	[-1.5976, 1.5976]
	(-1.2394, 4.1606)	(-0.7915, 2.4036)

*** TERMS IN SQUARE BRACKETS ARE GAMMAS AND ROUND BRACKETS ARE LAMBDAS ***

*** TEST STATISTIC FOR HO: DIAGONAL COVARIANCE MATRIX
USING THE LAGRANGE MULTIPLIER TEST (LM)

LM-STATISTIC = 119.3432 D.O.F. = 6

p-VALUE FOR CHI-SQUARE WITH 6 D.O.F. = 0.0000

TOTAL NO OF ITERATIONS TO CONVERGE = 33 WITH TOLERANCE FACTOR 10.0**(-4)
****

FIGURE E.10(e)

SURCON ANALYSIS

ESTIMATES OF TRANSFORMATION PARAMETERS FOR DATA: FISHER'S IRIS DATA (VIRGINICA)

LAMBDA = 1.07 0.01 -0.82 1.37

	SINGLE EQUATION ESTIMATES				SUR ESTIMATES			
VAR.	GAMMA	S.E.	T-VALUE	EST. LAMBDA	GAMMA	S.E.	T-VALUE	EST. LAMBDA
1	1.4208	2.1188	0.6706 (0.2528)	-0.3472	-0.0083	1.4827	-0.0056 (0.4978)	1.0819
2	-0.1939	1.7295	-0.1121 (0.4556)	0.2080	-0.0003	1.4016	-0.0002 (0.4999)	0.0143
3	1.8987	2.4542	0.7736 (0.2214)	-2.7234	0.0000	1.7201	0.0000 (0.5000)	-0.8247
4	0.2201	1.8394	0.1197 (0.4526)	1.1527	0.0000	1.5716	0.0000 (0.5000)	1.3728

*** TERMS IN BRACKETS ARE P-VALUES FOR T(N-1) ***

*** JOINT TEST STATISTIC FOR H0: G1 = G2 = ...= GP = 0

F-STATISTIC = 0.0023 D.O.F. = 4, 196

p-VALUE: F-DISTRIBUTION = 1.0000 D.O.F. = 4, 196

CHI-DISTRIBUTION/P = 0.2500 D.O.F. = 4

*** CONFIDENCE INTERVALS

VAR	95.0% C.I. LS	95.0% C.I. 2SLS
1	[-2.1315, 4.9731] (-3.8995, 3.2052)	[-2.4941, 2.4776] (-1.4040, 3.5677)
2	[-3.0936, 2.7057] (-2.6917, 3.1076)	[-2.3502, 2.3496] (-2.3356, 2.3642)
3	[-2.2159, 6.0132] (-6.8380, 1.3912)	[-2.8839, 2.8839] (-3.7086, 2.0592)
4	[-2.8637, 3.3039] (-1.9311, 4.2365)	[-2.6349, 2.6349] (-1.2622, 4.0077)

*** TERMS IN SQUARE BRACKETS ARE GAMMAS AND ROUND BRACKETS ARE LAMBDAS ***

*** TEST STATISTIC FOR H0: DIAGONAL COVARIANCE MATRIX
USING THE LAGRANGE MULTIPLIER TEST (LM)

LM-STATISTIC = 81.9329 D.O.F. = 6

p-VALUE FOR CHI-SQUARE WITH 6 D.O.F. = 0.0000

TOTAL NO OF ITERATIONS TO CONVERGE = 34 WITH TOLERANCE FACTOR 10.0**(-4)
****

FIGURE E.10(f)

SURCON ANALYSIS

ESTIMATES OF TRANSFORMATION PARAMETERS FOR DATA: FISHER'S IRIS DATA (SETOSA+VERSICOLOUR+VIRGINICA)

LAMBDA = -0.25 0.87 0.79 0.58

	SINGLE EQUATION ESTIMATES				SUR ESTIMATES			
VAR.	GAMMA	S.E.	T-VALUE	EST. LAMBDA	GAMMA	S.E.	T-VALUE	EST. LAMBDA
1	-0.3711	1.0495	-0.3536 (0.3621)	0.1176	-0.0271	0.4811	-0.0564 (0.4776)	-0.2264
2	1.7070	0.7859	2.1722 (0.0157)	-0.8412	0.0409	0.6652	0.0614 (0.4756)	0.8249
3	-0.8254	0.4640	-1.7788 (0.0387)	1.6179	-0.0024	0.1393	-0.0172 (0.4932)	0.7950
4	-0.3514	0.2684	-1.3092 (0.0962)	0.9318	0.0000	0.0952	0.0000 (0.5000)	0.5804

*** TERMS IN BRACKETS ARE P-VALUES FOR T(N-1) ***

*** JOINT TEST STATISTIC FOR HO: G1 = G2 = ...= GP = 0

F-STATISTIC = 0.0054 D.O.F. = 4, 596

p-VALUE: F-DISTRIBUTION = 0.9999 D.O.F. = 4, 596

CHI-DISTRIBUTION/P = 0.2500 D.O.F. = 4

*** CONFIDENCE INTERVALS

VAR	95.0% C.I. LS	95.0% C.I. 2SLS
1	[-2.1081, 1.3659]	[-0.8234, 0.7692]
	(-1.6194, 1.8546)	(-1.0227, 0.5699)
2	[0.4063, 3.0077]	[-1.0602, 1.1419]
	(-2.1419, 0.4595)	(-0.2761, 1.9260)
3	[-1.5934, -0.0574]	[-0.2329, 0.2282]
	(0.8499, 2.3859)	(0.5644, 1.0255)
4	[-0.7957, 0.0929]	[-0.1575, 0.1575]
	(0.4875, 1.3761)	(0.4229, 0.7379)

*** TERMS IN SQUARE BRACKETS ARE GAMMAS AND ROUND BRACKETS ARE LAMBDAS ***

*** TEST STATISTIC FOR HO: DIAGONAL COVARIANCE MATRIX
USING THE LAGRANGE MULTIPLIER TEST (LM)

LM-STATISTIC = 394.0351 D.O.F. = 6

p-VALUE FOR CHI-SQUARE WITH 6 D.O.F. = 0.0000

TOTAL NO OF ITERATIONS TO CONVERGE = 28 WITH TOLERANCE FACTOR 10.0**(-4)
****

Stalactite analysis is carried out and the corresponding Stalactite chart is displayed in Figure E.10(b). There appears to be five observations which tend to stand out although they do not appear at the full sample level. These are observations 15, 23, 25, 42 and 45. Figure E.10(c) contains the Index plots of the estimated λ's with case deletions. From Figures E.10(c)i. and E.10(c)ii. the estimated λ's have even scatter about the respective maximum likelihood estimates apart from cases 15 and 42 which seem to cause significant fluctuations especially in X_2. These observations are part of the outlier set and from these plots it can be concluded that their outlyingness is in these two variables. In X_3 (Figure E.10(c)iii.), 23 and 25 have the greatest influence although 45 also appears so these three observations are outlying in only this variable. There is almost non–existant changes in the parameter estimates for λ_4 with case deletion a fact which was already ascertained from the small variability in it so no individual observation has any influence on the parameter in this variable. Although the individual influence of the outliers has been highlighted the joint influence of two or more observations can not be deduced.

This effect of outliers shows how influential an observation can be on the parameter estimates especially if the sample size is not large. It is necessary, therefore, to ascertain the presence/absence of outliers and their identities using an outlier detection technique, like the Stalactite analyis, and then decide on what to do with these (either delete them or minimise their influence) before carrying out any transformations.

EXAMPLE E.11 Repeat Soil Sample Survey Data.

The data are a sample of 57 observations (one region) from the representative soil sampling survey of arable and grassland fields to study the pH nutrient status of the soils in England and Wales between 1969 and 1973 carried out by the Rothamsted Experimental Station. This example consists of five varibles X_1 and X_2 the pH values of water (H_2O) and calcium chloride ($CaCl_2$) respectively; X_3, X_4 and X_5 the available Phosphorus (P), Pottasium (k) and Magnesium (Mg) [Church & Skinner, 1986].

FIGURE E.11(a) SURCON ANALYSIS

ESTIMATES OF TRANSFORMATION PARAMETERS FOR DATA: REPEAT SOIL SAMPLING SURVEY DATA (RSSS) ALL VARS

LAMBDA = -0.55 -0.30 0.02 -0.89 -0.23

	SINGLE EQUATION ESTIMATES				SUR ESTIMATES			
VAR.	GAMMA	S.E.	T-VALUE	EST. LAMBDA	GAMMA	S.E.	T-VALUE	EST. LAMBDA
1	3.3025	2.0623	1.6013 (0.0575)	-3.8553	0.1507	1.4032	0.1074 (0.4574)	-0.7035
2	2.6929	1.7985	1.4974 (0.0700)	-2.9923	-0.0041	1.2086	-0.0034 (0.4987)	-0.2954
3	0.0390	0.2356	0.1654 (0.4346)	-0.0224	-0.0074	0.2092	-0.0352 (0.4860)	0.0239
4	0.0353	0.5808	0.0608 (0.4759)	-0.9227	0.0000	0.5149	0.0000 (0.5000)	-0.8874
5	0.1530	0.3786	0.4041 (0.3438)	-0.3863	0.0000	0.3534	-0.0001 (0.5000)	-0.2332

*** TERMS IN BRACKETS ARE P-VALUES FOR T(N-1) ***

*** JOINT TEST STATISTIC FOR HO: G1 = G2 = ...= GP = 0

F-STATISTIC = 0.0034 D.O.F. = 5, 280

p-VALUE: F-DISTRIBUTION = 1.0000 D.O.F. = 5, 280
CHI-DISTRIBUTION/P = 0.2000 D.O.F. = 5

*** CONFIDENCE INTERVALS

VAR	95.0% C.I. LS	95.0% C.I. 2SLS
1	[-0.1468, 6.7518]	[-2.1962, 2.4975]
	(-7.3047, -0.4060)	(-3.0504, 1.6434)
2	[-0.3150, 5.7009]	[-2.0255, 2.0174]
	(-6.0003, 0.0156)	(-2.3168, 1.7261)
3	[-0.3550, 0.4330]	[-0.3573, 0.3425]
	(-0.4164, 0.3716)	(-0.3260, 0.3738)
4	[-0.9361, 1.0068]	[-0.8612, 0.8612]
	(-1.8942, 0.0487)	(-1.7486, -0.0262)
5	[-0.4803, 0.7863]	[-0.5910, 0.5910]
	(-1.0196, 0.2470)	(-0.8243, 0.3578)

*** TERMS IN SQUARE BRACKETS ARE GAMMAS AND ROUND BRACKETS ARE LAMBDAS ***

*** TEST STATISTIC FOR HO: DIAGONAL COVARIANCE MATRIX
USING THE LAGRANGE MULTIPLIER TEST (LM)

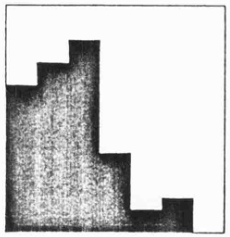
LM-STATISTIC = 80.9892 D.O.F. = 10
p-VALUE FOR CHI-SQUARE WITH 10 D.O.F. = 0.0000

TOTAL NO OF ITERATIONS TO CONVERGE = 38 WITH TOLERANCE FACTOR 10.0**(-4)
****

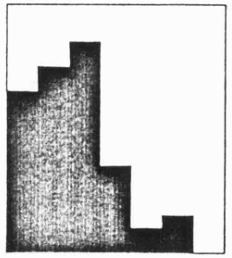
Figure E.11(b) Histograms for the Repeat Soil Sample Survey Data

a) X1 – pH(Water)

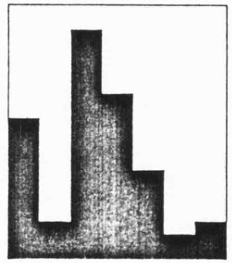
Original scale



Symmetry scale

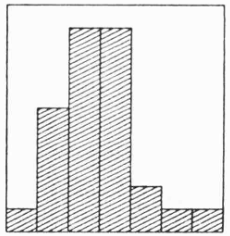


SURCON scale

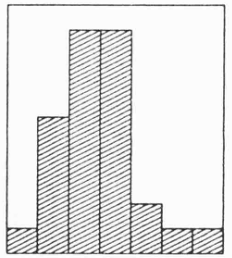


b) X2 – pH(CaCl2)

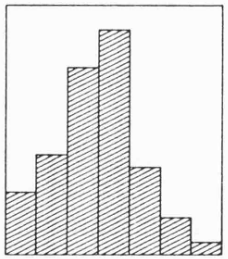
Original scale



Symmetry scale

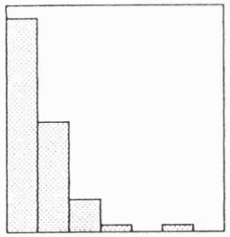


SURCON scale

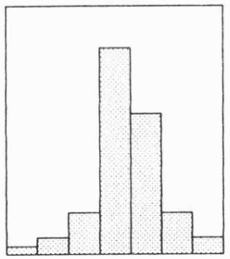


c) X3 – Available P

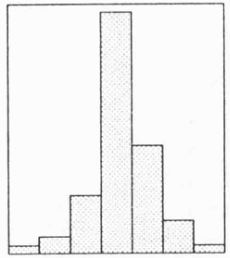
Original scale



Symmetry scale

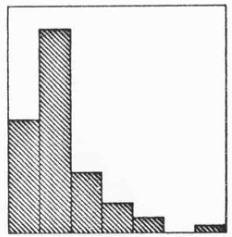


SURCON scale

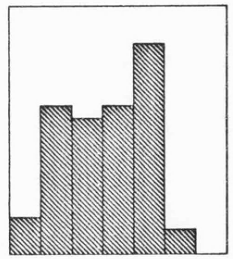


d) X4 – Available K

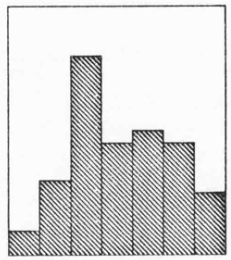
Original scale



Symmetry scale

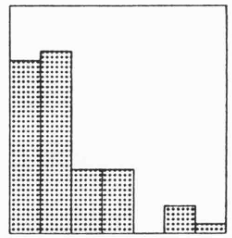


SURCON scale

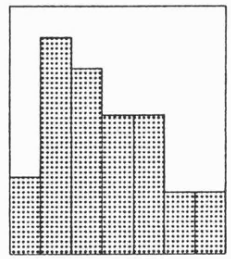


e) X5 – Available Mg

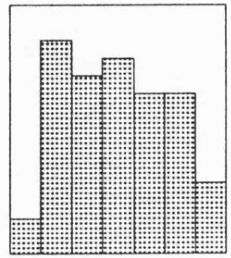
Original scale



Symmetry scale



SURCON scale



The first two variables X_1 and X_2 are very highly correlated and similar in magnitude and the remaining three are also similar in magnitude except that X_5 has small correlation with the others.

The symmetry transformations are no transformation for X_1 and X_2, the log, reciprocal and log transformations for X_3, X_4 and X_5 respectively. Rao's score test is not significant for the first two variables considered jointly although the "quick" parameter estimates are (just) significant. The SURCON estimates (Figure E.11(a)) suggest the reciprocal square root transformations for bith variables. For the next three variables Rao's score test is highly significant together with the "quick" estimates. The SURCON estimates are comparable to the marginal transformations to symmetry. When all the variables are considered jointly there is slight difference in the conclusions drawn. However, there are significant differences between the marginal and joint estimates for the first two variables whereas there is no noticeable difference for the remaining variables.

The 95% confidence intervals for the first two variables include unity and hence even with no transformation for these, the data would still exhibit overall joint normality.

The LM statistic is highly significant and hence joint estimates are worthwhile. The algorithm required 38 iterations to converge to the maximum likelihood estimates.

Figure E.11(b) displays histograms of each variable based on three different scales. The top line of histograms is the data as measured on the original scale (without any transformations). The second and third lines are based on Hinkley's quick transformations to symmetry and the joint transformations derived from the SURCON analysis, respectively.

The effect of transformations is not very significant in the first two variables apart from a curious peak on the left tail of X_1 on the SURCON scale making it bimodal. This could be explained by the fact that small (and large) values of X_1 tend to have large corresponding values of X_4. The correlation between X_1 and X_4 is negligible but on inspecting the scatter plot a non–linear relationship appears to exist. A similar relationship exists between X_1 and X_5. It can, therefore, be concluded that due to the joint influence of

X_4 and X_5 on the SURCON estimate for X_1 it leads to its not having marginal normality. There is significant normalising effect on the histograms of variables, X_3, X_4 and X_5. In all the three cases the strong positive skewness in the original scale is removed after the transformations. Variable X_3 which has the strongest skewness responded best to the transformations.

EXAMPLE E.12 Simulated Bivariate Normal Data (with induced correlation).

These data consist of 50 (=n) sets of bivariate normal deviates generated on a computer. Pairs of these random deviates were transformed to obtain 50 samples (X_1, X_2) with induced correlation as in Example E2.1. A range of values for ρ was used to provide a basis for comparing the different approaches discussed for transforming observations and seeing how correlation affects them.

Table E.6 displays the summary of the results. The two major approaches are the loglikelihood approach and the SURCON approach. In the loglikelihood approach both the marginal and joint estimates for λ are given. For the marginal case the transformation parameter is constant for X_1 (due to the scheme used in generating the data it does not change with ρ) but λ_2 ranges from 0.4 to 1.11 attaining these values at $\rho \approx 0.8$ and $\rho=0$ respectively. So as ρ increases λ_2 tends to reduce upto a certain value when it starts rising again. On considering the joint transformation parameter estimates the λ's range between 0.64 and 1.31 where these values are attained at around the mid–correlations with the actual minimum being again at $\rho \approx 0.8$.

The "quick" estimates from the SURCON analysis range between 0.33 and 1.99 and the SURCON maximum likelihood estimates are numerically identical to those obtained from the loglikelihood approach.

The points to note from this example are that if there is no correlation between the variables then the joint estimates will not be different from the marginal ones, hence, joint estimates procedures would not be worthwhile (the LM statistic is used to check for this). Secondly, as the correlation rises the parameter estimates for the "dependent" variable

tends to reduce upto the point of about 0.8 correlation from where it starts rising again.

The number of iterations required for the SURCON algorithm to converge to the maximum likelihood estimates also rises as the correlation increases upto a point from which they drop.

TABLE E.6 Effect of Correlation on Transformations to Joint Normality

	Loglikelihood				SURCON				
	Marginal		Joint		Qnick Estimates		MLE		No.of
ρ	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\tilde{\lambda}_1$	$\tilde{\lambda}_2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	Itrns
0.000	1.06	1.11	1.08	1.13	1.27	1.41	1.09	1.12	6
0.100	1.06	1.02	1.13	1.08	1.39	1.26	1.13	1.08	6
0.200	1.06	0.92	1.17	1.02	1.49	1.06	1.17	1.02	5
0.300	1.06	0.80	1.22	0.96	1.57	0.86	1.20	0.95	7
0.400	1.06	0.67	1.26	0.89	1.61	0.67	1.23	0.87	7
0.500	1.06	0.55	1.28	0.82	1.61	0.51	1.24	0.80	8
0.600	1.06	0.46	1.31	0.75	1.75	0.39	1.29	0.72	13
0.700	1.06	0.40	1.31	0.69	1.43	0.33	1.29	0.67	12
0.800	1.06	0.41	1.25	0.64	1.34	0.35	1.24	0.64	13
0.900	1.06	0.57	1.13	0.65	1.99	0.44	1.11	0.65	12
0.925	1.06	0.64	1.08	0.66	1.92	0.47	1.06	0.67	12
0.950	1.06	0.72	1.01	0.69	1.84	0.52	1.00	0.69	12
0.975	1.06	0.83	0.93	0.72	0.74	0.56	0.92	0.73	13
0.990	1.06	0.91	0.86	0.75	0.65	0.58	0.87	0.75	7

CHAPTER FOUR

4.0 The tSTAT PACKAGE

4.1 Introduction

The aim of this chapter is to discuss and present the computer package called **tSTAT** (short for *Transformation Statistics*) which implements the theory of the previous chapters. The package was developed not to serve as a comprehensive statistical package but as a complementary tool for data analysts in checking the validity and consistency of their data before embarking onto a full analysis using the well–known statistical packages like SPSS, GLIM, SAS, MINITAB etc. The process can be summarised as that of data screening. Apart from the proposed algorithms, the Stalactite Analysis and SURCON analysis, most of the statistical analyses included in the tSTAT package are readily available in many of the above packages; however, the ease of conducting the analyses varies. For example, one would require to write program–like modules (macros) using a special language syntax to obtain certain results in packages like GLIM and SAS.

The semi–programming languages allow for great flexibility in the sort of analysis which can be carried out but on the other hand cause inexperienced users to shy away or limit their analyses to only basic ones. Some packages like SPSS and MINITAB also have a semi–programming language syntax but it is mostly geared towards automating the execution of several commands than actual programming; for example programming constructs like DO–loops are not included. Latest micro–computer implementations of packages like SPSS go further to assist the inexperienced or casual user by providing a user–friendly menu–driven interface which greatly removes the burden of having to learn the language syntax and thus emphasis is on obtaining the results with the minimum of effort. The problem with such an environment is that the user is confined to the straitjacket of using only the options available and thus flexibility in the analysis is removed.

In designing the tSTAT package, amongst other things, consideration was given to the type of user–audience it would be useful for and we concluded that due to the purpose

of the package, data screening, every data analyst whether experienced or not would require the facility. Secondly, since the package is only used to screen the data before a comprehensive analysis is done it would, therefore, be essential that the analyst does not spend too much time and effort in carrying out the task. It is with these points in mind that the package was designed with a user–friendly menu–driven interface to cater for the inexperienced user, by not requiring him/her to learn any language syntax, and also the experienced user by removing the burden of having to develop the programs (macros) to carry out the exercise and thus providing him/her with a quick tool.

The tSTAT package is written in the C programming language. This language was chosen for a number of reasons. The main reason was due to its portability so even though the system is written for the IBM PC (and compatibles) under the PC DOS operating system it can easily be transfered with little or no modification to run under any other environment e.g. UNIX based systems like the SUN Workstations. The second attribute was the speed of execution and its immense capacity to control the hardware eg. changing the screen attributes, buffering, keyboard and mouse control etc. Finally, there is now a good collection of numerical routines available for the language e.g. Numerical Recipes in C, [Press et. al, 1988]. These have been used extensively in the package.

The chapter begins with the system design for the package in Section 4.2. This describes the overall structure of the package by specifying the different modules which form the complete system. There are several algorithms used in the analysis; however, a few of them require special mention and these are discussed in Section 4.3. The algorithms are presented in a variety of ways. Some are discussed by showing the formulae employed, others include flowcharts and source code in the C language. Section 4.4 describes the technical specifications of the package and outlines the general usage of the package in the form of a reference manual. Finally, Section 4.5 presents an example of a full session of the package on a typical data set.

4.2 System Design

This section describes the general structure of the package by providing an overview of all the modules in the package and their inter–dependencies. Figure 4.1 shows the System Flowchart for the package.

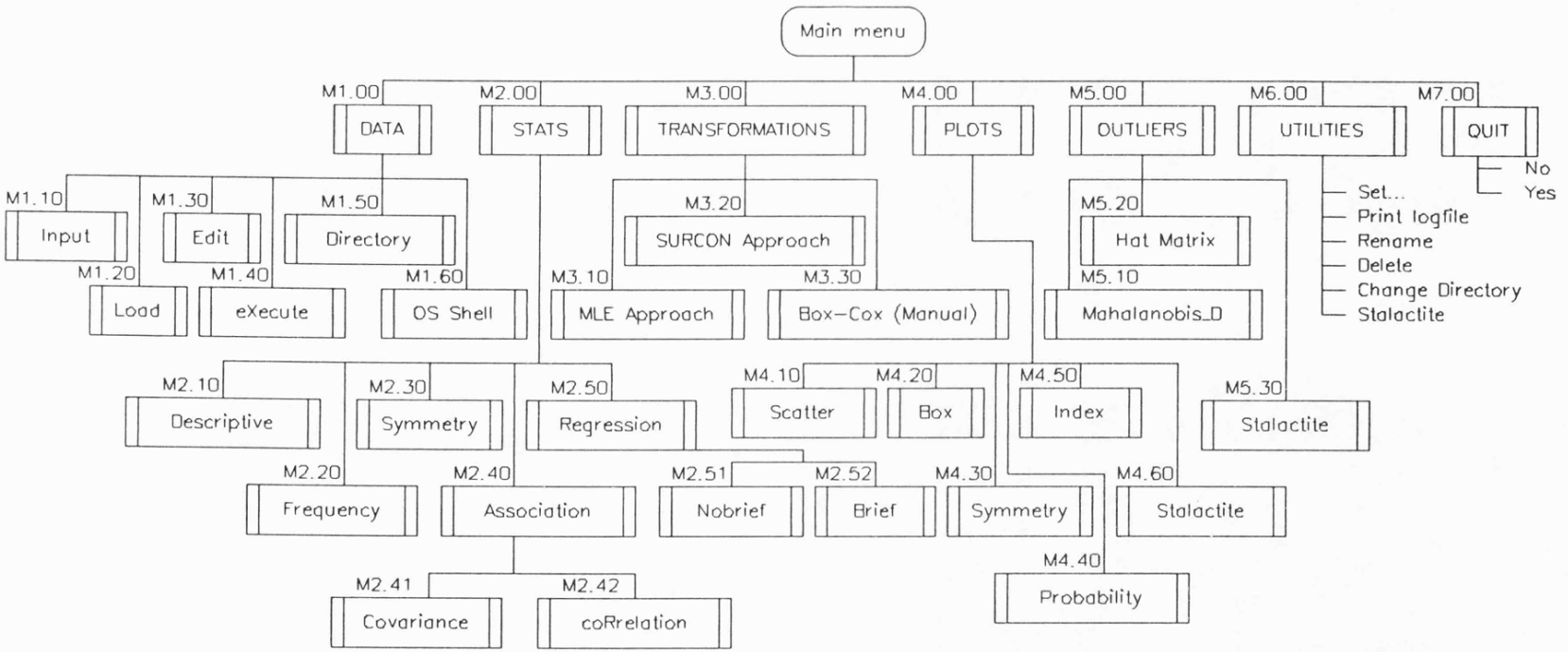
Each module is assigned a generic code of the form Mx.yz where x is the level one option number, y and z the level two and three option numbers e.g. M2.41 refers to option 1 at level three of option 4 at level two of option 2 at level one. The modules are physically located in a set of ten executable program files. There is some implicit grouping of the modules within these files such that each program file performs a specific task and is capable of running independently, however, for a structured run there is a controller file (*TSTAT.EXE*) which should always be the first file to run (See Section 4.4).

Table 4.1 is the summary of the executable file to module structure. An x in a module code indicates all the options within that level.

TABLE 4.1 Executable File to Module Structure

Menu Option	Program	Modules
Data	TSTAT.EXE	M1.20, M1.40, M1.50, M1.60
	TEDIT.EXE	M1.10, M1.30
Stats	TSTAT.EXE	M2.10, M2.40
	TFREQ.EXE	M2.20
	TSYMM.EXE	M2.30
	TREGR.EXE	M2.50
Transformations	TTRANS.EXE	M3.xx
Plots	TPLOTS.EXE	M4.10, M4.20, M4.30, M4.40, M4.50
	TSTALACT.EXE	M4.60
Outliers	TMAHALD.EXE	M5.10, M5.20
	TSTALACT.EXE	M5.30
Utilities	TSTAT.EXE	M6.xx
Quit	TSTAT.EXE	M7.xx

Figure 4.1 System Flowchart for the tSTAT Package



4.3 Main Algorithms

4.3.1 Summary Statistics

In carrying out an analysis of the data it is often revealing to first obtain the characteristics describing each of the variables. These are the descriptive (or summary) statistics and can be broken down into measures of location, spread and relative shape of the histogram. The measures included in the tSTAT package are:

a) Location – arithmetic mean (and its standard error), 10% a–trimmed mean, median

b) Dispersion – Variance, Standard deviation, minimum, maximum and range

c) Shape – skewness, kurtosis (and respective standard errors)

It is also necessary to study the influence of an observation on each of these statistics. A quick method of doing this is to delete an observation from the sample and observe the effect on the statistics. However, in implementing the algorithm of observation deletion the computational requirement can be greatly minimised by considering deletion formulae for the statistics. This means that having computed the full sample statistic, θ say, it is possible to re–compute the deletion statistic $\theta(k)$ when x_k the k–th observation is deleted, by using the relationship between the two. So, we have

$$\theta(k) = f(\theta, x_k) \tag{4.1}$$

Table 4.2 is a summary of some useful deletion statistics formulae. The formulae given in the table are based on a sample x_i, (i=1,2,...,n). Column (c) refers to the formula employed with observation k deleted and x[i] is the i–th ordered observation.

The skewness and kurtosis measures are not included in the deletion formulae. This is due to the fact that they are less robust than the other lower moments and so their use is purely for information and completeness. In fact in some texts their use is not recommended [Press, W.H. et al 1988, p.474]. The following are the formulae:

$$\text{Skew}(x_1,x_2,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i - \bar{x}}{\sigma}\right]^3 \tag{4.2}$$

TABLE 4.2 DELETION SUMMARY STATISTICS

(a) STATISTIC	(b) STANDARD FORMULA	(c) DELETION FORMULA	(d) REMARKS
1/ Mean $\bar{x}$	$\bar{x} = \frac{1}{n}\Sigma x_i$	$\bar{x}(k) = \frac{1}{n-m}\{n\bar{x} - x_k\}$ $= \frac{1}{n-m}\{\Sigma x_i - x_k\}$	$m = \begin{cases}1, & \text{obs. k deleted}\\ 0, & \text{otherwise}\end{cases}$ $x_k = \begin{cases}\text{k-th observation}\\ 0 \quad, \text{otherwise}\end{cases}$
2/ Variance s^2	$s^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$	$s^2(k) = \frac{1}{n-m-1}\left\{(n-1)s^2 - (x_k - \bar{x})^2\right\}$	$m = \begin{cases}1, & \text{obs. k deleted}\\ 0, & \text{otherwise}\end{cases}$ $x_k = \begin{cases}\text{k-th observation}\\ \bar{x} \quad, \text{otherwise}\end{cases}$
3/ Std. Dev. SD	$SD = \sqrt{(s^2)}$	$SD(k) = \sqrt{(s^2(k))}$	—
4/ Std. Err. SE	$SE = SD/\sqrt{n}$	$SE(k) = SD(k)/\sqrt{(n-m)}$	$m = \begin{cases}1, & \text{obs. k deleted}\\ 0, & \text{otherwise}\end{cases}$
5/ Min x	$\text{Min } x = \underset{i}{\text{Min}}\, x_i = x[1]$	$\text{Min } x(k) = \begin{cases}x[m+1], & x_k = \text{Min } x\\ x[1], & \text{otherwise}\end{cases}$	Requires data to be sorted. m as above.
6/ Max x	$\text{Max } x = \underset{i}{\text{Max}}\, x_i = x[n]$	$\text{Max } x(k) = \begin{cases}x[n-m], & x_k = \text{Max } x\\ x[n], & \text{otherwise}\end{cases}$	Requires data to be sorted. m as above.
7/ Range R	$R = \text{Max } x - \text{Min } x$	$R(k) = \text{Max } x(k) - \text{Min } x(k)$	—
8/ Median x	$\text{Med } x = \begin{cases}\frac{x[n/2] + x[n/2+1]}{2}\,\dagger\\ x[(n+1)/2]\,\ddagger\end{cases}$ † n even ‡‡ n odd	$\text{Med } x(k) = \begin{cases}x[n/2+1], & k = n/2\\ x[n/2], & k = n/2+1\\ (x[n/2] + x[n/2+1])/2, & \ddagger\end{cases}$ n even ‡ otherwise	Requires data to be sorted. $\text{Med } x(k) = \begin{cases}\frac{x[nm/2] + x[nm/2+2]}{2}\,*\\ x[(n+1)/2], \text{otherwise}\end{cases}$ n odd * k=(n+1)/2, nm = n-m, m as above.

$$\text{Kurt}(x_1,x_2,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i - \bar{x}}{\sigma}\right]^4 - 3 \qquad (4.3)$$

The –3 term in (4.3) is included to make the value of the kurtosis equal to zero for a normal distribution. The standard errors for the skewness and kurtosis for the idealised case of a normal distribution are approximately $\sqrt{(6/n)}$ and $\sqrt{(24/n)}$, respectively.

4.3.2 Stalactite Analysis

4.3.2.1 Initial Sub–sample Selection

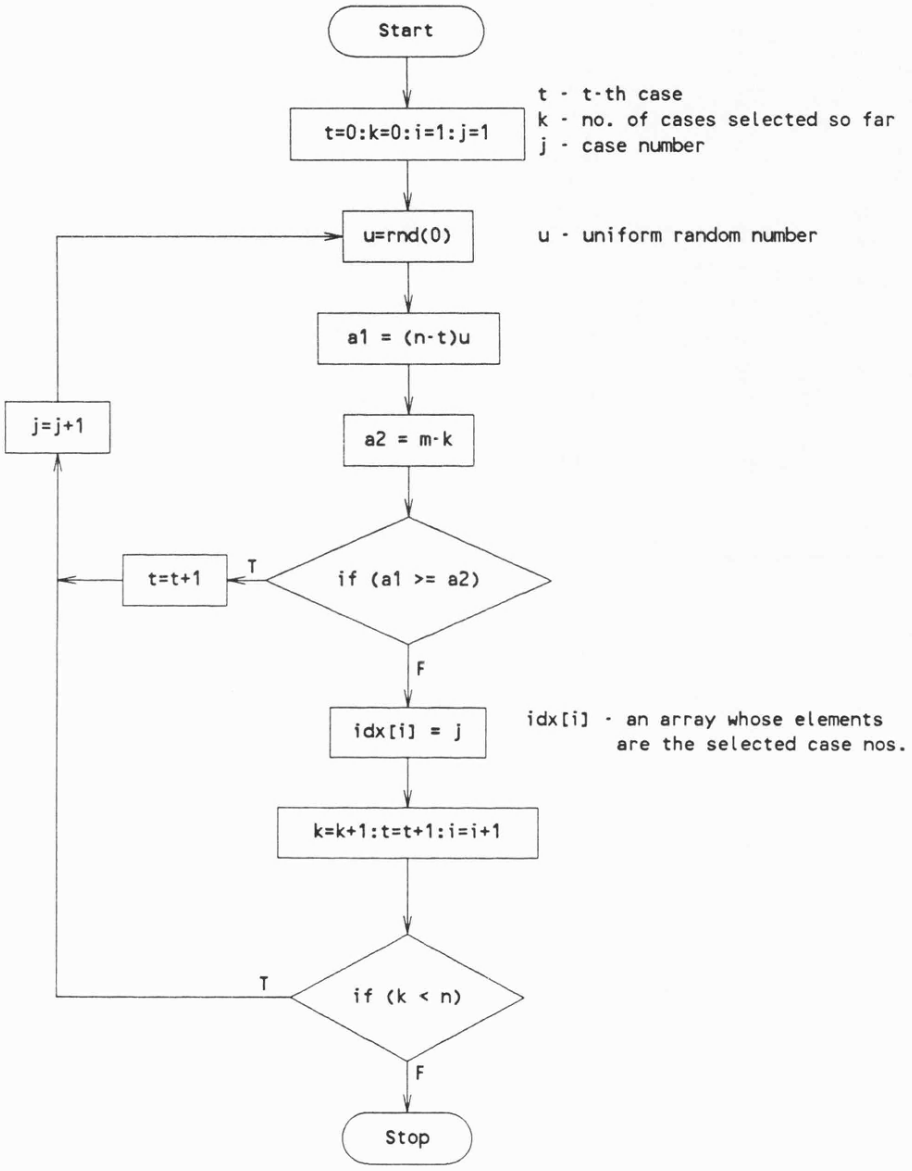
In the Stalactite Analysis algorithm the first sub–sample of p+1 observations is selected randomly from the data. The selection method has to be unbiased but at the same time ensuring that no observation is repeated within the chosen sub–sample. To achieve this the sampling of the data should be done without replacement. It is, therefore, desirable to use an efficient procedure for selecting the m=p+1 observations from the sample such that each observation has an equal probability of being chosen albeit once. The procedure should also ensure the efficient use of the computing facility by minimising on the storage required together with the speed of selection. The latter can be achieved by minimising the number of passes through the sample.

Several methods have been devised for this problem. The procedure adopted in the thesis is the *Selection Sampling Technique* (Algorithm S) [Knuth, D.E. 1981, p121.]. If we let the sample size be n and the sub–sample size be m=p+1 then the algorithm selects the (t+1)st observation with probability (m – k)/(n – t) if k observations have already been selected. This is the appropriate probability, since of all the possible ways to choose m items from n such that k values occur in the first t, exactly

$$\begin{bmatrix} n-t-1 \\ m-k-1 \end{bmatrix} \Big/ \begin{bmatrix} n-t \\ m-k \end{bmatrix} = \frac{m-k}{n-t} \qquad (4.4)$$

of these select the (t+1)st element. The technique does one pass through the data and uses a test based on the probability of selecting the (t+1)st observation in deciding whether or not to include the observation in the sub–sample. In particular, if this probability is

Figure 4.2 Flowchart for Initial Sub-sample Selection
for the Stalactite Analysis Algorithm (Selection Sampling)



greater than a generated uniform random number the observation is selected. Figure 4.2 displays a flowchart of the algorithm formulated from the procedure and Function 4.1 is the corresponding source code.

Function 4.1 Initial Sub–Sample Selection

```
/*  ** initial sub–sample selection **  */
/*     input: sample size n1, no.of vars. ip1 */
/*    output: idx[], vector of selected obs.nos */
/* variables: */
/*    double: a1, a2, u */
/*       int: i, j, k, t */

void t_pkobsk ( int n1, int ip1, int *idx)
{
  double a1, a2, u;
  int i=1, j=1, k=0, m1, t=0;

  m1 = ip1+1;
  for (;;)
   {
   u = random(0);
   a1 = (n1–t)*u;
   a2 = m–k;
   if ( a1 < a2)
     idx[i] = j;
   k += 1;
   t += 1;
   i += 1;
   if ( k >= n1) break;
   j += 1;
   }
}
```

The vector of selected observations idx[] can then be used to pick the values for these observations from which the sub–sample mean vector and covariance matrix can be computed.

4.3.2.2 Matrix Inversion

The matrix inversion algorithm adopted uses the Lower/Upper decomposition algorithm [Press, W.H., et al pp37–46]. The matrix is initially decomposed into a lower triangular (has elements only on the diagonal and below) and upper triangular (has elements on the diagonal and above). The inverse is then obtained by inverting the

decomposed matrix column by column.

4.3.2.3 Mahalanobis Distance

In computing the Mahalanobis distances, d_i (i=1,...,n) there are three main stages. Initially we need to compute the mean vector $\bar{x}$ and the covriance matrix S. The computation of these is carried out by a single function segment. The next stage is to compute the inverse of S. In the tSTAT package this is performed by a function using the algorithm described in Section 4.3.2.2. The final stage is to form the quadratic in equation (2.20) of Section 2.

Function 4.2 is the function segment which does the computations.

Function 4.2 Mahalanobis Distance

```
/*  ** mahalanobis distance **  */
/*     input: sample size n1, no.of vars. ip1 */
/*           data z[][], mean vector m1[], */
/*           inverse cov.matrix qi[][] */
/*    output: md[], mahalanobis distances vector */
/* variables: */
/*    double: t1, t2 */
/*       int: i, j, k */

void t_mahaldist( int n1, int ip1, double **z, double *m1,
           double **qi)
{
double t1, t2;
int i, j, k;

for (i=0; i<n1; i++)
  {
  t1 = 0.0;
  for (j=0; j<ip1; j++)
    {
    t2 = 0;
    for (k=0; k<ip1; k++)
      t2 = t2 + (z[i][k] - m1[k])*qi[k][j]*(z[i][j] - m1[j]);
    t1 = t1 + t2;
    }
  md[i] = sqrt(t1);
  }

}
```

4.3.2.4 Stalactite Analysis – The Complete Algorithm

For brevity, only flowcharts for both the Stalactite Analysis and SURCON analysis algorithms are displayed and not the full C source code as implemented in the tSTAT package. However, the use of flowcharts makes it possible to implement the algorithms in any desired language as opposed to having to translate them from C.

Figure 4.3 is the flowchart for the Stalactite analysis algorithm.

4.3.3 SURCON Transformations

Figure 4.4 is a flowchart of the SURCON analysis.

The next part of the thesis discusses the probability distributions used throughout the analyses considered and describes the algorithms adopted for them in the tSTAT package.

4.3.4 Probability Functions

In the scope of the thesis four probability distributions are used in many different contexts; these are the normal, Student's t, χ^2 and F distributions. This section contains the algorithms adopted within the tSTAT package for the computations of the cumulative distribution functions and their corresponding inverse functions.

The definition of a cumulative distribution function of a random variable x with density function $f(x)$ i.e.

$$\mathrm{F}(b) = \Pr(\mathrm{X} \leq b) = \int_{-\infty}^{b} f(x)\mathrm{d}x \qquad (4.5)$$

requires the integration of a function. The direct approach is to use numerical integration of the density function. There are several algorithms for carrying out numerical integration e.g.Simpson's Rule, Trapezoidal Rule, Romberg Integration, etc. [Press et al., 1988]. The idea is to obtain the integral as accurately as possible with the smallest number of function evaluations of the integrand. For the purposes of the tSTAT package approximations to the distribution functions are adequate. The following algorithms are, therefore, all based on approximations [See Cooke et al., 1982].

Figure 4.3 Flowchart for the Stalactite Analysis Algorithm

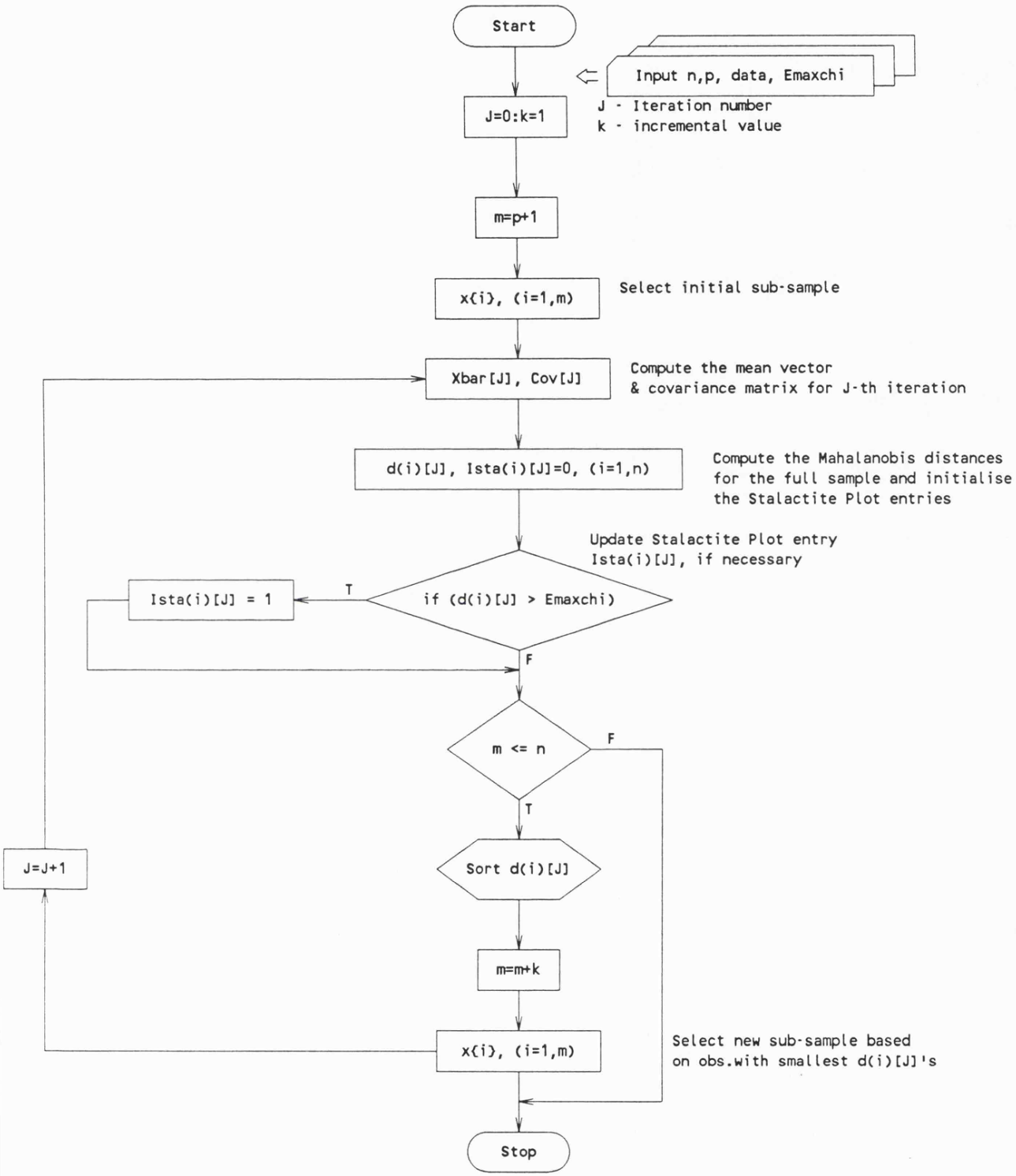
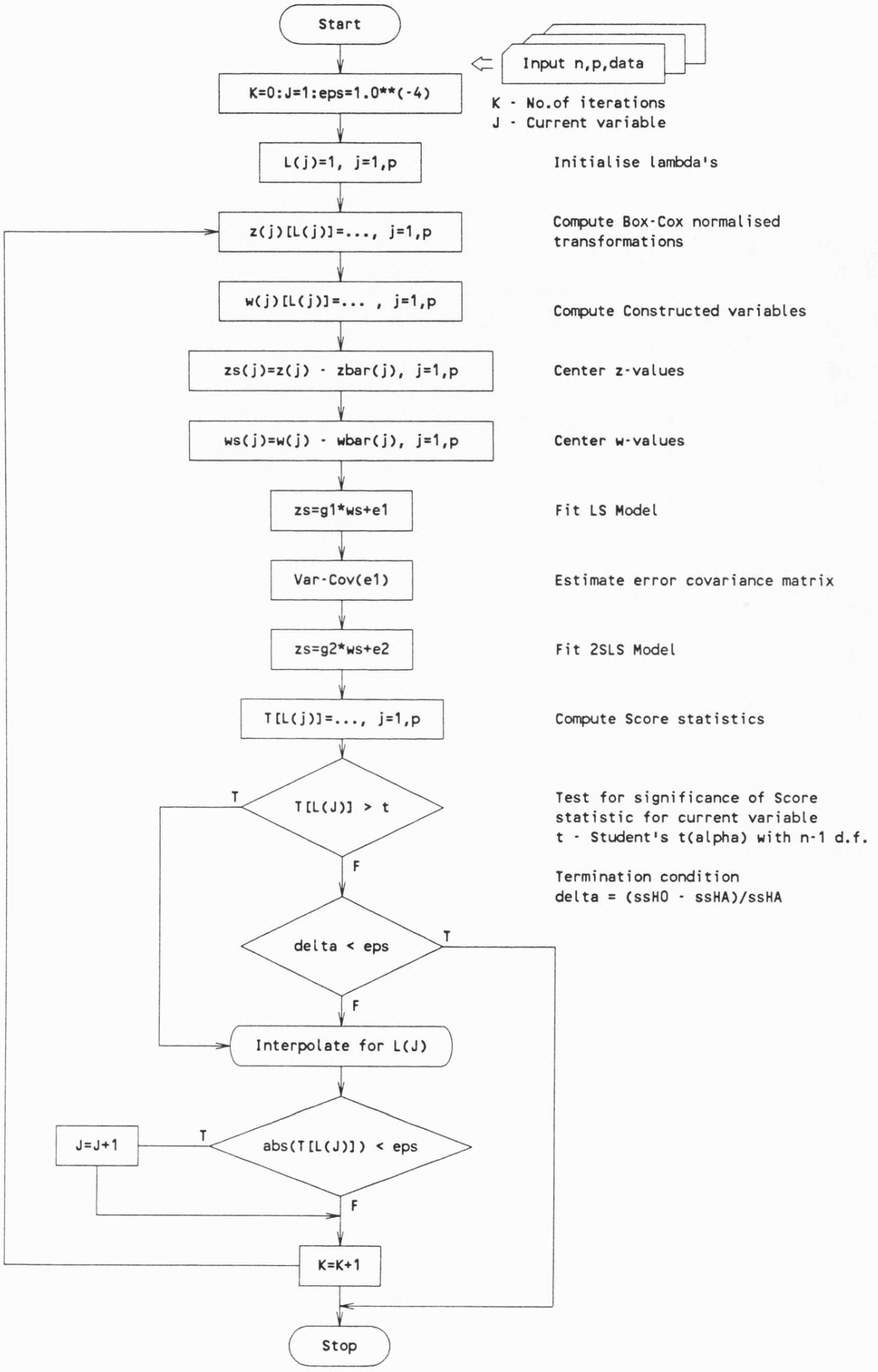


Figure 4.4 Flowchart for the SURCON Analysis Algorithm



a) The Standard Normal Distribution N[0,1]

The integral to be evaluated is

$$F(z) = Pr(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2).dt \qquad (4.6)$$

Since the integral cannot be evaluated over an infinite range using the numerical methods use is made of the fact that the integrand is symmetrical about t=0. So

$$F(z) = 0.5 + \int_{0}^{z} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2).dt \qquad (4.7)$$

Cooke et al. [1982] give an approximation of this integral which is an explicit function of z due to Hastings and quoted in Abramowitz and Stegun [1972]. For a non–negative z,

$$F(z) = 1 - 0.5(1 + a_1 z + a_2 z^2 + a_3 z^3 + a_4 z^4)^4 + \epsilon(z) \qquad (4.8)$$

where $a_1 = 0.196854$, $a_2 = 0.115194$, $a_3 = 0.000344$ and $a_4 = 0.019527$. Function 4.3 is the C implementation of the algorithm.

Function 4.3 Normal Distribution Function – Approximation

```
/*  ** normal distribution function – approx **  */
/*    input: argument z */
/*   output: p0, prob. x < z */
/* variables: */
/*   double: a1, a2, a3, a4, p0, w */

double t_normdf( double z)
{
  double a1, a2, a3, a4, p0, w;

  a1 = 0.196854;
  a2 = 0.115194;
  a3 = 0.000344;
  a4 = 0.019527;

  w = fabs( z);
  p0 = 1 + w*(a1 + w*(a2 + w*(a3 + w*a4)));
  p0 = pow( p0, 4);
  p0 = 1 – 0.5/p0;
  p0 = 0.5 + (p0 – 0.5)*sgn( z);

  return p0;

}
```

For negative z the expression is evaluated using $|z|$ to get the value p0 from which the required probability can be obtained as 1 − p0. It is stated that the error of approximation is of the order $|\epsilon(z)| < 2.5 \times 10^{-4}$.

Similarly, to obtain the inverse normal integral we use the following approximation. If we put $1 - F(z) = 1 - p = q$, then provided that $0 < q \leq 0.5$ the z corresponding to a particular value of q is given by

$$z = t - \frac{a_1 + a_2 t}{1 + a_1 t + a_2 t^2} + \epsilon(q) \qquad (4.9)$$

where $t = \sqrt{(-2\log_e q)}$. The absolute error $|\epsilon(q)| < 3.0 \times 10^{-3}$. Function 4.4 displays the algorithm.

Function 4.4 Inverse Normal – Approximation

```
/*  ** inverse normal function – approx **  */
/*    input: probability p0 */
/*   output: standard normal value z  */
/* variables:
/*  – double: a1, a2, a3, a4, q0, w, w1, w2, z */

double t_invnorm( double p0)
{
  double a1, a2, a3, a4, q0, w, w1, w2, z;

  a1 = 2.30753;
  a2 = 0.27061;
  a3 = 0.99229;
  a4 = 0.04481;

  q0 = 0.5 – fabs(p0 – 0.5);
  w  = sqr(– 2*log(q0));
  w1 = a1 + a2*w;
  w2 = 1.0 + w*(a3 + w*a4);
  z  = w – w1/w2;
  z  = z*sgn(p0 – 0.5);

  return z;
}
```

b) The Student's t Distribution t(k)

Abramowitz and Stegun [1972] provide an exact algorithm for the probability p that

corresponds to a given value of t with k degrees of freedom. The algorithm is constructed from a series summation. If we let $\theta = \tan^{-1}(t/\sqrt{k})$, then $p = \frac{1}{2}(1 + A)$, where

$$A = \begin{cases} 2\theta/\pi & \dagger \\ \frac{2}{\pi}\left[\theta + \sin\theta\left[\cos\theta + \frac{2}{3}\cos^3\theta + \ldots + \frac{2.4\ldots(k-3)}{1.3\ldots(k-2)}\cos^{k-2}\theta\right]\right] & \dagger\dagger \\ \sin\theta\left[1 + \frac{1}{2}\cos^2\theta + \frac{1.3}{2.4}\cos^4\theta + \ldots + \frac{1.3\ldots(k-3)}{2.4\ldots(k-2)}\cos^{k-2}\theta\right] & \dagger\dagger\dagger \end{cases} \quad (4.10)$$

† $k = 1$, †† $k > 1$ and odd, ††† k even.

The algorithm is implemented in Function 4.5.

Function 4.5 Student's t Distribution

```
/*  ** Student's t distribution function — approx **  */
/*    input: argument t0, degrees of freedom k1 */
/*   output: p0, probability x < t0 */
/* variables:
/* — double: a1, c0, p0, s0, t, t1, w */
/* — integer: i, j1, j2, k2, w1

double t_studentst( double t0, int k1)
{
  double a1, c0, p0, s0, t, t1, w;
  int i, j1, j2, k2, w1;

  a1 = 0.36338023;
  w  = atn(t0/sqr(k1));
  s0 = sin(w);
  c0 = cos(w);
  w1 = k1 — 2* ( int) (k1/2);

  if ( w1 != 0)
    {                  /* k1 odd */
    t1 = w;
    if ( k1 != 1)        /* k1 not 1 (special case) */
      {
      t = s0*c0;
      t1 += t;
      if ( k1 != 3)      /* k1 not 3 (special case) */
        {
        j1 = 0;
        j2 = 1;
        k2 = (k1 — 3)/2;
        }
      }
    }
```

```
else
  {                    /* k1 even */
  t1 = s0
  if ( k1 != 2)
    {
    t = s0;
    j1 = -1;
    j2 = 0;
    k2 = (k1 - 2)/2;
    }
  }

if ( k1 > 3)
  {
  for (i=1; i<k2; i++)  /* series summation */
     {
     j1 += 2;
     j2 += 2;
     t = t*c0*c0*j1/j2;
     t1 += t;
     }
  }

if ( ( k1 != 1) && ( k1 != 3))
  t1 = t1*(1 - a1*w1);

p0 = 0.5*(1 + t1);

return p0;
}
```

The inverse distribution of t is computed by making use of a transformation proposed by Wallace [1959]. For a given value of t with k degrees of freedom an approximation to the corresponding value of z (from the standard normal distribution) is given by

$$z = \frac{8k + 1}{8k + 3}\left[k \log_e\left[1 + \frac{t^2}{k} \right] \right]^{1/2} \qquad (4.11)$$

It, therefore, follows that from a given z value the corresponding t can be computed. This is achieved by rewritting (4.11) as follows:

$$t(k) = \sqrt{[k\,(\exp(w^2/k) - 1]} \qquad (4.12)$$

where k is the degrees of freedom and $w = z(8k + 3)/(8k + 1)$. Function 4.4 uses this approximation and it calls Function 4.3 (Inverse normal – approximation). It requires as

input the probability p0 and the degrees of freedom k1. It first calculates the corresponding standard normal variable z and then goes on to calculate the required t value from (4.12).

Function 4.6 Inverse Student's t Function

```
/*  ** Inverse Student's t function — approx **  */
/*    input: probability p0, degrees of freedom k1 */
/*    calls: t_invnorm (Function 4.2)
/*   output: t0, t—value */
/* variables:
/* — double: t0, w, z */

double t_invstudt( double p0, int k1)
{
  double t0, w, z;

  z = t_invnorm( p0);   /* find normal z corresponding to p0 */

  w  = z*(1 + 2/(1.0 + 8*k1));
  t0 = k1*(exp(w*w/k1) — 1.0);
  t0 = sqr( t0);

  return t0;
}
```

The algorithm is not reliable for fewer than four degrees of freedom. [See Cooke et al., 1982].

c) The χ^2 Distribution, $\chi^2(k)$

The algorithm adopted for the χ^2 distribution is based on an algorithm by Lau [1980] for the cumulative distribution of a gamma function. The gamma distribution with shape parameter a and scale parameter β , $G(a, \beta)$ has a probability density function

$$G(a, \beta) = \frac{1}{\Gamma(a)\beta^a} x^{a-1} e^{-x/\beta} \qquad (4.13)$$

$0 < x < \infty$. The χ^2 distribution with k degrees of freedom is the gamma distribution with shape parameter $k/2$ and scale parameter 2, i.e. $G(k/2, 2)$. The density function involves the evaluation of the gamma function $\Gamma(x)$. Using the relation between $\Gamma(x)$ and $\Gamma(x-1)$ then for a positive integer n we can write

$$\Gamma(n) = (n-\frac{1}{2}).(n-\frac{3}{2}).(n-\frac{5}{2})...\frac{3}{2}.\frac{1}{2}.\Gamma(\frac{1}{2}) \qquad (4.14)$$

where $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. This function can then be evaluated as a series of products. However, it can quickly cause numeric overflows, especially on a small computer, and so in its evaluation we consider its logarithm, $\log_e \Gamma(n)$, instead after which it can easily be reconverted. The algorithm is evaluated by Function 4.7.

Function 4.7 χ^2 Distribution

```
/*  ** Chi—squared distribution function — approx **  */
/*    input: argument x2, degrees of freedom k1 */
/*   output: p0, probability x < x2 */
/* variables:
/*    double: a2, g1, g2, p0, t, w1
/*   integer: k, w;

double t_chisq( double x2, int k1)
{
  double a2, g1, g2, p0, t, w1;
  int k, w;

  w1 = 0.5*x2;
  k1 = 0.5*k1;

  g1 = 0;        /* calculate log—gamma (k1 + 1) */
  for (;;)
   {
   w —= 1;        /* nb: (2*w) is an integer > 0  */
   if ( w > 0)
     g1 += log( w);   /* sum logs of factors */
   else
     break;
   }

  if ( w != 0)
    g1 += 0.57236494;      /* add ln gamma(0.5) */

  g2 = 0;
  a2 = exp(k1*log(w1) — g1 — w1);
  if ( a2 != 0)
    {
    t = 1;
    g2 = 1;
    k = k1;
    do
      {
      t *= w/k;
      g2 += t;
```

```
    }
    while ( t/g2 > 1.0e—6)
   p0 = g2*a2;
  }

 return p0;
```

To obtain the inverse χ^2 distribution we use the result by Wilson and Hilferty [1936] which states that for a χ^2 with k degrees of freedom the quantity $(\chi^2/k)^{1/3}$ is approximately normal with mean $(1 - 2/9k)$ and variance $(2/9k)$. Hence,

$$z = \frac{(\chi^2/k)^{1/3} - (1 - 2/9k)}{(2/9k)} \tag{4.15}$$

from which we may obtain $\chi^2(k)$ in terms of z:

$$\chi^2(k) = k\left[1 - \frac{2}{9k} + z\sqrt{\frac{2}{9k}}\right]^3 \tag{4.16}$$

Function 4.8 uses this equation.

Function 4.8 Inverse χ^2 Function

```
/*  ** Inverse Chi—squared function — approx **  */
/*    input: probability p0, degrees of freedom k1 */
/*    calls: t_invnorm (Function 4.2) */
/*   output: x2 Chi—squared value */
/* variables:
/*    double: a1, w, x2, z

double t_invchisq( double p0, int k1)
{
  double a1, w, x2, z;

  z = t_invnorm( p0);     /* find normal z corresponding to p0 */

  a1 = 2.0/(9*k1);
  w  = 1.0 — a1 + z*sqr(a1);
  x2 = k1*pow(w,3);

  return x2;
}
```

d) The F distribution, F(k_1, k_2)

The approximation (4.16) may be applied to derive one for F. The F distribution with k_1 and k_2 degrees of freedom is defined as the ratio of two χ^2 variables divided by their

degrees of freedom. Thus we may write

$$F = (\chi_1^2/k_1)/(\chi_2^2/k_2) \tag{4.17}$$

the ch–squared distributions χ_1^2 and χ_2^2 having k_1 and k_2 degrees of freedom. Hence,

$$F^{1/3} = (\chi_1^2/k_1)^{1/3}/(\chi_2^2/k_2)^{1/3} \tag{4.18}$$

is the ratio of two variables which are approximately normally distributed. Geary [1930] showed that if $v = z_1/z_2$ where z_1 and z_2 are normal variables with means μ_1, μ_2 and variances σ_1^2, σ_2^2 respectively then

$$z = \frac{\mu_1 - \mu_2 v}{\sqrt{(\sigma_1^2 + \sigma_2^2 v^2)}} \tag{4.19}$$

is approximately a standard normal variable. We can substitute $F^{1/3}$ for v and the appropriate values for μ_1, μ_2, σ_1^2, σ_2^2 i.e. $\mu_i = (1 - 2/9k_i)$ and $\sigma_i^2 = (2/9k_i)$, i=1,2 to obtain the corresponding z value [Paulson, 1942]. Function 4.7, makes use of this result. It calls the normal distribution function (Function 4.3). The algorithm is valid only for values of F $\geq$ 1. If F < 1, then F is replaced by 1/F, $k_1$and k_2 are interchanged and the resulting probability is subtracted from 1. A correction to the z value is required in the algorithm when $k_2 \leq 3$ in order to improve accuracy for these small values of k_2. If this correction is made the algorithm gives reasonably satisfactory results.

Function 4.9 F Distribution

```
/*  ** F distribution function — approx **  */
/*    input: argument f ( >= 1 )2, degrees of freedom k1, k2 */
/*    calls: t_normdf (Function 4.1) */
/*   output: p0, probability x < f */
/* variables:
/*   double: a1, a2, w, w1, w2, z, p0 */

double t_fdistn( double f, int k1, int k2)
{
  double a1, a2, w, w1, w2, z, p0;

  a1 = 2.0/(9.0*k1);
  a2 = 2.0/(9.0*k2);
  w  = pow(f, 1.0/3.0);
  w1 = w + a1 — w*a2 — 1.0
  w2 = a2*w*w + a1;
  z  = w1/sqr(w2);
```

```
  if (k2 < 3)          /* correction factor for small k2 */
    z = z*(1.0 + 0.08*(pow(z, 4.0)/(pow(k2, 3.0));
  p0 = t_normdf( z);     /* find p0 corresponding to z */

  return p0;
}
```

The relation between z and F in (4.19) ,substituted accordingly, yields a quadratic in $\sqrt[3]{F}$ when a particular value is substituted for z. One root is positive and the other negative, the latter being discarded. We can, therefore, obtain the inverse F distribution function from the result. Function 4.10 is the implementation of this algorithm.

Function 4.10 Inverse F Distribution – Approximation

```
/*  ** Inverse F distribution — approx **  */
/*    input: probability p0 ( >= 0.5), degrees of freedom k1, k2 */
/*    calls: t_invnorm (Function 4.2) */
/*   output: f, f—value */
/* variables:
/*   double: a1, a2, f, w, w1, w2, w3, w4, z */

double t_fdistn( double p0, int k1, int k2)
{
  double a1, a2, f, w, w1, w2, w3, w4, z;

  z = t_invnorm( p0);  /* find z corresponding to p0 */

  a1 = 2.0/(9.0*k1);
  a2 = 2.0/(9.0*k2);
  w  = z*z;

  w1 = 1.0 + a2*(a2 — w — 2.0);
  w2 = a1 + a2 — a1*a2 — 1.0;
  w3 = 1.0 + a1*(a1 — w — 2.0);
  w4 = sqr(w2*w2 — w1*w3);
  f  = (w4 — w2)/w1;
  f  = pow( f, 3.0);

  return f;
}
```

A substantial improvement in the algorithm, for $k_2 \leq 3$, can be made by replacing the z value obtained from the normal function with

$$z' = k_2^{3/4}\, u\,(1.1581 - 0.2296u - 0.0042u^2 - 0.0027u^3) \qquad (4.20)$$

where $u = z/k_2^{3/4}$.

Having discussed the main algorithms adopted in the package it is now possible to describe its usage. Section 4.4 is an outline of this usage. It is presented in the form of a user reference manual.

4.4 User Manual

4.4.1 Using the tSTAT package

The tSTAT package consists of several modules which are combined to form the system. Each module is designed to perform a specific part of the analysis. Each of the modules can be used independently but for a systematic and structured approach there is a controller module called *TSTAT.EXE* which contains a user–friendly interface with a menu system. So to start up the package type TSTAT at the DOS prompt ie.

C>TSTAT <CR>

The system will display the opening screen referred to as the "Main Screen" which contains the main menu options and various other items. The options in this menu call up all the other modules. Due to the numerous and involved sequence of parameters passed to each module to ensure the system's correct functioning it is always advisable to run the modules from this integrated environment.

a) Main Screen

The main screen is split up into four distinct areas (See Figure 4.5) as follows:

– Header/Title Area

This area displays the title of the package together with the version number.

– Main menu Area

The main menu area displays the "top–level" menu options. Each option loosely corresponds to a module within the system.

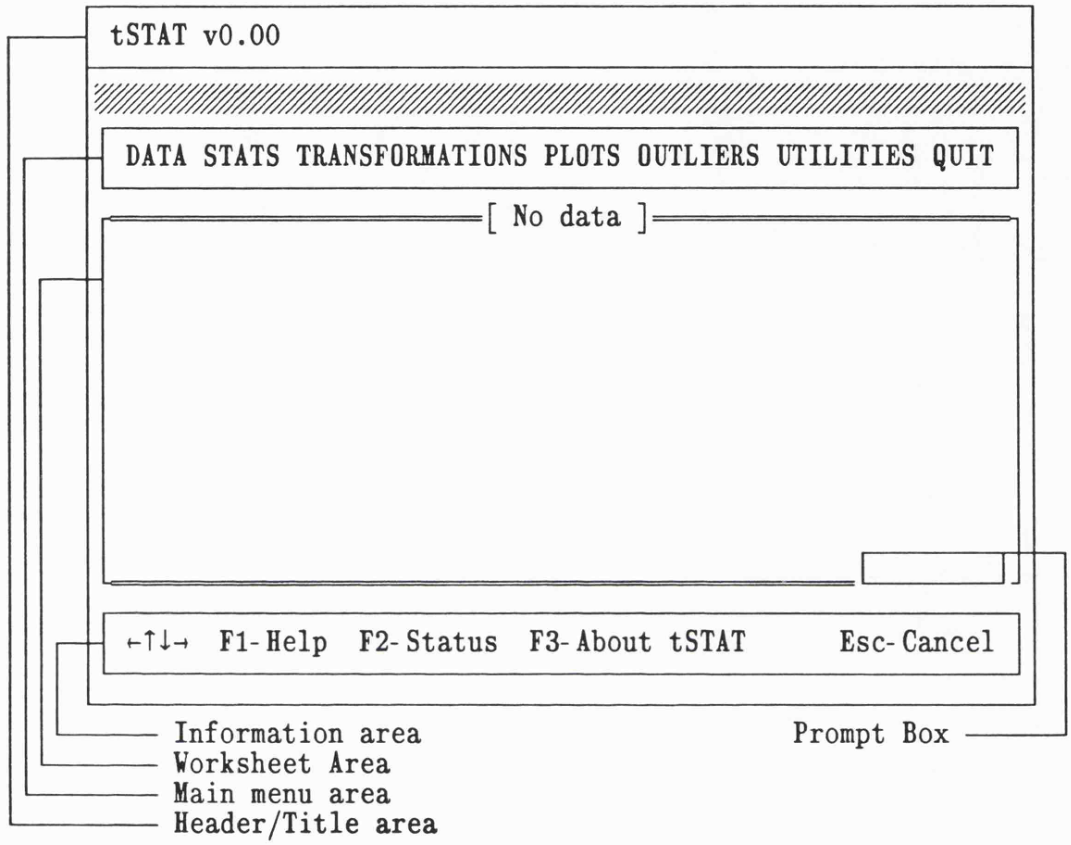
– Worksheet Area

This is where all the output is displayed from the analyses. All the output within this area is also written to disk into a logfile which can then be used to obtain a hard copy of the analyses.

– Footer Area

The Footer area indicates which keys may be pressed to select an option, obtain help or cancel an action.

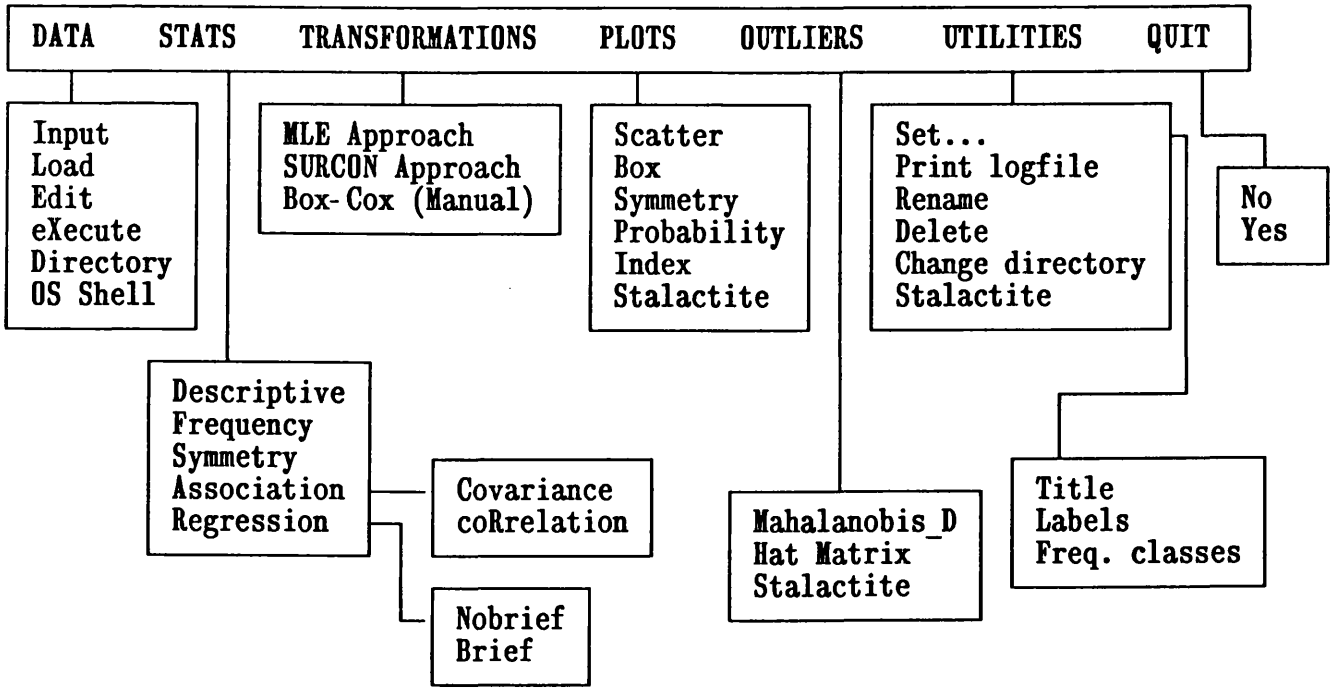
Figure 4.5 Main Screen



b) Menus

The package is operated by a series of menus which are divided into two levels (See Figure 4.6). These levels are the horizontal (options are aligned horizontally)) and dropdown (vertical alignment of the options). Some of the dropdown menu options contain a third level which is refered to as the side–menu. An option within any of the menu types is selected by moving the highlighted bar to it using the arrow keys, 'Home' or 'End' and then pressing 'Enter'. Alternatively, an option can be selected by pressing the highlighted letter within it.

Figure 4.6 Menu Structure



c) Messages

The package displays three main types of messages. These are:

– Error Messages

If an error occurs e.g. a particular computation cannot be performed like inverting a singular matrix, division by zero, an incorrect data file format, etc. the system displays a red box in the center of the screen which informs the user of the nature of the error. The red box will remain on the screen for a short while and is self–clearing.

– Wait Message

This type of message appears in the prompt box at the bottom right–hand corner of the worksheet area. It occurs when the system is busy e.g. when performing a particularly long series of computations. The prompt box displays the word 'WAIT' which flashes on a cyan background and it is self–clearing when the operation is over.

– Ready Message

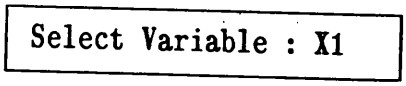
After each output to the worksheet the system flashes the word 'Ready' in the prompt box on a cyan background. This message suspends program execution and pressing any key will resume execution.

d) User Inquiry

When the system needs information a box appears on the screen. There two types of inquiry boxes depending on the nature of the inquiry. The first type deals with selecting the variables to be included in an operation and the second type with general inquiries.

– Variable selection

For variable selection, the system displays following box:

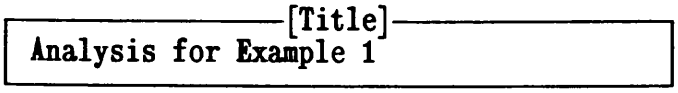


The default variable for analysis is the first variable, X1, and the up/down arrow keys scroll through the variable list. The variable list includes an extra item called 'ALL'

which is used to select all the variables. An item is selected by pressing Enter when it is displayed and to exit from the inquiry without selecting anything, press Esc.

– **General inquiry**

If the system requires information which has no preset list of responses e.g. setting the title for the analysis, execution of a DOS command, entering a file name, etc., a general inquiry box appears. This box has a title at the top which indicates the type of input required. The following is an example of setting the title for the analysis.



e) **Key Summary**

Key	Action
Enter	Selects highlighted item from a menu or confirms an action.
Esc	Exits from a menu or cancels an operation.
Ins	Toggles insert mode on or off.
Del	Deletes the character at the cursor position.
Home	Highlights the first option of a menu.
End	Highlights the last option of a menu.
Arrow keys	Move the menu highlight bar, scroll the variable list.

f) Data File Structure

A tSTAT data file is a row–column array which is the data matrix as defined in multivariate texts (sometimes refered to as a "flat file"), where each row is an observation and each column is a variable. The variables are seperated by spaces (i.e. delimited by spaces). The file has to be a text (ASCII) file so each line of data (observation) is terminated by a Carriage Return/Line Feed <CR–LF>. The data file has to be "even"

that is with the same number of data entries on each row. The file has to be terminated by an end–of–file character EOF.

Since tSTAT identifies an observation on a variable as a value separated from other values by spaces there cannot be any blank or missing values. If these occur tSTAT will run out of data to read before it reaches the end–of–file and will result into an error message.

A valid data file is structured as follows:

```
Obs —>    2.0    35.0    12.0<CR–LF>
          1.5    21.0    24.0
          .      .       .
          .      .       .
          EOF
```

The tSTAT package has a data entry/edit facility but the data file can be prepared outside the package using any standard text editor e.g.Edlin, Qedit, E, etc.. Most wordprocessing packages also have a facility to export ASCII files and so do spreadsheets like Lotus 1–2–3. It is possible to use Dbase III–Plus for creating the data file by copying the .DBF to a delimited file. The default extension for a data file is .DAT.

g) <u>The Data Editor</u>

The tSTAT data editor is one of the main modules within the system. Although it is primarily used to enter and edit the data for subsequent analyses it has numerous other functions. It has a spreadsheet–like mode of use (See Figure 4.7). The main difference is in the naming convention. We shall refer to the array of numbers displayed on the screen as a *datasheet.* In the spreadsheet environment the columns are normally named by letters of the alphabet, however, to maintain the structure of a data matrix the tSTAT data editor names the columns as variable numbers (e.g. column 1 is Var 1) these are termed as *variable names.* For ease of reference, each column also has a letter of the alphabet associated with it which we shall refer to as the *variable tag.* The rows refer to the

observation numbers. So a typical value within the datasheet is referenced by its variable tag and observation number (e.g. the first observation in the second variable is B1).

Arithmetic operations can be performed on the rows and columns of the datasheet. The valid operations are +, −, / and * for addition, subtraction, division and multiplication, respectively. The syntax for an arithmetic operation is:

Cell reference for 1st operand operator Cell reference for 2nd operand

Examples: A1 + C1, B2/A2, B10*A15.

In addition mathematical and statistical functions can be performed on the datasheet values. The following is a list of the available functions.

Function	Syntax	Example
Log x	LOG(c.r.)	LOG(B2)
$\sqrt{x}$	SQRT(c.r.)	SQRT(C10)
Sin x	SIN(c.r.)	SIN(A1)
Cos x	COS(c.r.)	COS(A2)
Tan x	TAN(c.r.)	TAN(D1)
e^x	EXP(c.r.)	EXP(B2)
Σx	c.r.1:c.r.2	a1:a20 (same column)
$\bar{x}$	MEAN(c.r.1:c.r.2)	MEAN(a1:a20) (same column)

Abbreviation: c.r. – cell reference.

The editor has an independent menu (See Figure 4.8) which is called up by pressing the F10 key.

Input/Edit Line

The Input/Edit line appears at the bottom of the editor screen and is used to type in the input to the datasheet together with entering the responses to certain actions.

Status Line

The Status line appears just above the Input/Edit line and shows the current status of the datasheet. The first two items on the status line, the observation number and

variable name, jointly give the current position (cell reference) of the cursor in the datasheet. The variable tag is displayed as the third item. The fourth item refers to the data type in the current position. There are three data types *value*, *text* and *formula* depending on whether the contents of the current position are numeric or alphabetic. If the position is empty then 'Empty' is displayed. The fifth item is a prompt to remind the user that if any formulae have been entered then any changes to the values included in the formulae will be automatically reflected in the formulae results. The last item on the Status line is the prompt for the key which invokes the editor's menu.

Figure 4.7 Data Editor

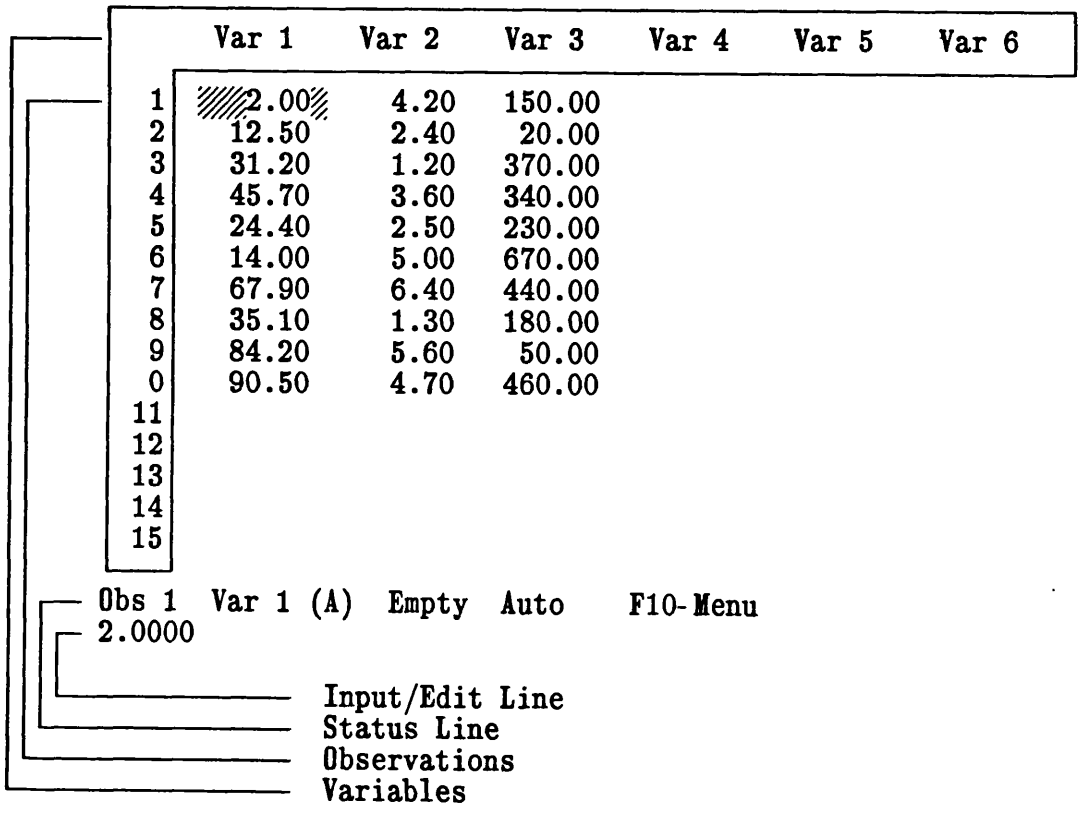
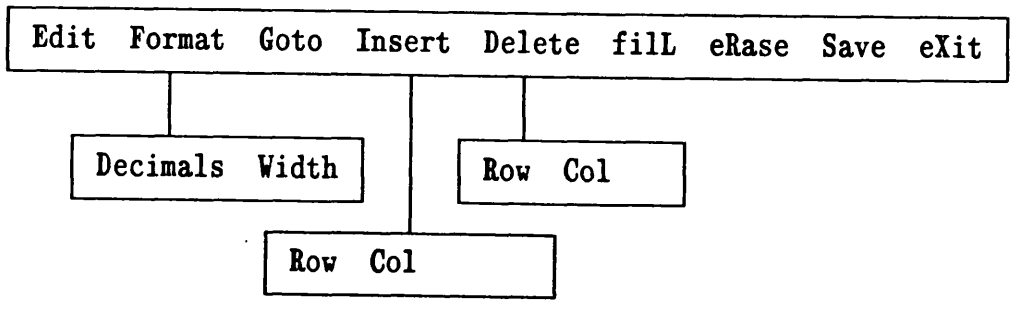


Figure 4.8 Data Editor Menu Structure



The following is a brief description of the options on the data editor menu.

Option	Function
Edit	To edit the contents of the current cursor position.
Format	The Format option is used to set the number of decimal places (the default value is 2) for a particular variable and to also set the column widths (the default value is 10).
Goto	Positions the cursor onto the specified cell reference. You can also use the 'End' key to move to the last observation in the last variable and the 'Home' key to move to the first observation in the first variable.
Insert	Inserts either a row/column before the current row/column.
Delete	Deletes the current row/column.
filL	Fills a column (variable) with an ascending sequence of numbers. To use it you specify the first and last cell references together with the start, increment and stop values for the numbers.
eRase	Clears the values in the current datasheet and places the cursor into the first observation and first variable position.
Save	This is used to save the current datasheet under a specified filename.
eXit	Returns to the tSTAT main menu.

N.B. The data editor allows the entry of both numeric and alphabetic characters but for analysis the saved data file MUST conform to the specification of a valid data file. If the specification is violated tSTAT displays an error message.

h) Technical Specifications

The tSTAT package has the following specifications:

Main memory (RAM)	$\geq$ 640KB
Disk Space	$\geq$ 700KB
Operating System	DOS 3.0+
Video Display Unit (VDU)	Mono, CGA, EGA, VGA+

Number of variables	$\leq$ 10 (prototype version)
Number of observations	$\leq$ 100 (prototype version)

The following section discusses and describes the use of all the tSTAT menu options.

4.4.2 tSTAT Functions

This section outlines the use of each of the tSTAT functions. The layout of the description of each function is broken down into the following parts:

PURPOSE	– a brief description of what the function does.
CALCULATIONS	– formulae for the calculations (if any).
INPUT	– describes the expected input to the function.
OPERATION	– describes the use of the function.
OUTPUT	– shows the format of the output from the function.

Appendix B is an example of a full analysis on a typical data set and so is used as a reference for each of the functions.

4.4.2.1 Main menu

Figure 4.9 shows the Mainmenu which calls up all the functions.

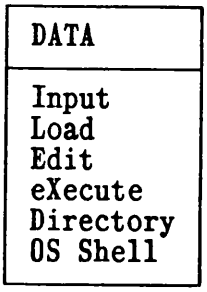
Figure 4.9 Mainmenu

The Data option deals with the data entry/load/edit facilities. It also includes some DOS related options namely execution of a DOS command from within tSTAT, the directory display and the shelling to DOS. The Stats option contains the general descriptive statistics and related analyses. It also includes the selection of symmetrising transformations using Hinkley's "quick" method and regression analysis. The Transformations option includes the two main approaches for transformations to

multivariate normality namely; the likelihood approach and the proposed SURCON approach. It also has an option for carrying out manual (i.e. user–defined) Box–Cox transformations on the data. The graphical output is done by the Plots option. This includes the scatter plots, box and whiskers plots, Gnanadesikan's symmetry plots, probability plots, index plots and the proposed Stalactite plots. The outliers option includes three tests for outliers; the classical Mahalanobis distance approach, the Hat Matrix approach and the proposed Stalactite Analysis approach. The Utilities option includes general housekeeping functions like setting the analysis title, printing the logfile, renaming/deleting files, changing directory and setting the Stalactite algorithm parameters. The Quit option is used to exit the package.

4.4.2.2 Data Option

Figure 4.10 Data Menu



Input

PURPOSE Used to enter new data.

CALCULATIONS None.

INPUT None.

OPERATION When this option is selected the Data Editor (See Section 4.4.1) is invoked so the data can be entered and edited. When the data entry is completed you can save the file using the Data Editor Save menu option.

OUTPUT Data file by the name specified in the Data Editor Save menu option.

Load

PURPOSE Loads an existing data file from disk.

CALCULATIONS None.

INPUT File name.

OPERATION Selecting this option produces a box with a default setting the default extension for data files (.DAT). You can type in the name of the file required or press Enter to obtain a list of all the data files in the current work directory. If the list is displayed move the highlighted bar to the required file and press Enter to select it. You can also move the highlighted bar by typing the first letter of the required file; this moves the bar to the first file with a match for the letter.

OUTPUT Data from the data file. The system automatically ascertains the number of observations and number of variables. If an error occurs in reading the data an error message box appears. The worksheet title is set to the data file name. The file name, the number of variables and number of observations is also displayed within the worksheet. The first five observations are also listed to allow the user to confirm that the right data is being used.

Edit

PURPOSE Edits data from the current data file.

CALCULATIONS None.

INPUT None.

OPERATION When this option is selected the current file is passed on to the Data Editor and can then be edited.

OUTPUT Edited version of the current data file.

eXecute

PURPOSE To execute a DOS command from within tSTAT.

CALCULATIONS None.

INPUT The DOS command to be executed.

OPERATION Selecting this option produces an empty inquiry box. Type in the DOS command and press Enter.

OUTPUT None.

Directory

PURPOSE To obtain a directory listing of files in the specified directory.

CALCULATIONS None.

INPUT The path for the required listing. The default is the current work directory.

OPERATION Selecting this option produces an inquiry box with a 'wild card' (*.*). You can specify the required path or press Enter to use the current path.

OUTPUT The list of files appears in a scrolling box on the screen.

OS Shell

PURPOSE To suspend program execution and exit to the DOS prompt.

CALCULATIONS None.

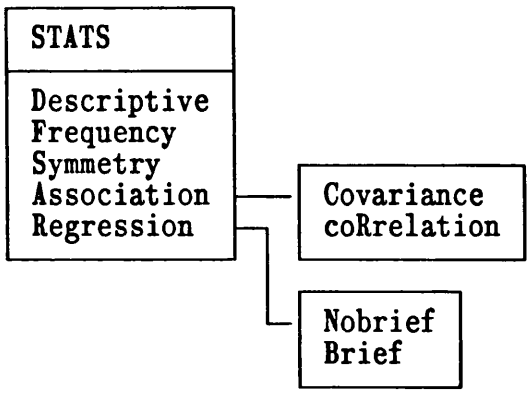
INPUT None.

OPERATION Selecting this option will pass execution to DOS. Type 'EXIT' at the DOS prompt to return to tSTAT.

OUTPUT None.

4.4.2.3. The Statistics Option

Figure 4.11 Statistics Menu



Descriptive

PURPOSE Computes descriptive statistics for the current data file.

CALCULATIONS See Section 4.3.1.

INPUT One or more variables.

OPERATION When this option is selected a variable selection inquiry box appears. You can use the up/down arrow keys to selected the required box item.

OUTPUT A table of descriptive statistics. The statistics computed are the mean (and its standard error), 10%–trimmed mean, variance, standard deviation, minimum, maximum, range, skewness and kurtosis (with the respective standard error). (See Appendix B).

Frequency

PURPOSE Computes and displays a frequency table and histogram.

CALCULATIONS Computes the maximum and minimum and groups the data into the number of specified class intervals (default is 6).

INPUT One or more variables.

OPERATION When this option is selected a Variable Selection inquiry box appears. You can use the up/down arrow keys to selected the required box item.

OUTPUT A frequency table with the class frequency (count, percentage of total frequency and cumulative percentage). A histogram based on the number of classes is also displayed. (See Appendix B).

Symmetry

PURPOSE Derives Hinkley's "quick" transformations to approximate symmetry.

CALCULATIONS See Section 3.2.2.

INPUT One or more variables.

OPERATION When this option is selected a variable selection inquiry box appears. You can use the up/down arrow keys to selected the required box item.

OUTPUT A summary of the calculations based on five transformation parameters t=–1, 0.5, 0, 1 and 2. The output also includes a recommended transformation. (See Appendix B).

Association

– Covariance

PURPOSE Computes the covariance matrix for the current data set.

CALCULATIONS For the data matrix {x} with n observations on p variables the i,j–th element of the covariance matrix is

$$s_{ij} = \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) \tag{4.21}$$

INPUT All the variables in the data file.

OPERATION Highlight the option and press Enter to select it.

OUTPUT The covariance matrix of the data set. (See Appendix B).

– Correlation

PURPOSE Computes the correlation matrix for the current data set.

CALCULATIONS For the data matrix {x} with n observations on p variables the i,j–th element of the correlation matrix is

$$r_{ij} = \frac{\sum_{k=1} (x_{ki}^{-} - x_i)(x_{kj}^{-} - x_j)}{\left\{ \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)^2 \sum_{k=1}^{n} (x_{kj} - \bar{x}_j)^2 \right\}^{1/2}} \tag{4.22}$$

$$= s_{ij} / \sqrt{(s_{ii} \times s_{jj})}$$

INPUT All the variables in the data file.

OPERATION Highlight the option and press Enter to select it.

OUTPUT The correlation matrix of the data set. (See Appendix B).

Regression

PURPOSE Fits a linear regression model of the form $y = \alpha + \beta x$.

CALCULATIONS Computes $\hat{\alpha}$ and $\hat{\beta}$ parameter estimates for α and β where

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (4.23)$$

and $\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$ where $\bar{x}$, $\bar{y}$ are the arithmetic means of x and y, respectively.

INPUT The independent variable y and dependent variable x.

OPERATION When this option is selected a variable selection inquiry box appears. You select the independent variable y and the dependent variable x from the list of variables in the data matrix. You cannot select the same data variable in both cases.

OUTPUT If the Brief option is used the output consists of the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ with the respective standard errors, t–statistics, the corresponding p values and an analysis of variance table. Using the Nobrief option additonally includes a listing of the fitted values $\hat{y}$, the residuals, the standardised residuals and the diagonal elements of the Hat matrix. The output also includes residual plots (the residuals vs the fitted values), a normal plot of the residuals, an index plot of the residuals and an index plot of the diagonal elements of the hat matrix. (See Appendix B).

4.4.2.4 Transformations

Figure 4.12 Transformations Menu

TRANSFORMATIONS
MLE Approach SURCON Approach Box-Cox (Manual)

MLE Approach

PURPOSE Computes the transformations to multivariate normality using the

likelihood approach.

CALCULATIONS See Section 3.3.

INPUT All the variables in the data file.

OPERATION When this option is selected an inquiry box appears which requests for the initial (hypothesised) values for the transformation parameters. The defaults for these values is 1, no transformation.

OUTPUT The maximum likelihood estimates for the transformation parameters.

SURCON Approach

PURPOSE Computes the transformations to multivariate normality using the proposed SURCON approach.

CALCULATIONS See Section 3.4.3.

INPUT All the variables in the data file.

OPERATION When this option is selected an inquiry box appears which requests for the initial (hypothesised) values for the transformation parameters. The defaults for these values is 1, no transformation.

OUTPUT The SURCON estimates for the transformation parameters together with the associated statitics. (See Appendix B).

Box–Cox (Manual)

PURPOSE To enter user–defined transformation parameters.

CALCULATIONS Transforms the selected variable(s) according to the Box–Cox transformation of Section 3.3.

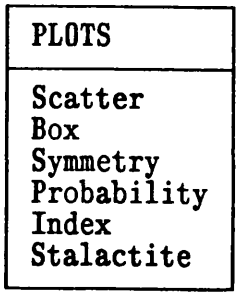
INPUT The transformation parameter.

OPERATION Selecting this option produces an inquiry box in which the desired transformation parameter value can be entered.

OUTPUT The transformed data.

4.4.2.5. Plots

Figure 4.13 Plots Menu



Scatter

PURPOSE To produce a scatter plot of two variables.

CALCULATIONS None.

INPUT The two selected variables.

OPERATION Selecting this option produces a variable selection inquiry box from which the two variables can be selected.

OUTPUT A scatter plot of the two variables.

Box

PURPOSE To produce a box and whiskers plot for a variable.

CALCULATIONS None.

INPUT The selected variable.

OPERATION A variable selection inquiry box appears from which the required variable can be selected.

OUTPUT A box and whiskers plot for a variable. The output also consists of values for quartiles of the variable.

Symmetry

PURPOSE To produce Gnanadesikan's symmetry plots for a selected variable. See Section 3.2.1.1.

CALCULATIONS None.

INPUT The selected variable.

OPERATION Selecting this option produces two inquiry boxes. The first is to select the plot type (Type I, Type II and Type III) and the second is to select the variable.

OUTPUT A Gnanadesikan's symmetry plot for the selected variable.

Probability

PURPOSE To produce a (normal/χ^2) probability plot.

CALCULATIONS None.

INPUT The variable/statistic for which the plot is required.

OPERATION Selecting this option produces two inquiry boxes. The first is to select the plot type (normal or χ^2) and the second is to select the variable or statistic for which the plot is required.

OUTPUT A (normal/χ^2) probability plot of the selected variable or statistic.

Index

PURPOSE To produce an index plot.

CALCULATIONS None.

INPUT The transformation parameter.

OPERATION Selecting this option produces an inquiry box from which the variable or statistic for which the plot is required can be selected.

OUTPUT An index plot of the selected variable or statistic.

Stalactite

PURPOSE To produce a Stalactite plot.

CALCULATIONS See Section 2.7.

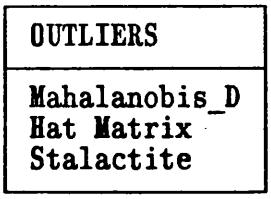
INPUT All the variables in the data file.

OPERATION Highlight the option and press enter to select it.

OUTPUT The Stalactite plot of the current data set.

4.4.2.6 Outliers

Figure 4.14 Outliers Menu



Mahalanobis_D

PURPOSE To test for outliers using the classical Mahalanobis distance approach.

CALCULATIONS See Section 2.4.

INPUT All the variables in the data file.

OPERATION Highlight the option and press enter to select it.

OUTPUT A list of the Mahalanobis distances for each observation. If an observation exceeds a given cut–off point (the default is the expected maximum χ^2 value) it has an arrow placed next to it. The output also consists of the discordancy tests for a single outlier from Section 2.2.1.

Hat Matrix

PURPOSE To test for outliers using the Hat matrix approach.

CALCULATIONS See Section 2.5.

INPUT All the variables in the data file.

OPERATION Highlight the option and press enter to select it.

OUTPUT A list of the diagonal elements of the Hat matrix. If an observation exceeds a given cut–off point (the default value is the 2p/n) it has an arrow placed next to it.

Stalactite

PURPOSE To test for outliers using the Stalactite analysis approach.

CALCULATIONS See Section 2.7.

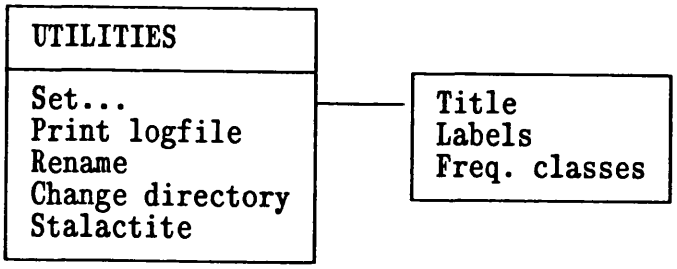
INPUT All the variables in the data file.

OPERATION Highlight the option and press enter to select it.

OUTPUT A list of the diagnostic quantities from the Stalactite analysis.

4.4.2.7. Utilities

Figure 4.15 Utilities Menu



Set...

- Title

PURPOSE To set the title for the analysis.

CALCULATIONS None.

INPUT None.

OPERATION Selecting this option produces an inquiry box. Type in a title of upto 72 characters and press Enter.

OUTPUT The title will appear at the top of the printout in the logfile. For clarity it is recommended that this option should be run before reading in a data file otherwise tSTAT will use the default title of "Untitled Analysis".

- Labels

PURPOSE To provide text labels for the data variables.

CALCULATIONS None.

INPUT Text label for each variable.

OPERATION An inquiry box appears. The box will scroll through the variable list to allow you tp type in the respective labels.

OUTPUT The tStat package uses X1, X2,...,Xp as the default variable labels. Although, this option will set new labels all reference to variables will be based on the default ones. A list of the labels can be obtained at any time by pressing the Status key (F2) and is printed at the begining of the logfile.

– Freq. classes

PURPOSE To set the number of classes (bins) used in the frequency table option.

CALCULATIONS None.

INPUT The number of classes.

OPERATION Type the number of required classes into the inquiry box and press Enter.

OUTPUT None.

Print logfile

PURPOSE To send a copy of the current logfile to the printer.

CALCULATIONS None.

INPUT None.

OPERATION Make sure the printer is on and ready then press Enter.

OUTPUT A hard copy of the logfile. N.B.The logfile is an ordinary text (ASCII) file and so can be printed outside tSTAT, this also means that it can be imported directly into many wordprocessing packages.

Rename

PURPOSE To rename a disk file.

CALCULATIONS None.

INPUT The old and new filenames.

OPERATION Type the old filename into the inquiry box (if tSTAT fails to find the file an error occurs) and then type in the new filename. If a file with the

same name exists on disk you are asked whether you would like to replace it. Press Esc to cancel the operation.

OUTPUT None.

Change directory

PURPOSE To set the DOS path from which all the input files are read from and all the output is sent to.

CALCULATIONS None.

INPUT DOS path.

OPERATION Type a valid DOS path (e.g. C:\MYDATA) into the inquiry box and press Enter. If the path does not exist an error occurs. The directory in which tSTAT is started from is the default path.

OUTPUT None.

Stalactite

PURPOSE To set the style of the Stalactite plot.

CALCULATIONS None.

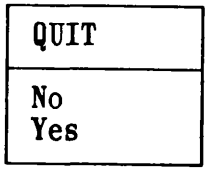
INPUT The plot style.

OPERATION Select the Short option to obtain a shortened version (first and last five iterations) of the Stalactite plot and the Long option for the full plot. The default value is Long.

OUTPUT None.

4.4.2.8. Ending a tSTAT Session

Figure 4.16 Quit Menu



PURPOSE To exit from tSTAT.

CALCULATIONS None.

INPUT None.

OPERATION Select the No option to continue with tSTAT and Yes to quit to the DOS prompt. It is always advisable to end a tSTAT session using this option because this ensures that all the files are properly closed. Do not switch the computer off in the middle of a session.

OUTPUT None.

4.5 Example

Appendix B shows an example of the output produced in the logfile from a typical analysis run from the tSTAT package. The data used is from Example E.2 of Section 2.

CHAPTER FIVE

5.0 CONCLUSIONS AND RECOMMENDATIONS

The thesis addresses the general problem of transformations of multivariate observations to multivariate normality. In order to carry out the transformations the data have to undergo a series of tests to check for conformity to the multivariate normal model. These tests intrinsically involve screening for any multivariate outliers; a problem which is not trivial, especially when there are several of them. In tackling these problems the thesis proposes three main tools; an outlier detection method for multivariate observations, a joint transformations method and a statistical computer package to perform the required analyses.

This chapter discusses some of the conclusions drawn from the study and proposals for future areas of work. Since the computer package is the combination of all the methods presented, it shall be discussed first.

The whole process of data screening is exploratory in nature. It is an important and integral part of data analysis. However, apart from experienced statistical analysts this phase of data analysis is seldom given enough emphasis. In most of the statistical analyses done by the inexperienced users effort is confined to carrying out the usual range and edit checks on the data. Tests for statistical consistency are hardly ever performed. There are several reasons for this amongst which are ignorance of how to perform the screening. Even with the knowledge, the computational requirements may seem prohibitive. In the absence of specialised software the time required to perform the screening is almost invariably relatively long and the process is neglected. The statistical package developed (tSTAT) attempts to provide solutions to these points. It is a specialised statistical consistency checker and transformations package which combines the complexities of the theory and computations with flexibility and ease of use. It is entirely user–friendly and menu–driven with a comprehensive online context–sensitive help system. The package, therefore, guides

the inexperienced user in testing for the statistical consistency of the data by providing the requisite options. Further the experienced user is spared the chore of adapting existing software. The list of tests included in the package is not exhaustive but consists of an adequate set to carry out a convincing screening analysis before conducting the comprehensive data analysis using the existing well–known software packages. Future work on the package would involve the extension of the number of possible tests and the inclusion of some of the standard statistical analyses.

The detection of outliers in multivariate observations is a well recognised problem and is increasingly becoming addressed. The main reason for its complexity is due to the well–known masking and swamping phenomena. Many solutions have been proposed but they generally suffer from the need to include prior information that is rarely available [Beckman and Cook 1983]. The classical approach using Mahalanobis distances may fail to indicate the presence of any outliers when many are present due to the influence of these outliers on the estimates of the means and covariance matrix used in calculating the distances. Rousseeuw and van Zomeren [1990] propose an approach using distances based on robust estimates of location and covariance. They emphasise that the approach does not require prior input of tuning constants. The outliers are judged relative to the metric of the minimum volume ellipsoid (MVE) containing the majority of the data. However, Cook and Hawkins [1990] show that this procedure may indicate an over–abundance of outliers, the identity of which can change dramatically with small changes in the parameters of the algorithm for robust estimation. They commend instead backward procedures in which outliers are sequentially detected and deleted, starting from all n observations. The approach proposed in the thesis, the Stalactite analysis, considers a forward procedure which starts by using a small random subset of the data for estimation of the means and covariances required for the calculation of the Mahalanobis distances. The size of the subset is then increased in such a way as to exclude outliers. The procedure unambiguously

identifies the outliers in the example studied by Cook and Hawkins [Atkinson and Mulira 1992], the 'notorious' wood gravity data [Rousseeuw and Leroy 1987, p.243]. It also has the advantage of computational modesty compared to the minimum volume elipsoid approach while yielding a simple graphical summary, the Stalactite plot.

The proposed method performs well even in the presence of appreciable masking. If the iterations were initiated from within an outlying cluster the procedure has been found to recover from the cluster and thus still be able to identify all the outlying observations. The procedure also produces a number of useful diagnostic quantities such as the stalactite scores which are a metric of the relative presence of an observation within the outlying set during the iterations and the contamination index which is the ratio of 'bad' observations to 'good' ones. In addition, a summary of the behaviour of the computed means during the iterations can be obtained using the means plot. In response to Cook and Hawkins general reactions to the MVE approach, the Stalactite analysis has some desirable properties. Regarding the point about tuning constants the Stalactite analysis does not require any such prior information. The second point is in connection with the fact that the MVE approach seems to find many outliers even in 'innocent' data. The Stalactite plot, together with the associated index and probability plots, has been shown to provide clear conclusions in which the observations are not too sparse. In Atkinson and Mulira [1992] two such examples were used with 100 replicates of the Stalactite analysis and they both led to the identification of the same outliers. The need for a graphical display showing how the outliers are distributed relative to the rest of the data is responded to by plotting the elements of the pull vector which are the signum functions of the differences between the full sample mean and the compound mean estimates. Finally, the computational requirements are minimal compared to the MVE approach and so the method can easily be implemented on PC–type machines.

The proposed method, although efficient for multivariate observations with an assumed underlying multivariate normal model, can be extended to other models eg. linear

regression models. It has been tried in this context and appears to be satisfactory.

Lastly, the problem of obtaining joint transformations to multivariate normality has not been accorded the same coverage. There are several techniques for data–based transformations for univariate observations, however, the major technique in the multivariate case is based on the numerical maximisation of the loglikelihood function. Since there are several parameters to be estimated the resulting maximisation problem is of high dimension and the choice of the maximisation algorithm can greatly affect the speed at which the results can be obtained. The proposed Surcon analysis provides a complementary procedure which requires far less computational time in terms of function evaluations. It, however, requires more calculations within an iteration, which may be seen as some kind of contradiction, but the results from these calculations implicitly provide several useful statistical quantities including confidence limits, single equation score tests and a test for the diagonality of the covariance matrix. The method can be extended from transforming only multivariate data to performing transformations on other types of data eg. individual residuals from linear regression and the residuals of constructed variables. A further extension could be to provide some graphical summaries. In this regard, analogues of the added variable plots can be sought eg. 3–dimensional equivalents or added variable plot matrices where each plot would correspond to a particular value of the transformation parameter.

In conclusion, the combination of the three tools developed provides yet another "stepping stone" on the path towards the much needed solutions to the problem of multivariate outlier detection and transformations to multivariate normality.

APPENDIX A(i)

Expected Maximum χ^2_p ($n < 50$)

n	Degrees of Freedom, p									
	1	2	3	4	5	6	7	8	9	10
5	2.635	4.554	6.208	7.741	9.201	10.612	11.986	13.331	14.655	15.959
6	2.919	4.915	6.622	8.197	9.693	11.135	12.539	13.912	15.260	16.588
7	3.164	5.222	6.970	8.579	10.104	11.572	13.000	14.394	15.763	17.111
8	3.379	5.488	7.271	8.908	10.457	11.947	13.394	14.807	16.193	17.557
9	3.572	5.724	7.536	9.197	10.767	12.275	13.739	15.167	16.568	17.945
10	3.746	5.935	7.773	9.454	11.042	12.567	14.045	15.487	16.900	18.289
11	4.050	6.302	8.183	9.898	11.515	13.067	14.569	16.034	17.468	18.878
13	4.186	6.464	8.362	10.092	11.722	13.285	14.798	16.272	17.716	19.133
14	4.312	6.614	8.528	10.271	11.913	13.486	15.008	16.491	17.943	19.369
15	4.430	6.753	8.683	10.438	12.090	13.673	15.204	16.695	18.154	19.587
16	4.541	6.884	8.827	10.594	12.256	13.847	15.386	16.884	18.350	19.790
17	4.646	7.007	8.963	10.740	12.411	14.010	15.556	17.062	18.534	19.980
18	4.746	7.124	9.091	10.877	12.556	14.163	15.716	17.228	18.707	20.158
19	4.840	7.234	9.212	11.007	12.694	14.308	15.867	17.385	18.870	20.326
20	4.930	7.338	9.327	11.130	12.825	14.445	16.010	17.534	19.024	20.485
21	5.016	7.438	9.436	11.247	12.948	14.575	16.146	17.675	19.170	20.636
22	5.098	7.533	9.540	11.359	13.066	14.699	16.275	17.809	19.308	20.779
23	5.177	7.624	9.639	11.465	13.179	14.817	16.399	17.937	19.441	20.915
24	5.252	7.711	9.735	11.567	13.287	14.930	16.516	18.059	19.567	21.046
25	5.325	7.795	9.826	11.665	13.390	15.038	16.629	18.176	19.688	21.171
26	5.395	7.875	9.914	11.758	13.489	15.141	16.737	18.288	19.804	21.290
27	5.463	7.953	9.998	11.848	13.584	15.241	16.841	18.396	19.915	21.405
28	5.528	8.027	10.079	11.935	13.675	15.337	16.941	18.500	20.023	21.516
29	5.591	8.100	10.158	12.019	13.764	15.429	17.037	18.600	20.126	21.622
30	5.653	8.169	10.234	12.100	13.849	15.519	17.130	18.696	20.225	21.725
31	5.712	8.237	10.307	12.178	13.931	15.605	17.220	18.789	20.321	21.824
32	5.770	8.302	10.378	12.254	14.011	15.688	17.306	18.879	20.414	21.919
33	5.826	8.366	10.447	12.327	14.088	15.769	17.390	18.966	20.504	22.012
34	5.880	8.428	10.514	12.398	14.163	15.847	17.472	19.050	20.591	22.102
35	5.933	8.488	10.579	12.467	14.236	15.923	17.551	19.132	20.676	22.189
36	5.984	8.546	10.642	12.534	14.306	15.997	17.627	19.211	20.758	22.273
37	6.035	8.602	10.703	12.599	14.375	16.069	17.702	19.289	20.837	22.355
38	6.084	8.658	10.763	12.663	14.441	16.138	17.774	19.364	20.915	22.435
39	6.131	8.711	10.821	12.724	14.506	16.206	17.845	19.437	20.990	22.512
40	6.178	8.764	10.877	12.784	14.570	16.272	17.914	19.508	21.064	22.588
41	6.224	8.815	10.933	12.843	14.631	16.336	17.980	19.577	21.135	22.661
42	6.268	8.865	10.986	12.900	14.691	16.399	18.046	19.644	21.205	22.733
43	6.312	8.914	11.039	12.956	14.750	16.460	18.109	19.710	21.273	22.803
44	6.354	8.962	11.091	13.011	14.807	16.520	18.171	19.774	21.339	22.871
45	6.396	9.008	11.141	13.064	14.863	16.579	18.232	19.837	21.404	22.938
46	6.437	9.054	11.190	13.116	14.918	16.636	18.291	19.899	21.467	23.003
47	6.477	9.099	11.238	13.167	14.972	16.692	18.349	19.959	21.529	23.067
48	6.516	9.143	11.285	13.217	15.024	16.746	18.406	20.017	21.589	23.129
49	6.555	9.186	11.332	13.266	15.075	16.800	18.462	20.075	21.649	23.190
50	6.592	9.228	11.377	13.314	15.126	16.852	18.516	20.131	21.707	23.250

For a sample of size n measured on p variables the expected maximum χ^2_p is

$$E[\text{Max } \chi^2] = \chi^2_p\left(\frac{n - 0.5}{n}\right).$$

Eg. n=31, p=2 , $E[\text{Max } \chi^2] = 8.237$.

APPENDIX A(ii)

Expected Maximum χ_p^2 (n > 60)

n	Degrees of Freedom, p									
	1	2	3	4	5	6	7	8	9	10
60	6.935	9.608	11.785	13.746	15.578	17.324	19.005	20.637	22.228	23.785
70	7.228	9.932	12.131	14.110	15.960	17.721	19.417	21.062	22.666	24.235
80	7.483	10.212	12.430	14.426	16.290	18.064	19.772	21.428	23.043	24.623
90	7.710	10.461	12.695	14.704	16.580	18.366	20.084	21.750	23.374	24.964
100	7.915	10.684	12.931	14.953	16.840	18.635	20.363	22.038	23.670	25.267
110	8.100	10.885	13.146	15.178	17.074	18.878	20.614	22.297	23.936	25.540
120	8.271	11.070	13.341	15.383	17.288	19.100	20.843	22.533	24.179	25.789
130	8.429	11.240	13.521	15.572	17.484	19.304	21.054	22.749	24.402	26.017
140	8.575	11.398	13.688	15.746	17.666	19.492	21.248	22.950	24.607	26.228
150	8.712	11.546	13.844	15.909	17.835	19.667	21.429	23.136	24.798	26.424
160	8.840	11.684	13.989	16.061	17.993	19.831	21.598	23.309	24.977	26.607
170	8.961	11.813	14.126	16.204	18.142	19.984	21.756	23.472	25.144	26.779
180	9.076	11.936	14.255	16.338	18.282	20.129	21.905	23.626	25.302	26.940
190	9.184	12.052	14.377	16.466	18.414	20.266	22.046	23.771	25.450	27.092
200	9.288	12.162	14.493	16.586	18.539	20.395	22.180	23.908	25.591	27.237

For a sample of size n measured on p variables the expected maximum χ_p^2 is

$$E[\text{Max } \chi^2] = \chi_p^2\left(\frac{n - 0.5}{n}\right).$$

Non tabulated values can be obtained by interpolation.

APPENDIX B

```
////////////////
// tSTAT v0.00 //
////////////////

/************** tSTAT Example Data Set **************/

Data File: D:\TURBOC\TSTAT.DAT

No. of obs 50    No. of vars 2

** Data Listing **

Case      X1        X2

   1     11.12     26.04
   2     12.44     10.53
   3      4.59     10.00
   4      5.65     18.00
   5      7.08     14.00
   .       .         .
   .       .         .
   .       .         .

** Summary Statistics ( All Variables ) **

         Variable      X1        X2
 Statistic
  Mean                11.90     21.74
  10%-Trim. Mean      12.75     23.68
  S.E.( Mean)          0.80      1.17
  Variance            32.29     68.84
  Std. Dev.            5.68      8.30

  Minimum              2.50      6.23
  Maximum             30.00     44.00
  Range               27.50     37.77

  Skewness             1.29      0.26
  S.E.(Skewness)       0.35      0.35
  Kurtosis             2.24     -0.25
  S.E.(Kurtosis)       0.69      0.69

 ** Rao's Score Test for Multivariate Normality **

 Multivariate Skewness (Full sample) = 0.0649
 Multivariate Kurtosis (Full sample) = 5.3819

                        ABS(Z)        T3        T4         T
                        2.6721    0.5410    5.5177    6.0587
 Degress of freedom                    4         5         9
 Chi-square p-Value                0.9694    0.3560    0.7340
```

** Frequency Distribution for variable X1 **

Class Interval	Count	%age	Cum.		Histogram
(2.50 , 8.00]	12	(24.0%)	24.0%	-	|**********
(8.00 , 13.50]	22	(44.0%)	68.0%	-	|**********2
(13.50 , 19.00]	12	(24.0%)	92.0%	-	|**********
(19.00 , 24.50]	1	(2.0%)	94.0%	-	|*
(24.50 , 30.00]	1	(2.0%)	96.0%	-	|*
(30.00 plus)	2	(4.0%)	100.0%	-	|**
Total	50	(100.0%)			

** Frequency Distribution for variable X2 **

Class Interval	Count	%age	Cum.		Histogram
(6.23 , 13.78]	10	(20.0%)	20.0%	-	|*********
(13.78 , 21.33]	15	(30.0%)	50.0%	-	|***************
(21.33 , 28.88]	15	(30.0%)	80.0%	-	|***************
(28.89 , 36.44]	8	(16.0%)	96.0%	-	|********
(36.44 , 43.99]	1	(2.0%)	98.0%	-	|*
(43.99 plus)	1	(2.0%)	100.0%	-	|*
Total	50	(100.0%)			

** Hinkley's Quick Transformations to Symmetry for variable X1 **

T =	-1	0	0.5	1	2
Sample Mean	-0.106	2.369	6.721	11.900	86.624
Sample Median	-0.089	2.414	6.687	11.180	62.498
s1 (Std.Dev)	0.122	2.417	6.903	13.187	126.860
s2 (IQ Range)	0.050	0.553	1.856	6.270	72.894
dt using s1	-0.132	-0.019	0.005	0.055	0.190
dt using s2	-0.323	-0.081	0.018	0.115	0.331

** Recommended transformation for symmetry **

Square root

** Hinkley's Quick Transformations to Symmetry for variable X2 **

T =	-1	0	0.5	1	2
Sample Mean	-0.055	2.997	9.147	21.741	269.736
Sample Median	-0.047	3.058	9.227	21.290	226.884
s1 (Std.Dev)	0.062	3.028	9.329	23.256	332.377
s2 (IQ Range)	0.023	0.489	2.271	10.600	234.472
dt using s1	-0.132	-0.020	-0.009	0.019	0.129
dt using s2	-0.356	-0.123	-0.035	0.043	0.183

** Recommended transformation for symmetry **

Square root

** Var-Covariance Matrix **

Var		
X1	32.29	
X2	26.71	68.18

** Correlation Matrix **

Var		
X1	1.000	
X2	0.569	1.000

** Regression Analysis **

Dependent Variable : X2 Independent variable(s) : X1

Regression equation : X2 = 11.896 + 0.827 X1

Predictor	Coeff	Std.Err	t-ratio	p-value
Constant	11.896	2.270	5.241	0.0000
X1	0.827	0.172	4.798	0.0000

Deviance = 2258.10 d.o.f. = 48 s = 6.86 R-Sq = 32.4%

** Analysis of Variance **

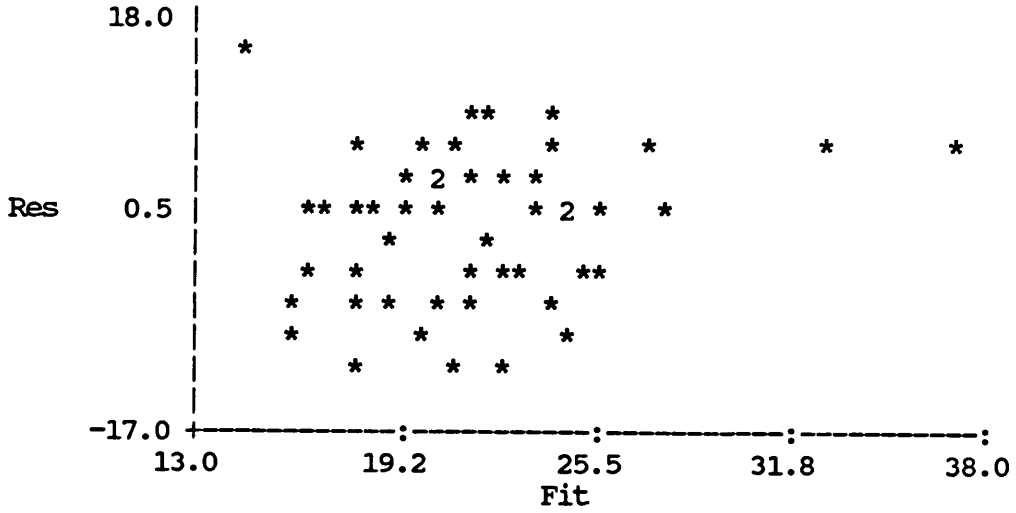
Source	DF	SS	MS	F	p-value
Regression	1	1082.838	1082.838	23.018	0.0001
Error	48	2258.099	47.044		
Total	49	3340.937			

** Residuals Listing **

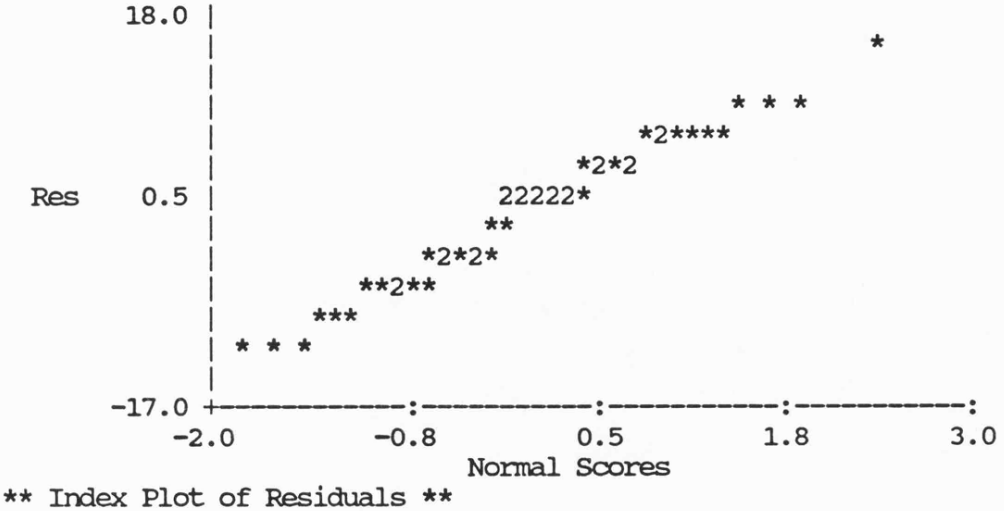
Case	X1	Obs X2	Fit	Residual	Std.Res	h
1	11.12	26.04	21.10	4.94	0.73	0.020
2	12.44	10.53	22.19	-11.66	-1.72	0.020
3	4.59	10.00	15.69	-5.69	-0.84	0.015
4	5.65	18.00	16.57	1.43	0.21	0.016
5	7.08	14.00	17.75	-3.75	-0.55	0.017
6	6.79	12.00	17.51	-5.51	-0.81	0.017
7	18.07	33.47	26.85	6.62	0.98	0.024
8	11.76	20.37	21.62	-1.25	-0.18	0.020
9	4.88	12.00	15.93	-3.93	-0.58	0.016
10	14.65	18.31	24.02	-5.71	-0.84	0.022
11	11.24	13.74	21.19	-7.45	-1.10	0.020
12	7.25	20.00	17.89	2.11	0.31	0.017
13	10.07	15.00	20.23	-5.23	-0.77	0.019
14	9.64	27.41	19.87	7.54	1.11	0.019
15	11.32	29.89	21.26	8.63	1.27	0.020
16	30.00	44.00	36.71	7.29	1.08	0.031

17	15.01	26.43	24.31	2.12	0.31	0.022
18	12.13	30.84	21.93	8.91	1.31	0.020
19	16.51	22.26	25.55	-3.29	-0.49	0.023
20	9.18	20.58	19.49	1.09	0.16	0.018
21	16.00	22.00	25.13	-3.13	-0.46	0.023
22	4.79	6.23	15.86	-9.63	-1.41	0.016
23	10.00	24.00	20.17	3.83	0.56	0.019
24	15.20	16.50	24.47	-7.97	-1.18	0.022
25	14.55	33.29	23.93	9.36	1.38	0.022
26	8.53	12.00	18.95	-6.95	-1.02	0.018
27	8.45	17.15	18.89	-1.74	-0.26	0.018
28	25.00	40.00	32.58	7.42	1.10	0.028
29	14.87	32.53	24.20	8.33	1.23	0.022
30	19.22	28.98	27.80	1.18	0.17	0.025
31	12.54	19.10	22.27	-3.17	-0.47	0.020
32	6.77	6.91	17.50	-10.59	-1.56	0.017
33	13.21	18.49	22.82	-4.33	-0.64	0.021
34	9.58	11.74	19.82	-8.08	-1.19	0.019
35	16.75	27.55	25.75	1.80	0.27	0.023
36	5.20	18.00	16.20	1.80	0.26	0.016
37	10.79	8.68	20.82	-12.14	-1.79	0.019
38	6.91	23.81	17.61	6.20	0.91	0.017
39	2.50	30.00	13.96	16.04	2.35	0.014
40	12.65	26.81	22.36	4.45	0.66	0.020
41	14.16	28.91	23.61	5.30	0.78	0.021
42	15.00	26.00	24.31	1.69	0.25	0.022
43	14.00	24.00	23.48	0.52	0.08	0.021
44	10.31	26.00	20.43	5.57	0.82	0.019
45	8.84	25.00	19.21	5.79	0.85	0.018
46	11.56	18.41	21.46	-3.05	-0.45	0.020
47	7.38	20.33	18.00	2.33	0.34	0.017
48	30.00	20.00	36.71	-16.71	-2.48	0.031
49	9.87	22.30	20.06	2.24	0.33	0.019
50	10.98	27.44	20.98	6.46	0.95	0.019

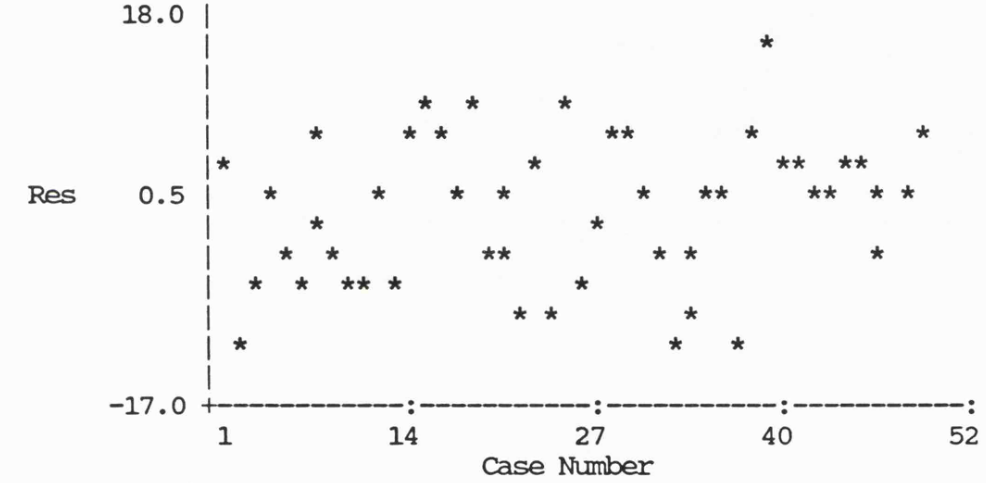
** Residual Plot (Residuals vs Fit) **



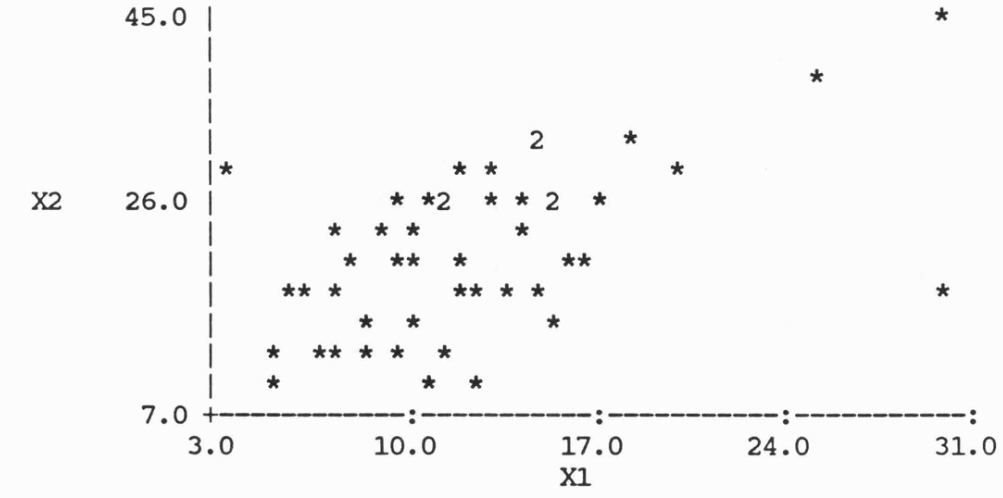
** Normal Plot of Residuals **



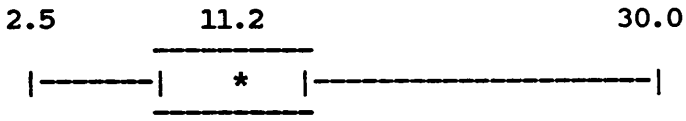
** Index Plot of Residuals **



** Scatter Plot (X1 vs X2) **



** Box Plot for variable X1 **



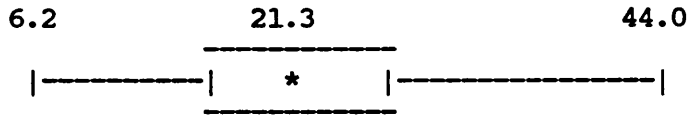
** Quartiles for variable X1 **

Sample size	50	
Median	11.18	
Quartiles	8.49	14.76
IQ range	6.27	
Extremes	2.50	30.00

** Outliers **

CASE 16, CASE 28, CASE 48,

** Box Plot for variable X2 **



** Quartiles for variable X2 **

Sample size	50	
Median	21.29	
Quartiles	16.83	27.42
IQ range	10.60	
Extremes	6.23	44.00

** Outliers **

CASE 16,

** Outliers - Classical Approach (Mahalanobis Distance) **

Case #	d^2	Case #	d^2	Case #	d^2	Case #	d^2
1	0.561	16	11.529<-	31	0.235	46	0.210
2	3.018	17	0.405	32	3.313	47	0.766
3	2.406	18	1.759	33	0.470	48	16.540<-
4	1.280	19	0.912	34	1.616	49	0.241
5	1.046	20	0.260	35	0.815	50	0.951
6	1.498	21	0.749	36	1.491		
7	2.175	22	3.650	37	3.304		
8	0.035	23	0.439	38	1.637		
9	1.900	24	1.751	39	8.486<-		
10	0.960	25	2.161	40	0.456		
11	1.244	26	1.429	41	0.783		
12	0.782	27	0.443	42	0.367		
13	0.711	28	6.643<-	43	0.145		

14	1.420	29	1.816	44	0.768
15	1.659	30	1.724	45	1.038

Exp [Max ChiSq] = 6.592 # 'sus' cases : 4

** Discordancy Tests for a single outlier **

T (Range) = 3.4334
b2 (Kurtosis) = 5.4227
Kimber's Z = 0.1873
Dmax = 16.5402 (Case 48)

** Outliers - Hat Matrix Approach **

Case #	h	Case #	h	Case #	h	Case #	h
1	0.031	16	0.112<-	31	0.019	46	0.016
2	0.046	17	0.027	32	0.010	47	0.025
3	0.004	18	0.050	33	0.023	48	0.346<-
4	0.025	19	0.038	34	0.015	49	0.022
5	0.007	20	0.018	35	0.032	50	0.039
6	0.006	21	0.035	36	0.028		
7	0.042	22	0.003	37	0.037		
8	0.016	23	0.028	38	0.049		
9	0.007	24	0.047	39	0.190<-		
10	0.034	25	0.049	40	0.029		
11	0.021	26	0.001	41	0.032		
12	0.024	27	0.011	42	0.026		
13	0.012	28	0.072	43	0.023		
14	0.048	29	0.044	44	0.035		
15	0.051	30	0.045	45	0.040		

Cut-off (2p/n) = 0.080 # 'sus' cases : 3

** SURCON Analysis (MLE) **

	X1	X2
H0 Lambda =	0.34	0.78

	Single Equation Estimates				SURCON Estimates			
Var.	Gamma	S.E.	t-value	Est. Lambda	Gamma	s.e.	t-value	Est. Lambda
1	0.3234	0.3840	0.8421 (0.2019)	0.0166	-0.0014	0.3415	-0.0042 (0.4984)	0.3414
2	0.2702	0.5518	0.4897 (0.3133)	0.5065	0.0000	0.4907	0.0000 (0.5000)	0.7767

[*** Terms in brackets are p-values for t(n-1) ***]

** Joint test statistic for H0: g1 = g2 = ...= gP = 0 **

F-statistic = 0.0000 d.o.f. = 2, 98

p-value: F-distn. = 1.0000 d.o.f. = 2, 98
Chi-square/P distn. = 0.5000 d.o.f. = 2

** Confidence Intervals

VAR	95.0% C.I. LS	95.0% C.I. 2SLS
X1	[-0.3205, 0.9672]	[-0.5740, 0.5711]
	(-0.6273, 0.6604)	(-0.2312, 0.9139)
X2	[-0.6549, 1.1953]	[-0.8227, 0.8227]
	(-0.4186, 1.4316)	(-0.0460, 1.5994)

[*** Square brackets are gammas and round brackest are lambdas ***]

** Lagrange Multiplier Test **

H0: Diagonal covariance matrix

LM-statistic = 13.1990 D.O.F. = 1
p-value for chi-square with 1 D.O.F. = 0.0003

** Total no. of iterations to converge = 12 with Tolerance Factor 10.0**(-4)

** Stalactite Plot (Shortened) **

Iteration VS Observation

```
Itrn Sub-sample          1         2         3         4         5
      size       12345678901234567890123456789012345678901234567890
   1    3(  6.0)   **  *  *      *     *   * *   * *  *          *
   2    4(  8.0)                 *           *                    *
   3    5( 10.0)   ************ ******** ************** * ***  ***
   4    6( 12.0)   ************  ******* ***************** ***  ***
   5    7( 14.0)   ************  **** ** ********************  ***
   .    .    .     .    .    .     .     .     .     .     .     .
   .    .    .     .    .    .     .     .     .     .     .     .
   .    .    .     .    .    .     .     .     .     .     .     .
  44   46( 92.0)                 *           *          *        *
  45   47( 94.0)                 *           *          *        *
  46   48( 96.0)                 *           *          *        *
  47   49( 98.0)                 *                      *        *
  48   50(100.0)                 *                                *

 Stalactite Score   03111132132120142131310322142322223232412220021400
                    12345678901234567890123456789012345678901234567890
                             1         2         3         4         5
```

** Stalactite Analysis **

Itrn	Sub-sample size	Observation Good #(%)	Observation Bad #(%)	Bad:Good Ratio	Total Sq. Distance Obs.	Total Sq. Distance Exp.
1	3(6.0)	38(76.0)	12(24.0)	0.32	4.00	4.00
2	4(8.0)	47(94.0)	3(6.0)	0.06	0.40	6.00
3	5(10.0)	9(18.0)	41(82.0)	4.56	7.48	8.00
4	6(12.0)	9(18.0)	41(82.0)	4.56	10.00	10.00
5	7(14.0)	9(18.0)	41(82.0)	4.56	12.00	12.00
6	8(16.0)	10(20.0)	40(80.0)	4.00	14.00	14.00
7	9(18.0)	11(22.0)	39(78.0)	3.55	16.00	16.00
8	10(20.0)	12(24.0)	38(76.0)	3.17	18.00	18.00
9	11(22.0)	14(28.0)	36(72.0)	2.57	20.00	20.00
10	12(24.0)	15(30.0)	35(70.0)	2.33	22.00	22.00
11	13(26.0)	15(30.0)	35(70.0)	2.33	24.00	24.00
12	14(28.0)	16(32.0)	34(68.0)	2.13	26.00	26.00
13	15(30.0)	19(38.0)	31(62.0)	1.63	28.00	28.00
14	16(32.0)	20(40.0)	30(60.0)	1.50	30.00	30.00
15	17(34.0)	20(40.0)	30(60.0)	1.50	32.00	32.00
16	18(36.0)	20(40.0)	30(60.0)	1.50	34.00	34.00
17	19(38.0)	23(46.0)	27(54.0)	1.17	36.00	36.00
18	20(40.0)	25(50.0)	25(50.0)	1.00	38.00	38.00
19	21(42.0)	27(54.0)	23(46.0)	0.85	40.00	40.00
20	22(44.0)	28(56.0)	22(44.0)	0.79	42.00	42.00
21	23(46.0)	28(56.0)	22(44.0)	0.79	44.00	44.00
22	24(48.0)	29(58.0)	21(42.0)	0.72	44.95	46.00
23	25(50.0)	30(60.0)	20(40.0)	0.67	47.36	48.00
24	26(52.0)	34(68.0)	16(32.0)	0.47	50.00	50.00
25	27(54.0)	35(70.0)	15(30.0)	0.43	51.98	52.00
26	28(56.0)	37(74.0)	13(26.0)	0.35	53.59	54.00
27	29(58.0)	37(74.0)	13(26.0)	0.35	56.00	56.00
28	30(60.0)	39(78.0)	11(22.0)	0.28	57.68	58.00
29	31(62.0)	39(78.0)	11(22.0)	0.28	59.43	60.00
30	32(64.0)	42(84.0)	8(16.0)	0.19	62.00	62.00
31	33(66.0)	42(84.0)	8(16.0)	0.19	64.00	64.00
32	34(68.0)	43(86.0)	7(14.0)	0.16	66.00	66.00
33	35(70.0)	45(90.0)	5(10.0)	0.11	68.00	68.00
34	36(72.0)	46(92.0)	4(8.0)	0.09	70.00	70.00
35	37(74.0)	46(92.0)	4(8.0)	0.09	72.00	72.00
36	38(76.0)	46(92.0)	4(8.0)	0.09	74.00	74.00
37	39(78.0)	46(92.0)	4(8.0)	0.09	76.00	76.00
38	40(80.0)	46(92.0)	4(8.0)	0.09	78.00	78.00
39	41(82.0)	46(92.0)	4(8.0)	0.09	80.00	80.00
40	42(84.0)	46(92.0)	4(8.0)	0.09	82.00	82.00
41	43(86.0)	46(92.0)	4(8.0)	0.09	84.00	84.00
42	44(88.0)	46(92.0)	4(8.0)	0.09	86.00	86.00
43	45(90.0)	46(92.0)	4(8.0)	0.09	87.97	88.00
44	46(92.0)	46(92.0)	4(8.0)	0.09	90.00	90.00
45	47(94.0)	46(92.0)	4(8.0)	0.09	92.00	92.00
46	48(96.0)	46(92.0)	4(8.0)	0.09	94.00	94.00
47	49(98.0)	47(94.0)	3(6.0)	0.06	96.00	96.00
48	50(100.0)	48(96.0)	2(4.0)	0.04	98.00	98.00

```
** Weighted mean vector **

            X1          X2
        13.257      22.308

/*** End Run ***/
15:12:01   19 Mar 92
tSTAT v0.00 (c) LSE, 1992.
```

APPENDIX C

a) Peruvian Data (Example E.8)

Case	X1	X2	Case	X1	X2	Case	X1	X2	Case	X1	X2
1	71.0	1629	11	66.5	1622	21	59.5	1513	31	69.0	1625
2	56.5	1569	12	59.1	1486	22	61.0	1653	32	73.0	1615
3	56.0	1561	13	64.0	1578	23	57.0	1566	33	64.0	1640
4	61.0	1619	14	69.5	1645	24	57.5	1580	34	65.0	1610
5	65.0	1566	15	64.0	1648	25	74.0	1647	35	71.0	1572
6	62.0	1639	16	56.5	1521	26	72.0	1620	36	60.2	1534
7	53.0	1494	17	57.0	1547	27	62.5	1637	37	55.0	1536
8	53.0	1568	18	55.0	1505	28	68.0	1528	38	70.0	1630
9	65.0	1540	19	57.0	1473	29	63.4	1647	39	87.0	1542
10	57.0	1530	20	58.0	1538	30	68.0	1605			

b) Minitab Tree Data (Example E.9)

Case	X1	Y	Case	X1	Y	Case	X1	Y	Case	X1	Y
1	70	10.3	11	79	24.2	21	78	34.5	31	87	77.0
2	65	10.3	12	76	21.0	22	80	31.7			
3	63	10.2	13	76	21.4	23	74	36.3			
4	72	16.4	14	69	21.3	24	72	38.3			
5	81	18.8	15	75	19.1	25	77	42.6			
6	83	19.7	16	74	22.2	26	81	55.4			
7	66	15.6	17	85	33.8	27	82	55.7			
8	75	18.2	18	86	27.4	28	80	58.3			
9	80	22.6	19	71	25.7	29	80	51.5			
10	75	19.9	20	64	24.9	20	80	51.0			

c) Repeat Soil Sample Survey Data (Example E.11)

Case	X1	X2	X3	X4	X5	Case	X1	X2	X3	X4	X5
1	6.3	5.8	31	88	130	31	5.9	5.3	22	80	68
2	6.6	6	40	68	76	32	5.8	5.3	23	78	79
3	6.6	6	32	93	79	33	7.6	7	50	315	370
4	6.3	5.8	40	128	106	34	7.1	6.5	16	177	686
5	6.6	5.9	22	77	61	35	6.5	6	42	207	358
6	6.9	6.3	11	77	45	36	7	6.4	29	147	348
7	5.8	5.1	22	175	91	37	6.2	5.6	8	74	150
8	5.5	5	12	221	106	38	6.2	5.5	33	315	148
9	6.2	5.7	17	77	103	39	6.3	5.6	17	102	125
10	6.2	5.6	18	114	225	40	5.5	4.8	17	105	180
11	6.6	6.1	14	86	275	41	5.6	5	14	171	144
12	6.5	6.1	30	270	245	42	5.9	5.3	22	270	239
13	7	6.5	18	72	180	43	5.8	5.2	15	74	330
14	5.8	5.1	5	136	118	44	6.3	5.9	31	350	574
15	6.5	5.7	17	86	193	45	6.8	6.2	19	136	353
16	6.3	5.7	16	134	158	46	7.2	6.7	21	147	506
17	8	7.4	21	134	109	47	6.9	6.3	18	225	551
18	7	6.3	18	77	61	48	6.2	5.9	27	142	89
19	8.3	7.7	13	102	70	49	5.5	5	14	112	110
20	8	7.5	117	61	70	50	5.5	5	14	112	110
21	5.8	5.1	13	102	165	51	6.3	5.7	16	84	77
22	6.8	6	5	69	214	52	5.8	5.1	14	81	91
23	7.2	6.4	28	82	176	53	6.9	6.2	11	76	73
24	6.8	6	3	56	138	54	6.6	6.1	32	128	46
25	6.2	5.6	10	82	275	55	7.5	6.8	70	481	88
26	6.6	6	10	197	325	56	7.1	6.4	57	334	68
27	6.6	6	12	100	308	57	6.2	5.6	13	74	62
28	5.6	4.9	14	88	224						
29	6.5	5.8	23	76	138						
30	6	5.5	16	187	96						

BIBLIOGRAPHY

Abramowitz, M. & Stegun, I.A., (1972). Handbook of Mathematical Functions, Dover, New York.

Afifi, A.A. & Azen, S.P., (1979). Statistical Analysis. A Computer Oriented Approach. (2nd ed.). Academic Press, Inc.

Ammeraal, L. (1989). Graphics Programming in Turbo C. John Wiley & Sons, Inc.

Anderson, E. (1935). The irises of the Gaspe Peninsula. Bull. Am. Iris Soc., 59, 2–5.

Andrews, D.F. (1972). Plots of high–dimensional data. Biometrics 28, 125–36.

Andrews, D.F. (1971). A note on the selection of data transformations. Biometrika, 58, 249–54.

Andrews, D.F., Gnanadesikan, R., & Warner, J.L. (1971). Transformations of Multivariate Data. Biometrics, 27, 825–40.

Anscombe, F.J. & Tukey, J.W. (1963). The examination and analysis of residuals. Technometrics 5, 141–60.

Ashby, J., (1968). A modification to Paulson's approximation to the variance ratio distribution, The Computer Journal, 11, 209–10.

Atkinson, A.C. (1986). Masking Unmasked. Biometrika, 73, 3, 533–41.

Atkinson, A.C. (1985). Plots, Transformations and Regression. Oxford University Press.

Atkinson, A.C. & Lawrance, A.J. (1989). A comparison of asymptotically equivalent test statistics for regression transformation., Biometrika, 76, 223–9.

Atkinson, A.C. & Mulira, H–M. (1992). The Stalactite Plot for the Detection of Multivariate Outliers. Submitted for publication.

Barnett, V. & Lewis, T. (1978). Outliers in Statistical Data. Wiley, New York.

Barnett, V. (1976). The ordering of multivariate data. J.R.Statist. Soc. A., 138, 318–44.

Box, G.E.P. (1980). Sampling and Bayes inference in scientific modeling and robustness (with discussion). Journal of the American Statistical Association, A., 143, 383–430.

Box,G.E.P. & Cox,D.R. (1964). An analysis of transformations. J.R.Stat.Soc, B26, 211–52.

Campbell, N.A. (1980). Robust procedures in multivariate analysis I. Robust covariance estimation. Appl. Stat., 29, 231–37.

Campbell, N.A. (1981). Improved Diagnostic Output from Statistical Packages (with discussion). Proc. of the meeting on Statistical Abilities of Computer Software, 279–92.

Chatterjee, S. & Price, B. (1977). Regression Analysis by Example, Wiley, New York.

Church, B.M & Kershaw, C.D. (1986). Survey work in the Statistics Department. Rothamsted Report for 1986, Part 2, Rothamsted Experimental Station.

Church, B.M. & Skinner, R.J. (1986). The pH and nutrient status of agricultural soils in England and Wales 1969–83. J. agric. Sci., Camb., 107, 21–28.

Cook, R.D. & Hawkins, D.M. (1990). Comment on Rousseeuw and van Zomeren (1990). Journal of the American Statistical Association, 85, 640–44.

Cook, R.D. & Weisberg, S. (1982). Residuals and Influence in Regression. Chapman and Hall, London.

Cooper, B.E. (1984). Comment on Nelder (1984). J.R.Statist.Soc. A., 147, Part 2, 159–60.

Cox, D.R. (1970). The Analysis of Binary Data. Methuen, London.

Cox, D.R. (1972). The analysis of multivariate binary data. Appl. Stat., 21, 113–20.

Cox, D.R. & Hinkley, D.V. (1974). Theoretical Statistics. Chapman & Hall, London.

David, H.A. (1981). Order Statistics. John Wiley & Sons, New York.

D'Agostino, R.B. & Pearson, E.S. (1973). Tests for departure from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$. Biometrika, 60, 613–22. [Correction: Bioemtrika, 61, 647].

D'Agostino, R.B. & Tietjen, G.L. (1971). Simulated probability points of b_2 for small samples. Biometrika, 58, 669–72.

Devlin, S.J., Gnanadesikan, R., & Kettenring, J.R. (1975). Robust estimation and outlier detection with correlation coefficients. Biometrika 62, 531–45.

Dixon, W.J. (1953). Processing data for outliers. Biometrics 9, 74–89.

Draper, N.R. & Cox, D.R. (1969). On distributions and their transformation to normality. J.R. Statist. Soc., B, 31, 472–76.

Draper, N.R. & Smith, H. (1965). Applied Regression Analysis. Wiley, New York.

du Toit, S.H.C., Steyn, A.G.W. & Stumpf, R.H. (1986). Graphical Exploratory Data Analysis. Springer–Verlag New York Inc.

Fellegi, I.P. & Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, 71, 17–35.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. Ann. Eugen. 7, 179–88.

Friedman, H.P. & Rubin, J. (1967). On some invariant criteria for grouping data. Journal of the American Statistical Association, 62, 1159–78.

Geary, R.C. (1930). The frequency distribution of the quotient of two normal variables. Journal of the Royal Statistical Association, 93, 442.

Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. John Wiley & Sons.

Gnanadesikan, R. & Kettenring, J.R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics, 28, 81–124.

Gower, J.C. (1985). The development of statisticl computing at Rothamsted. Rothamsted Report for 1985, Part 2, Rothamsted Experimental Station.

Grubbs, F.E. (1969). Procedures for detecting outlying observations in samples. Technometrics 11, 1–21.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. Journal of the American Statistical Association, 69, 383–93.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986). Robust Statistics. The Approach Based on Influence Functions. John Wiley & Sons.

Hawkins, D.M. (1974). The Detection of Errors in Mutlivariate Data Using Principal Components. Journal of the American Statistical Association, 69, 340–44.

Hawkins, D.M., Bradu, D. & Kass, G.V. (1984). Location of several outliers in multiple–regression data using elemental sets. Technometrics, 26, 197–208.

Healy, M.J.R. (1968). Multivariate normal plotting. Appl.Stat. 17, 157–61.

Hinkley, D. (1977). On Quick Choice of Power Transformations. Appl. Statist. 26, 67–69.

Hughes, J.K. & Michtom, J.I. (1977). A Structured Approach to Programming. Prentice Hall

Hutchinson, R.C. & Just, S.B. (1988). Programming Using the C Language. McGraw Hill.

Johnson,R.A. & Wichern,D.W. (1982). Applied Multivariate Statistical Analysis. Prentice Hall.

Judge,G.G., Griffiths,W.E., Hill,R.C., Lütkepohl, H. & Lee,T–C. (1985). The Theory and Practice of Econometrics, 2nd ed. John Wiley & Sons, Inc.

Kassab, V. (1989). Technical C programming. Prentice Hall.

Kernighan, B.W. & Ritchie, M. (1978). The C Programming Language. Prentice Hall.

Kimber, A.C. (1979). Tests for a single outlier in a gamma sample with unknown shape and scale parameters. Appl. Stat., 28, 243–50.

Knuth, D.E. (1981). The Art of Computer Programming, 2: Seminumerical Algorithms, 2nd ed. Reading, Mass: Addison–Wesley.

Kruskal, W.H. (1960). Some remarks on wild observations. Technometrics 2, 1–3.

Lau, C.L., (1980). Algorithm AS147, A simple series of the incomplete gamma integral, Applied Statistics, 2, 113–14.

Layard, M.W.J. (1974). A Monte Carlo comparison of tests of equality of covariance matrices. Biometrika, 61, 461–65.

Ling, R.F.(1974). Comparison of several algorithms for computing sample means and covariances. Journal of the American Statistical Association, 69, 859–66.

Lorenzen, T.J. (1980). Determining statistical characteristics of a vehicle emissions audit procedure. Technometrics, 22, 483–93.

Malkovitch J.F. & Afifi, A.A. (1973). On tests for multivariate normality. J.Am.Stat.Assoc. 68, 176–9.

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. Biometrika 57, 519–30.

Mardia, K.V. & Kent, J.T. (1991). Rao score tests for goodness of fit and independence. Biometrika, 78, 355–63.

Mardia,K.V., Kent,J.T. & Bibby,J.M. (1979). Multivariate Analysis. Academic Press.

Montgomery, D.C. & Peck, A.E. (1982). Introduction to Linear Regression Analysis. John Wiley & Sons, New York.

Moore, P.G. & Tukey, J.W. (1954). Answer to query 112. Biometrics 10, 562–8.

Nelder, J.A. (1984). Present Position and Potential Developments: Some Personal Views–Statistical Computing. J.R.Statist.Soc. A., 147, Part 2, 151–60.

Paulson, E. (1942). An approximate normalistation of the analysis of variance distribution. Annals of Mathematical Statistics, 13, 233–5.

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1988). Numerical Recipes in C. The Art of Scientific Computing. Cambridge University Press.

Puri, M.L. & Sen, P.K. (1971). Nonparametric Methods in Multivariate Analysis. Wiley, New York.

Rocke, D.M. (1989). Bootstrap Bartlett Adjustment in Seemingly Unrelated Regression. Journal of the American Statistical Association, 84, 598–601.

Rosenblatt, M. (1952). Remarks on a multivariate transformatiom. Ann.Math.Stat. 23, 470–2.

Rousseeuw, P.J. (1984). Least median of squares regression. Journal of the American Statistical Association, 79, 871–80.

Rousseeuw, P.J. & Leroy, A.M. (1987). Robust Regression and Outlier Diagnostics. John Wiley & Sons.

Rousseeuw, P.J. & van Zomeren, B.C. (1990). Unmasking Multivariate Outliers and Leverage Points. Journal of the American Statistical Association, 85, 633–39.

Ryan, T.A.,Jr., Joiner, B.L. & Ryan, B.F. (1976). Minitab: Student Handbook. Duxbury Press: North Scituate, Mass.

Schildt, H. (1988). TURBO C. The Complete Reference Guide. Borland–Osborne/McGraw–Hill.

Schwager, S.J. & Margolin, B.H. (1982). Detection of Multivariate Normal Outliers. The Annals of Statistics, 10, 943–54.

Seber, G.A.F. (1984). Multivariate Observations. John Wiley & Sons.

Tukey, J.W. (1970). Exploratory Data Analysis. Addison–Wesley, Reading, Mass.

Tukey, J.W. (1977). Exploratory Data Analysis. Addison–Wesley, Reading, Mass.

Tukey, J.W. (1957). On the comparative anatomy of transformations. Ann.Math.Stat., 28, 602–32.

Waite, M., Prata, S. & Martin, D. (1987). C Primer Plus. (Revised Edition). Howard W. Sams & Company.

Wallace, D.L. (1959). Bounds on normal approximations to Srudent's and chi–squared distributions, Annals of Mathematical Statistics, 30, 1121–30.

Wilk, M.B. & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. Biometrika, 55, 1–17.

Wilks, S.S. (1963). Multivariate statistical outliers. Sankhyā A25, 407–26.

Wilson, E.B. & Hilferty, M.M. (1931). The distribution of chi–square. Proc. Nat. Acad. Sci., 17, 684–88.

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated. Regressions and Tests for Aggregate Bias. Journal of the American Statistical Association, 57, 348–68.

Zellner, A. (1963). Estimators for Seemingly Unrelated Regression Equations: Some Exact Finite. Sample Results, Journal of the American Statistical Association, 58, 977–92.