Robust Multivariate Analysis

David J. Olive
Southern Illinois University
Department of Mathematics
Mailcode 4408
Carbondale, IL 62901-4408
dolive@siu.edu

February 1, 2013

Contents

Preface

Statistics is, or should be, about scientific investigation and how to do it better ....
Box (1990)

Statistics is the science of extracting useful information from data, and a statistical model is used to provide a useful approximation to some of the important characteristics of the population which generated the data.

A case or observation consists of the random variables measured for one person or thing. For multivariate location and dispersion the ith case is $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})^T$. There are n cases. Outliers are cases that lie far away from the bulk of the data, and they can ruin a classical analysis.

Olive (2013) and this book give a two volume presentation of robust statistics. Olive (2013) emphasized the location model, visualizing regression models, high breakdown regression, highly outlier resistant multivariate location and dispersion estimators such as the FCH estimator, and applications of the FCH estimator for visualizing regression models.

Robust Multivariate Analysis tries to find methods that give good results for multivariate analysis for a large group of underlying distributions and that are useful for detecting certain types of outliers. Plots for detecting outliers and prediction intervals and regions that work for large classes of distributions are also of interest.

This book covers robust multivariate analysis. Topics include applications of the easily computed robust estimators to multivariate analysis and when can multivariate procedures give good results if the data distribution is not multivariate normal.

Many of the most used estimators in statistics are semiparametric. For multivariate location and dispersion (MLD), the classical estimator is the sample mean and sample covariance matrix. Many classical procedures originally meant for the multivariate normal (MVN) distribution are semipara-

metric in that the procedures also perform well on a much larger class of elliptically contoured (EC) distributions.

An important goal of robust multivariate analysis is to produce easily computed semiparametric MLD estimators that perform well when the classical estimators perform well, but are also useful for detecting some important types of outliers.

Two paradigms appear in the robust literature. The "*perfect classification paradigm*" assumes that diagnostics or robust statistics can be used to perfectly classify the data into a "clean" subset and a subset of outliers. Then classical methods are applied to the clean data. These methods tend to be inconsistent, but this paradigm is widely used and can be very useful for a fixed data set that contains outliers.

The "*asymptotic paradigm*" assumes that the data are iid and develops the large sample properties of the estimators. Unfortunately, many robust estimators that have rigorously proven asymptotic theory are impractical to compute. In the robust literature for multivariate location and dispersion, often no distinction is made between the two paradigms: frequently the large sample properties for an impractical estimator are derived, but the examples and software use an inconsistent "perfect classification" procedure. In this text, some practical MLD estimators that have good statistical properties are developed (see Section 4.4), and some effort has been made to state whether the "perfect classification" or "asymptotic" paradigm is being used.

Olive (2013, ch. 10, 11) provides an introduction to robust multivariate analysis. Also see Atkinson, Riani and Cerioli (2004), and Wilcox (2012). Most work on robust multivariate analysis follows the Rousseeuw Yohai paradigm. See Maronna, Martin and Yohai (2006).

What is in the Book?

This book examines robust statistics for multivariate analysis. Robust statistics can be used to improve many of the most used statistical procedures. Often practical robust outlier resistant alternatives backed by large sample theory are also given, and may be used in tandem with the classical method. Emphasis is on the following topics. 1) The practical robust $\sqrt{n}$ consistent multivariate location and dispersion FCH estimator is developed, along with reweighted versions RFCH and RMVN. These estimators are useful for creating robust multivariate procedures such as robust principal components, for outlier detection and for determining whether the data is from a multivariate normal distribution or some other elliptically contoured distribution. 2) Practical asymptotically optimal prediction regions are de-

veloped.

Chapter 1 provides an introduction and some results that will be used later in the text. Chapters 2 and 3 cover multivariate distributions and limit theorems including the multivariate normal distribution, elliptically contoured distributions, and the multivariate central limit theorem. Chapter 4 considers classical and easily computed highly outlier resistant $\sqrt{n}$ consistent robust estimators of multivariate location and dispersion such as the FCH, RFCH and RMVN estimators. Chapter 5 considers DD plots and robust prediction regions. Chapters 6 through 13 consider principal component analysis, canonical correlation analysis, discriminant analysis, Hotelling's T^2 test, MANOVA, factor analysis, multivariate regression and clustering, respectively. Chapter 14 discusses other techniques while Chapter 15 provides information on software and suggests some projects for the students.

The text can be used for supplementary reading for courses in multivariate analysis and pattern recognition. See Duda, Hart and Stork (2000) and Bishop (2006). The text can also be used to present many statistical methods to students running a statistical consulting lab.

Some of the applications in this text include the following.

1) The first practical highly outlier resistant robust estimators of multivariate location and dispersion that are backed by large sample and breakdown theory are given with proofs. Section 4.4 provide the easily computed robust $\sqrt{n}$ consistent highly outlier resistant FCH, RFCH and RMVN estimators of multivariate location and dispersion. Applications are numerous, and *R* software for computing the estimators is provided.

2) Practical asymptotically optimal prediction regions are developed in Section 5.2, and should replace parametric prediction regions, which tend to be far too short when the parametric distribution is misspecified, and also replace bootstrap intervals that take too long to compute. These prediction regions are extended to multivariate regression in Section 12.4.

3) Throughout the book there are goodness of fit and lack of fit plots for examining the model. The main tool is the DD plot, and Section 5.1 shows that the DD plot can be used to detect multivariate outliers and as a diagnostic for whether the data is multivariate normal or from some other elliptically contoured distribution with second moments.

4) Applications for robust and resistant estimators are given. The basic idea is to replace the classical estimator or the inconsistent zero breakdown estimators (such as `cov.mcd`) used in the "robust procedure" with the easily

computed $\sqrt{n}$ consistent robust RFCH and RMVN estimators from Section 4.4. The resistant trimmed views methods for visualizing 1D regression models graphically are discussed in Section 14.3.

The website (www.math.siu.edu/olive/multbk.htm) for this book provides more than 20 data sets for *Arc*, and over 60 *R/Splus* programs in the file *mpack.txt*. The students should save the data and program files on a flash drive. Section 15.2 discusses how to get the data sets and programs into the software, but the following commands will work.

Downloading the book's R/Splus functions *mpack.txt* into *R* or *Splus*:

Download *mpack.txt* onto a flash drive G. Enter *R* and wait for the cursor to appear. Then go to the *File* menu and drag down *Source R Code*. A window should appear. Navigate the *Look in* box until it says *Removable Disk (G:)*. In the *Files of type* box choose *All files(*.*)* and then select *mpack.txt*. The following line should appear in the main *R* window.

```
> source("G:/mpack.txt")
```

If you use *Splus*, the above "source command" will enter the functions into *Splus*. Creating a special workspace for the functions may be useful.

Type *ls()*. Over 60 *R/Splus* functions from *mpack.txt* should appear. In *R*, enter the command *q()*. A window asking "*Save workspace image?*" will appear. Click on *No* to remove the functions from the computer (clicking on *Yes* saves the functions on *R*, but the functions are on your flash drive).

Similarly, to download the text's *R/Splus* data sets, save *mrobdata.txt* on a flash drive G, and use the following command.

```
> source("G:/mrobdata.txt")
```

Background

This course assumes that the student has had considerable exposure to statistics, but is at a much lower level than most texts on robust statistics. Calculus and a course in linear algebra are essential. The level of the text is similar to that of Johnson and Wichern (2007), Mardia, Kent, and Bibby (1979), Press (2005) and Rencher (2002). Anderson (2003) is at a much higher level.

Lower level texts on multivariate analysis include Flury and Riedwyl (1988), Grimm and Yarnold(1995, 2000), Hair, Black, Anderson and Tatham (2005), Kachigan (1991) and Tabachnick and Fidell (2006).

An advanced course in statistical inference, especially one that covered convergence in probability and distribution, is needed for several sections of the text. Casella and Berger (2002), Olive (2012b), Poor (1988) and White (1984) easily meet this requirement.

If the students have had only one calculus based course in statistics (eg Wackerly, Mendenhall and Scheaffer 2008), then skip the proofs of the theorems. Chapter 2, Sections 3.1-3.3, 4.4, and Chapter 5 are important. Then topics from the remaining chapters can be chosen.

Need for the book:

As a book on robust multivariate analysis, this book is an alternative to the Rousseeuw Yohai paradigm and attempts to find practical robust estimators that are backed by theory. As a book on multivariate analysis, this book provides large sample theory for the classical methods, showing that many of the methods are robust to nonnormality and work well on large classes of distributions.

The Rousseeuw Yohai paradigm for high breakdown multivariate robust statistics is to approximate an impractical brand name estimator by computing a fixed number of easily computed trial fits and then use the brand name estimator criterion to select the trial fit to be used in the final robust estimator. The resulting estimator will be called an F-brand name estimator where the F indicates that a fixed number of trial fits was used. For example, generate 500 easily computed estimators of multivariate location and dispersion as trial fits. Then choose the trial fit with the dispersion estimator that has the smallest determinant. Since the minimum covariance determinant (MCD) criterion is used, name the resulting estimator the FMCD estimator These practical estimators are typically not yet backed by large sample or breakdown theory. Most of the literature follows the Rousseeuw Yohai paradigm, using estimators like FMCD, FLTS, FMVE, F-S, FLMS, F-τ, F-Stahel-Donoho, F-Projection, F-MM, FLTA, F-Constrained M, ltsreg, lmsreg, cov.mcd, cov.mve or OGK that are not backed by theory. Maronna, Martin and Yohai (2006, ch. 2, 6) and Hubert, Rousseeuw and Van Aelst (2008) provide references for the above estimators.

The best papers from this paradigm either give large sample theory for impractical brand name estimators that take too long to compute, or give practical outlier resistant methods that could possibly be used as diagnostics but have not yet been shown to be consistent or high breakdown. As a rule of thumb, if $p > 2$ then the brand name estimators take too long to

compute, so researchers who claim to be using a practical implementation of an impractical brand name estimator are actually using a F-brand name estimator.

Some Theory and Conjectures for F-Brand Name Estimators

Some widely used F-brand name estimators are easily shown to be zero breakdown and inconsistent, but it is also easy to derive F-brand name estimators that have good theory. For example, suppose that the only trial fit is the classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ where $\overline{\boldsymbol{x}}$ is the sample mean and $\boldsymbol{S}$ is the sample covariance matrix. Computing the determinant of $\boldsymbol{S}$ does not change the classical estimator, so the resululting FMCD estimator is the classical estimator, which is $\sqrt{n}$ consistent on a large class of distributions. Now suppose there are two trial fits $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ and $(\boldsymbol{0}, \boldsymbol{I}_p)$ where $\boldsymbol{x}$ is a $p \times 1$ vector, $\boldsymbol{0}$ is the zero vector and $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. Since the determinant $det(I_p) = p$, the fit with the smallest determinant will not be the classical estimator if $det(\boldsymbol{S}) > p$. Hence this FMCD estimator is only consistent on a rather small class of distributions. Another FMCD estimator might use 500 trial fits, where each trial fit is the classical estimator applied to a subset of size $\lceil n/2 \rceil$ where n is the sample size and $\lceil 7.7 \rceil = 8$. If the subsets are randomly selected cases, then each trial fit is $\sqrt{n}$ consistent, so the resulting FMCD estimator is $\sqrt{n}$ consistent, but has little outlier resistance. Choosing trial fits so that the resulting estimator can be shown to be both consistent and outlier resistant is a very challenging problem.

Some theory for the F-brand name estimators actually used will be given after some notation. Let $p =$ the number of predictors. The elemental concentration and elemental resampling algorithms use K elemental fits where K is a fixed number that does not depend on the sample size n. To produce an elemental fit, randomly select h cases and compute the classical estimator $(T_i, \boldsymbol{C}_i)$ (or $T_i = \hat{\boldsymbol{\beta}}_i$ for regression) for these cases, where $h = p + 1$ for multivariate location and dispersion (and $h = p$ for multiple linear regression). The elemental resampling algorithm uses one of the K elemental fits as the estimator, while the elemental concentration algorithm refines the K elemental fits using all n cases. See Olive and Hawkins (2010, 2011) for more details.

Breakdown is computed by determining the smallest number of cases d_n that can be replaced by arbitrarily bad contaminated cases in order to make $\|T\|$ (or $\|\hat{\boldsymbol{\beta}}\|$) arbitrarily large or to drive the smallest or largest eigenvalues of the dispersion estimator $\boldsymbol{C}$ to 0 or ∞. High breakdown estimators have $\gamma_n = d_n/n \to 0.5$ and zero breakdown estimators have $\gamma_n \to 0$ as $n \to \infty$.

Note that an estimator can not be consistent for θ unless the number of randomly selected cases goes to ∞, except in degenerate situations. The following theorem shows the widely used elemental estimators are zero breakdown estimators. (If $K_n \to \infty$, then the elemental estimator is zero breakdown if $K_n = o(n)$. A necessary condition for the elemental basic resampling estimator to be consistent is $K_n \to \infty$.)

Theorem P.1: a) The elemental basic resampling algorithm estimators are inconsistent. b) The elemental concentration and elemental basic resampling algorithm estimators are zero breakdown.

Proof: a) Note that you can not get a consistent estimator by using Kh randomly selected cases since the number of cases Kh needs to go to ∞ for consistency except in degenerate situations.

b) Contaminating all Kh cases in the K elemental sets shows that the breakdown value is bounded by $Kh/n \to 0$, so the estimator is zero breakdown. QED

Theorem P.1 shows that the elemental basic resampling PROGRESS estimators of Rousseeuw (1984), Rousseeuw and Leroy (1987) and Rousseeuw and van Zomeren (1990) are zero breakdown and inconsistent. Yohai's two stage estimators, such as MM, need initial consistent high breakdown estimators such as LMS, MCD or MVE, but were implemented with the inconsistent zero breakdown elemental estimators such as `lmsreg`, Fake-LMS, Fake-MCD, MVEE or Fake-MVE. See Hawkins and Olive (2002, p. 157). You can get consistent estimators if $K_n \to \infty$ or $h_n \to \infty$ as $n \to \infty$. You can get high breakdown estimators and avoid singular starts if all $K_n = C(n,h) = O(n^h)$ elemental sets are used, but such an estimator is impractical.

Acknowledgments

Some of the research used in this text was partially supported by NSF grants DMS 0202922 and DMS 0600933. Collaboration with Douglas M. Hawkins was extremely valuable. I am very grateful to the developers of useful mathematical and statistical techniques and to the developers of computer software and hardware. A 1997 preprint of Rousseeuw and Van Driessen (1999) was the starting point for much of my work in multivariate analysis. An earlier version of this text was used in a robust multivariate analysis course in 2012.

Chapter 1

Introduction

1.1 Introduction

Multivariate analysis is a set of statistical techniques used to analyze correlated data containing observations on $p \geq 2$ random variables measured on a set of n cases. Let $\boldsymbol{x} = (x_1, ..., x_p)^T$ where $x_1, ..., x_p$ are p random variables. Usually context will be used to decide whether $\boldsymbol{x}$ is a random vector or the observed random vector. For multivariate location and dispersion the ith case is $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})^T$.

Notation: Typically lower case boldface letters such as $\boldsymbol{x}$ denote column vectors while upper case boldface letters such as $\boldsymbol{S}$ denote matrices with 2 or more columns. An exception may occur for random vectors which are usually denoted by $\boldsymbol{x}$, $\boldsymbol{y}$ or $\boldsymbol{z}$. If context is not enough to determine whether $\boldsymbol{x}$ is a random vector or an observed random vector, then $\boldsymbol{X} = (X_1, ..., X_p)^T$ and $\boldsymbol{Y}$ will be used for the random vectors, and $\boldsymbol{x} = (x_1, ..., x_p)^T$ for observed value of the random vector. This notation is used in Chapter 3 in order to study the conditional distribution of $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$. An upper case letter such as Y will usually be a random variable. A lower case letter such as x_1 will also often be a random variable. An exception to this notation is the generic multivariate location and dispersion estimator $(T, \boldsymbol{C})$ where the location estimator T is a $p \times 1$ vector such as $T = \overline{\boldsymbol{x}}$. $\boldsymbol{C}$ is a $p \times p$ dispersion estimator and conforms to the above notation. Another exception is in Chapter 3 where

Assume that the data $\boldsymbol{x}_i$ has been observed and stored in an $n \times p$ matrix

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix}$$

where the ith row of $\boldsymbol{W}$ is the ith case $\boldsymbol{x}_i^T$ and the jth column $\boldsymbol{v}_j$ of $\boldsymbol{W}$ corresponds to n measurements of the jth random variable x_j for $j = 1, ..., p$.

Often the n rows corresponding to the n cases are assumed to be iid or a random sample from some multivariate distribution. The p columns correspond to n measurements on the p correlated random variables $x_1, ..., x_p$. The n cases are $p \times 1$ vectors while the p columns are $n \times 1$ vectors.

Methods involving one response variable will not be covered in depth in this text. Such models include multiple linear regression, many experimental design models and generalized linear models. Discrete multivariate analysis = categorical data analysis will also not be covered.

Most of the multivariate techniques studied in this book will use estimators of multivariate location and dispersion. Typically the data will be assumed to come from a continuous distribution with a joint probability distribution function (pdf). Multivariate techniques that examine correlations among the p random variables $x_1, ..., x_p$ include principal component analysis, canonical correlation analysis and factor analysis. Multivariate techniques that compare the n cases $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ include discriminant analysis and cluster analysis. *Data reduction* attempts to simplify the multivariate data without losing important information. Since the data matrix $\boldsymbol{W}$ has np terms, *data reduction* is an important technique. Prediction and hypothesis testing are also important techniques. Hypothesis testing is important for multivariate regression, Hotelling's T^2 test, and MANOVA.

Robust multivariate analysis consists of i) techniques that are robust to nonnormality or ii) techniques that are robust to outliers. Techniques that are robust to outliers tend to have some robustness to nonnormality. The classical covariance matrix $\boldsymbol{S}$ is very robust to nonnormality, but is not robust to outliers. Large sample theory is useful for both robust techniques. See Section 3.4.

1.2 Things That Can Go Wrong with a Multivariate Analysis

In multivariate analysis, there is often a training data set used to predict or classify data in a future data set. Many things can go wrong. For classification and prediction, it is usually assumed that the data in the training set is from the same distribution as the data in the future set. Following Hand (2006), this crucial assumption is often not justified.

Population drift is a common reason why the above assumption, which assumes that the various distributions involved do not change over time, is violated. Population drift occurs when the population distribution does change over time. As an example, perhaps pot shards are classified after being sent to a lab for analysis. It is often the case that even if the shards are sent to the same lab twice, the two sets of lab measurements differ greatly. As another example, suppose there are several variables being used to produce greater yield of a crop or a chemical. If one journal paper out of 50 (the training set) finds a set of variables and variable levels that successfully increases yield, then the next 25 papers (the future set) are more likely to use variables and variable levels similar to the one successful paper than variables and variable levels of the 49 papers that did not succeed. Hand (2006) notes that classification rules used to predict whether applicants are likely to default on loans are updated every few months in the banking and credit scoring industries.

A second thing that can go wrong is that the training or future data set is distorted away from the population distribution. This could occur if outliers are present or if one of the data sets is not a random sample from the population. For example, the training data set could be drawn from three hospitals, and the future data set could be drawn from two more hospitals. These two data sets may not represent random samples from the same population of hospitals.

Often problems specific to the multivariate method can occur. Often simpler techniques can outperform sophisticated multivariate techniques because the user of the multivariate method does not have the expertise to get the most out of the sophisticated technique. For supervised classification, Hand (2006) notes that there can be error in class labels, arbitrariness in class definitions and data sets where different optimization criteria lead to very different classification rules. Hand (2006) suggests that simple rules such as linear discriminant analysis may perform almost as well or better

than sophisticated classification rules because of all of the possible problems. See Chapter 8.

1.3 Some Matrix Optimization Results

The following results will be useful throughout the text. Let $\boldsymbol{A} > 0$ denote that $\boldsymbol{A}$ is a positive definite matrix.

Theorem 1.1. Let $\boldsymbol{B} > 0$ be a $p \times p$ symmetric matrix with eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p > 0$ and the orthonormal eigenvectors satisfy $\boldsymbol{e}_i^T \boldsymbol{e}_i = 1$ while $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ for $i \neq j$. Let $\boldsymbol{d}$ be a given $p \times 1$ vector and let $\boldsymbol{a}$ be an arbitrary nonzero $p \times 1$ vector. See Johnson and Wichern (1988, p. 64-65, 184).

a) $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}^T \boldsymbol{d}\boldsymbol{d}^T \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a}} = \boldsymbol{d}^T \boldsymbol{B}^{-1} \boldsymbol{d}$ where the max is attained for $\boldsymbol{a} = c\boldsymbol{B}^{-1}\boldsymbol{d}$ for any constant $c \neq 0$. Note that the numerator $= (\boldsymbol{a}^T \boldsymbol{d})^2$.

b) $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{a}} = \max_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a} = \lambda_1$ where the max is attained for $\boldsymbol{a} = \boldsymbol{e}_1$.

c) $\min_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{a}} = \min_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a} = \lambda_p$ where the min is attained for $\boldsymbol{a} = \boldsymbol{e}_p$.

d) $\max_{\boldsymbol{a} \perp \boldsymbol{e}_1, ..., \boldsymbol{e}_k} \frac{\boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{a}} = \max_{\|\boldsymbol{a}\|=1, \boldsymbol{a} \perp \boldsymbol{e}_1, ..., \boldsymbol{e}_k} \boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a} = \lambda_{k+1}$ where the max is attained for $\boldsymbol{a} = \boldsymbol{e}_{k+1}$ for $k = 1, 2, ..., p-1$.

e) Let $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ be the observed sample mean and sample covariance matrix where $\boldsymbol{S} > 0$. Then $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}^T (\overline{\boldsymbol{x}} - \boldsymbol{\mu})(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{S}\boldsymbol{a}} = n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) = T^2$ where the max is attained for $\boldsymbol{a} = c\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu})$ for constant $c \neq 0$.

f) Let $\boldsymbol{A}$ be a $p \times p$ symmetric matrix. Then $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}^T \boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a}} = \lambda_1(\boldsymbol{B}^{-1}\boldsymbol{A})$, the largest eigenvalue of $\boldsymbol{B}^{-1}\boldsymbol{A}$.

1.4 The Location Model

The *location model*

$$Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{1.1}$$

is a special case of the multivariate location and dispersion model with $p = 1$. The location model is often summarized by obtaining point estimates and confidence intervals for a location parameter and a scale parameter. Assume that there is a sample $Y_1, \ldots, Y_n$ of size n where the Y_i are iid from a distribution with median MED(Y), mean $E(Y)$, and variance $V(Y)$ if they exist. Also assume that the Y_i have a cumulative distribution function (cdf) F that is known up to a few parameters. For example, Y_i could be normal, exponential, or double exponential. The location parameter μ is often the population mean or median while the scale parameter is often the population standard deviation $\sqrt{V(Y)}$. The ith *case* is Y_i.

Point estimation is one of the oldest problems in statistics and four of the most important statistics for the location model are the sample mean, median, variance, and the median absolute deviation (mad). Let $Y_1, \ldots, Y_n$ be the random sample; ie, assume that $Y_1, ..., Y_n$ are iid.

Definition 1.1. The *sample mean*

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}. \tag{1.2}$$

The sample mean is a measure of location and estimates the population mean (expected value) $\mu = E(Y)$. The sample mean is often described as the "balance point" of the data. The following alternative description is also useful. For any value m consider the data values $Y_i \leq m$, and the values $Y_i > m$. Suppose that there are n rods where rod i has length $|r_i(m)| = |Y_i - m|$ where $r_i(m)$ is the ith residual of m. Since $\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0$, $\overline{Y}$ is the value of m such that the sum of the lengths of the rods corresponding to $Y_i \leq m$ is equal to the sum of the lengths of the rods corresponding to $Y_i > m$. If the rods have the same diameter, then the weight of a rod is proportional to its length, and the weight of the rods corresponding to the $Y_i \leq \overline{Y}$ is equal to the weight of the rods corresponding to $Y_i > \overline{Y}$. The sample mean is drawn towards an outlier since the absolute residual corresponding to a single outlier is large.

If the data $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the ith order statistic and the $Y_{(i)}$'s are called the *order statistics*. Using this notation, the median

$$\text{MED}_c(n) = Y_{((n+1)/2)} \text{ if n is odd,}$$

and

$$\text{MED}_c(n) = (1-c)Y_{(n/2)} + cY_{((n/2)+1)} \text{ if n is even}$$

for $c \in [0,1]$. Note that since a statistic is a function, c needs to be fixed. The *low median* corresponds to $c = 0$, and the *high median* corresponds to $c = 1$. The choice of $c = 0.5$ will yield the sample median. For example, if the data $Y_1 = 1, Y_2 = 4, Y_3 = 2, Y_4 = 5$, and $Y_5 = 3$, then $\overline{Y} = 3$, $Y_{(i)} = i$ for $i = 1, ..., 5$ and $\text{MED}_c(n) = 3$ where the sample size $n = 5$.

Definition 1.2. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \text{ if n is odd,} \tag{1.3}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \text{ if n is even.}$$

The notation $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ will also be used.

Definition 1.3. The *sample variance*

$$S_n^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^n Y_i^2 - n(\overline{Y})^2}{n-1}, \tag{1.4}$$

and the *sample standard deviation* $S_n = \sqrt{S_n^2}$.

The sample median need not be unique and is a measure of location while the sample standard deviation is a measure of scale. In terms of the "rod analogy," the median is a value m such that at least half of the rods are to the left of m and at least half of the rods are to the right of m. Hence the number of rods to the left and right of m rather than the lengths of the rods determine the sample median. The sample standard deviation is vulnerable to outliers and is a measure of the average value of the rod lengths $|r_i(\overline{Y})|$. The sample mad, defined below, is a measure of the median value of the rod lengths $|r_i(\text{MED}(n))|$.

Definition 1.4. The *sample median absolute deviation* is

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n). \tag{1.5}$$

Since $\text{MAD}(n)$ is the median of n distances, at least half of the observations are within a distance $\text{MAD}(n)$ of $\text{MED}(n)$ and at least half of the observations are a distance of $\text{MAD}(n)$ or more away from $\text{MED}(n)$.

Example 1.1. Let the data be $1, 2, 3, 4, 5, 6, 7, 8, 9$. Then $\text{MED}(n) = 5$ and $\text{MAD}(n) = 2 = \text{MED}\{0, 1, 1, 2, 2, 3, 3, 4, 4\}$.

Since these estimators are nonparametric estimators of the corresponding population quantities, they are useful for a very wide range of distributions.

1.5 Mixture Distributions

Mixture distributions are often used as outlier models, and certain mixtures of elliptically contoured distributions have an elliptically contoured distribution. The following two definitions and proposition are useful for finding the mean and variance of a mixture distribution. Parts a) and b) of Proposition 1.2 below show that the definition of expectation given in Definition 1.6 is the same as the usual definition for expectation if Y is a discrete or continuous random variable.

Definition 1.5. The distribution of a random variable Y is a *mixture distribution* if the cdf of Y has the form

$$F_Y(y) = \sum_{i=1}^{k} \alpha_i F_{W_i}(y) \tag{1.6}$$

where $0 < \alpha_i < 1$, $\sum_{i=1}^{k} \alpha_i = 1$, $k \geq 2$, and $F_{W_i}(y)$ is the cdf of a continuous or discrete random variable W_i, $i = 1, ..., k$.

Definition 1.6. Let Y be a random variable with cdf $F(y)$. Let h be a function such that the expected value $Eh(Y) = E[h(Y)]$ exists. Then

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y) dF(y). \tag{1.7}$$

Proposition 1.2. a) If Y is a discrete random variable that has a pmf $f(y)$ with support $\mathcal{Y}$, then

$$Eh(Y) = \int_{-\infty}^{\infty} h(y) dF(y) = \sum_{y \in \mathcal{Y}} h(y) f(y).$$

b) If Y is a continuous random variable that has a pdf $f(y)$, then

$$Eh(Y) = \int_{-\infty}^{\infty} h(y) dF(y) = \int_{-\infty}^{\infty} h(y) f(y) dy.$$

c) If Y is a random variable that has a mixture distribution with cdf $F_Y(y) = \sum_{i=1}^{k} \alpha_i F_{W_i}(y)$, then

$$Eh(Y) = \int_{-\infty}^{\infty} h(y) dF(y) = \sum_{i=1}^{k} \alpha_i E_{W_i}[h(W_i)]$$

where $E_{W_i}[h(W_i)] = \int_{-\infty}^{\infty} h(y) dF_{W_i}(y)$.

Example 1.2. Proposition 1.2c implies that the pmf or pdf of W_i is used to compute $E_{W_i}[h(W_i)]$. As an example, suppose the cdf of Y is $F(y) = (1-\epsilon)\Phi(y) + \epsilon\Phi(y/k)$ where $0 < \epsilon < 1$ and $\Phi(y)$ is the cdf of $W_1 \sim N(0,1)$. Then $\Phi(y/k)$ is the cdf of $W_2 \sim N(0, k^2)$. To find EY, use $h(y) = y$. Then

$$EY = (1-\epsilon)EW_1 + \epsilon EW_2 = (1-\epsilon)0 + \epsilon 0 = 0.$$

To find EY^2, use $h(y) = y^2$. Then

$$EY^2 = (1-\epsilon)EW_1^2 + \epsilon EW_2^2 = (1-\epsilon)1 + \epsilon k^2 = 1 - \epsilon + \epsilon k^2.$$

Thus $\text{VAR}(Y) = E[Y^2] - (E[Y])^2 = 1 - \epsilon + \epsilon k^2$. If $\epsilon = 0.1$ and $k = 10$, then $EY = 0$, and $\text{VAR}(Y) = 10.9$.

To generate a random variable Y with the above mixture distribution, generate a uniform (0,1) random variable U which is independent of the W_i. If $U \leq 1-\epsilon$, then generate W_1 and take $Y = W_1$. If $U > 1-\epsilon$, then generate W_2 and take $Y = W_2$. Note that the cdf of Y is $F_Y(y) = (1-\epsilon)F_{W_1}(y) + \epsilon F_{W_2}(y)$.

Remark 1.1. Warning: Mixture distributions and linear combinations of random variables are very different quantities. As an example, let

$$W = (1-\epsilon)W_1 + \epsilon W_2$$

where W_1 and W_2 are independent random variables and $0 < \epsilon < 1$. Then the random variable W is a linear combination of W_1 and W_2, and W can be generated by generating two independent random variables W_1 and W_2. Then take $W = (1-\epsilon)W_1 + \epsilon W_2$.

If W_1 and W_2 are as in the previous example then the random variable W is a linear combination that has a normal distribution with mean

$$EW = (1-\epsilon)EW_1 + \epsilon EW_2 = 0$$

and variance

$$\text{VAR}(W) = (1-\epsilon)^2\text{VAR}(W_1) + \epsilon^2\text{VAR}(W_2) = (1-\epsilon)^2 + \epsilon^2 k^2 < \text{VAR}(Y)$$

where Y is given in the example above. Moreover, W has a unimodal normal distribution while Y does not follow a normal distribution. In fact, if $X_1 \sim N(0,1)$, $X_2 \sim N(10,1)$, and X_1 and X_2 are independent, then $(X_1+X_2)/2 \sim N(5, 0.5)$; however, if Y has a mixture distribution with cdf

$$F_Y(y) = 0.5F_{X_1}(y) + 0.5F_{X_2}(y) = 0.5\Phi(y) + 0.5\Phi(y-10),$$

then the pdf of Y is bimodal.

1.6 Summary

1) Given a small data set, find $\overline{Y}$, S, MED(n) and MAD(n). Recall that $\overline{Y} = \frac{\sum_{i=1}^n Y_i}{n}$ and the *sample variance*

$$\text{VAR}(n) = S^2 = S_n^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^n Y_i^2 - n(\overline{Y})^2}{n-1},$$

and the *sample standard deviation* (SD) $S = S_n = \sqrt{S_n^2}$.

If the data $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then the $Y_{(i)}$'s are called the *order statistics.* The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \text{ if n is odd,}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \text{ if n is even.}$$

The notation MED(n) = MED($Y_1, ..., Y_n$) will also be used. To find the sample median, sort the data from smallest to largest and find the middle value or values.

The *sample median absolute deviation*

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n).$$

To find MAD(n), find $D_i = |Y_i - \text{MED}(n)|$, then find the sample median of the D_i by ordering them from smallest to largest and finding the middle value or values.

1.7 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

1.1. Consider the data set 6, 3, 8, 5, and 2. Show work.

a) Find the sample mean $\overline{Y}$.

b) Find the standard deviation S

c) Find the sample median MED(n).

d) Find the sample median absolute deviation MAD(n).

1.2*. The Cushny and Peebles data set (see Staudte and Sheather 1990, p. 97) is listed below.

```
1.2  2.4  1.3  1.3  0.0  1.0  1.8  0.8  4.6  1.4
```

a) Find the sample mean $\overline{Y}$.

b) Find the sample standard deviation S.

c) Find the sample median MED(n).

d) Find the sample median absolute deviation MAD(n).

e) Plot the data. Are any observations unusually large or unusually small?

Chapter 2

Multivariate Distributions

2.1 Introduction

Definition 2.1. An important *multivariate location and dispersion model* is a joint distribution with joint probability density function (pdf)

$$f(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

for a $p \times 1$ random vector $\boldsymbol{x}$ that is completely specified by a $p \times 1$ population *location* vector $\boldsymbol{\mu}$ and a $p \times p$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. Thus $P(\boldsymbol{x} \in A) = \int_A f(\boldsymbol{z})d\boldsymbol{z}$ for suitable sets A.

Notation: Usually a vector $\boldsymbol{x}$ will be column vector, and a row vector $\boldsymbol{x}^T$ will be the transpose of the vector $\boldsymbol{x}$. However,

$$\int_A f(\boldsymbol{z})d\boldsymbol{z} = \int_A f(z_1, ..., z_p)dz_1 \cdots dz_p.$$

The notation $f(z_1, ..., z_p)$ will be used to write out the components z_i of a joint pdf $f(\boldsymbol{z})$ although in the formula for the pdf, eg $f(\boldsymbol{z}) = c \exp(\boldsymbol{z}^T\boldsymbol{z})$, $\boldsymbol{z}$ is a column vector.

Definition 2.2. A $p \times 1$ *random vector* $\boldsymbol{x} = (x_1, ..., x_p)^T = (X_1, ..., X_p)^T$ where $X_1, ..., X_p$ are p random variables. A *case* or *observation* consists of the p random variables measured for one person or thing. For multivariate location and dispersion the ith case is $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})^T$. There are n cases, and context will be used to determine whether $\boldsymbol{x}$ is the random vector or the

observed value of the random vector. *Outliers* are cases that lie far away from the bulk of the data, and they can ruin a classical analysis.

Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are n iid $p \times 1$ random vectors and that the joint pdf of $\boldsymbol{x}_i$ is $f(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Also assume that the data $\boldsymbol{x}_i$ has been observed and stored in an $n \times p$ matrix

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix}$$

where the ith row of $\boldsymbol{W}$ is the ith case $\boldsymbol{x}_i^T$ and the jth column $\boldsymbol{v}_j$ of $\boldsymbol{W}$ corresponds to n measurements of the jth random variable X_j for $j = 1, ..., p$. Hence the n rows of the data matrix $\boldsymbol{W}$ correspond to the n cases, while the p columns correspond to measurements on the p random variables $X_1, ..., X_p$. For example, the data may consist of n visitors to a hospital where the $p = 2$ variables *height* and *weight* of each individual were measured.

Notation: *In the theoretical sections of this text, $\boldsymbol{x}_i$ will sometimes be a random vector and sometimes the observed data.* Johnson and Wichern (1988, p. 7, 53) uses $\boldsymbol{X}$ to denote the $n \times p$ data matrix and a $n \times 1$ random vector, relying on the context to indicate whether $\boldsymbol{X}$ is a random vector or data matrix. Software tends to use different notation. For example, *R/Splus* will use commands such as

```
var(x)
```

to compute the sample covariance matrix of the data. Hence x corresponds to $\boldsymbol{W}$, x[,1] is the first column of x and $x[4,]$ is the 4th row of x.

2.2 The Sample Mean and Sample Covariance Matrix

Definition 2.3. If the second moments exist, the *population mean* of a random $p \times 1$ vector $\boldsymbol{x} = (X_1, ..., X_p)^T$ is

$$E(\boldsymbol{x}) = \boldsymbol{\mu} = (E(X_1), ..., E(X_p))^T,$$

and the $p \times p$ *population covariance matrix*

$$\text{Cov}(\boldsymbol{x}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T] = E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{x}^T] =$$

$$E(\boldsymbol{x}\boldsymbol{x}^T) - E(\boldsymbol{x})[E(\boldsymbol{x})]^T = ((\sigma_{i,j})) = \boldsymbol{\Sigma}_{\boldsymbol{x}}.$$

That is, the ij entry of $\text{Cov}(\boldsymbol{x})$ is $\text{Cov}(X_i, X_j) = \sigma_{i,j} = E([X_i - E(X_i)][X_j - E(X_j)])$. The $p \times p$ population correlation matrix $\text{Cor}(\boldsymbol{x}) = \boldsymbol{\rho} = ((\rho_{ij}))$. That is, the ij entry of $\text{Cor}(\boldsymbol{x})$ is $\text{Cor}(X_i, X_j) =$

$$\frac{\sigma_{i,j}}{\sigma_i \sigma_j} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

Let the $p \times p$ population standard deviation matrix

$$\boldsymbol{\Delta} = \text{diag}(\sqrt{\sigma_{11}}, ..., \sqrt{\sigma_{\text{pp}}}).$$

Then

$$\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Delta}\boldsymbol{\rho}\boldsymbol{\Delta}, \tag{2.1}$$

and

$$\boldsymbol{\rho} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Delta}^{-1}. \tag{2.2}$$

Let the population standardized random variables

$$Z_i = \frac{X_i - E(X_i)}{\sqrt{\sigma_{ii}}}$$

for $i = 1, ..., p$. Then $\text{Cor}(\boldsymbol{X}) = \boldsymbol{\rho}$ is the covariance matrix of $\boldsymbol{z} = (Z_1, ..., Z_p)^T$.

Definition 2.4. Let random vectors $\boldsymbol{x}$ be $p \times 1$ and $\boldsymbol{y}$ be $q \times 1$. The *population covariance matrix* of $\boldsymbol{x}$ with $\boldsymbol{y}$ is the $p \times q$ matrix

$$\text{Cov}(\boldsymbol{x}, \boldsymbol{y}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{y} - E(\boldsymbol{y}))^T] =$$

$$E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{y}^T] = E(\boldsymbol{x}\boldsymbol{y}^T) - E(\boldsymbol{x})[E(\boldsymbol{y})]^T = \boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{y}}$$

assuming the expected values exist. Note that the $q \times p$ matrix $\text{Cov}(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{y},\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{y}}^T$, and $\text{Cov}(\boldsymbol{x}) = \text{Cov}(\boldsymbol{x}, \boldsymbol{x})$.

A $p \times 1$ random vector $\boldsymbol{x}$ has an *elliptically contoured distribution,* if $\boldsymbol{x}$ has pdf

$$f(\boldsymbol{z}) = k_p|\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})], \tag{2.3}$$

and we say $\boldsymbol{x}$ has an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution. See Chapter 3. If second moments exist for this distribution, then

$$E(\boldsymbol{x}) = \boldsymbol{\mu} \text{ and } \text{Cov}(\boldsymbol{x}) = \text{c}_\text{x}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\boldsymbol{x}}$$

for some constant $c_x > 0$ where the ij entry is $\text{Cov}(X_i, X_j) = \sigma_{i,j}$.

Definition 2.5. Let $x_{1j}, ..., x_{nj}$ be measurements on the ith random variable X_j corresponding to the jth column of the data matrix $\boldsymbol{W}$. The jth *sample mean* is $\overline{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj}$. The *sample covariance* S_{ij} estimates $\text{Cov}(X_i, X_j) = \sigma_{ij}$, and

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j).$$

$S_{ii} = S_i^2$ is the *sample variance* that estimates the population variance $\sigma_{ii} = \sigma_i^2$. The *sample correlation* r_{ij} estimates the population correlation $\text{Cor}(X_i, X_j) = \rho_{ij}$, and

$$r_{ij} = \frac{S_{ij}}{S_i S_j} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}}.$$

Definition 2.6. The **sample mean** or *sample mean vector*

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i = (\overline{x}_1, ..., \overline{x}_p)^T = \frac{1}{n}\boldsymbol{W}^T\boldsymbol{1}$$

where $\boldsymbol{1}$ is the $n \times 1$ vector of ones. The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = ((S_{ij})).$$

That is, the ij entry of $\boldsymbol{S}$ is the sample covariance S_{ij}. The *classical estimator of multivariate location and dispersion* is $(\overline{\boldsymbol{x}}, \boldsymbol{S})$.

It can be shown that $(n-1)\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\ \overline{\boldsymbol{x}}^T =$

$$\boldsymbol{W}^T\boldsymbol{W} - \frac{1}{n}\boldsymbol{W}^T\mathbf{1}\mathbf{1}^T\boldsymbol{W}.$$

Hence if the *centering matrix* $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, then $(n-1)\boldsymbol{S} = \boldsymbol{W}^T\boldsymbol{H}\boldsymbol{W}$.

Definition 2.7. The **sample correlation matrix**

$$\boldsymbol{R} = ((r_{ij})).$$

That is, the ij entry of $\boldsymbol{R}$ is the sample correlation r_{ij}.

Let the standardized random variables

$$Z_i = \frac{x_i - \overline{x}_i}{\sqrt{S_{ii}}}$$

for $i = 1, ..., p$. Then $\boldsymbol{R}$ is the sample covariance matrix of $\boldsymbol{z} = (Z_1, ..., Z_p)^T$.

The population and sample correlation are measures of the strength of a **linear relationship** between two random variables, satisfying $-1 \leq \rho_{ij} \leq 1$ and $-1 \leq r_{ij} \leq 1$. Let the $p \times p$ sample standard deviation matrix

$$\boldsymbol{D} = \text{diag}(\sqrt{\text{S}_{11}}, ..., \sqrt{\text{S}_{\text{pp}}}).$$

Then

$$\boldsymbol{S} = \boldsymbol{D}\boldsymbol{R}\boldsymbol{D}, \tag{2.4}$$

and

$$\boldsymbol{R} = \boldsymbol{D}^{-1}\boldsymbol{S}\boldsymbol{D}^{-1}. \tag{2.5}$$

2.3 Distances

Definition 2.8. Let $\boldsymbol{A}$ be a positive definite symmetric matrix. Then the *Mahalanobis distance* of $\boldsymbol{x}$ from the vector $\boldsymbol{\mu}$ is

$$D_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{A}) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{A}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}.$$

Typically $\boldsymbol{A}$ is a dispersion matrix. The *population squared Mahalanobis distance*

$$D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}). \tag{2.6}$$

Estimators of multivariate location and dispersion $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ are of interest. The *sample squared Mahalanobis distance*

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}). \tag{2.7}$$

Notation: Recall that a square symmetric $p \times p$ matrix $\boldsymbol{A}$ has an *eigenvalue* λ with corresponding *eigenvector* $\boldsymbol{x} \neq \boldsymbol{0}$ if

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}. \tag{2.8}$$

The eigenvalues of $\boldsymbol{A}$ are real since $\boldsymbol{A}$ is symmetric. Note that if constant $c \neq 0$ and $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$, then $c\,\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$. Let $\boldsymbol{e}$ be an eigenvector of $\boldsymbol{A}$ with unit length $\|\boldsymbol{e}\| = \sqrt{\boldsymbol{e}^T\boldsymbol{e}} = 1$. Then $\boldsymbol{e}$ and $-\boldsymbol{e}$ are eigenvectors with unit length, and $\boldsymbol{A}$ has p eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$. Since $\boldsymbol{A}$ is symmetric, the eigenvectors are chosen such that the $\boldsymbol{e}_i$ are orthogonal: $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ for $i \neq j$. The symmetric matrix $\boldsymbol{A}$ is positive definite iff all of its eigenvalues are positive, and positive semidefinite iff all of its eigenvalues are nonnegative. If $\boldsymbol{A}$ is positive semidefinite, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. If $\boldsymbol{A}$ is positive definite, then $\lambda_p > 0$.

Theorem 2.1. Let $\boldsymbol{A}$ be a $p \times p$ symmetric matrix with eigenvector eigenvalue pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\boldsymbol{e}_i^T \boldsymbol{e}_i = 1$ and $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ for $i = 1, ..., p$. Then the *spectral decomposition* of $\boldsymbol{A}$ is

$$\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^T + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^T.$$

Using the same notation as Johnson and Wichern (1988, p. 50-51), let $\boldsymbol{P} = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_p]$ be the $p \times p$ orthogonal matrix with ith column $\boldsymbol{e}_i$. Then $\boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$. Let $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_p)$ and let $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_p})$. If $\boldsymbol{A}$ be is positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$, then $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T$ and

$$\boldsymbol{A}^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^T = \sum_{i=1}^{p} \frac{1}{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i^T.$$

Theorem 2.2. Let $\boldsymbol{A}$ be a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$. The *square root matrix* $\boldsymbol{A}^{1/2} = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}\boldsymbol{P}^T$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

Points $\boldsymbol{x}$ with the same distance $D_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{A}^{-1})$ lie on a hyperellipsoid. Let matrix $\boldsymbol{A}$ have determinant $\det(\boldsymbol{A}) = |\boldsymbol{A}|$. Recall that

$$|\boldsymbol{A}^{-1}| = \frac{1}{|\boldsymbol{A}|} = |\boldsymbol{A}|^{-1}.$$

See Johnson and Wichern (1988, p. 49-50, 102-103) for the following theorem.

Theorem 2.3. Let $h > 0$ be a constant, and let $\boldsymbol{A}$ be a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Then $\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{\mu}) \leq h^2\} =$

$$\{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{A}^{-1}) \leq h^2\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{A}^{-1}) \leq h\}$$

defines a hyperellipsoid centered at $\boldsymbol{\mu}$ with volume

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} |\boldsymbol{A}|^{-1/2} h^p.$$

Let $\boldsymbol{\mu} = \boldsymbol{0}$. Then the axes of the hyperellipsoid are given by the eigenvectors $\boldsymbol{e}_i$ of $\boldsymbol{A}$ with half length in the direction of $\boldsymbol{e}_i$ equal to $h/\sqrt{\lambda_i}$ for $i = 1, ..., p$.

In the following theorem, the shape of the hyperellipsoid is determined by the eigenvectors and eigenvalues of $\boldsymbol{\Sigma}$: $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Note $\boldsymbol{\Sigma}^{-1}$ has the same eigenvectors as $\boldsymbol{\Sigma}$ but eigenvalues equal to $1/\lambda_i$ since $\boldsymbol{\Sigma e} = \lambda \boldsymbol{e}$ iff $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma e} = \boldsymbol{e} = \boldsymbol{\Sigma}^{-1}\lambda \boldsymbol{e}$. Then divide both sides by $\lambda > 0$ since $\boldsymbol{\Sigma} > 0$ and is symmetric. Let $\boldsymbol{w} = \boldsymbol{x} - \boldsymbol{\mu}$. Then points at squared distance $\boldsymbol{w}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors of $\boldsymbol{\Sigma}$ where the half length in the direction of $\boldsymbol{e}_i$ is $h\sqrt{\lambda_i}$. Taking $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$ or $\boldsymbol{A} = \boldsymbol{S}^{-1}$ in Theorem 2.3 gives the volume results for the following two theorems.

Theorem 2.4. Let $\boldsymbol{\Sigma}$ be a positive definite symmetric matrix, eg a dispersion matrix. Let $U = D^2_{\boldsymbol{x}} = D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The hyperellipsoid $\{\boldsymbol{x} | D^2_{\boldsymbol{x}} \leq h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leq h^2\}$, where $h^2 = u_{1-\alpha}$ and $P(U \leq u_{1-\alpha}) = 1 - \alpha$, is the highest density region covering $1 - \alpha$ of the mass for an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution (see Definition 3.3) if g is continuous and decreasing. Let $\boldsymbol{w} = \boldsymbol{x} - \boldsymbol{\mu}$. Then points at squared distance $\boldsymbol{w}^T \boldsymbol{S}^{-1} \boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes

are given by the eigenvectors $\boldsymbol{e}_i$ where the half length in the direction of $\boldsymbol{e}_i$ is $h\sqrt{\lambda_i}$. The volume of the hyperellipsoid is

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)}|\boldsymbol{\Sigma}|^{1/2}h^p.$$

Theorem 2.5. Let the symmetric sample covariance matrix $\boldsymbol{S}$ be positive definite with eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p > 0$. The hyperellipsoid

$$\{\boldsymbol{x}|D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T\boldsymbol{S}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{x}}) \leq h^2\}$$

is centered at $\overline{\boldsymbol{x}}$. The volume of the hyperellipsoid is

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)}|\boldsymbol{S}|^{1/2}h^p.$$

Let $\boldsymbol{w} = \boldsymbol{x} - \overline{\boldsymbol{x}}$. Then points at squared distance $\boldsymbol{w}^T\boldsymbol{S}^{-1}\boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors $\hat{\boldsymbol{e}}_i$ where the half length in the direction of $\hat{\boldsymbol{e}}_i$ is $h\sqrt{\hat{\lambda}_i}$.

From Theorem 2.5, the volume of the hyperellipsoid $\{\boldsymbol{x}|D^2_{\boldsymbol{x}} \leq h^2\}$ is proportional to $|\boldsymbol{S}|^{1/2}$ so the squared volume is proportional to $|\boldsymbol{S}|$. Large $|\boldsymbol{S}|$ corresponds to large volume while small $|\boldsymbol{S}|$ corresponds to small volume.

Definition 2.9. The *generalized sample variance* $= |\boldsymbol{S}| = \det(\boldsymbol{S})$.

Following Johnson and Wichern (1988, p. 103-106), a generalized variance of zero is indicative of extreme degeneracy, and $|\boldsymbol{S}| = 0$ implies that at least one variable X_i is not needed given the other $p - 1$ variables are in the multivariate model. Two necessary conditions for $|\boldsymbol{S}| \neq 0$ are $n > p$ and that $\boldsymbol{S}$ has full rank p. If $\boldsymbol{1}$ is an $n \times 1$ vector of ones,then

$$(n-1)\boldsymbol{S} = (\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T)^T(\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T),$$

and $\boldsymbol{S}$ is of full rank p iff $\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T$ is of full rank p.

If $\boldsymbol{X}$ and $\boldsymbol{W}$ have dispersion matrices $\boldsymbol{\Sigma}$ and $c\boldsymbol{\Sigma}$ where $c > 0$, then the dispersion matrices have the same shape. The dispersion matrices determine the shape of the hyperellipsoid $\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq h^2\}$. Figure 2.1 was made with the *Arc* software of Cook and Weisberg (1999). The 10%,

30%, 50%, 70%, 90% and 98% highest density regions are shown for two multivariate normal (MVN) distributions. Both distributions have $\boldsymbol{\mu} = \mathbf{0}$. In Figure 2.1a),

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 4 \end{pmatrix}.$$

Note that the ellipsoids are narrow with high positive correlation. In Figure 2.1b),

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix}.$$

Note that the ellipsoids are wide with negative correlation. The highest density ellipsoids are superimposed on a scatterplot of a sample of size 100 from each distribution.

2.4 Predictor Transformations

Predictor transformations are used to remove gross nonlinearities in the predictors, and this technique is often very useful. Power transformations are particularly effective, and the techniques of this section are often useful for general regression problems, not just for multivariate analysis. A power transformation has the form $x = t_\lambda(w) = w^\lambda$ for $\lambda \neq 0$ and $x = t_0(w) = \log(w)$ for $\lambda = 0$. Often $\lambda \in \Lambda_L$ where

$$\Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\} \tag{2.9}$$

is called the *ladder of powers*. Often when a power transformation is needed, a transformation that goes "down the ladder," eg from $\lambda = 1$ to $\lambda = 0$ will be useful. If the transformation goes too far down the ladder, eg if $\lambda = 0$ is selected when $\lambda = 1/2$ is needed, then it will be necessary to go back "up the ladder." Additional powers such as ± 2 and ± 3 can always be added.

Definition 2.10. A **scatterplot** of x versus Y is used to visualize the conditional distribution of $Y|x$. A **scatterplot matrix** is an array of scatterplots. It is used to examine the marginal bivariate relationships between the predictors.

Often nine or ten variables can be placed in a scatterplot matrix. The names of the variables appear on the diagonal of the scatterplot matrix. The software *Arc* gives two numbers, the minimum and maximum of the variable,

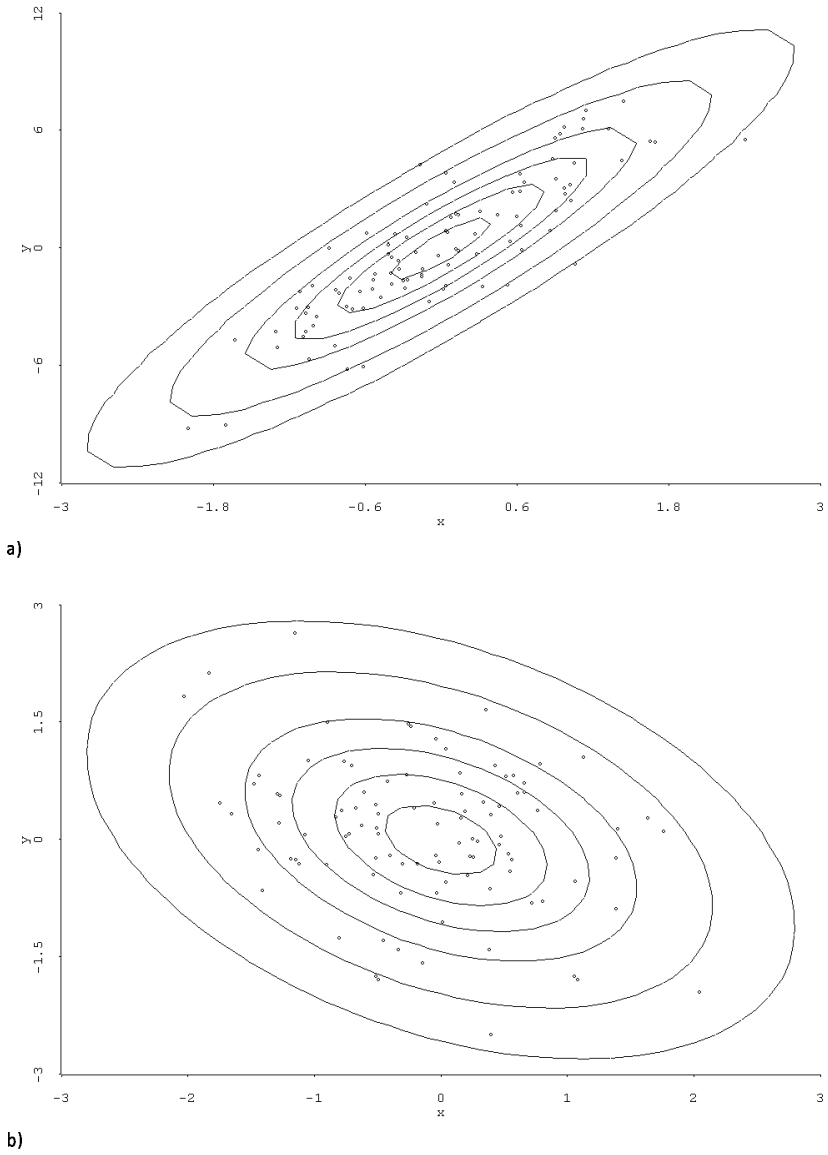

Figure 2.1: Highest Density Regions for 2 MVN Distributions

along with the name of the variable. The software *R/Splus* labels the values of each variable in two places, see Example 2.2 below. Let one of the variables be W. All of the marginal plots above and below W have W on the horizontal axis. All of the marginal plots to the left and the right of W have W on the vertical axis.

If n is large and the p random variables come from an elliptically contoured distribution, then the subplots in the scatterplot matrix should be linear. Nonlinearities suggest that data does not come from an elliptically contoured distribution. There are several rules of thumb that are useful for visually selecting a power transformation to remove nonlinearities from the predictors.

Notation. A *rule of thumb* is a rule that often but not always works well in practice.

Rule of thumb 2.1. a) If strong nonlinearities are apparent in the scatterplot matrix of the predictors $w_2, ..., w_p$, it is often useful to remove the nonlinearities by transforming the predictors using power transformations.

b) Use theory if available.

c) Suppose that variable X_2 is on the vertical axis and X_1 is on the horizontal axis and that the plot of X_1 versus X_2 is nonlinear. The *unit rule* says that if X_1 and X_2 have the same units, then try the same transformation for both X_1 and X_2.

Assume that all values of X_1 and X_2 are positive. Then the following six rules are often used.

d) The **log rule** states that a positive predictor that has the ratio between the largest and smallest values greater than ten should be transformed to logs. So $X > 0$ and $\max(X)/\min(X) > 10$ suggests using $\log(X)$.

e) The **range rule** states that a positive predictor that has the ratio between the largest and smallest values less than two should not be transformed. So $X > 0$ and $\max(X)/\min(X) < 2$ suggests keeping X.

f) The *bulging rule* states that changes to the power of X_2 and the power of X_1 can be determined by the direction that the bulging side of the curve points. If the curve is hollow up (the bulge points down), decrease the power

of X_2. If the curve is hollow down (the bulge points up), increase the power of X_2. If the curve bulges towards large values of X_1 increase the power of X_1. If the curve bulges towards small values of X_1 decrease the power of X_1. See Tukey (1977, p. 173–176).

g) The **ladder rule** appears in Cook and Weisberg (1999a, p. 86).
To spread *small* values of a variable, make λ *smaller*.
To spread *large* values of a variable, make λ *larger*.

h) If it is known that $X_2 \approx X_1^{\lambda}$ and the ranges of X_1 and X_2 are such that this relationship is one to one, then

$$X_1^{\lambda} \approx X_2 \text{ and } X_2^{1/\lambda} \approx X_1.$$

Hence either the transformation X_1^{λ} or $X_2^{1/\lambda}$ will linearize the plot. Note that $\log(X_2) \approx \lambda \log(X_1)$, so taking logs of both variables will also linearize the plot. This relationship frequently occurs if there is a volume present. For example let X_2 be the volume of a sphere and let X_1 be the circumference of a sphere.

i) The *cube root rule* says that if X is a volume measurement, then cube root transformation $X^{1/3}$ may be useful.

Theory, if available, should be used to select a transformation. Frequently more than one transformation will work. For example if W = weight and X_1 = volume = $(X_2)(X_3)(X_4)$, then W versus $X_1^{1/3}$ and $\log(W)$ versus $\log(X_1) = \log(X_2) + \log(X_3) + \log(X_4)$ may both work. Also if W is linearly related with X_2, X_3, X_4 and these three variables all have length units mm, say, then the units of X_1 are $(mm)^3$. Hence the units of $X_1^{1/3}$ are mm.

Suppose that all values of the variable w to be transformed are positive. The log rule says use $\log(w)$ if $\max(w_i)/\min(w_i) > 10$. This rule often works wonders on the data and the log transformation is the most used (modified) power transformation. If the variable w can take on the value of 0, use $\log(w + c)$ where c is a small constant like 1, 1/2, or 3/8.

To use the ladder rule, suppose you have a scatterplot of two variables $x_1^{\lambda_1}$ versus $x_2^{\lambda_2}$ where both $x_1 > 0$ and $x_2 > 0$. Also assume that the plotted points follow a nonlinear one to one function. Consider the ladder of powers

$$\Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1, \}.$$

To spread small values of the variable, make λ_i smaller. To spread large values of the variable, make λ_i larger.

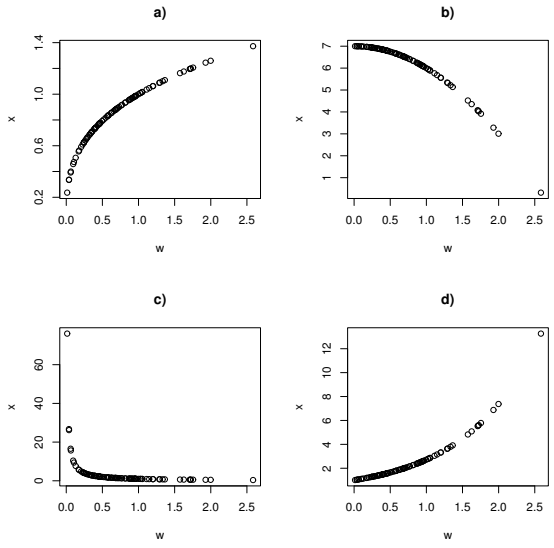

Figure 2.2: Plots to Illustrate the Bulging and Ladder Rules

For example, if both variables are **right skewed**, then there will be many more cases in the lower left of the plot than in the upper right. Hence small values of both variables need spreading.

Consider the ladder of powers. Often no transformation ($\lambda = 1$) is best, then the log transformation, then the square root transformation, then the reciprocal transformation.

Example 2.1. Examine Figure 2.2. Let $X_1 = w$ and $X_2 = x$. Since w is on the horizontal axis, mentally add a narrow vertical slice to the plot. If a large amount of data falls in the slice at the left of the plot, then small values need spreading. Similarly, if a large amount of data falls in the slice at the right of the plot (compared to the middle and left of the plot), then large values need spreading. For the variable on the vertical axis, make a narrow horizontal slice. If the plot looks roughly like the northwest corner of a square then small values of the horizontal and large values of the vertical variable need spreading. Hence in Figure 2.2a, small values of w need spreading. Notice that the plotted points bulge up towards small values of the horizontal variable. If the plot looks roughly like the northeast corner of a square, then large values of both variables need spreading. Hence in Figure 2.2b, large

values of x need spreading. Notice that the plotted points bulge up towards large values of the horizontal variable. If the plot looks roughly like the southwest corner of a square, as in Figure 2.2c, then small values of both variables need spreading. Notice that the plotted points bulge down towards small values of the horizontal variable. If the plot looks roughly like the southeast corner of a square, then large values of the horizontal and small values of the vertical variable need spreading. Hence in Figure 2.2d, small values of x need spreading. Notice that the plotted points bulge down towards large values of the horizontal variable.

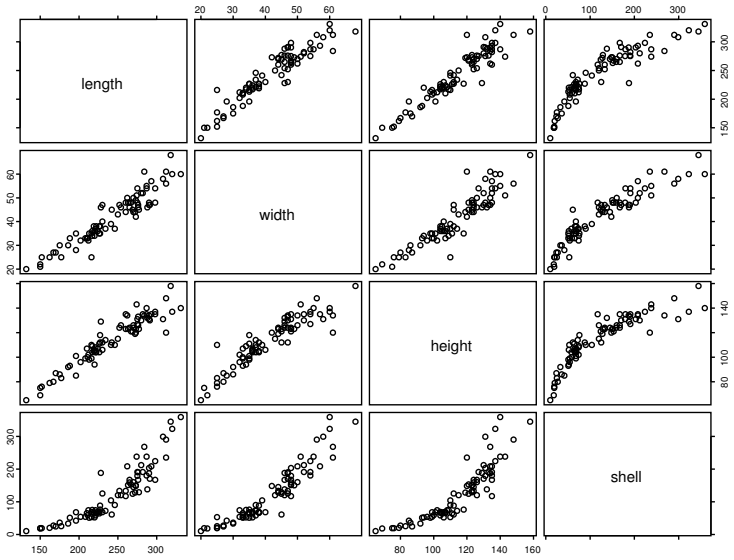

Figure 2.3: Scatterplot Matrix for Original Mussel Data Predictors

Example 2.2: Mussel Data. Cook and Weisberg (1999a, p. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. The response is *muscle mass* M in grams, and the predictors are a constant, the *length* L and *height* H of the shell in mm, the *shell width* W and the *shell mass* S. Figure 2.3 shows the scatterplot matrix of the predictors L, H, W and S. Examine the variable *length*. Length is on the vertical axis on the three top plots and the right of the scatterplot matrix labels this axis from 150 to 300. Length is on the horizontal axis on the three leftmost marginal plots, and this axis is labelled from 150 to 300 on the bottom of the

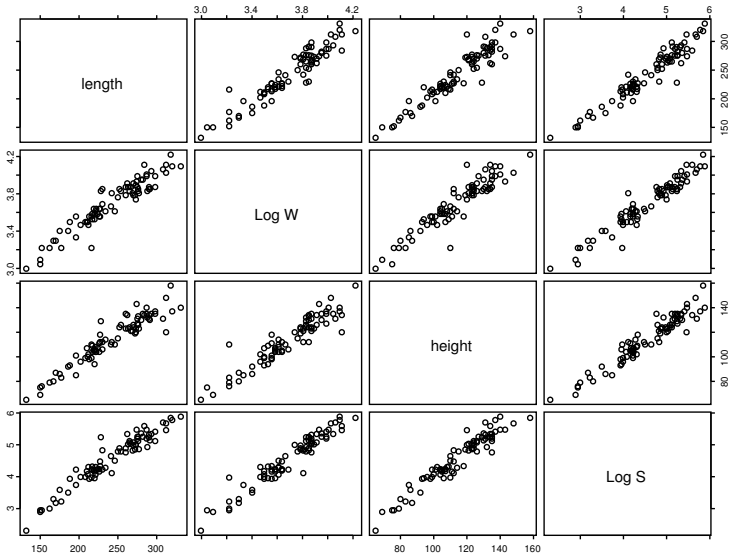

Figure 2.4: Scatterplot Matrix for Transformed Mussel Data Predictors

scatterplot matrix. The marginal plot in the bottom left corner has length on the horizontal and shell on the vertical axis. The marginal plot that is second from the top and second from the right has height on the horizontal and width on the vertical axis. If the data is stored in x, the plot can be made with the following command in R.

```
pairs(x,labels=c("length",‘"width","height","shell"))
```

Nonlinearity is present in several of the plots. For example, width and length seem to be linearly related while length and shell have a nonlinear relationship. The minimum value of shell is 10 while the max is 350. Since $350/10 = 35 > 10$, the log rule suggests that $\log S$ may be useful. If $\log S$ replaces S in the scatterplot matrix, then there may be some nonlinearity present in the plot of $\log S$ versus W with small values of W needing spreading. Hence the ladder rule suggests reducing λ from 1 and we tried $\log(W)$. Figure 2.4 shows that taking the log transformations of W and S results in a scatterplot matrix that is much more linear than the scatterplot matrix of Figure 2.3. Notice that the plot of W versus L and the plot of $\log(W)$ versus L both appear linear. This plot can be made with the following commands.

```
z <- x; z[,2] <- log(z[,2]); z[,4] <- log(z[,4])
pairs(z,labels=c("length","Log W","height","Log S"))
```

The plot of *shell* versus *height* in Figure 2.3 is nonlinear, and small values of *shell* need spreading since if the plotted points were projected on the horizontal axis, there would be too many points at values of *shell* near 0. Similarly, large values of *height* need spreading.

2.5 Summary

The following three quantities are important.

1) $E(\boldsymbol{x}) = \boldsymbol{\mu} = (E(x_1), ..., E(x_p))^T$.

2) The $p \times p$ *population covariance matrix* $\mathrm{Cov}(\boldsymbol{x}) = E(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T = ((\sigma_{i,j})) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$.

3) The $p \times p$ *population correlation matrix* $\mathrm{Cor}(\boldsymbol{x}) = \boldsymbol{\rho} = ((\rho_{ij}))$.

4) The *population covariance matrix* of $\boldsymbol{x}$ with $\boldsymbol{y}$ is $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{y}} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{y} - E(\boldsymbol{y}))^T]$.

5) Let the $p \times p$ matrix $\boldsymbol{\Delta} = \mathrm{diag}(\sqrt{\sigma_{11}}, ..., \sqrt{\sigma_{\mathrm{pp}}})$. Then $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Delta}\boldsymbol{\rho}\boldsymbol{\Delta}$, and $\boldsymbol{\rho} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Delta}^{-1}$.

6) The $n \times p$ data matrix

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix}$$

7) The **sample mean** or *sample mean vector*

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = (\overline{x}_1, ..., \overline{x}_p)^T = \frac{1}{n}\boldsymbol{W}^T\boldsymbol{1}$$

where $\boldsymbol{1}$ is the $p \times 1$ vector of ones.

8) The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = ((S_{ij})).$$

9) $(n-1)\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\ \overline{\boldsymbol{x}}^T = (\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T)^T(\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T) =$

$\boldsymbol{W}^T\boldsymbol{W} - \frac{1}{n}\boldsymbol{W}^T\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{W}$. Hence if the *centering matrix* $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$, then $(n-1)\boldsymbol{S} = \boldsymbol{W}^T\boldsymbol{H}\boldsymbol{W}$.

10) The **sample correlation matrix** $\boldsymbol{R} = ((r_{ij}))$.

11) Let the $p \times p$ sample standard deviation matrix $\boldsymbol{D} = \text{diag}(\sqrt{\text{S}_{11}}, ..., \sqrt{\text{S}_{\text{pp}}})$. Then $\boldsymbol{S} = \boldsymbol{D}\boldsymbol{R}\boldsymbol{D}$, and $\boldsymbol{R} = \boldsymbol{D}^{-1}\boldsymbol{S}\boldsymbol{D}^{-1}$.

12) The spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T =$
$\lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^T + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^T$.

13) Let $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$ be a positive definite $p \times p$ symmetric matrix. Let $\boldsymbol{P} = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_p]$ be the $p \times p$ orthogonal matrix with ith column $\boldsymbol{e}_i$. Let $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_p})$. The *square root matrix* $\boldsymbol{A}^{1/2} = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}\boldsymbol{P}^T$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

14) The population squared Mahalanobis distance
$D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$.

15) The sample squared Mahalanobis distance
$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}})^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}})$.

16) The *generalized sample variance* $= |\boldsymbol{S}| = \det(\boldsymbol{S})$.

17) The hyperellipsoid $\{\boldsymbol{x} | D^2_{\boldsymbol{x}} \leq h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T\boldsymbol{S}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{x}}) \leq h^2\}$ is centered at $\overline{\boldsymbol{x}}$ and has volume is

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)}|\boldsymbol{S}|^{1/2}h^p.$$

Let $\boldsymbol{S}$ have eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$. If $\overline{\boldsymbol{x}} = \boldsymbol{0}$, the axes are given by the eigenvectors $\hat{\boldsymbol{e}}_i$ where the half length in the direction of $\hat{\boldsymbol{e}}_i$ is $h\sqrt{\hat{\lambda}_i}$. Here $\hat{\boldsymbol{e}}_i^T\hat{\boldsymbol{e}}_j = 0$ for $i \neq j$ while $\hat{\boldsymbol{e}}_i^T\hat{\boldsymbol{e}}_i = 1$.

18) A **scatterplot** of x versus y is used to visualize the conditional distribution of $y|x$. A **scatterplot matrix** is an array of scatterplots. It is used to examine the bivariate relationships of the p random variables.

19) There are several guidelines for **choosing power transformations**. First, suppose you have a scatterplot of two variables $x_1^{\lambda_1}$ versus $x_2^{\lambda_2}$ where both $x_1 > 0$ and $x_2 > 0$. Also assume that the plotted points follow a nonlinear one to one function. The **ladder rule:** consider the **ladder of powers**

$$-1, -0.5, -1/3, 0, 1/3, 0.5, \quad \text{and} \quad 1.$$

To spread small values of the variable, make λ_i smaller. To spread large values of the variable, make λ_i larger.

20) Suppose that all values of the variable w to be transformed are positive. The **log rule** says use $\log(w)$ if $\max(w_i)/\min(w_i) > 10$.

21) If p random variables come from an elliptically contoured distribution, then the subplots in the scatterplot matrix should be linear.

2.6 Complements

Section 2.3 will be useful for principal component analysis and for prediction regions.

2.7 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

2.1. Assuming all relevant expectations exist, show
$\mathrm{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$.

2.2. Suppose $Z_i = \dfrac{X_i - E(X_i)}{\sqrt{\sigma_{ii}}}$. Show $\mathrm{Cov}(Z_i, Z_j) = \mathrm{Cor}(X_i, X_j)$.

2.3. i) Let $\boldsymbol{\Sigma}$ be a $p \times p$ matrix with eigenvalue eigenvector pair $(\lambda, \boldsymbol{x})$. Show that $c\boldsymbol{x}$ is also an eigenvector of $\boldsymbol{\Sigma}$ where $c \neq 0$ is a real number.

ii) Let $\boldsymbol{\Sigma}$ be a $p \times p$ matrix with eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$. Find the eigenvalue eigenvector pairs of $\boldsymbol{A} = c\boldsymbol{\Sigma}$ where $c \neq 0$ is a real number.

2.4. i) Let $\boldsymbol{\Sigma}$ be a $p \times p$ matrix with eigenvalue eigenvector pair $(\lambda, \boldsymbol{x})$. Show that $c\boldsymbol{x}$ is also an eigenvector of $\boldsymbol{\Sigma}$ where $c \neq 0$ is a real number.

ii) Let $\boldsymbol{\Sigma}$ be a $p \times p$ matrix with eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$. Find the eigenvalue eigenvector pairs of $\boldsymbol{A} = c\boldsymbol{\Sigma}$ where $c \neq 0$ is a real number.

2.5. Suppose $\boldsymbol{A}$ is a symmetric positive definite matrix with eigenvalue eigenvector pair $(\lambda, \boldsymbol{e})$. Then $\boldsymbol{A}\boldsymbol{e} = \lambda\boldsymbol{e}$ so $\boldsymbol{A}^2\boldsymbol{e} = \boldsymbol{A}\boldsymbol{A}\boldsymbol{e} = \boldsymbol{A}\lambda\boldsymbol{e}$. Find an eigenvalue eigenvector pair for $\boldsymbol{A}^2$.

2.6. Suppose $\boldsymbol{A}$ is a symmetric positive definite matrix with eigenvalue eigenvector pair $(\lambda, \boldsymbol{e})$. Then $\boldsymbol{Ae} = \lambda\boldsymbol{e}$ so $\boldsymbol{A}^{-1}\boldsymbol{Ae} = \boldsymbol{A}^{-1}\lambda\boldsymbol{e}$. Find an eigenvalue eigenvector pair for $\boldsymbol{A}^{-1}$.

Problems using ARC

To quit *Arc*, move the cursor to the **x** in the northeast corner and click.

2.7*. This problem makes plots similar to Figure 2.1. Data sets of $n = 100$ cases from two multivariate normal $N_2(\mathbf{0}, \boldsymbol{\Sigma}_i)$ distributions are generated and plotted in a scatterplot along with the 10%, 30%, 50%, 70%, 90% and 98% highest density regions where

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 4 \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix}.$$

Activate *Arc* (Cook and Weisberg 1999a). Generally this will be done by finding the icon for *Arc* or the executable file for *Arc*. Using the mouse, move the pointer (cursor) to the icon and press the leftmost mouse button twice, rapidly. This procedure is known as *double clicking* on the icon. A window should appear with a "greater than" > prompt. The menu *File* should be in the upper left corner of the window. Move the pointer to *File* and hold the leftmost mouse button down. Then the menu will appear. Drag the pointer down to the menu command *load.* Then click on *data* and then click on *demo-bn.lsp*. (You may need to use the *slider bar* in the middle of the screen to see the file *demo-bn.lsp*: click on the arrow pointing to the right until the file appears.) In the future these menu commands will be denoted by "File > Load > Data > demo-bn.lsp." These are the commands needed to activate the file *demo-bn.lsp.*

a) In the *Arc* dialog window, enter the numbers
0 0 1 4 0.9 and 100. Then click on *OK*.

The graph can be printed with the menu commands "File>Print," but it will generally save paper by placing the plots in the *Word* editor.

Activate *Word* (often by double clicking on the *Word* icon). Click on the screen and type "Problem 2.4a." In *Arc,* use the menu commands "Edit>Copy." In *Word,* click on the *Paste* icon near the upper left corner of *Word* and hold down the leftmost mouse button. This will cause a menu to appear. Drag the pointer down to *Paste*. The plot should appear on the screen. (Older versions of *Word,* use the menu commands "Edit>Paste.") **In the future**, "paste the output into *Word*" will refer to these mouse commands.

b) Either click on *new graph* on the current plot in *Arc* or reload *demo-bn.lsp*. In the *Arc* dialog window, enter the numbers
0 0 1 1 -0.4 and 100. Then place the plot in *Word.*

After editing your *Word* document, get a printout by clicking on the upper left *icon*, select "Print" then select "Print". (Older versions of *Word* use the menu commands "File>Print.")

To save your output on your flash drive G, click on the icon in the upper left corner of *Word*. Then drag the pointer to "Save as." A window will appear, click on the *Word Document* icon. A "Save as" screen appears. Click on the right "check" on the top bar, and then click on "Removable Disk (G:)". Change the file name to HW2d4.docx, and then click on "Save."

To exit from *Word* and *Arc*, click on the "X" in the upper right corner of the screen. In *Word* a screen will appear and ask whether you want to save changes made in your document. Click on *No*. In *Arc,* click on *OK*.

2.8*. In *Arc* enter the menu commands "File>Load>Data" and open the file *mussels.lsp*. Use the commands "Graph&Fit>Scatterplot Matrix of." In the dialog window select H, L, S, W and M (so select M last). Click on "OK" and include the scatterplot matrix in *Word*. The response M is the edible part of the mussel while the 4 predictors are shell measurements. Are any of the marginal predictor relationships nonlinear? Is $E(M|H)$ linear or nonlinear?

2.9*. Activate the McDonald and Schwing (1973) *pollution.lsp* data set with the menu commands "File > Load > Removable Disk (G:) > pollution.lsp." Scroll up the screen to read the data description. Often simply using the log rule on the predictors with $\max(x)/\min(x) > 10$ works wonders.

a) Make a scatterplot matrix of the first nine predictor variables and *Mort*. The commands "Graph&Fit > Scatterplot-Matrix of" will bring down a Dialog menu. Select DENS, EDUC, HC, HOUS, HUMID, JANT, JULT, NONW, NOX and MORT. Then click on *OK*.

A scatterplot matrix with slider bars will appear. Move the slider bars for NOX, NONW and HC to 0, providing the log transformation. In *Arc*, the diagonals have the min and max of each variable, and these were the three predictor variables satisfying the log rule. Open *Word.*

In *Arc*, use the menu commands "Edit > Copy." In *Word*, use the menu commands "Edit > Paste." This should copy the scatterplot matrix into the

Word document. Print the graph.

b) Make a scatterplot matrix of the last six predictor variables. The commands "Graph&Fit > Scatterplot-Matrix of" will bring down a Dialog menu. Select OVR65, POOR, POPN, PREC, SO, WWDRK and MORT. Then click on *OK*. Move the slider bar of SO to 0 and copy the plot into *Word*. Print the plot as described in a).

R/Splus Problems

2.10. Use the following *R/Splus* commands to make 100 multivariate normal (MVN) $N_3(\mathbf{0}, I_3)$ cases and 100 trivariate non-EC lognormal cases.

```
n3x <- matrix(rnorm(300),nrow=100,ncol=3)
ln3x <- exp(n3x)
```

In *R*, type the command *library(MASS).*

Using the commands *pairs(n3x)* and *pairs(ln3x)* and include both scatterplot matrices in *Word.* (Click on the plot and hit *Ctrl* and *c* at the same time. Then go to *file* in the *Word* menu and select *paste.*) Are strong nonlinearities present among the MVN predictors? How about the non-EC predictors? (Hint: a box or ball shaped plot is linear.)

Chapter 3

Elliptically Contoured Distributions

The multivariate location and dispersion model of Definition 2.1 is in many ways similar to the multiple linear regression model. The data are iid vectors from some distribution such as the multivariate normal (MVN) distribution. The location parameter $\boldsymbol{\mu}$ of interest may be the mean or the center of symmetry of an elliptically contoured distribution. Hyperellipsoids will be estimated instead of hyperplanes, and Mahalanobis distances will be used instead of absolute residuals to determine if an observation is a potential outlier. Review Section 2.1 for important notation.

Although usually random vectors in this text are denoted by $\boldsymbol{x}$, $\boldsymbol{y}$ or $\boldsymbol{z}$, this chapter will usually use the notation $\boldsymbol{X} = (X_1, ..., X_p)^T$ and $\boldsymbol{Y}$ for the random vectors, and $\boldsymbol{x} = (x_1, ..., x_p)^T$ for the observed value of the random vector. This notation will be useful to avoid confusion when studying conditional distributions such as $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$.

3.1 The Multivariate Normal Distribution

Definition 3.1: Rao (1965, p. 437). A $p \times 1$ random vector $\boldsymbol{X}$ has a $p-$dimensional *multivariate normal distribution* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ iff $\boldsymbol{t}^T\boldsymbol{X}$ has a univariate normal distribution for any $p \times 1$ vector $\boldsymbol{t}$.

If $\boldsymbol{\Sigma}$ is positive definite, then $\boldsymbol{X}$ has a pdf

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(1/2)(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})} \tag{3.1}$$

where $|\boldsymbol{\Sigma}|^{1/2}$ is the square root of the determinant of $\boldsymbol{\Sigma}$. Note that if $p = 1$, then the quadratic form in the exponent is $(z-\mu)(\sigma^2)^{-1}(z-\mu)$ and X has the univariate $N(\mu, \sigma^2)$ pdf. If $\boldsymbol{\Sigma}$ is positive semidefinite but not positive definite, then $\boldsymbol{x}$ has a degenerate distribution. For example, the univariate $N(0, 0^2)$ distribution is degenerate (the point mass at 0).

Definition 3.2. The *population mean* of a random $p \times 1$ vector $\boldsymbol{X} = (X_1, ..., X_p)^T$ is

$$E(\boldsymbol{X}) = (E(X_1), ..., E(X_p))^T$$

and the $p \times p$ *population covariance matrix*

$$\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = E(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T = ((\sigma_{i,j})).$$

That is, the ij entry of $\text{Cov}(\boldsymbol{X})$ is $\text{Cov}(X_i, X_j) = \sigma_{i,j}$.

The covariance matrix is also called the variance–covariance matrix and variance matrix. Sometimes the notation $\text{Var}(\boldsymbol{X})$ is used. Note that $\text{Cov}(\boldsymbol{X})$ is a symmetric positive semidefinite matrix. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $p \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{a} + \boldsymbol{X}) = \boldsymbol{a} + E(\boldsymbol{X}) \quad \text{and} \quad E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}) \tag{3.2}$$

and

$$E(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}E(\boldsymbol{X}) \quad \text{and} \quad E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}. \tag{3.3}$$

Thus

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}) = \text{Cov}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X})\boldsymbol{A}^T. \tag{3.4}$$

Some important properties of MVN distributions are given in the following three propositions. These propositions can be proved using results from Johnson and Wichern (1988, p. 127-132).

Proposition 3.1. a) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and

$$\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}.$$

b) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then any linear combination $\boldsymbol{t}^T\boldsymbol{X} = t_1X_1 + \cdots + t_pX_p \sim N_1(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$. Conversely, if $\boldsymbol{t}^T\boldsymbol{X} \sim N_1(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$ for every $p \times 1$ vector $\boldsymbol{t}$, then $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

c) **The joint distribution of independent normal random variables is MVN.** If $X_1, ..., X_p$ are independent univariate normal $N(\mu_i, \sigma_i^2)$ random vectors, then $\boldsymbol{X} = (X_1, ..., X_p)^T$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)$ and $\boldsymbol{\Sigma} = diag(\sigma_1^2, ..., \sigma_p^2)$ (so the off diagonal entries $\sigma_{i,j} = 0$ while the diagonal entries of $\boldsymbol{\Sigma}$ are $\sigma_{i,i} = \sigma_i^2$).

d) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{AX} \sim N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants, then $\boldsymbol{a} + \boldsymbol{X} \sim N_p(\boldsymbol{a} + \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

It will be useful to partition $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. Let $\boldsymbol{X}_1$ and $\boldsymbol{\mu}_1$ be $q \times 1$ vectors, let $\boldsymbol{X}_2$ and $\boldsymbol{\mu}_2$ be $(p-q) \times 1$ vectors, let $\boldsymbol{\Sigma}_{11}$ be a $q \times q$ matrix, let $\boldsymbol{\Sigma}_{12}$ be a $q \times (p-q)$ matrix, let $\boldsymbol{\Sigma}_{21}$ be a $(p-q) \times q$ matrix, and let $\boldsymbol{\Sigma}_{22}$ be a $(p-q) \times (p-q)$ matrix. Then

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Proposition 3.2. a) **All subsets of a MVN are MVN:** $(X_{k_1}, ..., X_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\boldsymbol{\mu}}_i = E(X_{k_i})$ and $\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}(X_{k_i}, X_{k_j})$. In particular, $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

b) If $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent, then $\text{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \boldsymbol{\Sigma}_{12} = E[(\boldsymbol{X}_1 - E(\boldsymbol{X}_1))(\boldsymbol{X}_2 - E(\boldsymbol{X}_2))^T] = \boldsymbol{0}$, a $q \times (p-q)$ matrix of zeroes.

c) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent iff $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

d) If $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ are independent, then

$$\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N_p \left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right).$$

Proposition 3.3. The conditional distribution of a MVN is MVN. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. That is,

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

Example 3.1. Let $p = 2$ and let $(Y, X)^T$ have a bivariate normal distribution. That is,

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \text{Cov}(Y, X) \\ \text{Cov}(X, Y) & \sigma_X^2 \end{pmatrix} \right).$$

Also recall that the population correlation between X and Y is given by

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{VAR}(X)}\sqrt{\mathrm{VAR}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

if $\sigma_X > 0$ and $\sigma_Y > 0$. Then $Y|X = x \sim N(E(Y|X = x), \mathrm{VAR}(Y|X = x))$ where the conditional mean

$$E(Y|X = x) = \mu_Y + \mathrm{Cov}(Y,X)\frac{1}{\sigma_X^2}(x - \mu_X) = \mu_Y + \rho(X,Y)\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}(x - \mu_X)$$

and the conditional variance

$$\mathrm{VAR}(Y|X = x) = \sigma_Y^2 - \mathrm{Cov}(X,Y)\frac{1}{\sigma_X^2}\mathrm{Cov}(X,Y)$$

$$= \sigma_Y^2 - \rho(X,Y)\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}\rho(X,Y)\sqrt{\sigma_X^2}\sqrt{\sigma_Y^2}$$

$$= \sigma_Y^2 - \rho^2(X,Y)\sigma_Y^2 = \sigma_Y^2[1 - \rho^2(X,Y)].$$

Also $aX + bY$ is univariate normal with mean $a\mu_X + b\mu_Y$ and variance

$$a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\ \mathrm{Cov}(X,Y).$$

Remark 3.1. There are several common misconceptions. First, **it is not true that every linear combination $\boldsymbol{t}^T\boldsymbol{X}$ of normal random variables is a normal random variable,** and **it is not true that all uncorrelated normal random variables are independent**. The key condition in Proposition 3.1b and Proposition 3.2c is that the joint distribution of $\boldsymbol{X}$ is MVN. It is possible that $X_1, X_2, ..., X_p$ each has a marginal distribution that is univariate normal, but the joint distribution of $\boldsymbol{X}$ is not MVN. See Seber and Lee (2003, p. 23), Kowalski (1973) and examine the following example from Rohatgi (1976, p. 229). Suppose that the joint pdf of X and Y is a mixture of two bivariate normal distributions both with $EX = EY = 0$ and $\mathrm{VAR}(X) = \mathrm{VAR}(Y) = 1$, but $\mathrm{Cov}(X,Y) = \pm\rho$. Hence $f(x,y) =$

$$\frac{1}{2}\ \frac{1}{2\pi\sqrt{1-\rho^2}}\exp(\frac{-1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)) +$$

$$\frac{1}{2}\ \frac{1}{2\pi\sqrt{1-\rho^2}}\exp(\frac{-1}{2(1-\rho^2)}(x^2 + 2\rho xy + y^2)) \equiv \frac{1}{2}f_1(x,y) + \frac{1}{2}f_2(x,y)$$

where x and y are real and $0 < \rho < 1$. Since both marginal distributions of $f_i(x, y)$ are N(0,1) for $i = 1$ and 2 by Proposition 3.2 a), the marginal distributions of X and Y are N(0,1). Since $\int\int xyf_i(x, y)dxdy = \rho$ for $i = 1$ and $-\rho$ for $i = 2$, X and Y are uncorrelated, but X and Y are not independent since $f(x, y) \neq f_X(x)f_Y(y)$.

Remark 3.2. In Proposition 3.3, suppose that $\boldsymbol{X} = (Y, X_2, ..., X_p)^T$. Let $X_1 = Y$ and $\boldsymbol{X}_2 = (X_2, ..., X_p)^T$. Then $E[Y|\boldsymbol{X}_2] = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ and VAR$[Y|\boldsymbol{X}_2]$ is a constant that does not depend on $\boldsymbol{X}_2$. Hence $Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e$ follows the multiple linear regression model.

3.2 Elliptically Contoured Distributions

Definition 3.3: Johnson (1987, p. 107-108). A $p \times 1$ random vector $\boldsymbol{X}$ has an *elliptically contoured distribution,* also called an *elliptically symmetric distribution,* if $\boldsymbol{X}$ has joint pdf

$$f(\boldsymbol{z}) = k_p|\boldsymbol{\Sigma}|^{-1/2}g[(\boldsymbol{z} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z} - \boldsymbol{\mu})], \tag{3.5}$$

and we say $\boldsymbol{X}$ has an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution.

If $\boldsymbol{X}$ has an elliptically contoured (EC) distribution, then the characteristic function of $\boldsymbol{X}$ is

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(i\boldsymbol{t}^T\boldsymbol{\mu})\psi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}) \tag{3.6}$$

for some function ψ. If the second moments exist, then

$$E(\boldsymbol{X}) = \boldsymbol{\mu} \tag{3.7}$$

and

$$\text{Cov}(\boldsymbol{X}) = c_X\boldsymbol{\Sigma} \tag{3.8}$$

where

$$c_X = -2\psi'(0).$$

Definition 3.4. The *population squared Mahalanobis distance*

$$U \equiv D^2 = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}). \tag{3.9}$$

For elliptically contoured distributions, U has pdf

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{3.10}$$

For $c > 0$, an $EC_p(\boldsymbol{\mu}, c\boldsymbol{I}, g)$ distribution is *spherical about* $\boldsymbol{\mu}$ where $\boldsymbol{I}$ is the $p \times p$ identity matrix. The *multivariate normal distribution* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has $k_p = (2\pi)^{-p/2}$, $\psi(u) = g(u) = \exp(-u/2)$ and $h(u)$ is the χ^2_p pdf.

The following lemma is useful for proving properties of EC distributions without using the characteristic function (10.6). See Eaton (1986) and Cook (1998, p. 57, 130).

Lemma 3.4. Let $\boldsymbol{X}$ be a $p \times 1$ random vector with 1st moments; ie, $E(\boldsymbol{X})$ exists. Let $\boldsymbol{B}$ be any constant full rank $p \times r$ matrix where $1 \leq r \leq p$. Then $\boldsymbol{X}$ is elliptically contoured iff for all such conforming matrices $\boldsymbol{B}$,

$$E(\boldsymbol{X}|\boldsymbol{B}^T\ \boldsymbol{X}) = \boldsymbol{\mu} + \boldsymbol{M}_B \boldsymbol{B}^T(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{a}_B + \boldsymbol{M}_B \boldsymbol{B}^T \boldsymbol{X} \tag{3.11}$$

where the $p \times 1$ constant vector $\boldsymbol{a}_B$ and the $p \times r$ constant matrix $\boldsymbol{M}_B$ both depend on $\boldsymbol{B}$.

A useful fact is that $\boldsymbol{a}_B$ and $\boldsymbol{M}_B$ do not depend on g:

$$\boldsymbol{a}_B = \boldsymbol{\mu} - \boldsymbol{M}_B \boldsymbol{B}^T \boldsymbol{\mu} = (\boldsymbol{I}_p - \boldsymbol{M}_B \boldsymbol{B}^T)\boldsymbol{\mu},$$

and

$$\boldsymbol{M}_B = \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1}.$$

See Problem 3.11. Notice that in the formula for $\boldsymbol{M}_B$, $\boldsymbol{\Sigma}$ can be replaced by $c\boldsymbol{\Sigma}$ where $c > 0$ is a constant. In particular, if the EC distribution has 2nd moments, $\mathrm{Cov}(\boldsymbol{X})$ can be used instead of $\boldsymbol{\Sigma}$.

To use Lemma 3.4 to prove interesting properties, partition $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. Let $\boldsymbol{X}_1$ and $\boldsymbol{\mu}_1$ be $q \times 1$ vectors, let $\boldsymbol{X}_2$ and $\boldsymbol{\mu}_2$ be $(p-q) \times 1$ vectors. Let $\boldsymbol{\Sigma}_{11}$ be a $q \times q$ matrix, let $\boldsymbol{\Sigma}_{12}$ be a $q \times (p-q)$ matrix, let $\boldsymbol{\Sigma}_{21}$ be a $(p-q) \times q$ matrix, and let $\boldsymbol{\Sigma}_{22}$ be a $(p-q) \times (p-q)$ matrix. Then

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Also assume that the $(p+1) \times 1$ vector $(Y, \boldsymbol{X}^T)^T$ is $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ where Y is a random variable, $\boldsymbol{X}$ is a $p \times 1$ vector, and use

$$\left(\begin{array}{c} Y \\ \boldsymbol{X} \end{array} \right), \ \boldsymbol{\mu} = \left(\begin{array}{c} \mu_Y \\ \boldsymbol{\mu}_X \end{array} \right), \text{ and } \boldsymbol{\Sigma} = \left(\begin{array}{cc} \Sigma_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{array} \right).$$

Proposition 3.5. Let $\boldsymbol{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ and assume that $E(\boldsymbol{X})$ exists.

a) Any subset of $\boldsymbol{X}$ is EC, in particular $\boldsymbol{X}_1$ is EC.

b) (Cook 1998 p. 131, Kelker 1970). If $\mathrm{Cov}(\boldsymbol{X})$ is nonsingular,

$$\mathrm{Cov}(\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X}) = d_g(\boldsymbol{B}^T\boldsymbol{X})[\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{\Sigma}]$$

where the real valued function $d_g(\boldsymbol{B}^T\boldsymbol{X})$ is constant iff $\boldsymbol{X}$ is MVN.

Proof of a). Let $\boldsymbol{A}$ be an arbitrary full rank $q \times r$ matrix where $1 \leq r \leq q$. Let

$$\boldsymbol{B} = \left(\begin{array}{c} \boldsymbol{A} \\ \boldsymbol{0} \end{array} \right).$$

Then $\boldsymbol{B}^T\boldsymbol{X} = \boldsymbol{A}^T\boldsymbol{X}_1$, and

$$E[\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X}] = E[\left(\begin{array}{c} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{array} \right) |\boldsymbol{A}^T\boldsymbol{X}_1] =$$

$$\left(\begin{array}{c} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{array} \right) + \left(\begin{array}{c} \boldsymbol{M}_{1B} \\ \boldsymbol{M}_{2B} \end{array} \right) \left(\begin{array}{cc} \boldsymbol{A}^T & \boldsymbol{0}^T \end{array} \right) \left(\begin{array}{c} \boldsymbol{X}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{X}_2 - \boldsymbol{\mu}_2 \end{array} \right)$$

by Lemma 3.4. Hence $E[\boldsymbol{X}_1|\boldsymbol{A}^T\boldsymbol{X}_1] = \boldsymbol{\mu}_1 + \boldsymbol{M}_{1B}\boldsymbol{A}^T(\boldsymbol{X}_1 - \boldsymbol{\mu}_1)$. Since $\boldsymbol{A}$ was arbitrary, $\boldsymbol{X}_1$ is EC by Lemma 3.4. Notice that $\boldsymbol{M}_B = \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1} =$

$$\left(\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array} \right) \left(\begin{array}{c} \boldsymbol{A} \\ \boldsymbol{0} \end{array} \right) [\left(\begin{array}{cc} \boldsymbol{A}^T & \boldsymbol{0}^T \end{array} \right) \left(\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array} \right) \left(\begin{array}{c} \boldsymbol{A} \\ \boldsymbol{0} \end{array} \right)]^{-1}$$

$$= \left(\begin{array}{c} \boldsymbol{M}_{1B} \\ \boldsymbol{M}_{2B} \end{array} \right).$$

Hence

$$\boldsymbol{M}_{1B} = \boldsymbol{\Sigma}_{11}\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{\Sigma}_{11}\boldsymbol{A})^{-1}$$

and $\boldsymbol{X}_1$ is EC with location and dispersion parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_{11}$. QED

Proposition 3.6. Let $(Y, \boldsymbol{X}^T)^T$ be $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ where Y is a random variable.

a) Assume that $E[(Y, \boldsymbol{X}^T)^T]$ exists. Then $E(Y|\boldsymbol{X}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{X}$ where $\alpha = \mu_Y - \boldsymbol{\beta}^T\boldsymbol{\mu}_X$ and

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}.$$

b) Even if the first moment does not exist, the conditional median

$$\text{MED}(Y|\boldsymbol{X}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{X}$$

where α and $\boldsymbol{\beta}$ are given in a).

Proof. a) The trick is to choose $\boldsymbol{B}$ so that Lemma 3.4 applies. Let

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{0}^T \\ \boldsymbol{I}_p \end{pmatrix}.$$

Then $\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{\Sigma}_{XX}$ and

$$\boldsymbol{\Sigma}\boldsymbol{B} = \begin{pmatrix} \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XX} \end{pmatrix}.$$

Now

$$E[\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \mid \boldsymbol{X}] = E[\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \mid \boldsymbol{B}^T\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix}]$$
$$= \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^T\begin{pmatrix} Y - \mu_Y \\ \boldsymbol{X} - \boldsymbol{\mu}_X \end{pmatrix}$$

by Lemma 3.4. The right hand side of the last equation is equal to

$$\boldsymbol{\mu} + \begin{pmatrix} \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XX} \end{pmatrix}\boldsymbol{\Sigma}_{XX}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_X) = \begin{pmatrix} \mu_Y - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\mu}_X + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{X} \\ \boldsymbol{X} \end{pmatrix}$$

and the result follows since

$$\boldsymbol{\beta}^T = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}.$$

b) See Croux, Dehon, Rousseeuw and Van Aelst (2001) for references.

Example 3.2. This example illustrates another application of Lemma 3.4. Suppose that $\boldsymbol{X}$ comes from a mixture of two multivariate normals with the same mean and proportional covariance matrices. That is, let

$$\boldsymbol{X} \sim (1-\gamma)N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \gamma N_p(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$$

where $c > 0$ and $0 < \gamma < 1$. Since the multivariate normal distribution is elliptically contoured (and see Proposition 1.2c),

$$E(\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X}) = (1-\gamma)[\boldsymbol{\mu} + \boldsymbol{M}_1\boldsymbol{B}^T(\boldsymbol{X}-\boldsymbol{\mu})] + \gamma[\boldsymbol{\mu} + \boldsymbol{M}_2\boldsymbol{B}^T(\boldsymbol{X}-\boldsymbol{\mu})]$$

$$= \boldsymbol{\mu} + [(1-\gamma)\boldsymbol{M}_1 + \gamma\boldsymbol{M}_2]\boldsymbol{B}^T(\boldsymbol{X}-\boldsymbol{\mu}) \equiv \boldsymbol{\mu} + \boldsymbol{M}\boldsymbol{B}^T(\boldsymbol{X}-\boldsymbol{\mu}).$$

Since $\boldsymbol{M}_B$ only depends on $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$, it follows that $\boldsymbol{M}_1 = \boldsymbol{M}_2 = \boldsymbol{M} = \boldsymbol{M}_B$. Hence $\boldsymbol{X}$ has an elliptically contoured distribution by Lemma 3.4.

Let $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $y \sim \chi^2_d$ be independent. Let $w_i = x_i/(y/d)^{1/2}$ for $i = 1, ..., p$. Then $\boldsymbol{w}$ has a multivariate t-distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and degrees of freedom d, an important elliptically contoured distribution. Cornish (1954) shows that the covariance matrix of $\boldsymbol{w}$ is $\text{Cov}(\boldsymbol{w}) = \dfrac{d}{d-2}\boldsymbol{\Sigma}$ for $d > 2$. The case $d = 1$ is known as a multivariate Cauchy distribution. The joint pdf of $\boldsymbol{w}$ is

$$f(\boldsymbol{z}) = \frac{\Gamma((d+p)/2))\ |\Sigma|^{-1/2}}{(\pi d)^{p/2}\Gamma(d/2)}[1 + d^{-1}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})]^{-(d+p)/2}.$$

See Mardia, Kent and Bibby (1979, p. 43, 57). See Johnson and Kotz (1972, p. 134) for the special case where the $x_i \sim N(0,1)$.

If $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $u_i = \exp(x_i)$ for $i = 1, ..., p$, then $\boldsymbol{u}$ has a multivariate lognormal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. This distribution is not an elliptically contoured distribution.

3.3 Sample Mahalanobis Distances

In the multivariate location and dispersion model, sample Mahalanobis distances play a role similar to that of residuals in multiple linear regression. The observed data $\boldsymbol{X}_i = \boldsymbol{x}_i$ for $i = 1, ..., n$ is collected in an $n \times p$ matrix $\boldsymbol{W}$ with n rows $\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T$. Let the $p \times 1$ column vector $T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{W})$ be a dispersion estimator.

Definition 3.5. The ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{X}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{X}_i - T(\boldsymbol{W})) \quad (3.12)$$

for each point $\boldsymbol{X}_i$. Notice that D_i^2 is a random variable (scalar valued).

Notice that the population squared Mahalanobis distance is

$$D_{\boldsymbol{X}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \quad (3.13)$$

and that the term $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$ is the $p-$dimensional analog to the z-score used to transform a univariate $N(\mu, \sigma^2)$ random variable into a $N(0, 1)$ random variable. Hence the sample Mahalanobis distance $D_i = \sqrt{D_i^2}$ is an analog of the absolute value $|Z_i|$ of the sample Z-score $Z_i = (X_i - \overline{X})/\hat{\sigma}$. Also notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

Example 3.3. The contours of constant density for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution are ellipsoids defined by $\boldsymbol{x}$ such that $(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = a^2$. An $\alpha-$density region R_α is a set such that $P(\boldsymbol{X} \in R_\alpha) = \alpha$, and for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, the regions of highest density are sets of the form

$$\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq \chi_p^2(\alpha)\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq \chi_p^2(\alpha)\}$$

where $P(W \leq \chi_p^2(\alpha)) = \alpha$ if $W \sim \chi_p^2$. If the $\boldsymbol{X}_i$ are n iid random vectors each with a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ pdf, then a scatterplot of $X_{i,k}$ versus $X_{i,j}$ should be ellipsoidal for $k \neq j$. Similar statements hold if $\boldsymbol{X}$ is $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, but the α-density region will use a constant U_α obtained from Equation (3.10).

The classical Mahalanobis distance corresponds to the sample mean and sample covariance matrix

$$T(\boldsymbol{W}) = \overline{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i,$$

and

$$\boldsymbol{C}(\boldsymbol{W}) = \boldsymbol{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T$$

and will be denoted by MD_i. When $T(\boldsymbol{W})$ and $\boldsymbol{C}(\boldsymbol{W})$ are estimators other than the sample mean and covariance, $D_i = \sqrt{D_i^2}$ will sometimes be denoted by RD_i.

3.4 Large Sample Theory

The first three subsections will review large sample theory for the univariate case, then multivariate theory will be given.

3.4.1 The CLT and the Delta Method

Large sample theory, also called asymptotic theory, is used to approximate the distribution of an estimator when the sample size n is large. This theory is extremely useful if the exact sampling distribution of the estimator is complicated or unknown. To use this theory, one must determine what the estimator is estimating, the rate of convergence, the asymptotic distribution, and how large n must be for the approximation to be useful. Moreover, the (asymptotic) standard error (SE), an estimator of the asymptotic standard deviation, must be computable if the estimator is to be useful for inference.

Theorem 3.7: the Central Limit Theorem (CLT). Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $\text{VAR}(Y) = \sigma^2$. Let the sample mean $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Then

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Hence

$$\sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) = \sqrt{n}\left(\frac{\sum_{i=1}^n Y_i - n\mu}{n\sigma}\right) \xrightarrow{D} N(0, 1).$$

Note that the sample mean is estimating the *population mean* μ with a $\sqrt{n}$ convergence rate, the asymptotic distribution is normal, and the SE $= S/\sqrt{n}$ where S is the *sample standard deviation.* For many distributions the central limit theorem provides a good approximation if the sample size $n > 30$. A special case of the CLT is proven after Theorem 3.20.

Notation. The notation $X \sim Y$ and $X \stackrel{D}{=} Y$ both mean that the random variables X and Y have the same distribution. Hence $F_X(x) = F_Y(y)$ for all real y. The notation $Y_n \xrightarrow{D} X$ means that for large n we can approximate the cdf of Y_n by the cdf of X. The distribution of X is the limiting distribution or asymptotic distribution of Y_n. For the CLT, notice that

$$Z_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) = \left(\frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}}\right)$$

is the z–score of $\overline{Y}$. If $Z_n \xrightarrow{D} N(0,1)$, then the notation $Z_n \approx N(0,1)$, also written as $Z_n \sim AN(0,1)$, means approximate the cdf of Z_n by the standard normal cdf. Similarly, the notation

$$\overline{Y}_n \approx N(\mu, \sigma^2/n),$$

also written as $\overline{Y}_n \sim AN(\mu, \sigma^2/n)$, means approximate the cdf of $\overline{Y}_n$ as if $\overline{Y}_n \sim N(\mu, \sigma^2/n)$.

The two main applications of the CLT are to give the limiting distribution of $\sqrt{n}(\overline{Y}_n - \mu)$ and the limiting distribution of $\sqrt{n}(Y_n/n - \mu_X)$ for a random variable Y_n such that $Y_n = \sum_{i=1}^n X_i$ where the X_i are iid with $E(X) = \mu_X$ and $\mathrm{VAR}(X) = \sigma_X^2$.

Example 3.4. a) Let $Y_1, ..., Y_n$ be iid Ber(ρ). Then $E(Y) = \rho$ and $\mathrm{VAR}(Y) = \rho(1-\rho)$. Hence

$$\sqrt{n}(\overline{Y}_n - \rho) \xrightarrow{D} N(0, \rho(1-\rho))$$

by the CLT.

b) Now suppose that $Y_n \sim BIN(n, \rho)$. Then $Y_n \stackrel{D}{=} \sum_{i=1}^n X_i$ where $X_1, ..., X_n$ are iid Ber(ρ). Hence

$$\sqrt{n}(\frac{Y_n}{n} - \rho) \xrightarrow{D} N(0, \rho(1-\rho))$$

since

$$\sqrt{n}(\frac{Y_n}{n} - \rho) \stackrel{D}{=} \sqrt{n}(\overline{X}_n - \rho) \xrightarrow{D} N(0, \rho(1-\rho))$$

by a).

c) Now suppose that $Y_n \sim BIN(k_n, \rho)$ where $k_n \to \infty$ as $n \to \infty$. Then

$$\sqrt{k_n}(\frac{Y_n}{k_n} - \rho) \approx N(0, \rho(1-\rho))$$

or

$$\frac{Y_n}{k_n} \approx N\left(\rho, \frac{\rho(1-\rho)}{k_n}\right) \quad \text{or} \quad \mathrm{Y_n} \approx \mathrm{N}\left(\mathrm{k_n}\rho, \mathrm{k_n}\rho(1-\rho)\right).$$

Theorem 3.8: the Delta Method. If $g'(\theta) \neq 0$ and

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2),$$

then

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{D} N(0, \sigma^2[g'(\theta)]^2).$$

Example 3.5. Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and VAR$(Y) = \sigma^2$. Then by the CLT,

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Let $g(\mu) = \mu^2$. Then $g'(\mu) = 2\mu \neq 0$ for $\mu \neq 0$. Hence

$$\sqrt{n}((\overline{Y}_n)^2 - \mu^2) \xrightarrow{D} N(0, 4\sigma^2\mu^2)$$

for $\mu \neq 0$ by the delta method.

Example 3.6. Let $X \sim$ Binomial(n, p) where the positive integer n is large and $0 < p < 1$. Find the limiting distribution of $\sqrt{n}\left[\left(\frac{X}{n}\right)^2 - p^2\right]$.

Solution. Example 3.4b gives the limiting distribution of $\sqrt{n}(\frac{X}{n} - p)$. Let $g(p) = p^2$. Then $g'(p) = 2p$ and by the delta method,

$$\sqrt{n}\left[\left(\frac{X}{n}\right)^2 - p^2\right] = \sqrt{n}\left(g(\frac{X}{n}) - g(p)\right) \xrightarrow{D}$$

$$N(0, p(1-p)(g'(p))^2) = N(0, p(1-p)4p^2) = N(0, 4p^3(1-p)).$$

Example 3.7. Let $X_n \sim$ Poisson$(n\lambda)$ where the positive integer n is large and $0 < \lambda$.

a) Find the limiting distribution of $\sqrt{n}\left(\frac{X_n}{n} - \lambda\right)$.

b) Find the limiting distribution of $\sqrt{n}\left[\sqrt{\frac{X_n}{n}} - \sqrt{\lambda}\right]$.

Solution. a) $X_n \stackrel{D}{=} \sum_{i=1}^n Y_i$ where the Y_i are iid Poisson(λ). Hence $E(Y) = \lambda = Var(Y)$. Thus by the CLT,

$$\sqrt{n}\left(\frac{X_n}{n} - \lambda\right) \stackrel{D}{=} \sqrt{n}\left(\frac{\sum_{i=1}^n Y_i}{n} - \lambda\right) \xrightarrow{D} N(0, \lambda).$$

b) Let $g(\lambda) = \sqrt{\lambda}$. Then $g'(\lambda) = \frac{1}{2\sqrt{\lambda}}$ and by the delta method,

$$\sqrt{n}\left[\sqrt{\frac{X_n}{n}} - \sqrt{\lambda}\right] = \sqrt{n}\left(g(\frac{X_n}{n}) - g(\lambda)\right) \xrightarrow{D}$$

$$N(0, \lambda\,(g'(\lambda))^2) = N(0, \lambda\frac{1}{4\lambda}) = N(0, \frac{1}{4}).$$

Example 3.8. Let $Y_1, ..., Y_n$ be independent and identically distributed (iid) from a Gamma(α, β) distribution.

a) Find the limiting distribution of $\sqrt{n}\left(\overline{Y} - \alpha\beta\right)$.

b) Find the limiting distribution of $\sqrt{n}\left((\overline{Y})^2 - c\right)$ for appropriate constant c.

Solution: a) Since $E(Y) = \alpha\beta$ and $V(Y) = \alpha\beta^2$, by the CLT $\sqrt{n}\left(\overline{Y} - \alpha\beta\right) \xrightarrow{D} N(0, \alpha\beta^2)$.

b) Let $\mu = \alpha\beta$ and $\sigma^2 = \alpha\beta^2$. Let $g(\mu) = \mu^2$ so $g'(\mu) = 2\mu$ and $[g'(\mu)]^2 = 4\mu^2 = 4\alpha^2\beta^2$. Then by the delta method, $\sqrt{n}\left((\overline{Y})^2 - c\right) \xrightarrow{D} N(0, \sigma^2[g'(\mu)]^2) = N(0, 4\alpha^3\beta^4)$ where $c = \mu^2 = \alpha^2\beta^2$.

3.4.2 Modes of Convergence and Consistency

Definition 3.6. Let $\{Z_n, n = 1, 2, ...\}$ be a sequence of random variables with cdfs F_n, and let X be a random variable with cdf F. Then Z_n **converges in distribution to** X, written

$$Z_n \xrightarrow{D} X,$$

or Z_n *converges in law to* X, written $Z_n \xrightarrow{L} X$, if

$$\lim_{n\to\infty} F_n(t) = F(t)$$

at each continuity point t of F. The distribution of X is called the **limiting distribution** or the **asymptotic distribution** of Z_n.

An important fact is that **the limiting distribution does not depend on the sample size** n. Notice that the CLT and delta method give the limiting distributions of $Z_n = \sqrt{n}(\overline{Y}_n - \mu)$ and $Z_n = \sqrt{n}(g(T_n) - g(\theta))$, respectively.

Convergence in distribution is useful because if the distribution of X_n is unknown or complicated and the distribution of X is easy to use, then for large n we can approximate the probability that X_n is in an interval by the probability that X is in the interval. To see this, notice that if $X_n \xrightarrow{D} X$, then $P(a < X_n \leq b) = F_n(b) - F_n(a) \to F(b) - F(a) = P(a < X \leq b)$ if F is continuous at a and b.

Warning: convergence in distribution says that the cdf $F_n(t)$ of X_n gets close to the cdf of F(t) of X as $n \to \infty$ provided that t is a continuity point of F. Hence for any $\epsilon > 0$ there exists N_t such that if $n > N_t$, then $|F_n(t) - F(t)| < \epsilon$. Notice that N_t depends on the value of t. Convergence in distribution does not imply that the random variables $X_n \equiv X_n(\omega)$ converge to the random variable $X \equiv X(\omega)$ for all ω.

Example 3.8. Suppose that $X_n \sim U(-1/n, 1/n)$. Then the cdf $F_n(x)$ of X_n is

$$F_n(x) = \begin{cases} 0, & x \leq \frac{-1}{n} \\ \frac{nx}{2} + \frac{1}{2}, & \frac{-1}{n} \leq x \leq \frac{1}{n} \\ 1, & x \geq \frac{1}{n}. \end{cases}$$

Sketching $F_n(x)$ shows that it has a line segment rising from 0 at $x = -1/n$ to 1 at $x = 1/n$ and that $F_n(0) = 0.5$ for all $n \geq 1$. Examining the cases $x < 0$, $x = 0$ and $x > 0$ shows that as $n \to \infty$,

$$F_n(x) \to \begin{cases} 0, & x < 0 \\ \frac{1}{2} & x = 0 \\ 1, & x > 0. \end{cases}$$

Notice that if X is a random variable such that $P(X = 0) = 1$, then X has cdf

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0. \end{cases}$$

Since $x = 0$ is the only discontinuity point of $F_X(x)$ and since $F_n(x) \to F_X(x)$ for all continuity points of $F_X(x)$ (ie for $x \neq 0$),

$$X_n \xrightarrow{D} X.$$

Example 3.9. Suppose $Y_n \sim U(0, n)$. Then $F_n(t) = t/n$ for $0 < t \leq n$ and $F_n(t) = 0$ for $t \leq 0$. Hence $\lim_{n\to\infty} F_n(t) = 0$ for $t \leq 0$. If $t > 0$ and

$n > t$, then $F_n(t) = t/n \to 0$ as $n \to \infty$. Thus $\lim_{n\to\infty} F_n(t) = 0$ for all t and Y_n does not converge in distribution to any random variable Y since $H(t) \equiv 0$ is not a cdf.

Definition 3.7. A sequence of random variables X_n converges in distribution to a constant $\tau(\theta)$, written

$$X_n \xrightarrow{D} \tau(\theta), \quad \text{if} \quad \mathrm{X_n} \xrightarrow{\mathrm{D}} \mathrm{X}$$

where $P(X = \tau(\theta)) = 1$. The distribution of the random variable X is said to be degenerate at $\tau(\theta)$.

Definition 3.8. A sequence of random variables X_n converges in probability to a constant $\tau(\theta)$, written

$$X_n \xrightarrow{P} \tau(\theta),$$

if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - \tau(\theta)| < \epsilon) = 1 \quad \text{or, equivalently,} \quad \lim_{\mathrm{n}\to\infty} \mathrm{P}(|\mathrm{X_n} - \tau(\theta)| \geq \epsilon) = 0.$$

The sequence X_n **converges in probability to** X, written

$$X_n \xrightarrow{P} X,$$

if $X_n - X \xrightarrow{P} 0$.

Notice that $X_n \xrightarrow{P} X$ if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1, \quad \text{or, equivalently,} \quad \lim_{\mathrm{n}\to\infty} \mathrm{P}(|\mathrm{X_n} - \mathrm{X}| \geq \epsilon) = 0.$$

Definition 3.9. A sequence of estimators T_n of $\tau(\theta)$ is **consistent** for $\tau(\theta)$ if

$$T_n \xrightarrow{P} \tau(\theta)$$

for every $\theta \in \Theta$. If T_n is consistent for $\tau(\theta)$, then T_n is a **consistent estimator** of $\tau(\theta)$.

Consistency is a weak property that is usually satisfied by good estimators. T_n is a consistent estimator for $\tau(\theta)$ if the probability that T_n falls in any neighborhood of $\tau(\theta)$ goes to one, regardless of the value of $\theta \in \Theta$.

Definition 3.10. For a real number $r > 0$, Y_n converges in rth mean to a random variable Y, written

$$Y_n \xrightarrow{r} Y,$$

if

$$E(|Y_n - Y|^r) \to 0$$

as $n \to \infty$. In particular, if $r = 2$, Y_n **converges in quadratic mean** to Y, written

$$Y_n \xrightarrow{2} Y \text{ or } \mathrm{Y_n} \xrightarrow{\mathrm{qm}} \mathrm{Y},$$

if

$$E[(Y_n - Y)^2] \to 0$$

as $n \to \infty$.

Lemma 3.9: Generalized Chebyshev's Inequality. Let $u : \Re \to [0, \infty)$ be a nonnegative function. If $E[u(Y)]$ exists then for any $c > 0$,

$$P[u(Y) \geq c] \leq \frac{E[u(Y)]}{c}.$$

If $\mu = E(Y)$ exists, then taking $u(y) = |y - \mu|^r$ and $\tilde{c} = c^r$ gives
Markov's Inequality: for $r > 0$ and any $c > 0$,

$$P(|Y - \mu| \geq c] = P(|Y - \mu|^r \geq c^r] \leq \frac{E[|Y - \mu|^r]}{c^r}.$$

If $r = 2$ and $\sigma^2 = \mathrm{VAR}(Y)$ exists, then we obtain
Chebyshev's Inequality:

$$P(|Y - \mu| \geq c] \leq \frac{\mathrm{VAR}(Y)}{c^2}.$$

Proof. The proof is given for pdfs. For pmfs, replace the integrals by sums. Now

$$E[u(Y)] = \int_{\Re} u(y)f(y)dy = \int_{\{y:u(y)\geq c\}} u(y)f(y)dy + \int_{\{y:u(y)<c\}} u(y)f(y)dy$$

$$\geq \int_{\{y:u(y)\geq c\}} u(y)f(y)dy$$

since the integrand $u(y)f(y) \geq 0$. Hence

$$E[u(Y)] \geq c \int_{\{y:u(y)\geq c\}} f(y)dy = cP[u(Y) \geq c]. \;\; QED$$

The following proposition gives sufficient conditions for T_n to be a consistent estimator of $\tau(\theta)$. Notice that $MSE_{\tau(\theta)}(T_n) \to 0$ for all $\theta \in \Theta$ is equivalent to $T_n \stackrel{qm}{\to} \tau(\theta)$ for all $\theta \in \Theta$.

Proposition 3.10. a) If

$$\lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) = 0$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

b) If

$$\lim_{n\to\infty} \mathrm{VAR}_\theta(T_n) = 0 \;\text{ and }\; \lim_{n\to\infty} E_\theta(T_n) = \tau(\theta)$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

Proof. a) Using Lemma 3.9 with $Y = T_n$, $u(T_n) = (T_n - \tau(\theta))^2$ and $c = \epsilon^2$ shows that for any $\epsilon > 0$,

$$P_\theta(|T_n - \tau(\theta)| \geq \epsilon) = P_\theta[(T_n - \tau(\theta))^2 \geq \epsilon^2] \leq \frac{E_\theta[(T_n - \tau(\theta))^2]}{\epsilon^2}.$$

Hence

$$\lim_{n\to\infty} E_\theta[(T_n - \tau(\theta))^2] = \lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) \to 0$$

is a sufficient condition for T_n to be a consistent estimator of $\tau(\theta)$.

b) Recall that

$$MSE_{\tau(\theta)}(T_n) = \mathrm{VAR}_\theta(T_n) + [\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n})]^2$$

where $\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n}) = \mathrm{E}_\theta(\mathrm{T_n}) - \tau(\theta)$. Since $MSE_{\tau(\theta)}(T_n) \to 0$ if both $\mathrm{VAR}_\theta(T_n) \to 0$ and $\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n}) = \mathrm{E}_\theta(\mathrm{T_n}) - \tau(\theta) \to 0$, the result follows from a). QED

The following result shows estimators that converge at a $\sqrt{n}$ rate are consistent. Use this result and the delta method to show that $g(T_n)$ is a consistent estimator of $g(\theta)$. Note that b) follows from a) with $X_\theta \sim N(0, v(\theta))$.

The WLLN shows that $\overline{Y}$ is a consistent estimator of $E(Y) = \mu$ if $E(Y)$ exists.

Proposition 3.11. a) Let X be a random variable and $0 < \delta \leq 1$. If

$$n^{\delta}(T_n - \tau(\theta)) \xrightarrow{D} X$$

then $T_n \xrightarrow{P} \tau(\theta)$.

b) If

$$\sqrt{n}(T_n - \tau(\theta)) \xrightarrow{D} N(0, v(\theta))$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

Definition 3.11. A sequence of random variables X_n *converges almost everywhere* (or *almost surely*, or *with probability 1*) to X if

$$P(\lim_{n\to\infty} X_n = X) = 1.$$

This type of convergence will be denoted by

$$X_n \xrightarrow{ae} X.$$

Notation such as "X_n converges to X ae" will also be used. Sometimes "ae" will be replaced with "as" or "wp1." We say that X_n *converges almost everywhere* to $\tau(\theta)$, written

$$X_n \xrightarrow{ae} \tau(\theta),$$

if $P(\lim_{n\to\infty} X_n = \tau(\theta)) = 1$.

Theorem 3.12. Let Y_n be a sequence of iid random variables with $E(Y_i) = \mu$. Then

a) **Strong Law of Large Numbers (SLLN):** $\overline{Y}_n \xrightarrow{ae} \mu$, and

b) **Weak Law of Large Numbers (WLLN):** $\overline{Y}_n \xrightarrow{P} \mu$.

Proof of WLLN when $V(Y_i) = \sigma^2$: By Chebyshev's inequality, for every $\epsilon > 0$,

$$P(|\overline{Y}_n - \mu| \geq \epsilon) \leq \frac{V(\overline{Y}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

as $n \to \infty$. QED

In proving consistency results, there is an infinite sequence of estimators that depend on the sample size n. Hence the subscript n will be added to the estimators.

Definition 3.12. Lehmann (1999, p. 53-54): a) A sequence of random variables W_n is *tight* or *bounded in probability*, written $W_n = O_P(1)$, if for every $\epsilon > 0$ there exist positive constants D_ϵ and N_ϵ such that

$$P(|W_n| \leq D_\epsilon) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$. Also $W_n = O_P(X_n)$ if $|W_n/X_n| = O_P(1)$.

b) The sequence $W_n = o_P(n^{-\delta})$ if $n^\delta W_n = o_P(1)$ which means that

$$n^\delta W_n \xrightarrow{P} 0.$$

c) W_n has the *same order as* X_n *in probability*, written $W_n \asymp_P X_n$, if for every $\epsilon > 0$ there exist positive constants N_ϵ and $0 < d_\epsilon < D_\epsilon$ such that

$$P(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon) = P(\frac{1}{D_\epsilon} \leq \left|\frac{X_n}{W_n}\right| \leq \frac{1}{d_\epsilon}) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$.

d) Similar notation is used for a $k \times r$ matrix $\boldsymbol{A} = [a_{i,j}]$ if each element $a_{i,j}$ has the desired property. For example, $\boldsymbol{A} = O_P(n^{-1/2})$ if each $a_{i,j} = O_P(n^{-1/2})$.

Definition 3.13. Let $W_n = \|\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}\|$.

a) If $W_n \asymp_P n^{-\delta}$ for some $\delta > 0$, then both W_n and $\hat{\boldsymbol{\mu}}_n$ have (tightness) **rate** n^δ.

b) If there exists a constant κ such that

$$n^\delta(W_n - \kappa) \xrightarrow{D} X$$

for some nondegenerate random variable X, then both W_n and $\hat{\boldsymbol{\mu}}_n$ have *convergence rate* n^δ.

If W_n has convergence rate n^δ, then W_n has tightness rate n^δ, and the term "tightness" will often be omitted. Notice that if $W_n \asymp_P X_n$, then $X_n \asymp_P W_n$, $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$. Notice that if $W_n = O_P(n^{-\delta})$, then n^δ is a lower bound on the rate of W_n.

Proposition 3.13. Suppose there exists a constant κ such that

$$n^\delta(W_n - \kappa) \xrightarrow{D} X.$$

a) Then $W_n = O_P(n^{-\delta})$.

b) If X is not degenerate, then $W_n \asymp_P n^{-\delta}$.

The above result implies that if W_n has convergence rate n^{δ}, then W_n has tightness rate n^{δ}, and the term "tightness" will often be omitted. Part a) is proved, for example, in Lehmann (1999, p. 67).

The following result shows that if $W_n \asymp_P X_n$, then $X_n \asymp_P W_n$, $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$. Notice that if $W_n = O_P(n^{-\delta})$, then n^{δ} is a lower bound on the rate of W_n. As an example, if the CLT holds then $\overline{Y}_n = O_P(n^{-1/3})$, but $\overline{Y}_n \asymp_P n^{-1/2}$.

Proposition 3.14. a) If $W_n \asymp_P X_n$ then $X_n \asymp_P W_n$.

b) If $W_n \asymp_P X_n$ then $W_n = O_P(X_n)$.

c) If $W_n \asymp_P X_n$ then $X_n = O_P(W_n)$.

d) $W_n \asymp_P X_n$ iff $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$.

Proof. a) Since $W_n \asymp_P X_n$,

$$P(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon) = P(\frac{1}{D_\epsilon} \leq \left|\frac{X_n}{W_n}\right| \leq \frac{1}{d_\epsilon}\ ) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$. Hence $X_n \asymp_P W_n$.

b) Since $W_n \asymp_P X_n$,

$$P(|W_n| \leq |X_n D_\epsilon|) \geq P(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$. Hence $W_n = O_P(X_n)$.

c) Follows by a) and b).

d) If $W_n \asymp_P X_n$, then $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$ by b) and c). Now suppose $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$. Then

$$P(|W_n| \leq |X_n| D_{\epsilon/2}) \geq 1 - \epsilon/2$$

for all $n \geq N_1$, and

$$P(|X_n| \leq |W_n| 1/d_{\epsilon/2}) \geq 1 - \epsilon/2$$

for all $n \geq N_2$. Hence

$$P(A) \equiv P(\left|\frac{W_n}{X_n}\right| \leq D_{\epsilon/2}) \geq 1 - \epsilon/2$$

and

$$P(B) \equiv P(d_{\epsilon/2} \leq \left|\frac{W_n}{X_n}\right|) \geq 1 - \epsilon/2$$

for all $n \geq N = \max(N_1, N_2)$. Since $P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1$,

$$P(A \cap B) = P(d_{\epsilon/2} \leq \left|\frac{W_n}{X_n}\right| \leq D_{\epsilon/2}) \geq 1 - \epsilon/2 + 1 - \epsilon/2 - 1 = 1 - \epsilon$$

for all $n \geq N$. Hence $W_n \asymp_P X_n$. QED

The following result is used to prove the following Theorem 3.16 which says that if there are K estimators $T_{j,n}$ of a parameter $\boldsymbol{\beta}$, such that $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ where $0 < \delta \leq 1$, and if T_n^* picks one of these estimators, then $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$.

Proposition 3.15: Pratt (1959). Let $X_{1,n}, ..., X_{K,n}$ each be $O_P(1)$ where K is fixed. Suppose $W_n = X_{i_n,n}$ for some $i_n \in \{1, ..., K\}$. Then

$$W_n = O_P(1). \tag{3.14}$$

Proof.

$$P(\max\{X_{1,n}, ..., X_{K,n}\} \leq x) = P(X_{1,n} \leq x, ..., X_{K,n} \leq x) \leq$$

$$F_{W_n}(x) \leq P(\min\{X_{1,n}, ..., X_{K,n}\} \leq x) = 1 - P(X_{1,n} > x, ..., X_{K,n} > x).$$

Since K is finite, there exists $B > 0$ and N such that $P(X_{i,n} \leq B) > 1 - \epsilon/2K$ and $P(X_{i,n} > -B) > 1 - \epsilon/2K$ for all $n > N$ and $i = 1, ..., K$. Bonferroni's inequality states that $P(\cap_{i=1}^K A_i) \geq \sum_{i=1}^K P(A_i) - (K-1)$. Thus

$$F_{W_n}(B) \geq P(X_{1,n} \leq B, ..., X_{K,n} \leq B) \geq$$

$$K(1 - \epsilon/2K) - (K-1) = K - \epsilon/2 - K + 1 = 1 - \epsilon/2$$

and

$$-F_{W_n}(-B) \geq -1 + P(X_{1,n} > -B, ..., X_{K,n} > -B) \geq$$

$$-1 + K(1 - \epsilon/2K) - (K-1) = -1 + K - \epsilon/2 - K + 1 = -\epsilon/2.$$

Hence

$$F_{W_n}(B) - F_{W_n}(-B) \geq 1 - \epsilon \text{ for n > N. QED}$$

Theorem 3.16. Suppose $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ for $j = 1, ..., K$ where $0 < \delta \leq 1$. Let $T_n^* = T_{i_n,n}$ for some $i_n \in \{1, ..., K\}$ where, for example, $T_{i_n,n}$ is the $T_{j,n}$ that minimized some criterion function. Then

$$\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{3.15}$$

Proof. Let $X_{j,n} = n^{\delta}\|T_{j,n} - \boldsymbol{\beta}\|$. Then $X_{j,n} = O_P(1)$ so by Proposition 3.15, $n^{\delta}\|T_n^* - \boldsymbol{\beta}\| = O_P(1)$. Hence $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$. QED

3.4.3 Slutsky's Theorem and Related Results

Theorem 3.17: Slutsky's Theorem. Suppose $Y_n \xrightarrow{D} Y$ and $W_n \xrightarrow{P} w$ for some constant w. Then

a) $Y_n + W_n \xrightarrow{D} Y + w$,

b) $Y_n W_n \xrightarrow{D} wY$, and

c) $Y_n/W_n \xrightarrow{D} Y/w$ if $w \neq 0$.

Theorem 3.18. a) If $X_n \xrightarrow{P} X$ then $X_n \xrightarrow{D} X$.

b) If $X_n \xrightarrow{ae} X$ then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

c) If $X_n \xrightarrow{r} X$ then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

d) $X_n \xrightarrow{P} \tau(\theta)$ iff $\mathrm{X_n} \xrightarrow{\mathrm{D}} \tau(\theta)$.

e) If $X_n \xrightarrow{P} \theta$ and τ is continuous at θ, then $\tau(X_n) \xrightarrow{P} \tau(\theta)$.

f) If $X_n \xrightarrow{D} \theta$ and τ is continuous at θ, then $\tau(X_n) \xrightarrow{D} \tau(\theta)$.

Suppose that for all $\theta \in \Theta$, $T_n \xrightarrow{D} \tau(\theta)$, $T_n \xrightarrow{r} \tau(\theta)$ or $T_n \xrightarrow{ae} \tau(\theta)$. Then T_n is a consistent estimator of $\tau(\theta)$ by Theorem 3.18.

Example 3.10. Let $Y_1, ..., Y_n$ be iid with mean $E(Y_i) = \mu$ and variance $V(Y_i) = \sigma^2$. Then the sample mean $\overline{Y}_n$ is a consistent estimator of μ since i) the SLLN holds (use Theorem 3.12 and 3.18), ii) the WLLN holds and iii) the CLT holds (use Proposition 3.11). Since

$$\lim_{n\to\infty} \mathrm{VAR}_{\mu}(\overline{Y}_n) = \lim_{n\to\infty} \sigma^2/n = 0 \quad \text{and} \quad \lim_{n\to\infty} E_{\mu}(\overline{Y}_n) = \mu,$$

$\overline{Y}_n$ is also a consistent estimator of μ by Proposition 3.10b. By the delta method and Proposition 3.11b, $T_n = g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if $g'(\mu) \neq 0$ for all $\mu \in \Theta$. By Theorem 3.18e, $g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if g is continuous at μ for all $\mu \in \Theta$.

Theorem 3.19. a) **Generalized Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is such that $P[X \in C(g)] = 1$ where $C(g)$ is the set of points where g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

b) **Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

Remark 3.3. For Theorem 3.18, a) follows from Slutsky's Theorem by taking $Y_n \equiv X = Y$ and $W_n = X_n - X$. Then $Y_n \xrightarrow{D} Y = X$ and $W_n \xrightarrow{P} 0$. Hence $X_n = Y_n + W_n \xrightarrow{D} Y + 0 = X$. The convergence in distribution parts of b) and c) follow from a). Part f) follows from d) and e). Part e) implies that if T_n is a consistent estimator of θ and τ is a continuous function, then $\tau(T_n)$ is a consistent estimator of $\tau(\theta)$. Theorem 3.19 says that convergence in distribution is preserved by continuous functions, and even some discontinuities are allowed as long as the set of continuity points is assigned probability 1 by the asymptotic distribution. Equivalently, the set of discontinuity points is assigned probability 0.

Example 3.11. (Ferguson 1996, p. 40): If $X_n \xrightarrow{D} X$ then $1/X_n \xrightarrow{D} 1/X$ if X is a continuous random variable since $P(X = 0) = 0$ and $x = 0$ is the only discontinuity point of $g(x) = 1/x$.

Example 3.12. Show that if $Y_n \sim t_n$, a t distribution with n degrees of freedom, then $Y_n \xrightarrow{D} Z$ where $Z \sim N(0, 1)$.

Solution: $Y_n \stackrel{D}{=} Z/\sqrt{V_n/n}$ where $Z \perp\!\!\!\perp V_n \sim \chi^2_n$. If $W_n = \sqrt{V_n/n} \xrightarrow{P} 1$, then the result follows by Slutsky's Theorem. But $V_n \stackrel{D}{=} \sum_{i=1}^{n} X_i$ where the iid $X_i \sim \chi^2_1$. Hence $V_n/n \xrightarrow{P} 1$ by the WLLN and $\sqrt{V_n/n} \xrightarrow{P} 1$ by Theorem 3.14e.

Theorem 3.20: Continuity Theorem. Let Y_n be sequence of random variables with characteristic functions $\phi_n(t)$. Let Y be a random variable with cf $\phi(t)$.

a)

$$Y_n \xrightarrow{D} Y \text{ iff } \phi_n(t) \to \phi(t) \ \forall t \in \Re.$$

b) Also assume that Y_n has mgf m_n and Y has mgf m. Assume that all of the mgfs m_n and m are defined on $|t| \leq d$ for some $d > 0$. Then if $m_n(t) \to m(t)$ as $n \to \infty$ for all $|t| < c$ where $0 < c < d$, then $Y_n \xrightarrow{D} Y$.

Application: Proof of a Special Case of the CLT. Following Rohatgi (1984, p. 569-9), let $Y_1, ..., Y_n$ be iid with mean μ, variance σ^2 and mgf $m_Y(t)$ for $|t| < t_o$. Then

$$Z_i = \frac{Y_i - \mu}{\sigma}$$

has mean 0, variance 1 and mgf $m_Z(t) = \exp(-t\mu/\sigma)m_Y(t/\sigma)$ for $|t| < \sigma t_o$. Want to show that

$$W_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1).$$

Notice that $W_n =$

$$n^{-1/2}\sum_{i=1}^{n} Z_i = n^{-1/2}\sum_{i=1}^{n}\left(\frac{Y_i - \mu}{\sigma}\right) = n^{-1/2}\frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma} = \frac{n^{-1/2}}{\frac{1}{n}}\ \frac{\overline{Y}_n - \mu}{\sigma}.$$

Thus

$$m_{W_n}(t) = E(e^{tW_n}) = E[\exp(tn^{-1/2}\sum_{i=1}^{n} Z_i)] = E[\exp(\sum_{i=1}^{n} tZ_i/\sqrt{n})]$$

$$= \prod_{i=1}^{n} E[e^{tZ_i/\sqrt{n}}] = \prod_{i=1}^{n} m_Z(t/\sqrt{n}) = [m_Z(t/\sqrt{n})]^n.$$

Set $\psi(x) = \log(m_Z(x))$. Then

$$\log[m_{W_n}(t)] = n\log[m_Z(t/\sqrt{n})] = n\psi(t/\sqrt{n}) = \frac{\psi(t/\sqrt{n})}{\frac{1}{n}}.$$

Now $\psi(0) = \log[m_Z(0)] = \log(1) = 0$. Thus by L'Hôpital's rule (where the derivative is with respect to n), $\lim_{n\to\infty}\log[m_{W_n}(t)] =$

$$\lim_{n\to\infty}\frac{\psi(t/\sqrt{n}\ )}{\frac{1}{n}} = \lim_{n\to\infty}\frac{\psi'(t/\sqrt{n}\ )[\frac{-t/2}{n^{3/2}}]}{(\frac{-1}{n^2})} = \frac{t}{2}\lim_{n\to\infty}\frac{\psi'(t/\sqrt{n}\ )}{\frac{1}{\sqrt{n}}}.$$

Now

$$\psi'(0) = \frac{m_Z'(0)}{m_Z(0)} = E(Z_i)/1 = 0,$$

so L'Hôpital's rule can be applied again, giving $\lim_{n\to\infty} \log[m_{W_n}(t)] =$

$$\frac{t}{2}\lim_{n\to\infty}\frac{\psi''(t/\sqrt{n}\,)[\frac{-t}{2n^{3/2}}]}{(\frac{-1}{2n^{3/2}})} = \frac{t^2}{2}\lim_{n\to\infty}\psi''(t/\sqrt{n}\,) = \frac{t^2}{2}\psi''(0).$$

Now

$$\psi''(t) = \frac{d}{dt}\frac{m_Z'(t)}{m_Z(t)} = \frac{m_Z''(t)m_Z(t) - (m_Z'(t))^2}{[m_Z(t)]^2}.$$

So

$$\psi''(0) = m_Z''(0) - [m_Z'(0)]^2 = E(Z_i^2) - [E(Z_i)]^2 = 1.$$

Hence $\lim_{n\to\infty} \log[m_{W_n}(t)] = t^2/2$ and

$$\lim_{n\to\infty} m_{W_n}(t) = \exp(t^2/2)$$

which is the N(0,1) mgf. Thus by the continuity theorem,

$$W_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1).$$

3.4.4 Multivariate Limit Theorems

Many of the univariate results of the previous 3 subsections can be extended to random vectors. For the limit theorems, the vector $\boldsymbol{X}$ is typically a $k \times 1$ column vector and $\boldsymbol{X}^T$ is a row vector. Let $\|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_k^2}$ be the Euclidean norm of $\boldsymbol{x}$.

Definition 3.14. Let $\boldsymbol{X}_n$ be a sequence of random vectors with joint cdfs $F_n(\boldsymbol{x})$ and let $\boldsymbol{X}$ be a random vector with joint cdf $F(\boldsymbol{x})$.

a) $\boldsymbol{X}_n$ **converges in distribution** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, if $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ as $n \to \infty$ for all points $\boldsymbol{x}$ at which $F(\boldsymbol{x})$ is continuous. The distribution of $\boldsymbol{X}$ is the **limiting distribution** or **asymptotic distribution** of $\boldsymbol{X}_n$.

b) $\boldsymbol{X}_n$ **converges in probability** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \to 0$ as $n \to \infty$.

c) Let $r > 0$ be a real number. Then $\boldsymbol{X}_n$ **converges in rth mean** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{r} \boldsymbol{X}$, if $E(\|\boldsymbol{X}_n - \boldsymbol{X}\|^r) \to 0$ as $n \to \infty$.

d) $\boldsymbol{X}_n$ **converges almost everywhere** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{ae} \boldsymbol{X}$, if $P(\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}) = 1$.

Theorems 3.21 and 3.22 below are the multivariate extensions of the limit theorems in subsection 3.4.1. When the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta}))$ is multivariate normal $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$, approximate the joint cdf of $\boldsymbol{Z}_n$ with the joint cdf of the $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$ distribution. Thus to find probabilities, manipulate $\boldsymbol{Z}_n$ as if $\boldsymbol{Z}_n \approx N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. To see that the CLT is a special case of the MCLT below, let $k = 1$, $E(X) = \mu$ and $V(X) = \boldsymbol{\Sigma}_{\boldsymbol{X}} = \sigma^2$.

Theorem 3.21: the Multivariate Central Limit Theorem (MCLT). If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}}$, then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$$

where the sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

To see that the delta method is a special case of the multivariate delta method, note that if T_n and parameter θ are real valued, then $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = g'(\theta)$.

Theorem 3.22: the Multivariate Delta Method. If

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

then

$$\sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{D} N_d(\boldsymbol{0}, \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}\boldsymbol{\Sigma}\boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})})$$

where the $d \times k$ Jacobian matrix of partial derivatives

$$\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} g_1(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial \theta_k} g_1(\boldsymbol{\theta}) \\ \vdots & & \vdots \\ \frac{\partial}{\partial \theta_1} g_d(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial \theta_k} g_d(\boldsymbol{\theta}) \end{bmatrix}.$$

Here the mapping $\boldsymbol{g} : \Re^k \to \Re^d$ needs to be differentiable in a neighborhood of $\boldsymbol{\theta} \in \Re^k$.

Definition 3.15. If the estimator $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$, then $\boldsymbol{g}(\boldsymbol{T}_n)$ is a **consistent estimator** of $\boldsymbol{g}(\boldsymbol{\theta})$.

Proposition 3.23. If $0 < \delta \leq 1$, $\boldsymbol{X}$ is a random vector, and

$$n^{\delta}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{D} \boldsymbol{X},$$

then $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$.

Theorem 3.24. If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid, $E(\|\boldsymbol{X}\|) < \infty$ and $E(\boldsymbol{X}) = \boldsymbol{\mu}$, then

a) WLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{P} \boldsymbol{\mu}$ and

b) SLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{ae} \boldsymbol{\mu}$.

Theorem 3.25: Continuity Theorem. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors with characteristic function $\phi_n(\boldsymbol{t})$ and let $\boldsymbol{X}$ be a $k \times 1$ random vector with cf $\phi(\boldsymbol{t})$. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \phi_{\mathrm{n}}(\boldsymbol{t}) \to \phi(\boldsymbol{t})$$

for all $\boldsymbol{t} \in \Re^k$.

Theorem 3.26: Cramér Wold Device. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors and let $\boldsymbol{X}$ be a $k \times 1$ random vector. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \boldsymbol{t}^{\mathrm{T}}\boldsymbol{X}_{\mathrm{n}} \xrightarrow{\mathrm{D}} \boldsymbol{t}^{\mathrm{T}}\boldsymbol{X}$$

for all $\boldsymbol{t} \in \Re^k$.

Theorem 3.27: a) If $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, then $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$.

b)

$$\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta}) \text{ iff } \boldsymbol{X}_{\mathrm{n}} \xrightarrow{\mathrm{D}} \boldsymbol{g}(\boldsymbol{\theta}).$$

Let $g(n) \geq 1$ be an increasing function of the sample size n: $g(n) \uparrow \infty$, eg $g(n) = \sqrt{n}$. See White (1984, p. 15). If a $k \times 1$ random vector $\boldsymbol{T}_n - \boldsymbol{\mu}$ converges to a nondegenerate multivariate normal distribution with convergence rate $\sqrt{n}$, then $\boldsymbol{T}_n$ has (tightness) rate $\sqrt{n}$.

Definition 3.16. Let $\boldsymbol{A}_n = [a_{i,j}(n)]$ be an $r \times c$ random matrix.

a) $\boldsymbol{A}_n = O_P(X_n)$ if $a_{i,j}(n) = O_P(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.

b) $\boldsymbol{A}_n = o_p(X_n)$ if $a_{i,j}(n) = o_p(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.

c) $\boldsymbol{A}_n \asymp_P (1/(g(n))$ if $a_{i,j}(n) \asymp_P (1/(g(n))$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.

d) Let $\boldsymbol{A}_{1,n} = \boldsymbol{T}_n - \boldsymbol{\mu}$ and $\boldsymbol{A}_{2,n} = \boldsymbol{C}_n - c\boldsymbol{\Sigma}$ for some constant $c > 0$. If

$\boldsymbol{A}_{1,n} \asymp_P (1/(g(n))$ and $\boldsymbol{A}_{2,n} \asymp_P (1/(g(n))$, then $(\boldsymbol{T}_n, \boldsymbol{C}_n)$ has (tightness) rate $g(n)$.

Recall that the smallest integer function $\lceil x \rceil$ rounds up, eg $\lceil 7.7 \rceil = 8$.

Theorem 3.28: Continuous Mapping Theorem. Let $\boldsymbol{X}_n \in \Re^k$. If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and if the function $\boldsymbol{g} : \Re^k \to \Re^j$ is continuous, then $\boldsymbol{g}(\boldsymbol{X}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{X})$.

The following two theorems are taken from Severini (2005, p. 345-349, 354).

Theorem 3.29: Let $\boldsymbol{X}_n = (X_{1n}, ..., X_{kn})^T$ be a sequence of $k \times 1$ random vectors, let $\boldsymbol{Y}_n$ be a sequence of $k \times 1$ random vectors and let $\boldsymbol{X} = (X_1, ..., X_k)^T$ be a $k \times 1$ random vector. Let $\boldsymbol{W}_n$ be a sequence of $k \times k$ nonsingular random matrices and let $\boldsymbol{C}$ be a $k \times k$ constant nonsingular matrix.

a) $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$ iff $X_{in} \xrightarrow{P} X_i$ for $i = 1, ..., k$.

b) **Slutsky's Theorem**: If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{Y}_n \xrightarrow{P} \boldsymbol{c}$ for some constant $k \times 1$ vector $\boldsymbol{c}$, then i) $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow{D} \boldsymbol{X} + \boldsymbol{c}$ and

ii) $\boldsymbol{Y}_n^T \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{c}^T \boldsymbol{X}$.

c) If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{W}_n \xrightarrow{D} \boldsymbol{C}$, then $\boldsymbol{W}_n \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}\boldsymbol{X}$, $\boldsymbol{X}_n^T \boldsymbol{W}_n \xrightarrow{D} \boldsymbol{X}^T \boldsymbol{C}$, $\boldsymbol{W}_n^{-1} \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}^{-1} \boldsymbol{X}$ and $\boldsymbol{X}_n^T \boldsymbol{W}_n^{-1} \xrightarrow{D} \boldsymbol{X}^T \boldsymbol{C}^{-1}$.

Theorem 3.30: Let W_n, X_n, Y_n and Z_n be sequences of random variables such that $Y_n > 0$ and $Z_n > 0$. (Often Y_n and Z_n are deterministic, eg $Y_n = n^{-1/2}$.)

a) If $W_n = O_P(1)$ and $X_n = O_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = O_P(1)$, thus $O_P(1) + O_P(1) = O_P(1)$ and $O_P(1) O_P(1) = O_P(1)$.

b) If $W_n = O_P(1)$ and $X_n = o_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = o_P(1)$, thus $O_P(1) + o_P(1) = O_P(1)$ and $O_P(1) o_P(1) = o_P(1)$.

c) If $W_n = O_P(Y_n)$ and $X_n = O_P(Z_n)$, then $W_n + X_n = O_P(\max(Y_n, Z_n))$ and $W_n X_n = O_P(Y_n Z_n)$, thus $O_P(Y_n) + O_P(Z_n) = O_P(\max(Y_n, Z_n))$ and $O_P(Y_n) O_P(Z_n) = O_P(Y_n Z_n)$.

Theorem 3.31. i) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A}\boldsymbol{\mu}) \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

ii) If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^δ where $s > 0$

is some constant and $0 < \delta \leq 0.5$, then $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) =$

$$s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta}).$$

iii) If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ where $s > 0$ is some constant, then $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_P(1)$, so $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a consistent estimator of $s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

iv) Let $\boldsymbol{\Sigma} > 0$. If $\sqrt{n}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and if $\boldsymbol{C}$ is a consistent estimator of $\boldsymbol{\Sigma}$, then $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}(T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. In particular,
$n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$.

Proof: ii) $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) =$
$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$
$= (\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu}) + (\boldsymbol{x} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - T)$
$+(\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu} - T) + (\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu})$
$+(\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta})$.

iii) Following the proof for ii), $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_P(1)$. Alternatively, $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a continuous function of $(T, \boldsymbol{C})$ if $\boldsymbol{C} > 0$ for $n > 10p$. Hence $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) \xrightarrow{P} D^2_{\boldsymbol{x}}(\mu, s\boldsymbol{\Sigma})$.

iv) Note that $\boldsymbol{Z}_n = \sqrt{n}\ \boldsymbol{\Sigma}^{-1/2}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Thus $\boldsymbol{Z}_n^T \boldsymbol{Z}_n = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. Now $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}(T - \boldsymbol{\mu}) =$
$n(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) +$
$n(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) + o_P(1) \xrightarrow{D} \chi^2_p$ since
$\sqrt{n}(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}] \sqrt{n}(T - \boldsymbol{\mu}) = O_P(1) o_P(1) O_P(1) = o_P(1)$.

3.5 Summary

1) If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $p \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}),\ E(\boldsymbol{a} + \boldsymbol{Y}) = \boldsymbol{a} + E(\boldsymbol{Y}),\ \&\ E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}.$$

Also

$$\mathrm{Cov}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}) = \mathrm{Cov}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{X})\boldsymbol{A}^T.$$

Note that $E(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}E(\boldsymbol{Y})$ and $\mathrm{Cov}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{Y})\boldsymbol{A}^T$.

2) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

3) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{AX} \sim N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants, then $\boldsymbol{X} + \boldsymbol{a} \sim N_p(\boldsymbol{\mu} + \boldsymbol{a}, \boldsymbol{\Sigma})$. See Q2, HW2 E.

$$\text{Let } \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

4) **All subsets of a MVN are MVN:** $(X_{k_1}, ..., X_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\boldsymbol{\mu}}_i = E(X_{k_i})$ and $\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}(X_{k_i}, X_{k_j})$. In particular, $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent iff $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

5)

$$\text{Let } \begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \text{Cov}(Y, X) \\ \text{Cov}(X, Y) & \sigma_X^2 \end{pmatrix} \right).$$

Also recall that the *population correlation* between X and Y is given by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

if $\sigma_X > 0$ and $\sigma_Y > 0$.

6) The conditional distribution of a MVN is MVN. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. That is,

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

7) Notation:

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

8) Be able to compute the above quantities if X_1 and X_2 are scalars.

9) A $p \times 1$ random vector $\boldsymbol{X}$ has an *elliptically contoured distribution,* if $\boldsymbol{X}$ has density

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})], \tag{3.16}$$

and we say $\boldsymbol{X}$ has an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution. If the second moments exist, then

$$E(\boldsymbol{X}) = \boldsymbol{\mu} \tag{3.17}$$

and

$$\mathrm{Cov}(\boldsymbol{X}) = c_X \boldsymbol{\Sigma} \tag{3.18}$$

for some constant $c_X > 0$.

10) The *population squared Mahalanobis distance*

$$U \equiv D^2 = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}). \tag{3.19}$$

For elliptically contoured distributions, U has pdf

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{3.20}$$

$U \sim \chi^2_p$ if $\boldsymbol{x}$ has a multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution.

11) The classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ of multivariate location and dispersion is the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \quad \text{and} \quad \boldsymbol{S} = \frac{1}{\mathrm{n}-1} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\boldsymbol{x}_\mathrm{i} - \overline{\boldsymbol{x}})(\boldsymbol{x}_\mathrm{i} - \overline{\boldsymbol{x}})^\mathrm{T}.$$

12) Let the $p \times 1$ column vector $T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{W})$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \tag{3.21}$$

for each observation $\boldsymbol{x}_i$. Notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$. The classical Mahalanobis distance uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$.

13) If p random variables come from an elliptically contoured distribution, then the subplots in the scatterplot matrix should be linear.

14) Let $\boldsymbol{X}_n$ be a sequence of random vectors with joint cdfs $F_n(\boldsymbol{x})$ and let $\boldsymbol{X}$ be a random vector with joint cdf $F(\boldsymbol{x})$.

a) $\boldsymbol{X}_n$ **converges in distribution** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, if $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ as $n \to \infty$ for all points $\boldsymbol{x}$ at which $F(\boldsymbol{x})$ is continuous. The distribution of $\boldsymbol{X}$ is the **limiting distribution** or **asymptotic distribution** of $\boldsymbol{X}_n$.

b) $\boldsymbol{X}_n$ **converges in probability** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \to 0$ as $n \to \infty$.

15) Multivariate Central Limit Theorem (MCLT): If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$, then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$$

where the sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

16) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A}\boldsymbol{\mu}) \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

17) Suppose $\boldsymbol{A}$ is a conformable constant matrix and $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$. Then $\boldsymbol{A}\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{A}\boldsymbol{X}$.

3.6 Complements

Johnson and Wichern (1988) and Mardia, Kent and Bibby (1979) are good references for multivariate statistical analysis based on the multivariate normal distribution. The elliptically contoured distributions generalize the multivariate normal distribution and are discussed (in increasing order of difficulty) in Johnson (1987), Fang, Kotz and Ng (1990), Fang and Anderson (1990), and Gupta and Varga (1993). Fang, Kotz and Ng (1990) sketch the history of elliptically contoured distributions while Gupta and Varga (1993) discuss matrix valued elliptically contoured distributions. Cambanis, Huang and Simons (1981), Chmielewski (1981) and Eaton (1986) are also important references. Also see Muirhead (1982, p. 30–42).

There are several PhD level texts on large sample theory including, in roughly increasing order of difficulty, Lehmann (1999), Ferguson (1996), Sen and Singer (1993), and Serfling (1980). Cramér (1946) is also an important reference, and White (1984) considers asymptotic theory for econometric applications. Also see DasGupta (2008), Davidson (1994), Jiang (2010), Polansky (2011), Sen, Singer and Pedrosa De Lima (2010) and van der Vaart (1998). Section 3.4 followed Olive (2012b, ch. 8) closely.

In analysis, convergence in probability is a special case of convergence in measure and convergence in distribution is a special case of weak convergence.

See Ash (1972, p. 322) and Sen and Singer (1993, p. 39). Almost sure convergence is also known as strong convergence. See Sen and Singer (1993, p. 34). Since $\overline{Y} \xrightarrow{P} \mu$ iff $\overline{Y} \xrightarrow{D} \mu$, the WLLN refers to weak convergence. Technically the X_n and X need to share a common probability space for convergence in probability and almost sure convergence.

3.7 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

3.1*. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 49 \\ 100 \\ 17 \\ 7 \end{pmatrix}, \begin{pmatrix} 3 & 1 & -1 & 0 \\ 1 & 6 & 1 & -1 \\ -1 & 1 & 4 & 0 \\ 0 & -1 & 0 & 2 \end{pmatrix} \right).$$

a) Find the distribution of X_2.

b) Find the distribution of $(X_1, X_3)^T$.

c) Which pairs of random variables X_i and X_j are independent?

d) Find the correlation $\rho(X_1, X_3)$.

3.2*. Recall that if $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

Let $\sigma_{12} = \text{Cov}(Y, X)$ and suppose Y and X follow a bivariate normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 49 \\ 100 \end{pmatrix}, \begin{pmatrix} 16 & \sigma_{12} \\ \sigma_{12} & 25 \end{pmatrix} \right).$$

a) If $\sigma_{12} = 0$, find $Y|X$. Explain your reasoning.

b) If $\sigma_{12} = 10$ find $E(Y|X)$.

c) If $\sigma_{12} = 10$, find $\text{Var}(Y|X)$.

3.3. Let $\sigma_{12} = \text{Cov}(Y, X)$ and suppose Y and X follow a bivariate normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 15 \\ 20 \end{pmatrix}, \begin{pmatrix} 64 & \sigma_{12} \\ \sigma_{12} & 81 \end{pmatrix} \right).$$

a) If $\sigma_{12} = 10$ find $E(Y|X)$.

b) If $\sigma_{12} = 10$, find $\text{Var}(Y|X)$.

c) If $\sigma_{12} = 10$, find $\rho(Y, X)$, the correlation between Y and X.

3.4. Suppose that

$$\boldsymbol{X} \sim (1-\gamma)EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g_1) + \gamma EC_p(\boldsymbol{\mu}, c\boldsymbol{\Sigma}, g_2)$$

where $c > 0$ and $0 < \gamma < 1$. Following Example 3.2, show that $\boldsymbol{X}$ has an elliptically contoured distribution assuming that all relevant expectations exist.

3.5. In Proposition 3.5b, show that if the second moments exist, then $\boldsymbol{\Sigma}$ can be replaced by $\text{Cov}(\boldsymbol{X})$.

crancap	hdlen	hdht	Data for 3.6
1485	175	132	
1450	191	117	
1460	186	122	
1425	191	125	
1430	178	120	
1290	180	117	
90	75	51	

3.6*. The table ($\boldsymbol{W}$) above represents 3 head measurements on 6 people and one ape. Let X_1 = *cranial capacity*, X_2 = *head length* and X_3 = *head height*. Let $\boldsymbol{x} = (X_1, X_2, X_3)^T$. Several multivariate location estimators, including the coordinatewise median and sample mean, are found by applying a univariate location estimator to each random variable and then collecting the results into a vector. a) Find the coordinatewise median MED($\boldsymbol{W}$).

b) Find the sample mean $\overline{\boldsymbol{x}}$.

3.7. Using the notation in Proposition 3.6, show that if the second moments exist, then

$$\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY} = [\mathrm{Cov}(\boldsymbol{X})]^{-1}\mathrm{Cov}(\boldsymbol{X}, Y).$$

3.8. Using the notation under Lemma 3.4, show that if $\boldsymbol{X}$ is elliptically contoured, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is also elliptically contoured.

3.9*. Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. Find the distribution of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ if $\boldsymbol{X}$ is an $n \times p$ full rank constant matrix and $\boldsymbol{\beta}$ is a $p \times 1$ constant vector.

3.10. Recall that $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = E[(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$. Using the notation of Proposition 3.6, let $(Y, \boldsymbol{X}^T)^T$ be $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ where Y is a random variable. Let the covariance matrix of $(Y, \boldsymbol{X}^T)$ be

$$\mathrm{Cov}((Y, \boldsymbol{X}^T)^T) = c\begin{pmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{pmatrix} = \begin{pmatrix} \mathrm{VAR}(Y) & \mathrm{Cov}(Y, \boldsymbol{X}) \\ \mathrm{Cov}(\boldsymbol{X}, Y) & \mathrm{Cov}(X) \end{pmatrix}$$

where c is some positive constant. Show that $E(Y|\boldsymbol{X}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{X}$ where

$$\alpha = \mu_Y - \boldsymbol{\beta}^T\boldsymbol{\mu}_X \quad \text{and}$$

$$\boldsymbol{\beta} = [\mathrm{Cov}(\boldsymbol{X})]^{-1}\mathrm{Cov}(\boldsymbol{X}, Y).$$

3.11. (Due to R.D. Cook.) Let $\boldsymbol{X}$ be a $p \times 1$ random vector with $E(\boldsymbol{X}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Let $\boldsymbol{B}$ be any constant full rank $p \times r$ matrix where $1 \leq r \leq p$. Suppose that for all such conforming matrices $\boldsymbol{B}$,

$$E(\boldsymbol{X}|\boldsymbol{B}^T\ \boldsymbol{X}) = \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{X}$$

where $\boldsymbol{M}_B$ a $p \times r$ constant matrix that depend on $\boldsymbol{B}$.

Using the fact that $\boldsymbol{\Sigma B} = \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{B}^{\mathrm{T}}\boldsymbol{X}) = \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{B}) = \mathrm{E}[\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{B}|\boldsymbol{B}^{\mathrm{T}}\boldsymbol{X})]$, compute $\boldsymbol{\Sigma B}$ and show that $\boldsymbol{M}_B = \boldsymbol{\Sigma B}(\boldsymbol{B}^T\boldsymbol{\Sigma B})^{-1}$. Hint: what acts as a constant in the inner expectation?

3.12. Let $\boldsymbol{x}$ be a $p \times 1$ random vector with covariance matrix $\mathrm{Cov}(\boldsymbol{x})$. Let $\boldsymbol{A}$ be an $r \times p$ constant matrix and let $\boldsymbol{B}$ be a $q \times p$ constant matrix. Find $\mathrm{Cov}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{B}\boldsymbol{x})$ in terms of $\boldsymbol{A}, \boldsymbol{B}$ and $\mathrm{Cov}(\boldsymbol{x})$.

3.13. The table $\boldsymbol{W}$ shown below represents 4 measurements on 5 people.

age	breadth	cephalic	size
39.00	149.5	81.9	3738
35.00	152.5	75.9	4261
35.00	145.5	75.4	3777
19.00	146.0	78.1	3904
0.06	88.5	77.6	933

a) Find the sample mean $\overline{\boldsymbol{x}}$.

b) Find the coordinatewise median MED($\boldsymbol{W}$).

3.14. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors from a multivariate t-distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with d degrees of freedom. Then $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{x}) = \dfrac{d}{d-2}\boldsymbol{\Sigma}$ for $d > 2$. Assuming $d > 2$, find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

3.15. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 9 \\ 16 \\ 4 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 & -0.4 & 0 \\ 0.8 & 1 & -0.56 & 0 \\ -0.4 & -0.56 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \right).$$

a) Find the distribution of X_3.

b) Find the distribution of $(X_2, X_4)^T$.

c) Which pairs of random variables X_i and X_j are independent?

d) Find the correlation $\rho(X_1, X_3)$.

3.16. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors where

$$\boldsymbol{x}_i \sim (1-\gamma)N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \gamma N_p(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$$

with $0 < \gamma < 1$ and $c > 0$. Then $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{x}_i) = [1 + \gamma(c-1)]\boldsymbol{\Sigma}$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

Let $\boldsymbol{X}$ be an $n \times p$ constant matrix and let $\boldsymbol{\beta}$ be a $p \times 1$ constant vector. Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. Find the distribution of $\boldsymbol{HY}$ if $\boldsymbol{H}^T = \boldsymbol{H} = \boldsymbol{H}^2$ is an $n \times n$ matrix and if $\boldsymbol{HX} = \boldsymbol{X}$. Simplify.

3.17. Recall that if $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix
$\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Let Y and X follow a bivariate normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2\left(\begin{pmatrix} 134 \\ 96 \end{pmatrix}, \begin{pmatrix} 24.5 & 1.1 \\ 1.1 & 23.0 \end{pmatrix} \right).$$

a) Find $E(Y|X)$.

b) Find $\mathrm{Var}(Y|X)$.

3.18. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4\left(\begin{pmatrix} 1 \\ 7 \\ 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 0 & 2 & 1 \\ 0 & 1 & 0 & 0 \\ 2 & 0 & 3 & 1 \\ 1 & 0 & 1 & 5 \end{pmatrix} \right).$$

a) Find the distribution of $(X_1, X_4)^T$.

b) Which pairs of random variables X_i and X_j are independent?

c) Find the correlation $\rho(X_1, X_4)$.

3.19. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4\left(\begin{pmatrix} 3 \\ 4 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 3 & 2 & 1 & 1 \\ 2 & 4 & 1 & 0 \\ 1 & 1 & 2 & 0 \\ 1 & 0 & 0 & 3 \end{pmatrix} \right).$$

a) Find the distribution of $(X_1, X_3)^T$.

b) Which pairs of random variables X_i and X_j are independent?

c) Find the correlation $\rho(X_1, X_3)$.

3.20. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors where $E(\boldsymbol{x}_i) = e^{0.5}\mathbf{1}$ and
$\text{Cov}(\boldsymbol{x}_i) = (e^2 - e)\boldsymbol{I}_p$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

3.21. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 49 \\ 25 \\ 9 \\ 4 \end{pmatrix}, \begin{pmatrix} 2 & -1 & 3 & 0 \\ -1 & 5 & -3 & 0 \\ 3 & -3 & 5 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} \right).$$

a) Find the distribution of $(X_1, X_3)^T$.

b) Which pairs of random variables X_i and X_j are independent?

c) Find the correlation $\rho(X_1, X_3)$.

3.22. Recall that if $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix
$\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Let Y and X follow a bivariate normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 49 \\ 17 \end{pmatrix}, \begin{pmatrix} 3 & -1 \\ -1 & 4 \end{pmatrix} \right).$$

a) Find $E(Y|X)$.

b) Find $\text{Var}(Y|X)$.

3.23. Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. Find the distribution of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ if $\boldsymbol{X}$ is an $n \times p$ full rank constant matrix and $\boldsymbol{\beta}$ is a $p \times 1$ constant vector. Simplify.

3.24. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid 2×1 random vectors from a multivariate lognormal
LN($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) distribution. Let $\boldsymbol{x}_i = (X_{i1}, X_{i2})^T$. Following Press (2005, p. 149-150),

$E(X_{ij}) = \exp(\mu_j + \sigma_j^2/2)$, $V(X_{ij}) = \exp(\sigma_j^2)[\exp(\sigma_j^2) - 1]\exp(2\mu_j)$ for $j = 1, 2$, and
$\mathrm{Cov}(X_{i1}, X_{i2}) = exp[\mu_1 + \mu_2 + 0.5(\sigma_1^2 + \sigma_2^2) + \sigma_{12}][exp(\sigma_{12}) - 1]$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

Chapter 4

MLD Estimators

Let $\boldsymbol{\mu}$ be a $p \times 1$ location vector and $\boldsymbol{\Sigma}$ a $p \times p$ symmetric dispersion matrix. Because of symmetry, the first row of $\boldsymbol{\Sigma}$ has p distinct unknown parameters, the second row has $p-1$ distinct unknown parameters, the third row has $p-2$ distinct unknown parameters, ..., and the pth row has one distinct unknown parameter for a total of $1+2+\cdots+p = p(p+1)/2$ unknown parameters. Since $\boldsymbol{\mu}$ has p unknown parameters, an estimator $(T, \boldsymbol{C})$ of multivariate location and dispersion (MLD), needs to estimate $p(p+3)/2$ unknown parameters when there are p random variables. If the p variables can be transformed into an uncorrelated set then there are only $2p$ parameters, the means and variances, while if the dimension can be reduced from p to $p-1$, the number of parameters is reduced by $p(p+3)/2 - (p-1)(p+2)/2 = p-1$.

The sample covariance or sample correlation matrices estimate these parameters very efficiently since $\boldsymbol{\Sigma} = ((\sigma_{ij}))$ where σ_{ij} is a population covariance or correlation. These quantities can be estimated with the sample covariance or correlation taking two variables X_i and X_j at a time. Note that there are $p(p+1)/2$ pairs that can be chosen from p random variables $X_1, ..., X_p$.

Rule of thumb 4.1. For the classical estimators of multivariate location and dispersion, $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ or $(\overline{\boldsymbol{z}}, \boldsymbol{R})$, want $n > 10p$. Want $n > 20p$ for the robust MLD estimators (FCH, RFCH or RMVN) described later in this chapter.

4.1 Affine Equivariance

Before defining an important equivariance property, some notation is needed. Again assume that the data is collected in an $n \times p$ data matrix $\boldsymbol{W}$. Let

$\boldsymbol{B} = \mathbf{1}\boldsymbol{b}^T$ where $\mathbf{1}$ is an $n \times 1$ vector of ones and $\boldsymbol{b}$ is a $p \times 1$ constant vector. Hence the ith row of $\boldsymbol{B}$ is $\boldsymbol{b}_i^T \equiv \boldsymbol{b}^T$ for $i = 1, ..., n$. For such a matrix $\boldsymbol{B}$, consider the affine transformation $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A} + \boldsymbol{B}$ where $\boldsymbol{A}$ is any nonsingular $p \times p$ matrix.

Definition 4.1. Then the multivariate location and dispersion estimator $(T, \boldsymbol{C})$ is *affine equivariant* if

$$T(\boldsymbol{Z}) = T(\boldsymbol{W}\boldsymbol{A} + \boldsymbol{B}) = \boldsymbol{A}^T T(\boldsymbol{W}) + \boldsymbol{b}, \tag{4.1}$$

and

$$\boldsymbol{C}(\boldsymbol{Z}) = \boldsymbol{C}(\boldsymbol{W}\boldsymbol{A} + \boldsymbol{B}) = \boldsymbol{A}^T \boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}. \tag{4.2}$$

The following proposition shows that the Mahalanobis distances are invariant under affine transformations. See Rousseeuw and Leroy (1987, p. 252-262) for similar results. Thus if $(T, \boldsymbol{C})$ is affine equivariant, so is $(T, D^2_{(c_n)}(T, \boldsymbol{C})\ \boldsymbol{C})$ where $D^2_{(j)}(T, \boldsymbol{C})$ is the jth order statistic of the D_i^2.

Proposition 4.1. If $(T, \boldsymbol{C})$ is affine equivariant, then

$$D_i^2(\boldsymbol{W}) \equiv D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) =$$
$$D_i^2(T(\boldsymbol{Z}), \boldsymbol{C}(\boldsymbol{Z})) \equiv D_i^2(\boldsymbol{Z}). \tag{4.3}$$

Proof. Since $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A} + \boldsymbol{B}$ has ith row

$$\boldsymbol{z}_i^T = \boldsymbol{x}_i^T \boldsymbol{A} + \boldsymbol{b}^T,$$

$$D_i^2(\boldsymbol{Z}) = [\boldsymbol{z}_i - T(\boldsymbol{Z})]^T \boldsymbol{C}^{-1}(\boldsymbol{Z})[\boldsymbol{z}_i - T(\boldsymbol{Z})]$$
$$= [\boldsymbol{A}^T(\boldsymbol{x}_i - T(\boldsymbol{W}))]^T [\boldsymbol{A}^T \boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}]^{-1} [\boldsymbol{A}^T(\boldsymbol{x}_i - T(\boldsymbol{W}))]$$
$$= [\boldsymbol{x}_i - T(\boldsymbol{W})]^T \boldsymbol{C}^{-1}(\boldsymbol{W})[\boldsymbol{x}_i - T(\boldsymbol{W})] = D_i^2(\boldsymbol{W}).\ QED$$

Warning: Estimators that use randomly chosen elemental sets or projections are not affine equivariant since these estimators change every time they are computed. Such estimators can sometimes be made affine equivariant by using the same fixed random number seed each time the estimator is used. Then the affine equivariance of the estimator depends on the random number seed, and such estimators are not as attractive as affine equivariant estimators that do not depend on a fixed random number seed.

4.2 Breakdown

This section gives a standard definition of breakdown for estimators of multivariate location and dispersion. The following notation will be useful. Let $\boldsymbol{W}$ denote the $n \times p$ data matrix with ith row $\boldsymbol{x}_i^T$ corresponding to the ith case. Let $\boldsymbol{w}_1, ...\boldsymbol{w}_n$ be the contaminated data after d_n of the $\boldsymbol{x}_i$ have been replaced by arbitrarily bad contaminated cases. Let $\boldsymbol{W}_d^n$ denote the $n \times p$ data matrix with ith row $\boldsymbol{w}_i^T$. Then the contamination fraction is $\gamma_n = d_n/n$. Let $(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ denote an estimator of multivariate location and dispersion where the $p \times 1$ vector $T(\boldsymbol{W})$ is an estimator of location and the $p \times p$ symmetric positive semidefinite matrix $\boldsymbol{C}(\boldsymbol{W})$ is an estimator of dispersion. Recall from Theorem 1.1 that if $\boldsymbol{C}(\boldsymbol{W}_d^n) > 0$, then $\max_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{C}(\boldsymbol{W}_d^n)\boldsymbol{a} = \lambda_1$ and $\min_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{C}(\boldsymbol{W}_d^n)\boldsymbol{a} = \lambda_p$. A high breakdown dispersion estimator $\boldsymbol{C}$ is positive definite if the amount of contamination is less than the breakdown value. Since $\boldsymbol{a}^T \boldsymbol{C}\boldsymbol{a} = \sum_{i=1}^p \sum_{j=1}^p c_{ij} a_i a_j$, the largest eigenvalue λ_1 is bounded as $\boldsymbol{W}_d^n$ varies iff $\boldsymbol{C}(\boldsymbol{W}_d^n)$ is bounded as $\boldsymbol{W}_d^n$ varies.

Definition 4.2. The *breakdown value* of the multivariate location estimator T at $\boldsymbol{W}$ is

$$B(T, \boldsymbol{W}) = \min\{\frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$ and $1 \leq d_n \leq n$. Let $\lambda_1(\boldsymbol{C}(\boldsymbol{W})) \geq \cdots \geq \lambda_p(\boldsymbol{C}(\boldsymbol{W})) \geq 0$ denote the eigenvalues of the dispersion estimator applied to data $\boldsymbol{W}$. The estimator $\boldsymbol{C}$ breaks down if the smallest eigenvalue can be driven to zero or if the largest eigenvalue can be driven to ∞. Hence the *breakdown value* of the dispersion estimator is

$$B(\boldsymbol{C}, \boldsymbol{W}) = \min\{\frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \max[\frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{W}_d^n))}, \lambda_1(\boldsymbol{C}(\boldsymbol{W}_d^n))] = \infty\}.$$

Definition 4.3. Let γ_n be the breakdown value of $(T, \boldsymbol{C})$. *High breakdown (HB) statistics* have $\gamma_n \to 0.5$ as $n \to \infty$ if the (uncontaminated) clean data are in *general position*: no more than p points of the clean data lie on any $(p-1)$-dimensional hyperplane. Estimators are *zero breakdown* if $\gamma_n \to 0$ and *positive breakdown* if $\gamma_n \to \gamma > 0$ as $n \to \infty$.

Note that if the number of outliers is less than the number needed to cause breakdown, then $\|T\|$ is bounded and the eigenvalues are bounded away from 0 and ∞. Also, the bounds do not depend on the outliers but do depend on the estimator $(T, \boldsymbol{C})$ and on the clean data $\boldsymbol{W}$.

The following result shows that a multivariate location estimator T basically "breaks down" if the d outliers can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|)$ arbitrarily large where $\boldsymbol{w}_i^T$ is the ith row of $\boldsymbol{W}_d^n$. Thus a multivariate location estimator T will not break down if T can not be driven out of some ball of (possibly huge) radius r about the origin.

Proposition 4.2. If nonequivariant estimators (that may have a breakdown value of greater than 1/2) are excluded, then a multivariate location estimator has a breakdown value of d_T/n iff d_T is the smallest number of arbitrarily bad cases that can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|)$ arbitrarily large.

Proof. Note that for a fixed data set $\boldsymbol{W}_d^n$ with ith row $\boldsymbol{w}_i$, if the multivariate location estimator $T(\boldsymbol{W}_d^n)$ satisfies $\|T(\boldsymbol{W}_d^n)\| \leq M$ for some constant M, then the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) \leq \max_{i=1,\ldots,n} \|\boldsymbol{x}_i - T(\boldsymbol{W}_d^n)\| \leq \max_{i=1,\ldots,n} \|\boldsymbol{x}_i\| + M$ if $d_n < n/2$. Similarly, if $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) \leq M$ for some constant M, then $\|T(\boldsymbol{W}_d^n)\|$ is bounded if $d_n < n/2$. QED

Since the coordinatewise median $\text{MED}(\boldsymbol{W})$ is a HB estimator of multivariate location, it is also true that a multivariate location estimator T will not break down if T can not be driven out of some ball of radius r about $\text{MED}(\boldsymbol{W})$. Hence $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ is a HB estimator of MLD.

If a high breakdown estimator $(T, \boldsymbol{C}) \equiv (T(\boldsymbol{W}_d^n), \boldsymbol{C}(\boldsymbol{W}_d^n))$ is evaluated on the contaminated data $\boldsymbol{W}_d^n$, then the location estimator T is contained in some ball about the origin of radius r, and $0 < a < \lambda_p \leq \lambda_1 < b$ where the constants a, r and b depend on the clean data and $(T, \boldsymbol{C})$, but not on $\boldsymbol{W}_d^n$ if the number of outliers d_n satisfies $0 \leq d_n \leq n\gamma_n < n/2$ where the breakdown value $\gamma_n \to 0.5$ as $n \to \infty$.

The following lemma will be used to show that if the classical estimator $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is applied to $c_n \approx n/2$ cases contained in a ball about the origin of radius r where r depends on the clean data but not on $\boldsymbol{W}_d^n$, then $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is a high breakdown estimator.

Lemma 4.3. If the classical estimator $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is applied to c_n cases that are contained in some bounded region where $p + 1 \leq c_n \leq n$, then the

maximum eigenvalue λ_1 of $\boldsymbol{S}_B$ is bounded.

Proof. The largest eigenvalue of a $p \times p$ matrix $\boldsymbol{A}$ is bounded above by $p \max |a_{i,j}|$ where $a_{i,j}$ is the (i,j) entry of $\boldsymbol{A}$. See Datta (1995, p. 403). Denote the c_n cases by $\boldsymbol{z}_1, ..., \boldsymbol{z}_{c_n}$. Then the (i,j)th element $a_{i,j}$ of $\boldsymbol{A} = \boldsymbol{S}_B$ is

$$a_{i,j} = \frac{1}{c_n - 1} \sum_{m=1}^{c_n} (z_{i,m} - \overline{z}_i)(z_{j,m} - \overline{z}_j).$$

Hence the maximum eigenvalue λ_1 is bounded. □

The determinant $det(\boldsymbol{S}) = |\boldsymbol{S}|$ of $\boldsymbol{S}$ is known as the *generalized sample variance*. Consider the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq D^2_{(c_n)}\} \tag{4.4}$$

where $D^2_{(c_n)}$ is the c_nth smallest squared Mahalanobis distance based on $(T, \boldsymbol{C})$. This ellipsoid contains the c_n cases with the smallest D^2_i. Suppose $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}_M, b\ \boldsymbol{S}_M)$ is the sample mean and scaled sample covariance matrix applied to some subset of the data where $b > 0$. The classical, RFCH and RMVN estimators satisfy this assumption. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)}.$$

If $h^2 = D^2_{(c_n)}$, then the volume is proportional to the square root of the determinant $|\boldsymbol{S}_M|^{1/2}$, and this volume will be positive unless extreme degeneracy is present among the c_n cases. See Johnson and Wichern (1988, p. 103-104).

4.3 The Concentration Algorithm

Definition 4.4. Consider the subset J_o of $c_n \approx n/2$ observations whose sample covariance matrix has the lowest determinant among all $C(n, c_n)$ subsets of size c_n. Let T_{MCD} and $\boldsymbol{C}_{MCD}$ denote the sample mean and sample covariance matrix of the c_n cases in J_o. Then the *minimum covariance determinant* MCD(c_n) estimator is $(T_{MCD}(\boldsymbol{W}), \boldsymbol{C}_{MCD}(\boldsymbol{W}))$.

The MCD estimator is a high breakdown (HB) estimator, and the value $c_n = \lfloor (n+p+1)/2 \rfloor$ is often used as the default. The MCD estimator is the pair

$$(\hat{\beta}_{LTS}, Q_{LTS}(\hat{\beta}_{LTS})/(c_n - 1))$$

in the location model where LTS stands for the least trimmed sum of squares estimator. The population analog of the MCD estimator is closely related to the ellipsoid of highest concentration that contains $c_n/n \approx$ half of the mass. The MCD estimator is a $\sqrt{n}$ consistent HB estimator for

$$(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$$

where a_{MCD} is some positive constant when the data $\boldsymbol{x}_i$ are elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, and T_{MCD} has a Gaussian limit. See Butler, Davies, and Jhun (1993) and Cator and Lopuhaä (2009, 2010).

Computing robust covariance estimators can be very expensive. For example, to compute the exact MCD(c_n) estimator (T_{MCD}, C_{MCD}), we need to consider the $C(n, c_n)$ subsets of size c_n. Woodruff and Rocke (1994, p. 893) note that if 1 billion subsets of size 101 could be evaluated per second, it would require 10^{33} millenia to search through all $C(200, 101)$ subsets if the sample size $n = 200$.

Hence algorithm estimators will be used to approximate the robust estimators. Elemental sets are the key ingredient for both *basic resampling* and *concentration* algorithms.

Definition 4.5. Suppose that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are $p \times 1$ vectors of observed data. For the multivariate location and dispersion model, an *elemental set* J is a set of $p+1$ cases. An elemental start is the sample mean and sample covariance matrix of the data corresponding to J. In a *concentration algorithm,* let $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ be the jth start (not necessarily elemental) and compute all n Mahalanobis distances $D_i(T_{-1,j}, \boldsymbol{C}_{-1,j})$. At the next iteration, the classical estimator $(T_{0,j}, \boldsymbol{C}_{0,j}) = (\overline{\boldsymbol{x}}_{0,j}, \boldsymbol{S}_{0,j})$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest distances. This iteration can be continued for k steps resulting in the sequence of estimators $(T_{-1,j}, \boldsymbol{C}_{-1,j}), (T_{0,j}, \boldsymbol{C}_{0,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. The result of the iteration $(T_{k,j}, \boldsymbol{C}_{k,j})$ is called the jth *attractor*. If K_n starts are used, then $j = 1, ..., K_n$. The *concentration attractor*, $(T_A, \boldsymbol{C}_A)$, is the attractor chosen by the algorithm. The attractor is used to obtain the final estimator. A common choice is the attractor that has the smallest determinant $det(\boldsymbol{C}_{k,j})$. The *basic resampling*

algorithm estimator is a special case where $k = -1$ so that the attractor is the start: $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j}) = (\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$.

This concentration algorithm is a simplified version of the algorithms given by Rousseeuw and Van Driessen (1999) and Hawkins and Olive (1999). Using $k = 10$ concentration steps often works well.

Proposition 4.4: Rousseeuw and Van Driessen (1999, p. 214). Suppose that the classical estimator $(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ is computed from c_n cases and that the n Mahalanobis distances $D_i \equiv D_i(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ are computed. If $(\overline{\boldsymbol{x}}_{t+1,j}, \boldsymbol{S}_{t+1,j})$ is the classical estimator computed from the c_n cases with the smallest Mahalanobis distances D_i, then $det(\boldsymbol{S}_{t+1,j}) \leq det(\boldsymbol{S}_{t,j})$ with equality iff $(\overline{\boldsymbol{x}}_{t+1,j}, \boldsymbol{S}_{t+1,j}) = (\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$.

Starts that use a consistent initial estimator could be used. K_n is the number starts and k is the number of concentration steps used in the algorithm. Suppose the algorithm estimator uses some criterion to choose an attractor as the final estimator where there are K attractors and K is fixed, eg $K = 500$, so K does not depend on n. A crucial observation is that the theory of the algorithm estimator depends on the theory of the attractors, not on the estimator corresponding to the criterion.

For example, let $(\mathbf{0}, \boldsymbol{I}_p)$ and $(\mathbf{1}, diag(1, 3, ..., p))$ be the high breakdown attractors where $\mathbf{0}$ and $\mathbf{1}$ are the $p \times 1$ vectors of zeroes and ones. If the minimum determinant criterion is used, then the final estimator is $(\mathbf{0}, \boldsymbol{I}_p)$. Although the MCD criterion is used, the algorithm estimator does not have the same properties as the MCD estimator.

Hawkins and Olive (2002) showed that if K randomly selected elemental starts are used with concentration to produce the attractors, then the resulting estimator is inconsistent and zero breakdown if K and k are fixed and free of n. Note that each elemental start can be made to breakdown by changing one case. Hence the breakdown value of the final estimator is bounded by $K/n \to 0$ as $n \to \infty$. Note that the classical estimator computed from h_n randomly drawn cases is an inconsistent estimator unless $h_n \to \infty$ as $n \to \infty$. Thus the classical estimator applied to a randomly drawn elemental set of $h_n \equiv p + 1$ cases is an inconsistent estimator, so the K starts and the K attractors are inconsistent.

This theory shows that the Maronna, Martin and Yohai (2006, p. 198-199) estimators that use $K = 500$ and one concentration step ($k = 0$) are inconsistent and zero breakdown. The following theorem is useful because

it does not depend on the criterion used to choose the attractor. If the algorithm needs to use many attractors to achieve outlier resistance, then the individual attractors have little outlier resistance. Such estimators include elemental concentration algorithms, heuristic and genetic algorithms and projection algorithms. Algorithms where all K of the attractors are inconsistent, such as elemental concentration algorithms that use k concentration steps, are especially untrustworthy. As another example, Stahel Donoho algorithms use randomly chosen projections and the attractor is a weighted mean and covariance matrix computed for each projection. If randomly chosen projections result in inconsistent attractors, then the Stahel Donoho algorithm is likely inconsistent.

Suppose there are K consistent estimators $(T_j, \boldsymbol{C}_j)$ of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ for some constant $a > 0$, each with the same rate n^δ. If $(T_A, \boldsymbol{C}_A)$ is an estimator obtained by choosing one of the K estimators, then $(T_A, \boldsymbol{C}_A)$ is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with rate n^δ by Pratt (1959). See Theorem 3.16.

Theorem 4.5. Suppose the algorithm estimator chooses an attractor as the final estimator where there are K attractors and K is fixed.

i) If all of the attractors are consistent estimators of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$, then the algorithm estimator is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$.

ii) If all of the attractors are consistent estimators of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with the same rate, eg, n^δ where $0 < \delta \leq 0.5$, then the algorithm estimator is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with the same rate as the attractors.

iii) If all of the attractors are high breakdown, then the algorithm estimator is high breakdown.

iv) Suppose the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid and $P(\boldsymbol{x}_i = \boldsymbol{\mu}) < 1$. The elemental basic resampling algorithm estimator ($k = -1$) is inconsistent.

v) The elemental concentration algorithm is zero breakdown.

Proof. i) Choosing from K consistent estimators for $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ results in a consistent estimator for of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$, and ii) follows from Pratt (1959). iii) Let $\gamma_{n,i}$ be the breakdown value of the ith attractor if the clean data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are in general position. The breakdown value γ_n of the algorithm estimator can be no lower than that of the worst attractor: $\gamma_n \geq \min(\gamma_{n,1}, ..., \gamma_{n,K}) \to 0.5$ as $n \to \infty$.

iv) Let $(\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$ be the classical estimator applied to a randomly drawn elemental set. Then $\overline{\boldsymbol{x}}_{-1,j}$ is the sample mean applied to $p+1$ iid cases. Hence $E[\overline{\boldsymbol{x}}_{-1,j}] = E(\boldsymbol{x}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\overline{\boldsymbol{x}}_{-1,j}) = \mathrm{Cov}(\boldsymbol{x})/(p+1) = \boldsymbol{\Sigma}_{\boldsymbol{x}}/(p+1)$ assuming second moments. So the $(\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$ are identically distributed

and inconsistent estimators of $(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Even without second moments, there exists $\epsilon > 0$ such that $P(\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = \delta_\epsilon > 0$ where the probability, ϵ and δ_ϵ do not depend on n since the distribution of $\overline{\boldsymbol{x}}_{-1,j}$ only depends on the distribution of the iid $\boldsymbol{x}_i$, not on n. Then $P(\min_j \|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = P(\text{all } \|\overline{\boldsymbol{x}}_{-1,\text{j}} - \boldsymbol{\mu}\| > \epsilon) \to \delta_\epsilon^{\text{K}} > 0$ as $n \to \infty$ where equality would hold if the $\overline{\boldsymbol{x}}_{-1,j}$ were iid. Hence the "best start" that minimizes $\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\|$ is inconsistent.

v) The classical estimator with breakdown $1/n$ is applied to each elemental start. Hence $\gamma_n \leq K/n \to 0$ as $n \to \infty$. □

Since the FMCD estimator is a zero breakdown elemental concentration algorithm, the Hubert, Rousseeuw and Van Aelst (2008) claim that "MCD can be efficiently computed with the FAST-MCD estimator" is false. Suppose K is fixed, but at least one randomly drawn start is iterated to convergence so that k is not fixed. Then it is not known whether the attractors are inconsistent or consistent estimators, so it is not known whether FMCD is consistent. It is possible to produce consistent estimators if $K \equiv K_n$ is allowed to increase to ∞.

Remark 4.1. Let γ_o be the highest percentage of large outliers that an elemental concentration algorithm can detect reliably. For many data sets,

$$\gamma_o \approx \min(\frac{n - c_n}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h})100\% \tag{4.5}$$

if n is large, $c_n \geq n/2$ and $h = p + 1$.

Equation (4.5) agrees very well with the Rousseeuw and Van Driessen (1999) simulation performed on the hybrid FMCD algorithm that uses both concentration and partitioning. Section 4.4 will provide theory for the useful practical algorithms and will show that there exists a useful class of data sets where the elemental concentration algorithm can tolerate up to 25% massive outliers.

4.4 Theory for Practical Estimators

It is convenient to let the $\boldsymbol{x}_i$ be random vectors for large sample theory, but the $\boldsymbol{x}_i$ are fixed clean observed data vectors when discussing breakdown. This section presents the FCH estimator to be used along with the classical

and FMCD estimators. Recall from Definition 4.5 that a *concentration algorithm* uses K_n *starts* $(T_{0,j}, \boldsymbol{C}_{0,j})$. Each start is refined with k concentration steps, resulting in K_n *attractors* $(T_{k,j}, \boldsymbol{C}_{k,j})$, and the concentration attractor $(T_A, \boldsymbol{C}_A)$ is the attractor that optimizes the criterion.

Concentration algorithms include the *basic resampling algorithm* as a special case with $k = -1$. Using $k = 10$ concentration steps works well, and iterating until convergence is usually fast. The DGK estimator (Devlin, Gnanadesikan and Kettenring 1975, 1981) defined below is one example. Gnanadesikan and Kettenring (1972, p. 94–95) provide a similar algorithm. The DGK estimator is affine equivariant since the classical estimator is affine equivariant and Mahalanobis distances are invariant under affine transformations by Proposition 4.1. This section will show that the Olive (2004) MB estimator is high breakdown estimator and that the DGK estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$, the same quantity estimated by the MCD estimator. Both estimators use the classical estimator computed from $c_n \approx n/2$ cases. The breakdown point of the DGK estimator has been conjectured to be "at most $1/p$." See Rousseeuw and Leroy (1987, p. 254). Gnanadesikan (1977, p. 134) provides an estimator somewhat similar to the MB estimator.

Definition 4.6. The *DGK estimator* $(T_{k,D}, \boldsymbol{C}_{k,D}) = (T_{DGK}, \boldsymbol{C}_{DGK})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start.

Definition 4.7. The *median ball (MB) estimator* $(T_{k,M}, \boldsymbol{C}_{k,M}) = (T_{MB}, \boldsymbol{C}_{MB})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ as the only start where $\text{MED}(\boldsymbol{W})$ is the coordinatewise median. So $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of data closest to $\text{MED}(\boldsymbol{W})$ in Euclidean distance.

The proof of the following theorem implies that a high breakdown estimator $(T, \boldsymbol{C})$ has $\text{MED}(D_i^2) \leq V$ and that the hyperellipsoid $\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \leq D_{(c_n)}^2\}$ that contains c_n of the cases is in some ball about the origin of radius r, where V and r do not depend on the outliers even if the number of outliers is close to $n/2$. Also the attractor of a high breakdown estimator is a high breakdown estimator if the number of concentration steps k is fixed, eg, $k = 10$. The theorem implies that the MB estimator $(T_{MB}, \boldsymbol{C}_{MB})$ is high breakdown.

Theorem 4.6. Suppose $(T, \boldsymbol{C})$ is a high breakdown estimator where $\boldsymbol{C}$ is a symmetric, positive definite $p \times p$ matrix if the contamination proportion

d_n/n is less than the breakdown value. Then the concentration attractor $(T_k, \boldsymbol{C}_k)$ is a high breakdown estimator if the coverage $c_n \approx n/2$ and the data are in general position.

Proof. Following Leon (1986, p. 280), if $\boldsymbol{A}$ is a symmetric positive definite matrix with eigenvalues $\tau_1 \geq \cdots \geq \tau_n$, then for any nonzero vector $\boldsymbol{x}$,

$$0 < \|\boldsymbol{x}\|^2 \ \tau_n \leq \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \leq \|\boldsymbol{x}\|^2 \ \tau_1. \tag{4.6}$$

Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of $\boldsymbol{C}$. By (4.6),

$$\frac{1}{\lambda_1}\|\boldsymbol{x} - T\|^2 \leq (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq \frac{1}{\lambda_n}\|\boldsymbol{x} - T\|^2. \tag{4.7}$$

By (4.7), if the $D^2_{(i)}$ are the order statistics of the $D^2_i(T, \boldsymbol{C})$, then $D^2_{(i)} < V$ for some constant V that depends on the clean data but not on the outliers even if i and d_n are near $n/2$. (Note that $1/\lambda_n$ and $\text{MED}(\|\boldsymbol{x}_i - T\|^2)$ are both bounded for high breakdown estimators even for d_n near $n/2$.)

Following Johnson and Wichern (1988, p. 50, 103), the boundary of the set $\{\boldsymbol{x} | D^2_{\boldsymbol{x}} \leq h^2\} = \{\boldsymbol{x} | (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq h^2\}$ is a hyperellipsoid centered at T with axes of length $2h\sqrt{\lambda_i}$. Hence $\{\boldsymbol{x} | D^2_{\boldsymbol{x}} \leq D^2_{(c_n)}\}$ is contained in some ball about the origin of radius r where r does not depend on the number of outliers even for d_n near $n/2$. This is the set containing the cases used to compute $(T_0, \boldsymbol{C}_0)$. Since the set is bounded, T_0 is bounded and the largest eigenvalue $\lambda_{1,0}$ of $\boldsymbol{C}_0$ is bounded by Lemma 4.3. The determinant $det(\boldsymbol{C}_{MCD})$ of the HB minimum covariance determinant estimator satisfies $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_0) = \lambda_{1,0} \cdots \lambda_{p,0}$, and $\lambda_{p,0} > \inf det(\boldsymbol{C}_{MCD})/\lambda^{p-1}_{1,0} > 0$ where the infinum is over all possible data sets with $n - d_n$ clean cases and d_n outliers. Since these bounds do not depend on the outliers even for d_n near $n/2$, $(T_0, \boldsymbol{C}_0)$ is a high breakdown estimator. Now repeat the argument with $(T_0, \boldsymbol{C}_0)$ in place of $(T, \boldsymbol{C})$ and $(T_1, \boldsymbol{C}_1)$ in place of $(T_0, \boldsymbol{C}_0)$. Then $(T_1, \boldsymbol{C}_1)$ is high breakdown. Repeating the argument iteratively shows $(T_k, \boldsymbol{C}_k)$ is high breakdown. □

The following corollary shows that it is easy to find a subset J of $c_n \approx n/2$ cases such that the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to J is a HB estimator of MLD.

Corollary 4.7. Let J consist of the c_n cases $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|)$. Then the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to J is a HB estimator of MLD.

To investigate the consistency and rate of robust estimators of multivariate location and dispersion, review Definition 3.16.

The following assumption (E1) gives a class of distributions where we can prove that the new robust estimators are $\sqrt{n}$ consistent. Cator and Lopuhaä (2009, 2010) show that MCD is consistent provided that the MCD functional is unique. Distributions where the functional is unique are called "unimodal," and rule out, for example, a spherically symmetric uniform distribution. Theorem 4.8 is crucial for theory and Theorem 4.9 shows that under (E1), both MCD and DGK are estimating $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$.

Assumption (E1): The $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from a "unimodal" $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with nonsingular covariance matrix $\mathrm{Cov}(\boldsymbol{x}_i)$ where g is continuously differentiable with finite 4th moment: $\int (\boldsymbol{x}^T\boldsymbol{x})^2 g(\boldsymbol{x}^T\boldsymbol{x}) d\boldsymbol{x} < \infty$.

Lopuhaä (1999) shows that if a start $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$, then the classical estimator applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where $a, s > 0$ are some constants. Affine equivariance is not used for $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Also, the attractor and the start have the same rate. If the start is inconsistent, then so is the attractor. The constant a depends on $h > 0$, s, p, and on the elliptically contoured distribution, but does not otherwise depend on the consistent start $(T, \boldsymbol{C})$. The weight function $I(D_i^2(T, \boldsymbol{C}) \leq h^2)$ is an indicator that is 1 if $D_i^2(T, \boldsymbol{C}) \leq h^2$ and 0 otherwise.

Theorem 4.8, Lopuhaä (1999). a) If the start $(T, \boldsymbol{C})$ is inconsistent, then so is the attractor.

b) Suppose $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{I}_p)$ with rate n^δ where $s > 0$ and $0 < \delta \leq 0.5$. Assume (E1) holds and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Then the classical estimator $(T_0, \boldsymbol{C}_0)$ applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{I}_p)$ with the same rate n^δ where $a > 0$.

c) Suppose $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where $s > 0$ and $0 < \delta \leq 0.5$. Assume (E1) holds. Then the classical estimator $(T_0, \boldsymbol{C}_0)$ applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with the same rate n^δ where $a > 0$. The constant a depends on the positive constants s, h, p and the elliptically contoured distribution, but does not otherwise depend on the consistent start $(T, \boldsymbol{C})$.

Let $\delta = 0.5$. Applying Theorem 4.8c) iteratively for a fixed number k of

steps produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_j\boldsymbol{\Sigma})$ where the constants $a_j > 0$ depend on s, h, p and the elliptically contoured distribution, but do not otherwise depend on the consistent start $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$.

The 4th moment assumption was used to simplify theory, but likely holds under 2nd moments. Affine equivariance is needed so that the attractor is affine equivariant, but probably is not needed to prove consistency.

Conjecture 4.1. Change the finite 4th moments assumption to a finite 2nd moments in assumption E1). Suppose $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where $s > 0$ and $0 < \delta \leq 0.5$. Then the classical estimator applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with the same rate n^δ where $a > 0$.

Remark 4.2. To see that the Lopuhaä (1999) theory extends to concentration where the weight function uses $h^2 = D^2_{(c_n)}(T, \boldsymbol{C})$, note that $(T, \tilde{\boldsymbol{C}}) \equiv (T, D^2_{(c_n)}(T, \boldsymbol{C})\, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$ where $b > 0$ is derived in (4.9), and weight function $I(D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1)$ is equivalent to the concentration weight function $I(D_i^2(T, \boldsymbol{C}) \leq D^2_{(c_n)}(T, \boldsymbol{C}))$. As noted above Proposition 4.1, $(T, \tilde{\boldsymbol{C}})$ is affine equivariant if $(T, \boldsymbol{C})$ is affine equivariant. Hence Lopuhaä (1999) theory applied to $(T, \tilde{\boldsymbol{C}})$ with $h = 1$ is equivalent to theory applied to affine equivariant $(T, \boldsymbol{C})$ with $h^2 = D^2_{(c_n)}(T, \boldsymbol{C})$.

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^δ where $0 < \delta \leq 0.5$, then $D^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$$

$$= s^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta}). \tag{4.8}$$

Thus the sample percentiles of $D_i^2(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $s^{-1}D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose $c_n/n \to \xi \in (0, 1)$ as $n \to \infty$, and let $D^2_\xi(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the ξth percentile of the population squared distances. Then $D^2_{(c_n)}(T, \boldsymbol{C}) \xrightarrow{P} s^{-1} D^2_\xi(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $b\boldsymbol{\Sigma} = s^{-1} D^2_\xi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) s\boldsymbol{\Sigma} = D^2_\xi(\boldsymbol{\mu}, \boldsymbol{\Sigma})\boldsymbol{\Sigma}$. Thus

$$b = D^2_\xi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{4.9}$$

does not depend on $s > 0$ or $\delta \in (0, 0.5]$. □

Concentration applies the classical estimator to cases with $D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C})$. Let $c_n \approx n/2$ and

$$b = D_{0.5}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

be the population median of the population squared distances. By Remark 4.2, if $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ then $(T, \tilde{\boldsymbol{C}}) \equiv (T, D_{(c_n)}^2(T, \boldsymbol{C})\ \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$, and $D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1$ is equivalent to $D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C}))$. Hence Lopuhaä (1999) theory applied to $(T, \tilde{\boldsymbol{C}})$ with $h = 1$ is equivalent to theory applied to the concentration estimator using the affine equivariant estimator $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$ as the start. Since b does not depend on s, concentration produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where the constant $a > 0$ is the same for $j = 0, 1, ..., k$.

Theorem 4.9 shows that $a = a_{MCD}$ where $\xi = 0.5$. Hence concentration with a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ as a start results in a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with rate n^δ. This result can be applied iteratively for a finite number of concentration steps. Hence DGK is a $\sqrt{n}$ consistent affine equivariant estimator of the same quantity that MCD is estimating. It is not known if the results hold if concentration is iterated to convergence. For multivariate normal data, $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi_p^2$.

Theorem 4.9. Assume that (E1) holds and that $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where the constants $s > 0$ and $0 < \delta \leq 0.5$. Then the classical estimator $(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ computed from the $c_n \approx n/2$ of cases with the smallest distances $D_i(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with the same rate n^δ.

Proof. By Remark 4.1 the estimator is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with rate n^δ. By the remarks above, a will be the same for any consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ and a does not depend on $s > 0$ or $\delta \in (0, 0.5]$. Hence the result follows if $a = a_{MCD}$. The MCD estimator is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ by Butler, Davies and Jhun (1993) and Cator and Lopuhaä (2009, 2010). If the MCD estimator is the start, then it is also the attractor by Rousseeuw and Van Driessen (1999) who show that concentration does not increase the MCD criterion. Hence $a = a_{MCD}$. □

Next we define the new easily computed robust $\sqrt{n}$ consistent FCH estimator, so named since it is fast, consistent and uses a high breakdown attractor. The FCH and MBA estimators use the $\sqrt{n}$ consistent DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$ and the high breakdown MB estimator $(T_{MB}, \boldsymbol{C}_{MB})$ as attractors.

Definition 4.8. Let the "median ball" be the hypersphere containing the "half set" of data closest to MED($\boldsymbol{X}$) in Euclidean distance. The *FCH estimator* uses the MB attractor if the DGK location estimator T_{DGK} is outside of the median ball, and the attractor with the smallest determinant, otherwise. Let $(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_{FCH}, \boldsymbol{C}_{FCH})$ takes $T_{FCH} = T_A$ and

$$\boldsymbol{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{4.10}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with p degrees of freedom.

Remark 4.3. The *MBA estimator* $(T_{MBA}, \boldsymbol{C}_{MBA})$ uses the attractor $(T_A, \boldsymbol{C}_A)$ with the smallest determinant. Hence the DGK estimator is used as the attractor if $det(\boldsymbol{C}_{DGK}) \leq det(\boldsymbol{C}_{MB})$, and the MB estimator is used as the attractor, otherwise. Then $T_{MBA} = T_A$ and $\boldsymbol{C}_{MBA}$ is computed using the right hand side of (4.10). The difference between the FCH and MBA estimators is that the FCH estimator also uses a location criterion to choose the attractor: if the DGK location estimator T_{DGK} has a greater Euclidean distance from MED($\boldsymbol{W}$) than half the data, then FCH uses the MB attractor. The FCH estimator only uses the attractor with the smallest determinant if $\|T_{DGK} - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p))$. Using the location criterion increases the outlier resistance of the FCH estimator for certain types of outliers, as will be seen in Section 4.5.

The following theorem shows the FCH estimator has good statistical properties. We conjecture that FCH is high breakdown. Note that the location estimator T_{FCH} is high breakdown and that $det(C_{FCH})$ is bounded away from 0 and ∞ if the data is in general position, even if nearly half of the cases are outliers.

Theorem 4.10. T_{FCH} is high breakdown if the clean data are in general position. Suppose (E1) holds. If $(T_A, \boldsymbol{C}_A)$ is the DGK or MB attractor

with the smallest determinant, then $(T_A, \boldsymbol{C}_A)$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. Hence the MBA and FCH estimators are outlier resistant $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c = u_{0.5}/\chi^2_{p,0.5}$, and $c = 1$ for multivariate normal data.

Proof. T_{FCH} is high breakdown since it is a bounded distance from MED($\boldsymbol{W}$) even if the number of outliers is close to $n/2$. Under (E1) the FCH and MBA estimators are asymptotically equivalent since $\|T_{DGK} - \text{MED}(\boldsymbol{W})\| \to 0$ in probability. The estimator satisfies $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_A) \leq det(\boldsymbol{S}_{0,M}) < \infty$ by Theorem 4.6 if up to nearly 50% of the cases are outliers. If the distribution is spherical about $\boldsymbol{\mu}$, then the result follows from Pratt (1959) and Theorem 4.9 since both starts are $\sqrt{n}$ consistent. Otherwise, the MB estimator $\boldsymbol{C}_{MB}$ is a biased estimator of $a_{MCD}\boldsymbol{\Sigma}$. But the DGK estimator $\boldsymbol{C}_{DGK}$ is a $\sqrt{n}$ consistent estimator of $a_{MCD}\boldsymbol{\Sigma}$ by Theorem 4.9 and $\|\boldsymbol{C}_{MCD} - \boldsymbol{C}_{DGK}\| = O_P(n^{-1/2})$. Thus the probability that the DGK attractor minimizes the determinant goes to one as $n \to \infty$, and $(T_A, \boldsymbol{C}_A)$ is asymptotically equivalent to the DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$.

Let $\boldsymbol{C}_F = \boldsymbol{C}_{FCH}$ or $\boldsymbol{C}_F = \boldsymbol{C}_{MBA}$. Let $P(U \leq u_\alpha) = \alpha$ where U is given by (3.9). Then the scaling in (4.10) makes $\boldsymbol{C}_F$ a consistent estimator of $c\boldsymbol{\Sigma}$ where $c = u_{0.5}/\chi^2_{p,0.5}$, and $c = 1$ for multivariate normal data. □

Many variants of the FCH and MBA estimators can be given where the algorithm gives a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$. One such variant uses K starts $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ that are affine equivariant $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, s_j\boldsymbol{\Sigma})$ where $s_j > 0$. The MCD criteria is used to choose the final attractor, and scaling is done as in (4.10). A second variant is the same as the first, but the Kth attractor is replaced by the MB estimator, and for $j < K$ the jth attractor $(T_{k,j}, \boldsymbol{C}_{k,j})$ is not used if $T_{k,j}$ has a greater Euclidean distance from MED($\boldsymbol{X}$) than half the data. Then the location estimator of the algorithm is high breakdown.

Suppose the attractor is $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j})$ computed from a subset of c_n cases. The MCD(c_n) criterion is the determinant $det(\boldsymbol{S}_{k,j})$. The volume of the hyperellipsoid $\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,j})^T \boldsymbol{S}_{k,j}^{-1}(\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,j}) \leq h^2\}$ is equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{S}_{k,j})}, \tag{4.11}$$

see Johnson and Wichern (1988, p. 103-104). The "MVE(c_n)" criterion is $h^p\sqrt{\det(\boldsymbol{S}_{\text{k,j}})}$ where $h = D_{(c_n)}(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j})$ (but does not actually correspond

to the minimum volume ellipsoid (MVE) estimator).

We also considered several estimators that use the MB and DGK estimators as attractors. CMVE is a concentration algorithm like FCH, but the "MVE" criterion is used in place of the MCD criterion. A standard method of reweighting can be used to produce the RMBA, RFCH and RCMVE estimators. RMVN uses a slightly modified method of reweighting so that RMVN gives good estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for multivariate normal data, even when certain types of outliers are present.

Definition 4.9. The *RFCH estimator* uses two standard reweighting steps. Let $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ be the classical estimator applied to the n_1 cases with $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH}) \leq \chi^2_{p,0.975}$, and let

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\text{MED}(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1))}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_1.$$

Then let $(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2)$ be the classical estimator applied to the cases with $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1) \leq \chi^2_{p,0.975}$, and let

$$\boldsymbol{C}_{RFCH} = \frac{\text{MED}(D_i^2(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2))}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_2.$$

RMBA and RFCH are $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ by Lopuhaä (1999) where the weight function uses $h^2 = \chi^2_{p,0.975}$, but the two estimators use nearly 97.5% of the cases if the data is multivariate normal. We conjecture CMVE and RMVE are also $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$.

Definition 4.10. The *RMVN estimator* uses $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ and n_1 as above. Let $q_1 = \min\{0.5(0.975)n/n_1, 0.995\}$, and

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\text{MED}(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1))}{\chi^2_{p,q_1}} \tilde{\boldsymbol{\Sigma}}_1.$$

Then let $(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2)$ be the classical estimator applied to the n_2 cases with $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1)) \leq \chi^2_{p,0.975}$. Let $q_2 = \min\{0.5(0.975)n/n_2, 0.995\}$, and

$$\boldsymbol{C}_{RMVN} = \frac{\text{MED}(D_i^2(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2))}{\chi^2_{p,q_2}} \tilde{\boldsymbol{\Sigma}}_2.$$

Table 4.1: Average Dispersion Matrices for Near Point Mass Outliers

RMVN	FMCD	OGK	MB
$\begin{bmatrix} 1.002 & -0.014 \\ -0.014 & 2.024 \end{bmatrix}$	$\begin{bmatrix} 0.055 & 0.685 \\ 0.685 & 122.5 \end{bmatrix}$	$\begin{bmatrix} 0.185 & 0.089 \\ 0.089 & 36.24 \end{bmatrix}$	$\begin{bmatrix} 2.570 & -0.082 \\ -0.082 & 5.241 \end{bmatrix}$

Table 4.2: Average Dispersion Matrices for Mean Shift Outliers

RMVN	FMCD	OGK	MB
$\begin{bmatrix} 0.990 & 0.004 \\ 0.004 & 2.014 \end{bmatrix}$	$\begin{bmatrix} 2.530 & 0.003 \\ 0.003 & 5.146 \end{bmatrix}$	$\begin{bmatrix} 19.67 & 12.88 \\ 12.88 & 39.72 \end{bmatrix}$	$\begin{bmatrix} 2.552 & 0.003 \\ 0.003 & 5.118 \end{bmatrix}$

The RMVN estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ by Lopuhaä (1999) where the weight function uses $h^2 = \chi^2_{p,0.975}$ and $d = u_{0.5}/\chi^2_{p,q}$ where $q_2 \to q$ in probability as $n \to \infty$. Here $0.5 \leq q < 1$ depends on the elliptically contoured distribution, but $q = 0.5$ and $d = 1$ for multivariate normal data.

If the bulk of the data is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the RMVN estimator can give useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for certain types of outliers where FCH and RFCH estimate $(\boldsymbol{\mu}, d_E\boldsymbol{\Sigma})$ for $d_E > 1$. To see this claim, let $0 \leq \gamma < 0.5$ be the outlier proportion. If $\gamma = 0$, then $n_i/n \xrightarrow{P} 0.975$ and $q_i \xrightarrow{P} 0.5$. If $\gamma > 0$, suppose the outlier configuration is such that the $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH})$ are roughly χ^2_p for the clean cases, and the outliers have larger D_i^2 than the clean cases. Then $\text{MED}(D_i^2) \approx \chi^2_{p,q}$ where $q = 0.5/(1-\gamma)$. For example, if $n = 100$ and $\gamma = 0.4$, then there are 60 clean cases, $q = 5/6$, and the quantile $\chi^2_{p,q}$ is being estimated instead of $\chi^2_{p,0.5}$. Now $n_i \approx n(1-\gamma)0.975$, and q_i estimates q. Thus $\boldsymbol{C}_{RMVN} \approx \boldsymbol{\Sigma}$. Of course consistency cannot generally be claimed when outliers are present.

Simulations suggested $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ gives useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a variety of outlier configurations. Using 20 runs and $n = 1000$, the averages of the dispersion matrices were computed when the bulk of the data are iid $N_2(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = diag(1, 2)$. For clean data, FCH, RFCH and RMVN give $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}$, while FMCD and the Maronna and Zamar (2002) OGK estimator seem to be approximately unbiased for $\boldsymbol{\Sigma}$. The median ball estimator was scaled using (4.10) and estimated $diag(1.13, 1.85)$.

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_2((0, 15)^T, 0.0001\boldsymbol{I}_2)$, a near point mass at the major axis. FCH, MB and RFCH estimated $2.6\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$. FMCD and OGK failed to estimate d $\boldsymbol{\Sigma}$. Note

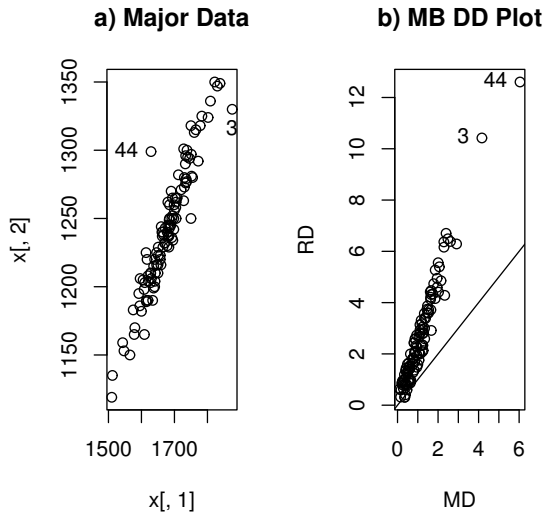

Figure 4.1: Plots for Major Data

that $\chi^2_{2,5/6}/\chi^2_{2,0.5} = 2.585$. See Table 4.1. The following R commands were used where `mldsim` is from `mpack`.

```
qchisq(5/6,2)/qchisq(.5,2) = 2.584963
mldsim(n=1000,p=2,outliers=6,pm=15)
```

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_2((20,20)^T, \boldsymbol{\Sigma})$, a mean shift with the same covariance matrix as the clean cases. Rocke and Woodruff (1996) suggest that outliers with mean shift are hard to detect. FCH, FMCD, MB and RFCH estimated $2.6\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$, and OGK failed. See Table 4.2. The *R command* is shown below.

```
mldsim(n=1000,p=2,outliers=3,pm=20)
```

Example 4.1. Tremearne (1911) recorded *height* = x[,1] and *height while kneeling* = x[,2] of 112 people. Figure 4.1a shows a scatterplot of the data. Case 3 has the largest Euclidean distance of 214.767 from MED($\boldsymbol{W}$) = $(1680, 1240)^T$, but if the distances correspond to the contours of a covering ellipsoid, then case 44 has the largest distance. For $k = 0$, $(\overline{\boldsymbol{x}}_{0,M}, \boldsymbol{S}_{0,M})$ is the classical estimator applied to the "half set" of cases closest to MED($\boldsymbol{W}$) in Euclidean distance. The hypersphere (circle) centered at MED($\boldsymbol{W}$) that

covers half the data is small because the data density is high near MED($\boldsymbol{W}$). The median Euclidean distance is 59.661 and case 44 has Euclidean distance 77.987. Hence the intersection of the sphere and the data is a highly correlated clean ellipsoidal region. Figure 4.1b shows the DD plot of the classical distances versus the MB distances. Notice that both the classical and MB estimators give the largest distances to cases 3 and 44. Notice that case 44 could not be detected using marginal methods.

As the dimension p gets larger, outliers that can not be detected by marginal methods (case 44 in Example 4.1) become harder to detect. When $p = 3$ imagine that the clean data is a baseball bat with one end at the SW corner of the bottom of the box (corresponding to the coordinate axes) and one end at the NE corner of the top of the box. If the outliers are a ball, there is much more room to hide them in the box than in a covering rectangle when $p = 2$.

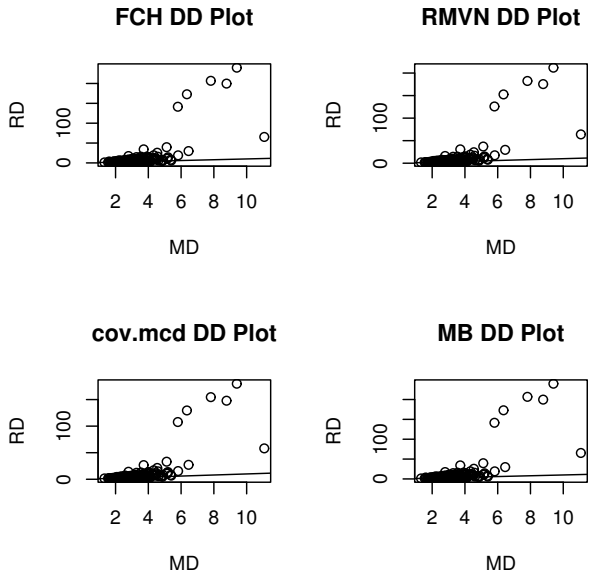

Figure 4.2: DD Plots for Gladstone Data

Example 4.2. The estimators can be useful when the data is not elliptically contoured. The Gladstone (1905-6) data has 11 variables on 267 persons after death. Head measurements were *breadth, circumference, head height, length* and *size* as well as *cephalic index* and *brain weight*. *Age*, *height*

and two categorical variables *ageclass* (0: under 20, 1: 20-45, 2: over 45) and *sex* were also given. Figure 4.2 shows the DD plots for the FCH, RMVN, `cov.mcd` and MB estimators. The DD plots from the DGK, MBA, CMVE, RCMVE and RFCH estimators were similar, and the six outliers in Figure 4.2 correspond to the six infants in the data set.

Chapter 5 shows that if a consistent robust estimator is scaled as in (4.10), then the plotted points in the DD plot will cluster about the identity line with unit slope and zero intercept if the data is multivariate normal, and about some other line through the origin if the data is from some other elliptically contoured distribution with a nonsingular covariance matrix. Since multivariate procedures tend to perform well for elliptically contoured data, the DD plot is useful even if outliers are not present.

4.5 Outlier Resistance and Simulations

Simulations were used to compare $(T_{FCH}, \boldsymbol{C}_{FCH})$, $(T_{RFCH}, \boldsymbol{C}_{RFCH})$, $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $(T_{FMCD}, \boldsymbol{C}_{FMCD})$. Shown below are the averages, using 20 runs and $n = 1000$, of the dispersion matrices when the bulk of the data are iid $N_4(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = diag(1, 2, 3, 4)$. The first pair of matrices used $\gamma = 0$. Here the FCH, RFCH and RMVN estimators are $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}$, while $\boldsymbol{C}_{FMCD}$ seems to be approximately unbiased for $0.94\boldsymbol{\Sigma}$.

```
RMVN                              FMCD
 0.996  0.014  0.002 -0.001      0.931  0.017  0.011 0.000
 0.014  2.012 -0.001  0.029      0.017  1.885 -0.003 0.022
 0.002 -0.001  2.984  0.003      0.011 -0.003  2.803 0.010
-0.001  0.029  0.003  3.994      0.000  0.022  0.010 3.752
```

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_4((0, 0, 0, 15)^T$, $0.0001\ \boldsymbol{I}_4)$, a near point mass at the major axis. FCH and RFCH estimated $1.93\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$. The FMCD estimator failed to estimate $d\ \boldsymbol{\Sigma}$. Note that $\chi^2_{4,5/6}/\chi^2_{4,0.5} = 1.9276$.

```
RMVN                                FMCD
 0.988 -0.023 -0.007  0.021        0.227  -0.016  0.002  0.049
-0.023  1.964 -0.022 -0.002       -0.016   0.435 -0.014  0.0130
```

Table 4.3: Scaled Variance $nS^2(T_p)$ and $nS^2(C_{p,p})$

p	n	V	FCH	RFCH	RMVN	DGK	OGK	CLAS	FMCD	MB
5	50	C	216.0	72.4	75.1	209.3	55.8	47.12	153.9	145.8
5	50	T	12.14	6.50	6.88	10.56	6.70	4.83	8.38	13.23
5	5000	C	307.6	64.1	68.6	325.7	59.3	48.5	60.4	309.5
5	5000	T	18.6	5.34	5.33	19.33	6.61	4.98	5.40	20.20
10	100	C	817.3	276.4	286.0	725.4	229.5	198.9	459.6	610.4
10	100	T	21.40	11.42	11.68	20.13	12.75	9.69	14.05	24.13
10	5000	C	955.5	237.9	243.8	966.2	235.8	202.4	233.6	975.0
10	5000	T	29.12	10.08	10.09	29.35	12.81	9.48	10.06	30.20

```
-0.007 -0.022  3.053  0.007      0.002  -0.014  0.673  0.179
 0.021 -0.002  0.007  3.870      0.049   0.013  0.179 55.648
```

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_4(15\ \mathbf{1}, \boldsymbol{\Sigma})$, a mean shift with the same covariance matrix as the clean cases. Again FCH and RFCH estimated $1.93\boldsymbol{\Sigma}$ while RMVN and FMCD estimated $\boldsymbol{\Sigma}$.

```
RMVN                                FMCD
 1.013  0.008  0.006 -0.026      1.024  0.002  0.003 -0.025
 0.008  1.975 -0.022 -0.016      0.002  2.000 -0.034 -0.017
 0.006 -0.022  2.870  0.004      0.003 -0.034  2.931  0.005
-0.026 -0.016  0.004  3.976     -0.025 -0.017  0.005  4.046
```

If $W_{in} \sim N(0, \tau^2/n)$ for $i = 1, ..., r$ and if S_W^2 is the sample variance of the W_{in}, then $E(nS_W^2) = \tau^2$ and $V(nS_W^2) = 2\tau^4/(r-1)$. So $nS_W^2 \pm \sqrt{5}SE(nS_W^2) \approx \tau^2 \pm \sqrt{10}\tau^2/\sqrt{r-1}$. So for $r = 1000$ runs, expect nS_W^2 to be between $\tau^2 - 0.1\tau^2$ and $\tau^2 + 0.1\tau^2$ with high confidence. Similar results hold for many estimators if W_{in} is $\sqrt{n}$ consistent and asymptotically normal and if n is large enough. If W_{in} has less than $\sqrt{n}$ rate, eg $n^{1/3}$ rate, then the scaled sample variance $nS_W^2 \to \infty$ as $n \to \infty$.

Table 4.3 considers $W = T_p$ and $W = C_{p,p}$ for eight estimators, $p = 5$ and 10 and $n = 10p$ and 5000 when $\boldsymbol{x} \sim N_p(\mathbf{0}, diag(1, ..., p))$. For the classical estimator, denoted by CLAS, $T_p = \overline{x}_p \sim N(0, p/n)$, and $nS^2(T_p) \approx p$

while $C_{p,p}$ is the sample variance of n iid $N(0,p)$ random variables. Hence $nS^2(C_{p,p}) \approx 2p^2$. RFCH, RMVN, FMCD and OGK use a "reweight for efficiency" concentration step that uses a random number of cases with percentage close to 97.5%. These four estimators had similar behavior. DGK, FCH and MB used about 50% of the cases and had similar behavior. By Lopuhaä (1999), estimators with less than $\sqrt{n}$ rate still have zero efficiency after the reweighting. Although FMCD, MB and OGK have not been proven to be $\sqrt{n}$ consistent, their values did not blow up even for $n = 5000$.

Geometrical arguments suggest that the MB estimator has considerable outlier resistance. Suppose the outliers are far from the bulk of the data. Let the "median ball" correspond to the half set of data closest to MED($\boldsymbol{W}$) in Euclidean distance. If the outliers are outside of the median ball, then the initial half set in the iteration leading to the MB estimator will be clean. Thus the MB estimator will tend to give the outliers the largest MB distances unless the initial clean half set has very high correlation in a direction about which the outliers lie. This property holds for very general outlier configurations. The FCH estimator tries to use the DGK attractor if the $det(\boldsymbol{C}_{DGK})$ is small and the DGK location estimator T_{DGK} is in the median ball. Distant outliers that make $det(\boldsymbol{C}_{DGK})$ small also drag T_{DGK} outside of the median ball. Then FCH uses the MB attractor.

Compared to OGK and FMCD, the MB estimator is vulnerable to outliers that lie within the median ball. If the bulk of the data is highly correlated with the major axis of an ellipsoidal region, then the distances based on the clean data can be very large for outliers that fall within the median ball. The outlier resistance of the MB estimator decreases as p increases since the volume of the median ball rapidly increases with p.

A simple simulation for outlier resistance is to count the number of times the minimum distance of the outliers is larger than the maximum distance of the clean cases. The simulation used 100 runs. If the count was 97, then in 97 data sets the outliers can be separated from the clean cases with a horizontal line in the DD plot, but in 3 data sets the robust distances did not achieve complete separation.

The clean cases had $\boldsymbol{x} \sim N_p(\boldsymbol{0}, diag(1, 2, ..., p))$. Outlier types were the mean shift $\boldsymbol{x} \sim N_p(pm\boldsymbol{1}, diag(1, 2, ..., p))$ where $\boldsymbol{1} = (1, ..., 1)^T$, and $\boldsymbol{x} \sim N_p((0, ..., 0, pm)^T, 0.0001\boldsymbol{I}_p)$, a near point mass at the major axis. Notice that the clean data can be transformed to a $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ distribution by multiplying $\boldsymbol{x}_i$ by $\text{diag}(1, 1/\sqrt{2}, ..., 1/\sqrt{p})$, and this transformation changes the location of the near point mass to $(0, ..., 0, pm/\sqrt{p})^T$.

Table 4.4: Number of Times Mean Shift Outliers had the Largest Distances

p	γ	n	pm	MBA	FCH	RFCH	RMVN	OGK	FMCD	MB
10	.1	100	4	49	49	85	84	38	76	57
10	.1	100	5	91	91	99	99	93	98	91
10	.4	100	7	90	90	90	90	0	48	100
40	.1	100	5	3	3	3	3	76	3	17
40	.1	100	8	36	36	37	37	100	49	86
40	.25	100	20	62	62	62	62	100	0	100
40	.4	100	20	20	20	20	20	0	0	100
40	.4	100	35	44	98	98	98	95	0	100
60	.1	200	10	49	49	49	52	100	30	100
60	.1	200	20	97	97	97	97	100	35	100
60	.25	200	25	60	60	60	60	100	0	100
60	.4	200	30	11	21	21	21	17	0	100
60	.4	200	40	21	100	100	100	100	0	100

For near point mass outliers, an ellipsoid with very small volume can cover half of the data if the outliers are at one end of the ellipsoid and some of the clean data are at the other end. This half set will produce a classical estimator with very small determinant by (4.11). In the simulations for large γ, as the near point mass is moved very far away from the bulk of the data, only the classical, MB and OGK estimators did not have numerical difficulties. Since the MCD estimator has smaller determinant than DGK while MVE has smaller volume than DGK, estimators like FMCD and MBA that use the MVE or MCD criterion without using location information will be vulnerable to these outliers. FMCD is also vulnerable to outliers if γ is slightly larger than γ_o given by (4.5).

Tables 4.4 and 4.5 help illustrate the results for the simulation. Large counts and small pm for fixed γ suggest greater ability to detect outliers. Values of p were 5, 10, 15, ..., 60. First consider the mean shift outliers and Table 4.4. For $\gamma = 0.25$ and 0.4, MB usually had the highest counts. For $5 \leq p \leq 20$ and the mean shift, the OGK estimator often had the smallest counts, although FMCD could not handle 40% outliers for $p = 20$. For $25 \leq p \leq 60$, OGK usually had the highest counts for $\gamma = 0.05$ and 0.1. For $p \geq 30$, FMCD could not handle 25% outliers even for enormous values of pm.

Table 4.5: Number of Times Near Point Mass Outliers had the Largest Distances

p	γ	n	pm	MBA	FCH	RFCH	RMVN	OGK	FMCD	MB
10	.1	100	40	73	92	92	92	100	95	100
10	.25	100	25	0	99	99	90	0	0	99
10	.4	100	25	0	100	100	100	0	0	100
40	.1	100	80	0	0	0	0	79	0	80
40	.1	100	150	0	65	65	65	100	0	99
40	.25	100	90	0	88	87	87	0	0	88
40	.4	100	90	0	91	91	91	0	0	91
60	.1	200	100	0	0	0	0	13	0	91
60	.25	200	150	0	100	100	100	0	0	100
60	.4	200	150	0	100	100	100	0	0	100
60	.4	200	20000	0	100	100	100	64	0	100

In Table 4.5, FCH greatly outperformed MBA although the only difference between the two estimators is that FCH uses a location criterion as well as the MCD criterion. OGK performed well for $\gamma = 0.05$ and $20 \leq p \leq 60$ (not tabled). For large γ, OGK often has large bias for $c\boldsymbol{\Sigma}$. Then the outliers may need to be enormous before OGK can detect them. Also see Table 4.2, where OGK gave the outliers the largest distances for all runs, but $\boldsymbol{C}_{OGK}$ does not give a good estimate of $c\boldsymbol{\Sigma} = c \ \ diag(1, 2)$.

The DD plot of MD_i versus RD_i is useful for detecting outliers. The resistant estimator will be useful if $(T, \boldsymbol{C}) \approx (\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c > 0$ since scaling by c affects the vertical labels of the RD_i but not the shape of the DD plot. For the outlier data, the MBA estimator is biased, but the mean shift outliers in the MBA DD plot will have large RD_i since $\boldsymbol{C}_{MBA} \approx 2\boldsymbol{C}_{FMCD} \approx 2\boldsymbol{\Sigma}$.

When p is increased to 8, the `cov.mcd` estimator was usually not useful for detecting the mean shift outliers. Figure 4.3 shows that now the FMCD RD_i are highly correlated with the MD_i. The DD plot based on the MBA estimator detects the outliers. See Figure 4.4.

For many data sets, equation (4.5) gives a rough approximation for the number of large outliers that concentration algorithms using K starts each consisting of h cases can handle. However, if the data set is multivariate and the bulk of the data falls in one compact ellipsoid while the outliers fall in another hugely distant compact ellipsoid, then a concentration algorithm using a single start can sometimes tolerate nearly 25% outliers. For example, sup-

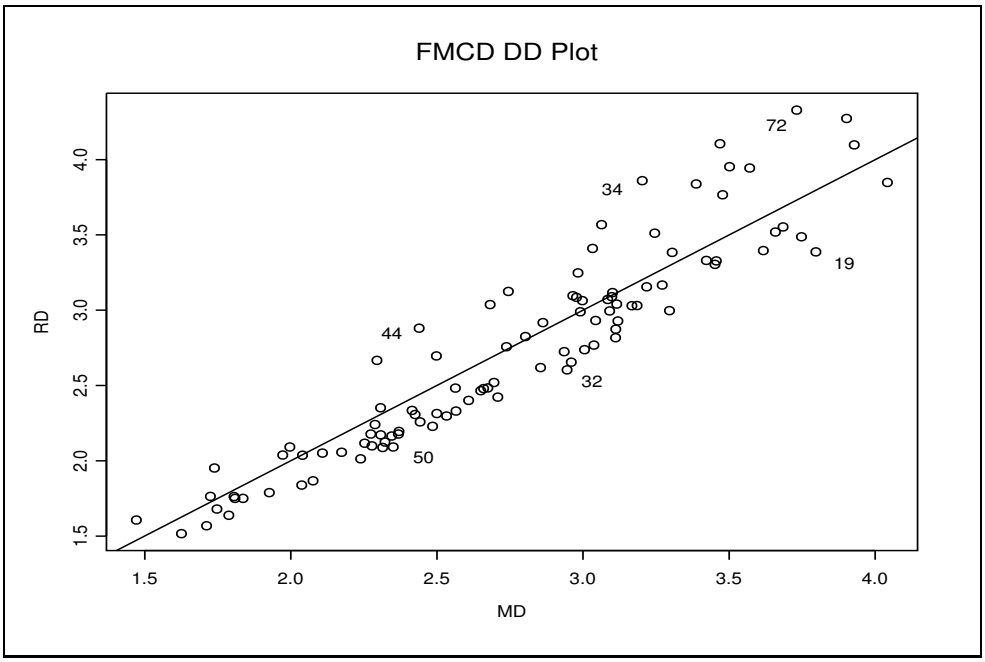

Figure 4.3: The FMCD Estimator Failed

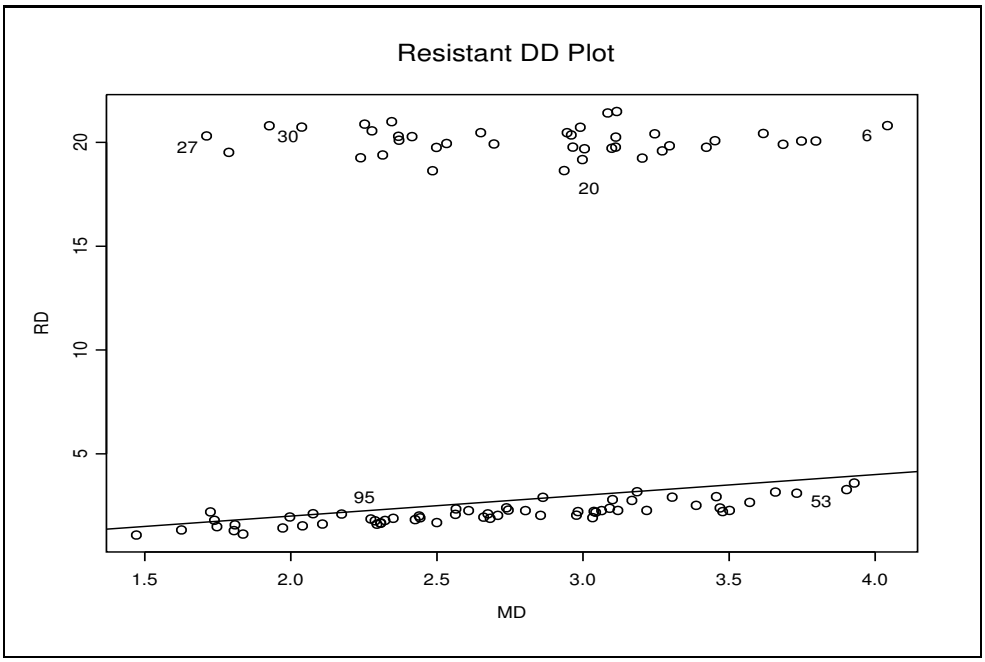

Figure 4.4: The Outliers are Large in the MBA DD Plot

pose that all $p+1$ cases in the elemental start are outliers but the covariance matrix is nonsingular so that the Mahalanobis distances can be computed. Then the classical estimator is applied to the $c_n \approx n/2$ cases with the smallest distances. Suppose the percentage of outliers is less than 25% and that all of the outliers are in this "half set." Then the sample mean applied to the c_n cases should be closer to the bulk of the data than to the cluster of outliers. Hence after a concentration step, the percentage of outliers will be reduced if the outliers are very far away. After the next concentration step the percentage of outliers will be further reduced and after several iterations, all c_n cases will be clean.

In a small simulation study, 20% outliers were planted for various values of p. If the outliers were distant enough, then the minimum DGK distance for the outliers was larger than the maximum DGK distance for the nonoutliers. Hence the outliers would be separated from the bulk of the data in a DD plot of classical versus robust distances. For example, when the clean data comes from the $N_p(\mathbf{0}, \boldsymbol{I}_p)$ distribution and the outliers come from the $N_p(2000\ \mathbf{1}, \boldsymbol{I}_p)$ distribution, the DGK estimator with 10 concentration steps was able to separate the outliers in 17 out of 20 runs when $n = 9000$ and $p = 30$. With 10% outliers, a shift of 40, $n = 600$ and $p = 50$, 18 out of 20 runs worked. Olive (2004a) showed similar results for the Rousseeuw and Van Driessen (1999) FMCD algorithm and that the MBA estimator could often correctly classify up to 49% distant outliers. The following proposition shows that it is very difficult to drive the determinant of the dispersion estimator from a concentration algorithm to zero.

Proposition 4.11. Consider the concentration and MCD estimators that both cover c_n cases. For multivariate data, if at least one of the starts is nonsingular, then the concentration attractor $\boldsymbol{C}_A$ is less likely to be singular than the high breakdown MCD estimator $\boldsymbol{C}_{MCD}$.

Proof. If all of the starts are singular, then the Mahalanobis distances cannot be computed and the classical estimator can not be applied to c_n cases. Suppose that at least one start was nonsingular. Then $\boldsymbol{C}_A$ and $\boldsymbol{C}_{MCD}$ are both sample covariance matrices applied to c_n cases, but by definition $\boldsymbol{C}_{MCD}$ minimizes the determinant of such matrices. Hence $0 \leq \det(\boldsymbol{C}_{MCD}) \leq \det(\boldsymbol{C}_A)$. QED

Software

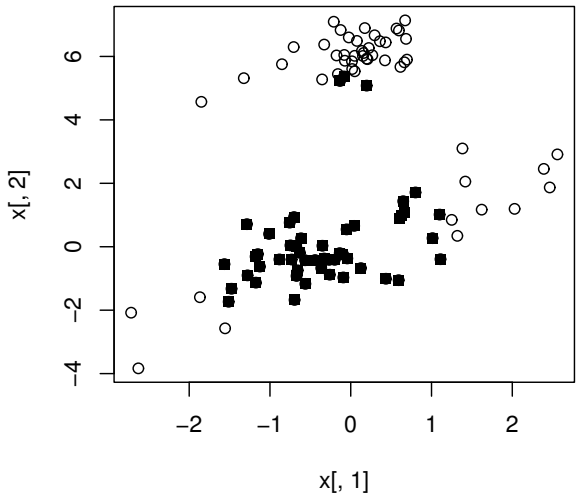

Figure 4.5: highlighted cases = half set with smallest RD = $(T_0, \boldsymbol{C}_0)$

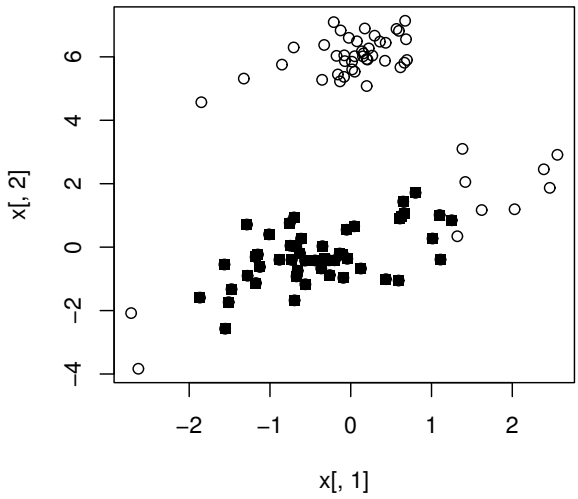

Figure 4.6: highlighted cases = half set with smallest RD = $(T_1, \boldsymbol{C}_1)$

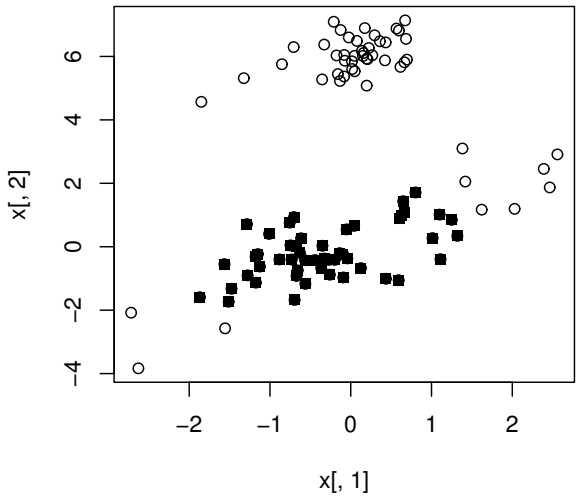

Figure 4.7: highlighted cases = half set with smallest RD = $(T_2, \boldsymbol{C}_2)$

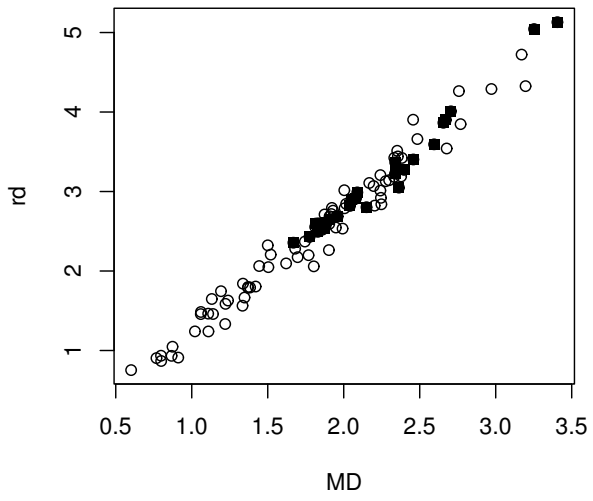

Figure 4.8: highlighted cases = outliers, RD = $(T_{0,D}, \boldsymbol{C}_{0,D})$

The `robustbase` library was downloaded from (www.r-project.org/#doc). § 15.2 explains how to use the source command to get the `mpack` functions in *R* and how to download a library from *R*. Type the commands `library(MASS)` and `library(robustbase)` to compute the FMCD and OGK estimators with the `cov.mcd` and `covOGK` functions.

The `mpack` function
mldsim(n=200,p=5,gam=.2,runs=100,outliers=1,pm=15)
can be used to produce Tables 4.1–4.5. Change outliers to 0 to examine the average of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. The function `mldsim6` is similar but does not need the `library` command since it compares the FCH, RFCH, CMVE, RCMVE and MB estimators. The command
sctplt(n=200,p=10,gam=.2,outliers=3, pm=5)
will make an outlier data set. Then the FCH and MB DD plots are made (click on the right mouse button and highlight stop to go to the next plot) and then the scatterplot matrix. The scatterplot matrix can be used to determine whether the outliers are hard to detect with bivariate or univariate methods. If $p > 10$ the bivariate plots may be too small. See Zhang (2011) for more simulations.

The function *covsim2* can be modified to show that the `R` implementation of FCH is usually much faster than OGK which is much faster than FMCD. The function *corrsim* can be used to simulate the correlations of robust distances with classical distances. RCMVE, RMBA and RFCH are reweighted versions of CMVE, MBA and FCH that may perform better for small n. For MVN data, the command
corrsim(n=200,p=20,nruns=100,type=5)
suggests that the correlation of the RFCH distances with the classical distances is about 0.97. Changing *type* to 4 suggests that FCH needs $n = 800$ before the correlation is about 0.97. The function `corrsim2` uses a wider variety of EC distributions. See Zhang (2011) for simulations.

The function *cmve* computes CMVE and RCMVE, function *covfch* computes FCH and RFCH while *covrmvn* computes the RMVN and MB estimators. The function *covrmb* computes MB and RMB where RMB is like RMVN except the MB estimator is reweighted instead of FCH. Functions *covdgk*, *covmba* and *rmba* compute the scaled DGK, MBA and RMBA estimators.

The *concmv* function described in Problem 4.5 illustrates concentration where the start is $(\text{MED}(\boldsymbol{W}), diag([MAD(X_i)]^2))$. In Figures 4.5, 4.6, and 4.7, the highlighted cases are the half set with the smallest distances, and

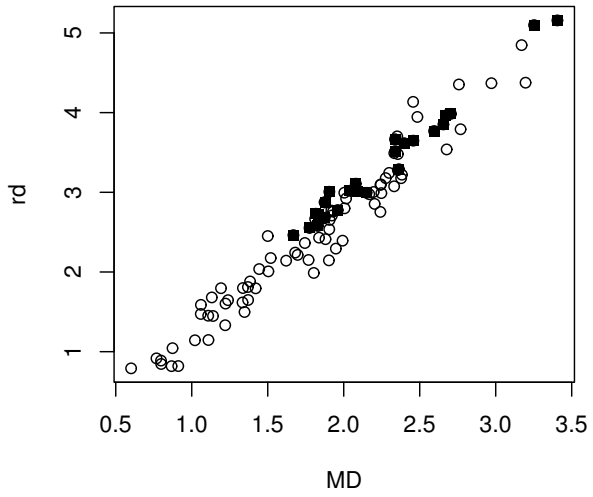

Figure 4.9: highlighted cases = outliers, RD = $(T_{1,D}, \boldsymbol{C}_{1,D})$

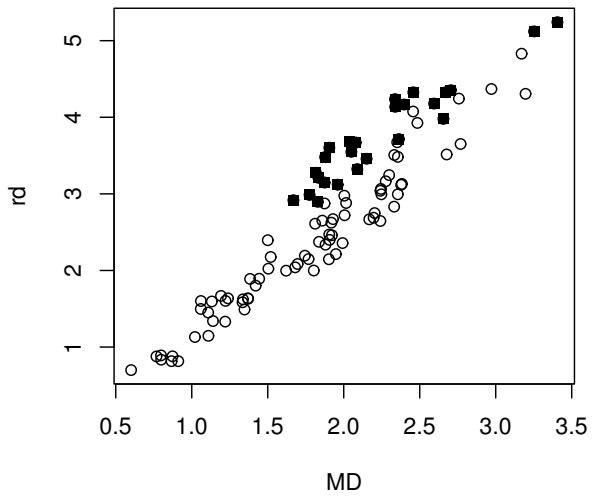

Figure 4.10: highlighted cases = outliers, RD = $(T_{2,D}, \boldsymbol{C}_{2,D})$

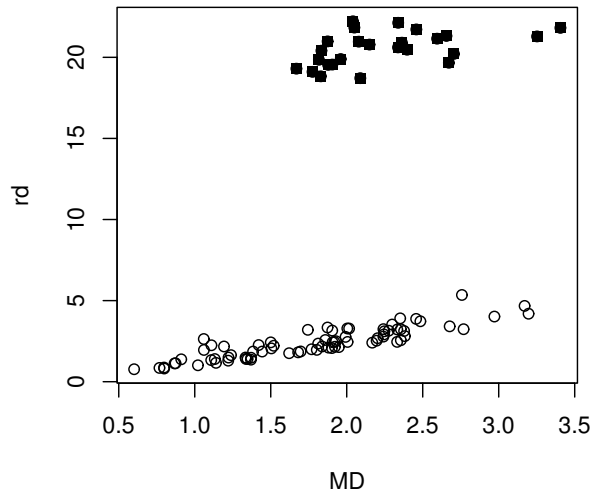

Figure 4.11: highlighted cases = outliers, RD = $(T_{3,D}, \boldsymbol{C}_{3,D})$

the initial half set shown in Figure 4.5 is not clean, where $n = 100$ and there are 40 outliers. The attractor shown in Figure 4.7 is clean. This type of data set has too many outliers for DGK while the MB starts and attractors are almost always clean.

The *ddmv* function in Problem 4.6 illustrates concentration for the DGK estimator where the start is the classical estimator. Now $n = 100, p = 4$ and there are 25 outliers. A DD plot of classical distances MD versus robust distances RD is shown. See Figures 4.8, 4.9, 4.10 and 4.11. The half set of cases with the smallest RDs is used, and the initial half set shown in Figure 4.8 is not clean. The attractor in Figure 4.11 is the DGK estimator which uses a clean half set. The clean cases $\boldsymbol{x}_i \sim N_4(\boldsymbol{0}, diag(1, 2, 3, 4))$ while the outliers $\boldsymbol{x}_i \sim N_4((10, 10\sqrt{2}, 10\sqrt{3}, 20)^T,\ diag(1, 2, 3, 4))$.

4.6 Summary

1) Given a table of data $\boldsymbol{W}$ for variables $X_1, ..., X_p$, be able to find the **coordinatewise median** MED($\boldsymbol{W}$) and the **sample mean** $\overline{\boldsymbol{x}}$. If $\boldsymbol{x} =$

$(X_1, X_2, ..., X_p)^T$ where X_j corresponds to the jth column of $\boldsymbol{W}$, then $\text{MED}(\boldsymbol{W}) = (\text{MED}_{X_1}(n), ..., \text{MED}_{X_p}(n))^T$ where $\text{MED}_{X_j}(n) = \text{MED}(X_{j,1}, ..., X_{j,n})$ is the sample median of the data in the jth column. Similarly, $\overline{\boldsymbol{x}} = (\overline{X}_1, ..., \overline{X}_p)^T$ where $\overline{X}_j$ is the sample mean of the data in the jth column. See Q3.

2) A **DD plot** is a plot of classical vs robust Mahalanobis distances. The DD plot is used to check i) if the data is MVN (plotted points follow the identity line), ii) if the data is EC but not MVN (plotted points follow a line through the origin with slope > 1), iii) if the data is not EC (plotted points do not follow a line through the origin) iv) if multivariate outliers are present (eg some plotted points are far from the bulk of the data or the plotted points follow two lines). See Q3.

3) Many practical "robust estimators" generate a sequence of K trial fits called *attractors*: $(T_1, \boldsymbol{C}_1), ..., (T_K, \boldsymbol{C}_K)$. Then the attractor $(T_A, \boldsymbol{C}_A)$ that minimizes some criterion is used to obtain the final estimator. One way to obtain attractors is to generate trial fits called *starts*, and then use the *concentration* technique. Let $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ be the jth start and compute all n Mahalanobis distances $D_i(T_{-1,j}, \boldsymbol{C}_{-1,j})$. At the next iteration, the classical estimator $(T_{0,j}, \boldsymbol{C}_{0,j})$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest distances. This iteration can be continued for k steps resulting in the sequence of estimators $(T_{-1,j}, \boldsymbol{C}_{-1,j}), (T_{0,j}, \boldsymbol{C}_{0,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. Then $(T_{k,j}, \boldsymbol{C}_{k,j})$ is the jth attractor for $j = 1, ..., K$. Using $k = 10$ often works well, and the basic resampling algorithm is a special case $k = -1$ where the attractors are the starts.

4) The DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start.

5) The median ball (MB) estimator $(T_{MB}, \boldsymbol{C}_{MB})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ as the only start where $\text{MED}(\boldsymbol{W})$ is the coordinatewise median. Hence $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of data closest to $\text{MED}(\boldsymbol{W})$ in Euclidean distance.

6) Elemental concentration algorithms use elemental starts: $(T_{-1,j}, \boldsymbol{C}_{-1,j}) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ is the classical estimator applied to a randomly selected "elemental set" of $p + 1$ cases. If the $\boldsymbol{x}_i$ are iid with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then the starts $(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ are identically distributed with $E(\overline{\boldsymbol{x}}_j) = E(\boldsymbol{x}_i)$ and $\text{Cov}(\overline{\boldsymbol{x}}_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}}/(p + 1)$.

7) Let the "median ball" be the hypersphere containing the half set of data closest to $\text{MED}(\boldsymbol{W})$ in Euclidean distance. The FCH estimator uses the MB attractor if the DGK location estimator $T_{DGK} = T_{k,D}$ is outside of

the median ball, and the attractor with the smallest determinant, otherwise. Let $(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_{FCH}, \boldsymbol{C}_{FCH})$ takes $T_{FCH} = T_A$ and

$$\boldsymbol{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{4.12}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with p degrees of freedom. The RFCH estimator uses two standard "reweight for efficiency steps" while the RMVN estimator uses a modified method for reweighting.

8) For a large class of elliptically contoured distributions, FCH, RFCH and RMVN are $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c_i\boldsymbol{\Sigma})$ for $c_1, c_2, c_3 > 0$ where $c_i = 1$ for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data.

9) An estimator $(T, \boldsymbol{C})$ of multivariate location and dispersion (MLD), needs to estimate $p(p+3)/2$ unknown parameters when there are p random variables. For $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ or $(\overline{\boldsymbol{z}}, \boldsymbol{R})$, want $n > 10p$. Want $n > 20p$ for FCH, RFCH or RMVN.

10) Brand name robust MLD estimators from the Rousseeuw and Yohai paradigm take too long to compute: F-brand name estimators that are not backed by breakdown or large sample theory are actually used. FMCD, F-MVE, F-S, F-MM, F-τ, F-constrained-M and F-Stahel-Donoho are especially common.

4.7 Complements

For concentration algorithms, note that $(T_{t,j}, \boldsymbol{C}_{t,j}) = (\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ is the classical estimator applied to the "half set" of cases satisfying $\{\boldsymbol{x}_i : D_i^2(\overline{\boldsymbol{x}}_{t-1,j}, \boldsymbol{S}_{t-1,j}) \leq D^2_{(c_n)}(\overline{\boldsymbol{x}}_{t-1,j}, \boldsymbol{S}_{t-1,j})\}$ for $t \geq 0$. Hence $(T_{t,j}, \boldsymbol{C}_{t,j})$ is estimating $(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, the population mean and covariance matrix of the truncated distribution covering half of the mass corresponding to $\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu}_{t-1})^T \boldsymbol{\Sigma}_{t-1}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{t-1}) \leq D^2_{0.5}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})\}$ where $D^2_{0.5}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$ is the population median of the population squared distances $D^2(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$. Here $(\boldsymbol{\mu}_{-1}, \boldsymbol{\Sigma}_{-1})$ is the population analog of $(T_{-1,j}, \boldsymbol{C}_{-1,j})$.

The DGK estimator $(T_{k,D}, \boldsymbol{C}_{k,D})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D})$ $= (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start. Thus $(\boldsymbol{\mu}_{-1,D}, \boldsymbol{\Sigma}_{-1,D})$ is the population mean and covariance matrix. For an elliptically contoured distribution with a nonsingular covariance matrix and for $t \geq 0$, $(\boldsymbol{\mu}_{t,D}, \boldsymbol{\Sigma}_{t,D})$ is the population mean and covariance matrix of the truncated distribution corresponding to the highest density region covering half the mass. Hence $\boldsymbol{\mu}_{t,D} = \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{t,D} = c\boldsymbol{\Sigma}$

for some $c > 0$. Riani, Atkinson and Cerioli (2009) find the population mean and covariance matrices for such truncated multivariate normal distributions, using results from Tallis (1963).

Conjecture 4.2. The DGK estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}_{k,D}, \boldsymbol{\Sigma}_{k,D})$ under mild conditions.

The median ball (MB) estimator $(T_{k,M}, \boldsymbol{C}_{k,M})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) = (\text{MED}(\boldsymbol{X}), \boldsymbol{I}_p)$ as the only start where $\text{MED}(\boldsymbol{X})$ is the coordinatewise median. Hence $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of data closest to $\text{MED}(\boldsymbol{X})$ in Euclidean distance while $(\boldsymbol{\mu}_{0,M}, \boldsymbol{\Sigma}_{0,M})$ is the population mean and covariance matrix of the truncated distribution corresponding to the hypersphere centered at the population median that contains half the mass. For a distribution that is spherical about $\boldsymbol{\mu}$ and for $t \geq 0$, $(\boldsymbol{\mu}_{t,M}, \boldsymbol{\Sigma}_{t,M}) = (\boldsymbol{\mu}, c\boldsymbol{I}_p)$ for some $c > 0$. For nonspherical elliptically contoured distributions, $\boldsymbol{\Sigma}_{t,M} \neq c\boldsymbol{\Sigma}$. However, the bias seems to be small even for $t = 0$, and to get smaller as k increases. If the median ball estimator is iterated to convergence, we do not know whether $\boldsymbol{\Sigma}_{\infty,M} = c\boldsymbol{\Sigma}$.

Conjecture 4.3. The MB estimator is a high breakdown $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}_{k,M}, \boldsymbol{\Sigma}_{k,M})$ under mild conditions. For elliptically contoured distributions, $\boldsymbol{\mu}_{k,M} = \boldsymbol{\mu}$.

Arcones (1995) and Kim (2000) showed that $\overline{\boldsymbol{x}}_{0,M}$ is a HB $\sqrt{n}$ consistent estimator of $\boldsymbol{\mu}$. Olive (2004a) showed that $(\overline{\boldsymbol{x}}_{0,M}, \boldsymbol{S}_{0,M})$ is a high breakdown estimator. If the data distribution is EC but not spherical about $\boldsymbol{\mu}$, then for $k \geq 0$, $\boldsymbol{S}_{k,M} = \boldsymbol{C}_{MB}$ under estimates the major axis and over estimates the minor axis of the highest density region. Concentration reduces but fails to eliminate this bias. Hence the estimated highest density region based on the attractor is "shorter" in the direction of the major axis and "fatter" in the direction of the minor axis than estimated regions based on consistent estimators.

Recall that the sample median $\text{MED}(Y_i) = Y((n+1)/2)$ is the middle order statistic if n is odd. Thus if $n = m + d$ where m is the number of clean cases and $d = m - 1$ is the number of outliers so $\gamma \approx 0.5$, then the sample median can be driven to the max or min of the clean cases. The jth element of $\text{MED}(\boldsymbol{W})$ is the sample median of the jth predictor. Hence with $m-1$ outliers, $\text{MED}(\boldsymbol{W})$ can be driven to the "coordinatewise covering box" of the m clean cases. The boundaries of this box are at the min and

max of the clean cases from each predictor, and the lengths of the box edges equal the ranges R_i of the clean cases for the ith variable. If $d \approx m/2$ so that $\gamma \approx 1/3$, then the MED($\boldsymbol{W}$) can be moved to the boundary of the much smaller "coordinatewise IQR box" corresponding the 25th and 75th percentiles of the clean date. Then the edge lengths are approximately equal to the interquartile ranges of the clean cases.

Note that $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p) = \|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|$ is the Euclidean distance of $\boldsymbol{x}_i$ from MED($\boldsymbol{W}$). Let $\mathcal{C}$ denote the set of m clean cases. If $d \leq m-1$, then the minimum distance of the outliers is larger than the maximum distance of the clean cases if the distances for the outliers satisfy $D_i > B$ where

$$B^2 = \max_{i \in \mathcal{C}} \|\boldsymbol{x}_i - \text{MED}(\boldsymbol{X})\|^2 \leq \sum_{i=1}^{p} R_i^2 \leq p(\max R_i^2).$$

One of the most effective methods for detecting outliers for large data sets or if $p > n$ is to use $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$.

The MB estimator has outlier resistance similar to (MED($\boldsymbol{W}$), $\boldsymbol{I}_p$) for distant outliers but, as shown in Example 4.1, can be much more effective for detecting certain types of outliers that can not be found by marginal methods. For EC data, the MB estimator is best if the data is spherical about $\boldsymbol{\mu}$ or if the data is highly correlated with the major axis of the highest density region $\{\boldsymbol{x}_i : D_i^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq d^2\}$.

If the DGK estimator is used by itself, we recommend $k = 10$ in the concentration algorithm. We use $k = 5$ when the DGK and MB estimators are used as attractors for the FCH, CMVE and MBA estimators. The scaling (4.10) makes $\boldsymbol{C}_{FCH}$ a better estimate of $\boldsymbol{\Sigma}$ if the data is multivariate normal MVN.

Concentration for the MB estimator begins with the "half set" of data closest to the coordinatewise median in Euclidean distance, resulting in the estimator $(T_{0,M}, \boldsymbol{C}_{0,M})$ that uses 50% trimming. $(T_{0,M}, \boldsymbol{C}_{0,M})$ is a high breakdown estimator by Corollary 4.7. Since only cases $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|)$ are used, the largest eigenvalue of $\boldsymbol{C}_{0,50}$ is bounded if fewer than half of the cases are outliers by Lemma 4.3.

The geometric behavior of $(T_{0,M}, \boldsymbol{C}_{0,M})$ is simple. If the data $\boldsymbol{x}_i$ are MVN (or EC) then the highest density regions of the data are hyperellipsoids. The set of $\boldsymbol{x}$ closest to the coordinatewise median in Euclidean distance is a hypersphere. For EC data the highest density ellipsoid and hypersphere will have approximately the same center as the hypersphere, and the hypersphere

will be drawn towards the longest axis of the hyperellipsoid. Hence too much data will be trimmed in that direction. For example, if the data are MVN with $\boldsymbol{\Sigma} = \text{diag}(1, 2, ..., p)$ then $\boldsymbol{C}_{0,M}$ will underestimate the largest variance and overestimate the smallest variance. Taking k concentration steps can greatly reduce but not eliminate the bias of the MB estimator $\boldsymbol{C}_{k,M}$ if the data is EC, and the determinant $|\boldsymbol{C}_{k,M}| < |\boldsymbol{C}_{0,M}|$ unless the attractor is equal $(T_{0,M}, \boldsymbol{C}_{0,M})$ by Proposition 4.4. The MB estimator $(T_{k,M}, \boldsymbol{C}_{k,M})$ is not affine equivariant but is resistant to gross outliers in that they will initially be given weight zero if they are further than the median Euclidean distance from the coordinatewise median. Gnanadesikan and Kettenring (1972, p. 94) suggest an estimator similar to the MB estimator, also see Croux and Van Aelst (2002). Another estimator similar to MB was suggested by Wilk, Gnanadesikan, Huyett and Lauh (1962). See Gnanadesikan (1977, p. 134).

Recall that the *population squared Mahalanobis distance*

$$U \equiv D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}). \tag{4.13}$$

For elliptically contoured distributions, U has pdf given by (3.10), and the 50% highest density region has the form of the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) \leq U_{0.5}\}$$

where $U_{0.5}$ is the median of the distribution of U. For example, if the $\boldsymbol{x}$ are MVN, then U has the χ^2_p distribution. Concentration estimators attempt to estimate the population mean and covariance matrix of the mass in this 50% highest density region. So it should not be surprising that good concentration attractors estimate the same quantity $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. See Theorem 4.9.

In regression, if the start is a consistent estimator for $\boldsymbol{\beta}$, then so is the attractor. Hence all attractors are estimating the *same* parameter $\boldsymbol{\beta}$. Theorem 4.9 showed that MLD concentration attractors with $k \geq 0$ are estimating the *same* parameter $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ even if the affine equivariant starts are estimating $(\boldsymbol{\mu}, s_i\boldsymbol{\Sigma})$ where the $s_i > 0$ can differ for $i = 1, ..., K$.

Olive (2002) showed the following result. Assume $(T_i, \boldsymbol{C}_i)$ are consistent estimators for $(\boldsymbol{\mu}, a_i\boldsymbol{\Sigma})$ where $a_i > 0$ for $i = 1, 2$. Let $D_{i,1}$ and $D_{i,2}$ be the corresponding distances and let R be the set of cases with distances $D_i(T_1, \boldsymbol{C}_1) \leq \text{MED}(D_i(T_1, \boldsymbol{C}_1))$. Let r_n be the correlation between $D_{i,1}$ and $D_{i,2}$ for the cases in R. Then $r_n \to 1$ in probability as $n \to \infty$.

The theory for concentration algorithms is due to Hawkins and Olive (2002) and Olive and Hawkins (2010). The MBA estimator is due to Olive

(2004a). The computational and theoretical simplicity of the FCH estimator makes it one of the most useful robust estimators ever proposed. An important application of the robust algorithm estimators and of case diagnostics is to detect outliers. Sometimes it can be assumed that the analysis for influential cases and outliers was completely successful in classifying the cases into outliers and good or "clean" cases. Then classical procedures can be performed on the good cases. This assumption of perfect classification is often unreasonable, and it is useful to have robust procedures, such as the FCH estimator, that have rigorous asymptotic theory and are practical to compute. Since the FCH estimator is about an order of magnitude faster than alternative robust estimators, the FCH estimator may be useful for computationally intensive applications.

The RFCH and RMVN estimators takes slightly longer to compute than the FCH estimator, and may have slightly less resistance to outliers.

In addition to concentration and randomly selecting elemental sets, three other algorithm techniques are important. He and Wang (1996) suggest computing the classical estimator and a consistent robust estimator. The final cross checking estimator is the classical estimator if both estimators are "close," otherwise the final estimator is the robust estimator. The second technique was proposed by Gnanadesikan and Kettenring (1972, p. 90). They suggest using the dispersion matrix $\boldsymbol{C} = ((c_{i,j}))$ where $c_{i,j}$ is a robust estimator of the covariance of X_i and X_j. Computing the classical estimator on a subset of the data results in an estimator of this form. The identity

$$c_{i,j} = \mathrm{Cov}(\mathrm{X_i}, \mathrm{X_j}) = [\mathrm{VAR}(\mathrm{X_i} + \mathrm{X_j}) - \mathrm{VAR}(\mathrm{X_i} - \mathrm{X_j})]/4$$

where $\mathrm{VAR}(\mathrm{X}) = \sigma^2(\mathrm{X})$ suggests that a robust estimator of dispersion can be created by replacing the sample standard deviation $\hat{\sigma}$ by a robust estimator of scale. Maronna and Zamar (2002) modify this idea to create a fairly fast (possibly high breakdown consistent) OGK estimator of multivariate location and dispersion. This estimator may be the leading competitor of the FCH estimator. Also see Alqallaf, Konis, Martin and Zamar (2002) and Mehrotra (1995). Woodruff and Rocke (1994) introduced the third technique, partitioning, which evaluates a start on a subset of the cases. Poor starts are discarded, and L of the best starts are evaluated on the entire data set. This idea is also used by Rocke and Woodruff (1996) and by Rousseeuw and Van Driessen (1999).

Billor, Hadi and Velleman (2000) have a BACON algorithm that uses $m_0 = 4p$ or $m_0 = 5p$ cases, computes the sample mean and covariance matrix of these cases, finds the m_1 cases with Mahalanobis distances less than some cutoff, then iterates until the subset of cases no longer changes. Version V1 uses the m_0 cases with the smallest classical Mahalanobis distances while version V2 uses the m_0 cases closest to the coordinatewise median.

Croux, Dehon and Yadine (2010) claim that the practical Sign Covariance Matrix is high breakdown and that their practical k-step Spatial Sign Covariance Matrix is high breakdown and consistently estimates the orientation of the scatter matrix. The Sign Covariance Matix $\hat{\boldsymbol{\Sigma}}_S = \frac{1}{n}\sum_{i=1}^{n} \frac{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n)^T}{\|\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n\|^2}$ which is similar to the classical covariance estimator computed from $\boldsymbol{z}_i = \frac{\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n}{\|\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n\|}$. Here $\hat{\boldsymbol{\mu}}_n$ is the L_1-median or spatial median, defined as

$$\hat{\boldsymbol{\mu}}_n = \text{argmin}_{\boldsymbol{\mu}} \quad \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|,$$

is a fairly practical high breakdown estimator of multivariate location.

There certainly exist types of outlier configurations where the FMCD estimator outperforms the robust FCH estimator. The FCH estimators is vulnerable to outliers that lie inside the hypersphere based on the median Euclidean distance from the coordinatewise median. Although the FCH estimator should not be viewed as a replacement for the FMCD estimator, the FMCD estimator should be modified so that it is backed by theory. Until this modification appears in the software, both estimators can be used for outlier detection by making a scatterplot matrix of the Mahalanobis distances from the FMCD, FCH and classical estimators.

The simplest version of the MBA estimator only has two starts. A simple modification would be to add additional starts as in Problem 4.7. The Det-MCD estimator of Hubert, Rousseeuw, and Verdonck (2012) is very similar, uses 6 starts, but is not yet backed by theory.

Rousseeuw (1984) introduced the MCD and the minimum volume ellipsoid MVE(c_n) estimator. For the MVE estimator, $T(\boldsymbol{W})$ is the center of the minimum volume ellipsoid covering c_n of the observations and $\boldsymbol{C}(\boldsymbol{W})$ is determined from the same ellipsoid. T_{MVE} has a cube root rate and the limiting distribution is not Gaussian. See Davies (1992). Bernholdt and Fisher (2004) show that the MCD estimator can be computed with $O(n^v)$

complexity where $v = 1 + p(p+3)/2$ if $\boldsymbol{x}$ is a $p \times 1$ vector.

Rocke and Woodruff (1996, p. 1050) claim that any affine equivariant location and shape estimation method gives an unbiased location estimator and a shape estimator that has an expectation that is a multiple of the true shape for elliptically contoured distributions. Hence there are many candidate robust estimators of multivariate location and dispersion. See Cook, Hawkins and Weisberg (1993) for an exact algorithm for the MVE. Other papers on robust algorithms include Hawkins (1993, 1994), Hawkins and Olive (1999a), Hawkins and Simonoff (1993), He and Wang (1996), Olive (2004a), Olive and Hawkins (2007, 2008), Rousseeuw and Van Driessen (1999), Rousseeuw and van Zomeren (1990), Ruppert (1992), and Woodruff and Rocke (1993). Rousseeuw and Leroy (1987, § 7.1) also describes many methods.

The discussion by Rocke and Woodruff (2001) and by Hubert (2001) of Peña and Prieto (2001) stresses the fact that no one estimator can dominate all others for every outlier configuration. These papers and Wisnowski, Simpson, and Montgomery (2002) give outlier configurations that can cause problems for the FMCD estimator.

Papers on robust distances include Olive (2002) and García-Escudero and Gordaliza (2005).

Huber and Ronchetti (2009, p. 214, 233) note that theory for M-estimators of multivariate location and dispersion is "not entirely satisfactory with regard to joint estimation of" $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ and that "so far we have neither a really fast, nor a demonstrably convergent, procedure for calculating simultaneous M-estimates of location and scatter."

If an exact algorithm exists but an approximate algorithm is also used, the two estimators should be distinguished in some manner. For example $(T_{MCD}, \boldsymbol{C}_{MCD})$ could denote the estimator from the exact algorithm while $(T_{AMCD}, \boldsymbol{C}_{AMCD})$ could denote the estimator from the approximate algorithm. In the literature this distinction is too seldomly made, but there are a few outliers. Cook and Hawkins (1990, p. 640) point out that the AMVE is not the minimum volume ellipsoid (MVE) estimator.

Where the Rousseeuw-Yohai Paradigm Goes Wrong

i) Estimators from this paradigm that have been shown to be both high breakdown and consistent take too long to compute.

Let the ith case $\boldsymbol{x}_i$ be a $p \times 1$ random vector, and suppose the n cases are collected in an $n \times p$ matrix $\boldsymbol{W}$ with rows $\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T$. The fastest estimators of multivariate location and dispersion that have been shown to be both consistent and high breakdown are the minimum covariance determinant (MCD)

estimator with $O(n^v)$ complexity where $v = 1+p(p+3)/2$ and possibly an all elemental subset estimator of He and Wang (1997). See Bernholt and Fischer (2004). The minimum volume ellipsoid complexity is far higher, and for $p > 2$ there may be no known method for computing S, τ, projection based, constrained M, MM, and Stahel-Donoho estimators. **These estimators have computational complexity is higher than $O(n^p)$.** See Maronna, Martin and Yohai (2006, ch. 6) for descriptions and references.

Estimators with complexity higher than $O[(n^3 + n^2p + np^2 + p^3)\log(n)]$ take too long to compute and will rarely be used. Reyen, Miller, and Wegman (2009) simulate the OGK and the Olive (2004a) median ball algorithm (MBA) estimators for $p = 100$ and n up to 50000, and note that the OGK complexity is $O[p^3 + np^2\log(n)]$ while that of MBA is $O[p^3 + np^2 + np\log(n)]$. FCH, RMBA, RMVN, CMVE and RCMVE have the same complexity as MBA. FMCD has the same complexity as FCH, but FCH roughly 100 to 200 times faster.

ii) No practical useful "high breakdown" estimator of multivariate location and dispersion from this paradigm has been shown to be consistent or high breakdown: to my knowledge, **if the complexity of the estimator is less than $O(n^4)$ for general p, and if the estimator has been claimed in the published literature to be both high breakdown and consistent, then the estimator has not been shown to be either high breakdown or consistent.** Also Hawkins and Olive (2002) showed that elemental concentration estimators using K starts are zero breakdown estimators. They are inconsistent if they use k concentration steps where k is fixed.

Papers with titles like Rousseeuw and Van Driessen (1999) "A Fast Algorithm for the Minimum Covariance Determinant Estimator" and Hubert, Rousseeuw and Van Aelst (2008) "High Breakdown Multivariate Methods" where the zero breakdown estimators have not been shown to be consistent are common, and very misleading to researchers who are not experts in robust statistics. Also see Olive (2012a).

iii) Many papers give theory for an impractical estimator such as MCD, then replace the estimator by a zero breakdown practical estimator such as FAST-MCD.

If an estimator can not be computed in a reasonable amount of time, then most of its theoretical properties are only of academic interest (consistency of MCD is needed for the practical FCH estimator). What is of interest are the theoretical properties of the estimator actually used.

The central thesis of Hawkins and Olive (2002) was that, given the disconnect between the theoretically defined estimator and what can actually be computed, the theoretical properties of the former do not necessarily give useful guidance on the properties of the latter. Nearly all of the literature appears to overlook this disconnect, including Hubert, Rousseeuw and Van Aelst (2008) and Maronna, Martin and Yohai (2006).

iv) Papers on breakdown and maximal bias are not useful.

Both these properties are weaker than asymptotic unbiasedness. Also the properties are derived for estimators that take far too long to compute.

Breakdown is a very weak property: having $\|T\|$ bounded and eigenvalues of $\boldsymbol{C}$ bounded away from 0 and ∞ does not mean that the estimator is good. All too often claims are made that "high breakdown estimators make outliers have large distances."

Sometimes the literature gives a claim similar to "the fact that FMCD is not the MCD estimator is unimportant since the algorithm that uses all elemental sets has the same high breakdown value as MCD." FMCD is not the MCD estimator and FMCD is not the estimator that uses all elemental sets. FMCD only uses a fixed number of elemental sets, hence FMCD is zero breakdown.

v) Too much emphasis is given on the property of affine equivariance since typically this is the only property that can be shown for a practical estimator of MLD.

Huber and Ronchetti (2009, p. 200, 283) note that "one ought to be aware that affine equivariance is a requirement deriving from mathematical aesthetics; it is hardly ever dictated by the scientific content of the underlying problem," and the lack of affine equivariance "may be less of a disadvantage than it first seems, since in statistics problems possessing genuine affine equivariance are quite rare." Also see the end of Section 4.1.

Being a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ is an important property, and the FCH estimator is asymptotically equivalent to the scaled DGK estimator, which is affine equivariant.

vi) The literature implies that the breakdown value is a measure of the global reliability of the estimator and is a lower bound on the amount of contamination needed to destroy an estimator.

These interpretations are not correct since the complement of complete and total failure is *not* global reliability. The breakdown value d_n/n is actually an upper bound on the amount of contamination that the estimator can tolerate since the estimator can be made arbitrarily bad with d_n mali-

ciously placed cases. In particular, the breakdown value of an estimator tells nothing about more important properties such as consistency or asymptotic normality.

4.8 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

R/Splus Problems

Use the command *source("G:/mpack.txt")* **to download the functions** and the command *source("G:/mrobdata.txt")* **to download the data. See Preface or Section 15.2.** Typing the name of the `mpack` function, eg *covmba*, will display the code for the function. Use the `args` command, eg *args(covmba)*, to display the needed arguments for the function.

4.1. a) Download the `maha` function that creates the classical Mahalanobis distances.

b) Enter the following commands and check whether observations 1–40 look like outliers.

```
> simx2 <- matrix(rnorm(200),nrow=100,ncol=2)
> outx2 <- matrix(10 + rnorm(80),nrow=40,ncol=2)
> outx2 <- rbind(outx2,simx2)
> maha(outx2)
```

4.2. Download the `rmaha` function that creates the robust Mahalanobis distances. Obtain outx2 as in Problem 4.1 b). *R* users need to enter the command *library(MASS)*. Enter the command *rmaha(outx2)* and check whether observations 1–40 look like outliers.

4.3. a) Download the `covmba` function.

b) Download the program `rcovsim`.

c) Enter the command `rcovsim(100)` three times and include the output in *Word.*

d) Explain what the output is showing.

4.4*. a) Assuming that you have done the two source commands above Problem 4.1 (and in *R* the *library(MASS)* command), type the command

ddcomp(buxx). This will make 4 DD plots based on the DGK, FCH, FMCD and median ball estimators. The DGK and median ball estimators are the two attractors used by the FCH estimator. With the leftmost mouse button, move the cursor to an outlier and click. This data is the Buxton (1920) data and cases with numbers 61, 62, 63, 64, and 65 were the outliers with head lengths near 5 feet. After identifying at least three outliers in each plot, hold the rightmost mouse button down (and in *R* click on *Stop*) to advance to the next plot. When done, hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

b) Repeat a) but use the command *ddcomp(cbrainx).* This data is the Gladstone (1905-6) data and some infants are multivariate outliers.

c) Repeat a) but use the command *ddcomp(museum[,-1]).* This data is the Schaaffhausen (1878) skull measurements and cases 48–60 were apes while the first 47 cases were humans.

4.5*. (Perform the *source("G:/mpack.txt")* command if you have not already done so.) The *concmv* function illustrates concentration with $p = 2$ and a scatterplot of X_1 versus X_2. The outliers are such that the MBA and FCH estimators can not always detect them. Type the command *concmv().* Hold the rightmost mouse button down (and in *R* click on *Stop*) to see the DD plot after one concentration step. The start uses the coordinatewise median and $diag([MAD(X_i)]^2)$. Repeat 4 more times to see the DD plot based on the attractor. The outliers have large values of X_2 and the highlighted cases have the smallest distances. Repeat the command *concmv()* several times. Sometimes the start will contain outliers but the attractor will be clean (none of the highlighted cases will be outliers), but sometimes concentration causes more and more of the highlighted cases to be outliers, so that the attractor is worse than the start. Copy one of the DD plots where none of the outliers are highlighted into *Word*.

4.6*. (Perform the *source("G:/mpack.txt")* command if you have not already done so.) The *ddmv* function illustrates concentration with the DD plot. The outliers are highlighted. The first graph is the DD plot after one concentration step. Hold the rightmost mouse button down (and in *R* click on *Stop*) to see the DD plot after two concentration steps. Repeat 4 more times to see the DD plot based on the attractor. In this problem, try to determine the proportion of outliers *gam* that the DGK estimator can detect for $p = 2, 4, 10$ and 20. Make a table of p and *gam*. For example the command

ddmv(p=2,gam=.4) suggests that the DGK estimator can tolerate nearly 40% outliers with $p = 2$, but the command *ddmv(p=4,gam=.4)* suggest that *gam* needs to be lowered (perhaps by 0.1 or 0.05). Try to make $0 < gam < 0.5$ as large as possible.

4.7. (Perform the *source("G:/mpack.txt")* command if you have not already done so.) A simple modification of the MBA estimator adds starts trimming M% of cases furthest from the coordinatewise median MED($\boldsymbol{x}$). For example use $M \in \{98, 95, 90, 80, 70, 60, 50\}$. Obtain the program *cmba2* from `mpack.txt` and try the MBA estimator on the data sets in Problem 4.4.

4.8. The *mpack* function `covesim` compares various ways to robustly estimate the covariance matrix. The estimators used are ccov: the classical estimator applied to the clean cases, RFCH and RMVN. The average dispersion matrix is reported over nruns = 20. Let diag(A) be the diagonal of the average dispersion matrix. Then diagdiff = diag(ccov) - diag(rmvne) and abssumd = sum(abs(diagdiff)). The clean data $N_p(0, diag(1, ..., p))$.

a) The *R* command covesim(n=100,p=4) gives output when there are no outliers. Copy and paste the output into *Word*.

b) The command covesim(n=100,p=4,outliers=1,pm=15) uses 40% outliers that are a tight cluster at major axis with mean $(0, ..., 0, pm)^T$. Hence *pm* determines how far the outliers are from the bulk of the data. Copy and paste the output into *Word*. The average dispersion matrices should be $\approx c$ $diag(1, 2, 3, 4)$ for this type of outlier configuration. What is c for RFCH and RMVN?

4.9. The *R* function `cov.mcd` is a FMCD estimator. If `cov.mcd` computed the minimum covariance determinant estimator, then the log determinant of the dispersion matrix would be a minimum and would not change when the rows of the data matrix are permuted. The *R commands* for this problem permute the rows of the Gladstone (1905-6) data matrix seven times. The log determinant is given for each of the resulting `cov.mcd` estimators.

a) Paste the output into *Word*.

b) How many distinct values of the log determinant were produced? (Only one if the MCD estimator is being computed.)

Chapter 5

DD Plots and Prediction Regions

5.1 DD Plots

A basic way of designing a graphical display is to arrange for reference situations to correspond to straight lines in the plot.
Chambers, Cleveland, Kleiner, and Tukey (1983, p. 322)

Definition 5.1: Rousseeuw and Van Driessen (1999). The *DD plot* is a plot of the classical Mahalanobis distances MD_i versus robust Mahalanobis distances RD_i.

The DD plot is used as a diagnostic for multivariate normality, elliptical symmetry and for outliers. Assume that the data set consists of iid vectors from an $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with second moments. Then the classical sample mean and covariance matrix $(T_M, \boldsymbol{C}_M) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ is a consistent estimator for $(\boldsymbol{\mu}, c_{\boldsymbol{X}}\boldsymbol{\Sigma}) = (E(\boldsymbol{X}), \mathrm{Cov}(\boldsymbol{X}))$. Assume that an alternative algorithm estimator $(T_A, \boldsymbol{C}_A)$ is a consistent estimator for $(\boldsymbol{\mu}, a_A\boldsymbol{\Sigma})$ for some constant $a_A > 0$. By scaling the algorithm estimator, the DD plot can be constructed to follow the identity line with unit slope and zero intercept. Let $(T_R, \boldsymbol{C}_R) = (T_A, \boldsymbol{C}_A/\tau^2)$ denote the scaled algorithm estimator where $\tau > 0$ is a constant to be determined. Notice that $(T_R, \boldsymbol{C}_R)$ is a valid estimator of location and dispersion. Hence the robust distances used in the DD plot are given by

$$\mathrm{RD}_i = \mathrm{RD}_i(T_R, \boldsymbol{C}_R) = \sqrt{(\boldsymbol{x}_i - T_R(\boldsymbol{W}))^T[\boldsymbol{C}_R(\boldsymbol{W})]^{-1}(\boldsymbol{x}_i - T_R(\boldsymbol{W}))}$$

$= \tau\ D_i(T_A, \boldsymbol{C}_A)$ for $i = 1, ..., n$.

The following proposition shows that if consistent estimators are used to construct the distances, then the DD plot will tend to cluster tightly about the line segment through $(0, 0)$ and $(\mathrm{MD}_{n,\alpha}, \mathrm{RD}_{n,\alpha})$ where $0 < \alpha < 1$ and $\mathrm{MD}_{n,\alpha}$ is the α sample percentile of the MD_i. Nevertheless, the variability in the DD plot may increase with the distances. Let $K > 0$ be a constant, eg the 99th percentile of the χ^2_p distribution.

Proposition 5.1. Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid observations from a distribution with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix. Let $a_j > 0$ and assume that $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ are consistent estimators of $(\boldsymbol{\mu}, a_j\boldsymbol{\Sigma})$ for $j = 1, 2$.

a) $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = o_P(1)$.

b) Let $0 < \delta \leq 0.5$. If $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - (\boldsymbol{\mu}, a_j\boldsymbol{\Sigma}) = O_p(n^{-\delta})$ and $a_j\hat{\boldsymbol{\Sigma}}_j^{-1} - \boldsymbol{\Sigma}^{-1} = O_P(n^{-\delta})$, then

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = O_P(n^{-\delta}).$$

c) Let $D_{i,j} \equiv D_i(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ be the ith Mahalanobis distance computed from $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$. Consider the cases in the region $R = \{i | 0 \leq D_{i,j} \leq K,\ j = 1, 2\}$. Let r_n denote the correlation between $D_{i,1}$ and $D_{i,2}$ for the cases in R (thus r_n is the correlation of the distances in the "lower left corner" of the DD plot). Then $r_n \to 1$ in probability as $n \to \infty$.

Proof. Let B_n denote the subset of the sample space on which both $\hat{\boldsymbol{\Sigma}}_{1,n}$ and $\hat{\boldsymbol{\Sigma}}_{2,n}$ have inverses. Then $P(B_n) \to 1$ as $n \to \infty$.

a) and b): $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) =$

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} - \frac{\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

$$= (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{-\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) + (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T (-\boldsymbol{\Sigma}^{-1} + a_j\ \hat{\boldsymbol{\Sigma}}_j^{-1}) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)\ \ +$$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

$$+\frac{2}{a_j}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_j)+\frac{1}{a_j}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_j)$$

$$+\frac{1}{a_j}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j)^T[a_j\hat{\boldsymbol{\Sigma}}_j^{-1}-\boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j) \quad (5.1)$$

on B_n, and the last three terms are $o_P(1)$ under a) and $O_P(n^{-\delta})$ under b).

c) Following the proof of a), $D_j^2 \equiv D_{\boldsymbol{x}}^2(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) \xrightarrow{P} (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/a_j$ for fixed $\boldsymbol{x}$, and the result follows.

QED

The above result implies that a plot of the MD_i versus the $D_i(T_A, \boldsymbol{C}_A) \equiv D_i(A)$ will follow a line through the origin with some positive slope since if $\boldsymbol{x} = \boldsymbol{\mu}$, then both the classical and the algorithm distances should be close to zero. We want to find τ such that $\mathrm{RD}_i = \tau\ D_i(T_A, \boldsymbol{C}_A)$ and the DD plot of MD_i versus RD_i follows the identity line. By Proposition 5.1, the plot of MD_i versus $D_i(A)$ will follow the line segment defined by the origin $(0, 0)$ and the point of observed median Mahalanobis distances, $(\mathrm{med}(\mathrm{MD}_i), \mathrm{med}(D_i(A)))$. This line segment has slope

$$\mathrm{med}(D_i(A))/\mathrm{med}(\mathrm{MD}_i)$$

which is generally not one. By taking $\tau = \mathrm{med}(\mathrm{MD}_i)/\mathrm{med}(D_i(A))$, the plot will follow the identity line if $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ is a consistent estimator of $(\boldsymbol{\mu}, c_{\boldsymbol{x}}\boldsymbol{\Sigma})$ and if $(T_A, \boldsymbol{C}_A)$ is a consistent estimator of $(\boldsymbol{\mu}, a_A\boldsymbol{\Sigma})$. (Using the notation from Proposition 5.1, let $(a_1, a_2) = (c_{\boldsymbol{x}}, a_A)$.) The classical estimator is consistent if the population has a nonsingular covariance matrix. The algorithm estimators $(T_A, \boldsymbol{C}_A)$ from Theorem 4.10 are consistent on a large class of EC distributions that have a nonsingular covariance matrix, but tend to be biased for non–EC distributions.

By replacing the observed median $\mathrm{med}(\mathrm{MD}_i)$ of the classical Mahalanobis distances with the target population analog, say MED, τ can be chosen so that the DD plot is *simultaneously* a diagnostic for elliptical symmetry and a diagnostic for the target EC distribution. That is, the plotted points follow the identity line if the data arise from a target EC distribution such as the multivariate normal distribution, but the points follow a line with non-unit slope if the data arise from an alternative EC distribution. In addition the

DD plot can often detect departures from elliptical symmetry such as outliers, the presence of two groups, or the presence of a mixture distribution. These facts make the DD plot a useful alternative to other graphical diagnostics for target distributions. See Easton and McCulloch (1990), Li, Fang, and Zhu (1997), and Liu, Parelius, and Singh (1999) for references.

Example 5.1. Rousseeuw and Van Driessen (1999) choose the multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution as the target. If the data are indeed iid MVN vectors, then the $(\mathrm{MD}_i)^2$ are asymptotically χ^2_p random variables, and $\mathrm{MED} = \sqrt{\chi^2_{p,0.5}}$ where $\chi^2_{p,0.5}$ is the median of the χ^2_p distribution. Since the target distribution is Gaussian, let

$$\mathrm{RD}_i = \frac{\sqrt{\chi^2_{p,0.5}}}{\mathrm{med}(D_i(A))} D_i(A) \quad \text{so that} \quad \tau = \frac{\sqrt{\chi^2_{\mathrm{p},0.5}}}{\mathrm{med}(\mathrm{D_i(A)})}. \tag{5.2}$$

Note that the DD plot can be tailored to follow the identity line if the data are iid observations from any target elliptically contoured distribution that has nonsingular covariance matrix. If it is known that $\mathrm{med}(\mathrm{MD}_i) \approx \mathrm{MED}$ where MED is the target population analog (obtained, for example, via simulation, or from the actual target distribution as in Equations (3.8), (3.9) and (3.10)), then use

$$\mathrm{RD}_i = \tau \; D_i(A) = \frac{\mathrm{MED}}{\mathrm{med}(D_i(A))} D_i(A). \tag{5.3}$$

The choice of the algorithm estimator $(T_A, \boldsymbol{C}_A)$ is important, and the $\sqrt{n}$ consistent RFCH estimator is a good choice. In this chapter we used the *R/Splus* function `cov.mcd` which is basically an implementation of the elemental FMCD concentration algorithm described in the previous chapter. The number of starts used was $K = \max(500, n/10)$ (the default is $K = 500$, so the default can be used if $n \leq 5000$).

Conjecture 5.1. If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ and an elemental FMCD concentration algorithm is used to produce the estimator $(T_{A,n}, \boldsymbol{C}_{A,n})$, then this algorithm estimator is consistent for $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ for some constant $a > 0$ (that depends on g) if the number of starts $K = K(n) \to \infty$ as the sample size $n \to \infty$.

Table 5.1: **Corr(RD_i, MD_i) for $N_p(\mathbf{0}, \boldsymbol{I}_p)$ Data, 100 Runs.**

p	n	mean	min	% < 0.95	% < 0.8
3	44	0.866	0.541	81	20
3	100	0.967	0.908	24	0
7	76	0.843	0.622	97	26
10	100	0.866	0.481	98	12
15	140	0.874	0.675	100	6
15	200	0.945	0.870	41	0
20	180	0.889	0.777	100	2
20	1000	0.998	0.996	0	0
50	420	0.894	0.846	100	0

Notice that if this conjecture is true, and if the data is EC with 2nd moments, then

$$\left[\frac{\text{med}(D_i(A))}{\text{med}(\text{MD}_i)}\right]^2 \boldsymbol{C}_A \tag{5.4}$$

estimates $\text{Cov}(\boldsymbol{x})$. For the DD plot, consistency is desirable but not necessary. It is necessary that the correlation of the smallest 99% of the MD_i and RD_i be very high. This correlation goes to 1 by Proposition 5.1 if consistent estimators are used.

The choice of using a concentration algorithm to produce $(T_A, \boldsymbol{C}_A)$ is certainly not perfect, and the `cov.mcd` estimator should be modified by adding the FCH starts to the 500 elemental starts. There exist data sets with outliers or two groups such that both the classical and robust estimators produce ellipsoids that are nearly concentric. We suspect that the situation worsens as p increases.

In a simulation study, $N_p(\mathbf{0}, \boldsymbol{I}_p)$ data were generated and `cov.mcd` was used to compute first the $D_i(A)$, and then the RD_i using Equation (5.2). The results are shown in Table 5.1. Each choice of n and p used 100 runs, and the 100 correlations between the RD_i and the MD_i were computed. The mean and minimum of these correlations are reported along with the percentage of correlations that were less than 0.95 and 0.80. The simulation shows that small data sets (of roughly size $n < 8p + 20$) yield plotted points that may not cluster tightly about the identity line even if the data distribution is

Gaussian.

Since every estimator of location and dispersion defines a hyperellipsoid, the DD plot can be used to examine which points are in the robust hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T_R)^T \boldsymbol{C}_R^{-1} (\boldsymbol{x} - T_R) \leq RD_{(h)}^2\} \tag{5.5}$$

where $RD_{(h)}^2$ is the hth smallest squared robust Mahalanobis distance, and which points are in a classical hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \leq MD_{(h)}^2\}. \tag{5.6}$$

In the DD plot, points below $RD_{(h)}$ correspond to cases that are in the hyperellipsoid given by Equation (5.5) while points to the left of $MD_{(h)}$ are in a hyperellipsoid determined by Equation (5.6).

The DD plot will follow a line through the origin closely if the two hyperellipsoids are nearly concentric, eg if the data is EC. The DD plot will follow the identity line closely if $\text{med}(\text{MD}_i) \approx \text{MED}$, and $\text{RD}_i^2 =$

$$(\boldsymbol{x}_i - T_A)^T [(\frac{\text{MED}}{\text{med}(D_i(A))})^2 \boldsymbol{C}_A^{-1}](\boldsymbol{x}_i - T_A) \approx (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) = \text{MD}_i^2$$

for $i = 1, ..., n$. When the distribution is not EC,

$$(T_A, \boldsymbol{C}_A) = (T_{RFCH}, \boldsymbol{C}_{RFCH}) \text{ or } (\text{T}_\text{A}, \boldsymbol{C}_\text{A}) = (\text{T}_\text{FMCD}, \boldsymbol{C}_\text{FMCD})$$

and $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ will often produce hyperellipsoids that are far from concentric.

Application 5.1. The DD plot can be used *simultaneously* as a diagnostic for whether the data arise from a multivariate normal (MVN or Gaussian) distribution or from another EC distribution with nonsingular covariance matrix. EC data will cluster about a straight line through the origin; MVN data in particular will cluster about the identity line. Thus the DD plot can be used to assess the success of numerical transformations towards elliptical symmetry. This application is important since many statistical methods assume that the underlying data distribution is MVN or EC.

For this application, the RFCH estimator may be best. For MVN data, the RD_i from the RFCH estimator tend to have a higher correlation with the MD_i from the classical estimator than the RD_i from the FCH estimator, and the `cov.mcd` estimator may be inconsistent.

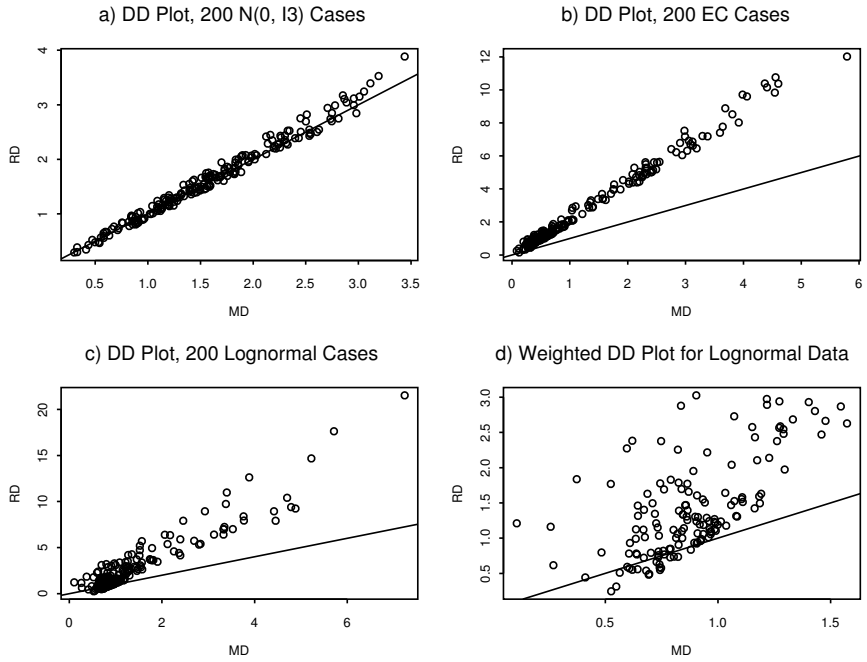

Figure 5.1: 4 DD Plots

Figure 5.1 shows the DD plots for 3 artificial data sets using `cov.mcd`. The DD plot for 200 $N_3(\mathbf{0}, \boldsymbol{I}_3)$ points shown in Figure 5.1a resembles the identity line. The DD plot for 200 points from the elliptically contoured distribution $0.6N_3(\mathbf{0}, \boldsymbol{I}_3) + 0.4N_3(\mathbf{0}, 25\ \boldsymbol{I}_3)$ in Figure 5.1b clusters about a line through the origin with a slope close to 2.0.

A *weighted DD plot* magnifies the lower left corner of the DD plot by omitting the cases with $\mathrm{RD}_i \geq \sqrt{\chi^2_{p,.975}}$. This technique can magnify features that are obscured when large RD_i's are present. If the distribution of $\boldsymbol{x}$ is EC with nonsingular $\boldsymbol{\Sigma}$, Proposition 5.1 implies that the correlation of the points in the weighted DD plot will tend to one and that the points will cluster about a line passing through the origin. For example, the plotted points in the weighted DD plot (not shown) for the non-MVN EC data of Figure 5.1b are highly correlated and still follow a line through the origin with a slope close to 2.0.

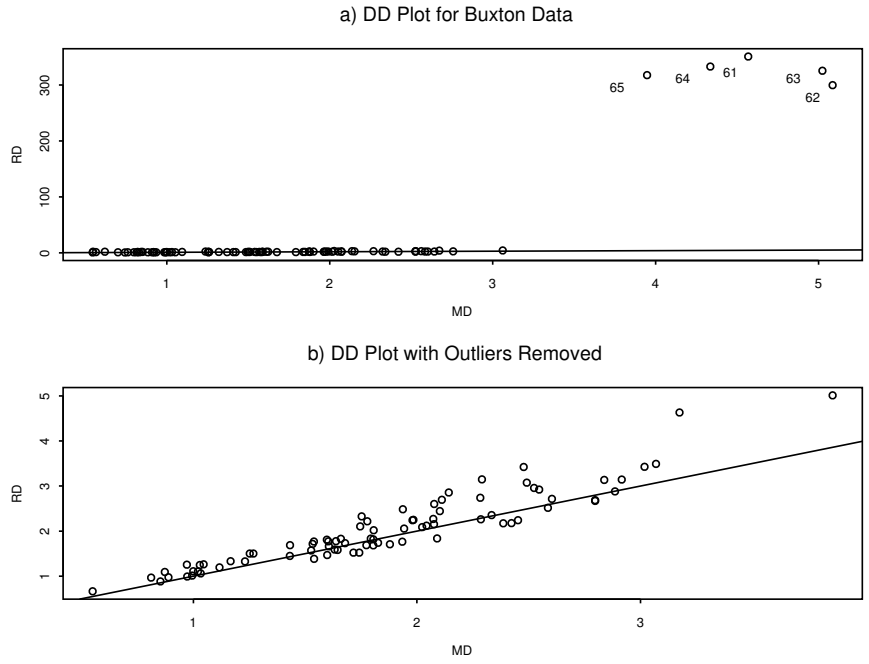

Figure 5.2: DD Plots for the Buxton Data

Figures 5.1c and 5.1d illustrate how to use the weighted DD plot. The ith case in Figure 5.1c is $(\exp(x_{i,1}), \exp(x_{i,2}), \exp(x_{i,3}))^T$ where $\boldsymbol{x}_i$ is the ith case in Figure 5.1a; ie, the marginals follow a lognormal distribution. The plot does not resemble the identity line, correctly suggesting that the distribution of the data is not MVN; however, the correlation of the plotted points is rather high. Figure 5.1d is the weighted DD plot where cases with $\mathrm{RD}_i \geq \sqrt{\chi^2_{3,.975}} \approx 3.06$ have been removed. Notice that the correlation of the plotted points is not close to one and that the best fitting line in Figure 5.1d may not pass through the origin. These results suggest that the distribution of $\boldsymbol{x}$ is not EC.

It is easier to use the DD plot as a diagnostic for a target distribution such as the MVN distribution than as a diagnostic for elliptical symmetry. If the data arise from the target distribution, then the DD plot will tend to be a useful diagnostic when the sample size n is such that the sample correlation coefficient in the DD plot is at least 0.80 with high probability.

As a diagnostic for elliptical symmetry, it may be useful to add the OLS line to the DD plot and weighted DD plot as a visual aid, along with numerical quantities such as the OLS slope and the correlation of the plotted points.

Numerical methods for transforming data towards a target EC distribution have been developed. Generalizations of the Box–Cox transformation towards a multivariate normal distribution are described in Velilla (1993). Alternatively, Cook and Nachtsheim (1994) offer a two-step numerical procedure for transforming data towards a target EC distribution. The first step simply gives zero weight to a fixed percentage of cases that have the largest robust Mahalanobis distances, and the second step uses Monte Carlo case reweighting with Voronoi weights.

Example 5.2. Buxton (1920, p. 232-5) gives 20 measurements of 88 men. We will examine whether the multivariate normal distribution is a plausible model for the measurements *head length, nasal height, bigonal breadth,* and *cephalic index* where one case has been deleted due to missing values. Figure 5.2a shows the DD plot. Five head lengths were recorded to be around 5 feet and are massive outliers. Figure 5.2b is the DD plot computed after deleting these points and suggests that the normal distribution is plausible. (The recomputation of the DD plot means that the plot is not a weighted DD plot which would simply omit the outliers and then rescale the vertical axis.)

The DD plot complements rather than replaces the numerical procedures. For example, if the goal of the transformation is to achieve a multivariate normal distribution and if the data points cluster tightly about the identity line, as in Figure 5.1a, then perhaps no transformation is needed. For the data in Figure 5.1c, a good numerical procedure should suggest coordinatewise log transforms. Following this transformation, the resulting plot shown in Figure 5.1a indicates that the transformation to normality was successful.

Application 5.2. The DD plot can be used to detect multivariate outliers. See Figures 4.2, 4.4, 5.2a and 5.7.

5.2 Robust Prediction Regions

Suppose that $(T_A, \boldsymbol{C}_A)$ is a good estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$. Section 5.1 showed that if $\boldsymbol{x}$ is multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, T_A estimates $\boldsymbol{\mu}$ and $\boldsymbol{C}_A/\tau^2$ estimates $\boldsymbol{\Sigma}$ where τ is given in Equation (5.2). Then $(T_R, \boldsymbol{C}_R) \equiv (T_A, \boldsymbol{C}_A/\tau^2)$ is an estimator of multivariate location and dispersion.

Suppose $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}_M, b\ \boldsymbol{S}_M)$ is the sample mean and scaled sample covariance matrix applied to some subset of the data. The classical and RMVN estimators satisfy this assumption. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\} \quad (5.7)$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)}. \quad (5.8)$$

A future observation (random vector) $\boldsymbol{x}_f$ is in the region (5.7) if $D_{\boldsymbol{x}_f} \leq h$.

A large sample $(1-\alpha)100\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n) \xrightarrow{P} 1 - \alpha$. Let $q_n = \min(1 - \alpha + 0.05, 1 - \alpha + p/n)$ for $\alpha > 0.1$ and

$$q_n = \min(1 - \alpha/2, 1 - \alpha + 10\alpha p/n), \quad \text{otherwise.}$$

$$\text{If } \ \mathrm{q_n} < 1 - \alpha + 0.001, \ \ \text{use } \ \mathrm{q_n} = 1 - \alpha. \quad (5.9)$$

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then (5.7) is a large sample $(1 - \alpha)100\%$ prediction regions if $h = D_{(up)}$ where $D_{(up)}$ is the q_nth sample quantile of the D_i where the D_i^2 are given by (3.12). If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ and $\boldsymbol{x}_f$ are iid from an EC distribution (with continuous decreasing g), then region (5.7) is asymptotically optimal in that its volume converges in probability to the volume of the minimum volume covering region $\{\boldsymbol{z} : (\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) \leq u_{1-\alpha}\}$ where $P(U \leq u_{1-\alpha}) = 1 - \alpha$ and U has pdf given by (3.10). The classical parametric MVN prediction region uses $MD_{\boldsymbol{x}_f} \leq \sqrt{\chi^2_{p,1-\alpha}}$.

Notice that for the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, if $\boldsymbol{C}^{-1}$ exists, then $100q_n\%$ of the n cases are in the prediction region, and $q_n \to 1 - \alpha$ even if $(T, \boldsymbol{C})$ is not a good estimator. Hence the coverage q_n of the data is robust to model assumptions. Of course the volume of the prediction region could be large if a poor estimator $(T, \boldsymbol{C})$ is used or if the $\boldsymbol{x}_i$ do not come from an elliptically contoured distribution. Also notice that $q_n = 1 - \alpha/2$ or $q_n = 1 - \alpha + 0.05$

for $n \leq 20p$ and $q_n \to 1 - \alpha$ as $n \to \infty$. If $q_n \equiv 1 - \alpha$, then (5.7) is a large sample prediction region, but taking q_n given by (5.9) improves the finite sample performance of the region. Taking $q_n \equiv 1 - \alpha$ does not take into account variability of $(T, \boldsymbol{C})$, and for small n the resulting prediction region tended to have undercoverage as high as $\min(0.05, \alpha/2)$. Using (5.9) helped reduce undercoverage for small n due to the unknown variability of $(T, \boldsymbol{C})$.

Three new prediction regions will be considered. The nonparametric region uses the classical estimator $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ and $h = D_{(up)}$. The semiparametric region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h = D_{(up)}$. The parametric MVN region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h^2 = \chi^2_{p,q_n}$ where $P(W \leq \chi^2_{p,\alpha}) = \alpha$ if $W \sim \chi^2_p$. All three regions are asymptotically optimal for MVN distributions with nonsingular $\boldsymbol{\Sigma}$. The first two regions are asymptotically optimal under the large class of EC distribution given by Assumption (E1) used in Theorem 4.8. For distributions with nonsingular covariance matrix $c_X\boldsymbol{\Sigma}$, the nonparametric region is a large sample $(1 - \alpha)100\%$ prediction region, but regions with smaller volume may exist.

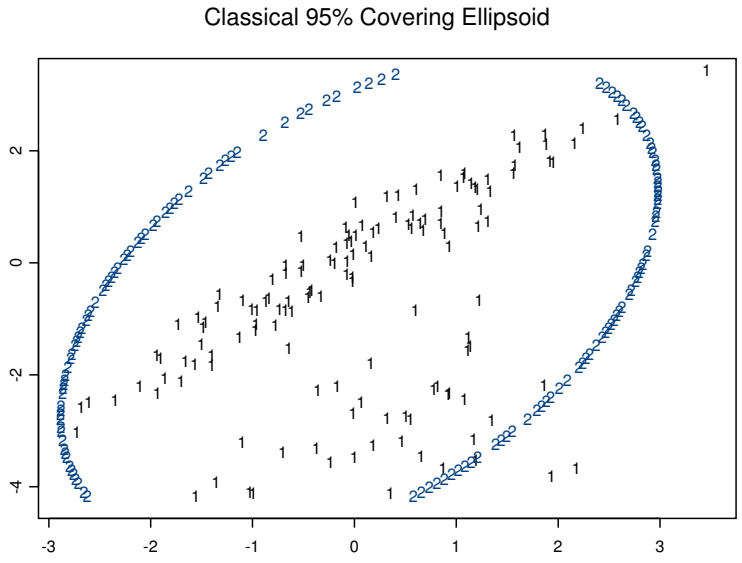

Figure 5.3: Artificial Bivariate Data

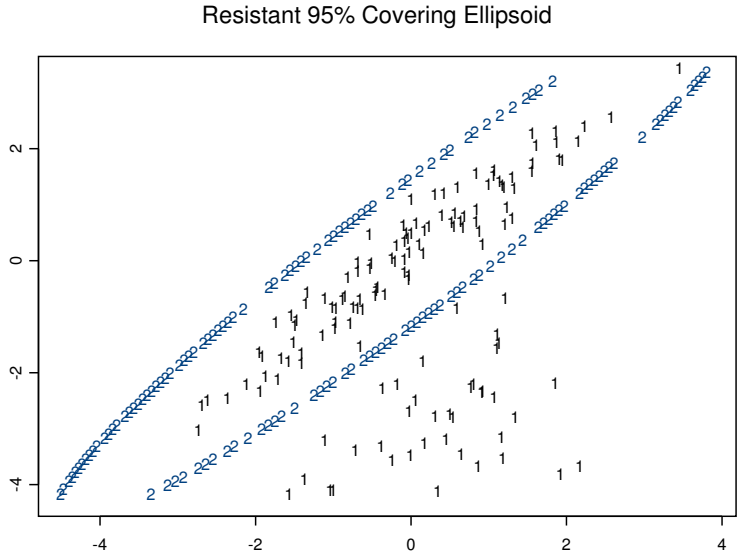

Figure 5.4: Artificial Data

Example 5.3. An artificial data set consisting of 100 iid cases from a

$$N_2\left(\left(\begin{array}{c}0\\0\end{array}\right), \left(\begin{array}{cc}1.49 & 1.4\\1.4 & 1.49\end{array}\right)\right)$$

distribution and 40 iid cases from a bivariate normal distribution with mean $(0, -3)^T$ and covariance $\boldsymbol{I}_2$. Figure 5.3 shows the classical ellipsoid (with $MD \leq \sqrt{\chi^2_{2,0.95}}$) that uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. The symbol "1" denotes the data while the symbol "2" is on the border of the covering ellipse. Notice that the classical parametric ellipsoid covers almost all of the data. Figure 5.4 displays the robust ellipsoid (using $RD \leq \sqrt{\chi^2_{2,0.95}}$) which contains most of the 100 "clean" cases and excludes the 40 outliers. Problem 5.5 recreates similar figures with the classical and RMVN estimators using $q_n = 0.95$.

Example 5.4. Buxton (1920) gives various measurements on 87 men including *height*, *head length*, *nasal height*, *bigonal breadth* and *cephalic index*. Five *heights* were recorded to be about 19mm and are massive outliers. First *height* and *nasal height* were used with $q_n = 0.95$. Figure 5.5 shows that the

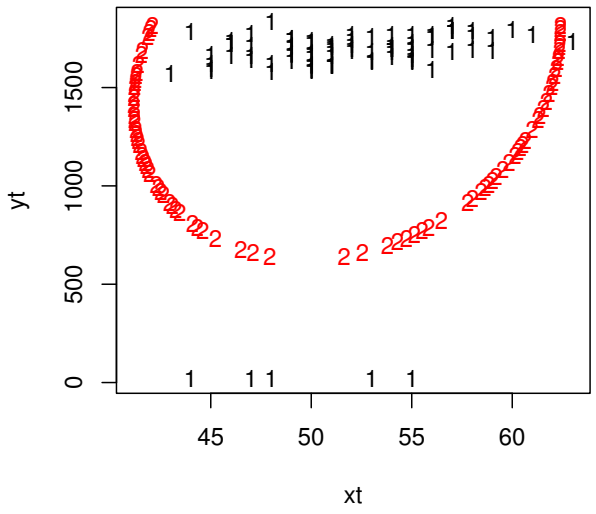

Figure 5.5: Ellipsoid is Inflated by Outliers

classical parametric prediction region (using $MD \leq \sqrt{\chi^2_{2,.95}}$) is quite large but does not include any of the outliers. Figure 5.6 shows that the parametric MVN prediction region (using $RD \leq \sqrt{\chi^2_{2,.95}}$) is not inflated by the outliers.

Next all 87 cases and 5 predictors were used. Figure 5.7 shows the RMVN DD plot with the identity line added as a visual aid. Points to the left of the vertical line are in the nonparametric large sample 90% prediction region. Points below the horizontal line are in the semiparametric region. The horizontal line at $RD = 3.33$ corresponding to the parametric MVN 90% region is obscured by the identity line. This region contains 78 of the cases. Since $n = 87$, the nonparametric and semiparametric regions used the 95th quantile. Since there were 5 outliers, this quantile was a linear combination of the largest clean distance and the smallest outlier distance. The semiparametric 90% region blows up unless the outlier proportion is small.

Figure 5.8 shows the DD plot and 3 prediction regions after the 5 outliers were removed. The classical and robust distances cluster about the identity line and the three regions are similar, with the parametric MVN region cutoff

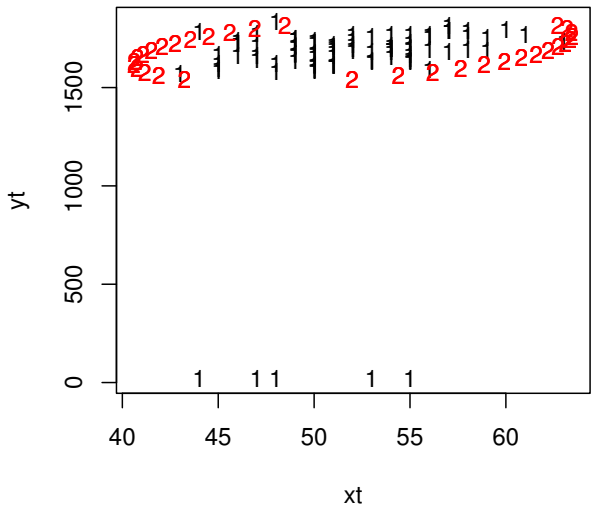

Figure 5.6: Ellipsoid Ignores Outliers

again at 3.33, slightly below the semiparametric region cutoff of 3.44.

Simulations for the prediction regions used $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{w}$ where $\boldsymbol{A} = diag(\sqrt{1}, ..., \sqrt{p})$, $\boldsymbol{w} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ (MVN), $\boldsymbol{w} \sim LN(\boldsymbol{0}, \boldsymbol{I}_p)$ where the marginals are iid lognormal(0,1), or $\boldsymbol{w} \sim MVT_p(1)$, a multivariate t distribution with 1 degree of freedom so the marginals are iid Cauchy(0,1). All simulations used 5000 runs and $\alpha = 0.1$.

For large n, the semiparametric and nonparametric regions are likely to have coverage near 0.90 because the coverage on the training sample is slightly larger than 0.9 and $\boldsymbol{x}_f$ comes from the same distribution as the $\boldsymbol{x}_i$. For $n = 10p$ and $2 \leq p \leq 40$, the semiparametric region had coverage near 0.9. The ratio of the volumes

$$\frac{h_i^p \sqrt{det(\boldsymbol{C}_i)}}{h_2^p \sqrt{det(\boldsymbol{C}_2)}}$$

was recorded where $i = 1$ was the nonparametric region, $i = 2$ was the semiparametric region, and $i = 3$ was the parametric MVN region. The volume ratio converges in probability to 1 for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data, and the ratio converges to 1 for $i = 1$ if Assumption (E1) holds. The parametric MVN

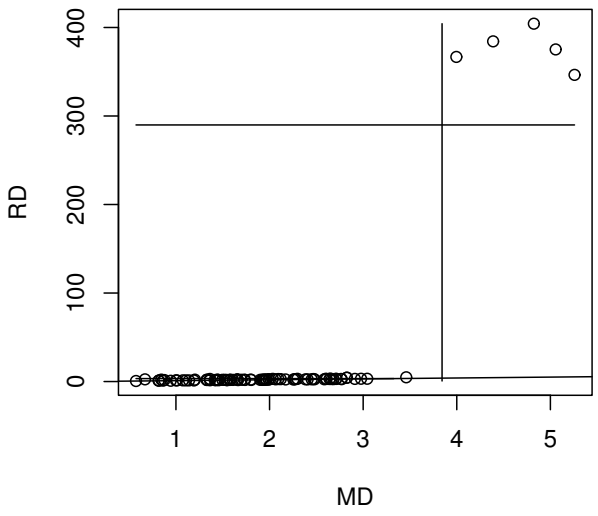

Figure 5.7: Prediction Regions for Buxton Data

region often had coverage much lower than 0.9 with a volume ratio near 0, recorded as 0+. The volume ratio tends to be tiny when the coverage is much less than the nominal value 0.9. For $10p \leq n \leq 20p$, the nonparametric region often had good coverage and volume ratio near 0.5.

Table 5.2: Coverages for 90% Prediction Regions

$\boldsymbol{w}$ dist	n	p	ncov	scov	mcov	voln	volm
MVN	600	30	0.906	0.919	0.902	0.503	0.512
MVN	1500	30	0.899	0.899	0.900	1.014	1.027
LN	1000	10	0.903	0.906	0.567	0.659	0+
MVT(1)	1000	10	0.914	0.914	0.541	22634.3	0+

Simulations and Table 5.2 suggest that for MVN data, the coverages (ncov, scov and mcov) for the 3 regions are near 90% for $n = 20p$ and that the volume ratios voln and volm are near 1 for $n = 50p$. With fewer than

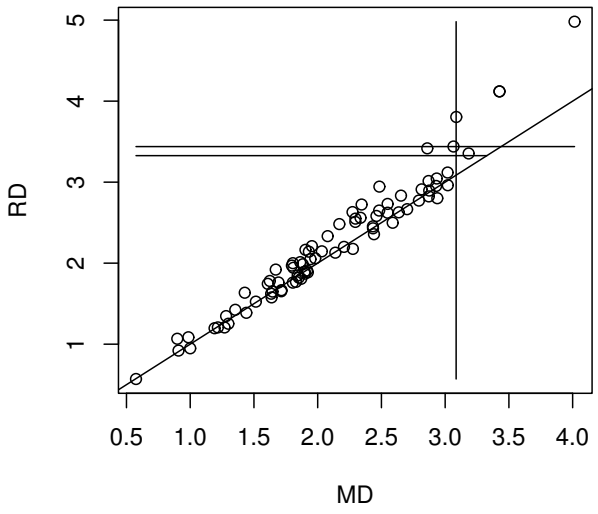

Figure 5.8: Prediction Regions for Buxton Data without Outliers

5000 runs, this result held for $2 \leq p \leq 80$. For the non–elliptically contoured LN data, the nonparametric region had voln well under 1, but the volume ratio blew up for $\boldsymbol{w} \sim MVT_p(1)$.

5.3 Summary

1) For $h > 0$, the hyperellipsoid $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$. A future observation (random vector) $\boldsymbol{x}_f$ is in this region if $D_{\boldsymbol{x}_f} \leq h$. A large sample $(1-\alpha)100\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n) \xrightarrow{P} 1 - \alpha$ where $0 < \alpha < 1$.

2) The classical $(1-\alpha)100\%$ large sample prediction region is $\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi^2_{p,1-\alpha}\}$ and works well if n is large and the data are iid MVN.

3) Let $q_n = \min(1 - \alpha + 0.05, 1 - \alpha + p/n)$ for $\alpha > 0.1$ and $q_n = \min(1 - \alpha/2, 1 - \alpha + 10\alpha p/n)$, otherwise. If $q_n < 1 - \alpha + 0.001$, set $q_n = 1 - \alpha$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then $\{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$ is a large sample $(1-\alpha)100\%$ prediction regions if $h = D_{(up)}$ where $D_{(up)}$ is the q_nth sample quantile of the D_i. The nonparametric prediction region uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$

and the semiparametric prediction region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$. The parametric MVN prediction region
$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(T, \boldsymbol{C}) \leq \chi^2_{p,q_n}\}$ also uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$.

4) These 3 regions can be displayed in an RMVN DD plot with cases in the nonparametric region corresponding to points to the left of the vertical line corresponding to $D_{(up)}(\overline{\boldsymbol{x}}, \boldsymbol{S})$. Cases in the semiparametric region correspond to points below the horizontal line corresponding to $D_{(up)}(T_{RMVN}, \boldsymbol{C}_{RMVN})$ while cases in the parametric MVN region correspond to points below the horizontal line corresponding to $\sqrt{\chi^2_{p,q_n}}$. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_x$. The three prediction regions are asymptotically optimal if the data is MVN. The semiparametric and nonparametric prediction regions are asymptotically optimal on a large class of EC distributions and the nonparametric prediction region is a large sample $100(1-\alpha)\%$ prediction region, although large sample prediction regions with smaller volume may exist.

5) Suppose m independent large sample $100(1-\alpha)\%$ prediction regions are made where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from the same distribution for each of the m runs. Let Y count the number of times $\boldsymbol{x}_f$ is in the prediction region. Then $Y \sim$ binomial $(m, 1-\alpha_n)$ where $1-\alpha_n$ is the true coverage and $1-\alpha_n \to 1-\alpha$ as $n \to \infty$. Simulation can be used to see if the true or actual coverage $1-\alpha_n$ is close to the nominal coverage $1-\alpha$. A prediction region with $1-\alpha_n < 1-\alpha$ is liberal and a region with $1-\alpha_n > 1-\alpha$ is conservative. It is better to be conservative by 5% than liberal by 5%. Parametric prediction regions tend to have large undercoverage and so are too liberal.

6) For prediction regions, want $n > 10p$ for the nonparametric prediction region and $n > 20p$ for the semiparametric prediction region.

5.4 Complements

The first section of this chapter followed Olive (2002) closely. The DD plot can be used to diagnose elliptical symmetry, to detect outliers, and to assess the success of numerical methods for transforming data towards an elliptically contoured distribution. Since many statistical methods assume that the underlying data distribution is Gaussian or EC, there is an enormous literature on numerical tests for elliptical symmetry. Bogdan (1999), Czörgö (1986) and Thode (2002) provide references for tests for multivariate normal-

ity while Koltchinskii and Li (1998) and Manzotti, Pérez and Quiroz (2002) have references for tests for elliptically contoured distributions.

There are few practical competitors for the Olive (2013b) prediction regions in Section 5.2. Parametric regions such as the classical region for multivariate normal data tend to have severe undercoverage because the data rarely follows the parametric distribution. Procedures that use brand name high breakdown multivariate location and dispersion estimators take too long to compute for $p > 2$.

5.5 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

5.1*. If X and Y are random variables, show that

$$\mathrm{Cov(X, Y)} = [\mathrm{Var(X + Y)} - \mathrm{Var(X - Y)}]/4.$$

R/Splus Problems

Warning: Use the command *source("G:/mpack.txt")* **to download the programs. See Preface or Section 15.2.** Typing the name of the `mpack` function, eg *ddplot*, will display the code for the function. Use the `args` command, eg *args(ddplot)*, to display the needed arguments for the function.

5.2. a) Download the program `ddsim`. (In *R*, type the command *library(MASS).*)

b) Using the function *ddsim* for $p = 2, 3, 4$, determine how large the sample size n should be in order for the RFCH DD plot of n $N_p(\mathbf{0}, \boldsymbol{I}_p)$ cases to cluster tightly about the identity line with high probability. Table your results. (Hint: type the command *ddsim(n=20,p=2)* and increase n by 10 until most of the 20 plots look linear. Then repeat for $p = 3$ with the n that worked for $p = 2$. Then repeat for $p = 4$ with the n that worked for $p = 3$.)

5.3. a) Download the program `corrsim`. (In *R*, type the command *library(MASS).*)

b) A numerical quantity of interest is the correlation between the MD_i and RD_i in a RFCH DD plot that uses n $N_p(\mathbf{0}, \boldsymbol{I}_p)$ cases. Using the function

corrsim for $p = 2, 3, 4$, determine how large the sample size n should be in order for 9 out of 10 correlations to be greater than 0.9. (Try to make n small.) Table your results. (Hint: type the command *corrsim(n=20,p=2,nruns=10)* and increase n by 10 until 9 or 10 of the correlations are greater than 0.9. Then repeat for $p = 3$ with the n that worked for $p = 2$. Then repeat for $p = 4$ with the n that worked for $p = 3$.)

5.4***.** a) Download the `ddplot` function. (In *R*, type the command *library(MASS).*)

b) Using the following commands to make generate data from the EC distribution $(1 - \epsilon)N_p(\mathbf{0}, \boldsymbol{I}_p) + \epsilon N_p(\mathbf{0}, 25\ \boldsymbol{I}_p)$ where $p = 3$ and $\epsilon = 0.4$.

```
n <- 400
p <- 3
eps <- 0.4
x <- matrix(rnorm(n * p), ncol = p, nrow = n)
zu <- runif(n)
x[zu < eps,] <- x[zu < eps,]*5
```

c) Use the command `ddplot(x)` to make a DD plot and include the plot in *Word.* What is the slope of the line followed by the plotted points?

5.5. a) Download the `ellipse` function.

b) Use the following commands to create a bivariate data set with outliers and to obtain a classical and robust RMVN covering ellipsoid. Include the two plots in *Word.*

```
> simx2 <- matrix(rnorm(200),nrow=100,ncol=2)
> outx2 <- matrix(10 + rnorm(80),nrow=40,ncol=2)
> outx2 <- rbind(outx2,simx2)
> ellipse(outx2)

> zout <- covrmvn(outx2)
> ellipse(outx2,center=zout$center,cov=zout$cov)
```

5.6. a) Download the function `mplot`.

b) Enter the commands in Problem 5.4b to obtain a data set x. The function `mplot` makes a plot without the RD_i and the slope of the resulting line is of interest.

c) Use the command `mplot(x)` and place the resulting plot in *Word*.

d) Do you prefer the DD plot or the mplot? Explain.

5.7 a) Download the function `wddplot`.

b) Enter the commands in Problem 5.4b to obtain a data set x.

c) Use the command `wddplot(x)` and place the resulting plot in *Word*.

5.8. Use the *R* command *source("G:/mrobdata.txt")* then *ddplot4(buxx,alpha=0.2)* and put the plot in *Word*. The Buxton data has 5 outliers, $p = 4$, and $n = 87$, so the 80% prediction regions use $1 - \alpha + p/n = 0.846$ percentiles. The output shows that the cutoffs are 2.527, 2.734 and 2.583 for the nonparametric, semiparametric and robust parametric prediction regions. The two horizontal lines that correspond to the robust distances are obscured by the identity line.

5.9. a) Type the *R* command `predsim()` and paste the output into *Word*.

This computes $\boldsymbol{x}_i \sim N_4(\boldsymbol{0}, diag(1, 2, 3, 4))$ for $i = 1, ..., 100$ and $\boldsymbol{x}_f = \boldsymbol{x}_{101}$. One hundred such data sets are made, and ncvr, scvr, mcvr counts the number of times $\boldsymbol{x}_f$ was in the nonparametric, semiparametric and parametric MVN 90% prediction regions. The volumes of the prediction regions are computed and voln, vols and volm are the average ratio of the volume of the ith prediction region over that of the semiparametric region. Hence vols is always equal to 1. For multivariate normal data, these ratios should converge to 1 as $n \to \infty$. Were the three coverages near 90%?

5.10. Tests for covariance matrices are very nonrobust to nonnormality. Let a plot of x versus y have x on the horizontal axis and y on the vertical axis. A good diagnostic is to use the DD plot. So a diagnostic for $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_0$ is to plot $D_i(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i(\overline{\boldsymbol{x}}, \boldsymbol{\Sigma}_0)$ for $i = 1, ..., n$. If $n > 10p$ and H_0 is true, then the plotted points in the DD plot should cluster tightly about the identity line.

a) A test for sphericity is a test of $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = d\boldsymbol{I}_p$ for some unknown constant $d > 0$. Make a "DD plot" of $D_i^2(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i^2(\overline{\boldsymbol{x}}, \boldsymbol{I}_p)$. If $n > 10p$ and H_0 is true, then the plotted points in the "DD plot" should cluster tightly about the line through the origin with slope d. Use the *R* commands for this part and paste the plot into *Word*. The simulated data set has $\boldsymbol{x}_i \sim N_{10}(\boldsymbol{0}, 100\boldsymbol{I}_{10})$ where $n = 100$ and $p = 10$. Do the plotted points follow

a line through the origin with slope 100?

b) Now suppose there are k samples, and want to test $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}_1} = \cdots = \boldsymbol{\Sigma}_{\boldsymbol{x}_k}$, that is, all k populations have the same covariance matrix. As a diagnostic, make a DD plot of $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ versus $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_{pool})$ for $j = 1, ..., k$ and $i = 1, ..., n_i$. If each $n_i > 10p$ and H_0 is true, what line will the plotted points cluster about in each of the k DD plots?

Chapter 6

Principal Component Analysis

6.1 Introduction

Principal component analysis (PCA) is used to explain the dispersion structure with a few linear combinations of the original variables, called principal components. These linear combinations are uncorrelated if $\boldsymbol{S}$ or $\boldsymbol{R}$ is used as the dispersion matrix. The analysis is used for data reduction and interpretation. The notation $\boldsymbol{e}_j$ will be used for orthonormal eigenvectors: $\boldsymbol{e}_j^T \boldsymbol{e}_j = 1$ and $\boldsymbol{e}_j^T \boldsymbol{e}_k = 0$ for $j \neq k$. The eigenvalue eigenvector pairs of a matrix $\boldsymbol{\Sigma}$ will be $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. The eigenvalue eigenvector pairs of a matrix $\hat{\boldsymbol{\Sigma}}$ will be $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), ..., (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. The generalized correlation matrix defined below is the correlation matrix when second moments exist if $\boldsymbol{\Sigma} = c \text{ Cov}(\boldsymbol{x})$ for some constant $c > 0$.

Definition 6.1. Let $\boldsymbol{\Sigma} = ((\sigma_{ij}))$ be a positive definite symmetric $p \times p$ dispersion matrix. A *generalized correlation matrix* $\boldsymbol{\rho} = \big(\big(\rho_{ij}\big)\big)$ where

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

The following theorem holds since the eigenvalues and generalized correlation matrix are continuous functions of $\boldsymbol{\Sigma}$. Also see Theorem 3.29. When the distribution of the $\boldsymbol{x}_i$ is unknown, then a good dispersion estimator estimates $c\boldsymbol{\Sigma}$ on a large class of distributions where $c > 0$ depends on the unknown distribution of $\boldsymbol{x}_i$. For example, if the $\boldsymbol{x}_i \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, then the sample covariance matrix $\boldsymbol{S}$ estimates $\text{Cov}(\boldsymbol{x}) = c_X \boldsymbol{\Sigma}$.

Theorem 6.1. Suppose the dispersion matrix $\boldsymbol{\Sigma}$ has eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Suppose $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ for some constant $c > 0$. Let the eigenvalue eigenvector pairs of $\hat{\boldsymbol{\Sigma}}$ be $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), ..., (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. Then $\hat{\lambda}_j(\hat{\boldsymbol{\Sigma}}) \xrightarrow{P} c\lambda_j(\boldsymbol{\Sigma}) = c\lambda_j$, $\hat{\boldsymbol{\rho}} \xrightarrow{P} \boldsymbol{\rho}$ and $\hat{\lambda}_j\left(\hat{\boldsymbol{\rho}}\right) \xrightarrow{P} \lambda_j\left(\boldsymbol{\rho}\right)$ where $\lambda_j(\boldsymbol{A})$ is the jth eigenvalue of $\boldsymbol{A}$ for $j = 1, ..., p$.

Eigenvectors $\boldsymbol{e}_j$ are not continuous functions of $\boldsymbol{\Sigma}$, and if $\boldsymbol{e}_j$ is an eigenvector of $\boldsymbol{\Sigma}$ then so is $-\boldsymbol{e}_j$. The software produces $\hat{\boldsymbol{e}}_j$ which sometimes approximates $\boldsymbol{e}_j$ and sometimes approximates $-\boldsymbol{e}_j$ if the eigenvalue λ_j is unique, since then the set of eigenvectors corresponding to λ_j has the form $a\boldsymbol{e}_j$ for any nonzero constant a. The situation becomes worse if some of the eigenvalues are equal, since the possible eigenvectors then span a space of dimension equal to the multiplicity of the eigenvalue. Hence if the multiplicity is two and both $\boldsymbol{e}_j$ and $\boldsymbol{e}_k$ are eigenvectors corresponding to the eigenvalue λ_i, then $\boldsymbol{e}_i = \boldsymbol{x}_i/\|\boldsymbol{x}_i\|$ is also an eigenvector corresponding to λ_i where $\boldsymbol{x}_i = a_j\boldsymbol{e}_j + a_k\boldsymbol{e}_k$ for constants a_j and a_k which are not both equal to 0. The software produces $\hat{\boldsymbol{e}}_j$ and $\hat{\boldsymbol{e}}_k$ that are approximately in the span of $\boldsymbol{e}_j$ and $\boldsymbol{e}_k$ for large n by the following theorem, which also shows that $\hat{\boldsymbol{e}}_i$ is asymptotically an eigenvector of $\boldsymbol{\Sigma}$ in that $(\boldsymbol{\Sigma} - \lambda_i)\hat{\boldsymbol{e}}_i \xrightarrow{P} \boldsymbol{0}$. It is possible that $\hat{\boldsymbol{e}}_{i,n}$ is arbitrarily close to $\boldsymbol{e}_i$ for some values of n and arbitrarily close to $-\boldsymbol{e}_i$ for other values of n so that $\hat{\boldsymbol{e}}_i \equiv \hat{\boldsymbol{e}}_{i,n}$ oscillates and does not converge in probability to either $\boldsymbol{e}_i$ or $-\boldsymbol{e}_i$.

Theorem 6.2. Assume the $p \times p$ symmetric dispersion matrix $\boldsymbol{\Sigma}$ is positive definite.

a) If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\Sigma}$, then $\hat{\boldsymbol{\Sigma}}\boldsymbol{e}_i - \hat{\lambda}_i\boldsymbol{e}_i \xrightarrow{P} \boldsymbol{0}$.

b) If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\Sigma}$, then $\boldsymbol{\Sigma}\hat{\boldsymbol{e}}_i - \lambda_i\hat{\boldsymbol{e}}_i \xrightarrow{P} \boldsymbol{0}$.

If $\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} = O_P(n^{-\delta})$ where $0 < \delta \leq 0.5$, then

c) $\lambda_i\boldsymbol{e}_i - \hat{\boldsymbol{\Sigma}}\boldsymbol{e}_i = O_P(n^{-\delta})$, and

d) $\hat{\lambda}_i\hat{\boldsymbol{e}}_i - \boldsymbol{\Sigma}\hat{\boldsymbol{e}}_i = O_P(n^{-\delta})$.

e) If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ for some constant $c > 0$, and if the eigenvalues $\lambda_1 > \cdots > \lambda_p > 0$ of $\boldsymbol{\Sigma}$ are unique, then the absolute value of the correlation of $\hat{\boldsymbol{e}}_j$ with $\boldsymbol{e}_j$ converges to 1 in probability: $|\text{corr}(\hat{\boldsymbol{e}}_j, \boldsymbol{e}_j)| \xrightarrow{P} 1$.

Proof. a) $\hat{\boldsymbol{\Sigma}}\boldsymbol{e}_i - \hat{\lambda}_i\boldsymbol{e}_i \xrightarrow{P} \boldsymbol{\Sigma}\boldsymbol{e}_i - \lambda_i\boldsymbol{e}_i = \boldsymbol{0}$.

b) Note that $(\boldsymbol{\Sigma} - \lambda_i\boldsymbol{I})\hat{\boldsymbol{e}}_i = [(\boldsymbol{\Sigma} - \lambda_i\boldsymbol{I}) - (\hat{\boldsymbol{\Sigma}} - \hat{\lambda}_i\boldsymbol{I})]\hat{\boldsymbol{e}}_i = o_P(1)O_P(1) \xrightarrow{P} \boldsymbol{0}$.

c) $\lambda_i \boldsymbol{e}_i - \hat{\boldsymbol{\Sigma}}\boldsymbol{e}_i = \boldsymbol{\Sigma}\boldsymbol{e}_i - \hat{\boldsymbol{\Sigma}}\boldsymbol{e}_i = O_P(n^{-\delta})$.
d) $\hat{\lambda}_i \hat{\boldsymbol{e}}_i - \boldsymbol{\Sigma}\hat{\boldsymbol{e}}_i = \hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{e}}_i - \boldsymbol{\Sigma}\hat{\boldsymbol{e}}_i = O_P(n^{-\delta})$.
e) Note that a) and b) hold if $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\Sigma}$ is replaced by $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$. Hence for large n, $\hat{\boldsymbol{e}}_i \equiv \hat{\boldsymbol{e}}_{i,n}$ is arbitrarily close to either $\boldsymbol{e}_i$ or $-\boldsymbol{e}_i$, and the result follows.

Rule of thumb 6.1. To use PCA, assume the DD plot and subplots of the scatterplot matrix are linear. Want $n > 10p$ for classical PCA and $n > 20p$ for robust PCA that uses FCH, RFCH or RMVN. For classical PCA, use the correlation matrix $\boldsymbol{R}$ instead of the covariance matrix $\boldsymbol{S}$ if $\max_{i=1,...,p} S_i^2 / \min_{i=1,...,p} S_i^2 > 2$. If $\boldsymbol{S}$ is used, also do a PCA using $\boldsymbol{R}$.

The trace of a matrix $\boldsymbol{A}$ is the sum of the diagonal elements of $\boldsymbol{A}$ and the sum of the eigenvalues of $\boldsymbol{A}$. If $\boldsymbol{A}$ is a $p \times p$ matrix, then $\text{trace}(\boldsymbol{A}) = tr(\boldsymbol{A}) = \sum_{i=1}^{p} \boldsymbol{A}_{ii} = \sum_{i=1}^{p} \lambda_i$. Note that $tr(\text{Cov}(\boldsymbol{x})) = \sigma_1^2 + \cdots + \sigma_p^2$ and $tr(\hat{\boldsymbol{\rho}}) = p$.

Definition 6.2. Let dispersion estimator $\hat{\boldsymbol{\Sigma}}$ have eigenvalue eigenvector pairs $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), ..., (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. Then the p *principal components* corresponding to the jth case $\boldsymbol{x}_j$ are $Z_{j1} = \hat{\boldsymbol{e}}_1^T \boldsymbol{x}_j, ..., Z_{jp} = \hat{\boldsymbol{e}}_p^T \boldsymbol{x}_j$. Let the vector $\boldsymbol{z}_j = (Z_{j1}, ..., Z_{jp})^T$. The *proportion of the trace explained* by the first kth principal components is $\sum_{i=1}^{k} \hat{\lambda}_i / \sum_{j=1}^{p} \hat{\lambda}_j = \sum_{i=1}^{k} \hat{\lambda}_i / tr(\hat{\boldsymbol{\Sigma}})$. When a correlation or covariance matrix is being estimated, "trace" is replaced by "variance." The population analogs use the dispersion matrix $\boldsymbol{\Sigma}$ with eigenvalue eigenvector pairs $(\lambda_i, \boldsymbol{e}_i)$ for $i = 1, ..., p$. The population principal components corresponding to the j case are $Y_{ji} = \boldsymbol{e}_i^T \boldsymbol{x}_j$, and $Z_{ji} = \hat{Y}_{ji}$ for $i = 1, ..., p$.

Note that the principal components can be collected into an $n \times p$ data matrix

$$\boldsymbol{Z} = \begin{bmatrix} Z_{1,1} & Z_{1,2} & \dots & Z_{1,p} \\ Z_{2,1} & Z_{2,2} & \dots & Z_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n,1} & Z_{n,2} & \dots & Z_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{z}_1^T \\ \vdots \\ \boldsymbol{z}_n^T \end{bmatrix}.$$

Then $\boldsymbol{u}_i$ corresponds to the ith principal component. A plot of the second principal component versus the first principal component can be useful.

The data matrix $\boldsymbol{W}$ corresponds to the usual axes where $\boldsymbol{e}_i$ is a vector of zeroes except for a one in the ith position. Hence the ith axis corresponds to

the ith variable X_i. The data matrix $\boldsymbol{Z}$ corresponds to axes that are parallel to the axes of the hyperellipsoid corresponding to the dispersion matrix $\hat{\boldsymbol{\Sigma}}$. These axes are a rotation of the usual axes about the origin.

If $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}$, then the definition of the estimated proportion of the total population variance may make little sense if the variables are measured on different scales. Assume the population covariance matrix is I_2. Then $\lambda_j/(\lambda_1 + \lambda_2) = 0.5$, but if x_j is multiplied by 3 then $V(x_j) = 9 = \lambda_j$, and $\lambda_j/(\lambda_1 + \lambda_2) = 0.9$. Then x_j seems much more important than the other variable just by scaling. This is why rule of thumb 6.1 says $\boldsymbol{R}$ should be used instead of $\boldsymbol{S}$ if $\max_{i=1,...,p} S_i^2 / \min_{i=1,...,p} S_i^2 > 2$.

Examine Theorems 2.4, 2.5 and Figure 2.1. The hyperellipsoid $\{\boldsymbol{x}|D_{\boldsymbol{x}}^2 \leq h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq h^2\}$, where $h^2 = u_{1-\alpha}$ and $P(U \leq u_{1-\alpha}) = 1 - \alpha$, is the highest density region covering $1 - \alpha$ of the mass for an elliptically contoured distribution. The hyperellipsoid is centered at $\boldsymbol{\mu}$. If $\boldsymbol{\mu} = \boldsymbol{0}$, then points at squared distance $\boldsymbol{w}^T\boldsymbol{S}^{-1}\boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors $\boldsymbol{e}_i$ where the half length in the direction of $\boldsymbol{e}_i$ is $h\sqrt{\lambda_i}$.

The projection vector of a vector $\boldsymbol{x}$ onto a vector $\boldsymbol{e}$ is

$$\frac{\boldsymbol{e}\boldsymbol{e}^T\boldsymbol{x}}{\boldsymbol{e}^T\boldsymbol{e}}.$$

Hence if $\boldsymbol{e}^T\boldsymbol{e} = 1$, the projection vector is $\boldsymbol{v} = [\boldsymbol{e}^T\boldsymbol{x}]\boldsymbol{e}$ and $\|\boldsymbol{v}\| = |\boldsymbol{e}^T\boldsymbol{x}|$. So $\boldsymbol{e}^T\boldsymbol{x}$ is the signed length of the projection vector of $\boldsymbol{x}$ onto $\boldsymbol{e}$, and $\boldsymbol{e}^T\boldsymbol{x}$ is called the (scalar) projection of $\boldsymbol{x}$ onto $\boldsymbol{e}$.

The $\boldsymbol{e}_i$ are the directions of the axes through the origin that are parallel to the axes of the hyperellipsoid. Suppose $\boldsymbol{\mu} = \boldsymbol{0}$. Then the ith principle component is the linear combination of the predictors that is the projection on the ith axis of the hyperellipsoid. That is, get the projection vectors of the $\boldsymbol{x}_i$ onto $\boldsymbol{e}_i$ and find their signed lengths $\boldsymbol{e}_i^T\boldsymbol{x}_i$ from the origin. Then these scalars form the ith principal components corresponding to the n data cases $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. So the first principal component is the projection on the major axis, the second principal component is the projection on the next longest axis, ..., the pth principal component is the projection on the minor axis. The axes are orthogonal, so the directions $\boldsymbol{e}_i$ are orthogonal.

When $\boldsymbol{\mu} \neq \boldsymbol{0}$ the projections on $\boldsymbol{e}_i$ are projections on the axes through the origin that are parallel to the axes of the hyperellipsoid. Figure 2.1 shows two ellipsoids where $p = 2$.

The first k principal components can be regarded as a good k dimensional approximation to the p dimensional data. Suppose the data cloud approximates the hyperellipsoid $\{\boldsymbol{x}|D^2_{\boldsymbol{x}} \leq h^2\}$ where $h^2 = D^2_{(n)}$, the largest squared distance, so the hyperellipsoid contains all of the data. Then a good one dimensional approximation is the projection on the major axis since this captures the dimension with the greatest variability or dispersion as measured by $\boldsymbol{\Sigma}$. A good two dimensional approximation uses the projection on the major axis and the projection on the next largest axis since these are the two orthogonal directions where the two projections have the greatest variability. Following Mardia, Kent and Bibby (1979, p. 220), if $\boldsymbol{S}$ (with centered data) or $\boldsymbol{R}$ is used as the dispersion matrix, then the vector space spanned by the first k principal components has smaller mean square deviation from the p variables than any other $k-$dimensional subspace.

Since $\boldsymbol{Z}$ represents a new coordinate system, the ith case $\boldsymbol{x}_i = (\boldsymbol{x}_i^T \hat{\boldsymbol{e}}_i)\hat{\boldsymbol{e}}_1 + \cdots + (\boldsymbol{x}_i^T \hat{\boldsymbol{e}}_p)\hat{\boldsymbol{e}}_p = Z_{i,1}\hat{\boldsymbol{e}}_1 + \cdots + Z_{i,p}\hat{\boldsymbol{e}}_p$. Also $\boldsymbol{x}_i = \tilde{\boldsymbol{x}}_i(k) + \boldsymbol{r}_i(k)$ where $\tilde{\boldsymbol{x}}_i(k) = \sum_{j=1}^{k} Z_{i,j}\hat{\boldsymbol{e}}_j$ and the residual vector $\boldsymbol{r}_i(k) = \sum_{j=k+1}^{p} Z_{i,j}\hat{\boldsymbol{e}}_j$. The squared length of the residual vector is $\|\boldsymbol{r}_i(k)\|^2 = \boldsymbol{r}_i(k)^T\boldsymbol{r}_i(k) = Z^2_{i,k+1} + \cdots + Z^2_{i,p}$.

Suppose $\boldsymbol{S}$ or $\boldsymbol{R}$ is used as the as the dispersion matrix and that $T = \boldsymbol{0}$ so the hyperellisoid is centered at the origin. Following Kendall (1980, p. 17), the eigenvector corresponding to the largest eigenvalue determines the major axis of the hyperellipsoid. This axis forms the line through the origin such that the sum of squared distances from the n data points $\boldsymbol{x}_i$ to this line is a minimum. If the data points are projected onto a hyperplane perpendicular to the major axis line, then the eigenvector corresponding to the next largest eigenvalue determines the second longest axis of the hyperellipsoid, and this axis is the line through the origin in the hyperplane that minimizes the sum of squared distances, and so on.

When the covariance matrix is used, that the first principal component $\boldsymbol{e}_1^T\boldsymbol{x}$ is the linear combination $\boldsymbol{g}_1^T\boldsymbol{x}$ that maximizes $\mathrm{Var}(\boldsymbol{g}_1^T\boldsymbol{x})$ subject to $\boldsymbol{g}_1^T\boldsymbol{g}_1 = 1$, while the jth principal component is the linear combination $\boldsymbol{g}_j^T\boldsymbol{x}$ that maximizes $\mathrm{Var}(\boldsymbol{g}_j^T\boldsymbol{x})$ subject to $\boldsymbol{g}_j^T\boldsymbol{g}_j = 1$ and $\mathrm{Cov}(\boldsymbol{g}_j^T\boldsymbol{x}, \boldsymbol{g}_k^T\boldsymbol{x}) = 0$ for $k < j$. This result can be proved using Theorem 1.1.

Definition 6.3. A *scree plot* is a plot of component number versus eigenvalue.

Dimension reduction involves using the first k principal components to

approximate the data matrix without losing much important information. Want the proportion of the trace explained by the first k principal components to be higher than 0.8 or 0.9.

Rule of thumb 6.2. The value of k should be such that

$$\frac{\sum_{i=1}^{k} \hat{\lambda}_i}{\sum_{i=1}^{p} \hat{\lambda}_i} \geq 0.9.$$

The scree plot is also useful for choosing k since often there is a sharp bend in the scree plot when the components are no longer important. See Cattell (1966).

Following Johnson and Wichern (1988, p. 343, 347), let $\boldsymbol{x} = (X_1, ..., X_p)$ be a random vector such that the $\boldsymbol{x}_i$ and $\boldsymbol{x}$ have the same distribution. Let $Y_i = \boldsymbol{e}_i^T \boldsymbol{x}$ be the population principal components based on the covariance matrix $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$. Let $\boldsymbol{e}_i = (e_{1i}, ..., e_{pi})^T$. Then e_{ki} is proportional to the correlation between Y_i and X_k, in fact,

$$\text{corr}(Y_i, X_k) = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

for $i, k = 1, ..., p$. If the correlation matrix $\boldsymbol{\rho}$ is used instead of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then $\text{corr}(Y_i, X_k) = e_{ki}\sqrt{\lambda_i}$.

Following Johnson and Wichern (1988, p. 252-253), some software that uses $\boldsymbol{S}$ or $\boldsymbol{R}$ centers the data by using $\boldsymbol{x}_i - \overline{\boldsymbol{x}}$. Centering does not change $\boldsymbol{S}$ or $\boldsymbol{R}$ but makes the ith principal component equal to $\hat{\boldsymbol{e}}_i^T(\boldsymbol{x} - \overline{\boldsymbol{x}})$ for observation $\boldsymbol{x}$.

Warning: If $\hat{\lambda}_p \approx 0$, then $\hat{\boldsymbol{\Sigma}}$ is nearly singular, and there could be an unnoticed linear dependency in the data set, eg $X_p \approx \sum_{i=1}^{p-1} c_i X_i$. Then one or more of the variables is redundant and should be deleted. Following Johnson and Wichern (1988, p. 360), suppose $p = 4$ and X_1, X_2 and X_3 are midterm exam scores while X_4 is the total of the midterm scores so that $X_4 = X_1 + X_2 + X_3$. Due to rounding, $\hat{\lambda}_4$ could be nonzero, but very close to zero.

6.2 Robust Principal Component Analysis

A robust "plug in" method uses an analysis based on the $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ computed from a robust dispersion estimator $\boldsymbol{C}$. The RPCA method performs the

classical principal component analysis on the RMVN subset, using either the sample covariance matrix $\boldsymbol{C}_U = \boldsymbol{S}_U$ or the sample correlation matrix $\boldsymbol{R}_U$. Under assumption (E1) from Chapter 4, $\boldsymbol{C}_U$ and $\boldsymbol{R}_U$ are $\sqrt{n}$ consistent highly outlier resistant estimators of $c\boldsymbol{\Sigma} = d\text{Cov}(\boldsymbol{x})$ and the population correlation matrix $\boldsymbol{D}\text{Cov}(\boldsymbol{x})\boldsymbol{D} = \boldsymbol{\rho}$, respectively, where $\boldsymbol{D} = \text{diag}(1/\sqrt{\sigma}_{11}, ..., 1/\sqrt{\sigma}_{\text{pp}})$ and the σ_{ii} are the diagonal entries of $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = c_X\boldsymbol{\Sigma}$. Let $\lambda_i(\boldsymbol{A})$ be the eigenvalues of $\boldsymbol{A}$ where $\lambda_1(\boldsymbol{A}) \geq \lambda_2(\boldsymbol{A}) \geq \cdots \geq \lambda_p(\boldsymbol{A})$. Let $\hat{\lambda}_i(\hat{\boldsymbol{A}})$ be the eigenvalues of $\hat{\boldsymbol{A}}$ where $\hat{\lambda}_1(\hat{\boldsymbol{A}}) \geq \hat{\lambda}_2(\hat{\boldsymbol{A}}) \geq \cdots \geq \hat{\lambda}(\hat{\boldsymbol{A}})$.

Theorem 6.3. Under (E1), the correlation of the eigenvalues computed from the classical PCA and RPCA converges to 1 in probability.

Proof: The eigenvalues are continuous functions of the dispersion estimator, hence consistent estimators of dispersion give consistent estimators of the population eigenvalues. See Eaton and Tyler (1991) and Bhatia, Elsner and Krause (1990). Let $\lambda_i(\boldsymbol{\Sigma}) = \lambda_i$ be the eigenvalues of $\boldsymbol{\Sigma}$ so $c_X\lambda_i$ are the eigenvalues of $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$. Under (E1), $\lambda_i(\boldsymbol{S}) \xrightarrow{P} c_X\lambda_i$ and $\lambda_i(\boldsymbol{C}_U) \xrightarrow{P} c\lambda_i = \frac{c}{c_X}c_X\lambda_i = d\ c_X\ \lambda_i$. Hence the population eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $d\ \boldsymbol{\Sigma}_{\boldsymbol{x}}$ differ by the positive multiple d, and the population correlation of the two sets of eigenvalues is equal to one.

Now let $\lambda_i(\boldsymbol{\rho}) = \lambda_i$. Under (E1), both $\boldsymbol{R}$ and $\boldsymbol{R}_U$ converge to $\boldsymbol{\rho}$ in probability, so $\hat{\lambda}_i(\boldsymbol{R}) \xrightarrow{P} \lambda_i$ and $\hat{\lambda}_i(\boldsymbol{R}_U) \xrightarrow{P} \lambda_i$ for $i = 1, ..., p$. Hence the two population sets of eigenvalues are the same and thus have population correlation equal to one. □

Note that if $\boldsymbol{\Sigma}_{\boldsymbol{x}}\ \boldsymbol{e} = \lambda\boldsymbol{e}$, then

$$d\ \boldsymbol{\Sigma}_{\boldsymbol{x}}\ \boldsymbol{e} = d\lambda\boldsymbol{e}.$$

Thus $\hat{\lambda}_i(\boldsymbol{S}) \xrightarrow{P} \lambda_i(\boldsymbol{\Sigma}_{\boldsymbol{x}})$ and $\hat{\lambda}_i(\boldsymbol{C}_U) \xrightarrow{P} d\lambda_i(\boldsymbol{\Sigma}_{\boldsymbol{x}})$ for $i = 1, ..., p$. Since plotting software fills space, two scree plots of two sets of eigenvalues that differ by a constant positive multiple will look nearly the same, except for the labels of the vertical axis, and the "trace explained" by the largest k eigenvalues will be the same for the two sets of eigenvalues. Theorem 6.2 implies that for a large class of elliptically contoured distributions and for large n, the classical and robust scree plots should be similar visually, and the "trace explained" by the classical PCA and the robust PCA should also be similar.

The eigenvectors are not continuous functions of the dispersion estimator, and the sample size may need to be massive before the robust and classical

eigenvectors or principal components have high absolute correlation. In the software, sign changes in the eigenvectors are common, since $\boldsymbol{\Sigma}_{\boldsymbol{x}}\ \boldsymbol{e} = \lambda \boldsymbol{e}$ implies that $\boldsymbol{\Sigma}_{\boldsymbol{x}}\ (-\boldsymbol{e}) = \lambda(-\boldsymbol{e})$.

Table 6.1: Estimation of $\boldsymbol{\Sigma}$ with $\gamma = 0.4$, $n = 35p$

p	type	n	pm	Q
5	1	135	16	0.153
5	2	135	6	0.213
10	1	350	21	0.326
10	2	350	6	0.326
15	1	525	26	0.856
15	2	525	7	0.675
20	1	700	33	0.798
20	2	700	8	0.792
25	1	875	39	1.014
25	2	875	10	1.867

A simulation was done to check that RMVN estimates $\boldsymbol{\Sigma}$ if the clean data is MVN and γ is the percentage of outliers. The clean cases were MVN: $\boldsymbol{x} \sim N_p(\boldsymbol{0}, diag(1, 2, ..., p))$. Outlier types were $\boldsymbol{x} \sim N_p((0, ..., 0, pm)^T, 0.0001\boldsymbol{I}_p)$, a near point mass at the major axis, and the mean shift $\boldsymbol{x} \sim N_p(pm\boldsymbol{1}, diag$ $(1, 2, ..., p))$ where $\boldsymbol{1} = (1, ..., 1)^T$. On clean MVN data, $n \geq 20p$ gave good results for $2 \leq p \leq 100$. For the contaminated MVN data, the first $n\gamma$ cases were outliers, and the classical estimator $\boldsymbol{S}_c$ was computed on the clean cases. The diagonal elements of $\boldsymbol{S}_c$ and $\hat{\boldsymbol{\Sigma}}_{RMVN}$ should both be estimating $(1, 2, ..., p)^T$. The average diagonal elements of both matrices were computed for 20 runs, and the criterion Q was the sum of the absolute differences of the p diagonal elements from the two averaged matrices. Since $\gamma = 0.4$ and the initial subsets for the RMVN estimator are half sets, the simulations used $n = 35p$. The values of Q shown in Table 6.1 correspond to good estimation of the diagonal elements. Values of pm slightly smaller than the tabled values led to poor estimation of the diagonal elements.

Example 6.1. Buxton (1920) gives various measurements on 87 men including *height*, *head length*, *nasal height*, *bigonal breadth* and *cephalic index*. Five *heights* were recorded to be about 19mm with the true heights

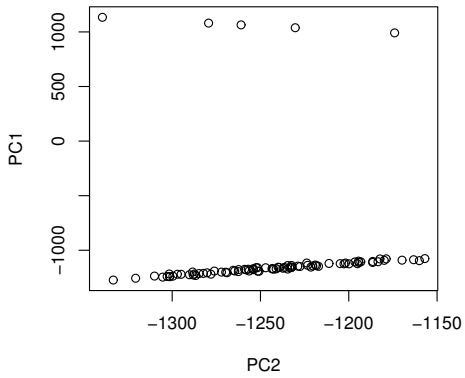

Figure 6.1: First Two Principal Components for Buxton data

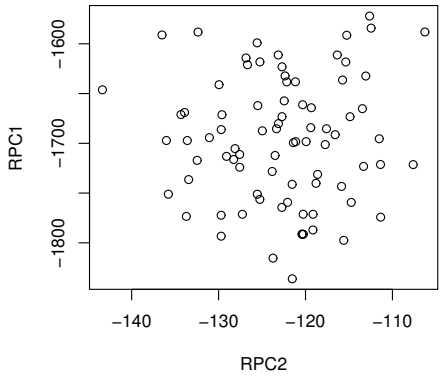

Figure 6.2: First Two Robust Principal Components with Outliers Omitted

recorded under head length. Performing a classical principal components analysis on these five variables using the covariance matrix resulted in a first principal component corresponding to a major axis that passed through the outliers. See Figure 6.1 where the second principal component is plotted versus the first. The robust PCA, or the classical PCA performed after the outliers are removed, resulted in a first principal component that was approximately $-$ *height* with $\hat{\boldsymbol{e}}_1 \approx (-1.000, 0.002, -0.023, -0.002, -0.009)^T$ while the second robust principal component was based on the eigenvector $\hat{\boldsymbol{e}}_2 \approx (-0.005, 0.848, -0.054, -0.048, 0.525)^T$. The plot of the first two robust principal components, with the outliers deleted, is shown in Figure 6.2. These two components explain about 86% of the variance.

The *R* function `prcomp` can be used to compute output. Suppose the data matrix is z. The commands

```
zz <- prcomp(z)
zz
```

will create and display output. The term *zz$sd* gives the square roots of the eigenvalues while the term *zz$rot* displays the eigenvectors using the covariance matrix. Hence Figure 6.1 can be made with the following commands.

```
z <- cbind(buxy,buxx)
zz <- prcomp(z)
PC1 <- z%*%zz$rot[,1]
PC2 <- z%*%zz$rot[,2]
plot(PC2,PC1)
```

It usually makes more sense to use the correlation matrix. the *mpack* function `rprcomp` does robust principal components. The two functions use "scale=T" or "cor=T" to use a correlation matrix.

```
zzcor <- prcomp(z,scale=T)
zrcor <- rprcomp(z,cor=T)
```

Then

```
zrcor$out$sd^2
```

gives the eigenvalues and *zrcoroutrot* gives the eigenvectors. Scree plots can be made with the following commands, and Figure 6.3 shows the robust scree plot which suggests that the last principal component can be deleted.

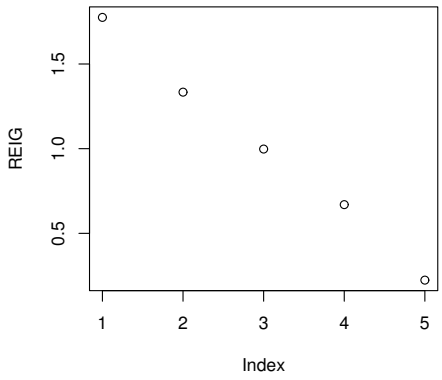

Figure 6.3: Robust Scree Plot

```
EIG <- zzcor$sd^2
plot(EIG)
#robust scree plot
REIG <- zrcor$out$sd^2
plot(REIG)
```

The outliers are known from the DD plot so the robust principal component analysis can be done with and without the outliers. The data matrix *zw* is the clean data without the outliers.

```
zw <-z[-c(61,62,63,64,65),]
zzcorc <- prcomp(zw,scale=T)
# clean data with corr matrix
> zzcorc
Standard deviations:
[1] 1.3184358 1.1723991 1.0155266 0.7867349 0.4867867
Rotation:
          PC1      PC2     PC3      PC4      PC5
buxy      0.01551  0.71466 0.02247 -0.68890 -0.11806
len       0.70308 -0.06778 0.07744 -0.16901  0.68302
nasal     0.15038  0.68868 0.02042  0.70385  0.08539
bigonal   0.11646 -0.04882 0.96504  0.02261 -0.22855
cephalic -0.68502  0.08950 0.24854 -0.03071  0.67825
```

```
zrcor <- rprcomp(z,cor=T)
> zrcor
$out
Standard deviations:
[1] 1.3323400 1.1548879 0.9988643 0.8182741 0.4730769
Rotation:
          PC1      PC2      PC3      PC4      PC5
buxy     -0.10724 -0.69431 -0.11325  0.69184 -0.12238
len       0.69909 -0.06324  0.02560  0.17129  0.69085
nasal     0.04094 -0.70310 -0.08718 -0.70093  0.07123
bigonal   0.02638 -0.13994  0.98660  0.01120 -0.07884
cephalic -0.70527 -0.00317  0.07443  0.02432  0.70460

> zrcorc <- rprcomp(zw,cor=T)
> zrcorc
$out
Standard deviations:
[1] 1.3369152 1.1466891 1.0016463 0.8123854 0.4842482
Rotation:
          PC1      PC2      PC3      PC4      PC5
buxy     -0.21306  0.67557 -0.01727 -0.68852 -0.15446
len       0.67272  0.21639  0.05560 -0.15178  0.68884
nasal    -0.22213  0.66958  0.05174  0.68978  0.15441
bigonal  -0.01374 -0.02995  0.99668 -0.03546 -0.06543
cephalic -0.67270 -0.21807  0.02363 -0.16076  0.68813
```

Note that the square roots of the eigenvalues, given by "Standard deviations," do not change much for the following three estimators: the classical estimator applied to the clean data, and the robust estimator applied to the full data or the clean data. The first eigenvector is roughly proportional to *length* − *cephalic* while the second eigenvector is roughly proportional to *buxy* + *nasal*. The third principal component is highly correlated with *bigonal*, the fourth principal component is proportional to *buxy* − *nasal*, and the fifth principal component to *length* + *cephalic*.

In simulations for principal component analysis, FCH, RMVN, OGK and Fake-MCD seem to estimate $c\boldsymbol{\Sigma}_{\boldsymbol{x}}$ if $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu}$ where $\boldsymbol{z} = (z_1, ..., z_p)^T$ and the z_i are iid from a continuous distribution with variance σ^2. Here

$\boldsymbol{\Sigma}_{\boldsymbol{x}} = \mathrm{Cov}(\boldsymbol{x}) = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T$. The bias for the MB estimator seemed to be small. It is known that affine equivariant estimators give unbiased estimators of $c\boldsymbol{\Sigma}_{\boldsymbol{x}}$ if the distribution of z_i is also symmetric. DGK and Fake-MCD (with fixed random number seed) are affine equivariant. FCH and RMVN are asymptotically equivalent to a scaled DGK estimator. But in the simulations the results also held for skewed distributions.

The simulations used 1000 runs where $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z}$ and $\boldsymbol{z} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$, $\boldsymbol{z} \sim LN(\boldsymbol{0}, \boldsymbol{I}_p)$ where the marginals are iid lognormal(0,1), or $\boldsymbol{z} \sim MVT_p(1)$, a multivariate t distribution with 1 degree of freedom so the marginals are iid Cauchy(0,1). The choice $\boldsymbol{A} = diag(\sqrt{1}, ..., \sqrt{p})$ results in $\boldsymbol{\Sigma} = diag(1, ..., p)$. Note that the population eigenvalues will be proportional to $(p, p-1, ..., 1)^T$ and the population "variance explained" by the ith principal component is $\lambda_i / \sum_{j=1}^{p} \lambda_j = 2(p+1-i)/[p(p+1)]$. For $p = 4$, these numbers are 0.4, 0.3 and 0.2 for the first three principal components. If the "correlation" option is used, then the population "correlation matrix" is the identity matrix $\boldsymbol{I}_p$, the ith population eigenvalue is proportional to $1/p$ and the population "variance explained" by the ith principal component is $1/p$.

Table 6.2 shows the mean "variance explained" along with the standard deviations for the first three principal components. Also a_i and p_i are the average absolute value of the correlation between the ith eigenvectors or the ith principal components of the classical and robust methods. Two rows were used for each "n–data type" combination. The a_i are shown in the top row while the p_i are in the lower row. The values of a_i and p_i were similar. The standard deviations were slightly smaller for the classical PCA for normal data. The classical method failed to estimate (0.4,0.3,0.2) for the Cauchy data. For the lognormal data, RPCA gave better estimates, and the p_i were not high except for $n = 10000$.

To compare affine equivariant and non-equivariant estimators, Maronna and Zamar (2002) suggest using $\boldsymbol{A}_{i,i} = 1$ and $\boldsymbol{A}_{i,j} = \rho$ for $i \neq j$ and $\rho = 0, 0.5, 0.7, 0.9$, and 0.99. Then $\boldsymbol{\Sigma} = \boldsymbol{A}^2$. If ρ is high, or if p is high and $\rho \geq 0.5$, then the data are concentrated about the line with direction $\boldsymbol{1} = (1, ..., 1)^T$. For $p = 50$ and $\rho = 0.99$, the population variance explained by the first principal component is 0.999998. If the "correlation" option is used, then there is still one extremely dominant principal component unless both p and ρ are small.

Table 6.3 shows the mean "variance explained" along with the standard deviations multiplied by 10^7 for the first principal component. The a_1 value is given but p_1 was always 1.0 to many decimal places even with Cauchy data.

Table 6.2: Variance Explained by PCA and RPCA, $p = 4$

n	type	M/S	vexpl	rvexpl	a_1/p_1	a_2/p_2	a_3/p_3
40	N	M	0.445,0.289,0.178	0.472,0.286,0.166	0.895	0.821	0.825
		S	0.050,0.037,0.032	0.062,0.043,0.037	0.912	0.813	0.804
100	N	M	0.419,0.295,0.191	0.425,0.293,0.189	0.952	0.926	0.963
		S	0.033,0.030,0.024	0.040,0.032,0.027	0.956	0.923	0.953
400	N	M	0.404,0.298,0.198	0.406,0.298,0.198	0.994	0.991	0.996
		S	0.019,0.017,0.014	0.021,0.019,0.015	0.995	0.990	0.994
40	C	M	0.765,0.159,0.056	0.514,0.275,0.147	0.563	0.519	0.511
		S	0.165,0.112,0.051	0.078,0.055,0.040	0.776	0.383	0.239
100	C	M	0.762,0.156,0.060	0.455,0.286,0.173	0.585	0.527	0.528
		S	0.173,0.112,0.055	0.054,0.041,0.034	0.797	0.377	0.269
400	C	M	0.756,0.162,0.060	0.413,0.296,0.194	0.608	0.562	0.575
		S	0.172,0.113,0.054	0.030,0.025,0.022	0.796	0.397	0.308
40	L	M	0.539,0.256,0.139	0.521,0.268,0.146	0.610	0.509	0.530
		S	0.127,0.075,0.054	0.099,0.061,0.047	0.643	0.439	0.398
100	L	M	0.482,0.270,0.165	0.459,0.279,0.172	0.647	0.555	0.566
		S	0.180,0.063,0.052	0.077,0.047,0.041	0.654	0.492	0.474
400	L	M	0.437,0.282,0.185	0.416,0.290,0.194	0.748	0.639	0.739
		S	0.080,0.048,0.044	0.049,0.035,0.033	0.727	0.594	0.690
10000	L	M	0.400,0.301,0.200	0.402,0.300,0.199	0.982	0.967	0.991
		S	0.027,0.023,0.018	0.013,0.011,0.009	0.976	0.967	0.989

Table 6.3: Variance Explained by PCA and RPCA, SSD = 10^7 SD, $p = 50$

n	type	vexpl	SSD	rvexpl	SSD	a_1
200	N	0.999998	1.958	0.999998	2.867	0.687
1000	N	0.999998	0.917	0.999998	0.971	0.944
1000	C	0.999996	161.3	0.999998	1.482	0.112
1000	L	0.999998	0.919	0.999998	1.508	0.175

Hence the eigenvectors from the robust and classical methods could have low absolute correlation, but the data was so tightly clustered that the first principal components from the robust and classical methods had absolute correlation near 1.

6.3 Summary

1) Let $\boldsymbol{\Sigma} = ((\sigma_{ij}))$ be a positive definite symmetric $p \times p$ dispersion matrix. A *generalized correlation matrix* $\boldsymbol{\rho} = \big(\big(\rho_{ij}\big)\big)$ where

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

The generalized correlation matrix is the correlation matrix when second moments exist if $\boldsymbol{\Sigma} = c\ \mathrm{Cov}(\boldsymbol{x})$ for some constant $c > 0$.

2) Classical principal component analysis (PCA) gets the eigenvalues and eigenvectors $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ of the sample covariance matrix $\boldsymbol{S}$ or of the sample correlation matrix $\boldsymbol{R}$.

3) Let U be the subset of at least half of the cases from which the robust estimator is computed. Let $\boldsymbol{S}_U$ and $\boldsymbol{R}_U$ denote the sample covariance matrix and sample correlation matrix computed from the cases in $\boldsymbol{U}$. Then the robust estimator $\boldsymbol{C} = d\boldsymbol{S}_U$ for some constant $d > 0$ and $\boldsymbol{R}_U$ is the generalized correlation matrix corresponding to $\boldsymbol{C}$. The robust PCA uses U corresponding to the RMVN estimator.

4) Want $n > 10p$ for the classical PCA and $n > 20p$ for the robust PCA.

5) Both R and SAS output give the eigenvectors as shown in symbols for the following table.

PC1	PC2	$\cdots$	PCp
$\hat{\boldsymbol{e}}_1$	$\hat{\boldsymbol{e}}_2$	$\cdots$	$\hat{\boldsymbol{e}}_p$

R output shows the square roots of the eigenvalues

$$\sqrt{\hat{\lambda}_1}, \sqrt{\hat{\lambda}_2}, ..., \sqrt{\hat{\lambda}_p}$$

while SAS output gives the eigenvalues $\hat{\lambda}_i$.

6) Given the eigenvalues or square roots of the eigenvalues, be able to sketch a
scree plot of i versus $\hat{\lambda}_i$.

7) The *trace explained* or *variance explained* by the first k principal components is $\frac{\sum_{i=1}^{k} \hat{\lambda}_i}{\sum_{i=1}^{p} \hat{\lambda}_i}$ where the denominator is equal to p if the correlation option $\boldsymbol{R}$ or $\boldsymbol{R}_U$ is used, as recommended in point 10).

8) Use k principal components if the trace explained is bigger than some percentage like 90%, 80% or 70%. There is often a sharp bend in the scree plot when the components are no longer useful.

9) When $\boldsymbol{R}$ or $\boldsymbol{R}_U$ is used, the correlation of the ith variable with the jth principal component is proportional to the ith entry of the jth eigenvector $\hat{\boldsymbol{e}}_j$. To try to explain the jth principal component, look at entries in $\hat{\boldsymbol{e}}_j$ that are large in magnitude and ignore entries close to zero. Sometimes only one entry is large. Sometimes all of the large entries have approximately the same size and sign, then the principal component is interpreted as an average of these entrees. If exactly two entries are of similar large magnitude but of different sign, the principal component is interpreted as a difference of the two entrees. If there are $j \geq 2$ large entrees that differ in magnitude, then the principal component is interpreted as a linear combination of the corresponding variables.

10) PCA based on $\boldsymbol{R}$ or $\boldsymbol{R}_U$ is easier to interpret than PCA based on $\boldsymbol{S}$ or $\boldsymbol{S}_U$.

i) If $\boldsymbol{S}$ is used, the variance explained by the first principal component could be large because one variable has much larger variance than the other variables.

ii) If $\boldsymbol{S}$ is used, the correlation of the ith variable with the jth principal component is proportional to the ith entry of the jth eigenvector $\hat{\boldsymbol{e}}_j$ divided by the standard deviation of ith variable: $e_{ij}/\sqrt{S_{ii}}$.

Hence PCA based on $\boldsymbol{S}$ is harder to interpret if p random variables do not have similar sample variances. The variances could differ if different units are used or if some variables are transformed while others are not. Hence PCA based on $\boldsymbol{R}$ or $\boldsymbol{R}_U$ is recommended.

```
11) Typical Routput is shown. Standard deviations:
[1] 1.3369152 1.1466891 1.0016463 0.8123854 0.4842482
Rotation: PC1         PC2         PC3         PC4         PC5
len       0.67271620 -0.21639022  0.05559575  0.15178244 -0.68883916
nasal    -0.22213361 -0.66957907  0.05173705 -0.68978370 -0.15440936
bigonal  -0.01373814  0.02995162  0.99668240  0.03545927  0.06542933
```

```
cephalic -0.67269993  0.21806615  0.02362841  0.16076405 -0.68812686
buxy     -0.21306252 -0.67556583 -0.01727087  0.68851877  0.15446292
```

12) Let $\hat{\boldsymbol{\Sigma}}$ be a consistent estimator of $\boldsymbol{\Sigma}$. The following theorems show that asymptotically, the eigenvalues and eigenvectors of $\hat{\boldsymbol{\Sigma}}$ act as those of $\boldsymbol{\Sigma}$ and vice verca. This result is useful since eigenvectors are not continuous functions of the dispersion matrix. The following theorem holds because eigenvalues and the generalized correlation matrix are continuous functions of the dispersion matrix.

i) **Theorem 6.1.** Suppose the dispersion matrix $\boldsymbol{\Sigma}$ has eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Suppose $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ for some constant $c > 0$. Let the eigenvalue eigenvector pairs of $\hat{\boldsymbol{\Sigma}}$ be $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), ..., (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. Then $\hat{\lambda}_j(\hat{\boldsymbol{\Sigma}}) \xrightarrow{P} c\lambda_j(\boldsymbol{\Sigma}) = c\lambda_j$, $\hat{\boldsymbol{\rho}} \xrightarrow{P} \boldsymbol{\rho}$ and $\hat{\lambda}_j\left(\hat{\boldsymbol{\rho}}\right) \xrightarrow{P} \lambda_j\left(\boldsymbol{\rho}\right)$ where $\lambda_j(\boldsymbol{A})$ is the jth eigenvalue of $\boldsymbol{A}$ for $j = 1, ..., p$.

ii) **Theorem 6.2.** Assume the $p \times p$ symmetric dispersion matrix $\boldsymbol{\Sigma}$ is positive definite. a) If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\Sigma}$, then $\hat{\boldsymbol{\Sigma}}\boldsymbol{e}_i - \hat{\lambda}_i\boldsymbol{e}_i \xrightarrow{P} \boldsymbol{0}$.

b) If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\Sigma}$, then $\boldsymbol{\Sigma}\hat{\boldsymbol{e}}_i - \lambda_i\hat{\boldsymbol{e}}_i \xrightarrow{P} \boldsymbol{0}$.

If $\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} = O_P(n^{-\delta})$ where $0 < \delta \leq 0.5$, then

c) $\lambda_i\boldsymbol{e}_i - \hat{\boldsymbol{\Sigma}}\boldsymbol{e}_i = O_P(n^{-\delta})$, and

d) $\hat{\lambda}_i\hat{\boldsymbol{e}}_i - \boldsymbol{\Sigma}\hat{\boldsymbol{e}}_i = O_P(n^{-\delta})$.

e) If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ for some constant $c > 0$, and if the eigenvalues $\lambda_1 > \cdots > \lambda_p > 0$ of $\boldsymbol{\Sigma}$ are unique, then the absolute value of the correlation of $\hat{\boldsymbol{e}}_j$ with $\boldsymbol{e}_j$ converges to 1 in probability: $|\text{corr}(\hat{\boldsymbol{e}}_j, \boldsymbol{e}_j)| \xrightarrow{P} 1$.

iii) **Theorem 6.3.** Under (E1), the correlation of the eigenvalues computed from the classical PCA and robust PCA converges to 1 in probability.

13) Centering uses $\boldsymbol{w}_i = \boldsymbol{x}_i - T$ where T is the sample mean or the sample mean of the standardized data for the full data set or for the set U used to compute the robust estimator. Centering does not change $\boldsymbol{S}, \boldsymbol{S}_U, \boldsymbol{R}$ or $\boldsymbol{R}_U$, but the jth principal component is $\hat{\boldsymbol{e}}_j^T\boldsymbol{w}_i = \hat{\boldsymbol{e}}_j^T(\boldsymbol{x}_i - T)$.

14) For PCA, the `summary(out)` statement shows

Importance of components:	PC1	PC2	$\cdots$	PCk	$\cdots$	PCp
Standard deviation	$\sqrt{\hat{\lambda}_1}$	$\sqrt{\hat{\lambda}_2}$	$\cdots$	$\sqrt{\hat{\lambda}_k}$	$\cdots$	$\sqrt{\hat{\lambda}_p}$
Proportion of variance	$\frac{\hat{\lambda}_1}{\sum_{i=1}^p \hat{\lambda}_i}$	$\frac{\hat{\lambda}_2}{\sum_{i=1}^p \hat{\lambda}_i}$	$\cdots$	$\frac{\hat{\lambda}_k}{\sum_{i=1}^p \hat{\lambda}_i}$	$\cdots$	$\frac{\hat{\lambda}_p}{\sum_{i=1}^p \hat{\lambda}_i}$
Cumulative Proportion	$\frac{\hat{\lambda}_1}{\sum_{i=1}^p \hat{\lambda}_i}$	$\frac{\sum_{j=1}^2 \hat{\lambda}_j}{\sum_{i=1}^p \hat{\lambda}_i}$	$\cdots$	$\frac{\sum_{j=1}^k \hat{\lambda}_j}{\sum_{i=1}^p \hat{\lambda}_i}$	$\cdots$	1

Recall that if $\boldsymbol{R}$ or $\boldsymbol{R}_U$ is used, then $\sum_{i=1}^p \hat{\lambda}_i = p$. Typically want to keep the first m principal components where $\frac{\sum_{j=1}^m \hat{\lambda}_j}{\sum_{i=1}^p \hat{\lambda}_i} > a$ where the threshold a is a number like 0.9, 0.8 or 0.7.

15) For PCA, a *biplot* is a plot of the first principal component versus the second principal component. The plotted points are $\hat{\boldsymbol{e}}_j^T \boldsymbol{x}_i$ for $j = 1, 2$ where the classical biplot uses $i = 1, ..., n$ and the robust plot uses cases in the RMVN set U. Let $\hat{\boldsymbol{e}}_j = (\hat{e}_{1j}, \hat{e}_{2j}, ..., \hat{e}_{pj})^T$. Then $\hat{e}_{kj}$ is called the *loading* of the kth variable on the jth principal component. An arrow with the kth variable name is the vector from the origin $(0, 0)^T$ to the loadings $(\hat{e}_{k1}, \hat{e}_{k2})^T$. So if the arrow is in the first quadrant, both loadings are positive, etc. If the arrow is long to the right but short down, then the loading with the first principal component is large and positive while the loading with the second principal component is small and negative. Be able to interpret the classical and robust biplots.

6.4 Complements

Suppose $\boldsymbol{Z}$ is the standardized $n \times p$ data matrix and $\boldsymbol{Y} = \boldsymbol{Z}/\sqrt{n-1}$. If $n < p$, then the correlation matrix $\boldsymbol{R} = \boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{Z}^T\boldsymbol{Z}/(n-1)$ does not have full rank. By singular value decomposition (SVD) theory, the SVD of $\boldsymbol{Y}$ is $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$ where the positive singular values are square roots of the positive eigenvalues of both $\boldsymbol{Y}^T\boldsymbol{Y}$ and of $\boldsymbol{Y}\boldsymbol{Y}^T$. Also $\boldsymbol{V} = (\hat{\boldsymbol{e}}_1\ \hat{\boldsymbol{e}}_2\ \cdots\ \hat{\boldsymbol{e}}_p)$, and $\boldsymbol{Y}^T\boldsymbol{Y}\hat{\boldsymbol{e}}_i = \sigma_i^2\hat{\boldsymbol{e}}_i$. Hence classical principal component analysis on the standardized data can be done using $\hat{\boldsymbol{e}}_i$ and $\hat{\lambda}_i = \sigma_i^2$. The SVD of $\boldsymbol{Y}^T$ is

$\boldsymbol{V}\boldsymbol{\Lambda}^T\boldsymbol{U}^T$, and

$$\boldsymbol{Y}\boldsymbol{Y}^T = \frac{1}{n-1}\begin{bmatrix} \boldsymbol{z}_1^T\boldsymbol{z}_1 & \boldsymbol{z}_1^T\boldsymbol{z}_2 & \cdots & \boldsymbol{z}_1^T\boldsymbol{z}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{z}_n^T\boldsymbol{z}_1 & \boldsymbol{z}_n^T\boldsymbol{z}_2 & \cdots & \boldsymbol{z}_n^T\boldsymbol{z}_n \end{bmatrix}$$

which is the matrix of scalar products divided by $(n-1)$. For more information about the SVD, see Datta (1995, p. 552-556).

It may be possible to do robust PCA when $n < p$ by standardizing the data with the MED(X_i) and MAD(X_i). Then plot the Euclidean distaces of the standardized data from the coordinatewise median MED$(\boldsymbol{Z})$ and delete outliers, leaving m cases in an $m \times p$ matrix $\boldsymbol{Y}$. Then use the SVD of $\boldsymbol{Y}$ to perform a "robust" PCA.

Jolliffe (2010) is an authoritative text on PCA. Cattell (1966) and Bentler and Yuan (1998) are good references for scree plots. Mϕller, von Frese and Bro (2005) discuss PCA, principal component regression and drawbacks of M estimators. Waternaux (1976) and Tyler (1983) give some large sample theory for PCA. In particular, if the $\boldsymbol{x}_i$ are iid from a multivariate distribution with fourth moments and a covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ such that the eigenvalues are distinct and positive, then $\sqrt{n}(\hat{\lambda}_i - \lambda_i) \xrightarrow{D} N(0, \kappa_i + 2\lambda_i^2)$ where κ_i is the kurtosis of the marginal distribution of x_i, for $i = 1, ..., p$.

The literature for robust PCA is large, but the "high breakdown" methods are impractical or not backed by theory. Some of these methods may be useful as outlier diagnostics. The theory of Boente (1987) for mildly outlier resistant principal components is not based on DGK estimators since the weighting function on the D_i is continuous. Spherical principal components is a mildly outlier resistant bounded influence approach suggested by Locantore, Marron, Simpson, Tripoli, Zhang and Cohen (1999). Boente and Fraiman (1999) claim that basis of the eigenvectors is consistently estimated by spherical principal components for elliptically contoured distributions. Also see Maronna, Martin and Yohai (2006, p. 212-213) and Taskinen, Koch and Oja (2012).

Bali, Boente, Tyler and Wang (2011) gave possibly impressive theory for infinite complexity impractical robust projection estimators, but should have given theory for the practical Fake-projection estimator actually used. This "bait and switch hoax" occurs far too often in multivariate "robust statistics" papers.

To estimate the first principal direction for principal component analysis, the Fake-projection (CR) estimator uses n projections $\boldsymbol{z}_i = \boldsymbol{w}_i/\|\boldsymbol{w}_i\|$ where $\boldsymbol{w}_i = \boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_n$. Note that for $p = 2$ one can select 360 projections through the origin and a point on the unit circle that are one degree apart. Then there is a projection that is highly correlated with any projection on the unit circle. If $p = 3$, then 360 projections are not nearly enough to adequately approximate all projections through the unit sphere. Since the surface area of a unit hypersphere is proportional to n^{p-1}, approximations rapidly get worse as p increases.

Theory for the Fake-projection (CR) estimator may be simple. Suppose the data is multivariate normal $N_p(\boldsymbol{0}, diag(p, 1, ..., 1))$. Then $\boldsymbol{\beta} = (1, 0, ..., 0)^T$ (or $-\boldsymbol{\beta}$) is the population first direction. Heuristically, assume $\hat{\boldsymbol{\mu}}_n = \boldsymbol{0}$, although in general $\hat{\boldsymbol{\mu}}_n$ should be a good $\sqrt{n}$ consistent estimator of $\boldsymbol{\mu}$ such as the coordinatewise median. Let $\boldsymbol{b}_o$ be the "best" estimated projection $\boldsymbol{z}_j$ that minimizes $\|\boldsymbol{z}_i - \boldsymbol{\beta}\|$ for $i = 1, ..., n$. "Good" projections will have a $\boldsymbol{y}_i$ that lies in one of two "hypercones" with a vertex at the origin and centered about a line through the origin and $\pm\boldsymbol{\beta}$ with radius r at $\pm\boldsymbol{\beta}$. So for $p = 2$ the two "cones" are determined by the two lines through the origin with slopes $\pm\ r$. The probability that a randomly selected $\boldsymbol{y}_i$ falls in one of the two "hypercones" is proportional to r^{p-1}, and for $\boldsymbol{b}_o$ to be consistent for $\boldsymbol{\beta}$ need $r \to 0$, P(at least one $\boldsymbol{y}_i$ falls in "hypercone") $\to 1$ and $n \to \infty$. If these heuristics are correct, need $r \propto n^{\frac{-1}{p-1}}$ for $\|\boldsymbol{b}_o - \boldsymbol{\beta}\| = O_P(n^{\frac{1}{p-1}})$. Note that $\boldsymbol{b}_o$ is not an estimator since $\boldsymbol{\beta}$ is not known, but the rate of the "best" projection $\boldsymbol{b}_o$ gives an upper bound on the rate of the Fake-projection estimator $\boldsymbol{v}_1$ since $\|\boldsymbol{v}_1 - \boldsymbol{\beta}\| \geq \|\boldsymbol{b}_o - \boldsymbol{\beta}\|$. If the scale estimator is $\sqrt{n}$ consistent, then for a large class of elliptically contoured distributions, a conjecture is that $\|\boldsymbol{v}_1 - \boldsymbol{\beta}\| = O_P(n^{\frac{1}{2(p-1)}})$ for $p > 1$.

Simulations were done in *R*. The `MASS` library was used to compute FMCD and the `robustbase` library was used to compute OGK. The *mpack* function `covrmvn` computes the FCH, RMVN and MB estimators while `covfch` computes the FCH, RFCH and MB estimators. The following functions were used in the three simulations and have more outlier configurations than the two described in the text. Function `covesim` was used to produce Table 6.1 and `pcasim` for Tables 6.2 and 6.3. See Zhang (2011) for more extensive simulations.

For a nonsingular matrix, the inverse of the matrix, the determinant of the matrix and the eigenvalues of the matrix are continuous functions of

the matrix. Hence if $\hat{\mathbf{\Sigma}}$ is a consistent estimator of $\mathbf{\Sigma}$, then the inverse, determinant and eigenvalues of $\hat{\mathbf{\Sigma}}$ are consistent estimators of the inverse, determinant and eigenvalues of $\mathbf{\Sigma}$. See, for example, Bhatia, Elsner and Krause (1990), Stewart (1969) and Severini (2005, p. 348-349).

6.5 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

6.1*. Assume the $p \times p$ dispersion matrix $\mathbf{\Sigma}$ is positive definite. If $\hat{\mathbf{\Sigma}} \xrightarrow{P} c\mathbf{\Sigma}$ for some constant $c > 0$, prove that $\mathbf{\Sigma}\hat{\boldsymbol{e}}_i - \lambda_i \hat{\boldsymbol{e}}_i \xrightarrow{P} \mathbf{0}$.

6.2. Shown below is PCA output using the correlation matrix for the Buxton data where 5 outliers were deleted. The variables were *length, nasal height, bigonal breadth, cephalic* and $buxy = height/20$. The "standard deviations" line corresponds to the square roots of the eigenvalues. The Rotation matrix gives the 5 principal components.

a) For the robust `rprcomp` output make a scree plot. What proportion of the trace is explained by the first 4 principal components?

b) Which principal component corresponds to i) bigonal, ii) nasal + buxy, iii) length + cephalic, iv) length − cephalic and v) nasal − buxy?

```
rprcomp(z)
$out
Standard deviations:
[1] 1.3369152 1.1466891 1.0016463 0.8123854 0.4842482

Rotation:
                PC1         PC2         PC3         PC4         PC5
len       0.67271620 -0.21639022  0.05559575  0.15178244 -0.68883916
nasal    -0.22213361 -0.66957907  0.05173705 -0.68978370 -0.15440936
bigonal  -0.01373814  0.02995162  0.99668240  0.03545927  0.06542933
cephalic -0.67269993  0.21806615  0.02362841  0.16076405 -0.68812686
buxy     -0.21306252 -0.67556583 -0.01727087  0.68851877  0.15446292

prcomp(z,scale=T)
Standard deviations:
```

```
[1] 1.3184358 1.1723991 1.0155266 0.7867349 0.4867867

Rotation:
                PC1         PC2        PC3         PC4         PC5
len      -0.70308364 -0.06777853 0.07743938  0.16900791  0.6830219
nasal    -0.15038248  0.68867720 0.02042098 -0.70384733  0.0853859
bigonal  -0.11646120 -0.04882199 0.96504341 -0.02261327 -0.2285455
cephalic  0.68502160  0.08950469 0.24854103  0.03070660  0.6782468
buxy     -0.01551443  0.71465734 0.02246533  0.68889840 -0.1180614
```

6.3. Let $Y_j = \boldsymbol{e}_j^T \boldsymbol{x}$ be the first population principal component where $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma_x}$.

a) Using $\text{Cov}(\boldsymbol{Ax}, \boldsymbol{Bx}) = \boldsymbol{A\Sigma_x B}^T$, show $\text{Cov}(\boldsymbol{x}, Y_j) = \boldsymbol{\Sigma_x e}_j = \lambda_j \boldsymbol{e}_j$.

b) Now $V(Y_j) = \text{Cov}(\boldsymbol{e}_j^T \boldsymbol{x}, \boldsymbol{e}_j^T \boldsymbol{x})$. Show that $V(Y_j) = \lambda_j$.

c) Let $\boldsymbol{x} = (X_1, ..., X_p)^T$ where X_i is the ith random variable with $V(X_i) = \sigma_{ii}$ and by a) $\text{Cov}(X_i, Y_j) = \lambda_j e_{ij}$ where $\boldsymbol{e}_j = (e_{1j}, ..., e_{ij}, ..., e_{pj})^T$. Find $\text{corr}(X_i, Y_j)$.

6.4. The classical PCA output below is for the Buxton data described in Problem 6.2 where 5 cases have massive outliers in the height and length variables. Interpret PC1 and PC2.

```
prcomp(z,scale=T)
[1] 1.431 1.074 0.964 0.926 0.106
       PC1    PC2    PC3    PC4    PC5
len   0.685  0.037  0.004 -0.189 -0.702
nas  -0.199  0.568  0.153 -0.783  0.047
big  -0.049 -0.569  0.783 -0.247 -0.007
ceph -0.100 -0.594 -0.603 -0.523  0.008
ht   -0.692 -0.000 -0.008  0.131 -0.710
```

6.5. SAS output for PCA using the correlation matrix is shown below. The Khattree and Naik (1999, p. 11) cork data gives the weights of cork borings in four directions for 28 trees in a block of plantations.

a) What is the variance explained by the first two principal components?

b) Interpret the first principal component.

```
                Eigenvalues of the Covariance Matrix
      Eigenvalue    Difference    Proportion   Cumulative
1       3.5967        3.3431        0.8992        0.8992
2       0.2536        0.1735        0.0634        0.9626
3       0.0801        0.0107        0.0200        0.9826
4       0.0694                      0.0174        1.0000
                        Eigenvectors
       Prin1      Prin2      Prin3        Prin4
north -0.5108992  0.1267234  0.803287920  0.2786606
east  -0.4829921  0.7604818 -0.328918253 -0.2831940
south -0.5082783 -0.3006659 -0.496526386  0.6361719
west  -0.4973468 -0.5614345  0.001687729 -0.6613884

Rotation:  PC1        PC2        PC3
length 0.5771831 -0.5884323 -0.5662218
width  0.5811769 -0.1910978  0.7910215
height 0.5736663  0.7856393 -0.2316848

> summary(out$out)
Importance of components:PC1     PC2     PC3
Standard deviation     1.7065 0.25601 0.14961
Proportion of Variance 0.9707 0.02185 0.00746
Cumulative Proportion  0.9707 0.99254 1.00000
```

6.6. The Johnson and Wichern (1988, p. 262) turtle data has $X_1 = length$, $X_2 = width$ and $X_3 = height$ for painted turtle shells with 48 cases. Principal component analysis output is shown above based on the (robust) correlation matrix.

a) How many principal components are needed?

b) Interpret the first principal component.

6.7. The output below describes lawyers' ratings of state judges in the US Superior Court with 43 observations on 12 numeric variables: CONT Number of contacts of lawyer with judge, INTG Judicial integrity, DMNR Demeanor, DILG Diligence, CFMG Case flow managing, DECI Prompt decisions, PREP Preparation for trial, FAMI Familiarity with law, ORAL Sound oral rulings, WRIT Sound written rulings, PHYS Physical ability, RTEN Worthy of retention.

```
> rprcomp(USJudgeRatings)
Standard deviations:
 [1] 3.22195231 1.03832823 0.51049711 0.41049221 0.22797980 0.16242562
 [7] 0.11155709 0.09407153 0.07441343 0.05595849 0.04492358 0.03805913

Rotation:
             PC1         PC2
CONT  0.09651014  0.90089601
INTG -0.29727192 -0.19029004
DMNR -0.28269055 -0.21697647
DILG -0.30634676  0.01963176
CFMG -0.29804314  0.19297945
DECI -0.30227359  0.18417871
PREP -0.30428044  0.10879296
FAMI -0.30144067  0.11286037
ORAL -0.30874784  0.05751148
WRIT -0.30769444  0.06085970
PHYS -0.28368257 -0.03718180
RTEN -0.30728474 -0.02411832
```

a) Interpret the first principal component.

b) Interpret the second principal component.

6.8. From the SAS output shown below, what is the variance explained by the second principal component?

```
          Eigenvalues of the Covariance Matrix
      Eigenvalue    Difference    Proportion    Cumulative
1    154.310607    145.147647        0.9439        0.9439
2      9.162960                      0.0561        1.0000
                    Eigenvectors
                         Prin1         Prin2
           July       0.343532      0.939141
           January    0.939141      -.343532
```

R/Splus Problems

Warning: Use the command *source("G:/mpack.txt")* **to download the programs. See Preface or Section 15.2.** Typing the name of the

`mpack` function, eg *ddplot*, will display the code for the function. Use the `args` command, eg *args(pcasim)*, to display the needed arguments for the function.

6.9. a) Type the *R* command `pcasim()` and paste the output into *Word*.

This command computes the first 3 eigenvalues and eigenvectors for the classical and robust PCA using the $\boldsymbol{R}$ and $\boldsymbol{R}_U$. The multivariate normal data is such that the cases cluster tightly about the eigenvector $c(1, 1, ..., 1)^T$ corresponding to the largest eigenvalue. The term mncor gives the mean correlation between the classical and robust eigenvalues while the terms vexpl and rvexpl give the average variance explained by the largest 3 eigenvalues. The terms abscoreigvi give the absolute correlation between the i classical and robust eigenvector for $i = 1, ..., 3$ while the term abscorpc gives the absolute correlations of the first 3 principal components.

b) Are the robust and classical eigenvalues highly correlated? Is the absolute correlation for first classical principal component and the robust principal component high?

6.10. The Venables and Ripley (2003) CPU data has variables syct = cycle time,
mmin = minimum main memory,
chmin = minimum number of channels,
chmax = maximum number of channels,
perf = published performance, and
estperf = estimated performance.

a) There are nonlinear relationships among the variables and 1 is added to each variable to make them positive. Read more about the data set and make a scatterplot matrix with the *R commands* for this part. You can make the help window small by clicking the box with the $-$ in the upper right corner. Include the scatterplot matrix in *Word*.

b) The log rule suggests using the log transformation on all of the variables. Make the log transformations, scatterplot matrix and DD plot with the *R commands* for this part. Right click "Stop" to go from the DD plot to the *R* prompt. Wait until part d) until you put plots in *Word*.

c) You might be able to get a better scatterplot matrix and DD plot by doing alternative transformations on the last two variables. The commands for this part give the log transformation for the first 4 variables and possible

transformations for the last variables. Clearly state which transformations you use for the 5th and 6th variable. For example if you decide logs are ok, write down the following transformations.

```
zz[,5] <- log(z[,5])
zz[,6] <- log(z[,6])
```

d) For your data set zz of transformed variables, make the scatterplot matrix and DD plot and put the two plots in *Word*.

e) Put the classical PCA output using the correlation matrix into *Word* with the command for this problem.

f) Put the robust PCA output using the correlation matrix into *Word* with the command for this problem.

g) Comment on the similarities or differences of the classical and robust PCA.

6.11. The *R* data set USArrests contains statistics, in arrests per 100,000 residents, for assault, murder, and rape in each of the 50 US states in 1973. The fourth variable, UrbanPop, is the percent urban population in each state. For PCA, the *R* `summary` command can be used to get proportion of variance explained and cumulative proportion of variance explained, similar to *SAS* output.

a) Use the *R commands* for this part to get the classical and robust PCA summaries where $\boldsymbol{S}$ or $\boldsymbol{S}_U$ is used. Paste the summaries into *Word*.

i) Are the summaries similar?

ii) Using the 0.9 threshold, how many principal components are needed?

a) Use the *R commands* for this part to get the classical and robust PCA summaries where $\boldsymbol{R}$ or $\boldsymbol{R}_U$ is used. Paste the summaries into *Word*.

i) Are the summaries similar?

ii) using the 0.9 threshold, how many principal components are needed?

6.12. For PCA, a *biplot* is a plot of the first principal component versus the second principal component. The plotted points are $\hat{\boldsymbol{e}}_j^T \boldsymbol{x}_i$ for $j = 1, 2$ where the classical biplot uses $i = 1, ..., n$ and the robust plot uses cases in the RMVN set U. Let $\hat{\boldsymbol{e}}_j = (\hat{e}_{1j}, \hat{e}_{2j}, ..., \hat{e}_{pj})^T$. Then $\hat{e}_{kj}$ is called the *loading* of the kth variable on the jth principal component. An arrow with the kth variable name is the vector from the origin $(0, 0)^T$ to the loadings $(\hat{e}_{k1}, \hat{e}_{k2})^T$. So if the arrow is in the first quadrant, both loadings are positive, etc. If the arrow is long to the right but short down, then the loading with the first

principal component is large and positive while the loading with the second principal component is small and negative.

The Buxton (1920) data has a cluster of 5 massive outliers. The first classical principal component tends to go right through a cluster of large outliers.

a) These *R* commands make the classical scree plot and biplot. Paste the plots into *Word*.

b) These *R* commands make the robust scree plot and biplot. Paste the plots into *Word*.

c) From the classical scree plot, how many principal components are needed? From the robust scree plot, how many principal components are needed?

d) The four variables used were *len, nasal, bigonal*, and *cephalic* . From the classical biplot, which variable had the 5 massive outliers.

e) From the robust biplot, which two variables loaded highest with the first principal component?

Chapter 7

Canonical Correlation Analysis

7.1 Introduction

Let $\boldsymbol{x}$ be the $p \times 1$ vector of predictors, and partition $\boldsymbol{x} = (\boldsymbol{w}^T, \boldsymbol{y}^T)^T$ where $\boldsymbol{w}$ is $m \times 1$ and $\boldsymbol{y}$ is $q \times 1$ where $m = p - q \leq q$ and $m, q \geq 2$. Canonical correlation analysis (CCA) seeks m pairs of linear combinations $(\boldsymbol{a}_1^T\boldsymbol{w}, \boldsymbol{b}_1^T\boldsymbol{y}), ..., (\boldsymbol{a}_m^T\boldsymbol{w}, \boldsymbol{b}_m^T\boldsymbol{y})$ such that $\text{corr}(\boldsymbol{a}_i^T\boldsymbol{w}, \boldsymbol{b}_i^T\boldsymbol{y})$ is large under some constraints on the $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ where $i = 1, ..., m$. The first pair $(\boldsymbol{a}_1^T\boldsymbol{w}, \boldsymbol{b}_1^T\boldsymbol{y})$ has the largest correlation. The next pair $(\boldsymbol{a}_2^T\boldsymbol{w}, \boldsymbol{b}_2^T\boldsymbol{y})$ has the largest correlation among all pairs uncorrelated with the first pair and the process continues so that $(\boldsymbol{a}_m^T\boldsymbol{w}, \boldsymbol{b}_m^T\boldsymbol{y})$ is the pair with the largest correlation that is uncorrelated with the first $m - 1$ pairs. The correlations are called *canonical correlations* while the pairs of linear combinations are called *canonical variables*.

Some notation is needed to explain CCA. Let the $p \times p$ positive definite symmetric dispersion matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Let $\boldsymbol{J} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$. Let $\boldsymbol{\Sigma}_a = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$, $\boldsymbol{\Sigma}_A = \boldsymbol{J}\boldsymbol{J}^T = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$, $\boldsymbol{\Sigma}_b = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}_B = \boldsymbol{J}^T\boldsymbol{J} = \boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$. Let $\boldsymbol{e}_i$ and $\boldsymbol{g}_i$ be sets of orthonormal eigenvectors, so $\boldsymbol{e}_i^T\boldsymbol{e}_i = 1$, $\boldsymbol{e}_i^T\boldsymbol{e}_j = 0$ for $i \neq j$, $\boldsymbol{g}_i^T\boldsymbol{g}_i = 1$ and $\boldsymbol{g}_i^T\boldsymbol{g}_j = 0$ for $i \neq j$. Let the $\boldsymbol{e}_i$ be $m \times 1$ while the $\boldsymbol{g}_i$ are $q \times 1$.

Let $\boldsymbol{\Sigma}_a$ have eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{a}_1), ..., (\lambda_m, \boldsymbol{a}_m)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. Let $\boldsymbol{\Sigma}_A$ have eigenvalue eigenvector pairs $(\lambda_i, \boldsymbol{e}_i)$ for $i =$

$1, ..., m$. Let $\boldsymbol{\Sigma}_b$ have eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{b}_1), ..., (\lambda_q, \boldsymbol{b}_q)$. Let $\boldsymbol{\Sigma}_B$ have eigenvalue eigenvector pairs $(\lambda_i, \boldsymbol{g}_i)$ for $i = 1, ..., q$. It can be shown that the m largest eigenvalues of the four matrices are the same. Hence $\lambda_i(\boldsymbol{\Sigma}_a) = \lambda_i(\boldsymbol{\Sigma}_A) = \lambda_i(\boldsymbol{\Sigma}_b) = \lambda_i(\boldsymbol{\Sigma}_B) \equiv \lambda_i$ for $i = 1, ..., m$. It can be shown that $\boldsymbol{a}_i = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_i$ and $\boldsymbol{b}_i = \boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{g}_i$. The eigenvectors $\boldsymbol{a}_i$ are not necessarily orthonormal and the eigenvectors $\boldsymbol{b}_i$ are not necessarily orthonormal.

Theorem 7.1. Assume the $p \times p$ dispersion matrix $\boldsymbol{\Sigma}$ is positive definite. Assume $\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_b$ are positive definite and that $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ for some constant $c > 0$. Let $\boldsymbol{d}_i$ be an eigenvector of the corresponding matrix. Hence $\boldsymbol{d}_i = \boldsymbol{a}_i, \boldsymbol{b}_i, \boldsymbol{e}_i$ or $\boldsymbol{g}_i$. Let $(\hat{\lambda}_i, \hat{\boldsymbol{d}}_i)$ be the ith eigenvalue eigenvector pair of $\hat{\boldsymbol{\Sigma}}_\gamma$.

a) $\hat{\boldsymbol{\Sigma}}_\gamma \xrightarrow{P} \boldsymbol{\Sigma}_\gamma$ and $\hat{\lambda}_i(\hat{\boldsymbol{\Sigma}}_\gamma) \xrightarrow{P} \lambda_i(\boldsymbol{\Sigma}_\gamma) = \lambda_i$ where $\gamma = A, a, B$ or b.

b) $\boldsymbol{\Sigma}_\gamma \hat{\boldsymbol{d}}_i - \lambda_i \hat{\boldsymbol{d}}_i \xrightarrow{P} \boldsymbol{0}$ and $\hat{\boldsymbol{\Sigma}}_\gamma \boldsymbol{d}_i - \hat{\lambda}_i \boldsymbol{d}_i \xrightarrow{P} \boldsymbol{0}$.

c) If the jth eigenvalue λ_j is unique where $j \leq m$, then the absolute value of the correlation of $\hat{\boldsymbol{d}}_j$ with $\boldsymbol{d}_j$ converges to 1 in probability: $|\text{corr}(\hat{\boldsymbol{d}}_j, \boldsymbol{d}_j)| \xrightarrow{P} 1$.

Proof. a) $\hat{\boldsymbol{\Sigma}}_\gamma \xrightarrow{P} \boldsymbol{\Sigma}_\gamma$ since matrix multiplication is a continuous function of the relevant matrices and matrix inversion is a continuous function of a positive definite matrix. Then $\hat{\lambda}_i(\hat{\boldsymbol{\Sigma}}_\gamma) \xrightarrow{P} \lambda_i$ since an eigenvalue is a continuous function of its associated matrix.

b) Note that $(\boldsymbol{\Sigma}_\gamma - \lambda_i \boldsymbol{I})\hat{\boldsymbol{d}}_i = [(\boldsymbol{\Sigma}_\gamma - \lambda_i \boldsymbol{I}) - (\hat{\boldsymbol{\Sigma}}_\gamma - \hat{\lambda}_i \boldsymbol{I})]\hat{\boldsymbol{d}}_i = o_P(1)O_P(1) \xrightarrow{P} \boldsymbol{0}$, and $\hat{\boldsymbol{\Sigma}}_\gamma \boldsymbol{d}_i - \hat{\lambda}_i \boldsymbol{d}_i \xrightarrow{P} \boldsymbol{\Sigma}_\gamma \boldsymbol{d}_i - \lambda_i \boldsymbol{d}_i = \boldsymbol{0}$.

c) If n is large, then $\hat{\boldsymbol{d}}_i \equiv \hat{\boldsymbol{d}}_{i,n}$ is arbitrarily close to either $\boldsymbol{d}_i$ or $-\boldsymbol{d}_i$, and the result follows.

Rule of thumb 7.1. To use CCA, assume the DD plot and subplots of the scatterplot matrix are linear. Want $n > 10p$ for classical CCA and $n > 20p$ for robust CCA that uses FCH, RFCH or RMVN. Also make the DD plot for the $\boldsymbol{y}$ variables and the DD plot for the $\boldsymbol{z}$ variables.

Definition 7.1. Let the dispersion matrix be $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$. Let $(\lambda_i, \boldsymbol{e}_i)$ and $(\lambda_i, \boldsymbol{g}_i)$ be the eigenvalue eigenvector pairs of $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_B$. The kth pair of *population canonical variables* is

$$U_k = \boldsymbol{a}_k^T \boldsymbol{w} = \boldsymbol{e}_k^T \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{w} \text{ and } V_k = \boldsymbol{b}_k^T \boldsymbol{y} = \boldsymbol{g}_k^T \boldsymbol{\Sigma}_{22}^{-1/2} \boldsymbol{y}$$

for $k = 1, ..., m$. Then the *population canonical correlations* $\rho_k = corr(U_k, V_k)$

$= \sqrt{\lambda_k}$ for $k = 1, ..., m$. The vectors $\boldsymbol{a}_k = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_k$ and $\boldsymbol{b}_k = \boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{g}_k$ are the kth *canonical correlation coefficient vectors* for $\boldsymbol{w}$ and $\boldsymbol{y}$.

Theorem 7.2. Johnson and Wichern (1988, p. 440-441): Let the dispersion matrix be $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$. Then $V(U_k) = V(V_k) = 1$, $\text{Cov}(C_k, D_j) = corr(C_k, D_j) = 0$ for $k \neq j$ where $C_k = U_k$ or $C_k = V_k$, and $D_j = U_j$ or $D_j = V_j$ and $j, k = 1, ..., m$. That is, U_k is uncorrelated with V_j and U_j for $j \neq k$, and V_k is uncorrelated with V_j and U_j for $j \neq k$. The first pair of canonical variables is the pair of linear combinations (U, V) having unit variances that maximizes $\text{corr}(U, V)$ and this maximum is $\text{corr}(U_1, V_1) = \rho_1$. The ith pair of canonical variables are the linear combinations (U, V) with unit variances that maximize $\text{corr}(U, V)$ among all choices uncorrelated with the previous $k - 1$ canonical variable pairs.

Definition 7.2. Suppose standardized data $\boldsymbol{z} = (\boldsymbol{w}^T, \boldsymbol{y}^T)^T$ is used and the dispersion matrix is the correlation matrix $\boldsymbol{\Sigma} = \boldsymbol{\rho}$. Hence $\boldsymbol{\Sigma}_{ii} = \boldsymbol{\rho}_{ii}$ for $i = 1, 2$. Let $(\lambda_i, \boldsymbol{e}_i)$ and $(\lambda_i, \boldsymbol{g}_i)$ be the eigenvalue eigenvector pairs of $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_B$. The kth pair of *population canonical variables* is

$$U_k = \boldsymbol{a}_k^T\boldsymbol{w} = \boldsymbol{e}_k^T\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{w} \text{ and } V_k = \boldsymbol{b}_k^T\boldsymbol{y} = \boldsymbol{g}_k^T\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{y}$$

for $k = 1, ..., m$ for $k = 1, ..., m$. Then the *population canonical correlations* $\rho_k = corr(U_k, V_k) = \sqrt{\lambda_k}$ for $k = 1, ..., m$.

Then Theorem 7.2 holds for the standardized data and the canonical correlations are unchanged by the standardization.

Let

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{11} & \hat{\boldsymbol{\Sigma}}_{12} \\ \hat{\boldsymbol{\Sigma}}_{21} & \hat{\boldsymbol{\Sigma}}_{22} \end{pmatrix}.$$

Define estimators $\hat{\boldsymbol{\Sigma}}_a$, $\hat{\boldsymbol{\Sigma}}_A$, $\hat{\boldsymbol{\Sigma}}_b$ and $\hat{\boldsymbol{\Sigma}}_B$ in the same manner as their population analogs but using $\hat{\boldsymbol{\Sigma}}$ instead of $\boldsymbol{\Sigma}$. For example, $\hat{\boldsymbol{\Sigma}}_a = \hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{\Sigma}}_{22}^{-1}\hat{\boldsymbol{\Sigma}}_{21}$.

Let $\hat{\boldsymbol{\Sigma}}_a$ have eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{a}}_i)$, and let $\hat{\boldsymbol{\Sigma}}_A$ have eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ for $i = 1, ..., m$. Let $\hat{\boldsymbol{\Sigma}}_b$ have eigenvalue eigenvector pairs $(\hat{\lambda}_1, \hat{\boldsymbol{b}}_1)$, and let $\hat{\boldsymbol{\Sigma}}_B$ have eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{g}}_i)$ for $i = 1, ..., q$. For these four matrices $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_m$.

Definition 7.3. Let $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}$ if data $\boldsymbol{x} = (\boldsymbol{w}^T, \boldsymbol{y}^T)^T$ is used, and let $\hat{\boldsymbol{\Sigma}} = \boldsymbol{R}$ if standardized data $\boldsymbol{z} = (\boldsymbol{w}^T, \boldsymbol{y}^T)^T$ is used. The kth pair of *sample*

canonical variables is

$$\hat{U}_k = \hat{\boldsymbol{a}}_k^T \boldsymbol{w} = \hat{\boldsymbol{e}}_k^T \hat{\boldsymbol{\Sigma}}_{11}^{-1/2} \boldsymbol{w} \text{ and } \hat{V}_k = \hat{\boldsymbol{b}}_k^T \boldsymbol{y} = \hat{\boldsymbol{g}}_k^T \hat{\boldsymbol{\Sigma}}_{22}^{-1/2} \boldsymbol{y}$$

for $k = 1, ..., m$. Then the *sample canonical correlations* $\hat{\rho}_k = corr(\hat{U}_k, \hat{V}_k)$ $= \sqrt{\hat{\lambda}_k}$ for $k = 1, ..., m$. The vectors $\hat{\boldsymbol{a}}_k = \hat{\boldsymbol{\Sigma}}_{11}^{-1/2} \hat{\boldsymbol{e}}_k$ and $\hat{\boldsymbol{b}}_k = \hat{\boldsymbol{\Sigma}}_{22}^{-1/2} \hat{\boldsymbol{g}}_k$ are the kth *sample canonical correlation vectors* for $\boldsymbol{w}$ and $\boldsymbol{y}$.

Theorem 7.3. Under the conditions of Definition 7.3, the first pair of canonical variables $(\hat{U}_1, \hat{V}_1)$ is the pair of linear combinations $(\hat{U}, \hat{V})$ having unit sample variances that maximizes the sample correlation $\text{corr}(\hat{U}, \hat{V})$ and this maximum is $\text{corr}(\hat{U}_1, \hat{V}_1) = \hat{\rho}_1$. The ith pair of canonical variables are the linear combinations $(\hat{U}, \hat{V})$ with unit sample variances that maximize the sample $\text{corr}(\hat{U}, \hat{V})$ among all choices uncorrelated with the previous $k-1$ canonical variable pairs.

7.2 Robust CCA

The *R* function `cancor` does classical CCA and the *mpack* function `rcancor` does robust CCA by applying `cancor` on the RMVN set: the subset of the data used to compute RMVN.

Some theory is simple: the FCH, RFCH and RMVN methods of RCCA produce consistent estimators of the kth canonical correlation ρ_k on a large class of elliptically contoured distributions.

To see this, suppose $\text{Cov}(\boldsymbol{x}) = c_{\boldsymbol{x}} \boldsymbol{\Sigma}$ and $\boldsymbol{C} \equiv \boldsymbol{C}(\boldsymbol{X}) \xrightarrow{P} c\boldsymbol{\Sigma}$ where $c_x > 0$ and $c > 0$ are some constants. Then $\boldsymbol{C}_{XX}^{-1} \boldsymbol{C}_{XY} \boldsymbol{C}_{YY}^{-1} \boldsymbol{C}_{YX} \xrightarrow{P} \boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY} \boldsymbol{\Sigma}_{YY}^{-1} \boldsymbol{\Sigma}_{YX}$, and $\boldsymbol{C}_{YY}^{-1} \boldsymbol{C}_{YX} \boldsymbol{C}_{XX}^{-1} \boldsymbol{C}_{XY} \xrightarrow{P} \boldsymbol{\Sigma}_B = \boldsymbol{\Sigma}_{YY}^{-1} \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY}$. Note that $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_B$ only depend on $\boldsymbol{\Sigma}$ and do not depend on the constants c or $c_{\boldsymbol{x}}$.

(If $\boldsymbol{C}$ is also the classical covariance matrix applied to some subset of the data, then the correlation matrix $\boldsymbol{G} \equiv \boldsymbol{R}_{\boldsymbol{C}}$ applied to the same subset satisfies $\boldsymbol{G}_{XX}^{-1} \boldsymbol{G}_{XY} \boldsymbol{G}_{YY}^{-1} \boldsymbol{G}_{YX} \xrightarrow{P} \boldsymbol{R}_A = \boldsymbol{R}_{XX}^{-1} \boldsymbol{R}_{XY} \boldsymbol{R}_{YY}^{-1} \boldsymbol{R}_{YX}$, and $\boldsymbol{G}_{YY}^{-1} \boldsymbol{G}_{YX} \boldsymbol{G}_{XX}^{-1} \boldsymbol{G}_{XY} \xrightarrow{P} \boldsymbol{R}_B = \boldsymbol{R}_{YY}^{-1} \boldsymbol{R}_{YX} \boldsymbol{R}_{XX}^{-1} \boldsymbol{R}_{XY}$.)

Since eigenvalues are continuous functions of the associated matrix, and the FCH, RFCH and RMVN estimators are consistent estimators of $c_1\boldsymbol{\Sigma}, c_2\boldsymbol{\Sigma}$ and $c_3\boldsymbol{\Sigma}$ on a large class of elliptically contoured distributions, Theorem

7.1 holds, so these three RCCA methods and `rcancor` produce consistent estimators the kth canonical correlation ρ_k on that class of distributions.

Example 7.1. Example 2.2 describes the mussel data. Log transformation were taken on *muscle mass* M, *shell width* W and on the *shell mass* S. Then x contained the two log mass measurements while y contains L, H and $\log(W)$. The robust and classical CCAs were similar, but the canonical coefficients were difficult to interpret since $\log(W)$ has different units than L and H. Hence the log transformation were taken on all five variables and output is shown below.

The data set zm contains x and y, and the DD plot showed case 48 was separated from the bulk of the data, but near the identity line. The DD plot for x showed two cases, 8 and 48, were separated from the bulk of the data. Also the plotted points did not cluster tightly about the identity line. The DD plot for y looked fine. The classical CCA produces output \$cor, \$xcoef and \$ycoef. These are the canonical correlations, the $\boldsymbol{a}_i$ and the $\boldsymbol{b}_i$. The labels for the RCCA are \$out\$cor, \$out\$xcoef and \$out\$ycoef.

Note that the first correlation was about 0.98 while the second correlation was small. The RCCA is the CCA on the RMVN data set, which is contained in a compact ellipsoidal region. The variability of the truncated data set is less than that of the entire data set, hence expect the robust $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ to be larger in magnitude, ignoring sign, than that of the classical $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$, since the variance of each canonical variate is equal to one, and RCCA uses the truncated data. Note that $\boldsymbol{a}_1$ was roughly proportional to $\log(S)$ while $\boldsymbol{b}_1$ gave slightly higher weight for $\log(H)$ then $\log(W)$ and then $\log(L)$. Note that the five variables have high pairwise correlations, so $\log(M)$ was not important given that $\log(S)$ was in x. The second pair $(\boldsymbol{a}_2, \boldsymbol{b}_2)$ might be ignored since the second canonical correlation was very low.

```
> cancor(x,y)
$cor
[1] 0.9818605 0.1555381

$xcoef
        [,1]       [,2]
S 0.12650486  0.4077765
M 0.01897332 -0.4872522
```

```
$ycoef
       [,1]       [,2]       [,3]
L 0.1567463  0.7277888  2.1935890
W 0.1605139  0.8650480 -1.0676419
H 0.2143781 -2.0634587 -0.8303862

$xcenter
       S        M
4.563856 2.850187

$ycenter
       L        W        H
5.472944 3.697654 4.723295

> rcancor(x,y)
$out
$out$cor
[1] 0.98596703 0.06797587

$out$xcoef
        [,1]       [,2]
S 0.14966183  0.6460117
M 0.03236328 -0.8543387

$out$ycoef
       [,1]       [,2]       [,3]
L 0.1625452  0.4237524 -2.8492678
W 0.2369692  1.5379681  0.9356495
H 0.2530324 -2.6806462  1.7785931

$out$xcenter
       S        M
4.651941 2.948571

$out$ycenter
       L        W        H
5.496255 3.728292 4.745839
```

7.3 Summary

1) Let $\boldsymbol{x}$ be the $p \times 1$ vector of predictors, and partition $\boldsymbol{x} = (\boldsymbol{w}^T, \boldsymbol{y}^T)^T$ where $\boldsymbol{w}$ is $m \times 1$ and $\boldsymbol{y}$ is $q \times 1$ where $m = p - q \leq q$ and $m, q \geq 2$. Canonical correlation analysis (CCA) seeks m pairs of linear combinations $(\boldsymbol{a}_1^T\boldsymbol{w}, \boldsymbol{b}_1^T\boldsymbol{y}), ..., (\boldsymbol{a}_m^T\boldsymbol{w}, \boldsymbol{b}_m^T\boldsymbol{y})$ such that $\text{corr}(\boldsymbol{a}_i^T\boldsymbol{w}, \boldsymbol{b}_i^T\boldsymbol{y})$ is large under some constraints on the $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ where $i = 1, ..., m$. The first pair $(\boldsymbol{a}_1^T\boldsymbol{w}, \boldsymbol{b}_1^T\boldsymbol{y})$ has the largest correlation. The next pair $(\boldsymbol{a}_2^T\boldsymbol{w}, \boldsymbol{b}_2^T\boldsymbol{y})$ has the largest correlation among all pairs uncorrelated with the first pair and the process continues so that $(\boldsymbol{a}_m^T\boldsymbol{w}, \boldsymbol{b}_m^T\boldsymbol{y})$ is the pair with the largest correlation that is uncorrelated with the first $m - 1$ pairs. The correlations are called *canonical correlations* while the pairs of linear combinations are called *canonical variables*.

2) *R* output is shown in symbols for the following table.

corr					
$\hat{\rho}_1$	$\cdots$	$\hat{\rho}_1$			
wcoef					
$\boldsymbol{w}$	$\hat{\boldsymbol{a}}_1$	$\cdots$	$\hat{\boldsymbol{a}}_m$		
ycoef					
$\boldsymbol{y}$	$\hat{\boldsymbol{b}}_1$	$\cdots$	$\hat{\boldsymbol{b}}_m$	$\cdots$	$\hat{\boldsymbol{b}}_q$

```
64) $out$cor
[1] 0.98596703 0.06797587   $out$ycoef
$out$xcoef                        [,1]       [,2]       [,3]
        [,1]       [,2]   L 0.1625452  0.4237524 -2.8492678
S 0.14966183  0.6460117   W 0.2369692  1.5379681  0.9356495
M 0.03236328 -0.8543387   H 0.2530324 -2.6806462  1.7785931
```

3) Some notation is needed to explain CCA. Let the $p \times p$ positive definite symmetric dispersion matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Let $\boldsymbol{J} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$. Let $\boldsymbol{\Sigma}_a = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$, $\boldsymbol{\Sigma}_A = \boldsymbol{J}\boldsymbol{J}^T = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$, $\boldsymbol{\Sigma}_b = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}_B = \boldsymbol{J}^T\boldsymbol{J} = \boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$. Let $\boldsymbol{e}_i$ and $\boldsymbol{g}_i$ be sets of orthonormal eigenvectors, so $\boldsymbol{e}_i^T\boldsymbol{e}_i = 1$, $\boldsymbol{e}_i^T\boldsymbol{e}_j = 0$ for $i \neq j$, $\boldsymbol{g}_i^T\boldsymbol{g}_i = 1$ and $\boldsymbol{g}_i^T\boldsymbol{g}_j = 0$ for $i \neq j$. Let the $\boldsymbol{e}_i$ be $m \times 1$ while the $\boldsymbol{g}_i$ are $q \times 1$.

Let $\boldsymbol{\Sigma}_a$ have eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{a}_1), ..., (\lambda_m, \boldsymbol{a}_m)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. Let $\boldsymbol{\Sigma}_A$ have eigenvalue eigenvector pairs $(\lambda_i, \boldsymbol{e}_i)$ for $i =$

$1, ..., m$. Let $\boldsymbol{\Sigma}_b$ have eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{b}_1), ..., (\lambda_q, \boldsymbol{b}_q)$. Let $\boldsymbol{\Sigma}_B$ have eigenvalue eigenvector pairs $(\lambda_i, \boldsymbol{g}_i)$ for $i = 1, ..., q$. It can be shown that the m largest eigenvalues of the four matrices are the same. Hence $\lambda_i(\boldsymbol{\Sigma}_a) = \lambda_i(\boldsymbol{\Sigma}_A) = \lambda_i(\boldsymbol{\Sigma}_b) = \lambda_i(\boldsymbol{\Sigma}_B) \equiv \lambda_i$ for $i = 1, ..., m$. It can be shown that $\boldsymbol{a}_i = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_i$ and $\boldsymbol{b}_i = \boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{g}_i$. The eigenvectors $\boldsymbol{a}_i$ are not necessarily orthonormal and the eigenvectors $\boldsymbol{b}_i$ are not necessarily orthonormal.

Theorem 7.1. Assume the $p \times p$ dispersion matrix $\boldsymbol{\Sigma}$ is positive definite. Assume $\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_b$ are positive definite and that $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ for some constant $c > 0$. Let $\boldsymbol{d}_i$ be an eigenvector of the corresponding matrix. Hence $\boldsymbol{d}_i = \boldsymbol{a}_i, \boldsymbol{b}_i, \boldsymbol{e}_i$ or $\boldsymbol{g}_i$. Let $(\hat{\lambda}_i, \hat{\boldsymbol{d}}_i)$ be the ith eigenvalue eigenvector pair of $\hat{\boldsymbol{\Sigma}}_\gamma$.

a) $\hat{\boldsymbol{\Sigma}}_\gamma \xrightarrow{P} \boldsymbol{\Sigma}_\gamma$ and $\hat{\lambda}_i(\hat{\boldsymbol{\Sigma}}_\gamma) \xrightarrow{P} \lambda_i(\boldsymbol{\Sigma}_\gamma) = \lambda_i$ where $\gamma = A, a, B$ or b.

b) $\boldsymbol{\Sigma}_\gamma \hat{\boldsymbol{d}}_i - \lambda_i \hat{\boldsymbol{d}}_i \xrightarrow{P} \boldsymbol{0}$ and $\hat{\boldsymbol{\Sigma}}_\gamma \boldsymbol{d}_i - \hat{\lambda}_i \boldsymbol{d}_i \xrightarrow{P} \boldsymbol{0}$.

c) If the jth eigenvalue λ_j is unique where $j \leq m$, then the absolute value of the correlation of $\hat{\boldsymbol{d}}_j$ with $\boldsymbol{d}_j$ converges to 1 in probability: $|\text{corr}(\hat{\boldsymbol{d}}_j, \boldsymbol{d}_j)| \xrightarrow{P} 1$.

7.4 Complements

Muirhead and Waternaux (1980) shows that if the population canonical correlations ρ_k are distinct and if the underlying population distribution has a finite fourth moments, then the limiting joint distribution of $\sqrt{n}(\hat{\rho}_k^2 - \rho_k^2)$ is multivariate normal where the $\hat{\rho}_k$ are the classical sample canonical correlations and $k = 1, ..., p$. If the data are iid from an elliptically contoured distribution with kurtosis 3κ, then the limiting joint distribution of

$$\sqrt{n}\ \frac{\hat{\rho}_k^2 - \rho_k^2}{2\rho_k(1 - \rho_k^2)}$$

for $k = 1, ..., p$ is $N_p(\boldsymbol{0}, (\kappa + 1)\boldsymbol{I}_p)$. Note that $\kappa = 0$ for multivariate normal data.

Alkenani and Yu (2012), Zhang (2011) and Zhang, Olive and Ye (2012) develop robust CCA based on FCH, RFCH and RMVN.

7.5 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USE-

FUL.

7.1*. Examine the R output in Example 7.1. a) What is the first canonical correlation $\hat{\rho}_1$?

b) What is $\hat{\boldsymbol{a}}_1$?

c) What is $\hat{\boldsymbol{b}}_1$?

7.2. The R output below is for a canonical correlation analysis on Venables and Ripley (2003) CPU data. The variables were syct = log(cycle time + 1),
mmin = log(minimum main memory + 1),
chmin = log(minimum number of channels + 1),
chmax = log(maximum number of channels + 1),
perf = log(published performance + 1) and
estperf = $20/\sqrt{(}$estimated performance$+1)$. These six variables had a linear scatterplot matrix and DD plot and similar variances. Want to compare the two performance variables with the four remaining variables.

a) What is the first canonical correlation $\hat{\rho}_1$?

b) What is $\hat{\boldsymbol{a}}_1$?

c) What is $\hat{\boldsymbol{b}}_1$?

d) Interpret the second canonical variable $U_2 = \hat{\boldsymbol{a}}_2^T \boldsymbol{w}$.

```
> cancor(w,y)
$cor
[1] 0.8769433 0.2278554

$xcoef
```

```
              [,1]      [,2]
perf     0.02536432 0.1558717
estperf -0.04121870 0.1431100

$ycoef
              [,1]        [,2]         [,3]         [,4]
syct  -0.013613254  0.05700360  0.089757416 -0.011423664
mmin   0.037485282 -0.01874858  0.084442460  0.005859654
chmin  0.006932264  0.09843612 -0.021782624  0.090756713
chmax  0.019998948  0.01159728  0.007855559 -0.094198608
```

7.3. Edited SAS output for SAS Institute (1985, p. 146) Fitness Club Data is given below for CCA. Three physiological and three exercise variables measured on 20 middle aged men at a fitness club.

a) What is the first canonical correlation $\hat{\rho}_1$?

b) What is $\hat{\boldsymbol{a}}_1$?

c) What is $\hat{\boldsymbol{b}}_1$?

```
 Canonical
Correlation
0.7956
0.2006
0.0726
```

```
Raw Canonical Coefficients for the Physiological Variables
        PHYS1    PHYS2    PHYS3
weight -0.0314  -0.0763  -0.0077
waist   0.0493   0.3687   0.1580
pulse  -0.0082  -0.0321   0.1457

Raw Canonical Coefficients for the Exercise Variables
         Exer1    Exer2   Exer3
chinups -0.0661  -0.0714  -0.2428
situps  -0.0168   0.0020   0.0198
jumps    0.0140   0.0207  -0.0082
```

7.4. The output below is for a canonical correlations analysis on the R Seatbelts data set where $y_1 = drivers$ = number of drivers killed or seriously injured, $y_2 = front$ = number of front seat passengers killed or seriously injured, and $y_3 = rear$ = number of back seat passengers killed or seriously injured, $x_1 = kms$ = distance driven, $x_2 = PetrolPrice$ = petrol price and $x_3 = VanKilled$ = number of van drivers killed. The data consists of 192 monthly totals in Great Britain from January 1969 to December 1984.

a) What is the first canonical correlation $\hat{\rho}_1$?

b) What is $\hat{\boldsymbol{a}}_1$?

c) What is $\hat{\boldsymbol{b}}_1$?

d) Let $\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T$. The from the DD plot, the $\boldsymbol{z}_i$ appeared to follow a multivariate normal distribution. Sketch the DD plot.

```
> rcancor(x,y)
$out
$out$cor
```

```
[1] 0.8116953 0.5064619 0.1376399

$out$xcoef
                       [,1]          [,2]          [,3]
x.kms         -2.080206e-05 -0.0000233873 -2.259723e-06
x.PetrolPrice -1.847967e+00  3.7173715818  5.292041e+00
x.VanKilled    1.597620e-03 -0.0168450843  1.673662e-02

$out$ycoef
                    [,1]          [,2]          [,3]
y.drivers  1.678751e-06 -2.487259e-05  0.0004717902
y.front    5.594715e-04 -7.797027e-05 -0.0008157585
y.rear    -9.964980e-04 -7.521578e-04  0.0005045756
```

7.5. The R output below is for a canonical correlation analysis on some iris data. An iris is a flower, and there were 50 observations with 4 variables sepal length, sepal width, petal length and petal width.

a) What is the first canonical correlation $\hat{\rho}_1$?

b) What is $\hat{\boldsymbol{a}}_1$?

c) What is $\hat{\boldsymbol{b}}_1$?

```
w<-iris3[,,3]
x <- w[,1:2]
y <- w[,3:4]
cancor(x,y)

$cor
[1] 0.8642869 0.4836991

$xcoef
                 [,1]       [,2]
Sepal L. -0.223034210 -0.1186117
```

```
Sepal W. -0.006920448  0.4980378

$ycoef
                 [,1]        [,2]
Petal L. -0.257853414 -0.09094352
Petal W. -0.006108292  0.54939125
```

Chapter 8

Discriminant Analysis

8.1 Introduction

Definition 8.1. In *supervised classification*, there are k known groups and m cases. Each case is assigned to exactly one group based on its measurements $\boldsymbol{w}_i$.

Suppose there are k populations or groups where $k \geq 2$. Assume that for each population there is a probability density function (pdf) $f_j(\boldsymbol{z})$ where $\boldsymbol{z}$ is a $p \times 1$ vector and $j = 1, ..., k$. Hence if the random vector $\boldsymbol{x}$ comes from population j, then $\boldsymbol{x}$ has pdf $f_j(\boldsymbol{z})$. Assume that there is a random sample of n_j cases $\boldsymbol{x}_{1,j}, ..., \boldsymbol{x}_{n_j,j}$ for each group. Let $(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ denote the sample mean and covariance matrix for each group. Let $\boldsymbol{w}_i$ be a new $p \times 1$ random vector from one of the k groups, but the group is unknown. Usually there are many $\boldsymbol{w}_i$, and *discriminant analysis* attempts to allocate the $\boldsymbol{w}_i$ to the correct groups.

Definition 8.2. The *maximum likelihood discriminant rule* allocates case $\boldsymbol{w}$ to group a if $\hat{f}_a(\boldsymbol{w})$ maximizes $\hat{f}_j(\boldsymbol{w})$ for $j = 1, ..., k$.

For the following rules, assume that costs of correct and incorrect allocation are unknown or equal, and assume that the probabilities $\rho_a(\boldsymbol{w}_i)$ that $\boldsymbol{w}_i$ is in group a are unknown or equal: $\rho_a(\boldsymbol{w}_i) = 1/k$ for $a = 1, ..., k$. Often it is assumed that the k groups have the same covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. Then

the pooled covariance matrix estimator is

$$\boldsymbol{S}_{pool} = \frac{1}{n-k}\sum_{j=1}^{k}(n_j - 1)\boldsymbol{S}_j$$

where $n = \sum_{j=1}^{k} n_j$. Let $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$ be the estimator of multivariate location and dispersion for the jth group, eg the sample mean and sample covariance matrix $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$.

Definition 8.3. Assume the population dispersion matrices are equal: $\boldsymbol{\Sigma}_j \equiv \boldsymbol{\Sigma}$ for $j = 1, ..., k$. Let $\hat{\boldsymbol{\Sigma}}_{pool}$ be an estimator of $\boldsymbol{\Sigma}$. Then the *linear discriminant rule* is allocate $\boldsymbol{w}$ to the group with the largest value of

$$d_j(\boldsymbol{w}) = \hat{\boldsymbol{\mu}}_j^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1}\boldsymbol{w} - \frac{1}{2}\hat{\boldsymbol{\mu}}_j^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1}\hat{\boldsymbol{\mu}}_j = \hat{\alpha}_j + \hat{\boldsymbol{\beta}}_j^T \boldsymbol{w}$$

where $j = 1, ..., k$. *Linear discriminant analysis* (LDA) uses $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_{pool}) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_{pool})$.

Definition 8.4. The *quadratic discriminant rule* is allocate $\boldsymbol{w}$ to the group with the largest value of

$$Q_j(\boldsymbol{w}) = \frac{-1}{2}\log(|\hat{\boldsymbol{\Sigma}}_j|) - \frac{1}{2}(\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1}(\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)$$

where $j = 1, ..., k$. *Quadratic discriminant analysis* (QDA) uses $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$.

Definition 8.5. The *distance discriminant rule* allocates $\boldsymbol{w}$ to the group with the smallest squared distance $D^2_{\boldsymbol{w}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1}(\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)$ where $j = 1, ..., k$.

Definition 8.6. Assume that $k = 2$ and that there is a group 0 and a group 1. Let $\rho(\boldsymbol{w}) = P(\boldsymbol{w} \in \text{ group 1})$. Let $\hat{\rho}(\boldsymbol{w})$ be the logistic regression estimate of $\rho(\boldsymbol{w})$. The *logistic regression discriminant rule* allocates $\boldsymbol{w}$ to group 1 if $\hat{\rho}(\boldsymbol{w}) \geq 0.5$ and allocates $\boldsymbol{w}$ to group 0 if $\hat{\rho}(\boldsymbol{w}) < 0.5$. Logistic regression produces an estimated sufficient predictor $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w}$. Then

$$\hat{\rho}(\boldsymbol{w}) = \frac{e^{ESP}}{1 + e^{ESP}} = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w})}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w})}.$$

Let $Y_i = j$ if case i is in group j for $j = 0, 1$. Then a *response plot* is a plot of ESP versus Y_i (on the vertical axis) with $\hat{\rho}(\boldsymbol{x}_i) \equiv \hat{\rho}(ESP)$ added as a visual aid where $\boldsymbol{x}_i$ is the vector of predictors for case i. Also divide the ESP into J slices with approximately the same number of cases in each slice. Then compute the sample mean = sample proportion in slice s: $\hat{\rho}_s = \overline{Y}_s = \sum_s Y_i/m_s$ where m_s is the number of cases in slice s. Then plot the resulting step function as a visual aid. If n_0 and n_1 are the sample sizes of both groups and $n_i > 5p$, then the logistic regression model was useful if the step function of observed slice proportions scatter fairly closely about the logistic curve $\hat{\rho}(ESP)$.

Examining some of the rules for $k = 2$ and one predictor w is informative. First, assume group 2 has a uniform(−10,10) distribution and group 1 has a uniform$(a-1, a+1)$ distribution. If $a = 0$ is known, then the maximum likelihood discriminant rule assigns w to group 1 if $-1 < w < 1$ and assigns w to group 2, otherwise. This occurs since $f_2(w) = 1/20$ for $-10 < w < 10$ and $f_2(w) = 0$, otherwise, while $f_1(w) = 1/2$ for $-1 < w < 1$ and $f_1(w) = 0$, otherwise. For the distance rule, the distances are basically the absolute value of the z-score. Hence $D_1(w) \approx 1.732|w - a|$ and $D_2(w) \approx 0.1732|w|$. If w is from group 1, then w will not be classified very well unless $|a| \geq 10$ or if w is very close to a. In particular, if $a = 0$ then expect nearly all w to be classified to group 2 if w is used to classify the groups. On the other hand, if $a = 0$, then $D_1(w)$ is small for w in group 1 but large for w in group 2. Hence using $z = D_1(w)$ in the distance rule would result in classification with low error rates.

Similarly if group 2 comes from a $N_p(\boldsymbol{0}, 10\boldsymbol{I}_p)$ distribution and group 1 comes from a $N_p(\boldsymbol{\mu}, \boldsymbol{I}_p)$ distribution, the maximum likelihood rule will tend to classify $\boldsymbol{w}$ in group 1 if $\boldsymbol{w}$ is close to $\boldsymbol{\mu}$ and to classify $\boldsymbol{w}$ in group 2 otherwise. The two misclassification error rates should both be low. For the distance rule, the distances D_i have an approximate χ^2_p distribution if $\boldsymbol{w}$ is from group i. If covering ellipsoids from the two groups have little overlap, then the distance rule does well. If $\boldsymbol{\mu} = \boldsymbol{0}$, then expect all $\boldsymbol{w}$ to be classified to group 2 with the distance rule, but $D_1(\boldsymbol{w})$ will be small for $\boldsymbol{w}$ from group 1 and large for $\boldsymbol{w}$ from group 2, so using the single predictor $z = D_1(\boldsymbol{w})$ in the distance rule would result in classification with low error rates. More generally, if group 1 has a covering ellipsoid that has little overlap with the observations from group 2, using the single predictor $z = D_1(\boldsymbol{w})$ in the distance rule should result in classification with low error rates even if the

observations from group 2 do not fall in an ellipsoidal region.

Now suppose the k groups come from the same family of elliptically contoured $EC(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g)$ distributions where g is a decreasing function that does not depend on j for $j = 1, ..., k$. For example, could have $\boldsymbol{w} \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Using Equation (3.5), $\log(f_j(\boldsymbol{w})) =$

$$\log(k_p) - \frac{1}{2}\log(|\boldsymbol{\Sigma}_j|) + \log(g[(\boldsymbol{w} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_j)]) =$$

$$\log(k_p) - \frac{1}{2}\log(|\boldsymbol{\Sigma}_j|) + \log(g[D^2_{\boldsymbol{w}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]).$$

Hence the maximum likelihood rule leads to the quadratic rule if the k groups have $N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ distributions, and the maximum likelihood rule leads to the distance rule if the groups have dispersion matrices that have the same determinant: $\det(\boldsymbol{\Sigma}_j) = |\boldsymbol{\Sigma}_j| \equiv |\boldsymbol{\Sigma}|$ for $j = 1, ..., k$. This is a much weaker assumption that of equal dispersion matrices. For example, let $c_X \boldsymbol{\Sigma}_j$ be the covariance matrix of $\boldsymbol{x}$, and let $\boldsymbol{\Gamma}_j$ be an orthogonal matrix. Then $\boldsymbol{y} = \boldsymbol{\Gamma}_j \boldsymbol{x}$ corresponds to rotating $\boldsymbol{x}$, and $c_X \boldsymbol{\Gamma}_j \boldsymbol{\Sigma}_j \boldsymbol{\Gamma}_j^T$ is the covariance matrix of $\boldsymbol{y}$ with $|\mathrm{Cov}(\boldsymbol{x})| = |\mathrm{Cov}(\boldsymbol{y})|$.

Note that if the k groups come from the same family of elliptically contoured $EC(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g)$ distributions with nonsingular covariance matrices $c_X \boldsymbol{\Sigma}_j$, then $D^2_{\boldsymbol{w}}(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ is a consistent estimator of $D^2_{\boldsymbol{w}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)/c_X$. Hence the distance rule using $(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ is a maximum likelihood rule if the $\boldsymbol{\Sigma}_j$ have the same determinant.

Now $D^2_{\boldsymbol{w}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \boldsymbol{w}^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{w} - \boldsymbol{w}^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j - \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{w} + \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j = \boldsymbol{w}^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{w} - 2\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{w} + \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j = \boldsymbol{w}^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{w} + \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1}(-2\boldsymbol{w} + \boldsymbol{\mu}_j)$. Hence if $\boldsymbol{\Sigma}_j \equiv \boldsymbol{\Sigma}$ for $j = 1, ..., k$, then want to minimize $\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1}(-2\boldsymbol{w} + \boldsymbol{\mu}_j)$ or maximize $\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1}(2\boldsymbol{w} - \boldsymbol{\mu}_j)$, which is leads to the linear discriminant rule.

The maximum likelihood rule is robust to nonnormality, but it is difficult to estimate $\hat{f}_j(\boldsymbol{w})$ if $p > 1$. The linear discriminant rule and distance rule are robust to nonnormality, as is the logistic regression discriminant rule if $k = 2$. Expect the distance rule to be best when the ellipsoidal covering regions of the k groups have little overlap.

Rule of thumb 8.1. Use the distance rule if $n_j > 10p$ for $j = 1, ..., k$. Make the k DD plots using the $\boldsymbol{x}_{i,j}$ for each group to check for outliers, which could be cases that were incorrectly classified. If the distance rule error rates are very poor for some groups and very good for others, compute $z_j = D_j$, the distances for all n cases based on the jth group, where $j = 1, ..., k$. Since the

z_j may be highly correlated, use no more than $k-1$ of the z_j as predictors. The error rates computed using the data $\boldsymbol{x}_{i,j}$ with known groups give a lower bound on the error rates for the $\boldsymbol{w}_i$. That is, the error rates computed on the training data $\boldsymbol{x}_{i,j}$ are optimistic. When the discriminant rule is applied to the m $\boldsymbol{w}_i$ where the groups are unknown, the error rates will be higher. If equal covariance matrices are assumed, plot $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ versus $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{\Sigma}_{pool})$ for each of the k groups, where the $\boldsymbol{x}_{i,j}$ are used for $i = 1, ..., n_j$. The plotted points should cluster tightly about the identity line if n_j is large in each of the k plots if the assumption is reasonable. The linear discriminant rule has some robustness against the assumption of equal covariance matrices.

8.2 Two New Methods

Assume the k groups come from k distributions where the prediction regions from Section 5.2 are reasonable. For example, the jth group may have n_j cases that are iid $EC_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g_j)$ for $j = 1, ..., k$. That is, there may be k different elliptically contoured distributions with different location vectors and dispersion matrices.

Two new methods of discriminant analysis will be considered. For each group, compute $D_i(j) \equiv D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ and the maximum distance $D_{(n_j)}(j)$ where $i = 1, ..., n_j$ and $j = 1, ..., k$. Then $\{\boldsymbol{z} : D_{\boldsymbol{z}}(j) \leq D_{(n_j)}(j)\}$ is a covering region for the jth group since the hyperellipsoid contains all n_j cases $\boldsymbol{x}_{i,j}$ from the jth group.

Let $\boldsymbol{w}$ be a new case to be classified. If $D_{\boldsymbol{w}}(j) > D_{(n_j)}(j)$ for all $j = 1, ..., k$, then both Methods 1 and 2 allocate $\boldsymbol{w}$ to the group a with the smallest value of

$$\frac{D_{\boldsymbol{w}}(j)}{D_{(n_j)}(j)}. \tag{8.1}$$

Now consider the groups where $D_{\boldsymbol{w}}(j) \leq D_{(n_j)}(j)$ for at least one j. Hence $\boldsymbol{w}$ is in at least one of the k covering regions.

For Method 1, allocate $\boldsymbol{w}$ to group a with the smallest $D_{\boldsymbol{w}}(a)$ for the groups with $D_{\boldsymbol{w}}(j) \leq D_{(n_j)}(j)$. Method 1 is very similar to the distance rule, but when $\boldsymbol{w}$ is in at least one of the k covering regions, distances are only computed for the groups that have covering regions that contain $\boldsymbol{w}$. Also, Equation (8.1) is used instead of the smallest distance if $\boldsymbol{w}$ is not in any of the k covering regions.

Method 2 combines Method 1 with a maximum likelihood rule based on a kernel density estimator of $\hat{f}_j$. For Method 2, if there is only one group a where $D_{\boldsymbol{w}}(a) \leq D_{(n_a)}(a)$, allocate $\boldsymbol{w}$ to group a. Otherwise compute $\hat{f}_j(\boldsymbol{w})$ for the groups where $D_{\boldsymbol{w}}(j) \leq D_{(n_j)}(j)$ and allocate $\boldsymbol{w}$ to the group a with the largest $\hat{f}_a(\boldsymbol{w})$.

Note: To find the z_j of Rule of thumb 8.1, find $D_h(j)$ using all n of the $\boldsymbol{x}_{i,j}$, eg stack the $\boldsymbol{x}_{i,j}$ into an $n \times 1$ vector $\boldsymbol{x}$ and compute the $D_h(j)$ for $h = 1, ..., n$. These k new predictor variables still have known groups. Find $D_{\boldsymbol{w}_i}(j)$ for $i = 1, ..., m$ and $j = 1, ..., k$ to create k new predictor variables for the ith case to be classified. Then input up to $k - 1$ of these variables, with or without some of the p original predictor variables, into Method 1 or 2. Section 8.3 will give an example.

8.2.1 The Kernel Density Estimator

Definition 8.7. Let $K(\boldsymbol{z})$ be a multivariate probability density function. Then a *kernel density estimator* is

$$\hat{f}(\boldsymbol{z}) = \frac{1}{n} \frac{1}{h^p} \sum_{i=1}^{n} K\left(\frac{1}{h}(\boldsymbol{z} - \boldsymbol{x}_i)\right)$$

where there are n iid cases $\boldsymbol{x}_i$ that come from a population with unknown pdf $f(\boldsymbol{z})$.

For example, the uniform distribution on the unit hypersphere has

$$K(\boldsymbol{z}) = \frac{p\Gamma(p/2)}{2\pi^{p/2}} I(\boldsymbol{z}^T \boldsymbol{z} \leq 1)$$

so

$$\hat{f}(\boldsymbol{z}) = \frac{p\Gamma(p/2)}{2\pi^{p/2}} \frac{1}{n} \frac{1}{h^p} \sum_{i=1}^{n} I(\|\boldsymbol{z} - \boldsymbol{x}_i\|^2 \leq h^2).$$

Following Silverman (1986, p. 84), want the bias and variance of $\hat{f}$ to go to 0 as $n \to \infty$, and this will happen if $h \to 0$ and $nh^p \to \infty$. The asymptotically optimal value of h satisfies $h_{opt} \propto \frac{1}{n^{\frac{1}{p+4}}}$.

Now suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from a multivariate distribution with pdf f, and consider a hypersphere of radius r centered at $\boldsymbol{w}$ where r is small

enough so that if $\boldsymbol{z}$ is in the hypersphere, then $f(\boldsymbol{z}) \approx f(\boldsymbol{w})$. Then the probability that an observation $\boldsymbol{x}_i$ falls in the hypersphere $\approx f(\boldsymbol{w})$ (volume of the hypersphere) $= f(\boldsymbol{w})\frac{2\pi^{p/2}}{p\Gamma(p/2)}r^p \propto r^p$. Hence the number of $\boldsymbol{x}_i$ in the hypersphere $\propto nr^p$. If $r = h_{opt}$ then this number is $\propto n^{\frac{4}{4+p}}$. If $r = h \propto n^{\frac{1}{2p}}$, then the number of cases that fall in the hypersphere is proportional to $\sqrt{n}$.

To define the kernel density estimator used in Method 2, let $v_j = \lceil 2\sqrt{n_j} \rceil$ and let $r_j^2 = \|\boldsymbol{x}_{i,j} - \overline{\boldsymbol{x}}_j\|_{(v_j)}^2 = D_{(v_j)}^2(\overline{\boldsymbol{x}}_j, \boldsymbol{I}_p)$ where the n_h $\boldsymbol{x}_{i,j}$ are in group j. Hence the hypersphere centered at $\overline{\boldsymbol{x}}_j$ with radius r_j contains $\approx 2\sqrt{n}$ of the $\boldsymbol{x}_{i,j}$ in group j. Then the kernel density estimator used in Method 2 is

$$\hat{f}_j(\boldsymbol{w}) = \frac{p\Gamma(p/2)}{2\pi^{p/2}} \ \frac{1}{n_j} \ \frac{1}{(r_j)^p} \ \sum_{i=1}^{n_j} I(\|\boldsymbol{w} - \boldsymbol{x}_{i,j}\|^2 \leq r_j^2)$$

which is equal to the number of the $\boldsymbol{x}_{i,j}$ in the hypersphere of radius r_j centered at $\boldsymbol{w}$ divided by $n_j V_{r_j}$ where V_{r_j} is the volume of the hypersphere.

The main reasons for using this kernel density estimator are that it is simple to explain, fast to compute and does not use too few observations when $p > 4$. Since kernel density estimators do not work well for $p > 1$, speed is more important than asymptotic optimality. Also only need a crude estimator since if $f_a(\boldsymbol{w})$ is the pdf that maximizes $f_j(\boldsymbol{w})$, only need $\hat{f}_a(\boldsymbol{w})$ to maximize the $\hat{f}_j(\boldsymbol{w})$: hence extremely accurate estimators of the $f_j(\boldsymbol{w})$ are not needed. Using good predictors with p small is important since the performance of kernel density estimators decreases very rapidly as the number of predictors increases. See Silverman (1986, p. 94).

8.3 Some Examples

The *mpack* functions `ddiscr` and `ddiscr2` do discriminant analysis using Methods 1 and 2. The functions need x: the training data that has been classified into k groups, w: the data to be classified, *group*: a vector of integers where the ith element is j if the ith element of x is in group j, and $xwflag$ which is set equal to T if $w = x$ and to F if $w \neq x$. Each row of w and x corresponds to a case. The functions return the distances of the $\boldsymbol{x}$ and $\boldsymbol{w}$ computed for the k groups, the classifications for the $\boldsymbol{x}$ and $\boldsymbol{w}$, the error rates for the $\boldsymbol{x}$ classifications for each group, and the total error rate.

Example 8.1. Generated n random $N_p(\mathbf{0}, \boldsymbol{I}_p)$ random variables $\boldsymbol{x}_i$. Then $\boldsymbol{x}$ was put in group 1 if $D^2_{\boldsymbol{x}_i} \leq \chi^2_{p,0.5}$ and in group 2 otherwise. Expect group 2 to have smaller distances than group 1 so error rate will be near 1 for group 1 and near 0 for group 2. Output is shown below with $p = 2$ and shows that this was the case. Then the predictor $D_i(1)$ was used in *out2*, reducing the dimension from $p = 2$ to 1. The error rates were low since group 1 falls in an ellipsoidal region so the distances are a good predictor. Method 2 worked much better on the raw data and about the same as Method 1 when the predictor $D_i(1)$ was used.

```
n <- 100
p <- 2
x <- matrix(rnorm(n*p),nrow=n,ncol=p)
group <- 1 + 0*1:n
covv <- diag(p)
mns<- apply(x, 2, mean)
md2 <- mahalanobis(x, center = mns, covv)
group[md2>qchisq(0.5,p)] <- 2

out1 <- ddiscr(x,w=x,group,xwflag=T)
out2<-ddiscr(x=out1$mdx[,1],w=out1$mdw[,1],group,xwflag=T)
out3 <- ddiscr2(x,w=x,group,xwflag=T)
out4<-ddiscr2(x=out1$mdx[,1],w=out1$mdw[,1],group,xwflag=T)

out1$err
[1] 0.9787234 0.0000000
out2$err
[1] 0.08510638 0.01886792
out3$err
[1] 0.0000000 0.1320755
out4$err
[1] 0.04255319 0.05660377

out1$toterr
[1] 0.46
out2$toterr
[1] 0.05
out3$toterr
```

```
[1] 0.07
out4$toterr
[1] 0.05
```

Example 8.2. Now groups 1 and 2 had $n_i = 50$, and group 1 used $\boldsymbol{x} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ while group 2 used $\boldsymbol{x} \sim N_p(2 \ \ \boldsymbol{1}, \ \boldsymbol{I}_p)$. Output is shown below for $p = 2$. Now the single predictor $D_i^2(1)$ was slightly worse than using the raw data, and Method 1 was about as good as Method 2, which is not surprising since both methods approximate the maximum likelihood discriminant rule when the groups are multivariate normal with the same covariance matrix.

```
 n <- 100
 p <- 2
 x <- matrix(rnorm(n*p),nrow=n,ncol=p)
 group <- 1 + 0*1:n
 group[1:50] <- 1
 group[51:100] <- 2
 x[51:100,] <- x[51:100,] + c(2,2)
 out1 <- ddiscr(x,w=x,group,xwflag=T)
 out2<-ddiscr(x=out1$mdx[,1],w=out1$mdw[,1],group,xwflag=T)
 out3 <- ddiscr2(x,w=x,group,xwflag=T)
 out4<-ddiscr2(x=out1$mdx[,1],w=out1$mdw[,1],group,xwflag=T)

out1$err
[1] 0.12 0.08
out2$err
[1] 0.14 0.10
out3$err
[1] 0.08 0.12
out4$err
[1] 0.14 0.10

library(MASS)
group <- pottery[pottery[,1]!=5,1]
group <- (as.integer(group!=1)) + 1
x <- pottery[pottery[,1]!=5,-1]
```

```
out<-lda(x,group)
1-mean(predict(out,x)$class==group)
[1] 0.03571429
out<-lda(x[,-c(1)],group)
1-mean(predict(out,x[,-c(1)])$class==group)
out<-lda(x[,-c(1,2)],group)
1-mean(predict(out,x[,-c(1,2)])$class==group)
out<-lda(x[,-c(1,2,3)],group)
1-mean(predict(out,x[,-c(1,2,3)])$class==group)
out<-lda(x[,-c(1,2,3,4)],group)
1-mean(predict(out,x[,-c(1,2,3,4)])$class==group)
out<-lda(x[,-c(1,2,3,4,5)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5)])$class==group)
[1] 0.03571429
out<-lda(x[,-c(1,2,3,4,5,6)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,6)])$class==group)
[1] 0.07142857
out<-lda(x[,-c(1,2,3,4,5,7)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7)])$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8)])$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9)])$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10)])$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,11)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,11)])$class==group)
[1] 0.07142857
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12)])$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13)])$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,14)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,14)])$class==
group)
[1] 0.07142857
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,15)])$class==
```

```
group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16)])$
class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17)])$
class==group)
[1] 0.03571429
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17,18)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17,18)])
$class==group)
[1] 0.07142857
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17,19)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17,19)])
$class==group)
[1] 0.03571429
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17,19,20)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17,19,
20)])$class==group)
[1] 0

#x6,x11,x14,x18 seem good for LDA
```

8.4 Summary

1) In *supervised classification*, there are k known groups or populations and m cases. Each case is assigned to exactly one group based on its measurements $\boldsymbol{w}_i$. Assume that for each population there is a probability density function (pdf) $f_j(\boldsymbol{z})$ where $\boldsymbol{z}$ is a $p \times 1$ vector and $j = 1, ..., k$. Hence if the random vector $\boldsymbol{x}$ comes from population j, then $\boldsymbol{x}$ has pdf $f_j(\boldsymbol{z})$. Assume that there is a random sample of n_j cases $\boldsymbol{x}_{1,j}, ..., \boldsymbol{x}_{n_j,j}$ for each group. Let $(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ denote the sample mean and covariance matrix for each group. Let $\boldsymbol{w}_i$ be a new $p \times 1$ random vector from one of the k groups, but the group is unknown. Usually there are many $\boldsymbol{w}_i$, and *discriminant analysis* attempts to allocate the $\boldsymbol{w}_i$ to the correct groups.

2) The *maximum likelihood discriminant rule* allocates case $\boldsymbol{w}$ to group a if $\hat{f}_a(\boldsymbol{w})$ maximizes $\hat{f}_j(\boldsymbol{w})$ for $j = 1, ..., k$. This rule is robust to nonnormality

and the assumption of equal population dispersion matrices, but $\hat{f}_j$ is hard to compute for $p > 1$.

3) Given the $\hat{f}_j(\boldsymbol{w})$ or a plot of the $\hat{f}_j(\boldsymbol{w})$, determine the maximum likelihood discriminant rule.

For the following rules, assume that costs of correct and incorrect allocation are unknown or equal, and assume that the probabilities $\rho_a(\boldsymbol{w}_i)$ that $\boldsymbol{w}_i$ is in group a are unknown or equal: $\rho_a(\boldsymbol{w}_i) = 1/k$ for $a = 1, ..., k$. Often it is assumed that the k groups have the same covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. Then the pooled covariance matrix estimator is

$$\boldsymbol{S}_{pool} = \frac{1}{n-k}\sum_{j=1}^{k}(n_j - 1)\boldsymbol{S}_j$$

where $n = \sum_{j=1}^{k} n_j$. Let $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$ be the estimator of multivariate location and dispersion for the jth group, eg the sample mean and sample covariance matrix $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$.

4) Assume the population dispersion matrices are equal: $\boldsymbol{\Sigma}_j \equiv \boldsymbol{\Sigma}$ for $j = 1, ..., k$. Let $\hat{\boldsymbol{\Sigma}}_{pool}$ be an estimator of $\boldsymbol{\Sigma}$. Then the *linear discriminant rule* is allocate $\boldsymbol{w}$ to the group with the largest value of

$$d_j(\boldsymbol{w}) = \hat{\boldsymbol{\mu}}_j^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1}\boldsymbol{w} - \frac{1}{2}\hat{\boldsymbol{\mu}}_j^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1}\hat{\boldsymbol{\mu}}_j = \hat{\alpha}_j + \hat{\boldsymbol{\beta}}_j^T \boldsymbol{w}$$

where $j = 1, ..., k$. *Linear discriminant analysis* (LDA) uses $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_{pool}) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_{pool})$. LDA is robust to nonnormality and somewhat robust to the assumption of equal population covariance matrices.

5) The *quadratic discriminant rule* is allocate $\boldsymbol{w}$ to the group with the largest value of

$$Q_j(\boldsymbol{w}) = \frac{-1}{2}\log(|\hat{\boldsymbol{\Sigma}}_j|) - \frac{1}{2}(\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1}(\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)$$

where $j = 1, ..., k$. *Quadratic discriminant analysis* (QDA) uses $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$. QDA has some robustness to nonnormality.

6) The *distance discriminant rule* allocates $\boldsymbol{w}$ to the group with the smallest squared distance $D^2_{\boldsymbol{w}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1}(\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)$ where $j = 1, ..., k$. This rule is robust to nonnormality and the assumption of equal $\boldsymbol{\Sigma}_j$, but needs $n_j > 10p$ for $j = 1, ..., k$.

7) Assume that $k = 2$ and that there is a group 0 and a group 1. Let $\rho(\boldsymbol{w}) = P(\boldsymbol{w} \in \text{group 1})$. Let $\hat{\rho}(\boldsymbol{w})$ be the logistic regression (LR) estimate of $\rho(\boldsymbol{w})$. Logistic regression produces an estimated sufficient predictor $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w}$. Then

$$\hat{\rho}(\boldsymbol{w}) = \frac{e^{ESP}}{1 + e^{ESP}} = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w})}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w})}.$$

The *logistic regression discriminant rule* allocates $\boldsymbol{w}$ to group 1 if $\hat{\rho}(\boldsymbol{w}) \geq 0.5$ and allocates $\boldsymbol{w}$ to group 0 if $\hat{\rho}(\boldsymbol{w}) < 0.5$. Equivalently, the LR rule allocates $\boldsymbol{w}$ to group 1 if $ESP > 0$ and allocates $\boldsymbol{w}$ to group 0 if $ESP < 0$.

8) Let $Y_i = j$ if case i is in group j for $j = 0, 1$. Then a *response plot* is a plot of ESP versus Y_i (on the vertical axis) with $\hat{\rho}(\boldsymbol{x}_i) \equiv \hat{\rho}(ESP)$ added as a visual aid where $\boldsymbol{x}_i$ is the vector of predictors for case i. Also divide the ESP into J slices with approximately the same number of cases in each slice. Then compute the sample mean = sample proportion in slice s: $\hat{\rho}_s = \overline{Y}_s = \sum_s Y_i / m_s$ where m_s is the number of cases in slice s. Then plot the resulting step function as a visual aid. If n_0 and n_1 are the sample sizes of both groups and $n_i > 5p$, then the logistic regression model was useful if the step function of observed slice proportions scatter fairly closely about the logistic curve $\hat{\rho}(ESP)$. If the LR response plot is good, $n_0 > 5p$ and $n_1 > 5p$, then the LR rule is robust to nonnormality and the assumption of equal population dispersion matrices. Know how to tell a good LR response plot from a bad one.

9) Given LR output, as shown below in symbols and for a real data set, and given $\boldsymbol{x}$ to classify, be able to a) compute ESP, b) classify $\boldsymbol{x}$ in group 0 or group 1, c) compute $\hat{\rho}(\boldsymbol{x})$.

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\alpha}$	$se(\hat{\alpha})$	$z_{o,0}$	for Ho: $\alpha = 0$
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1 / se(\hat{\beta}_1)$	for Ho: $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p / se(\hat{\beta}_p)$	for Ho: $\beta_p = 0$

```
Binomial Regression Kernel mean function = Logistic
Response = Status  Terms = (Bottom Left) Trials = Ones
Coefficient Estimates
```

```
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -389.806         104.224        -3.740     0.0002
Bottom     2.26423         0.333233        6.795     0.0000
Left       2.83356         0.795601        3.562     0.0004
```

10) Suppose there is training data $\boldsymbol{x}_{ij}$ for $i = 1, ..., n_j$ for group j. Hence it is known that $\boldsymbol{x}_{ij}$ came from group j where there are $k \geq 2$ groups. Use the discriminant analysis method to classify the training data. If m_j of the n_j group j cases are correctly classified, then the *apparent error rate for group* j is $1 - m_j/n_j$. If $m_A = \sum_{j=1}^{k} m_j$ of the $n = \sum_{j=1}^{k} n_j$ cases were correctly classified. Then the *apparent error rate* $\text{AER} = 1 - m_A/n$.

11) For the `ddiscr` method, get the apparent error rate for each of the k groups with the following commands. Replace `ddiscr` by `ddiscr2` for the `ddiscr2` method.

```
out1 <- ddiscr(x,w=x,group,xwflag=T)
out1$err
```

Get apparent error rates for `ddiscr, LDA` and `QDA` with the following commands.

```
out1 <- ddiscr(x,w=x,group,xwflag=T)
out1$toterr

out2  <- lda(x,group)
1-mean(predict(out2,x)$class==group)

out3  <- qda(x,group)
1-mean(predict(out3,x)$class==group)
```

Get the AERs for the methods that use variables x_1, x_3 and x_7 with the following commands.

```
out <- ddiscr(x[,c(1,3,7)],w=x[,c(1,3,7)],group,xwflag=T)
out$toterr

out <- lda(x[,c(1,3,7)],group)
1-mean(predict(out,x[,c(1,3,7)])$class==group)
```

```
out <- qda(x[,c(1,3,7)],group)
1-mean(predict(out,x[,c(1,3,7)])$class==group)
```

Get the AERs for the methods that leave out variables x_1, x_4 and x_5 with the following commands.

```
out <- ddiscr(x[,-c(1,4,5)],w=x[,-c(1,4,5)],group,xwflag=T)
out$toterr

out <- lda(x[,-c(1,4,5)],group)
1-mean(predict(out,x[,-c(1,4,5)])$class==group)

out <- qda(x[,-c(1,4,5)],group)
1-mean(predict(out,x[,-c(1,4,5)])$class==group)
```

12) Expect the apparent error rate to be too low: the method works better on the training data than on the new data to be classified.

13) Cross validation (CV): for $i = 1, ..., n$ where the training data has n cases, compute the discriminant rule with case i left out and see if the rule correctly classifies case i. Let m_C be the number of cases correctly classified. Then the CV error rate is $1 - m_C/n$.

14) Suppose the training data has n cases. Randomly select a subset L of m cases to be left out when computing the discriminant rule. Hence $n - m$ cases are used to compute the discriminant rule. Let m_L be the number of cases from subset L that are correctly classified. Then the "leave a subset out" error rate is $1 - m_L/m$. Here m should be large enough to get a good rate. Often m uses between $0.1n$ and $0.5n$.

15) Variable selection is the search for a subset of variables that does a good job of classification.

16) Forward selection: suppose $X_1, ..., X_p$ are variables.

Step 1) Choose variable $W_1 = X_1$ that minimizes the AER.

Step 2) Keep W_1 in the model, and add variable W_2 that minimizes the AER. So W_1 and W_2 are in the model at the end of Step 2).

Step k) Have $W_1, ..., W_{k-1}$ in the model. Add variable W_k that minimizes the AER. So $W_1, ..., W_k$ are in the model at the end of Step k).

Step p) $W_1, ..., W_p = X_1, ..., X_p$, so all p variables are in the model.

17) Backward elimination: suppose $X_1, ..., X_p$ are variables.

Step 1) $W_1, ..., W_p = X_1, ..., X_p$, so all p variables are in the model.

Step 2) Delete variable $W_p = X_j$ such that the model with $p-1$ variables $W_1, ..., W_{p-1}$ minimizes the AER.

Step 3) Delete variable $W_{p-1} = X_j$ such that the model with $p-2$ variables $W_1, ..., W_{p-2}$ minimizes the AER.

Step k) $W_1, ..., W_{p-k+2}$ are in the model. Delete variable $W_{p-k+2} = X_j$ such that the model with $p-k+1$ variables $W_1, ..., W_{p-k+1}$ minimizes the AER.

Step p) Have W_1 and W_2 in the model. Delete variable W_2 such that the model with 1 variable W_1 minimizes the AER.

18) Other criterion can be used and `proc stepdisc` in *SAS* does variable selection.

19) In *R*, using LDA, leave one variable out at a time as long as the AER does not increase much, to find a good subset quickly.

8.5 Complements

For $k = 2$, an alternative to the logistic regression model is the discriminant function model. See Hosmer and Lemeshow (2000, p. 43–44). Assume that $\rho_j = P(Y = j)$ and that $\boldsymbol{x}|Y = j \sim N_k(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j = 0, 1$. That is, the conditional distribution of $\boldsymbol{x}$ given $Y = j$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}$ which does not depend on j. Notice that $\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{x}|Y) \neq \text{Cov}(\boldsymbol{x})$. Then as for the logistic regression model,

$$P(Y = 1|\boldsymbol{x}) = \rho(\boldsymbol{x}) = \frac{\exp(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})}{1 + \exp(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})}.$$

Definition 8.8. Under the conditions above, the **discriminant function** parameters are given by

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{8.2}$$

and

$$\alpha = \log\left(\frac{\rho_1}{\rho_0}\right) - 0.5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0).$$

To use Definition 8.8 to simulate logistic regression data, set $\rho_0 = \rho_1 = 0.5$, $\boldsymbol{\Sigma} = \boldsymbol{I}$, and $\boldsymbol{\mu}_0 = \boldsymbol{0}$. Then $\alpha = -0.5\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1$ and $\boldsymbol{\beta} = \boldsymbol{\mu}_1$. The discriminant function estimators $\hat{\alpha}_D$ and $\hat{\boldsymbol{\beta}}_D$ are found by replacing the population quantities ρ_1, ρ_0, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}$ by sample quantities. Alternatively, generate n values of the $SP_i = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i$, then generate a binomial$(1, \rho(SP_i))$ case for $i = 1, ..., n$. This alternative method is useful since the $\boldsymbol{x}_i$ need not be from a multivariate normal distribution.

See Olive (2010: ch. 10, 2013) for more information about logistic regression and response plots for logistic regression.

Huberty and Olejnik (2006) and McLachlan (2004) are useful references for discriminant analysis. Silverman (1986, § 6.1) and Raveh (1989) are good references for nonparametric discriminant analysis. Discrimination when $p > n$ is interesting. See Cai and Liu (2011) and Mai, Zou and Yuan (2012).

Logistic regression is a useful alternative to discriminant analysis when there are two groups. The distance rule and Methods 1 and 2 can use RFCH or RMVN to compute $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$.

Hand (2006) notes that supervised classification is a research area in statistics, machine learning, pattern recognition, computational learning theory and data mining. Hand (2006) argues that simple classification methods, such as linear discriminant analysis, are almost as good as more sophisticated methods such as neural networks and support vector machines.

8.6 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

8.1*. Assume the cases in each of the k groups are iid from a population with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}(j)$ Find $E(\boldsymbol{S}_{pool})$ assuming that the k groups have the same covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}(j) \equiv \boldsymbol{\Sigma}_{\boldsymbol{x}}$ for $j = 1, ..., k$.

```
Logistic Regression Output,
Response = nodal involvement, Terms = (acid size xray)
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -3.57564         1.18002        -3.030     0.0024
acid       2.06294         1.26441         1.632     0.1028
size       1.75556         0.738348        2.378     0.0174
```

```
xray        2.06178          0.777103         2.653      0.0080

Number of cases: 53, Degrees of freedom: 49, Deviance: 50.660
```

8.2. Following Collett (1999, p. 11), treatment for prostate cancer depends on whether the cancer has spread to the surrounding lymph nodes. Let the response variable = group $y = nodal\ involvement$ (0 for absence, 1 for presence). Let $x_1 = acid$ (serum acid phosphatase level), $x_2 = size$ (= tumor size: 0 for small, 1 for large) and $x_3 = xray$ (xray result: 0 for negative, 1 for positive). Assume the case to be classified has $\boldsymbol{x}$ with $x_1 = acid = 0.65$, $x_2 = 0$ and $x3 = 0$.

a) Find ESP for $\boldsymbol{x}$.

b) Is $\boldsymbol{x}$ classified in group 0 or group 1?

c) Find $\hat{\rho}(\boldsymbol{x})$.

8.3. Recall that X comes from a uniform(a,b) distribution, written $x \sim U(a, b)$, if the pdf of x is $f(x) = \dfrac{1}{b-a}$ for $a < x < b$ and $f(x) = 0$, otherwise. Suppose group 1 has $X \sim U(-3, 3)$, group 2 has $X \sim U(-5, 5)$, and group 3 has $X \sim U(-1, 1)$. Find the maximum likelihood discriminant rule for classifying a new observation x.

```
out<-prcomp(state[,1:4],scale=T)
summary(out)
Importance of components: PC1    PC2    PC3     PC4
Standard deviation     1.6040 0.8803 0.6879 0.42318
Proportion of Variance 0.6432 0.1937 0.1183 0.04477
Cumulative Proportion  0.6432 0.8369 0.9552 1.00000

> out<-rprcomp(state[,1:4])
summary(out$out)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.6705 0.8216 0.59362 0.42645
Proportion of Variance 0.6977 0.1688 0.08809 0.04546
Cumulative Proportion  0.6977 0.8664 0.95454 1.00000

Rotation:PC1          PC2        PC3         PC4
gdp     0.4525021  0.688328888 -0.5429877 -0.1631243
```

```
povrt  -0.5563898 -0.016929402 -0.2468286 -0.7932335
unins  -0.4442238  0.725197372  0.5076082  0.1381588
lifexp  0.5369706  0.002347129  0.6217506 -0.5701607

out <- lda(state[,1:4],state[,5])
1-mean(predict(out,state[,1:4])$class==state[,5])
[1] 0.3
```

8.4. The PCA and LDA output above is for the Minor (2012) state data where gdp = GDP per capita, povrt = poverty rate, unins = 3 year average uninsured rate 2007-9, and lifexp = life expectancy for the 50 states.

a) How many principal components are needed? Use a 0.9 threshold.

b) Which principal component corresponds to 9 gdp -9 unins -11 povrt $+11$ lifeexp?

c) The fifth variable was a 1 if the state was not worker friendly and a 2 if the state was worker friendly. With these two groups, what was the apparent error rate (AER) for LDA?

```
> out <- lda(x,group)
> 1-mean(predict(out,x)$class==group)
[1] 0.02
>
> out<-lda(x[,-c(1)],group)
> 1-mean(predict(out,x[,-c(1)])$class==group)
[1] 0.02
> out<-lda(x[,-c(1,2)],group)
> 1-mean(predict(out,x[,-c(1,2)])$class==group)
[1] 0.04
> out<-lda(x[,-c(1,3)],group)
> 1-mean(predict(out,x[,-c(1,3)])$class==group)
[1] 0.03333333
> out<-lda(x[,-c(1,4)],group)
> 1-mean(predict(out,x[,-c(1,4)])$class==group)
[1] 0.04666667
>
> out<-lda(x[,c(2,3,4)],group)
> 1-mean(predict(out,x[,c(2,3,4)])$class==group)
[1] 0.02
```

8.5. The above output is for LDA on the famous iris data set. the variables are x_1 = sepal length, x_2 = sepal width, x_3 = petal length and x_4 = petal width. These four predictors are in the x data matrix. There are three groups corresponding to types of iris: setosa versicolor virginica.

a) What is the AER using all 4 predictors?

b) Which variables, if any, can be deleted without increasing the AER in a)?

R/Splus Problems

Warning: Use the command *source("G:/mpack.txt")* **to download the programs. See Preface or Section 15.2.** Typing the name of the `mpack` function, eg *ddplot*, will display the code for the function. Use the `args` command, eg *args(ddplot)*, to display the needed arguments for the function.

8.5. Wisseman, Hopke and Schindler-Kaudelka (1987) pottery data has 36 pottery shards of Roman earthware produced between second century B.C. and fourth century A.D. Often the pottery was stamped by the manufacturer. A chemical analysis was done for 20 chemicals (variables), and 28 cases were classified as Arrentine (group 1) or nonArrentine (group 2), while 8 cases were of questionable origin. So the training data has $n = 28$ and $p = 20$.

a) Copy and paste the R commands for this part into R to make the data set.

b) Because of the small sample size, LDA should be used instead of QDA, as in the handout. Nonetheless, variable selection using QDA will be done. Copy and paste the R commands for this part into R. The first 9 variables result in no misclassification errors.

c) Now use commands like those shown in this section to delete variables whose deletion does not result in a classification error. Should get four variables are needed for perfect classification. What are they (eg X1, X2, X3 and X4)?

8.6. The distance discriminant rule is attractive theoretically as a maximum likelihood discriminant rule, but the distance rule does not work well for groups that have similar means. The `ddiscr` rule is a modification of the distance rule, and the `ddiscr2` rule tries to use the maximum likelihood rule where the $\hat{f}_j$ are estimated with a kernel density estimator.

The R code for this problem generates $N_2(\mathbf{0}, \boldsymbol{I})$ data where group 1 consists of the half set of cases closes to $\mathbf{0}$ in Mahalanobis distance (an ellipse

about the origin), and group 2 consists of the remaining cases (the covering ellipse with inner ellipse removed).

a) Copy and paste the commands to make the data.

b) The commands for this part give the error rate for the ddiscr method that uses $\boldsymbol{x}$ as the two predictors. Put this output in *Word*.

c) The commands for this part give the error rate for the ddiscr method that uses the distances based on group 1 applied to all of the cases as the predictor. Put this output in *Word*.

d) The commands for this part give the error rate for the ddiscr2 method that uses $\boldsymbol{x}$ as the two predictors. Put this output in *Word*.

e) The commands for this part give the error rate for the ddiscr2 method that uses the distances based on group 1 applied to all of the cases as the predictor. Put this output in *Word*.

f) The commands for this part get the error rate for LDA using $\boldsymbol{x}$ as the two predictors.

g) The commands for this part get the error rate for QDA using $\boldsymbol{x}$ as the two predictors.

h) Which method worked the best?

Chapter 9

Hotelling's T^2 Test

9.1 One Sample

The one sample Hotelling's T^2 test is used to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. The test rejects H_0 if

$$T_H^2 = n(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0) > \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$$

where $P(Y \leq F_{p,d,\alpha}) = \alpha$ if $Y \sim F_{p,d}$.

If a multivariate location estimator T satisfies

$$\sqrt{n}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{D}),$$

then a competing test rejects H_0 if

$$T_C^2 = n(T - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{D}}^{-1} (T - \boldsymbol{\mu}_0) > \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$$

if H_0 holds and $\hat{\boldsymbol{D}}$ is a consistent estimator of $\boldsymbol{D}$. The scaled F cutoff can be used since $T_C^2 \xrightarrow{D} \chi_p^2$ if H_0 holds, and

$$\frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha} \to \chi^2_{p,1-\alpha}$$

as $n \to \infty$. This idea is used for small p by Srivastava and Mudholkar (2001) where T is the coordinatewise trimmed mean. The one sample Hotelling's T^2 test uses $T = \overline{\boldsymbol{x}}$, $\boldsymbol{D} = \boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{D}} = \boldsymbol{S}$.

The Hotelling's T^2 test is a large sample level α test in that if $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from a distribution with mean $\boldsymbol{\mu}_0$ and nonsingular covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then the type I error = P(reject H_0 when H_0 is true) $\rightarrow \alpha$ as $n \rightarrow \infty$. Want $n > 10p$ if the DD plot is linear through the origin and subplots in the scatterplot matrix all look ellipsoidal. For any n, there are distributions with nonsingular covariance matrix where the χ^2_p approximation to T^2_H is poor.

Let pval be an estimate of the pvalue. Typically use $T^2_C = T^2_H$ in the following 4 step test. i) State the hypotheses $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$.
ii) Find the test statistic $T^2_C = n(T - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{D}}^{-1}(T - \boldsymbol{\mu}_0)$.
iii) Find pval =

$$P\left(T^2_C < \frac{(n-1)p}{n-p} F_{p,n-p}\right) = P\left(\frac{n-p}{(n-1)p} \; T^2_C < F_{p,n-p}\right).$$

iv) State whether you fail to reject H_0 or reject H_0. If you reject H_0 then conclude that $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ while if you fail to reject H_0 conclude that the population mean $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ or that there is not enough evidence to conclude that $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. Reject H_0 if pval $< \alpha$ and fail to reject H_0 if pval $\geq \alpha$. As a benchmark for this text, use $\alpha = 0.05$ if α is not given.

If $\boldsymbol{W}$ is the data matrix, then $R(\boldsymbol{W})$ is a large sample $100(1-\alpha)\%$ confidence region for $\boldsymbol{\mu}$ if $P[\boldsymbol{\mu} \in R(\boldsymbol{W})] \rightarrow 1 - \alpha$ as $n \rightarrow \infty$. If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from a distribution with mean $\boldsymbol{\mu}$ and nonsingular covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then

$$R(\boldsymbol{W}) = \{\boldsymbol{\mu} | n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}\}$$

is a large sample $100(1-\alpha)\%$ confidence region for $\boldsymbol{\mu}$. This region is a hyperellipsoid centered at $\overline{\boldsymbol{x}}$. Note that the estimated covariance matrix for $\overline{\boldsymbol{x}}$ is $\boldsymbol{S}/n$ and $n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) = D^2_{\boldsymbol{\mu}}(\overline{\boldsymbol{x}}, \boldsymbol{S}/n)$. Hence $\boldsymbol{\mu}$ that are close to $\overline{\boldsymbol{x}}$ with respect to the Mahalanobis distance based on dispersion matrix $\boldsymbol{S}/n$ are in the confidence region.

Recall from Theorem 1.1e that $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T(\overline{\boldsymbol{x}} - \boldsymbol{\mu})(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{S} \boldsymbol{a}} =$
$n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) = T^2$. This fact can be used to derive large sample simultaneous confidence intervals for $\boldsymbol{a}^T \boldsymbol{\mu}$ in that separate confidence statements using different choices of $\boldsymbol{a}$ all hold simultaneously with probability

tending to $1-\alpha$. Let $\boldsymbol{x}_1, ...\boldsymbol{x}_n$ be iid with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma_x} > 0$. Then simultaneously for all $\boldsymbol{a} \neq \boldsymbol{0}$, $P(L_{\boldsymbol{a}} < \boldsymbol{a}^T\boldsymbol{\mu} < U_{\boldsymbol{a}}) \rightarrow 1-\alpha$ as $n \rightarrow \infty$ where

$$(L_{\boldsymbol{a}}, U_{\boldsymbol{a}}) = \boldsymbol{a}^T\overline{\boldsymbol{x}} \pm \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p,1-\alpha} \boldsymbol{a}^T\boldsymbol{S}\boldsymbol{a}}.$$

Simultaneous confidence intervals (CIs) can be made after collecting data and hence are useful for "data snooping." Following Johnson and Wichern (1988, p. 184-5), the p confidence intervals (CIs) for μ_i and $p(p-1)/2$ CIs for $\mu_i - \mu_k$ can be made such that they all hold simultaneously with confidence $\rightarrow 1-\alpha$. Hence if $\alpha = 0.05$, then in 100 samples, expect all $p + p(p-1)/2$ CIs to contain μ_i and $\mu_i - \mu_k$ about 95 times while about 5 times at least one of the CIs will fail to contain its parameter. The CIs for μ_i are

$$(L, U) = \overline{x}_i \pm \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p,1-\alpha}} \sqrt{\frac{S_{ii}}{n}}$$

while the CIs for $\mu_i - \mu_k$ are

$$(L, U) = \overline{x}_i - \overline{x}_k \pm \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p,1-\alpha}} \sqrt{\frac{S_{ii} - 2S_{ik} + S_{kk}}{n}}.$$

9.1.1 A diagnostic for the Hotelling's T^2 test

Now the RMVN estimator is asymptotically equivalent to a scaled DGK estimator that uses $k = 5$ concentration steps and two "reweight for efficiency" steps. Lopuhaä (1999, p. 1651-1652) shows that if (E1) holds, then the classical estimator applied to cases with $D_i(\overline{\boldsymbol{x}}, S) \leq h$ is asymptotically normal with

$$\sqrt{n}(T_{0,D} - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \kappa_p \boldsymbol{\Sigma}).$$

Here h is some fixed positive number, such as $h = \chi^2_{p,0.975}$, so this estimator is not quite the DGK estimator after one concentration step.

We conjecture that a similar result holds after concentration:

$$\sqrt{n}(T_{RMVN} - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \tau_p \boldsymbol{\Sigma})$$

for a wide variety of elliptically contoured distributions where τ_p depends on both p and the underlying distribution. Since the "test" is based on a

conjecture, it is ad hoc, and should be used as an outlier diagnostic rather than for inference.

For MVN data, simulations suggest that τ_p is close to 1. The ad hoc test that rejects H_0 if

$$T_R^2/f_{n,p} = n(T_{RMVN} - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{C}}_{RMVN}^{-1}(T_{RMVN} - \boldsymbol{\mu}_0)/f_{n,p} > \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$$

where $f_{n,p} = 1.04 + 0.12/p + (40+p)/n$ gave fair results in the simulations described later in this subsection for $n \geq 15p$ and $2 \leq p \leq 100$.

The correction factor $f_{n,p}$ was found by simulating the "robust" and classical test statistics for 100 runs, plotting the test statistics, then finding a correction factor so that the identity line passed through the data. The following R commands were used to make Figure 9.1, which shows that the plotted points of the scaled "robust" test statistic versus the classical test statistic scatter about the identity line.

```
zout <- rhotsim(n=4000,p=30)
SRHOT <- zout$rhot/(1.04 + 0.12/p + (40+p)/n)
HOT <- zout$hot
plot(SRHOT,HOT)
abline(0,1)
```

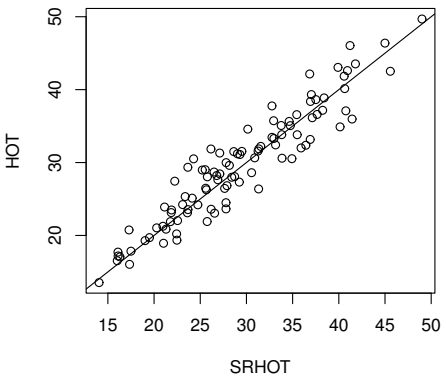

Figure 9.1: Scaled "Robust" Statistic Versus T_H^2 Statistic

For the Hotelling's T_H^2 simulation, the data is $N_p(\delta \mathbf{1}, diag(1, 2, ..., p))$ where $H_0 : \boldsymbol{\mu} = \mathbf{0}$ is being tested with 5000 runs at a nominal level of 0.05. In Table 9.1, $\delta = 0$ so H_0 is true, while hcv and rhcv are the proportion of rejections by the T_H^2 test and by the ad hoc robust test. Sample sizes are $n = 15p, 20p$ and $30p$. The robust test is not recommended for $n < 15p$ and appears to be conservative (number of rejections is less than the nominal 0.05) except when $n = 15p$ and $75 \leq p \leq 100$. See Zhang (2011).

If $\delta > 0$, then H_0 is false and the proportion of rejections estimates the power of the test. Table 9.2 shows that T_H^2 has more power than the robust test, but suggests that the power of both tests rapidly increases to one as δ increases.

Table 9.1: Hotelling simulation

p	n=15p	hcv	rhcv	n=20p	hcv	rhcv	n=30p	hcv	rhcv
10	150	0.0476	0.0300	200	0.0516	0.0304	300	0.0498	0.0286
15	225	0.0474	0.0318	300	0.0506	0.0308	450	0.0492	0.0320
20	300	0.0540	0.0368	400	0.0548	0.0314	600	0.0520	0.0354
25	375	0.0444	0.0334	500	0.0462	0.0296	750	0.0456	0.0288
30	450	0.0472	0.0324	600	0.0516	0.0358	900	0.0484	0.0342
35	525	0.0490	0.0384	700	0.0522	0.0358	1050	0.0502	0.0374
40	600	0.0534	0.0440	800	0.0486	0.0354	1200	0.0526	0.0336
45	675	0.0406	0.0390	900	0.0544	0.0390	1350	0.0512	0.0366
50	750	0.0498	0.0430	1000	0.0522	0.0394	1500	0.0512	0.0364
55	825	0.0504	0.0502	1100	0.0496	0.0392	1650	0.0510	0.0374
60	900	0.0482	0.0514	1200	0.0488	0.0404	1800	0.0474	0.0376
65	975	0.0568	0.0602	1300	0.0524	0.0414	1950	0.0548	0.0410
70	1050	0.0462	0.0530	1400	0.0558	0.0432	2100	0.0522	0.0424
75	1125	0.0474	0.0632	1500	0.0502	0.0486	2250	0.0490	0.0370
80	1200	0.0524	0.0620	1600	0.0524	0.0432	2400	0.0468	0.0356
85	1275	0.0482	0.0758	1700	0.0496	0.0456	2550	0.0520	0.0404
90	1350	0.0504	0.0746	1800	0.0484	0.0454	2700	0.0484	0.0398
95	1425	0.0524	0.0892	1900	0.0472	0.0506	2850	0.0538	0.0424
100	1500	0.0554	0.0808	2000	0.0452	0.0506	3000	0.0488	0.0392

9.2 Matched Pairs

Assume that there are $k = 2$ treatments, and both treatments are given to the same n cases or units. For example, systolic and diastolic blood pressure

Table 9.2: Hotelling power simulation

p	n	hcv	rhcv	δ	n	hcv	rhcv	δ	n	hcv	rhcv	δ
5	75	0.459	0.245	0.20	100	0.366	0.184	0.15	150	0.333	0.208	0.12
5	75	0.682	0.416	0.25	100	0.599	0.368	0.20	150	0.577	0.394	0.16
5	75	0.840	0.588	0.30	100	0.816	0.587	0.30	150	0.860	0.708	0.40
10	150	0.221	0.113	0.10	200	0.312	0.182	0.10	300	0.469	0.340	0.10
10	150	0.621	0.400	0.17	200	0.655	0.467	0.15	300	0.647	0.504	0.12
10	150	0.888	0.729	0.22	200	0.848	0.692	0.18	300	0.872	0.767	0.15
15	225	0.314	0.188	0.10	300	0.442	0.294	0.10	450	0.317	0.228	0.07
15	225	0.714	0.543	0.15	300	0.623	0.449	0.12	450	0.648	0.522	0.10
15	225	0.881	0.738	0.18	300	0.858	0.755	0.15	450	0.853	0.762	0.12
20	300	0.408	0.276	0.10	400	0.341	0.230	0.08	600	0.291	0.216	0.06
20	300	0.691	0.525	0.13	400	0.674	0.534	0.11	600	0.554	0.433	0.08
20	300	0.935	0.852	0.17	400	0.858	0.742	0.13	600	0.790	0.701	0.10
25	375	0.304	0.214	0.08	500	0.434	0.319	0.08	750	0.354	0.266	0.06
25	375	0.728	0.580	0.12	500	0.676	0.531	0.10	750	0.660	0.556	0.08
25	375	0.926	0.837	0.15	500	0.868	0.771	0.12	750	0.887	0.815	0.10
30	450	0.374	0.264	0.08	600	0.395	0.290	0.07	900	0.290	0.217	0.05
30	450	0.602	0.467	0.10	600	0.639	0.517	0.09	900	0.743	0.642	0.08
30	450	0.883	0.763	0.13	600	0.867	0.770	0.11	900	0.876	0.808	0.09

could be compared before and after the patient (case) receives blood pressure medication. Then $p = 2$. Alternatively use m correlated pairs, for example, pairs of animals from the same litter or neighboring farm fields. Then use randomization to decide whether the first member of the pair gets treatment 1 or treatment 2. Let $n_1 = n_2 = n$ and assume $n - p$ is large.

Let $\boldsymbol{y}_i = (Y_{i1}, Y_{i2}, ..., Y_{ip})^T$ denote the p measurements from the 1st treatment, and $\boldsymbol{z}_i = (Z_{i1}, Z_{i2}, ..., Z_{ip})^T$ denote the p measurements from the 2nd treatment. Let $\boldsymbol{d}_i \equiv \boldsymbol{x}_i = \boldsymbol{y}_i - \boldsymbol{z}_i$ for $i = 1, ..., n$. Assume that the $\boldsymbol{x}_i$ are iid with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma_x}$. Let $T^2 = n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu})$. Then $T^2 \xrightarrow{P} \chi^2_p$ and $pF_{p,n-p} \xrightarrow{P} \chi^2_p$. Let $P(F_{p,n} \leq F_{p,n,\delta}) = \delta$. Then the one sample Hotelling's T^2 inference is done on the differences $\boldsymbol{x}_i$ using m instead of n and using $\boldsymbol{\mu}_0 = \boldsymbol{0}$. If the p random variables are continuous, make 3 DD plots: one for the $\boldsymbol{x}_i$, one for the $\boldsymbol{y}_i$ and one for the $\boldsymbol{z}_i$ to detect outliers.

Let pval be an estimate of the pvalue. The **large sample multivariate matched pairs test** has 4 steps.
i) State the hypotheses $H_0 : \boldsymbol{\mu} = \boldsymbol{0} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{0}$.
ii) Find the test statistic $T^2 = n\overline{\boldsymbol{x}}^T \boldsymbol{S}^{-1}\overline{\boldsymbol{x}}$.
iii) Find pval =

$$P\left(T^2 < \frac{(n-1)p}{n-p} F_{p,n-p}\right) = P\left(\frac{n-p}{(n-1)p}\ T^2 < F_{p,n-p}\right).$$

iv) State whether you fail to reject H_0 or reject H_0. If you reject H_0 then conclude that $\boldsymbol{\mu} \neq \boldsymbol{0}$ while if you fail to reject H_0 conclude that the population mean $\boldsymbol{\mu} = \boldsymbol{0}$ or that there is not enough evidence to conclude that $\boldsymbol{\mu} \neq \boldsymbol{0}$. Reject H_0 if pval $< \alpha$ and fail to reject H_0 if pval $\geq \alpha$. As a benchmark for this text, use $\alpha = 0.05$ if α is not given.

A large sample $100(1-\alpha)\%$ confidence region for $\boldsymbol{\mu}$ is

$$\{\boldsymbol{\mu} |\ m(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}\},$$

and the p large sample simultaneous confidence intervals (CIs) for μ_i are

$$(L, U) = \overline{x}_i \pm \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p,1-\alpha}} \sqrt{\frac{S_{ii}}{n}}$$

where $S_{ii} = S_i^2$ is the ith diagonal element of $\boldsymbol{S}$.

9.3 Repeated Measurements

Repeated measurements = longitudinal data analysis. Take p measurements on the same unit, often the same measurement, eg blood pressure, at several time periods. The variables are $X_1, ..., X_p$ where often X_k is the measurement at the kth time period. The $E(\boldsymbol{x}) = (\mu_1, ..., \mu_p)^T = (\mu + \tau_1, ..., \mu + \tau_p)^T$. Let $y_{ij} = x_{ij} - x_{i,j+1}$ for $i = 1, ..., n$ and $j = 1, ..., p-1$. Then $\overline{\boldsymbol{y}} = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2, \overline{\boldsymbol{x}}_2 - \overline{\boldsymbol{x}}_3, ..., \overline{\boldsymbol{x}}_{p-1} - \overline{\boldsymbol{x}}_p)^T$. If $\boldsymbol{\mu}_Y = E(\boldsymbol{y}_i)$, then $\boldsymbol{\mu}_Y = \boldsymbol{0}$ is equivalent to $\mu_1 = \cdots = \mu_p$ where $E(X_k) = \mu_k$. Let $\boldsymbol{S}_y$ be the sample covariance matrix of the $\boldsymbol{y}_i$.

The **large sample repeated measurements test** has 4 steps.
i) State the hypotheses $H_0 : \boldsymbol{\mu}_y = \boldsymbol{0}$ $\quad H_1 : \boldsymbol{\mu}_y \neq \boldsymbol{0}$.
ii) Find the test statistic $T_R^2 = n\overline{\boldsymbol{y}}^T \boldsymbol{S}_y^{-1} \overline{\boldsymbol{y}}$.
iii) Find pval =

$$P\left(\frac{n-p+1}{(n-1)(p-1)}\, T_R^2 < F_{p-1,n-p+1}\right).$$

iv) State whether you fail to reject H_0 or reject H_0. If you reject H_0 then conclude that $\boldsymbol{\mu}_y \neq \boldsymbol{0}$ while if you fail to reject H_0 conclude that the population mean $\boldsymbol{\mu}_y = \boldsymbol{0}$ or that there is not enough evidence to conclude that $\boldsymbol{\mu}_y \neq \boldsymbol{0}$. Reject H_0 if pval $< \alpha$ and fail to reject H_0 if pval $\geq \alpha$. Give a nontechnical sentence, if possible.

9.4 Two Samples

Suppose there are two independent random samples $X_{1,1}, ..., X_{n_1,1}$ and $X_{1,2}, ..., X_{n_2,2}$ from populations with mean and covariance matrices $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\boldsymbol{x}_i})$ for $i = 1, 2$. Assume the $\boldsymbol{\Sigma}_{\boldsymbol{x}_i}$ are positive definite and that it is desired to test $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ where the $\boldsymbol{\mu}_i$ are $p \times 1$ vectors. To simplify large sample theory, assume $n_1 = kn_2$ for some positive real number k.

By the multivariate central limit theorem,

$$\begin{pmatrix} \sqrt{n_1}\,(\overline{X}_1 - \boldsymbol{\mu}_1) \\ \sqrt{n_2}\,(\overline{X}_2 - \boldsymbol{\mu}_2) \end{pmatrix} \xrightarrow{D} N_{2p}\left[\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x}_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{\boldsymbol{x}_2} \end{pmatrix}\right],$$

or

$$\begin{pmatrix} \sqrt{n_2}\,(\overline{X}_1 - \boldsymbol{\mu}_1) \\ \sqrt{n_2}\,(\overline{X}_2 - \boldsymbol{\mu}_2) \end{pmatrix} \xrightarrow{D} N_{2p}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}_1}}{k} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\boldsymbol{x}_2} \end{pmatrix}\right].$$

Hence

$$\sqrt{n_2}\,[(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \xrightarrow{D} N_p(\mathbf{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}_1}}{k} + \boldsymbol{\Sigma}_{\boldsymbol{x}_2}).$$

Using $n\boldsymbol{B}^{-1} = \left(\frac{\boldsymbol{B}}{n}\right)^{-1}$ and $n_2 k = n_1$, if $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, then

$$n_2(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \left(\frac{\boldsymbol{\Sigma}_{\boldsymbol{x}_1}}{k} + \boldsymbol{\Sigma}_{\boldsymbol{x}_2}\right)^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) =$$

$$(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \left(\frac{\boldsymbol{\Sigma}_{\boldsymbol{x}_1}}{n_1} + \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}_2}}{n_2}\right)^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) \xrightarrow{D} \chi^2_p.$$

Hence

$$T_0^2 = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \left(\frac{\boldsymbol{S}_1}{n_1} + \frac{\boldsymbol{S}_2}{n_2}\right)^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) \xrightarrow{D} \chi^2_p.$$

If the sequence of positive integer $d_n \to \infty$ and $Y_n \sim F_{p,d_n}$, then $Y_n \xrightarrow{D} \chi^2_p/p$. Using an F_{p,d_n} distribution instead of a χ^2_p distribution is similar to using a t_{d_n} distribution instead of a standard normal $N(0,1)$ distribution for inference. Instead of rejecting H_0 when $T_0^2 > \chi^2_{p,1-\alpha}$, reject H_0 when

$$T_0^2 > pF_{p,d_n,1-\alpha} = \frac{pF_{p,d_n,1-\alpha}}{\chi^2_{p,1-\alpha}} \chi^2_{p,1-\alpha}.$$

The term $\frac{pF_{p,d_n,1-\alpha}}{\chi^2_{p,1-\alpha}}$ can be regarded as a small sample correction factor that improves the test's performance for small samples. We will use $d_n = \min(n_1 - p, n_2 - p)$. Here $P(Y_n \leq \chi^2_{p,\alpha}) = \alpha$ if Y_n has a χ^2_p distribution, and $P(Y_n \leq F_{p,d_n,\alpha}) = \alpha$ if Y_n has an F_{p,d_n} distribution.

Let pval denote the estimated pvalue. The 4 step test is
i) State the hypotheses $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.
ii) Find the test statistic $t_0 = T_0^2/p$.
iii) Find pval $= P(t_0 < F_{p,d_n})$.
iv) State whether you fail to reject H_0 or reject H_0. If you reject H_0 then conclude that the population means are not equal while if you fail to reject

H_0 conclude that the population means are equal or that there is not enough evidence to conclude that the population means differ. Reject H_0 if pval $< \alpha$ and fail to reject H_0 if pval $\geq \alpha$. Give a nontechnical sentence if possible. As a benchmark for this text, use $\alpha = 0.05$ if α is not given.

9.5 Summary

1) The one sample Hotelling's T^2 test is used to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. The test rejects H_0 if $T_H^2 = n(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0) > \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$ where $P(Y \leq F_{p,d,\alpha}) = \alpha$ if $Y \sim F_{p,d}$.

If a multivariate location estimator T satisfies $\sqrt{n}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{D})$, then a competing test rejects H_0 if $T_C^2 = n(T-\boldsymbol{\mu}_0)^T \hat{\boldsymbol{D}}^{-1}(T-\boldsymbol{\mu}_0) > \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$ if H_0 holds and $\hat{\boldsymbol{D}}$ is a consistent estimator of $\boldsymbol{D}$. The scaled F cutoff can be used since $T_C^2 \xrightarrow{D} \chi_p^2$ if H_0 holds, and $\frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha} \to \chi^2_{p,1-\alpha}$ as $n \to \infty$.

2) Let pval be an estimate of the pvalue. As a benchmark for hypothesis testing, use $\alpha = 0.05$ if α is not given.

3) Typically use $T_C^2 = T_H^2$ in the following 4 step **one sample Hotelling's T_C^2 test**. i) State the hypotheses $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$.
ii) Find the test statistic $T_C^2 = n(T - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{D}}^{-1}(T - \boldsymbol{\mu}_0)$.
iii) Find pval =

$$P\left(\frac{n-p}{(n-1)p} \ \ T_C^2 < F_{p,n-p}\right).$$

iv) State whether you fail to reject H_0 or reject H_0. If you reject H_0 then conclude that $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ while if you fail to reject H_0 conclude that the population mean $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ or that there is not enough evidence to conclude that $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. Reject H_0 if pval $< \alpha$ and fail to reject H_0 if pval $\geq \alpha$.

4) The multivariate matched pairs test is used when there are $k = 2$ treatments applied to the same n cases with the same p variables used for each treatment. Let $\boldsymbol{y}_i$ be the p variables measured for treatment 1 and $\boldsymbol{z}_i$ be the p variables measured for treatment 2. Let $\boldsymbol{x}_i = \boldsymbol{y}_i - \boldsymbol{z}_i$. Let $\boldsymbol{\mu} = E(\boldsymbol{x}) = E(\boldsymbol{y}) - E(\boldsymbol{z})$. Want to test if $\boldsymbol{\mu} = \boldsymbol{0}$, so $E(\boldsymbol{y}) = E(\boldsymbol{z})$. The test can also be used if $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ are matched (highly dependent) in some way. For example if identical twins are in the study, $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ could be the

measurements on each twin. Let $(\overline{\boldsymbol{x}}, \boldsymbol{S}_x)$ be the sample mean and covariance matrix of the $\boldsymbol{x}_i$.

5) The **large sample multivariate matched pairs test** has 4 steps.
i) State the hypotheses $H_0 : \boldsymbol{\mu} = \boldsymbol{0}$ $\quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{0}$.
ii) Find the test statistic $T_M^2 = n\overline{\boldsymbol{x}}^T \boldsymbol{S}_x^{-1} \overline{\boldsymbol{x}}$.
iii) Find pval =

$$P\left(\frac{n-p}{(n-1)p}\ T_M^2 < F_{p,n-p}\right).$$

iv) State whether you fail to reject H_0 or reject H_0. If you reject H_0 then conclude that $\boldsymbol{\mu} \neq \boldsymbol{0}$ while if you fail to reject H_0 conclude that the population mean $\boldsymbol{\mu} = \boldsymbol{0}$ or that there is not enough evidence to conclude that $\boldsymbol{\mu} \neq \boldsymbol{0}$. Reject H_0 if pval $< \alpha$ and fail to reject H_0 if pval $\geq \alpha$. Give a nontechnical sentence if possible.

6) Repeated measurements = longitudinal data analysis. Take p measurements on the same unit, often the same measurement, eg blood pressure, at several time periods. The variables are X_1, ..., X_p where often X_k is the measurement at the kth time period. The $E(\boldsymbol{x}) = (\mu_1, ..., \mu_p)^T = (\mu + \tau_1, ..., \mu + \tau_p)^T$. Let $y_{ij} = x_{ij} - x_{i,j+1}$ for $i = 1, ..., n$ and $j = 1, ..., p-1$. Then $\overline{\boldsymbol{y}} = (\overline{x}_1 - \overline{x}_2, \overline{x}_2 - \overline{x}_3, ..., \overline{x}_{p-1} - \overline{x}_p)^T$. If $\boldsymbol{\mu}_Y = E(\boldsymbol{y}_i)$, then $\boldsymbol{\mu}_Y = \boldsymbol{0}$ is equivalent to $\mu_1 = \cdots = \mu_p$ where $E(X_k) = \mu_k$. Let $\boldsymbol{S}_y$ be the sample covariance matrix of the $\boldsymbol{y}_i$.

7) The **large sample repeated measurements test** has 4 steps.
i) State the hypotheses $H_0 : \boldsymbol{\mu}_y = \boldsymbol{0}$ $\quad H_1 : \boldsymbol{\mu}_y \neq \boldsymbol{0}$.
ii) Find the test statistic $T_R^2 = n\overline{\boldsymbol{y}}^T \boldsymbol{S}_y^{-1} \overline{\boldsymbol{y}}$.
iii) Find pval =

$$P\left(\frac{n-p+1}{(n-1)(p-1)}\ T_R^2 < F_{p-1,n-p+1}\right).$$

iv) State whether you fail to reject H_0 or reject H_0. If you reject H_0 then conclude that $\boldsymbol{\mu}_y \neq \boldsymbol{0}$ while if you fail to reject H_0 conclude that the population mean $\boldsymbol{\mu}_y = \boldsymbol{0}$ or that there is not enough evidence to conclude that $\boldsymbol{\mu}_y \neq \boldsymbol{0}$. Reject H_0 if pval $< \alpha$ and fail to reject H_0 if pval $\geq \alpha$. Give a nontechnical sentence, if possible.

8) The F tables give left tail area and the pval is a right tail area. Table 15.5 gives $F_{k,d,0.95}$. If $\alpha = 0.05$ and $\dfrac{n-p}{(n-1)p}\ T_C^2 < F_{k,d,0.95}$, then fail to reject

H_0. If $\frac{n-p}{(n-1)p}\ T_C^2 \geq F_{k,d,0.95}$ then reject H_0.

a) For the one sample Hotelling's T_C^2 test, and the matched pairs T_M^2 test, $k = p$ and $d = n - p$.

b) For the repeated measures T_R^2 test, $k = p - 1$ and $d = n - p + 1$.

9) If $n > 10p$, the tests in 89), 91) and 93) are robust to nonnormality. For the one sample Hotelling's T_C^2 test and the repeated measurements test, make a DD plot. For the multivariate matched pairs test, make a DD plot of the $\boldsymbol{x}_i$, of the $\boldsymbol{y}_i$ and of the $\boldsymbol{z}_i$.

10) Suppose there are two independent random samples $X_{1,1}, ..., X_{n_1,1}$ and $X_{1,2}, ..., X_{n_2,2}$ from populations with mean and covariance matrices $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\boldsymbol{x}_i})$ for $i = 1, 2$ where the $\boldsymbol{\mu}_i$ are $p \times 1$ vectors. Let $d_n = \min(n_1 - p, n_2 - p)$. The **large sample two sample Hotelling's T_0^2 test** is a 4 step test:
i) State the hypotheses $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.
ii) Find the test statistic $t_0 = T_0^2/p$.
iii) Find pval $= P(t_0 < F_{p,d_n})$.
iv) State whether you fail to reject H_0 or reject H_0. If you reject H_0 then conclude that the population means are not equal while if you fail to reject H_0 conclude that the population means are equal or that there is not enough evidence to conclude that the population means differ. Reject H_0 if pval $< \alpha$ and fail to reject H_0 if pval $\geq \alpha$. Give a nontechnical sentence if possible.

11) Tests for covariance matrices are very nonrobust to nonnormality. Let a plot of x versus y have x on the horizontal axis and y on the vertical axis. A good diagnostic is to use the DD plot. So a diagnostic for $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_0$ is to plot $D_i(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i(\overline{\boldsymbol{x}}, \boldsymbol{\Sigma}_0)$ for $i = 1, ..., n$. If $n > 10p$ and H_0 is true, then the plotted points in the DD plot should cluster tightly about the identity line.

12) A test for sphericity is a test of $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = d\boldsymbol{I}_p$ for some unknown constant $d > 0$. As a diagnostic, make a "DD plot" of $D_i^2(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i^2(\overline{\boldsymbol{x}}, \boldsymbol{I}_p)$. If $n > 10p$ and H_0 is true, then the plotted points in the "DD plot" should cluster tightly about the line through the origin with slope d.

13) Now suppose there are k samples, and want to test $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}_1} = \cdots = \boldsymbol{\Sigma}_{\boldsymbol{x}_k}$, that is, all k populations have the same covariance matrix. As a diagnostic, make a DD plot of $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ versus $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_{pool})$ for $j = 1, ..., k$ and $i = 1, ..., n_i$.

9.6 Complements

The *mpack* function `rhotsim` is useful for simulating the robust diagnostic for the one sample Hotelling's T^2 test. See Zhang (2011) for more simulations.

Willems, Pison, Rousseeuw, and Van Aelst (2002) use similar reasoning to present a diagnostic based on the FMCD estimator.

Yao (1965) suggests a more complicated denominator degrees of freedom than $d_n = \min(n_1 - p, n_2 - p)$ for the two sample Hotelling's T^2 test. Good (2012, p. 55-57) provides randomization tests as competitors for the two sample Hotelling's T^2 test.

9.7 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

R/Splus Problems

Warning: Use the command *source("G:/mpack.txt")* **to download the programs. See Preface or Section 15.2.** Typing the name of the `mpack` function, eg *ddplot*, will display the code for the function. Use the `args` command, eg *args(ddplot)*, to display the needed arguments for the function.

9.1*. Use the *R* commands in Subsection 1.1.1 to make a plot similar to Figure 9.1.

9.2. Conjecture:

$$\sqrt{n}(T_{RMVN} - \boldsymbol{\mu}) \xrightarrow{D} N_p(\mathbf{0}, \tau_p \boldsymbol{\Sigma})$$

for a wide variety of elliptically contoured distributions where τ_p depends on both p and the underlying distribution. The following "test" is based on a conjecture, and should be used as an outlier diagnostic rather than for inference. The ad hoc "test" that rejects H_0 if

$$T_R^2/f_{n,p} = n(T_{RMVN} - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{C}}_{RMVN}^{-1}(T_{RMVN} - \boldsymbol{\mu}_0)/f_{n,p} > \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$$

where $f_{n,p} = 1.04 + 0.12/p + (40 + p)/n$. The simulations use $n = 150$ and $p = 10$.

a) The *R* commands for this part use simulated data is $\boldsymbol{x}_i \sim N_p(\mathbf{0}, diag(1, 2, ..., p))$ where $H_0 : \boldsymbol{\mu} = \mathbf{0}$ is being tested with 5000 runs at a nominal level of 0.05. So H_0 is true, and hcv and rhcv are the proportion of rejections by the T_H^2 test and by the ad hoc robust test. Want hcv and rhcv near 0.05. THIS SIMULATION WILL TAKE ABOUT 5 MINUTES. Record hcv and rhcv. Were hcv and rhcv near 0.05?

b) The *R* commands for this part use simulated data is $\boldsymbol{x}_i \sim N_p(\delta\mathbf{1}, diag(1, 2, ..., p))$ where $H_0 : \boldsymbol{\mu} = \mathbf{0}$ is being tested with 5000 runs at a nominal level of 0.05. In the simulation, $\delta = 0.2$, so H_0 is false, and hcv and rhcv are the proportion of rejections by the T_H^2 test and by the ad hoc robust test. Want hcv and rhcv near 1 so that the power is high. Paste the output into *Word*. THIS SIMULATION WILL TAKE ABOUT 5 MINUTES. Record hcv and rhcv. Were hcv and rhcv near 1?

Chapter 10

MANOVA

10.1 Introduction

Definition 10.1. The **response variables** are the variables that you want to predict. The **predictor variables** are the variables used to predict the response variables.

Notation. The **MANOVA model** $\boldsymbol{y}_i = \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and d predictor variables $X_1, X_2, ..., X_d$. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, ..., x_{id}, Y_{i1}, ..., Y_{im})$. If a constant $x_{i1} = 1$ is in the model, then x_{i1} could be omitted from the case.

For the multivariate analysis of variance (MANOVA) model, the predictors are not quantitative variables, so the predictors are indicator variables. Sometimes the trivial predictor $\mathbf{1}$ is also in the model. The multivariate regression model of Chapter 12 has at least one quantitative variable.

In matrix form, the MANOVA model is $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$, and the data matrix $\boldsymbol{W} = [\boldsymbol{X} \;\; \boldsymbol{Y}]$. The $n \times m$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,m} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n,1} & Y_{n,2} & \dots & Y_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \dots & \boldsymbol{Y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix}.$$

The $n \times d$ matrix $\boldsymbol{X}$ is not necessarily of full rank d, and

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,d} \\ x_{2,1} & x_{2,2} & \dots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,d} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_d \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

where $\boldsymbol{v}_1 = \boldsymbol{1}$.

The $d \times m$ matrix

$$\boldsymbol{B} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \dots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \dots & \beta_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{d,1} & \beta_{d,2} & \dots & \beta_{d,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \dots & \boldsymbol{\beta}_m \end{bmatrix}.$$

The $n \times m$ matrix

$$\boldsymbol{E} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \dots & \epsilon_{1,m} \\ \epsilon_{2,1} & \epsilon_{2,2} & \dots & \epsilon_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \dots & \epsilon_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \dots & \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix}.$$

Warning: The $\boldsymbol{e}_i$ are error vectors, not orthonormal eigenvectors.

Definition 10.2. Models in which a single response variable Y is quantitative, but all of the predictor variables are qualitative are called *analysis of variance* (ANOVA) models, *experimental design* models or *design of experiments* (DOE) models. Each combination of the levels of the predictors gives a different distribution for Y, and there are p different distributions or treatments. A predictor variable W is often called a factor and a factor level a_i is one of the categories W can take. In an ANOVA model,

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,d}\beta_d + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{10.1}$$

for $i = 1, \dots, n$. In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{10.2}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}$ is an $n \times d$ matrix of predictors, $\boldsymbol{\beta}$ is a $d \times 1$ vector of unknown coefficients, $\boldsymbol{e}$ is an $n \times 1$ vector

of unknown errors, and $d \geq p$. Equivalently,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,d} \\ x_{2,1} & x_{2,2} & \dots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,d} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{10.3}$$

The e_i are iid with zero mean and variance σ^2, and a linear model estimator such as least squares is used to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2.

Each response variable in a MANOVA model follows an ANOVA model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$. Hence the errors corresponding to the jth response are uncorrelated with variance $\sigma_j^2 = \sigma_{jj}$. Notice that the **same design matrix** $\boldsymbol{X}$ of predictors is used for each of the m models, but the jth response variable vector $\boldsymbol{Y}_j$, coefficient vector $\boldsymbol{\beta}_j$ and error vector $\boldsymbol{e}_j$ change and thus depend on j. Hence for a one way MANOVA model, each response variable follows a one way ANOVA model, while for a two way MANOVA model, each response variable follows a two way ANOVA model for $j = 1, ..., m$.

Once the ANOVA model is fixed, eg a one way ANOVA model, the design matrix $\boldsymbol{X}$ depends on the parameterization of the ANOVA model. The fitted values and residuals are the same for each parameterization, but the interpretation of the parameters depend on the parameterization.

Now consider the ith case $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T)$ which corresponds to the ith row of $\boldsymbol{Z}$ and the ith row of $\boldsymbol{X}$. Then

$$\begin{bmatrix} Y_{i1} = \beta_{11}x_{i1} + \cdots + \beta_{d1}x_{id} + \epsilon_{i1} = \boldsymbol{x}_i^T\boldsymbol{\beta}_1 + \epsilon_{i1} \\ Y_{i2} = \beta_{12}x_{i1} + \cdots + \beta_{d2}x_{id} + \epsilon_{i2} = \boldsymbol{x}_i^T\boldsymbol{\beta}_2 + \epsilon_{i2} \\ \vdots \\ Y_{im} = \beta_{1m}x_{i1} + \cdots + \beta_{dm}x_{id} + \epsilon_{im} = \boldsymbol{x}_i^T\boldsymbol{\beta}_m + \epsilon_{im} \end{bmatrix}$$

or $\boldsymbol{y}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i$ where

$$E(\boldsymbol{y}_i) = \boldsymbol{B}^T\boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{x}_i^T\boldsymbol{\beta}_1 \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_m \end{bmatrix}.$$

The notation $\boldsymbol{y}_i|\boldsymbol{x}_i$ and $E(\boldsymbol{y}_i|\boldsymbol{x}_i)$ is more accurate, but usually the conditioning is suppressed. Taking $E(\boldsymbol{y}_i|\boldsymbol{x}_i)$ to be a constant, $\boldsymbol{y}_i$ and $\boldsymbol{\epsilon}_i$ have

the same covariance matrix. In the MANOVA model, this covariance matrix $\boldsymbol{\Sigma_\epsilon}$ does not depend on i. Observations from different cases are uncorrelated (often independent), but the m errors for the m different response variables for the *same case* are correlated.

Definition 10.3. The **MANOVA model** $\boldsymbol{y}_k = \boldsymbol{B}^T\boldsymbol{x}_k + \boldsymbol{\epsilon}_k$ for $k = 1, ..., n$ is written in matrix form as $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma_\epsilon} = ((\sigma_{ij}))$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$. Then $\boldsymbol{B}$ and $\boldsymbol{\Sigma_\epsilon}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$. Considering the kth row of $\boldsymbol{Z}, \boldsymbol{X}$ and $\boldsymbol{E}$ shows that $\boldsymbol{y}_k^T = \boldsymbol{x}_k^T\boldsymbol{B} + \boldsymbol{\epsilon}_k^T$.

10.2 One Way ANOVA

Before describing the one way MANOVA model, it is useful to give a brief description on the one way ANOVA model.

Definition 10.4. A **lurking variable** is not one of the variables in the study, but may affect the relationships among the variables in the study. A **unit** is the experimental material assigned **treatments**, which are the conditions the investigator wants to study. The unit is *experimental* if it was randomly assigned to a treatment, and the unit is *observational* if it was not randomly assigned to a treatment.

Definition 10.5. In an **experiment**, the investigators use **randomization** to assign treatments to units. To assign p treatments to $n = n_1 + \cdots + n_p$ experimental units, draw a random permutation of $\{1, ..., n\}$. Assign the first n_1 units treatment 1, the next n_2 units treatment 2, ..., and the final n_p units treatment p.

Randomization allows one to do valid inference such as F tests of hypotheses and confidence intervals. Randomization also washes out the effects of lurking variables and makes the p treatment groups similar except for the treatment. The effects of lurking variables are present in observational studies defined in Definition 10.6.

Definition 10.6. In an **observational study**, investigators simply observe the response, and the treatment groups need to be p random samples

from p populations (the levels) for valid inference.

Example 10.1. Consider using randomization to assign the following nine people (units) to three treatment groups.

Carroll, Collin, Crawford, Halverson, Lawes,
Stach, Wayman, Wenslow, Xumong

Balanced designs have the group sizes the same: $n_i \equiv h = n/p$. Label the units alphabetically so Carroll gets 1, ..., Xumong gets 9. The *R/Splus* function `sample` can be used to draw a random permutation. Then the first 3 numbers in the permutation correspond to group 1, the next 3 to group 2 and the final 3 to group 3. Using the output shown below, gives the following 3 groups.

group 1: Stach, Wayman, Xumong
group 2: Lawes, Carroll, Halverson
group 3: Collin, Wenslow, Crawford

```
> sample(9)
[1] 6 7 9 5 1 4 2 8 3
```

Often there is a table or computer file of units and related measurements, and it is desired to add the unit's group to the end of the table. The *mpack* function `rand` reports a random permutation and the quantity groups[i] = treatment group for the ith person on the list. Since persons 6, 7 and 9 are in group 1, groups[7] = 1. Since Carroll is person 1 and is in group 2, groups[1] = 2, et cetera.

```
> rand(9,3)
$perm
[1] 6 7 9 5 1 4 2 8 3

$groups
[1] 2 3 3 2 2 1 1 3 1
```

Definition 10.7. Replication means that for each treatment, the n_i response variables $Y_{i,1}, ..., Y_{i,n_i}$ are approximately iid random variables.

Example 10.2. a) If ten students work two types of paper mazes three times each, then there are 60 measurements that are not replicates. Each

student should work the six mazes in random order since speed increases with practice. For the ith student, let Z_{i1} be the average time to complete the three mazes of type 1, let Z_{i2} be the average time for mazes of type 2 and let $D_i = Z_{i1} - Z_{i2}$. Then $D_1, ..., D_{10}$ are replicates.

b) Cobb (1998, p. 126) states that a student wanted to know if the shapes of sponge cells depends on the color (green or white). He measured hundreds of cells from one white sponge and hundreds of cells from one green sponge. There were only two units so $n_1 = 1$ and $n_2 = 1$. The student should have used a sample of n_1 green sponges and a sample of n_2 white sponges to get more replicates.

c) Replication depends on the goals of the study. Box, Hunter and Hunter (2005, p. 215-219) describes an experiment where the investigator times how long it takes him to bike up a hill. Since the investigator is only interested in his performance, each run up a hill is a replicate (the time for the ith run is a sample from all possible runs up the hill by the investigator). If the interest had been on the effect of eight treatment levels on student bicyclists, then replication would need $n = n_1 + \cdots + n_8$ student volunteers where n_i ride their bike up the hill under the conditions of treatment i.

Definition 10.8. Let $f_Z(z)$ be the pdf of Z. Then the family of pdfs $f_Y(y) = f_Z(y - \mu)$ indexed by the *location parameter* μ, $-\infty < \mu < \infty$, is the *location family* for the random variable $Y = \mu + Z$ with *standard pdf* $f_Z(z)$.

Definition 10.9. A *one way fixed effects ANOVA model* has a single qualitative predictor variable W with p categories a_1, ..., a_p. There are p different distributions for Y, one for each category a_i. The distribution of

$$Y|(W = a_i) \sim f_Z(y - \mu_i)$$

where the location family has second moments. Hence all p distributions come from the same location family with different location parameter μ_i and the same variance σ^2.

Definition 10.10. The *one way fixed effects normal ANOVA model* is the special case where

$$Y|(W = a_i) \sim N(\mu_i, \sigma^2).$$

Example 10.3. The pooled 2 sample t–test is a special case of a one way ANOVA model with $p = 2$. For example, one population could be ACT

scores for men and the second population ACT scores for women. Then $W =$ *gender* and $Y =$ *score.*

Notation. It is convenient to relabel the response variable $Y_1, ..., Y_n$ as the vector $\boldsymbol{Y} = (Y_{11}, ..., Y_{1,n_1}, Y_{21}, ..., Y_{2,n_2}, ..., Y_{p1}, ..., Y_{p,n_p})^T$ where the Y_{ij} are independent and $Y_{i1}, ..., Y_{i,n_i}$ are iid. Here $j = 1, ..., n_i$ where n_i is the number of cases from the ith level where $i = 1, ..., p$. Thus $n_1 + \cdots + n_p = n$. Similarly use double subscripts on the errors. Then there will be many equivalent parameterizations of the one way fixed effects ANOVA model.

Definition 10.11. The *cell means model* is the parameterization of the one way fixed effects ANOVA model such that

$$Y_{ij} = \mu_i + e_{ij}$$

where Y_{ij} is the value of the response variable for the jth trial of the ith factor level. The μ_i are the unknown means and $E(Y_{ij}) = \mu_i$. The e_{ij} are iid from the location family with pdf $f_Z(z)$ and unknown variance $\sigma^2 = \text{VAR}(Y_{ij}) = \text{VAR}(e_{ij})$. For the normal cell means model, the e_{ij} are iid $N(0, \sigma^2)$ for $i = 1, ..., p$ and $j = 1, ..., n_i$.

The cell means model is a linear model (without intercept) of the form $\boldsymbol{Y} = \boldsymbol{X}_c \boldsymbol{\beta}_c + \boldsymbol{e} =$

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1,n_1} \\ Y_{21} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} e_{11} \\ \vdots \\ e_{1,n_1} \\ e_{21} \\ \vdots \\ e_{2,n_2} \\ \vdots \\ e_{p,1} \\ \vdots \\ e_{p,n_p} \end{bmatrix}. \tag{10.4}$$

Notation. Let $Y_{i0} = \sum_{j=1}^{n_i} Y_{ij}$ and let

$$\hat{\mu}_i = \overline{Y}_{i0} = Y_{i0}/n_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}. \tag{10.5}$$

Hence the "dot notation" means sum over the subscript corresponding to the 0, eg j. Similarly, $Y_{00} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij}$ is the sum of all of the Y_{ij}.

Notice that the indicator variables used in the cell means model (10.4) are $v_{h,k} = 1$ if the hth case has $W = a_k$, and $v_{h_k} = 0$, otherwise, for $k = 1, ..., p$ and $h = 1, ..., n$. So Y_{ij} has $v_{h,k} = 1$ only if $i = k$ and $j = 1, ..., n_i$. Here $\boldsymbol{v}_k$ is the kth column of $\boldsymbol{X}_c$. The model can use p indicator variables for the factor instead of $p - 1$ indicator variables because the model does not contain an intercept. Also notice that

$$E(\boldsymbol{Y}) = \boldsymbol{X}_c \boldsymbol{\beta}_c = (\mu_1, ..., \mu_1, \mu_2, ..., \mu_2, ..., \mu_p, ..., \mu_p)^T,$$

$(\boldsymbol{X}_c^T \boldsymbol{X}_c) = \text{diag}(\text{n}_1, ..., \text{n}_\text{p})$ and $\boldsymbol{X}_c^T \boldsymbol{Y} = (Y_{10}, ..., Y_{10}, Y_{20}, ..., Y_{20}, ..., Y_{p0}, ..., Y_{p0})^T$. Hence $(\boldsymbol{X}_c^T \boldsymbol{X}_c)^{-1} = \text{diag}(1/\text{n}_1, ..., 1/\text{n}_\text{p})$ and the OLS estimator

$$\hat{\boldsymbol{\beta}}_c = (\boldsymbol{X}_c^T \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c^T \boldsymbol{Y} = (\overline{Y}_{10}, ..., \overline{Y}_{p0})^T = (\hat{\mu}_1, ..., \hat{\mu}_p)^T.$$

Thus $\hat{\boldsymbol{Y}} = \boldsymbol{X}_c \hat{\boldsymbol{\beta}}_c = (\overline{Y}_{10}, ..., \overline{Y}_{10}, ..., \overline{Y}_{p0}, ..., \overline{Y}_{p0})^T$. Hence the ijth fitted value is

$$\hat{Y}_{ij} = \overline{Y}_{i0} = \hat{\mu}_i \tag{10.6}$$

and the ijth residual is

$$r_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \hat{\mu}_i. \tag{10.7}$$

Since the cell means model is a linear model, there is an associated response plot and residual plot. However, many of the interpretations of the OLS quantities for ANOVA models differ from the interpretations for multiple linear regression (MLR) models. First, for MLR models, the conditional distribution $Y|\boldsymbol{x}$ makes sense even if $\boldsymbol{x}$ is not one of the observed $\boldsymbol{x}_i$ provided that $\boldsymbol{x}$ is not far from the $\boldsymbol{x}_i$. This fact makes MLR very powerful. For MLR, at least one of the variables in $\boldsymbol{x}$ is a continuous predictor. For the one way fixed effects ANOVA model, the p distributions $Y|\boldsymbol{x}_i$ make sense where $\boldsymbol{x}_i^T$ is a row of $\boldsymbol{X}_c$.

Also, the OLS MLR ANOVA F test for the cell means model tests $H_0 : \boldsymbol{\beta} = 0 \equiv H_0 : \mu_1 = \cdots = \mu_p = 0$, while the one way fixed effects ANOVA F test given after Definition 10.15 tests $H_0 : \mu_1 = \cdots = \mu_p$.

Definition 10.12. Consider the one way fixed effects ANOVA model. The *response plot* is a plot of $\hat{Y}_{ij} \equiv \hat{\mu}_i$ versus Y_{ij} and the *residual plot* is a

plot of $\hat{Y}_{ij} \equiv \hat{\mu}_i$ versus r_{ij}. Add the identity line to the response plot and $r = 0$ line to the residual plot as visual aids.

The points in the response plot scatter about the identity line and the points in the residual plot scatter about the $r = 0$ line, but the scatter need not be in an evenly populated band. A *dot plot* of $Z_1, ..., Z_m$ consists of an axis and m points each corresponding to the value of Z_i. The response plot consists of p dot plots, one for each value of $\hat{\mu}_i$. The dot plot corresponding to $\hat{\mu}_i$ is the dot plot of $Y_{i1}, ..., Y_{i,n_i}$. The p dot plots should have roughly the same amount of spread, and each $\hat{\mu}_i$ corresponds to level a_i. If a new level a_f corresponding to $\boldsymbol{x}_f$ was of interest, hopefully the points in the response plot corresponding to a_f would form a dot plot at $\hat{\mu}_f$ similar in spread to the other dot plots, but it may not be possible to predict the value of $\hat{\mu}_f$. Similarly, the residual plot consists of p dot plots, and the plot corresponding to $\hat{\mu}_i$ is the dot plot of $r_{i1}, ..., r_{i,n_i}$.

Assume that each $n_i \geq 10$. Under the assumption that the Y_{ij} are from the same location scale family with different parameters μ_i, each of the p dot plots should have roughly the same shape and spread. This assumption is easier to judge with the residual plot. If the response plot looks like the residual plot, then a horizontal line fits the p dot plots about as well as the identity line, and there is not much difference in the μ_i. If the identity line is clearly superior to any horizontal line, then at least some of the means differ.

Definition 10.13. An **outlier** corresponds to a case that is far from the bulk of the data. Look for a large vertical distance of the plotted point from the identity line or the $r = 0$ line.

Rule of thumb 10.1. Mentally add 2 lines parallel to the identity line and 2 lines parallel to the $r = 0$ line that cover most of the cases. Then a case is an outlier if it is well beyond these 2 lines.

This rule often fails for large outliers since often the identity line goes through or near a large outlier so its residual is near zero. A response that is far from the bulk of the data in the response plot is a "large outlier" (large in magnitude). Look for a large gap between the bulk of the data and the large outlier.

Suppose there is a dot plot of n_j cases corresponding to level a_j that is far from the bulk of the data. This dot plot is probably not a cluster of "bad outliers" if $n_j \geq 4$ and $n \leq 50$. If $n_j = 1$, such a case may be a large outlier.

Rule of thumb 10.2. Often an outlier is very good, but more often an outlier is due to a measurement error and is very bad.

The assumption of the Y_{ij} coming from the same location scale family with different location parameters μ_i and the same constant variance σ^2 is a big assumption and often does not hold. Another way to check this assumption is to make a box plot of the Y_{ij} for each i. The box in the box plot corresponds to the lower, middle and upper quartiles of the Y_{ij}. The middle quartile is just the sample median of the data m_{ij}: at least half of the $Y_{ij} \geq m_{ij}$ and at least half of the $Y_{ij} \leq m_{ij}$. The p boxes should be roughly the same length and the median should occur in roughly the same position (eg in the center) of each box. The "whiskers" in each plot should also be roughly similar. Histograms for each of the p samples could also be made. All of the histograms should look similar in shape.

Example 10.4. Kuehl (1994, p. 128) gives data for counts of hermit crabs on 25 different transects in each of six different coastline habitats. Let Z be the count. Then the response variable $Y = \log_{10}(Z+1/6)$. Although the counts Z varied greatly, each habitat had several counts of 0 and often there were several counts of 1, 2 or 3. Hence Y is not a continuous variable. The cell means model was fit with $n_i = 25$ for $i = 1, ..., 6$. Each of the six habitats was a level. Figure 10.1a and b shows the response plot and residual plot. There are 6 dot plots in each plot. Because several of the smallest values in each plot are identical, it does not always look like the identity line is passing through the six sample means $\overline{Y}_{i0}$ for $i = 1, ..., 6$. In particular, examine the dot plot for the smallest mean (look at the 25 dots furthest to the left that fall on the vertical line FIT ≈ 0.36). Random noise (jitter) has been added to the response and residuals in Figure 10.1c and d. Now it is easier to compare the six dot plots. They seem to have roughly the same spread.

The plots contain a great deal of information. The response plot can be used to explain the model, check that the sample from each population (treatment) has roughly the same shape and spread, and to see which populations have similar means. Since the response plot closely resembles the residual plot in Figure 10.1, there may not be much difference in the six populations. Linearity seems reasonable since the samples scatter about the identity line. The residual plot makes the comparison of "similar shape" and "spread" easier.

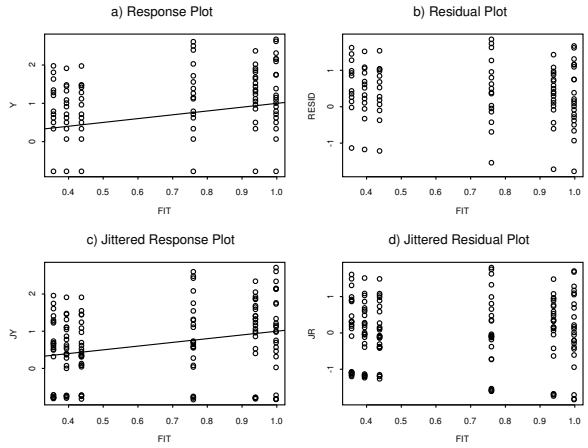

Figure 10.1: Plots for One Way ANOVA Model for Crab Data

Definition 10.14. a) The *total sum of squares*

$$SSTO = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{00})^2.$$

b) The *treatment sum of squares*

$$SSTR = \sum_{i=1}^{p} n_i(\overline{Y}_{i0} - \overline{Y}_{00})^2.$$

c) The residual sum of squares or *error sum of squares*

$$SSE = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{io})^2.$$

Definition 10.15. Associated with each SS in Definition 10.14 is a degrees of freedom (df) and a mean square $= SS/df$. For SSTO, df $= n-1$ and $MSTO = SSTO/(n-1)$. For SSTR, df $= p-1$ and $MSTR = SSTR/(p-1)$. For SSE, $df = n-p$ and $MSE = SSE/(n-p)$.

Let $S_i^2 = \sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i0})^2/(n_i - 1)$ be the sample variance of the ith group. Then the MSE is a weighted sum of the S_i^2:

$$\hat{\sigma}^2 = MSE = \frac{1}{n-p}\sum_{i=1}^{p}\sum_{j=1}^{n_i} r_{ij}^2 = \frac{1}{n-p}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i0})^2 =$$

$$\frac{1}{n-p}\sum_{i=1}^{p}(n_i-1)S_i^2 = S_{pool}^2$$

where S_{pool}^2 is known as the pooled variance estimator.

The ANOVA table is the same as that for MLR, except that SSTR replaces the regression sum of squares. The MSE is again an estimator of σ^2. The ANOVA F test tests whether all p means μ_i are equal. Shown below is an ANOVA table given in symbols. Sometimes "Treatment" is replaced by "Between treatments," "Between Groups," "Model," "Factor" or "Groups." Sometimes "Error" is replaced by "Residual," or "Within Groups." Sometimes "p-value" is replaced by "P", "$Pr(> F)$" or "PR > F."

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Treatment	p-1	SSTR	MSTR	Fo=MSTR/MSE	for Ho:
Error	n-p	SSE	MSE		$\mu_1 = \cdots = \mu_p$

Note that the software output uses pvalue for pval, an estimate of the pvalue.

Be able to perform the 4 step fixed effects one way ANOVA F test of hypotheses:
i) State the hypotheses Ho: $\mu_1 = \mu_2 = \cdots = \mu_p$ and Ha: not Ho.
ii) Find the test statistic $F_o = MSTR/MSE$ or obtain it from output.
iii) Find the pval from output or use the F–table: pval =

$$P(F_{p-1,n-p} > F_o).$$

iv) State whether you reject Ho or fail to reject Ho. If the pval $< \delta$, reject Ho and conclude that the mean response depends on the level of the factor. Otherwise fail to reject Ho and conclude that the mean response does not depend on the level of the factor. Give a nontechnical sentence.

Rule of thumb 10.3. If

$$\max(S_1, ..., S_p) \leq 2\min(S_1, ..., S_p),$$

then the one way ANOVA F test results will be approximately correct if the response and residual plots suggest that the remaining one way ANOVA model assumptions are reasonable. See Moore (2000, p. 512). If all of the

$n_i \geq 5$, replace the standard deviations by the ranges of the dot plots when examining the response and residual plots.

Remark 10.1. If the units are a representative sample of some population of interest, then randomization of units into groups makes the assumption that $Y_{i1}, ..., Y_{i,n_i}$ are iid hold to a useful approximation for large sample theory. Random sampling from populations also induces the iid assumption. Linearity can be checked with the response plot, and similar shape and spread of the location families can be checked with both the response and residual plots. Also check that outliers are not present. If the p dot plots in the response plot are approximately symmetric, then the sample sizes n_i can be smaller than if the dot plots are skewed.

Remark 10.2. When the assumption that the p groups come from the same location family with finite variance σ^2 is violated, the one way ANOVA F test may not make much sense because unequal means may not imply the superiority of one category over another. Suppose Y is the time in minutes until relief from a headache and that $Y_{1j} \sim N(60, 1)$ while $Y_{2j} \sim N(65, \sigma^2)$. If $\sigma^2 = 1$, then the type 1 medicine gives headache relief 5 minutes faster, on average, and is superior, all other things being equal. But if $\sigma^2 = 100$, then many patients taking medicine 2 experience much faster pain relief than those taking medicine 1, and many experience much longer time until pain relief. In this situation, predictor variables that would identify which medicine is faster for a given patient would be very useful.

```
fat1  fat2  fat3  fat4 One way Anova for Fat1 Fat2 Fat3 Fat4
64    78    75    55   Source     DF SS     MS     F     P
72    91    93    66   treatment  3  1636.5 545.5  5.41  0.0069
68    97    78    49   error      20 2018.0 100.9
77    82    71    64
56    85    63    70
95    77    76    68
```

Example 10.5. The output above represents grams of fat (minus 100 grams) absorbed by doughnuts using 4 types of fat. See Snedecor and Cochran (1967, p. 259). Let μ_i denote the mean amount of fat$i$ absorbed by doughnuts, $i = 1, 2, 3$ and 4. a) Find $\hat{\mu}_1$. b) Perform a 4 step ANOVA F test.

Solution: a) $\hat{\boldsymbol{\beta}}_{1c} = \hat{\mu}_1 = \overline{Y}_{10} = Y_{10}/n_1 = \sum_{j=1}^{n_1} Y_{1j}/n_1 =$
$(64 + 72 + 68 + 77 + 56 + 95)/6 = 432/6 = 72.$

b) i) $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ H_a: not H_0
ii) F = 5.41
iii) pval = 0.0069
iv) Reject H_0, the mean amount of fat absorbed by doughnuts depends on the type of fat.

Definition 10.16. A **contrast** $C = \sum_{i=1}^{p} k_i \mu_i$ where $\sum_{i=1}^{p} k_i = 0$. The estimated contrast is $\hat{C} = \sum_{i=1}^{p} k_i \overline{Y}_{i0}$.

If the null hypothesis of the fixed effects one way ANOVA test is not true, then not all of the means μ_i are equal. Researchers will often have hypotheses, before examining the data, that they desire to test. Often such a hypothesis can be put in the form of a contrast. For example, the contrast $C = \mu_i - \mu_j$ is used to compare the means of the ith and jth groups while the contrast $\mu_1 - (\mu_2 + \cdots + \mu_p)/(p-1)$ is used to compare the last $p-1$ groups with the 1st group. This contrast is useful when the 1st group corresponds to a standard or control treatment while the remaining groups correspond to new treatments.

Assume that the normal cell means model is a useful approximation to the data. Then the $\overline{Y}_{i0} \sim N(\mu_i, \sigma^2/n_i)$ are independent, and

$$\hat{C} = \sum_{i=1}^{p} k_i \overline{Y}_{i0} \sim N\left(C, \sigma^2 \sum_{i=1}^{p} \frac{k_i^2}{n_i}\right).$$

Hence the standard error

$$SE(\hat{C}) = \sqrt{MSE \sum_{i=1}^{p} \frac{k_i^2}{n_i}}.$$

The degrees of freedom is equal to the MSE degrees of freedom $= n - p$.

Consider a family of null hypotheses for contrasts $\{Ho : \sum_{i=1}^{p} k_i \mu_i = 0$ where $\sum_{i=1}^{p} k_i = 0$ and the k_i may satisfy other constraints$\}$. Let δ_S denote the probability of a type I error for a single test from the family where a type I error is a false rejection. The **family level** δ_F is an upper bound on the (usually unknown) size δ_T. Know how to interpret $\delta_F \approx \delta_T =$
P(of making at least one type I error among the family of contrasts).

Two important families of contrasts are the family of all possible contrasts and the family of pairwise differences $C_{ij} = \mu_i - \mu_j$ where $i \neq j$. The

Scheffé multiple comparisons procedure has a δ_F for the family of all possible contrasts while the Tukey multiple comparisons procedure has a δ_F for the family of all $\binom{p}{2}$ pairwise contrasts.

To interpret output for multiple comparisons procedures, the underlined means or blocks of letters besides groups of means indicate that the group of means are not significantly different.

Example 10.6. The output below uses data from SAS Institute (1985, p. 126-129). The mean nitrogen content of clover depends on the strain of clover (3dok1, 3dok5, 3dok7, compos, 3dok4, 3dok13). Recall that means μ_1 and μ_2 are significantly different if you can conclude that $\mu_1 \neq \mu_2$ while μ_1 and μ_2 are not significantly different if there is not enough evidence to conclude that $\mu_1 \neq \mu_2$ (perhaps because the means are approximately equal or perhaps because the sample sizes are not large enough).

Notice that the strain of clover 3dok1 appears to have the highest mean nitrogen content. There are 4 pairs of means that are not significantly different. The letter B suggests 3dok5 and 3dok7, the letter C suggests 3dok7 and compos, the letter D suggests compos and 3dok4, while the letter E suggests 3dok4 and 3dok13 are not significantly different.

```
Means with the same letter are not significantly different.
     Waller Grouping          Mean      N    strain
                   A        28.820      5    3dok1

                   B        23.980      5    3dok5
                   B
              C    B        19.920      5    3dok7
              C
              C    D        18.700      5    compos
                   D
              E    D        14.640      5    3dok4
              E
              E             13.260      5    3dok13
```

Definition 10.17. Graphical Anova for the one way model uses the residuals as a reference set instead of a t, F or normal distribution. The scaled treatment deviations or scaled effect $c(\overline{Y}_{i0} - \overline{Y}_{00}) = c(\hat{\mu}_i - \overline{Y}_{00})$ are scaled to have the same variability as the residuals. A dot plot of the scaled deviations is placed above the dot plot of the residuals. Assume that

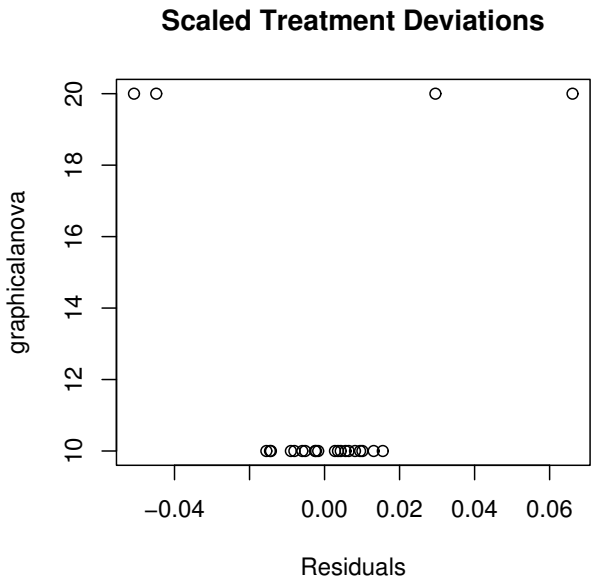

Figure 10.2: Graphical Anova

$n_i \equiv h = n/p$ for $i = 1, ..., p$. For small $n \leq 40$, suppose the distance between two scaled deviations (A and B, say) is greater than the range of the residuals $= \max(r_{ij}) - \min(r_{ij})$. Then declare μ_A and μ_B to be significantly different. If the distance is less than the range, do not declare μ_A and μ_B to be significantly different. Scaled deviations that lie outside the range of the residuals are significant (so significantly different from the overall mean).

For $n \geq 100$, let $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}$ be the order statistics of the residuals. Then instead of the range, use $r_{(\lceil 0.975n \rceil)} - r_{(\lceil 0.025n \rceil)}$ as the distance where $\lceil x \rceil$ is the smallest integer $\geq x$, eg $\lceil 7.7 \rceil = 8$. So effects outside of the interval $(r_{(\lceil 0.025n \rceil)}, r_{(\lceil 0.975n \rceil)})$ are significant. See Box, Hunter and Hunter (2005, p. 136, 166). A derivation of the scaling constant $c = \sqrt{(n-p)/(p-1)}$ is given in Section 10.5.

```
ganova(x,y)
sdev    0.02955502  0.06611268 -0.05080048 -0.04486722
Treatments "A"      "B"        "C"          "D"
```

Example 10.7. Cobb (1998, p. 160) describes a one way ANOVA design used to study the amount of calcium in the blood. For many animals, the

body's ability to use calcium depends on the level of certain hormones in the blood. The response was 1/(level of plasma calcium). The four groups were A: Female controls, B: Male controls, C: Females given hormone and D: Males given hormone. There were 10 birds of each gender, and five from each gender were given the hormone. The output above uses the `mpack` function *ganova* to produce Figure 10.2.

In Figure 10.2, the top dot plot has the scaled treatment deviations. From left to right, these correspond to C, D, A and B since the output shows that the deviation corresponding to C is the smallest with value -0.050. Since the deviations corresponding to C and D are much closer than the range of the residuals, the C and D effects yielded similar mean response values. A and B appear to be significantly different from C and D. The distance between the scaled A and B treatment deviations is about the same as the distance between the smallest and largest residuals, so there is only marginal evidence that the A and B effects are significantly different.

Since all 4 scaled deviations lie outside of the range of the residuals, all effects A, B, C and D appear to be significant.

10.2.1 Response Transformations for ANOVA Models

A model for an experimental design is $Y_i = E(Y_i) + e_i$ for $i = 1, ..., n$ where the error $e_i = Y_i - E(Y_i)$ and $E(Y_i) \equiv E(Y_i|\boldsymbol{x}_i)$ is the expected value of the response Y_i for a given vector of predictors $\boldsymbol{x}_i$. Many models can be fit with least squares (OLS or LS) and are linear models of the form

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$$

for $i = 1, \ldots, n$. Often $x_{i,1} \equiv 1$ for all i. In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ design matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. If the fitted values are $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$, then $Y_i = \hat{Y}_i + r_i$ where the residuals $r_i = Y_i - \hat{Y}_i$.

The applicability of an experimental design model can be expanded by allowing response transformations. An important class of *response transformation models* adds an additional unknown transformation parameter λ_o,

such that

$$Y_i = t_{\lambda_o}(Z_i) \equiv Z_i^{(\lambda_o)} = E(Y_i) + e_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i.$$

If λ_o was known, then $Y_i = t_{\lambda_o}(Z_i)$ would follow the linear model for the experimental design.

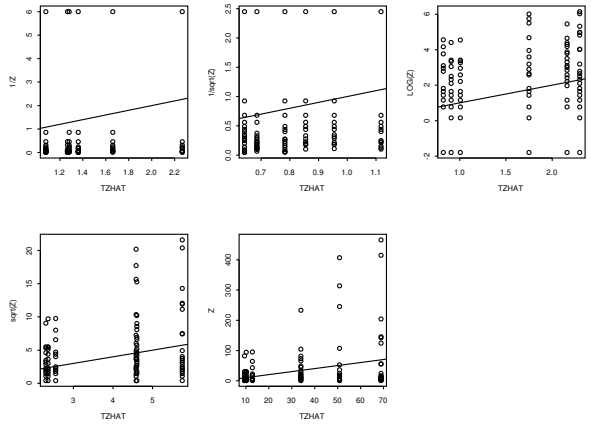

Figure 10.3: Transformation Plots for Crab Data

Definition 10.20. Assume that **all** of the values of the "response" Z_i are **positive**. A *power transformation* has the form $Y = t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$ where $\lambda \in \Lambda_L = \{-1, -1/2, 0, 1/2, 1\}$.

A graphical method for response transformations computes the fitted values $\hat{W}_i$ from the experimental design model using $W_i = t_\lambda(Z_i)$ as the "response." Then a plot of the $\hat{W}$ versus W is made for each of the five values of $\lambda \in \Lambda_L$. For many experimental design models, the plotted points follow the identity line in a (roughly) evenly populated band if the experimental design model is reasonable for $(\hat{W}, W)$. An exception is the one way ANOVA model where there will be p dot plots of roughly the same shape and spread that scatter about the identity line. If more than one value of $\lambda \in \Lambda_L$ gives a linear plot, consult subject matter experts and use the simplest or most reasonable transformation. Note that Λ_L has 5 models, and the graphical method selects the model with the best response plot. After selecting the transformation, the usual checks should be made. In particular, the transformation plot is also the response plot, and a residual plot should be made.

Definition 10.21. A *transformation plot* is a plot of $(\hat{W}, W)$ with the identity line added as a visual aid.

In the following example, the plots show $t_\lambda(Z)$ on the vertical axis. The label "TZHAT" of the horizontal axis are the fitted values that result from using $t_\lambda(Z)$ as the "response" in the software.

For one way ANOVA models with $n_i \equiv m \geq 5$, look for a transformation plot that satisfies the following conditions. i) The p dot plots scatter about the identity line with similar shape and spread. ii) Dot plots with more skew are worse than dot plots with less skew or dot plots that are approximately symmetric. iii) Spread that increases or decreases with TZHAT is bad.

Example 10.4, continued. Following Kuehl (1994, p. 128), let C be the count of crabs and let the "response" $Z = C + 1/6$. Figure 10.3 shows the five *transformation plots*. The transformation $\log(Z)$ results in dot plots that have roughly the same shape and spread. The transformations $1/Z$ and $1/\sqrt{Z}$ do not handle the 0 counts well, and the dot plots fail to cover the identity line. The transformations $\sqrt{Z}$ and Z have variance that increases with the mean.

Remark 10.4. The graphical method for response transformations can be used for design models that are linear models, not just one way ANOVA models. The method is nearly identical to that of Chapter 12, but Λ_L only has 5 values. The **log rule** states that if all of the $Z_i > 0$ and if $\frac{\max(Z_i)}{\min(Z_i)} \geq 10$, then the response transformation $Y = \log(Z)$ will often work.

10.3 One Way MANOVA

Using double subscripts will be useful for describing the one way MANOVA model. Suppose there independent random samples from p different populations (treatments), or $n = \sum_{i=1}^{p} n_i$ and n_i cases are randomly assigned to p treatment groups. Then the group sample sizes are n_i for $i = 1, ..., p$. Assume that m response variables $\boldsymbol{y}_{ij} = (Y_{ij1}, ..., Y_{ijm})^T$ are measured for the ith treatment. Hence $i = 1, ..., p$ and $j = 1, ..., n_i$. The Y_{ijk} follow different one way ANOVA models for $k = 1, ..., m$. Assume $E(\boldsymbol{y}_{ij}) = \boldsymbol{\mu}_i$ and $\text{Cov}(\boldsymbol{y}_{ij}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Hence the p treatments have different mean vectors $\boldsymbol{\mu}_i$, but common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. (This assumption can be relaxed for $p = 2$

with the appropriate 2 sample Hotelling's T^2 test.)

The one way MANOVA is used to test $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_p$. Often $\boldsymbol{\mu}_i = \boldsymbol{\mu} + \boldsymbol{\tau}_i$, so H_0 becomes $H_0 : \boldsymbol{\tau}_1 = \cdots = \boldsymbol{\tau}_p$. If $m = 1$, the one way MANOVA model is the one way ANOVA model. MANOVA is useful since it takes into account the correlations between the m response variables. Performing m ANOVA tests fails to account for these correlations, but can be a useful diagnostic. The Hotelling's T^2 test that uses a common covariance matrix is a special case of the one way MANOVA model with $m = 2$.

Let $\boldsymbol{\mu}_i = \boldsymbol{\mu} + \boldsymbol{\tau}_i$ where $\sum_{i=1}^{p} n_i \boldsymbol{\tau}_i = 0$. The jth case from the ith population or treatment group is $\boldsymbol{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_j + \boldsymbol{e}_{ij}$ where $\boldsymbol{e}_{ij}$ is an error vector, $i = 1, ..., p$ and $j = 1, ..., n_i$. Let $\overline{\boldsymbol{y}} = \hat{\boldsymbol{\mu}} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \boldsymbol{y}_{ij}/n$ be the overall mean. Let $\overline{\boldsymbol{y}}_i = \sum_{j=1}^{n_i} \boldsymbol{y}_{ij}/n_i$ so $\hat{\boldsymbol{\tau}}_i = \overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}}$. Let the residual $\hat{\boldsymbol{\epsilon}}_{ij} = \boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i = \boldsymbol{y}_{ij} - \hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\tau}}_i$. Then $\boldsymbol{y}_{ij} = \overline{\boldsymbol{y}} + (\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}}) + (\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i) = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\tau}}_i + \hat{\boldsymbol{\epsilon}}_{ij}$.

Let $\boldsymbol{S}_i$ be the sample covariance matrix corresponding to the ith treatment group. Then the within sum of squares and cross products matrix is $\boldsymbol{W} = (n_1 - 1)\boldsymbol{S}_1 + \cdots + (n_p - 1)\boldsymbol{S}_p = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i)(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i)^T$. Then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \boldsymbol{W}/(n - p)$. The treatment or between sum of squares and cross products matrix is $\boldsymbol{B} = \sum_{i=1}^{p} n_i (\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}})(\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}})^T$. The total corrected (for the mean) sum of squares and cross products matrix is $\boldsymbol{T} = \boldsymbol{B} + \boldsymbol{W} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}})(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}})^T$. Note that $\boldsymbol{T}/(n - 1)$ is the usual sample covariance matrix if it is assumed that all n of the $\boldsymbol{y}_{ij}$ are iid so that the $\boldsymbol{\mu}_i \equiv \boldsymbol{\mu}$ for $i = 1, ..., p$.

The one way MANOVA model is $\boldsymbol{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{\epsilon}_{ij}$ where the $\boldsymbol{\epsilon}_{ij}$ are iid with $E(\boldsymbol{\epsilon}_{ij}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_{ij}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. The MANOVA table is shown below.

Summary One Way MANOVA Table

Source	matrix	df
Treatment or Between	$\boldsymbol{B}$	$p - 1$
Residual or Error or Within	$\boldsymbol{W}$	$n - p$
Total (corrected)	$\boldsymbol{T}$	$n - 1$

If all n of the $\boldsymbol{y}_{ij}$ are iid with $E(\boldsymbol{y}_{ij}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{y}_{ij}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, it can be shown that $\boldsymbol{A}/df \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ where $\boldsymbol{A} = \boldsymbol{W}, \boldsymbol{B}$ or $\boldsymbol{T}$ and df is the corresponding degrees of freedom. Let t_0 be the test statistic. Although Pillai's trace is robust to nonnormality, often Wilk's lambda is used. Wilk's lambda

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{B} + \boldsymbol{W}|} = \frac{|\boldsymbol{W}|}{|\boldsymbol{T}|}$$

is good if the iid $\boldsymbol{\epsilon}_{ij} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Then $t_o = -[n-1-(m+p)/2]\log(\Lambda)$ and pval $= P(\chi^2_{m(p-1)} > t_0)$. Hence reject H_0 if $t_0 > \chi^2_{m(p-1)}(1-\alpha)$. See Johnson and Wichern (1988, p. 238).

The four steps of the one way MANOVA test follow.
i) State the hypotheses $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ and H_1 : not H_0.
ii) Get t_0 from output.
iii) Get pval from output.
iv) State whether you reject H_0 or fail to reject H_0. If pval $\leq \alpha$, reject H_0 and conclude that not all of the p treatment means are equal. If pval $> \alpha$, fail to reject H_0 and conclude that all p treatment means are equal or that there is not enough evidence to conclude that not all of the p treatment means are equal. As a textbook convention, use $\alpha = 0.05$ if α is not given.

Rule of thumb 10.4. In the one way MANOVA model, $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ is a one way ANOVA model for $j = 1, ..., m$. To check the one way MANOVA model, make the m response and residual plots corresponding to the m one way ANOVA models. Make a DD plot of the n residual vectors. Response transformations can be done as in Section 10.2.1. If the n_i are large, make p DD plots of the $\boldsymbol{y}_{ij}$ for $i = 1, ..., p$. Also if the n_i are large, make p plots of $D_{ij}(\overline{\boldsymbol{y}}_i, \boldsymbol{S}_i)$ versus $D_{ij}(\overline{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ to check that the common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ is an adequate assumption. The plotted points in these p plots should cluster tightly about the identity line if n_i is large and the covariance matrix of the ith treatment group is approximately $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$.

10.4 Summary

1) The **fixed effects one way ANOVA** model has one qualitative explanatory variable called a **factor** and a quantitative response variable Y_{ij}. The factor variable has p levels, $E(Y_{ij}) = \mu_i$ and $V(Y_{ij}) = \sigma^2$ for $i = 1, ..., p$ and $j = 1, ..., n_i$. **Experimental units** are randomly assigned to the treatment levels.

2) Let $n = n_1 + \cdots + n_p$. In an **experiment**, the investigators use randomization to randomly assign n units to treatments. Draw a random permutation of $\{1, ..., n\}$. Assign the first n_1 units to treatment 1, the next n_2 units to treatment 2, ..., and the final n_p units to treatment p. Use $n_i \equiv h = n/p$ if possible. Randomization washes out the effect of lurking variables.

3) The 4 step fixed effects one way ANOVA F test has steps
i) Ho: $\mu_1 = \mu_2 = \cdots = \mu_p$ and Ha: not Ho.
ii) Fo = MSTR/MSE is usually given by output.
iii) The pval = P($F_{p-1,n-p} > Fo$) is usually given by output.
iv) If the pval $< \delta$, reject Ho and conclude that the mean response depends on the level of the factor. Otherwise fail to reject Ho and conclude that the mean response does not depend on the level of the factor. Give a nontechnical sentence.

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Treatment	p-1	SSTR	MSTR	Fo=MSTR/MSE	for Ho:
Error	n-p	SSE	MSE		$\mu_1 = \cdots = \mu_p$

4) Shown is an ANOVA table given in symbols. Sometimes "Treatment" is replaced by "Between treatments," "Between Groups," "Model," "Factor" or "Groups." Sometimes "Error" is replaced by "Residual," or "Within Groups." Sometimes "p-value" is replaced by "P", "$Pr(> F)$" or "PR > F."

5) A *dot plot* of $Z_1, ..., Z_h$ consists of an axis and h points each corresponding to the value of Z_i. The *response plot* is a plot of $\hat{Y}$ versus Y. For the one way ANOVA model, the response plot is a plot of $\hat{Y}_{ij} = \hat{\mu}_i$ versus Y_{ij}. Often the identity line with unit slope and zero intercept is added as a visual aid. Vertical deviations from the identity line are the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \hat{\mu}_i$. The plot will consist of p dot plots that scatter about the identity line with similar shape and spread if the fixed effects one way ANOVA model is appropriate. The ith dot plot is a dot plot of $Y_{i,1}, ..., Y_{i,n_i}$. Assume that each $n_i \geq 10$. If the response plot looks like the residual plot, then a horizontal line fits the p dot plots about as well as the identity line, and there is not much difference in the μ_i. If the identity line is clearly superior to any horizontal line, then at least some of the means differ.

6) The *residual plot* is a plot of $\hat{Y}$ versus residual $r = Y - \hat{Y}$. The plot will consist of p dot plots that scatter about the $r = 0$ line with similar shape and spread if the fixed effects one way ANOVA model is appropriate. The ith dot plot is a dot plot of $r_{i,1}, ..., r_{i,n_i}$. Assume that each $n_i \geq 10$. Under the assumption that the Y_{ij} are from the same location scale family with

different parameters μ_i, each of the p dot plots should have roughly the same shape and spread. This assumption is easier to judge with the residual plot than with the response plot.

7) Rule of thumb: If $\max(S_1, ..., S_p) \leq 2\min(S_1, ..., S_p)$, then the one way ANOVA F test results will be approximately correct if the response and residual plots suggest that the remaining one way ANOVA model assumptions are reasonable.

8) The **cell means model** for the fixed effects one way ANOVA is $Y_{ij} = \mu_i + e_{ij}$ where Y_{ij} is the value of the response variable for the jth trial of the ith factor level for $i = 1, ..., p$ and $j = 1, ..., n_i$. The μ_i are the unknown means and $E(Y_{ij}) = \mu_i$. The e_{ij} are iid from the location family with pdf $f_Z(z)$, zero mean and unknown variance $\sigma^2 = V(Y_{ij}) = V(e_{ij})$. For the normal cell means model, the e_{ij} are iid $N(0, \sigma^2)$. The estimator $\hat{\mu}_i = \overline{Y}_{i0} = \sum_{j=1}^{n_i} Y_{ij}/n_i = \hat{Y}_{ij}$. The ith residual is $r_{ij} = Y_{ij} - \overline{Y}_{i0}$, and $\overline{Y}_{00}$ is the sample mean of all of the Y_{ij} and $n = \sum_{i=1}^{p} n_i$. The total sum of squares SSTO $= \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{00})^2$, the treatment sum of squares SSTR $= \sum_{i=1}^{p} n_i(\overline{Y}_{i0} - \overline{Y}_{00})^2$, and the error sum of squares SSE $= \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i0})^2$. The MSE is an estimator of σ^2. In the ANOVA table, SSTO, SSTR and SSE have $n-1$, $p-1$ and $n-p$ degrees of freedom.

9) Let $Y_{i0} = \sum_{j=1}^{n_i} Y_{ij}$ and let

$$\hat{\mu}_i = \overline{Y}_{i0} = Y_{i0}/n_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}.$$

Hence the "dot notation" means sum over the subscript corresponding to the 0, eg j. Similarly, $Y_{00} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij}$ is the sum of all of the Y_{ij}. Be able to find $\hat{\mu}_i$ from data.

10) The applicability of a DOE (design of experiments) model can be expanded by allowing response transformations. An important class of *response transformation models* is

$$Y = t_{\lambda_o}(Z) = E(Y) + e = \boldsymbol{x}^T\boldsymbol{\beta} + e$$

where the subscripts (eg Y_{ij}) have been suppressed. If λ_o was known, then $Y = t_{\lambda_o}(Z)$ would follow the DOE model. Assume that **all** of the values of the "response" Z are **positive**. A **power transformation** has the form $Y = t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$ where $\lambda \in \Lambda_L = \{-1, -1/2, 0, 1/2, 1\}$.

11) A graphical method for response transformations computes the fitted values $\hat{W}$ from the DOE model using $W = t_\lambda(Z)$ as the "response" for each of the five values of $\lambda \in \Lambda_L$. Let $\hat{T} = \hat{W}$ = TZHAT and plot TZHAT vs $t_\lambda(Z)$ for $\lambda \in \{-1, -1/2, 0, 1/2, 1\}$. These plots are called **transformation plots**. The residual or error degrees of freedom used to compute the MSE should not be too small. Choose the transformation $Y = t_{\lambda^*}(Z)$ that has the best plot. Consider the one way ANOVA model with $n_i > 4$ for $i = 1, ..., p$. i) The dot plots should spread about the identity line with similar shape and spread. ii) Dot plots that are approximately symmetric are better than skewed dot plots. iii) Spread that increases or decreases with TZHAT (the shape of the plotted points is similar to a right or left opening megaphone) is bad.

12) The transformation plot for the selected transformation is also the response plot for that model (eg for the model that uses $Y = \log(Z)$ as the response). Make all of the usual checks on the DOE model (residual and response plots) after selecting the response transformation.

13) The **log rule** says try $Y = \log(Z)$ if $\max(Z)/\min(Z) > 10$ where $Z > 0$ and the subscripts have been suppressed (so $Z \equiv Z_{ij}$ for the one way ANOVA model).

14) **Graphical Anova** for the **one way ANOVA** model makes a dot plot of scaled treatment deviations (effects) above a dot plot of the residuals. For small $n \leq 40$, suppose the distance between two scaled deviations (A and B, say) is greater than the range of the residuals $= \max(r_{ij}) - \min(r_{ij})$. Then declare μ_A and μ_B to be significantly different. If the distance is less than the range, do not declare μ_A and μ_B to be significantly different. Assume the $n_i \equiv m$ for $i = 1, ..., p$. Then the ith scaled deviation is $c(\overline{Y}_{i0} - \overline{Y}_{00}) = c\hat{\alpha}_i = \tilde{\alpha}_i$ where $c = \sqrt{df_e/df_{treat}} = \sqrt{\dfrac{n-p}{p-1}}$.

15) Assume that the residual degrees of freedom are large enough for testing. Then the response and residual plots contain much information. Linearity and constant variance may be reasonable if the p dot plots have roughly the same shape and spread, and the dot plots scatter about the identity line. The p dot plots of the residuals should have similar shape and spread, and the dot plots scatter about the $r = 0$ line. It is easier to check linearity with the response plot and constant variance with the residual plot. Curvature is often easier to see in a residual plot, but the response plot can be used to check whether the curvature is monotone or not. The response

plot is more effective for determining whether the signal to noise ratio is strong or weak, and for detecting outliers or influential cases.

16) In a MANOVA model, $\boldsymbol{y}_k = \boldsymbol{B}^T\boldsymbol{x}_k + \boldsymbol{\epsilon}_k$ for $k = 1, ..., n$ is written in matrix form as $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma_\epsilon} = ((\sigma_{ij}))$ for $k = 1, ..., n$. Each response variable in a MANOVA model follows an ANOVA model $\boldsymbol{Y}_j = \boldsymbol{X\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$.

17) The **one way MANOVA** model is as above where $\boldsymbol{Y}_j = \boldsymbol{X\beta}_j + \boldsymbol{e}_j$ is a one way ANOVA model for $j = 1, ..., m$. Check the model by making m response and residual plots and a DD plot of the residuals $\hat{\boldsymbol{\epsilon}}_i$.

18) The four steps of the one way MANOVA test follow.
i) State the hypotheses $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ and H_1 : not H_0.
ii) Get t_0 from output.
iii) Get pval from output.
iv) State whether you reject H_0 or fail to reject H_0. If pval $\leq \alpha$, reject H_0 and conclude that not all of the p means are equal. If pval $> \alpha$, fail to reject H_0 and conclude that all p means are equal or that there is not enough evidence to conclude that not all of the p means are equal. As a textbook convention, use $\alpha = 0.05$ if α is not given.

10.5 Summary

1) The **multivariate linear model** $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $X_1, X_2, ..., X_p$. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$. If a constant $x_{i1} = 1$ is in the model, then x_{i1} could be omitted from the case. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma_\epsilon} = ((\sigma_{ij}))$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$. Then $\boldsymbol{B}$ and $\boldsymbol{\Sigma_\epsilon}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$.

The data matrix $\boldsymbol{W} = [\boldsymbol{X} \;\; \boldsymbol{Y}]$ except usually the first column $\boldsymbol{1}$ of $\boldsymbol{X}$ is omitted if $X_1 = 1$. The $n \times m$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,m} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n,1} & Y_{n,2} & \dots & Y_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \dots & \boldsymbol{Y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix}.$$

The $n \times p$ matrix

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

where often $\boldsymbol{v}_1 = \mathbf{1}$.

The $p \times m$ matrix

$$\boldsymbol{B} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \dots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \dots & \beta_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p,1} & \beta_{p,2} & \dots & \beta_{p,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \dots & \boldsymbol{\beta}_m \end{bmatrix}.$$

The $n \times m$ matrix

$$\boldsymbol{E} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \dots & \epsilon_{1,m} \\ \epsilon_{2,1} & \epsilon_{2,2} & \dots & \epsilon_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \dots & \epsilon_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \dots & \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix}.$$

Warning: The $\boldsymbol{e}_i$ are error vectors, not orthonormal eigenvectors.

2) The univariate linear model is $Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i = \boldsymbol{\beta}^T\boldsymbol{x}_i + e_i$ for $i = 1, \dots, n$. In matrix notation, these n equations become $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, where $\boldsymbol{Y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors.

3) Each response variable in a multivariate linear model follows a univariate linear model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \mathbf{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$.

4) The one way MANOVA model is a generalization of the Hotelling's T^2 test from 2 groups to $p \geq 2$ groups, assumed to have different means but a common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Want to test $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$. This model is a multivariate linear model so there are m response variables $Y_1, ..., Y_m$ measured for each group. Each Y_i follows a one way ANOVA model for $i = 1, ..., m$.

5) For the one way MANOVA model, make a DD plot of the residuals $\hat{\boldsymbol{\epsilon}}_i$ where $i = 1, ..., n$. Use the plot to check whether the $\boldsymbol{\epsilon}_i$ follow a multivariate

normal distribution or some other elliptically contoured distribution. Want $n > 10p$.

6) For the one way MANOVA model, write the data as Y_{ijk} where $i = 1, ..., p$ and $j = 1, ..., n_i$. So k corresponds to the kth variable Y_k for $k = 1, ..., m$. Then $\hat{Y}_{ijk} = \hat{\mu}_{ik} = \overline{Y}_{i0k}$ for $i = 1, ..., p$. So for the kth variable, mean $\mu_{1k}, ..., \mu_{pk}$ are of interest. The residuals are $r_{ijk} = Y_{ijk} - \hat{Y}_{ijk}$. For each variable Y_k make a response plot of $\overline{Y}_{i0k}$ versus Y_{ijk} and a residual plot of $\overline{Y}_{i0k}$ versus r_{ijk}. Both plots will consist of p dot plots of n_k cases located at the $\overline{Y}_{i0k}$. The dot plots should follow the identity line in the response plot and the horizontal $r = 0$ line in the residual plot for each of the m response variables $Y_1, ..., Y_m$. For each variable Y_k, let R_{ik} be the range of the ith dot plot. If each $n_i \geq 5$, want $\max(R_{1k}, ..., R_{pk}) \leq 2 \min(R_{1k}, ..., R_{pk})$. The one way MANOVA model may be reasonable if the m response and residual plots satisfy the above graphical checks.

7) The four steps of the one way MANOVA test follow.
i) State the hypotheses $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ and H_1 : not H_0.
ii) Get t_0 from output.
iii) Get pval from output.
iv) State whether you reject H_0 or fail to reject H_0. If pval $\leq \alpha$, reject H_0 and conclude that not all of the p treatment means are equal. If pval $> \alpha$, fail to reject H_0 and conclude that all p treatment means are equal or that there is not enough evidence to conclude that not all of the p treatment means are equal. Give a nontechnical sentence as the conclusion, if possible.

8) The one way MANOVA test assumes that $\boldsymbol{\Sigma}_{\boldsymbol{x}_1} = \cdots = \boldsymbol{\Sigma}_{\boldsymbol{x}_p}$, but has some resistance to this assumption. See point 6).

9) Know how to use randomization to assign units to treatment groups with the *R/Splus* function `sample` that is used to draw a random permutation of $\{1, 2, ..., n\}$. If the units are $a_1, ..., a_9$ and the `sample(9)` command gives 6 7 9 5 1 4 2 8 3, then a_6, a_7 and a_9 are assigned treatment 1, a_5, a_1 and a_4 are assigned treatment 2, and a_2, a_8 and a_3 are assigned treatment 3.

10.6 Complements

Four good tests on the design and analysis of experiments (ANOVA) are Box, Hunter and Hunter (2005), Cobb (1998), Kuehl (1994) and Ledolter and Swersey (2007). Also see Olive (2010, ch. 5-9). Section 10.2 followed Olive (2010, ch. 5) closely.

All of the parameterizations of the one way fixed effects ANOVA model yield the same predicted values, residuals and ANOVA F test, but the interpretations of the parameters differ. The cell means model is a linear model (without intercept) of the form $\boldsymbol{Y} = \boldsymbol{X}_c\boldsymbol{\beta}_c + \boldsymbol{e} =$ that can be fit using OLS. The OLS MLR output gives the correct fitted values and residuals but an incorrect ANOVA table. An equivalent linear model (with intercept) with correct OLS MLR ANOVA table as well as residuals and fitted values can be formed by replacing any column of the cell means model by a column of ones **1**. Removing the last column of the cell means model and making the first column **1** gives the model $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1}x_{p-1} + e$ given in matrix form by (10.8).

It can be shown that the OLS estimators corresponding to (10.8) are $\hat{\beta}_0 = \overline{Y}_{p0} = \hat{\mu}_p$, and $\hat{\beta}_i = \overline{Y}_{i0} - \overline{Y}_{p0} = \hat{\mu}_i - \hat{\mu}_p$ for $i = 1, ..., p-1$. The cell means model has $\hat{\beta}_i = \hat{\mu}_i = \overline{Y}_{i0}$.

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1,n_1} \\ Y_{21} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} e_{11} \\ \vdots \\ e_{1,n_1} \\ e_{21} \\ \vdots \\ e_{2,n_2} \\ \vdots \\ e_{p,1} \\ \vdots \\ e_{p,n_p} \end{bmatrix}. \quad (10.8)$$

Graphical Anova uses scaled treatment effects = scaled treatment deviations $\tilde{d}_i = cd_i = c(\overline{Y}_{i0} - \overline{Y}_{00})$ for $i = 1, ..., p$. Following Box, Hunter and Hunter (2005, p. 166), suppose $n_i \equiv m = n/p$ for $i = 1, ..., n$. If Ho $\mu_1 = \cdots = \mu_p$ is true, want the sample variance of the scaled deviations to be approximately equal to the sample variance of the residuals. So want

$$1 \approx \frac{\frac{1}{p}\sum_{i=1}^{p} c^2 d_i^2}{\frac{1}{n}\sum_{i=1}^{n} r_i^2} = F_0 = \frac{MSTR}{MSE} = \frac{SSTR/(p-1)}{SSE/(n-p)} = \frac{\sum_{i=1}^{p} m d_i^2/(p-1)}{\sum_{i=1}^{n} r_i^2/(n-p)}$$

since $SSTR = \sum_{i=1}^{p} m(\overline{Y}_{i0} - \overline{Y}_{00})^2 = \sum_{i=1}^{p} md_i^2$. So

$$F_0 = \frac{\sum_{i=1}^{p} c^2 \frac{n}{p} d_i^2}{\sum_{i=1}^{n} r_i^2} = \frac{\sum_{i=1}^{p} \frac{m(n-p)}{p-1} d_i^2}{\sum_{i=1}^{n} r_i^2}.$$

Equating numerators gives

$$c^2 = \frac{mp}{n} \quad \frac{(n-p)}{(p-1)} = \frac{(n-p)}{(p-1)}$$

since $mp/n = 1$. Thus $c = \sqrt{(n-p)/(p-1)}$.

For Graphical Anova, see Box, Hunter and Hunter (2005, p. 136, 150, 164, 166) and Hoaglin, Mosteller, and Tukey (1991). The R package `granova`, available from (http://streaming.stat.iastate.edu/CRAN/) and authored by R.M. Pruzek and J.E. Helmreich, may be useful.

The *modified power transformation family*

$$Y_i = t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda}$$

for $\lambda \neq 0$ and $t_0(Z_i) = \log(Z_i)$ for $\lambda = 0$ where $\lambda \in \Lambda_L$.

Box and Cox (1964) give a numerical method for selecting the response transformation for the modified power transformations. Although the method gives a point estimator $\hat{\lambda}_o$, often an interval of "reasonable values" is generated (either graphically or using a profile likelihood to make a confidence interval), and $\hat{\lambda} \in \Lambda_L$ is used if it is also in the interval.

There are several reasons to use a coarse grid Λ_L of powers. First, several of the powers correspond to simple transformations such as the log, square root, and reciprocal. These powers are easier to interpret than $\lambda = .28$, for example. Secondly, if the estimator $\hat{\lambda}_n$ can only take values in Λ_L, then sometimes $\hat{\lambda}_n$ will converge in probability to $\lambda^* \in \Lambda_L$. Thirdly, Tukey (1957) showed that neighboring modified power transformations are often very similar, so restricting the possible powers to a coarse grid is reasonable.

The graphical method for response transformations is due to Olive (2004, 2010: ch. 5). A variant of the method would plot the residual plot or both the response and the residual plot for each of the five values of λ. Residual plots are also useful, but they do not distinguish between nonlinear monotone relationships and nonmonotone relationships. See Fox (1991, p. 55). Alternative methods are given by Cook and Olive (2001) and Box, Hunter and Hunter (2005, p. 321).

A **randomization test** for the one way ANOVA model has H_0: *the different treatments have no effect*. This null hypothesis is also true if all p pdfs $Y|(W = a_i) \sim f_Z(y - \mu)$ are the same. An impractical randomization test uses all $M = \frac{n!}{n_1! \cdots n_p!}$ ways of assigning n_i of the Y_{ij} to treatment i for $i = 1, ..., p$. Let F_0 be the usual F statistic. The F statistic is computed for each of the M permutations and H_0 is rejected if the proportion of the M F statistics that are larger than F_0 is less than δ. The distribution of the M F statistics is approximately $F_{p-1,n-p}$ for large n when H_0 is true. The power of the randomization test is also similar to that of the usual F test. See Hoeffding (1952). These results suggest that the usual F test is semiparametric: the pvalue is approximately correct if n is large and if all p pdfs $Y|(W = a_i) \sim f_Z(y - \mu)$ are the same.

Let $[x]$ be the integer part of x, eg $[7.7] = 7$. Olive (2011b) shows that practical randomization tests that use a random sample of $\max(1000, [n \log(n)])$ permutations have level and power similar to the tests that use all M possible permutations. See Ernst (2009) and the *mpack* function `rand1way` for R code.

Another alternative to one way ANOVA is to use feasible weighted least squares (FWLS) on the cell means model with $\sigma^2 \boldsymbol{V} = diag(\sigma_1^2, ..., \sigma_p^2)$ where σ_i^2 is the variance of the ith group for $i = 1, ..., p$. Then $\hat{\boldsymbol{V}} = diag(S_1^2, ..., S_p^2)$ where $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i0})^2$ is the sample variance of the Y_{ij}. Hence the estimated weights for FWLS are $\hat{w}_{ij} \equiv \hat{w}_i = 1/S_i^2$. Then the FWLS cell means model has $Y = \boldsymbol{X}_c \boldsymbol{\beta}_c + \boldsymbol{e}$ as in (10.4) except $\mathrm{Cov}(\boldsymbol{e}) = diag(\sigma_1^2, ..., \sigma_p^2)$.

Hence $\boldsymbol{Z} = \boldsymbol{U}_c \boldsymbol{\beta}_c + \boldsymbol{\epsilon}$. Then $\boldsymbol{U}_c^T \boldsymbol{U}_c = diag(n_1 \hat{w}_1, ..., n_p \hat{w}_p)$, $(\boldsymbol{U}_c^T \boldsymbol{U}_c)^{-1} = diag(S_1^2/n_1, ..., S_p^2/n_p) = (\boldsymbol{X} \hat{\boldsymbol{V}}^{-1} \boldsymbol{X}^T)^{-1}$, and $\boldsymbol{U}_c^T \boldsymbol{Z} = (\hat{w}_1 Y_{10}, ..., \hat{w}_p Y_{p0})^T$. Thus

$$\hat{\boldsymbol{\beta}}_{FWLS} = (\overline{Y}_{10}, ..., \overline{Y}_{p0})^T = \hat{\boldsymbol{\beta}}_c.$$

That is, the FWLS estimator equals the one way ANOVA estimator of $\boldsymbol{\beta}$ based on OLS applied to the cell means model. The ANOVA F test generalizes the pooled t test in that the two tests are equivalent for $p = 2$. The FWLS procedure is also known as the Welch one way ANOVA and generalizes the Welch t test. The Welch t test is thought to be much better than the pooled t test. See Brown and Forsythe (1974ab), Kirk (1982, p. 100, 101, 121, 122) and Welch (1947, 1951).

In matrix form $\boldsymbol{Z} = \boldsymbol{U}_c\boldsymbol{\beta}_c + \boldsymbol{\epsilon}$ becomes

$$\begin{bmatrix} \sqrt{\hat{w}_1}Y_{1,1} \\ \vdots \\ \sqrt{\hat{w}_1}Y_{1,n_1} \\ \sqrt{\hat{w}_2}Y_{21} \\ \vdots \\ \sqrt{\hat{w}_2}Y_{2,n_2} \\ \vdots \\ \sqrt{\hat{w}_p}Y_{p,1} \\ \vdots \\ \sqrt{\hat{w}_p}Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} \sqrt{\hat{w}_1} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \sqrt{\hat{w}_1} & 0 & 0 & \dots & 0 \\ 0 & \sqrt{\hat{w}_2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \sqrt{\hat{w}_2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & \sqrt{\hat{w}_p} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & \sqrt{\hat{w}_p} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1,n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2,n_2} \\ \vdots \\ \epsilon_{p,1} \\ \vdots \\ \epsilon_{p,n_p} \end{bmatrix}. \tag{10.9}$$

Four tests for $Ho : \mu_1 = \cdots = \mu_p$ can be used if Rule of Thumb 10.3: $\max(S_1, ..., S_p) \leq 2\min(S_1, ..., S_p)$ fails. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, and let $Y_{(1)} \leq Y_{(2)} \cdots \leq Y_{(n)}$ be the order statistics. Then the rank transformation of the response is $\boldsymbol{Z} = rank(\boldsymbol{Y})$ where $Z_i = j$ if $Y_i = Y_{(j)}$ is the jth order statistic. For example, if $\boldsymbol{Y} = (7.7, 4.9, 33.3, 6.6)^T$, then $\boldsymbol{Z} = (3, 1, 4, 2)^T$. The first test performs the one way ANOVA F test with $\boldsymbol{Z}$ replacing $\boldsymbol{Y}$. See Montgomery (1984, p. 117-118). Two of the next three tests are described in Brown and Forsythe (1974b). Let $\lceil x \rceil$ be the smallest integer $\geq x$, eg $\lceil 7.7 \rceil = 8$. Then the Welch (1951) ANOVA F test uses test statistic

$$F_W = \frac{\sum_{i=1}^p w_i(\overline{Y}_{i0} - \tilde{Y}_{00})^2/(p-1)}{1 + \frac{2(p-2)}{p^2-1}\sum_{i=1}^p (1 - \frac{w_i}{u})^2/(n_i - 1)}$$

where $w_i = n_i/S_i^2, u = \sum_{i=1}^p w_i$ and $\tilde{Y}_{00} = \sum_{i=1}^p w_i\overline{Y}_{i0}/u$. Then the test statistic is compared to an F_{p-1,d_W} distribution where $d_W = \lceil f \rceil$ and

$$1/f = \frac{3}{p^2-1}\sum_{i=1}^p (1 - \frac{w_i}{u})^2/(n_i - 1).$$

For the modified Welch (1947) test, the test statistic is compared to an $F_{p-1,d_{MW}}$ distribution where $d_{MW} = \lceil f \rceil$ and

$$f = \frac{\sum_{i=1}^p (S_i^2/n_i)^2}{\sum_{i=1}^p \frac{1}{n_i-1}(S_i^2/n_i)^2} = \frac{\sum_{i=1}^p (1/w_i)^2}{\sum_{i=1}^p \frac{1}{n_i-1}(1/w_i)^2}.$$

Some software uses f instead of d_W or d_{MW}, and variants on the denominator degrees of freedom d_W or d_{MW} are common.

The modified ANOVA F test uses test statistic

$$F_M = \frac{\sum_{i=1}^p n_i(\overline{Y}_{i0} - \overline{Y}_{00})^2}{\sum_{i=1}^p (1 - \frac{n_i}{n})S_i^2}$$

The test statistic is compared to an F_{p-1,d_M} distribution where $d_M = \lceil f \rceil$ and

$$1/f = \sum_{i=1}^p c_i^2/(n_i - 1)$$

where

$$c_i = (1 - \frac{n_i}{n})S_i^2/[\sum_{i=1}^p (1 - \frac{n_i}{n})S_i^2].$$

The `mpack` function *anovasim* can be used to compare the five tests.

Huberty and Olejnik (2006) and Khattree and Naik (1999, ch. 4) are useful reference for MANOVA. Mardia (1971) notes that the one way MANOVA test based on Pillai's trace V is robust to nonnormality, especially when all of the treatment sample sizes are the same: $n_i \equiv h$. Permutation tests offer an alternative. See, for example, Anderson (2001).

10.7 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

10.1*. In the MANOVA model, $\hat{\boldsymbol{\beta}}_i = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_i$, and $\boldsymbol{Y}_i = \boldsymbol{X}\boldsymbol{\beta}_i + \boldsymbol{e}_i$. Treating $\boldsymbol{X}\boldsymbol{\beta}_i$ as a constant, $\mathrm{Cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = \mathrm{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$. Using this information, show $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\beta}}_j) = \sigma_{ij}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

10.2. SAS Institute (1985, p. 498 - 501) describes a one way MANOVA model. There are two groups for gender: female and male. There were $p = 4$ (skull measurements) variables $X_1 = length$ $X_2 = basilar$, $X_3 = zygomat$ and $X_4 = postorb$. There were $n_1 = 18$ females and $n_2 = 22$ males measured. Suppose $t_0 = 0.9567$ and pvalue $= 0.6566$. Here t_o was Wilk's lambda, but the other three test statistics gave the same pvalue. Do a 4 step one way MANOVA test.

10.3. Suppose the 15 units are 1 Adatorwovor, 2 Adhikari, 3 Alanzi, 4 Alsibiani, 5 AlTalib, 6 Fan, 7 Kuo, 8 Lamsal, 9 Liu, 10 Meyer, 11 Peiris, 12 Rathnayake, 13 Rupasinghe, 14 Schroeppel and 15 Watagoda. Use the following output to allocate the 15 units to three groups of 5. Show the three groups.

```
> sample(15)
 [1]  6  3  4  2  1 10  7  5 12 15 13  8 14 11  9
```

R/Splus Problems

Warning: Use the command *source("G:/mpack.txt")* **to download the programs. See Preface or Section 15.2.** Typing the name of the `mpack` function, eg *ddplot*, will display the code for the function. Use the `args` command, eg *args(ddplot)*, to display the needed arguments for the function.

10.4. The Johnson and Wichern (1988, p. 262) turtle data gives the length, width and height of painted turtle shells. There is a sample of 24 female and a sample of 24 male turtles.

a) The *R* command for this part make the response and residual plots for each of the three variables. Click the rightmost mouse button and highlight *Stop* to advance the plot. When you have the response and residual plots for one variable on the screen, copy and paste the two plots into *Word*. Do this three times, once for each variable. The male turtles are smaller than the female turtles.

b) The *R* command for this plot makes a DD plot of the residuals and adds the lines corresponding to the three prediction regions of Section 5.2. The robust cutoff is larger than the semiparametric cutoff. Place the plot in *Word*. Do the residuals appear to follow a multivariate normal distribution?

Chapter 11

Factor Analysis

11.1 Introduction

Factor analysis gives an approximation of the dispersion matrix in terms of $m < p$ unobservable random quantities called *factors*. Typically factor analysis is useful if the p random variables can be placed into a few groups of variables with fairly high correlation such that the variables within the group are not highly correlated with variables outside of the group. Let m be the number of groups. Then the hope is that the kth group can be explained by the kth factor. For example, if the $p = 6$ random variables consist of three head measurements and height, arm length and leg length, then perhaps the three head measurements are highly correlated and the three other measurements are highly correlated. Then there would be $m = 2$ groups corresponding to a "head measurement" factor and a "length" factor.

Some notation is needed before presenting the model. When the eigenvalue λ_i of $\boldsymbol{\Sigma}$ is unique, there are two standardized eigenvectors: $\boldsymbol{e}_i$ and $-\boldsymbol{e}_i$. The literature sometimes states that the standardized eigenvectors are "unique up to sign." Assume $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$. If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ for some positive constant c, then by the spectral decomposition theorem, $\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{p} \hat{\lambda}_i \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^T \xrightarrow{P} c \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T = c\boldsymbol{\Sigma}$, and $\hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^T \xrightarrow{P} \boldsymbol{e}_i \boldsymbol{e}_i^T$ for $i = 1, ..., p$ by Theorem 6.2 since $\boldsymbol{e}_i \boldsymbol{e}_i^T = (-\boldsymbol{e}_i)(-\boldsymbol{e}_i)^T$.

The factor analysis approximation of the dispersion matrix $\boldsymbol{\Sigma} \approx \boldsymbol{\Sigma}_P$ uses the first m terms of the spectral decomposition of $\boldsymbol{\Sigma}$ and a diagonal matrix $\boldsymbol{\Psi}$ so that the approximation is exact for the diagonal elements: $\boldsymbol{\Sigma}_{ii} = \boldsymbol{\Sigma}_{P,ii}$. Let the ith column of the $p \times m$ matrix $\boldsymbol{L}$ be $\sqrt{\lambda_i}\boldsymbol{e}_i$ where $m < p$.

Then $\boldsymbol{L} = \left[\sqrt{\lambda_1}\boldsymbol{e}_1 \;\; \sqrt{\lambda_2}\boldsymbol{e}_2 \;\; \ldots \;\; \sqrt{\lambda_m}\boldsymbol{e}_m \right]$. Then $\boldsymbol{\Sigma} = \sum_{i=1}^{m} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T + \sum_{i=m+1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T = \boldsymbol{L}\boldsymbol{L}^T + \sum_{i=m+1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T \approx \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi} \equiv \boldsymbol{\Sigma}_P$ where $\boldsymbol{\Psi} = diag(\psi_1, ..., \psi_p)$ and $\boldsymbol{\Sigma}_{ii} = \boldsymbol{\Sigma}_{P,ii}$. Hence $(\boldsymbol{L}\boldsymbol{L}^T)_{ii} + \psi_i = \boldsymbol{\Sigma}_{ii}$.

Definition 11.1. The *orthogonal factor analysis model* is $\boldsymbol{x} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\epsilon}$ where the $p \times 1$ random vector $\boldsymbol{x} = (X_1, ..., X_p)$, the $p \times m$ *matrix of factor loadings* $\boldsymbol{L} = ((l_{ij}))$, the $m \times 1$ random vector of *common factors* is $\boldsymbol{F} = (F_1, ..., F_m)^T$ and the $p \times 1$ error vector is $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_p)^T$. The ϵ_i are called errors or *specific factors*. The dispersion structure is $\boldsymbol{\Sigma} \approx \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi} = \boldsymbol{\Sigma}_P$ with equality for the diagonal elements. Hence $\boldsymbol{\Sigma}_{ii} = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2 + \psi_i = h_i^2 + \psi_i$ where $h_i^2 = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2$ is called the ith *communality*. The model has $X_i - \mu_i = l_{i1}F_1 + l_{i2}F_2 + \cdots + l_{im}F_m + \epsilon_m$ for $i = 1, ..., p$. The *loading* of the ith variable on the jth factor $= l_{ij}$.

Data often does not have this structure, so an important question in whether the factor analysis structure is reasonable. Note that if $\boldsymbol{\Sigma}$ is the covariance matrix, then $V(X_i) = \sigma_{ii} = \boldsymbol{\Sigma}_{ii} = h_i^2 + \psi_i$. $\boldsymbol{L}, \boldsymbol{F}, \boldsymbol{\epsilon}$ and $\boldsymbol{\mu}$ are unobservable. When $\boldsymbol{\Sigma}$ is the covariance matrix, assume that $E(\boldsymbol{F}) = \boldsymbol{0}$, $\mathrm{Cov}(\boldsymbol{F}) = \boldsymbol{I}_m$, $E(\boldsymbol{\epsilon} = \boldsymbol{0}$, $\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi}$ and that $\boldsymbol{F}$ and $\boldsymbol{\epsilon}$ are independent. Then $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{F}) = \boldsymbol{L}$ or $\mathrm{Cov}(X_i, F_j) = l_{ij}$, and $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi} = \boldsymbol{\Sigma}_P$.

Let the ith column of the $p \times m$ matrix $\hat{\boldsymbol{L}}$ be $\sqrt{\hat{\lambda}_i}\hat{\boldsymbol{e}}_i$ where $m < p$. Then $\hat{\boldsymbol{L}} = \left[\sqrt{\hat{\lambda}_1}\hat{\boldsymbol{e}}_1 \;\; \sqrt{\hat{\lambda}_2}\hat{\boldsymbol{e}}_2 \;\; \ldots \;\; \sqrt{\hat{\lambda}_m}\hat{\boldsymbol{e}}_m \right]$. Then $\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{m} \hat{\lambda}_i \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^T + \sum_{i=m+1}^{p} \hat{\lambda}_i \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^T = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \sum_{i=m+1}^{p} \hat{\lambda}_i \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^T \approx \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}} \equiv \hat{\boldsymbol{\Sigma}}_P$ where $\hat{\boldsymbol{\Psi}} = diag(\hat{\psi}_1, ..., \hat{\psi}_p)$ and $\hat{\boldsymbol{\Sigma}}_{ii} = \hat{\boldsymbol{\Sigma}}_{P,ii}$. Hence $(\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T)_{ii} + \hat{\psi}_i = \hat{\boldsymbol{\Sigma}}_{ii}$.

Definition 11.2. The *principal component factor analysis* uses the approximation $\hat{\boldsymbol{\Sigma}} \approx \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}}$. $\hat{\boldsymbol{L}}$ is called the *matrix of estimated factor loadings*. The ith *estimated communality* $\hat{h}_i^2 = \hat{l}_{i1}^2 + \hat{l}_{i2}^2 + \cdots + \hat{l}_{im}^2$ for $i = 1, .., p$. The kth column $\sqrt{\hat{\lambda}_k}\hat{\boldsymbol{e}}_k$ of $\hat{\boldsymbol{L}}$ gives the estimated factor loadings for factor F_k. These estimated factor loadings do not change as m is increased. If $\boldsymbol{\Gamma}$ is an orthogonal matrix, then $\hat{\boldsymbol{L}}^* = \hat{\boldsymbol{L}}\boldsymbol{\Gamma}$ is also a matrix of estimated factor loadings, and $\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T = \hat{\boldsymbol{L}}^*(\hat{\boldsymbol{L}}^*)^T$. The communalities are unaffected by the choice of $\boldsymbol{\Gamma}$.

Rule of thumb 11.1. To use factor analysis, assume the DD plot and subplots of the scatterplot matrix are linear. Want $n > 10p$ for classical factor analysis and $n > 20p$ for robust factor analysis that uses FCH, RFCH

or RMVN. For classical factor analysis, use the correlation matrix $\boldsymbol{R}$ instead of the covariance matrix $\boldsymbol{S}$ if $\max_{i=1,...,p} S_i^2 / \min_{i=1,...,p} S_i^2 > 2$. If $\boldsymbol{S}$ is used, also do a factor analysis using $\boldsymbol{R}$. Want the *proportion of the trace explained* by the first m factors $= \sum_{i=1}^{m} \hat{\lambda}_i / \sum_{j=1}^{p} \hat{\lambda}_j = \sum_{i=1}^{m} \hat{\lambda}_i / tr(\hat{\boldsymbol{\Sigma}}) > 0.7$. Want $m < \min(10, p)$. Suppose $(T, \hat{\boldsymbol{\Sigma}})$ is the estimator of multivariate location and dispersion. Make a plot of $D_i(T, \hat{\boldsymbol{\Sigma}}_P)$ versus $D_i(T, \hat{\boldsymbol{\Sigma}})$ with the identity line that has unit slope and zero intercept added as a visual aid. If $\hat{\boldsymbol{\Sigma}}_P$ is an adequate approximation of $\hat{\boldsymbol{\Sigma}}$, then the plotted points should cluster tightly about the identity line.

11.2 Robust Factor Analysis

Robust factor analysis can be done using the FCH, RFCH or RMVN dispersion estimator as $\hat{\boldsymbol{\Sigma}}$. Under (E1) the robust factor analysis has $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ while $\boldsymbol{S} \xrightarrow{P} c_X \boldsymbol{\Sigma}$. If the generalized correlation matrix is used as $\hat{\boldsymbol{\Sigma}}$, then the classical and robust methods both satisfy $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\rho}$. The RMVN method is easy to program since it is the classical factor analysis applied to the RMVN subset.

11.3 Summary

1) Factor analysis is use to write $\hat{\boldsymbol{\Sigma}} \approx \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}} = \hat{\boldsymbol{\Sigma}}_F$. Factor analysis clusters variables into groups called factors and suggests that the $m < p$ factors explain the dispersion more simply than $X_1, ..., X_p$. $\hat{\boldsymbol{L}} = [\boldsymbol{L}_1, ..., \boldsymbol{L}_m]$ is the matrix of factor loadings.

2) Factor analysis output is a lot like PCA output, but replace PC1, ..., PCp by Factor 1, ..., Factor m:

Factor 1	Factor 2	$\cdots$	Factor m
$\hat{\boldsymbol{L}}_1$	$\hat{\boldsymbol{L}}_2$	$\cdots$	$\hat{\boldsymbol{L}}_m$

3) To try to explain Factor j, look at entries in $\hat{\boldsymbol{L}}_j$ that are large in magnitude and ignore entries close to zero. Sometimes only one entry is large. Sometimes all of the large entries have approximately the same size and sign, then the Factor is interpreted as an average of these entrees. If all of the large entries have approximately the same size but different signs then the Factor is interpreted as the sum of the variables with the positive sign $-$

the sum of the variables with a minus sign. Thus if exactly two entries are of similar large magnitude but of different sign, the Factor is interpreted as a difference of the two entrees. If there are $k \geq 2$ large entrees that differ in magnitude, then the Factor is interpreted as a linear combination of the corresponding variables.

4) The proportion of variance explained and cumulative proportion of variance explained are interpreted as for PCA. Use the k factor model if the proportion of the variance explained by the first k Factors is larger than some percentage such as 50%, 60%, 70%, 80% or 90%.

5) For a k factor model, want the degrees of freedom $d \geq 0$ where $d = 0.5(p-k)^2 - 0.5(p+k)$.

6) If the 1 factor model is not adequate, R will give a test for whether a k factor model is sufficient. A k factor model with $pval < 0.05$ is not sufficient: more factors are needed. A k factor model with $pval > 0.05$ is sufficient.

7) Let $\hat{\boldsymbol{\Gamma}}$ be an orthogonal matrix. The $\hat{\boldsymbol{L}}_\Gamma \hat{\boldsymbol{L}}_\Gamma^T = \hat{\boldsymbol{L}}\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Gamma}}^T\hat{\boldsymbol{L}}^T = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T$. The varimax and promax rotations seek $\hat{\boldsymbol{\Gamma}}$ such that $\hat{\boldsymbol{L}}_\Gamma = \hat{\boldsymbol{L}}\hat{\boldsymbol{\Gamma}}$ has loadings that are easier to interpret than the loadings of $\hat{\boldsymbol{L}}$. The promax rotation attempts to produce loading with a lot of zeroes.

11.4 Complements

Kosfeld (1996) does factor analysis with the DGK estimator.

11.5 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

```
Loadings:
               Factor1 Factor2
height          0.872
arm.span        0.973
forearm         0.938
lower.leg       0.876
weight                  0.961
bitro.diameter          0.803
```

```
chest.girth             0.796
chest.width     0.125   0.611

               Factor1 Factor2
SS loadings      3.375   2.589
Proportion Var   0.422   0.324
Cumulative Var   0.422   0.745
```

11.1***.** The above output is for the factor analysis using a correlation matrix of eight physical measurements on 305 girls between ages seven and seventeen.

a) What is the cumulative variance explained by the 2 factors?

b) Which factor has a nonzero loading for weight?

c) Explain Factor 2.

```
factanal(marry,factors=2,rotation="promax")

Uniquenesses:  pop    mmen    mwmn   mmilmen  milwmn
              0.010   0.005   0.005   0.005   0.005

Loadings:Factor1 Factor2
pop      0.986
mmen     1.003
mwmn     1.003
mmilmen          0.965
milwmn           0.958

               Factor1 Factor2
SS loadings      2.995   1.850
Proportion Var   0.599   0.370
Cumulative Var   0.599   0.969
```

11.2. The above output is for a factor analysis of the Hebbler (1847) data from the the 1843 Prussia census. Sometimes if the wife or husband was not at the household, then s/he would not be counted. $X_1 = pop =$ population of the district in 1843, $X_2 = mmen =$ number of married civilian men in the district, $X_3 = mwmn =$ number of women married to civilians in the district, $X_4 = mmilmen =$ number of married military men in the

district, and $x_5 = milwmn$ = number of women married to military men in the district.

a) What is the cumulative variance explained by the 2 factors?

b) Explain Factor 1.

c) Explain Factor 2.

```
Uniquenesses:
     age  breadth cephalic   circum   headht   height      len     size  cbrainy
   0.005    0.005    0.005    0.142    0.005    0.303    0.005    0.005    0.366

Loadings:
         Factor1 Factor2 Factor3 Factor4
log(age)          1.026
breadth   0.874           0.461  -0.142
cephalic -0.115           1.020
circum    0.849   0.113
headht                             0.965
height    0.202   0.597            0.204
len       1.109          -0.363  -0.156
size      0.805                    0.231
brainwt   0.642  -0.262            0.296

               Factor1 Factor2 Factor3 Factor4
SS loadings      3.833   1.491   1.389   1.161
Proportion Var   0.426   0.166   0.154   0.129
Cumulative Var   0.426   0.592   0.746   0.875
```

11.3. The above output is for the factor analysis of the Gladstone (1905-6) data. The variables included $log(age)$ and *height* and 7 head measurements breadth, cephalic, circum, headht, len, size, and brain weight.

a) What is the cumulative variance explained by the 4 factors?

b) Which factor has a nonzero loading for log(age)?

c) Explain Factor 3.

R/Splus Problems

Warning: Use the command *source("G:/mpack.txt")* **to download the programs. See Preface or Section 15.2.** Typing the name of the `mpack` function, eg *ddplot*, will display the code for the function. Use the

`args` command, eg *args(ddplot)*, to display the needed arguments for the function.

11.4. The Buxton data has 5 massive outliers in variables len and buxy = height.

a) The *R* commands for this part do a factor analysis on the Buxton data using the sample covariance matrix. Copy and paste the output into *Word*.

i) Which variables have nonzero loadings for factor 1?

ii) Which variables have nonzero loadings for factor 2?

iii) What is the cumulative variance explained by the two factors?

b) The *R* commands for this part do a factor analysis on the Buxton data using the RMVN dispersion matrix. Copy and paste the output into *Word*.

i) Which variables have nonzero loadings for factor 1?

ii) Which variables have nonzero loadings for factor 2?

iii) What is the cumulative variance explained by the two factors?

Chapter 12

Multivariate Linear Regression

12.1 Introduction

Definition 12.1. The **response variables** are the variables that you want to predict. The **predictor variables** are the variables used to predict the response variables.

Notation. The **multivariate linear regression model** $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $X_1, X_2, ..., X_p$ where $X_1 = 1$ is the trivial predictor. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (1, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$ where the 1 could be omitted.

In matrix form, the model is $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$, and the data matrix $\boldsymbol{W} = [\boldsymbol{X} \;\; \boldsymbol{Y}]$ except usually the first column $\boldsymbol{1}$ of $\boldsymbol{X}$ is omitted. The $n \times m$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,m} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n,1} & Y_{n,2} & \dots & Y_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \dots & \boldsymbol{Y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix}.$$

The $n \times p$ matrix

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

where $\boldsymbol{v}_1 = \mathbf{1}$.

The $p \times m$ matrix

$$\boldsymbol{B} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \dots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \dots & \beta_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p,1} & \beta_{p,2} & \dots & \beta_{p,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \dots & \boldsymbol{\beta}_m \end{bmatrix}.$$

The $n \times m$ matrix

$$\boldsymbol{E} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \dots & \epsilon_{1,m} \\ \epsilon_{2,1} & \epsilon_{2,2} & \dots & \epsilon_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \dots & \epsilon_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \dots & \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix}.$$

Warning: The $\boldsymbol{e}_i$ are error vectors, not orthonormal eigenvectors.

Definition 12.2. In the *multiple linear regression model*,

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{12.1}$$

for $i = 1, \dots, n$. In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{12.2}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Equivalently,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{12.3}$$

The e_i are iid with zero mean and variance σ^2, and multiple linear regression is used to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2.

Each response variable in a multivariate linear regression model follows a multiple linear regression model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is

assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$. Hence the errors corresponding to the jth response are uncorrelated with variance $\sigma_j^2 = \sigma_{jj}$. Notice that the **same design matrix** $\boldsymbol{X}$ of predictors is used for each of the m models, but the jth response variable vector $\boldsymbol{Y}_j$, coefficient vector $\boldsymbol{\beta}_j$ and error vector $\boldsymbol{e}_j$ change and thus depend on j.

Now consider the ith case $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T)$ which corresponds to the ith row of $\boldsymbol{Z}$ and the ith row of $\boldsymbol{X}$. Then

$$\begin{bmatrix} Y_{i1} = \beta_{11}x_{i1} + \cdots + \beta_{p1}x_{ip} + \epsilon_{i1} = \boldsymbol{x}_i^T\boldsymbol{\beta}_1 + \epsilon_{i1} \\ Y_{i2} = \beta_{12}x_{i1} + \cdots + \beta_{p2}x_{ip} + \epsilon_{i2} = \boldsymbol{x}_i^T\boldsymbol{\beta}_2 + \epsilon_{i2} \\ \vdots \\ Y_{im} = \beta_{1m}x_{i1} + \cdots + \beta_{pm}x_{ip} + \epsilon_{im} = \boldsymbol{x}_i^T\boldsymbol{\beta}_m + \epsilon_{im} \end{bmatrix}$$

or $\boldsymbol{y}_i = \boldsymbol{\mu}_{\boldsymbol{x}_i} + \boldsymbol{\epsilon}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i$ where

$$E(\boldsymbol{y}_i) = \boldsymbol{\mu}_{\boldsymbol{x}_i} = \boldsymbol{B}^T\boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{x}_i^T\boldsymbol{\beta}_1 \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_m \end{bmatrix}.$$

The notation $\boldsymbol{y}_i|\boldsymbol{x}_i$ and $E(\boldsymbol{y}_i|\boldsymbol{x}_i)$ is more accurate, but usually the conditioning is suppressed. Taking $\boldsymbol{\mu}_{\boldsymbol{x}_i}$ to be a constant (or condition on $\boldsymbol{x}_i$ if the predictor variables are random variables), $\boldsymbol{y}_i$ and $\boldsymbol{\epsilon}_i$ have the same covariance matrix. In the multivariate regression model, this covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ does not depend on i. Observations from different cases are uncorrelated (often independent), but the m errors for the m different response variables for the *same case* are correlated. If $\boldsymbol{X}$ is a random matrix, then assume $\boldsymbol{X}$ and $\boldsymbol{E}$ are independent and that expectations are conditional on $\boldsymbol{X}$.

Definition 12.3. The **multivariate linear regression model** $\boldsymbol{y}_k = \boldsymbol{B}^T\boldsymbol{x}_k + \boldsymbol{\epsilon}_k$ for $k = 1, ..., n$ is written in matrix form as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = ((\sigma_{ij}))$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$. Then $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{X}\boldsymbol{B}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$. Considering the kth row of $\boldsymbol{Z}, \boldsymbol{X}$ and $\boldsymbol{E}$ shows that $\boldsymbol{y}_k^T = \boldsymbol{x}_k^T\boldsymbol{B} + \boldsymbol{\epsilon}_k^T$.

Example 12.1. Suppose it is desired to predict the response variables $Y_1 =$ *height* and $Y_2 =$ *height at shoulder* of a person from partial skeletal

remains. A model for prediction can be built from nearly complete skeletons or from living humans, depending on the population of interest (eg ancient Egyptians or modern US citizens). The predictor variables might be $x_1 \equiv 1$, $x_2 = \textit{femur length}$ and $x_3 = \textit{ulna length}$. The two heights of individuals with $x_2 = 200mm$ and $x_3 = 140mm$ should be shorter on average than the two heights of individuals with $x_2 = 500mm$ and $x_3 = 350mm$. In this example Y_1, Y_2, x_2 and x_3 are quantitative variables. If $x_4 = \textit{gender}$ is a predictor variable, then gender (coded as male = 1 and female = 0) is qualitative.

Definition 12.4. Least squares is the classical method for fitting multivariate linear regression. The **least squares estimators** are $\hat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z} = \left[\begin{array}{cccc} \hat{\boldsymbol{\beta}}_1 & \hat{\boldsymbol{\beta}}_2 & \ldots & \hat{\boldsymbol{\beta}}_m \end{array} \right]$. The *predicted values* or *fitted values*

$$\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \left[\begin{array}{cccc} \hat{\boldsymbol{Y}}_1 & \hat{\boldsymbol{Y}}_2 & \ldots & \hat{\boldsymbol{Y}}_m \end{array} \right] = \left[\begin{array}{cccc} \hat{Y}_{1,1} & \hat{Y}_{1,2} & \ldots & \hat{Y}_{1,m} \\ \hat{Y}_{2,1} & \hat{Y}_{2,2} & \ldots & \hat{Y}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{Y}_{n,1} & \hat{Y}_{n,2} & \ldots & \hat{Y}_{n,m} \end{array} \right].$$

The *residuals* $\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} =$

$$\left[\begin{array}{c} \hat{\boldsymbol{\epsilon}}_1^T \\ \hat{\boldsymbol{\epsilon}}_2^T \\ \vdots \\ \hat{\boldsymbol{\epsilon}}_n^T \end{array} \right] = \left[\begin{array}{cccc} \hat{\boldsymbol{r}}_1 & \hat{\boldsymbol{r}}_2 & \ldots & \hat{\boldsymbol{r}}_m \end{array} \right] = \left[\begin{array}{cccc} \hat{\epsilon}_{1,1} & \hat{\epsilon}_{1,2} & \ldots & \hat{\epsilon}_{1,m} \\ \hat{\epsilon}_{2,1} & \hat{\epsilon}_{2,2} & \ldots & \hat{\epsilon}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\epsilon}_{n,1} & \hat{\epsilon}_{n,2} & \ldots & \hat{\epsilon}_{n,m} \end{array} \right].$$

These quantities can be found from the m multiple linear regressions of Y_j on the predictors: $\widehat{\boldsymbol{\beta}}_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_j$, $\widehat{\boldsymbol{Y}}_j = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j$ and $\hat{\boldsymbol{r}}_j = \boldsymbol{Y}_j - \hat{\boldsymbol{Y}}_j$ for $j = 1, ..., m$. Hence $\hat{\epsilon}_{i,j} = Y_{i,j} - \hat{Y}_{i,j}$ where $\hat{\boldsymbol{Y}}_j = (\hat{Y}_{1,j}, ..., \hat{Y}_{n,j})^T$. Finally, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} =$

$$\frac{(\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T(\boldsymbol{Z} - \hat{\boldsymbol{Z}})}{n-d} = \frac{(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})^T(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})}{n-d} = \frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-d} = \frac{1}{n-d}\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T.$$

The choices $d = 0$ and $d = p$ are common. If $d = 1$, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=1} = \boldsymbol{S}_r$, the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ since the sample mean of the $\hat{\boldsymbol{\epsilon}}_i$ is $\boldsymbol{0}$. Let $\hat{\boldsymbol{\Sigma}}_\epsilon = \hat{\boldsymbol{\Sigma}}_{\epsilon,p}$ be the unbiased estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Also,

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} = (n-d)^{-1}\boldsymbol{Z}^T[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}]\boldsymbol{Z},$$

and

$$\hat{\boldsymbol{E}} = [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}]\boldsymbol{Z}.$$

Theorem 12.1, (Johnson and Wichern (1988, p. 304): Suppose $\boldsymbol{X}$ has full rank $p < n$ and the covariance structure of Definition 12.3 holds. Then $E(\hat{\boldsymbol{B}}) = \boldsymbol{B}$ so $E(\hat{\boldsymbol{\beta}}_j) = \boldsymbol{\beta}_j$, $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_k) = \sigma_{jk}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ for $j, k = 1, ..., p$. Also $\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{B}}$ are uncorrelated, $E(\hat{\boldsymbol{E}}) = \boldsymbol{0}$ and

$$E(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = E\left(\frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-p}\right) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}.$$

Theorem 12.2. $\boldsymbol{S}_r = \boldsymbol{\Sigma}_\epsilon + O_P(n^{-1/2})$ if $\boldsymbol{B} - \hat{\boldsymbol{B}} = O_P(n^{-1/2})$, $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{x}_i^T = O_P(1)$, $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T = O_P(n^{1/2})$ and $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + O_P(n^{-1/2})$.

Proof. Note that $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i = \hat{\boldsymbol{B}}^T\boldsymbol{x}_i + \hat{\boldsymbol{\epsilon}}_i$. Hence $\hat{\boldsymbol{\epsilon}}_i = (\boldsymbol{B} - \hat{\boldsymbol{B}})^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$. Thus

$$\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T = \sum_{i=1}^n (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i)(\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i)^T = \sum_{i=1}^n [\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T + \boldsymbol{\epsilon}_i(\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i)^T + (\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i)\hat{\boldsymbol{\epsilon}}_i^T] =$$

$$\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T + (\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{x}_i^T)(\boldsymbol{B} - \hat{\boldsymbol{B}}) + (\boldsymbol{B} - \hat{\boldsymbol{B}})^T(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{\epsilon}_i^T) + (\boldsymbol{B} - \hat{\boldsymbol{B}})^T(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T)(\boldsymbol{B} - \hat{\boldsymbol{B}}).$$

Thus $\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T +$

$$O_P(1)O_P(n^{-1/2}) + O_P(n^{-1/2})O_P(1) + O_P(n^{-1/2})O_P(n^{1/2})O_P(n^{-1/2}),$$

and the result follows since $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + O_P(n^{-1/2})$ and

$$\boldsymbol{S}_r = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T.$$

12.2 Checking the Model

12.2.1 Plots

Notation. Plots will be used to simplify regression analysis, and in this text a plot of W versus Z uses W on the horizontal axis and Z on the vertical axis.

Definition 12.5. A **response plot** for the jth response variable is a plot of the fitted values $\widehat{Y}_{ij}$ versus the response Y_{ij}. The identity line with slope one and zero intercept is added to the plot as a visual aid. A **residual plot** corresponding to the jth response variable is a plot of $\hat{Y}_{ij}$ versus r_{ij}.

Remark 12.1. Make the m response and residual plots for any multivariate linear regression. In a response plot, the vertical deviations from the identity line are the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij}$. If the model is appropriate, then the plotted points should cluster about the identity line in each of the m response plots. If outliers are present or if the plot is not linear, then the current model or data need to be changed or corrected. If the model is good, then the each of the m residual plots should be ellipsoidal with no trend and should be centered about the $r = 0$ line. There should not be any pattern in the residual plot: as a narrow vertical strip is moved from left to right, the behavior of the residuals within the strip should show little change. Outliers and patterns such as curvature or a fan shaped plot are bad.

Notation. A *rule of thumb* is a rule that often but not always works well in practice.

Rule of thumb 12.1. Use multivariate linear regression if $n > 10 \max(p, m)$. The m response and residual plots should all look good. Make the DD plot of the $\hat{\boldsymbol{\epsilon}}_i$. If a residual plot would look good after several points have been deleted, and if these deleted points were not gross outliers (points far from the point cloud formed by the bulk of the data), then the residual plot is probably good. Beginners often find too many things wrong with a good model. For practice, use the computer to generate several multivariate linear regression data sets, and make the m response and residual plots for these data sets. This exercise will help show that the plots can have considerable variability even when the multivariate linear regression model is good.

Rule of thumb 12.2. If the plotted points in the residual plot look like a left or right opening megaphone, the first model violation to check is the assumption of nonconstant variance. (This is a rule of thumb because it is possible that such a residual plot results from another model violation such as nonlinearity, but nonconstant variance is much more common.)

Remark 12.2. Residual plots *magnify departures* from the model while the response plots emphasizes *how well the multivariate linear regression model fits the data.*

Definition 12.6. An **RR plot** is a scatterplot matrix of the m sets of residuals $\boldsymbol{r}_1, ..., \boldsymbol{r}_m$.

Definition 12.7. An **FF plot** is a scatterplot matrix of the m sets of fitted values of response variables $\hat{\boldsymbol{Y}}_1, ..., \hat{\boldsymbol{Y}}_m$. The m response variables $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_m$ can be added to the plot.

Remark 12.3. Multivariate linear regression makes the most sense if the m errors are linearly related, eg from an elliptically contoured distribution. Make the RR plot and a DD plot of the residuals $\hat{\boldsymbol{\epsilon}}_i$ to check that the errors are linearly related. Make a DD plot of the continuous predictor variables to check for x-outliers. Make a DD plot of $Y_1, ..., Y_m$ to check for outliers, especially if it is assumed that the response variables come from an elliptically contoured distribution.

Example 12.2. Tremearne (1911) presents a data set of about 17 measurements on 115 people of Hausa nationality. We deleted 3 cases (107, 108 and 109) because of missing values and used *height* as the response variable Y_1. Suppose Y_2 is the other response variable and that the response and residual plots for Y_2 are well behaved. Along with a constant $x_{i,1} \equiv 1$, the five additional predictor variables used were *height when sitting*, *height when kneeling*, *head length, nasal breadth*, and *span* (perhaps from left hand to right hand). Figure 12.1 presents the response and residual plots corresponding the response variable Y_1 = *height* for this data set. These plots show that the model should be useful for the data since the plotted points in the response plot are linear and follow the identity line while the plotted points in the residual plot follow the $r = 0$ line with no other pattern (except for a possible outlier marked 44).

To use the response plot to visualize the conditional distribution of $Y_1|\boldsymbol{x}^T\boldsymbol{\beta}_1$, use the fact that the fitted values $\hat{Y}_1 = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}_1$. For example, suppose the height given fit = 1700 is of interest. Mentally examine the plot about a narrow vertical strip about fit = 1700, perhaps from 1675 to 1725. The cases in the narrow strip have a mean close to 1700 since they fall close to the identity line. Similarly, when the fit = w for w between 1500 and 1850, the cases have heights near w, on average.

Cases 3, 44 and 63 are highlighted. The 3rd person was very tall while the 44th person was rather short. Beginners often label too many points as outliers. Mentally draw a box about the bulk of the data ignoring any outliers. Double the width of the box (about the identity line for the response

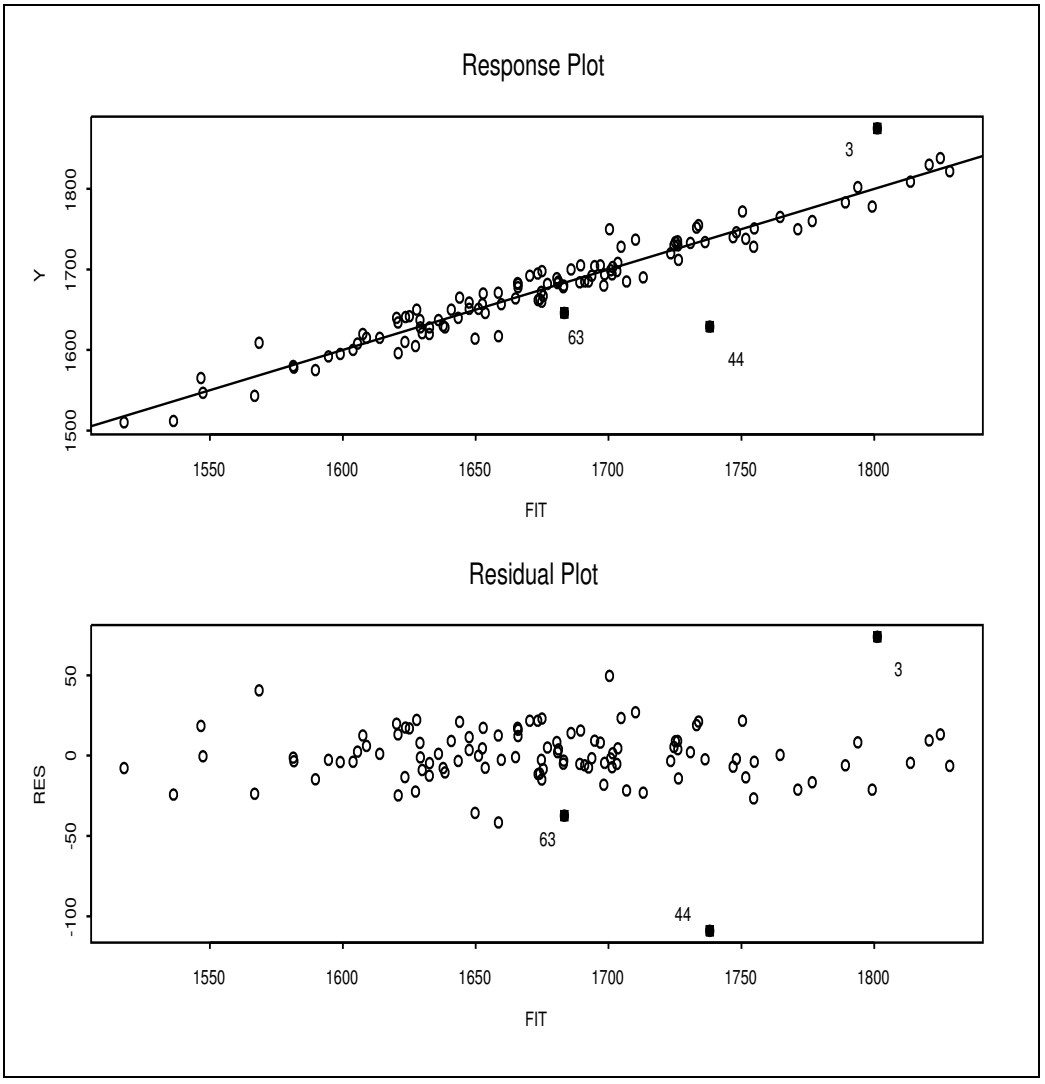

Figure 12.1: Residual and Response Plots for the Response Variable Height

plot and about the horizontal $r = 0$ line for the residual plot). Cases outside of this imaginary doubled box are potential outliers. Alternatively, visually estimate the standard deviation of the residuals in both plots. In the residual plot look for residuals that are more than 5 standard deviations from the $r = 0$ line. In Figure 12.1, the standard deviation of the residuals appears to be around 10. Hence cases 3 and 44 are certainly worth examining.

The plots corresponding to Y_1 can be made with the following commands. In general store $Y1, Y2, ..., Ym$ and make the *MLRplot(X, Y)* command m times for $Y = Y1, ..., Ym$.

```
source("G:/mpack.txt")
#assume the data is stored in R matrix major
X<-major[,-6]; Y1 <- major[,6]; MLRplot(X,Y1)
```

12.2.2 Predictor and Response Transformations

Predictor transformations for the continuous predictors can be made exactly as in Section 2.4.

Warning: The Rule of thumb 2.1 does not always work. For example, the log rule may fail. If the relationships in the scatterplot matrix are already linear or if taking the transformation does not increase the linearity, then no transformation may be better than taking a transformation. For the *Arc* data set `evaporat.lsp`, the log rule suggests transforming the response variable *Evap*, but no transformation works better.

Response transformations can also be made as in Section 2.4, but there is an alternative graphical method for response transformations once the predictors are fixed. Discussion will first be given for multiple linear regression with response variable Y. Then for multivariate regression, simply use the transformation plots for each of the m response variables $Y_1, ..., Y_m$.

An important class of *response transformation models* adds an additional unknown transformation parameter λ_o, such that

$$Y_i = t_{\lambda_o}(Z_i) \equiv Z_i^{(\lambda_o)} = E(Y_i|\boldsymbol{x}_i) + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i. \tag{12.4}$$

If λ_o was known, then $Y_i = t_{\lambda_o}(Z_i)$ would follow a multiple linear regression model with p predictors including the constant. Here, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients depending on λ_o, $\boldsymbol{x}$ is a $p \times 1$ vector of predictors

that are assumed to be measured with negligible error, and the errors e_i are assumed to be iid with zero mean.

Definition 12.8. Assume that **all** of the values of the "response" Z_i are **positive**. A *power transformation* has the form $Y = t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$ where

$$\lambda \in \Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\}.$$

Definition 12.9. Assume that **all** of the values of the "response" Z_i are **positive**. Then the *modified power transformation family*

$$t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda} \tag{12.5}$$

for $\lambda \neq 0$ and $Z_i^{(0)} = \log(Z_i)$. Often $Z_i^{(1)}$ is replaced by Z_i for $\lambda = 1$. Generally $\lambda \in \Lambda$ where Λ is some interval such as $[-1, 1]$ or a coarse subset such as Λ_L. This family is a special case of the response transformations considered by Tukey (1957).

A graphical method for response transformations refits the model using the same fitting method: changing only the "response" from Z to $t_\lambda(Z)$. Compute the "fitted values" $\hat{W}_i$ using $W_i = t_\lambda(Z_i)$ as the "response." Then a *transformation plot* of $\hat{W}_i$ versus W_i is made for each of the seven values of $\lambda \in \Lambda_L$ with the identity line added as a visual aid. Vertical deviations from the identity line are the "residuals" $r_i = W_i - \hat{W}_i$. Then a candidate response transformation $Y = t_{\lambda^*}(Z)$ is reasonable if the plotted points follow the identity line in a roughly evenly populated band. Curvature from the identity line suggests that the candidate response transformation is inappropriate.

Definition 12.10. A *transformation plot* is a plot of $\hat{W}$ versus W with the identity line added as a visual aid.

There are several reasons to use a coarse grid of powers. First, several of the powers correspond to simple transformations such as the log, square root, and cube root. These powers are easier to interpret than $\lambda = .28$, for example. According to Mosteller and Tukey (1977, p. 91), the **most commonly used power transformations** are the $\lambda = 0$ (log), $\lambda = 1/2$, $\lambda = -1$ and $\lambda = 1/3$ transformations in decreasing frequency of use. Secondly, if the estimator $\hat{\lambda}_n$ can only take values in Λ_L, then sometimes $\hat{\lambda}_n$ will

converge (eg in probability) to $\lambda^* \in \Lambda_L$. Thirdly, Tukey (1957) showed that neighboring power transformations are often very similar, so restricting the possible powers to a coarse grid is reasonable. Note that powers can always be added to the grid Λ_L. Useful powers are $\pm 1/4, \pm 2/3, \pm 2$, and ± 3. Powers from numerical methods can also be added.

Application 12.1. This graphical method for selecting a response transformation is very simple. Let $W_i = t_\lambda(Z_i)$. Then for each of the seven values of $\lambda \in \Lambda_L$, perform least squares (OLS) on $(W_i, \boldsymbol{x}_i)$ and make the transformation plot of $\hat{W}_i$ versus W_i. If the plotted points follow the identity line for λ^*, then take $\hat{\lambda}_o = \lambda^*$, that is, $Y = t_{\lambda^*}(Z)$ is the response transformation. (Note that this procedure can be modified to create a graphical diagnostic for a numerical estimator $\hat{\lambda}$ of λ_o by adding $\hat{\lambda}$ to Λ_L.) Note that for multivariate regression, use $W = Y_j$ for $j = 1, ..., m$. Hence $7m$ plots will be made.

If more than one value of $\lambda \in \Lambda_L$ gives a linear plot, take the simplest or most reasonable transformation or the transformation that makes the most sense to subject matter experts. Also check that the corresponding "residual plots" of $\hat{W}$ versus $W - \hat{W}$ look reasonable. The values of λ in decreasing order of importance are $1, 0, 1/2, -1$ and $1/3$. So the log transformation would be chosen over the cube root transformation if both transformation plots look equally good.

After selecting the transformations, the usual checks on the multivariate regression model should be made. In particular, make the m response and residual plots. In particular, the transformation plot for the selected transformation is the response plot, and a residual plot should also be made.

The following two examples illustrates the procedure for a single response variable $Y = Y_1$, and the plots show $t_\lambda(Z)$ on the vertical axis. The label "TZHAT" of the horizontal axis are the "fitted values" that result from using $t_\lambda(Z)$ as the "response" in the OLS software. In general for multivariate regression, the plots would be made for $Z_1, ..., Z_m$ resulting in response variables $Y_1 = t_1(Z_1), ..., Y_m = t_m(Z_m)$.

Example 12.3: Textile Data. In their pioneering paper on response transformations, Box and Cox (1964) analyze data from a 3^3 experiment on the behavior of worsted yarn under cycles of repeated loadings. The "response" Z is the *number of cycles to failure* and a constant is used along with the three predictors *length, amplitude* and *load.* Using the normal profile log likelihood for λ_o, Box and Cox determine $\hat{\lambda}_o = -0.06$ with approximate 95

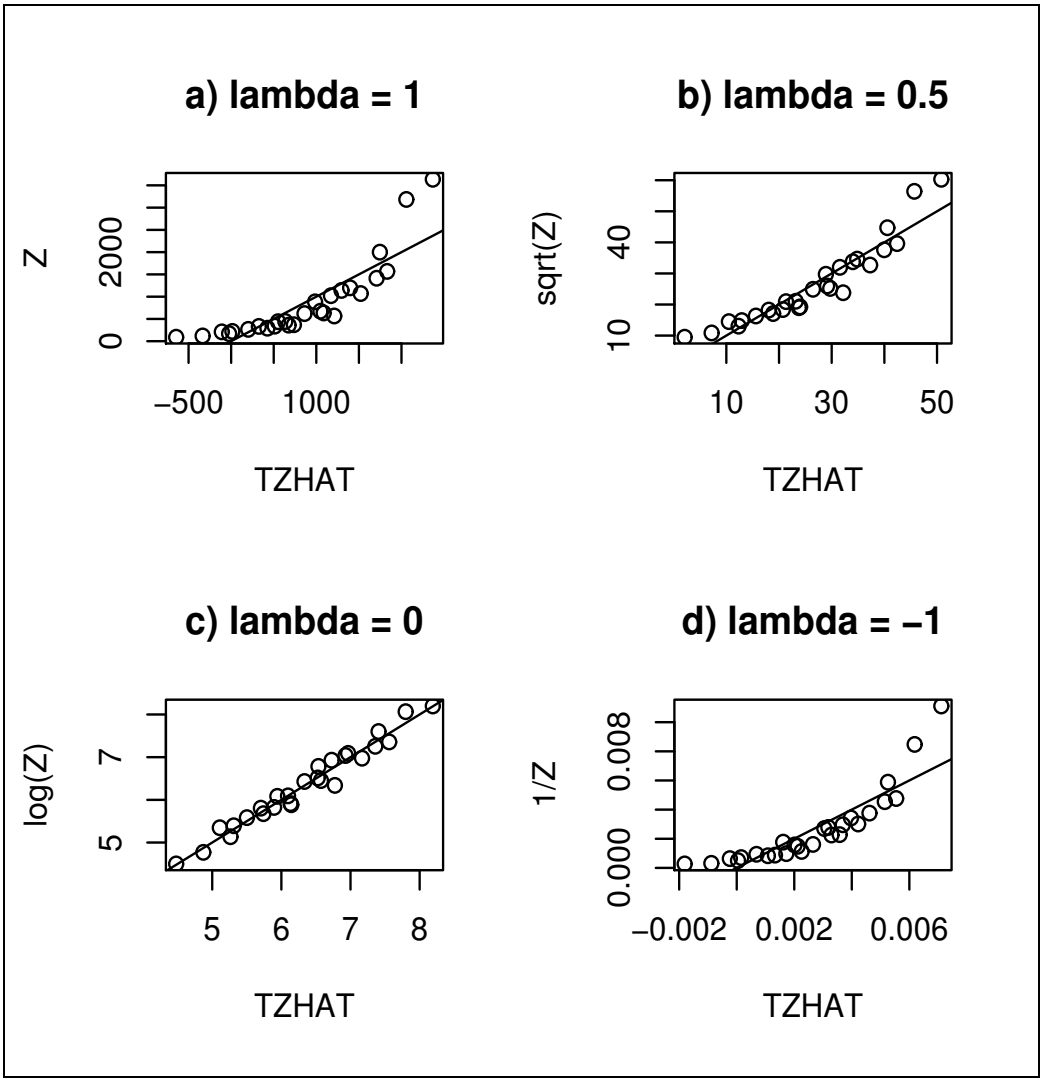

Figure 12.2: Four Transformation Plots for the Textile Data

percent confidence interval -0.18 to 0.06. These results give a strong indication that the log transformation may result in a relatively simple model, as argued by Box and Cox. Nevertheless, the numerical Box–Cox transformation method provides no direct way of judging the transformation against the data.

Shown in Figure 12.2 are transformation plots of $\hat{Z}$ versus Z^λ for four values of λ except $\log(Z)$ is used if $\lambda = 0$. The plots show how the transformations bend the data to achieve a homoscedastic linear trend. Perhaps more importantly, they indicate that the information on the transformation is spread throughout the data in the plot since changing λ causes all points along the curvilinear scatter in Figure 12.2a to form along a linear scatter in Figure 12.2c. Dynamic plotting using λ as a control seems quite effective for judging transformations against the data and the log response transformation does indeed seem reasonable.

Note the simplicity of the method: Figure 12.2a shows that a response transformation is needed since the plotted points follow a nonlinear curve while Figure 12.2c suggests that $Y = \log(Z)$ is the appropriate response transformation since the plotted points follow the identity line. If all 7 plots were made for $\lambda \in \Lambda_L$, then $\lambda = 0$ would be selected since this plot is linear. Also, Figure 12.2a suggests that the log rule is reasonable since $\max(Z)/\min(Z) > 10$.

The essential point of the next example is that observations that influence the choice of the usual Box–Cox numerical power transformation are often easily identified in the transformation plots. The transformation plots are especially useful if the bivariate relationships of the predictors, as seen in the scatterplot matrix of the predictors, are linear.

Example 12.4: Mussel Data. Cook and Weisberg (1999a, p. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. Suppose the response Z is *muscle mass* M in grams, and the predictors are the *length* L and *height* H of the shell in mm, the logarithm $\log W$ of the *shell width* W, the logarithm $\log S$ of the *shell mass* S and a constant. With this starting point, we might expect a log transformation of M to be needed because M and S are both mass measurements and $\log S$ is being used as a predictor. Using $\log M$ would essentially reduce all measurements to the scale of length. The Box–Cox likelihood method gave $\hat{\lambda}_0 = 0.28$ with approximate 95 percent confidence interval 0.15 to 0.4. The log transformation

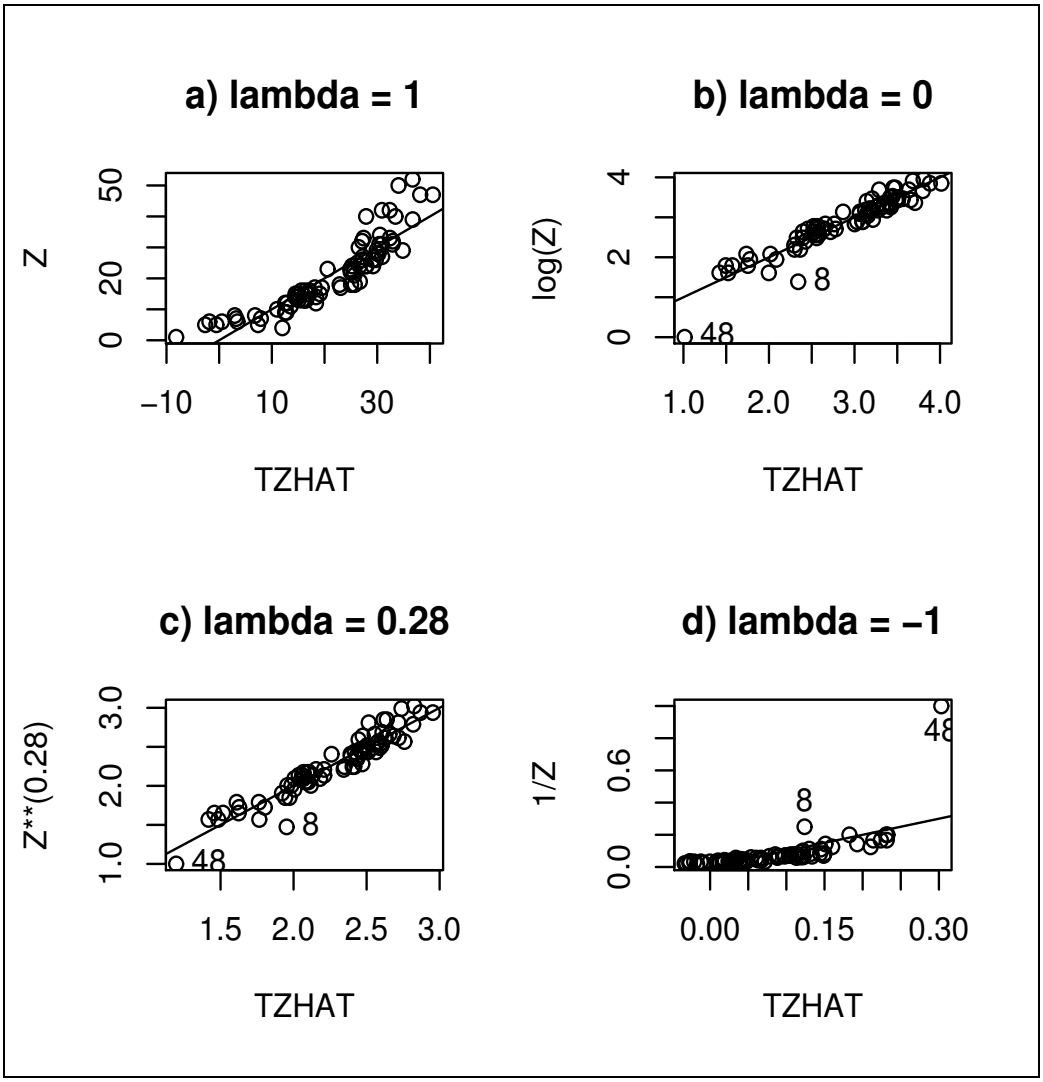

Figure 12.3: Transformation Plots for the Mussel Data

is excluded under this inference leading to the possibility of using different transformations of the two mass measurements.

Shown in Figure 12.3 are transformation plots for four values of λ. A striking feature of these plots is the two points that stand out in three of the four plots (cases 8 and 48). The Box–Cox estimate $\hat{\lambda} = 0.28$ is evidently influenced by the two outlying points and, judging deviations from the identity line in Figure 12.3c, the mean function for the remaining points is curved. In other words, the Box–Cox estimate is allowing some visually evident curvature in the bulk of the data so it can accommodate the two outlying points. Recomputing the estimate of λ_o without the highlighted points gives $\hat{\lambda}_o = -0.02$, which is in good agreement with the log transformation anticipated at the outset. Reconstruction of the transformation plots indicated that now the information for the transformation is consistent throughout the data on the horizontal axis of the plot.

Note that in addition to helping visualize $\hat{\lambda}$ against the data, the transformation plots can also be used to show the curvature and heteroscedasticity in the competing models indexed by $\lambda \in \Lambda_L$. Example 12.3 shows that the plot can also be used as a diagnostic to assess the success of numerical methods such as the Box–Cox procedure for estimating λ_o.

12.3 Variable Selection

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted without important loss of information. First we review variable selection for the multiple linear regression (MLR) model, and then adapt the techniques for multivariate linear regression.

12.3.1 Variable Selection for the MLR Model

This subsection follows Olive and Hawkins (2005) closely. A *model for variable selection* in multiple linear regression can be described by

$$Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = \boldsymbol{\beta}^T\boldsymbol{x} + e = \boldsymbol{\beta}_S^T\boldsymbol{x}_S + \boldsymbol{\beta}_E^T\boldsymbol{x}_E + e = \boldsymbol{\beta}_S^T\boldsymbol{x}_S + e \qquad (12.6)$$

where e is an error, Y is the response variable, $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is a $k_S \times 1$ vector and $\boldsymbol{x}_E$ is a $(p - k_S) \times 1$ vector.

Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \mathbf{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of k terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Then

$$Y = \boldsymbol{\beta}_I^T \boldsymbol{x}_I + \boldsymbol{\beta}_O^T \boldsymbol{x}_O + e. \tag{12.7}$$

Definition 12.11. The model $Y = \boldsymbol{\beta}^T \boldsymbol{x} + e$ that uses all of the predictors is called the *full model*. A model $Y = \boldsymbol{\beta}_I^T \boldsymbol{x}_I + e$ that only uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel*. The *sufficient predictor* (SP) is the linear combination of the predictor variables used in the model. Hence the full model has $SP = \boldsymbol{\beta}^T \boldsymbol{x}$ and the submodel has $SP = \boldsymbol{\beta}_I^T \boldsymbol{x}_I$.

Notice that the full model is a submodel. The estimated sufficient predictor (ESP) is $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ and the following remarks suggest that *a submodel I is worth considering if the correlation* $\text{corr}(ESP, ESP(I)) \geq 0.95$. Suppose that S is a subset of I and that model (12.6) holds. Then

$$SP = \boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}_S^T \boldsymbol{x}_S = \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T \boldsymbol{x}_{I/S} + \mathbf{0}^T \boldsymbol{x}_O = \boldsymbol{\beta}_I^T \boldsymbol{x}_I \tag{12.8}$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \mathbf{0}$ and the sample correlation $\text{corr}(\boldsymbol{\beta}^T \boldsymbol{x}_i, \boldsymbol{\beta}_I^T \boldsymbol{x}_{I,i}) = 1.0$ for the population model if $S \subseteq I$.

This subsection proposes a graphical method for evaluating candidate submodels. Let $\hat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$ obtained from the regression of Y on all of the terms $\boldsymbol{x}$. Denote the residuals and fitted values from the *full model* by $r_i = Y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i = Y_i - \hat{Y}_i$ and $\hat{Y}_i = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ respectively. Similarly, let $\hat{\boldsymbol{\beta}}_I$ be the estimate of $\boldsymbol{\beta}_I$ obtained from the regression of Y on $\boldsymbol{x}_I$ and denote the corresponding residuals and fitted values by $r_{I,i} = Y_i - \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ and $\hat{Y}_{I,i} = \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ where $i = 1, ..., n$. Two important summary statistics for a multiple linear regression model are R^2, the proportion of the variability of Y explained by the nontrivial predictors in the model, and the estimate $\hat{\sigma}$ of the error standard deviation σ.

Definition 12.12. The "fit–fit" or *FF plot* is a plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ while a "residual–residual" or *RR plot* is a plot $r_{I,i}$ versus r_i. A *response plot* is a plot of $\hat{Y}_{I,i}$ versus Y_i. A *residual plot* is a plot of $\hat{Y}_{I,i}$ versus $r_{I,i}$.

Many numerical methods such as forward selection, backward elimination, stepwise and all subset methods using the $C_p(I)$ criterion (Jones 1946, Mallows 1973), have been suggested for variable selection. We will use the FF plot, RR plot, the response plots from the full and submodel, and the residual plots (of the fitted values versus the residuals) from the full and submodel. These six plots will contain a great deal of information about the candidate subset provided that Equation (12.6) holds and that a good estimator for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_I$ is used.

For these plots to be useful, it is crucial to verify that a multiple linear regression (MLR) model is appropriate for the full model. **Both the response plot and the residual plot for the full model need to be used to check this assumption**. The plotted points in the response plot should cluster about the *identity line* (that passes through the origin with unit slope) while the plotted points in the residual plot should cluster about the line $r = 0$. Any nonlinear patterns or outliers in either plot suggests that an MLR relationship does not hold. Similarly, before accepting the candidate model, use the response plot and the residual plot from the candidate model to verify that an MLR relationship holds for the response Y and the predictors $\boldsymbol{x}_I$. If the submodel is good, then the residual and response plots of the submodel should be nearly identical to the corresponding plots of the full model. Assume that all submodels contain a constant.

Remark 12.4. To visualize whether a candidate submodel using predictors $\boldsymbol{x}_I$ is good, use the fitted values and residuals from the submodel and full model to make an RR plot of the $r_{I,i}$ versus the r_i and an FF plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$. Add the OLS line to the RR plot and identity line to both plots as visual aids. The subset I is good if the plotted points cluster tightly about the identity line in *both plots*. In particular, the OLS line and the identity line should "nearly coincide" so that it is difficult to tell that the two lines intersect at the origin in the RR plot.

The following notation will be useful. Suppose that all submodels include a constant and that $\boldsymbol{X}$ is the full rank $n \times p$ design matrix for the full model. Let the corresponding vectors of OLS fitted values and residuals be $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$ and $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, respectively. Suppose that $\boldsymbol{X}_I$ is the $n \times k$ design matrix for the candidate submodel and that the corresponding vectors of OLS fitted values and residuals are $\hat{\boldsymbol{Y}}_I = \boldsymbol{X}_I(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y} = \boldsymbol{H}_I\boldsymbol{Y}$ and $\boldsymbol{r}_I = (\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$, respectively. For

multiple linear regression, recall that if the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then the F_I statistic for testing whether the $p-k$ predictor variables in $\boldsymbol{x}_O$ can be deleted is

$$F_I = \frac{SSE(I) - SSE}{(n-k)-(n-p)} / \frac{SSE}{n-p} = \frac{n-p}{p-k}[\frac{SSE(I)}{SSE} - 1]$$

where SSE is the error sum of squares from the full model and SSE(I) is the error sum of squares from the candidate submodel. Then

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k$$

where MSE is the error mean square for the full model. Notice that $C_p(I) \leq 2k$ if and only if $F_I \leq p/(p-k)$. Remark 12.7 below suggests that for subsets I with k terms, submodels with $C_p(I) \leq \min(2k, p)$ are especially interesting.

Olive (2013, proposition 5.1) shows that

$$\text{corr}(r, r_I) = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}, \tag{12.9}$$

and that the plotted points in the FF, RR and response plots will cluster about the identity line. This proposition is a property of OLS and holds even if the data does not follow an MLR model.

Remark 12.5. Note that for large n, $C_p(I) < k$ or $F_I < 1$ will force corr(ESP,ESP(I)) to be high (≥ 0.95). Let d be a lower bound on $\text{corr}(r, r_I)$. If

$$C_p(I) \leq 2k + n\left[\frac{1}{d^2} - 1\right] - \frac{p}{d^2},$$

then $\text{corr}(r, r_I) \geq d$. The simple screen $C_p(I) \leq 2k$ corresponds to

$$d_n \equiv \sqrt{1 - \frac{p}{n}}.$$

To reduce the chance of overfitting, use the screen $C_p(I) \leq \min(2k, p)$.

A standard model selection procedure will often be needed to suggest models. For example, forward selection or backward elimination could be

used. If $p < 30$, Furnival and Wilson (1974) provide a technique for selecting a few candidate subsets after examining all possible subsets.

Rule of thumb 12.3 (assuming that the cost of each predictor is the same): a) After using a numerical method such as forward selection or backward elimination, let I_{min} correspond to the submodel with the smallest C_p. Find the submodel I_I with the fewest number of predictors such that $C_p(I_I) \leq C_p(I_{min}) + 1$. Then I_I is the initial submodel that should be examined. It is possible that $I_I = I_{min}$ or that I_I is the full model. Do not use more predictors than model I_I to avoid overfitting.

b) Models I with fewer predictors than I_I such that $C_p(I) \leq C_p(I_{min}) + 4$ are interesting and should also be examined.

c) Models I with k predictors, including a constant and with fewer predictors than I_I such that $C_p(I_{min}) + 4 < C_p(I) \leq \min(2k, p)$ should be checked but often underfit: important predictors are deleted from the model. Underfit is especially likely to occur if a predictor with one degree of freedom is deleted and the jump in C_p is large, greater than 4, say. (A factor has $c - 1$ degrees of freedom corresponding to the $c - 1$ indicator variables used to define the factor, and usually either all of the indicator variables are kept or deleted by variable selection software.)

d) If there are no models I with fewer predictors than I_I such that $C_p(I) \leq \min(2k, p)$, then model I_I is a good candidate for the best subset found by the numerical procedure.

Variable selection seeks a subset I of the variables to keep in the model. The submodel I will always contain a constant and will have $k - 1$ nontrivial predictors where $1 \leq k \leq p$.

Forward selection starts with a constant $= W_1 = X_1$. Step 1) $k = 2$: compute C_p for all models containing the constant and a single predictor X_i. Keep the predictor $W_2 = X_j$, say, that corresponds to the model with the smallest value of C_p.
Step 2) $k = 3$: Fit all models with $k = 3$ that contain W_1 and W_2. Keep the predictor W_3 that minimizes C_p. ...
Step j) $k = j + 1$: Fit all models with $k = j + 1$ that contains $W_1, W_2, ..., W_j$. Keep the predictor W_{j+1} that minimizes C_p. ...
Step $p - 1$): Fit the full model.

Backward elimination: All models contain a constant $= U_1 = X_1$.

Step 1) $k = p$: Start with the full model that contains $X_1, ..., X_p$. We will also say that the full model contains $U_1, ..., U_p$ where $U_1 = X_1$ but U_i need not equal X_i for $i > 1$.
Step 2) $k = p - 1$: fit each model with $p - 1$ predictors including a constant. Delete the predictor U_p, say, that corresponds to the model with the smallest C_p. Keep $U_1, ..., U_{p-1}$.
Step 3) $k = p-2$: fit each model with $p-2$ predictors and a constant. Delete the predictor U_{p-1} that corresponds to the smallest C_p. Keep $U_1, ..., U_{p-2}$. ...
Step j) $k = p - j + 1$: fit each model with $p - j + 1$ predictors and a constant. Delete the predictor U_{p-j+2} that corresponds to the smallest C_p. Keep $U_1, ..., U_{p-j+1}$. ...
Step $p - 1$) $k = 2$. The current model contains U_1, U_2 and U_3. Fit the model U_1, U_2 and the model U_1, U_3. Assume that model U_1, U_2 minimizes C_p. Then delete U_3 and keep U_1 and U_2.

Assume that the full model has p predictors including a constant and that the submodel I has k predictors including a constant. Assume that the full model has good response and residual plots and that $n > 5p$. Then we would like following properties i) – xi) (roughly in order of decreasing importance) to hold. Often we can not find a submodel where i) – xi) all hold simultaneously. Do not use more predictors than model I_I to avoid overfitting.

Then the submodel I is good if
i) the response and residual plots for the submodel looks like the response and residual plots for the full model.
ii) corr(ESP,ESP(I)) = corr($\hat{\mathrm{Y}}, \hat{\mathrm{Y}}_{\mathrm{I}}$) ≥ 0.95.
iii) The plotted points in the FF plot cluster tightly about the identity line.
iv) Want the p-value ≥ 0.01 for the partial F test that uses I as the reduced model.
v) Want $k \leq n/10$.
vi) The plotted points in the RR plot cluster tightly about the identity line.
vii) Want $R^2(I) > 0.9R^2$ and $R^2(I) > R^2 - 0.07$ ($R^2(I) \leq R^2(full)$ since adding predictors to I does not decrease $R^2(I)$).
viii) Want $C_p(I_{min}) \leq C_p(I) \leq \min(2k, p)$ with no big jumps in C_p (the increase should be less than four) as variables are deleted.
ix) Want hardly any predictors with p-values > 0.05.
x) Want few predictors with p-values between 0.01 and 0.05.
xi) Want MSE(I) to be smaller than or not much larger than the MSE from

the full model.

Example 12.5. The FF and RR plots can be used as a diagnostic for whether a given numerical method is including too many variables. Gladstone (1905-1906) attempts to estimate the *weight* of the human brain (measured in grams after the death of the subject) using simple linear regression with a variety of predictors including *age* in years, *height* in inches, *head height* in mm, *head length* in mm, *head breadth* in mm, *head circumference* in mm, and *cephalic index*. The *sex* (coded as 0 for females and 1 for males) of each subject was also included. The variable *cause* was coded as 1 if the cause of death was acute, 3 if the cause of death was chronic, and coded as 2 otherwise. A variable *ageclass* was coded as 0 if the age was under 20, 1 if the age was between 20 and 45, and as 3 if the age was over 45. Head *size*, the product of the *head length*, *head breadth*, and *head height*, is a volume measurement, hence $(size)^{1/3}$ was also used as a predictor with the same physical dimensions as the other lengths. Thus there are 11 nontrivial predictors and one response, and all models will also contain a constant. Nine cases were deleted because of missing values, leaving 267 cases.

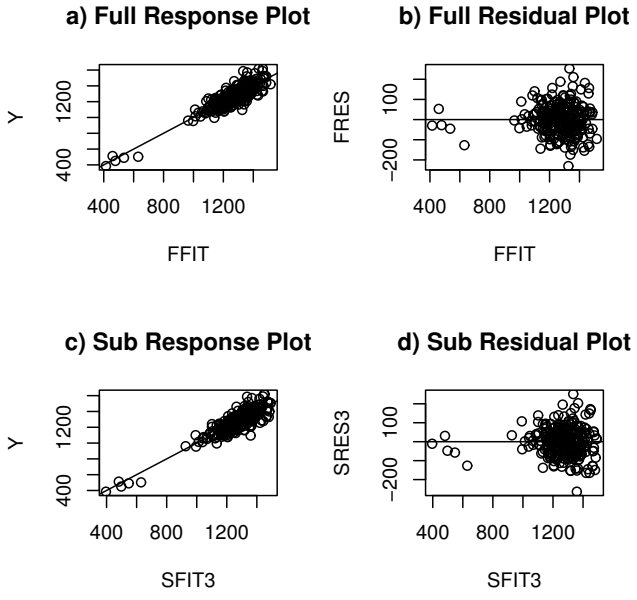

Figure 12.4: Gladstone data: comparison of the full model and the submodel.

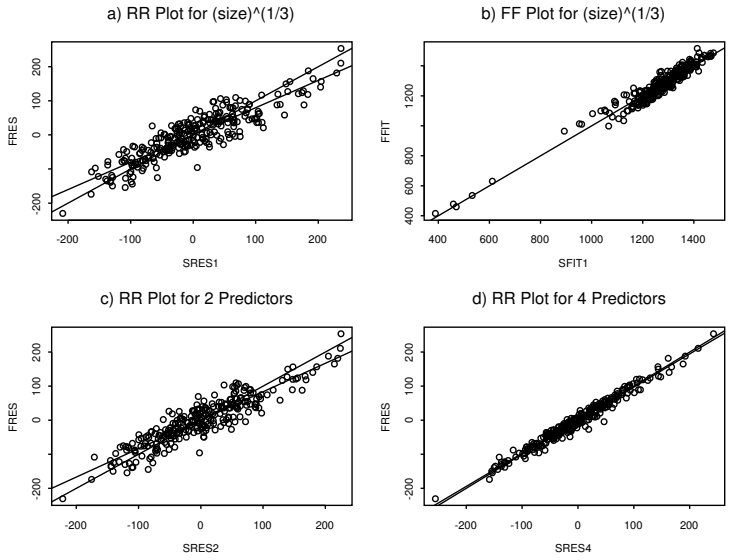

Figure 12.5: Gladstone data: submodels added $(size)^{1/3}$, sex, age and finally $breadth$.

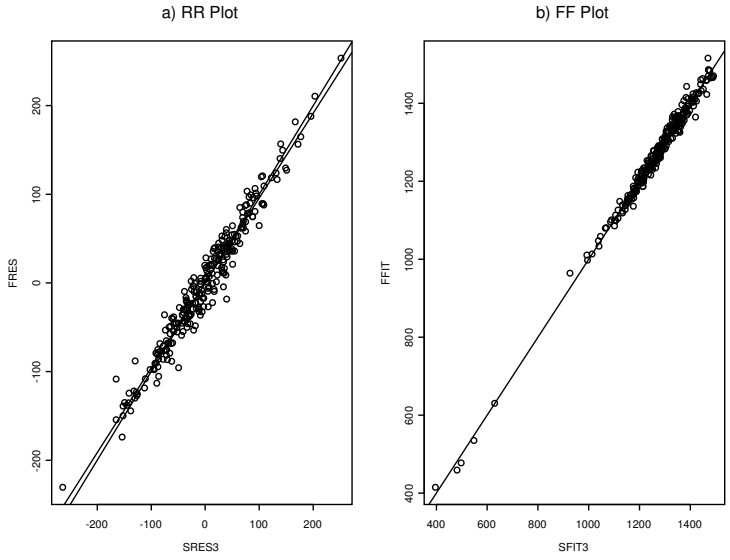

Figure 12.6: Gladstone data with Predictors $(size)^{1/3}$, sex, and age

Figure 12.4 shows the response plots and residual plots for the full model and the final submodel that used a constant, $size^{1/3}$, *age* and *sex*. The five cases separated from the bulk of the data in each of the four plots correspond to five infants. These may be outliers, but the visual separation reflects the small number of infants and toddlers in the data. A purely numerical variable selection procedure would miss this interesting feature of the data. We will first perform variable selection with the entire data set, and then examine the effect of deleting the five cases. Using forward selection and the C_p statistic on the Gladstone data suggests the subset I_5 containing a constant, $(size)^{1/3}$, *age*, *sex*, *breadth*, and *cause* with $C_p(I_5) = 3.199$. The p–values for breadth and cause were 0.03 and 0.04, respectively. The subset I_4 that deletes *cause* has $C_p(I_4) = 5.374$ and the p–value for *breadth* was 0.05. Figure 12.5d shows the RR plot for the subset I_4. Note that the correlation of the plotted points is very high and that the OLS and identity lines nearly coincide.

A scatterplot matrix of the predictors and response suggests that $(size)^{1/3}$ might be the best single predictor. First we regressed $Y =$ *brain weight* on the eleven predictors described above (plus a constant) and obtained the residuals r_i and fitted values $\hat{Y}_i$. Next, we regressed Y on the subset I containing $(size)^{1/3}$ and a constant and obtained the residuals $r_{I,i}$ and the fitted values $\hat{Y}_{I,i}$. Then the RR plot of $r_{I,i}$ versus r_i, and the FF plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ were constructed.

For this model, the correlation in the FF plot (Figure 12.5b) was very high, but in the RR plot the OLS line did not coincide with the identity line (Figure 12.5a). Next *sex* was added to I, but again the OLS and identity lines did not coincide in the RR plot (Figure 12.5c). Hence *age* was added to I. Figure 12.6a shows the RR plot with the OLS and identity lines added. These two lines now nearly coincide, suggesting that a constant plus $(size)^{1/3}$, *sex*, and *age* contains the relevant predictor information. This subset has $C_p(I) = 7.372$, $R_I^2 = 0.80$, and $\hat{\sigma}_I = 74.05$. The full model which used 11 predictors and a constant has $R^2 = 0.81$ and $\hat{\sigma} = 73.58$. Since the C_p criterion suggests adding *breadth* and *cause*, the C_p criterion may be leading to an overfit.

Figure 12.6b shows the FF plot. The five cases in the southwest corner correspond to five infants. Deleting them leads to almost the same conclusions, although the full model now has $R^2 = 0.66$ and $\hat{\sigma} = 73.48$ while the submodel has $R_I^2 = 0.64$ and $\hat{\sigma}_I = 73.89$.

12.3.2 Variable Selection for Multivariate Linear Regression

We still have the full model $\boldsymbol{x} = (\boldsymbol{x}_I^T, \boldsymbol{x}_O)^T$ where $\boldsymbol{x}_I$ is a candidate submodel. It is crucial to verify that a multivariate regression model is appropriate for the full model. **For each of the m response variables, use the response plot and the residual plot for the full model to check this assumption**.

To obtain the candidate subset for multivariate regression, do numerical variable selection such as forward selection or backward elimination for multiple linear regression for each response variable Y_j. Very often predictor variables are highly correlated and often similar sets of predictor variables will be used by each of the m multiple linear regressions. See if there is a pattern to the most important and least important predictors. Try to get rid of predictors that are not needed in any of the m multiple linear regressions. It is better to keep too many predictors than to possible delete a predictor that is needed in at least one of the m multiple linear regression, but want $n > 10p$.

Check the submodel $\boldsymbol{x}_I$ for multivariate linear regression with the FF, RR plots and the response and residual plots for the full model and for the candidate model for each of the m response variables $Y_1, ..., Y_m$. The submodels use Y_{Ij} for $j = 1, ..., m$.

12.4 Prediction

12.4.1 Prediction Intervals for Multiple Linear Regression

This subsection gives estimators for predicting a future or new value Y_f of the vector of response variables given the predictors $\boldsymbol{x}_f$. The following subsection will extend the results to multivariate regression.

Warning: All too often the MLR model seems to fit the data

$$(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_n, Y_n)$$

well, but when new data is collected, a very different MLR model is needed to fit the new data well. In particular, the MLR model seems to fit the data $(\boldsymbol{x}_i, Y_i)$ well for $i = 1, ..., n$, but when the researcher tries to predict Y_f for a

new vector of predictors $\boldsymbol{x}_f$, the prediction is very poor in that $\hat{Y}_f$ is not close to the Y_f actually observed. **Wait until after the MLR model has been shown to make good predictions before claiming that the model gives good predictions!**

There are several reasons why the MLR model may not fit new data well. i) The model building process is usually iterative. Data Z, $w_1, ..., w_k$ is collected. If the model is not linear, then functions of Z are used as a potential response and functions of the w_i as potential predictors. After trial and error, the functions are chosen, resulting in a final MLR model using Y and $x_1, ..., x_p$. Since the same data set was used during the model building process, biases are introduced and the MLR model fits the "training data" better than it fits new data. Suppose that Y, $x_1, ..., x_p$ are specified before collecting data and that the residual and response plots from the resulting MLR model look good. Then predictions from the prespecified model will often be better for predicting new data than a model built from an iterative process.

ii) If $(\boldsymbol{x}_f, Y_f,)$ come from a different population than the population of $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_n, Y_n)$, then prediction for Y_f can be arbitrarily bad.

iii) Even a good MLR model may not provide good predictions for an $\boldsymbol{x}_f$ that is far from the $\boldsymbol{x}_i$ (extrapolation).

iv) The MLR model may be missing important predictors (underfitting).

v) The MLR model may contain unnecessary predictors (overfitting).

Two remedies for i) are a) use previously published studies to select an MLR model before gathering data. b) Do a trial study. Collect some data, build an MLR model using the iterative process. Then use this model as the prespecified model and collect data for the main part of the study. Better yet, do a trial study, specify a model, collect more trial data, improve the specified model and repeat until the latest specified model works well. Unfortunately, trial studies are often too expensive or not possible because the data is difficult to collect. Also, often the population from a published study is quite different from the population of the data collected by the researcher. Then the MLR model from the published study is not adequate.

Definition 12.13. Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ and the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Let $h_i = h_{ii}$ be the ith diagonal element of $\boldsymbol{H}$ for $i = 1, ..., n$. Then h_i is called the ith **leverage** and $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$. Suppose new data is to be collected with predictor vector $\boldsymbol{x}_f$. Then the

leverage of $\boldsymbol{x}_f$ is $h_f = \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f$. **Extrapolation** occurs if $\boldsymbol{x}_f$ is far from the $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$.

Rule of thumb 12.4. Predictions based on extrapolation are not reliable. A rule of thumb is that extrapolation occurs if $h_f > \max(h_1, ..., h_n)$. This rule works best if the predictors are linearly related in that a plot of x_i versus x_j should not have any strong nonlinearities. If there are strong nonlinearities among the predictors, then $\boldsymbol{x}_f$ could be far from the $\boldsymbol{x}_i$ but still have $h_f < \max(h_1, ..., h_n)$.

Example 12.6. Consider predicting $Y = \textit{weight}$ from $x = \textit{height}$ and a constant from data collected on men between 18 and 24 where the minimum height was 57 and the maximum height was 79 inches. The OLS equation was $\hat{Y} = -167 + 4.7x$. If $x = 70$ then $\hat{Y} = -167 + 4.7(70) = 162$ pounds. If $x = 1$ inch, then $\hat{Y} = -167 + 4.7(1) = -162.3$ pounds. It is impossible to have negative weight, but it is also impossible to find a 1 inch man. This MLR model should not be used for x far from the interval $(57, 79)$.

The following theorem is analogous to the central limit theorem and the theory for the t–interval for μ based on $\overline{Y}$ and the sample standard deviation (SD) S_Y. If the data $Y_1, ..., Y_n$ are iid with mean 0 and variance σ^2, then $\overline{Y}$ is asymptotically normal and the t–interval will perform well if the sample size is large enough. The result below suggests that the OLS estimators $\hat{Y}_i$ and $\hat{\boldsymbol{\beta}}$ are good if the sample size is large enough. The condition $\max h_i \to 0$ in probability usually holds if the researcher picked the design matrix $\boldsymbol{X}$ or if the $\boldsymbol{x}_i$ are iid random vectors from a well behaved population. Outliers can cause the condition to fail.

Theorem 12.3: Huber (1981, p. 157-160). Consider the MLR model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ and assume that the errors are independent with zero mean and the same variance: $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$. Also assume that $\max_i(h_1, ..., h_n) \to 0$ in probability as $n \to \infty$. Then

a) $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} \to E(Y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$ in probability for $i = 1, ..., n$ as $n \to \infty$.

b) All of the least squares estimators $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ are asymptotically normal where $\boldsymbol{a}$ is any fixed constant $p \times 1$ vector.

Theorem 12.4. The least squares estimator satisfies $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = o_P(1)$ if

$$\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{X}^T\boldsymbol{e}}{n}\right) = o_P(1).$$

Proof:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{X}^T\boldsymbol{e}}{n}\right).$$

Definition 12.14. A large sample $100(1-\delta)\%$ prediction interval (PI) has the form $(\hat{L}_n, \hat{U}_n)$ where $P(\hat{L}_n < Y_f < \hat{U}_n) \xrightarrow{P} 1-\delta$ as the sample size $n \to \infty$.

The interpretation of a 100 $(1-\delta)\%$ PI for a random variable Y_f is similar to that of a confidence interval (CI). Collect data, then form the PI, and repeat for a total of k times where k trials are independent from the same population. If Y_{fi} is the ith random variable and PI_i is the ith PI, then the probability that $Y_{fi} \in PI_i$ for m of the PIs follows a binomial$(k, \rho = 1-\delta)$ distribution. Hence if 100 95% PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times.

The length of the CI goes to 0 as the sample size n goes to ∞ while the length of the PI converges to some nonzero number L, say. To see this, consider $\boldsymbol{x}_f$ such that the heights Y of women between 18 and 24 is normal with a mean of 66 inches and an SD of 3 inches. A 95% CI for $E(Y|\boldsymbol{x}_f)$ should be centered at about 66 and the length should go to zero as n gets large. But a 95% PI needs to contain about 95% of the heights so the PI should converge to the interval $66 \pm 1.96(3)$. This result follows because if $Y \sim N(66, 9)$ then $P(Y < 66 - 1.96(3)) = P(Y > 66 + 1.96(3)) = 0.025$. In other words, the endpoints of the PI estimate the 97.5 and 2.5 percentiles of the normal distribution. However, the percentiles of a parametric error distribution depend heavily on the parametric distribution and the parametric formulas are violated if the assumed error distribution is incorrect.

Let ξ_δ be the δ percentile of the error e, ie, $P(e \leq \xi_\delta) = \delta$. Let $\hat{\xi}_\delta$ be the sample δ percentile of the residuals. The percentiles of the residuals are consistent estimators, $\hat{\xi}_\delta \xrightarrow{P} \xi_\delta$, under "mild" regularity conditions, and this consistency is the basis for using QQ plots. For multiple linear regression with iid errors with constant variance σ^2, sufficient conditions are $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ and the $\boldsymbol{x}_i$ are bounded in probability. See Olive (2011), Olive and Hawkins (2003), Welsh (1986) and Rousseeuw and Leroy (1987, p. 128).

For many error distributions,

$$E(MSE) = E\left(\sum_{i=1}^{n} \frac{r_i^2}{n-p}\right) = \sigma^2 = E\left(\sum_{i=1}^{n} \frac{e_i^2}{n}\right).$$

This result suggests that

$$\sqrt{\frac{n}{n-p}} r_i \approx e_i.$$

Let

$$a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}}\sqrt{(1+h_f)}. \tag{12.10}$$

Following Olive (2007), a PI is asymptotically optimal if it has the shortest asymptotic length that gives the desired asymptotic coverage. If the error distribution is unimodal, an asymptotically optimal PI can be created by applying the shorth(c) estimator to the residuals where $c = \lceil n(1-\delta) \rceil$ and $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. That is, let $r_{(1)}, ..., r_{(n)}$ be the order statistics of the residuals. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, ..., r_{(n)} - r_{(n-c+1)}$. Let $(r_{(d)}, r_{(d+c-1)}) = (\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2})$ correspond to the interval with the smallest distance. Then the large sample 100 $(1-\delta)$% PI for Y_f is

$$(\hat{Y}_f + a_n \tilde{\xi}_{\delta_1}, \hat{Y}_f + a_n \tilde{\xi}_{1-\delta_2}). \tag{12.11}$$

12.4.2 Prediction Intervals for Multivariate linear Regression

For multivariate linear regression, want to predict a future or new value $\boldsymbol{Y}_f = (Y_{1f}, ..., Y_{mf})^T$ of the vector of m response variables given the vector of predictors $\boldsymbol{x}_f$.

The collection of m prediction intervals $(L_{1n}, U_{1n}), ..., (L_{mn}, U_{mn})$ are *large sample simultaneous conservative* $100(1-\delta)$% *prediction intervals* for Y_{jf} if the m prediction intervals all hold simultaneously, that is all m PIs (L_{jn}, U_{jn}) contain Y_{jf}, with probability $1 - \gamma_n$ where $1 - \gamma_n \to 1 - \gamma \geq 1 - \delta$ as $n \to \infty$.

The *Bonferroni* simultaneous PIs are made by increasing the coverage of a single PI from $1 - \delta$ to $(1 - \delta/m)$. Hence 90% large sample simultaneous PIs will use coverage 0.95 if $m = 2$ and coverage 0.99 if $m = 10$. Let E_j be an event with $P(E_j) = 1 - \delta_j$. Let $\overline{E}_j$ be the compliment of E_j so $P(\overline{E}_j) = \delta_j$.

Then Bonferroni's inequality is

$$P(\cap_{j=1}^{m} E_j) = 1 - P(\overline{\cap_{j=1}^{m} E_j}) = 1 - P(\cup_{j=1}^{m} \overline{E_j}) \geq 1 - \sum_{j=1}^{m} P(\overline{E}_j) =$$

$= 1 - \sum_{j=1}^{m} \delta_j = 1 - \delta$ if $\delta_j = \delta/m$. To use this inequality for simultaneous intervals, let E_j be the event that the jth PI contains Y_{jf}. Then $P(\cap_{j=1}^{m} E_j)$ is the probability that all m PIs contain Y_{jf} for $j = 1, ..., m$.

Let $\tau = \delta/m$. Then the m large sample simultaneous conservative $100(1-\delta)\%$ PIs are

$$(\hat{Y}_{jf} + a_n \tilde{\xi}_{\tau_1}, \hat{Y}_{jf} + a_n \tilde{\xi}_{1-\tau_2}) \tag{12.12}$$

for $j = 1, ..., m$ using Equation (12.11) and residuals $r_{1,j}, ..., r_{n,j}$. That is, make the $100(1-\tau)\%$ PI (12.11) for Y_{jf} for $j = 1, ..., m$ corresponding to the multiple linear regression of the jth response variable Y_j on $\boldsymbol{X}$.

These PIs make no use of the fact that $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, but no parametric distribution for the $\boldsymbol{\epsilon}_i$ is needed. The classical simultaneous prediction region for $\boldsymbol{y}_f$ assumes that the $\boldsymbol{\epsilon}_i$ are iid $N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ and tend to have large undercoverage (are too liberal) when the normality assumption is violated, which is usually the case.

12.4.3 Prediction Regions

Suppose a prediction region for $\boldsymbol{y}_f$ given a vector of predictors $\boldsymbol{x}_f$ is desired. If we had many cases $\boldsymbol{z}_i = \boldsymbol{B}^T \boldsymbol{x}_f + \boldsymbol{\epsilon}_i$, then we could make a prediction region for $\boldsymbol{z}_i$ using Section 5.2. Instead, use $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{B}}^T \boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$. This takes the data cloud of the n residual vectors $\hat{\boldsymbol{\epsilon}}_i$ and centers the cloud at $\hat{\boldsymbol{y}}_f$. Note that $\hat{\boldsymbol{z}}_i = (\boldsymbol{B} - \boldsymbol{B} + \hat{\boldsymbol{B}})^T \boldsymbol{x}_f + (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i) = \boldsymbol{z}_i + (\hat{\boldsymbol{B}} - \boldsymbol{B})^T \boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i = \boldsymbol{z}_i + O_P(n^{-1/2})$. Hence the distances based on the $\boldsymbol{z}_i$ and the distances based on the $\hat{\boldsymbol{z}}_i$ should have the same quantiles, asymptotically.

Theorem 12.5. Suppose $\boldsymbol{y}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i = \hat{\boldsymbol{y}}_i + \hat{\boldsymbol{\epsilon}}_i$ where $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, and where $\boldsymbol{\epsilon}_f$ and the $\boldsymbol{\epsilon}_i$ are iid for $i = 1, ..., n$. Suppose the fitted model produces $\hat{\boldsymbol{y}}_f$ and nonsingular $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Let $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ and

$$D_i^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$$

for $i = 1, ..., n$. Let $q_n = \min(1 - \alpha + 0.05, 1 - \alpha + m/n)$ for $\alpha > 0.1$ and

$$q_n = \min(1 - \alpha/2, 1 - \alpha + 10\alpha m/n), \quad \text{otherwise.}$$

If $q_n < 1 - \alpha + 0.001$, set $q_n = 1 - \alpha$. Let $0 < \alpha < 1$ and $h = D_{(U_n)}$ where $D_{(U_n)}$ is the q_nth sample quantile of the D_i. Consider the nominal $100(1 - \alpha)\%$ prediction region for $\boldsymbol{y}_f$

$$\{\boldsymbol{z} : (\boldsymbol{z} - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{z} - \hat{\boldsymbol{y}}_f) \leq D^2_{(U_n)}\} =$$

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D^2_{(U_n)}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}\}. \tag{12.13}$$

a) Consider the n prediction regions for the data where $(\boldsymbol{y}_{f,i}, \boldsymbol{x}_{f,i}) = (\boldsymbol{y}_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. If the order statistic $D_{(U_n)}$ is unique, then U_n of the n prediction regions contain $\boldsymbol{y}_i$ where $U_n/n \to 1 - \alpha$ as $n \to \infty$.

b) If $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ then (12.13) is a large sample $100(1 - \alpha)\%$ prediction region for $\boldsymbol{y}_f$.

c) If $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and the $\boldsymbol{\epsilon}_i$ come from an elliptically contoured distribution such that the highest density region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\alpha}\}$, then the prediction region (12.13) is asymptotically optimal.

Proof. a) Suppose $(\boldsymbol{x}_f, \boldsymbol{y}_f) = (\boldsymbol{x}_i, \boldsymbol{y}_i)$. Then

$$D^2_{\boldsymbol{y}_i}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i) = \hat{\boldsymbol{\epsilon}}_i^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \hat{\boldsymbol{\epsilon}}_i = D^2_{\hat{\boldsymbol{\epsilon}}_i}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}).$$

Hence $\boldsymbol{y}_i$ is in the ith prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ iff $\hat{\boldsymbol{\epsilon}}_i$ is in prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$, but exactly U_n of the $\hat{\boldsymbol{\epsilon}}_i$ are in the latter region by construction, if $D_{(U_n)}$ is unique. Since $D_{(U_n)}$ is the $(1 - \alpha)$ percentile of the D_i asymptotically, $U_n/n \to 1 - \alpha$.

b) Let $P[D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\alpha}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})] = 1 - \alpha$. Since $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, Proposition 5.1 shows that if $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} (E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ then $D(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Hence the percentiles of the distances also converge in probability, and the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{1-\alpha}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ converges to $1 - \alpha =$ the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\alpha}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$.

c) The asymptotically optimal prediction region is the region with the smallest volume (hence highest density) such that the coverage is $1 - \alpha$, as $n \to \infty$. This region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\alpha}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$ if the

asymptotically optimal region for the $\boldsymbol{\epsilon}_i$ is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\alpha}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$. Hence the result follows by b). $\square$

Multivariate linear regression satisfies Theorem 12.5, and applying a prediction region from Section 5.2 on the $\hat{\boldsymbol{z}}_i$ results in a large sample $100(1-\alpha)\%$ prediction region for $\boldsymbol{y}_f$ given the vector of predictors $\boldsymbol{x}_f$. The prediction region is asymptotically optimal if the $\boldsymbol{\epsilon}_i$ are iid from an $EC_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$ distribution for a large class of elliptically contoured distributions.

To see the above claim, note that if the $\boldsymbol{\epsilon}_i$ are iid from an elliptically contoured distribution with nonsingular covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, then the population asymptotically optimal prediction region is $\{\boldsymbol{y} : D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) < D_{1-\alpha}\}$ where $P(D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) < D_{1-\alpha}) = 1 - \alpha$. For example, if the iid $\boldsymbol{\epsilon}_i \sim N_m(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then $D_{1-\alpha} = \sqrt{\chi^2_{m,1-\alpha}}$. If the error distribution is not elliptically contoured, then the above region still has $100(1-\alpha)\%$ coverage, but prediction regions with smaller volume may exist. In general these quantities need to be estimated. If many errors $\boldsymbol{\epsilon}_i$ were available and $\boldsymbol{B}$ was known, could estimate $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ with $\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T/n$, compute $\boldsymbol{z}_i = \boldsymbol{B}^T\boldsymbol{x}_F + \boldsymbol{\epsilon}_i$ and estimate $D_{1-\alpha}$ with $D_{(\lceil n(1-\alpha)\rceil)}$, the sample $(1-\alpha)$ percentile of the $D_{\boldsymbol{z}_i}$. These quantities are unavailable, but the plug in estimators are $\hat{\boldsymbol{y}}_f = \hat{\boldsymbol{B}}^T\boldsymbol{x}_f$, $\boldsymbol{S}_r = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = (n-1)^{-1}\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T$, $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ and $\hat{D}_{1-\alpha}$, the sample $(1-\alpha)$ percentile of the $D_{\hat{\boldsymbol{z}}_i}$.

Following Section 5.2, suppose $(T, \boldsymbol{C})$ is the sample mean and scaled sample covariance matrix applied to the $\hat{\boldsymbol{z}}_i$ where the multivariate linear regression used least squares. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{y} : (\boldsymbol{y}-T)^T\boldsymbol{C}^{-1}(\boldsymbol{y}-T) \leq h^2\} = \{\boldsymbol{y} : D^2_{\boldsymbol{y}} \leq h^2\} = \{\boldsymbol{y} : D_{\boldsymbol{y}} \leq h\}. \quad (12.14)$$

A future observation (random vector) $\boldsymbol{y}_f$ is in the region (12.14) if $D_{\boldsymbol{y}_f} \leq h$. Set up the prediction region (12.14) using $h = D_{(U_n)}$ as described in Theorem 2.5. Following Section 5.2, this prediction region (12.14) will be called the nonparametric prediction region.

The nonparametic prediction region has some interesting properties. Let $\boldsymbol{S}_r$ be the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$. The sample mean of the residual vectors is $\mathbf{0}$ since least squares was used. Hence the $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ have sample covariance matrix $\boldsymbol{S}_r$, and sample mean $\hat{\boldsymbol{y}}_f$. Hence $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$, and the $D_i(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ are used to compute $D_{(U_n)}$. So if there are 100 different values $(\boldsymbol{x}_{jf}, \boldsymbol{y}_{jf})$ to be predicted, only need to update $\hat{\boldsymbol{y}}_{jf}$

for $j = 1, ..., 100$, do not need to update the covariance matrix $\boldsymbol{S}_r$.

The geometry of the nonparametric region is simple. Let R_r be the nonparametric prediction region applied to the residuals $\hat{\boldsymbol{\epsilon}}_i$, and let (12.14) be the nonparametric prediction region using $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ when the multivariate regression is fit by least squares. Then R_r is a hyperellipsoid with center $\mathbf{0}$, and the nonparametric prediction region (12.14) is the hyperellipsoid R_r translated to have center $\hat{\boldsymbol{y}}_f$.

It is common practice to examine how well the prediction regions work on the data. That is, for $i = 1, ..., n$, set $\boldsymbol{x}_f = \boldsymbol{x}_i$ and see if $\boldsymbol{y}_i$ is in the region with probability near to $1 - \alpha$ with a simulation study. Note that $\hat{\boldsymbol{y}}_f = \hat{\boldsymbol{y}}_i$ if $\boldsymbol{x}_f = \boldsymbol{x}_i$. Simulation is not needed for the nonparametric prediction region (12.14) for the data since the prediction region (12.14) centered at $\hat{\boldsymbol{y}}_i$ contains $\boldsymbol{y}_i$ iff R_r, the prediction region centered at $\mathbf{0}$, contains $\hat{\boldsymbol{\epsilon}}_i$ since $\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i = \hat{\boldsymbol{\epsilon}}_i$. Thus $100q_n\%$ of prediction regions corresponding to the data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ contain $\boldsymbol{y}_i$, and $100q_n\% \rightarrow 100(1 - \alpha)\%$. Hence the prediction regions work well on the data and should work well on $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ similar to the data. Of course simulation should be done for $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ that are not equal to data cases.

This result holds provided that the multivariate linear regression using least squares is such that the sample covariance matrix $\boldsymbol{S}_r$ of the residual vectors is nonsingular, **the multivariate regression model need not be correct**. Hence the coverage at the n data cases $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is very robust to model misspecification. Of course, the prediction regions may be very large if the model is severely misspecified, but severity of misspecification can be checked with the response and residual plots. Coverage can also be arbitrarily bad if there is extrapolation or if $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ comes from a different population than that of the data.

Example 12.5. Consider the Mussel data described in Example 2.2 with response variables $Y_1 = \log(S)$ and $Y_2 = \log(M)$ with predictors $X_2 = L, X_3 = \log(W)$, and $X_4 = height$. Figure 12.7 shows a scatterplot matrix of the data and Figure 12.8 shows a DD plot of the data with multivariate prediction regions added. These plots suggest that the data may come from an elliptically contoured distribution that is not multivariate normal. The semiparametric and nonparametric 90% prediction regions of Section 5.2 consist of the cases below the $RD = 5.86$ line and to the left of the $MD = 4.41$ line. These two lines intersect on a line through the origin that is followed by the plotted points. The parametric MVN prediction region is given by the points below the $RD = 3.33$ line and does not contain enough cases.

Figures 12.9 and 12.10 give the response and residual plots for Y_1 and Y_2.

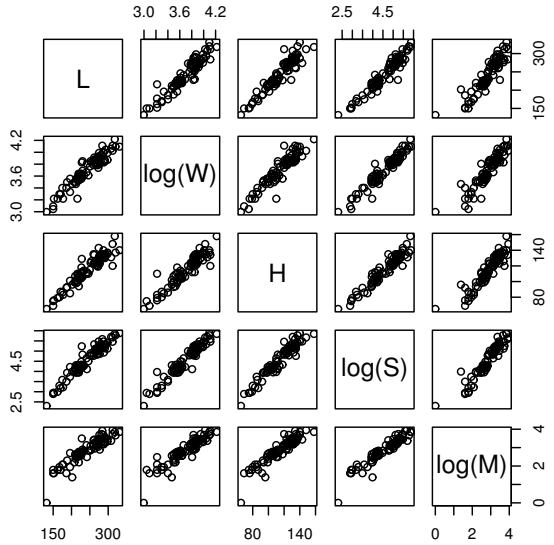

Figure 12.7: Scatterplot Matrix of the Mussels Data.

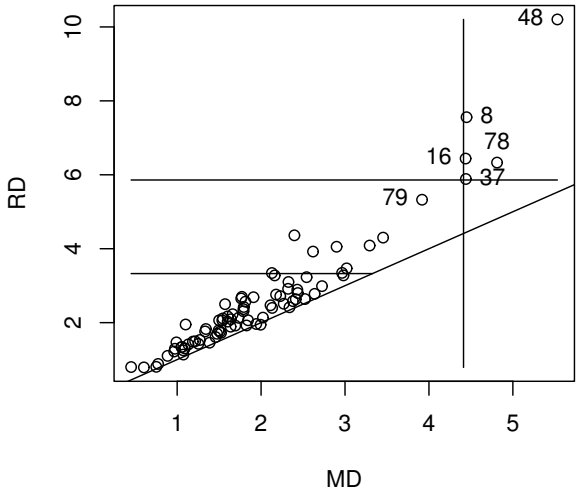

Figure 12.8: DD Plot of the Mussels Data.

For Y_2, cases 8, 25 and 48 are not fit well. A residual vector $\boldsymbol{r} = (\boldsymbol{r} - \boldsymbol{e}) + \boldsymbol{e}$ is a combination of $\boldsymbol{e}$ and a discrepancy $\boldsymbol{r} - \boldsymbol{e}$ that tends to have an approximate multivariate normal distribution. The $\boldsymbol{r} - \boldsymbol{e}$ term can dominate for small to moderate n when $\boldsymbol{e}$ is not multivariate normal, incorrectly suggesting that the distribution of the error $\boldsymbol{e}$ is closer to a multivariate normal distribution than is actually the case. Figure 12.11 shows the DD plot of the residual vectors. The nonparametric prediction region for the residuals consists of the points to the left of the vertical line $MD = 2.27$. Comparing Figure 12.8 and 12.11, the residual distribution is closer to a multivariate normal distribution. Cases 8, 48 and 79 have especially large distances. R code for producing the five figures is shown below.

```
y <- log(mussels)[,4:5]
x <- mussels[,1:3]
x[,2] <- log(x[,2])
z<-cbind(x,y)
pairs(z, labels=c("L","log(W)","H","log(S)","log(M)"))
ddplot4(z)
out <- mltreg(x,y)
ddplot4(out$res)
```

12.5 Testing Hypotheses

This section follows Khattree and Naik (1999, p. 66-67) closely.

Definition 12.15. Assume rank$(\boldsymbol{X}) = p$. The total corrected (for the mean) sum of squares and cross products matrix is

$$\boldsymbol{T} = \boldsymbol{R} + \boldsymbol{W} = \boldsymbol{Z}^T(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T)\boldsymbol{Z}.$$

Note that $\boldsymbol{T}/(n-1)$ is the usual sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}}$ if all n of the $\boldsymbol{y}_i$ are iid so that $\boldsymbol{B} = \boldsymbol{0}$. The regression sum of squares and cross products matrix is

$$\boldsymbol{R} = \boldsymbol{Z}^T[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T]\boldsymbol{Z} = \boldsymbol{Z}^T\boldsymbol{X}\hat{\boldsymbol{B}} - \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{1}\boldsymbol{1}^t\boldsymbol{Z}.$$

The error or residual sum of squares and cross products matrix is

$$\boldsymbol{W}_e = (\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T(\boldsymbol{Z} - \hat{\boldsymbol{Z}}) = \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{Z}^T[\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Z}.$$

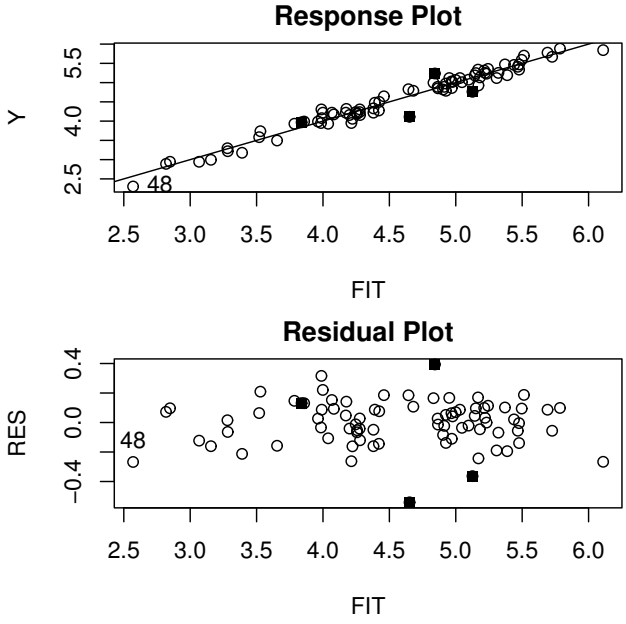

Figure 12.9: Plots for $Y_1 = \log(W)$.

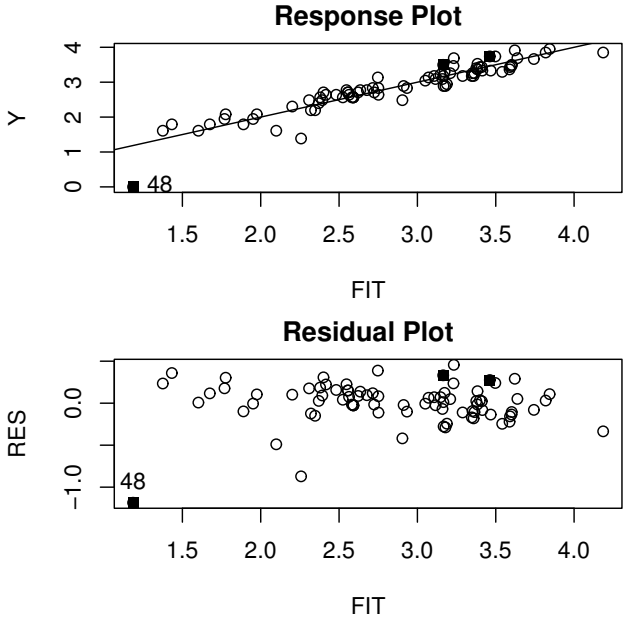

Figure 12.10: Plots for $Y_2 = \log(M)$.

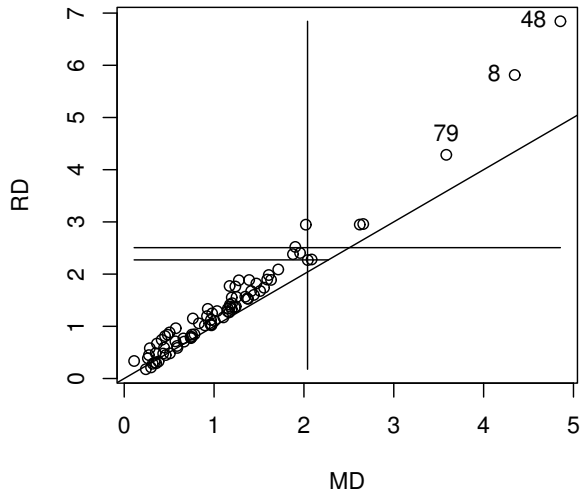

Figure 12.11: DD Plot of the Residual Vectors.

Note that $\boldsymbol{W}_e = \hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}$ and $\boldsymbol{W}_e/(n-p) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$.

Warning: *SAS* output uses $\boldsymbol{E}$ instead of $\boldsymbol{W}_e$.

The MANOVA table is shown below.

Summary MANOVA Table

Source	matrix	df
Regression or Treatment	$\boldsymbol{R}$	$p-1$
Error or Residual	$\boldsymbol{W}_e$	$n-p$
Total (corrected)	$\boldsymbol{T}$	$n-1$

Consider testing a linear hypothesis $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{LB} \neq \boldsymbol{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix. Let $\boldsymbol{H} = \hat{\boldsymbol{B}}\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the ordered eigenvalues of $\boldsymbol{W}_e^{-1}\boldsymbol{H}$. Then there are four commonly used test statistics.

The Wilk's Λ statistic is $\Lambda(\boldsymbol{L}) = |(\boldsymbol{H}+\boldsymbol{W}_e)^{-1}\boldsymbol{W}_e| = |\boldsymbol{W}_e^{-1}\boldsymbol{H}+\boldsymbol{I}|^{-1} =$

$\prod_{i=1}^{m}(1+\lambda_i)^{-1}$.

The Pillai's trace statistic is $V(\boldsymbol{L}) = tr[(\boldsymbol{H}+\boldsymbol{W}_e)^{-1}\boldsymbol{H}] = \sum_{i=1}^{m}\frac{\lambda_i}{1+\lambda_i}$.

The Hotelling-Lawley trace statistic is $U(\boldsymbol{L}) = tr[\boldsymbol{W}_e^{-1}\boldsymbol{H}] = \sum_{i=1}^{m}\lambda_i$.

The Roy's maximum root statistic is $\lambda_{max}(\boldsymbol{L}) = \lambda_1$.

Typically some function of one of the four above statistics is used to get pval, the estimated pvalue. Output often gives the pvals for all four test statistics. Be cautious about inference if the four test statistics do not lead to the same conclusions. Pillai's trace statistic is thought to be the most robust against nonnormality of the $\boldsymbol{\epsilon}_i$.

The four steps of the MANOVA test of linear hypotheses follow.
i) State the hypotheses $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ and $H_1 : \boldsymbol{LB} \neq \boldsymbol{0}$.
ii) Get test statistic from output.
iii) Get pval from output.
iv) State whether you reject H_0 or fail to reject H_0. If pval $\leq \alpha$, reject H_0 and conclude that $\boldsymbol{LB} \neq \boldsymbol{0}$. If pval $> \alpha$, fail to reject H_0 and conclude that $\boldsymbol{LB} = \boldsymbol{0}$ or that there is not enough evidence to conclude that $\boldsymbol{LB} \neq \boldsymbol{0}$. As a textbook convention, use $\alpha = 0.05$ if α is not given.

The MANOVA test of $H_0 : \boldsymbol{B} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{B} \neq \boldsymbol{0}$ is the special case corresponding to $\boldsymbol{L} = \boldsymbol{I}$ and $\boldsymbol{H} = \hat{\boldsymbol{B}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{B}} = \hat{\boldsymbol{Z}}^T\hat{\boldsymbol{Z}}$.

12.6 Justification of the Hotelling Lawley Test

Some notation is needed. Following Henderson and Searle (1979), let matrix $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_p \end{bmatrix}$. Then the vec operator stacks the columns of $\boldsymbol{A}$ on top of one another so

$$vec(\boldsymbol{A}) = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_p \end{pmatrix}.$$

Let $\boldsymbol{A} = ((a_{ij}))$ be an $m \times n$ matrix and $\boldsymbol{B}$ a $p \times q$ matrix. Then the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the $mp \times nq$ matrix

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2n}\boldsymbol{B} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix}.$$

An important fact is that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonsingular square matrices, then $[\boldsymbol{A} \otimes \boldsymbol{B}]^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$.

Consider testing a linear hypothesis $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{LB} \neq \boldsymbol{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix. For now assume the error distribution is multivariate normal $N_p(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon})$. Then

$$vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m \end{pmatrix} \sim N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

where

$$\boldsymbol{C} = \boldsymbol{\Sigma_\epsilon} \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} \sigma_{11}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{12}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{1p}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \sigma_{21}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{22}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{2p}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{p1}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{p2}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{pp}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{bmatrix}.$$

Now let $\boldsymbol{A}$ be a $rm \times pm$ block diagonal matrix: $\boldsymbol{A} = diag(\boldsymbol{L}, ..., \boldsymbol{L})$. Then $\boldsymbol{A}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = vec(\boldsymbol{L}(\hat{\boldsymbol{B}} - \boldsymbol{B})) =$

$$\begin{pmatrix} \boldsymbol{L}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \\ \boldsymbol{L}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \\ \vdots \\ \boldsymbol{L}(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m) \end{pmatrix} \sim N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)$$

where $\boldsymbol{D} = \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T = \boldsymbol{ACA}^T =$

$$\begin{bmatrix} \sigma_{11}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{12}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{1p}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \\ \sigma_{21}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{22}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{2p}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{p1}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{p2}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{pp}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \end{bmatrix}.$$

Under H_0, $vec(\boldsymbol{LB}) = \boldsymbol{A}$ $vec(\boldsymbol{B}) = \boldsymbol{0}$, and

$$vec(\boldsymbol{L}\hat{\boldsymbol{B}}) = \begin{pmatrix} \boldsymbol{L}\hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_m \end{pmatrix} \sim N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T).$$

Hence under H_0,

$$[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \sim \chi^2_{rm},$$

and

$$T = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}. \quad (12.15)$$

A large sample level δ test will reject H_0 if $pval < \delta$ where

$$pval = P(\frac{T}{rm} < F_{rm,n-mp}). \quad (12.16)$$

Since least squares estimators are asymptotically normal, for a large class of distributions,

$$\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = \sqrt{n}\ \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m \end{pmatrix} \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{W})$$

where

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{W}^{-1}.$$

Then under H_0,

$$\sqrt{n}\ \ vec(\boldsymbol{L}\hat{\boldsymbol{B}}) = \sqrt{n}\ \begin{pmatrix} \boldsymbol{L}\hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_m \end{pmatrix} \xrightarrow{D} N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{LWL}^T),$$

and

$$n\ \ [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{LWL}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}.$$

Hence (12.15) holds, and (12.16) gives a large sample level δ test if the least squares estimators are asymptotically normal.

Multivariate analogs of tests for multiple linear regression can be derived with appropriate choice of $\boldsymbol{L}$. Using $\boldsymbol{L} = [\boldsymbol{0} \ \boldsymbol{I}_{p-1}]$ tests whether the nontrivial predictors are needed in the multivariate linear regression model, an analog of the Anova F test. Using $\boldsymbol{L} = [\boldsymbol{0} \ \boldsymbol{I}_k]$ tests whether the last k predictors are needed in the multivariate linear regression model given that the remaining prredictors are in the model, an analog of the partial F test. Using $\boldsymbol{L} = (0, ..., 0, 1, 0, ..., 0)$, a row vector with a 1 in the jth position, tests whether the jth variable is needed in the multivariate linear regression model given that the other $p-1$ predictors are in the model, an analog to the t tests for multiple linear regression. This statistic has the form

$$T_j = \frac{1}{d_j}(\hat{\beta}_{j1}, \hat{\beta}_{j2}, ..., \hat{\beta}_{jm})\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \begin{pmatrix} \hat{\beta}_{j1} \\ \hat{\beta}_{j2} \\ \vdots \\ \hat{\beta}_{jm} \end{pmatrix}$$

where $d_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}$, the jth diagonal entry of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. The statistic T_j could be used for forward selection and backward elimination in variable selection.

12.7 Seemingly Unrelated Regressions

Each response variable in a multivariate linear regression model follows a multiple linear regression model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$. Hence the errors corresponding to the jth response are uncorrelated with variance $\sigma_j^2 = \sigma_{jj}$. Notice that **the same design matrix $\boldsymbol{X}$** of predictors is used for each of the m models, but the response variable vector $\boldsymbol{Y}_j$, coefficient vector $\boldsymbol{\beta}_j$ and error vector $\boldsymbol{e}_j$ change and thus depend on j.

The seemingly related regressions (SUR) model differs from the multivariate linear regression model in that each response model follows a multiple linear regression model $\boldsymbol{Y}_j = \boldsymbol{X}_j\boldsymbol{\beta}_j + \boldsymbol{e}_j$ with a different design matrix $\boldsymbol{X}_j$ and the $\boldsymbol{\beta}_j$ are $k_j \times 1$ vectors. Let $\boldsymbol{x}_{i,j} = (1, x_{2,j}, ..., x_{k_j,j})^T$. Then the ith case in the SUR model is $(Y_{i,1}, ..., Y_{i,m}, x_{2,1}, ..., x_{k_1,1}, x_{2,2}, ..., x_{k_2,2}, ..., x_{2,m}, ..., x_{k_m,m})$. That is, string $\boldsymbol{y}_i$ and the $\boldsymbol{x}_{i,j}$ into a vector, omitting the m ones.

The multivariate linear regression model can be regarded as the special case of the SUR model where all of the design matrices are equal $\boldsymbol{X}_j \equiv \boldsymbol{X}$ for $j = 1, ..., m$, and the SUR model can be regarded as a special case of the multivariate linear regression model where the design matrix $\boldsymbol{X}$ has columns corresponding to the constant $1, x_{2,1}, ..., x_{k_m,m}$. Hence if $k = \sum_{i=1}^{m} k_i$, then $\boldsymbol{X}$ is an $n \times (k - m + 1)$ matrix. Then the $(k - m + 1) \times 1$ vector $\boldsymbol{\beta}_j^* = (\beta_{1,j}, 0, ..., 0, \beta_{2,j}, ..., \beta_{k_j,j}, 0, ..., 0)^T$. Here $\boldsymbol{\beta}_j^*$ is the jth column of $\boldsymbol{B}$, and only k_j of the entries of $\boldsymbol{\beta}_j^*$ are nonzero. Hence most of the entries in $\boldsymbol{B}$ are zeroes.

A competitor of the SUR model would be the multivariate linear regression model where there are no restrictions on $\boldsymbol{B}$, so the columns $\boldsymbol{\beta}_j$ of $\boldsymbol{B}$ are estimated using least squares and $\boldsymbol{X}$. The SUR model says that the $Y_{i,1}, ..., Y_{i,m}$ are correlated, but only $\boldsymbol{x}_{i,j}$ is needed in the model for predicting the $Y_{i,j}$ when $\boldsymbol{x}_{i,1}, ..., \boldsymbol{x}_{i,m}$ are possible vectors of predictors. If this assumption is wrong, then the SUR model could be throwing away a lot of information from relevant predictors.

Definition 12.15. In the *seemingly unrelated regressions model*,

$$\boldsymbol{y}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i = \begin{pmatrix} \boldsymbol{x}_{i,1}^T \boldsymbol{\beta}_1 \\ \boldsymbol{x}_{i,2}^T \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{x}_{i,m}^T \boldsymbol{\beta}_m \end{pmatrix} + \begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \\ \vdots \\ \epsilon_{i,m} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_{i,1}^T \hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{x}_{i,2}^T \hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \boldsymbol{x}_{i,m}^T \hat{\boldsymbol{\beta}}_m \end{pmatrix} + \begin{pmatrix} \hat{\epsilon}_{i,1} \\ \hat{\epsilon}_{i,2} \\ \vdots \\ \hat{\epsilon}_{i,m} \end{pmatrix}$$

$= \hat{\boldsymbol{y}}_i + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$, where $\text{Cov}(\boldsymbol{\epsilon}_i) \equiv \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ is $m \times m$ and $E(\boldsymbol{\epsilon}_i) \equiv \boldsymbol{0}$. Here $\boldsymbol{x}_{i,j}$, $\boldsymbol{\beta}_j$ and $\hat{\boldsymbol{\beta}}_j$ are $k_j \times 1$ vectors where $\sum_{j=1}^{m} k_j = k$, and $\boldsymbol{y}_i = (y_{i1}, ..., y_{im})^T$.

There are several ways to estimate the $\hat{\boldsymbol{\beta}}_j$. First, estimate $\hat{\boldsymbol{\beta}}_j$ using least squares on the m multiple linear regression models $\boldsymbol{Y}_j = \boldsymbol{X}_j \boldsymbol{\beta}_j + \boldsymbol{e}_j$. This method should be equivalent to using the multivariate regression model where the $\boldsymbol{\beta}_j^*$ are the columns of $\boldsymbol{B}$ and the nonzero entries of $\hat{\boldsymbol{\beta}}_j^*$ are collected into the $k_j \times 1$ vectors $\hat{\boldsymbol{\beta}}_j$. Another method uses the seemingly unrelated regressions estimator (SURE) which uses the multivariate linear regression estimator as an initial estimator, and then uses generalized least squares. See Press (2005, § 8.5). In the discussion that follows, $\hat{\boldsymbol{\beta}}$ will be the SUR estimator which is thought to be more efficient than the alternatives. See White (1984, p. 166-171) for large sample theory of the SUR estimator.

Model checking and prediction for the SUR model is very similar to that for the multivariate regression model, but use the fitted values and residuals from the SUR model.

1) Make the m response and residual plots, and make the DD plot of the $\hat{\boldsymbol{\epsilon}}_i$.

2) Transformation plots and variable selection can be done using least squares on each of the m multiple linear regression models $\boldsymbol{Y}_j = \boldsymbol{X}_j = \boldsymbol{e}_j$ for $j = 1, ..., m$.

3) Simultaneous prediction intervals using (12.11) and (12.12) can be made using either least squares fits for each of the m models or using the fitted values and residuals from the SUR model.

4) A prediction region for $\boldsymbol{y}_f$ is made as in Section 12.4.3 using $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$ where $\hat{\boldsymbol{y}}_f = (\boldsymbol{x}_{f,1}^T \hat{\boldsymbol{\beta}}_1, ..., \boldsymbol{x}_{f,m}^T \hat{\boldsymbol{\beta}}_m)^T$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ and the $\hat{\boldsymbol{\beta}}_j$ are the SUR estimators.

```
mltreg(x,y,indices=c(3,4))

$partial
      partialF      Pval
[1,] 0.2001622 0.9349877

$Ftable
                Fj      pvals
[1,]   4.35326807 0.02870083
[2,] 600.57002201 0.00000000
[3,]   0.08819810 0.91597268
[4,]   0.06531531 0.93699302

$MANOVA
     MANOVAF         pval
[1,] 295.071 1.110223e-16
```

Example 12.2. The above output is for the Hebbler (1847) data from the the 1843 Prussia census. Sometimes if the wife or husband was not at the household, then s/he would not be counted. Y_1 = number of married civilian men in the district, Y_2 = number of women married to civilians in the district, x_2 = population of the district in 1843, x_3 = number of married

military men in the district, x_4 = number of women married to military men in the district. The reduced model deletes x_3 and x_4.

a) Do the MANOVA F test.

b) Do the F_2 test.

c) Do the F_4 test.

d) Do an appropriate 4 step test for the reduced model that deletes x_3 and x_4.

e) The output for the reduced model that deletes x_1 and x_2 is shown below. Do an appropriate 4 step test.

```
$partial
     partialF Pval
[1,] 569.6429    0
```

12.8 Summary

1) The multivariate linear regression model is a special case of the multivariate linear model where at least one predictor variable X_j is continuous. The MANOVA model is a multivariate linear model where all of the predictors are categorical variables so the X_j are coded and are often indicator variables.

2) The **multivariate linear regression model** $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $X_1, X_2, ..., X_p$. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$. The constant $x_{i1} = 1$ is in the model, and is often omitted from the case and the data matrix. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = ((\sigma_{ij}))$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$. Then $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$.

3) Each response variable in a multivariate linear regression model follows a univariate linear regression model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$.

4) For each variable Y_k make a response plot of $\hat{Y}_{ik}$ versus Y_{ik} and a residual plot of $\hat{Y}_{ik}$ versus $r_{ik} = Y_{ik} - \hat{Y}_{ik}$. If the multivariate linear regression

model is appropriate, then the plotted points should cluster about the identity line in each of the m response plots. If outliers are present or if the plot is not linear, then the current model or data need to be changed or corrected. If the model is good, then the each of the m residual plots should be ellipsoidal with no trend and should be centered about the $r = 0$ line. There should not be any pattern in the residual plot: as a narrow vertical strip is moved from left to right, the behavior of the residuals within the strip should show little change. Outliers and patterns such as curvature or a fan shaped plot are bad.

5) Make a scatterplot matrix of $Y_1, ..., Y_m$ and of the continuous predictors. Use power transformations to remove strong nonlinearities.

6) Consider testing $\boldsymbol{LB} = \boldsymbol{0}$ where $\boldsymbol{L}$ is a $r \times p$ full rank matrix. Let $\boldsymbol{W}_e = \hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}$ and $\boldsymbol{W}_e/(n-p) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Let $\boldsymbol{H} = \hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the ordered eigenvalues of $\boldsymbol{W}_e^{-1}\boldsymbol{H}$. Then there are four commonly used test statistics.

The Wilk's Λ statistic is $\Lambda(\boldsymbol{L}) = |(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{W}_e| = |\boldsymbol{W}_e^{-1}\boldsymbol{H} + \boldsymbol{I}|^{-1} = \prod_{i=1}^{m}(1+\lambda_i)^{-1}$.

The Pillai's trace statistic is $V(\boldsymbol{L}) = tr[(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{H}] = \sum_{i=1}^{m}\frac{\lambda_i}{1+\lambda_i}$.

The Hotelling-Lawley trace statistic is $U(\boldsymbol{L}) = tr[\boldsymbol{W}_e^{-1}\boldsymbol{H}] = \sum_{i=1}^{m}\lambda_i = \frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$.

The Roy's maximum root statistic is $\lambda_{max}(\boldsymbol{L}) = \lambda_1$.

7) Under regularity conditions, $-[n-p+1-0.5(m-r+3)]\log(\Lambda(\boldsymbol{L})) \xrightarrow{D} \chi^2_{rm}$,
$(n-p)V(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, $\quad (n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, and if $h = \max(r, m)$,

$$\frac{n-p-h+r}{h}\lambda_{max}(\boldsymbol{L}) \approx F(h, n-p-h+r).$$

The Hotelling Lawley statistic is robust against nonnormality.

8) For the Wilk's Lambda test,

$$pval = P\left(\frac{-[n-p+1-0.5(m-r+3)]}{rm} \quad \log(\Lambda(\boldsymbol{L})) < F_{rm,n-rm}\right).$$

For the Pillai's trace test, $pval = P\left(\frac{n-p}{rm} \ \ V(\boldsymbol{L}) < F_{rm,n-rm}\right)$.

For the Hotelling Lawley trace test, $pval = P\left(\frac{n-p}{rm} \ \ U(\boldsymbol{L}) < F_{rm,n-rm}\right)$.

The above three tests are large sample tests, P(reject $H_0|H_0$ is true) $\rightarrow \alpha$ as $n \rightarrow \infty$, under regularity conditions.

For the Roy's largest root test, use
$pval = P\left(\frac{n-p-h+r}{h} \ \lambda_{max}(\boldsymbol{L}) < F_{h,n-p-h+r}\right)$.

The F statistic is an upper bound on the F statistic that provides a lower bound on the nominal level of significance, α, under regularity conditions.

9) The 4 step MANOVA F test of hypotheses uses $\boldsymbol{L} = [\boldsymbol{0} \ \ \boldsymbol{I}_{p-1}]$:
i) State the hypotheses H_0: the nontrivial predictors are not needed in the mreg model H_1: at least one of the nontrivial predictors is needed
ii) Find the test statistic F_o from output.
iii) Find the pval from output.
iv) If pval $< \alpha$, reject H_0. If pval $\geq \alpha$, fail to reject H_0. If H_0 is rejected, conclude that there is a mreg relationship between the response variables $Y_1, ..., Y_m$ and the predictors X_2, ..., X_p. If you fail to reject H_0, conclude that there is a not a mreg relationship between $Y_1, ..., Y_m$ and the predictors X_2, ..., X_p. (Get the variable names from the story problem.)

10) The 4 step F_j test of hypotheses uses $\boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$ where the 1 is in the jth position. Let $\boldsymbol{b}_j^T$ be the jth row of $\boldsymbol{B}$. i) State the hypotheses H_0 :
$bb_j^T = \boldsymbol{0} \quad H_1 : \ \ \boldsymbol{b}_j^T \neq \boldsymbol{0}$
ii) Find the test statistic F_j from output.
iii) Find pval from output.
iv) If pval $< \alpha$, reject H_0. If pval $\geq \alpha$, fail to reject H_0. Give a nontechnical sentence restating your conclusion in terms of the story problem. If H_0 is rejected, then conclude that X_j is needed in the mreg model for $Y_1, ..., Y_m$ given that the other predictors are in the model. If you fail to reject H_0, then conclude that X_j is not needed in the mreg model for $Y_1, ..., Y_m$ given that the other predictors are in the model. (Get the variable names from the story problem.)

11) The 4 step **MANOVA partial F test** of hypotheses has a full model using all of the variables and a reduced model where r of the variables are deleted. The ith row of $\boldsymbol{L}$ has a 1 in the position corresponding to the ith variable to be deleted. Omitting the jth variable corresponds to the F_j test

while omitting variables $X_2, ..., X_p$ corresponds to the MANOVA F test.
i) State the hypotheses H_0: the reduced model is good H_1: use the full model.
ii) Find the test statistic F_R from output.
iii) Find the pval from output.
iv) If pval $< \alpha$, reject H_0 and conclude that the full model should be used. If pval $\geq \alpha$, fail to reject H_0 and conclude that the reduced model is good.

12) The 4 step MANOVA F test should reject H_0 if the response and residual plots look good, n is large enough and at least one response plot does not look like the corresponding residual plot. A response plot for Y_j will look like a residual plot if the identity line appears almost horizontal, hence the range of $\hat{Y}_j$ is small.

13) The *mpack* function `mltreg` produces the m response and residual plots, gives $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$, the MANOVA partial F test statistic and pval corresponding to the reduced model that leaves out the variables given by indices (so X_2 and X_4 in the output below with $F = 0.77$ and pval $= 0.614$), F_j and the pval for the F_j test for variables 1, 2, ..., p (where $p = 4$ in the output below so $F_2 = 1.51$ with pval $= 0.284$) and F_0 and pval for the MANOVA F test (in the output below $F_0 = 3.15$ and pval$= 0.06$). The command `out <- mltreg(x,y,indices=c(2))` would produce a MANOVA partial F test corresponding to the F_2 test while the command `out <- mltreg(x,y,indices=c(2,3,4))` would produce a MANOVA partial F test corresponding to the MANOVA F test for a data set with $p = 4$ predictor variables. The Hotelling Lawley trace statistic is used in the tests.

```
out <- mltreg(x,y,indices=c(2,4))

$Bhat
            [,1]         [,2]         [,3]
[1,] 47.96841291 623.2817463 179.8867890
[2,]  0.07884384   0.7276600  -0.5378649
[3,] -1.45584256 -17.3872206   0.2337900
[4,] -0.01895002   0.1393189  -0.3885967

$Covhat
          [,1]      [,2]     [,3]
[1,]  21.91591  123.2557  132.339
[2,] 123.25566 2619.4996 2145.780
```

```
[3,] 132.33902 2145.7797 2954.082

$partial
      partialF      Pval
[1,] 0.7703294 0.6141573

$Ftable
             Fj      pvals
[1,] 6.30355375 0.01677169
[2,] 1.51013090 0.28449166
[3,] 5.61329324 0.02279833
[4,] 0.06482555 0.97701447

$MANOVA
      MANOVAF       pval
[1,] 3.150118 0.06038742
```

14) Given $\hat{\boldsymbol{B}} = [\hat{\boldsymbol{\beta}}_1 \;\; \hat{\boldsymbol{\beta}}_2 \;\; \cdots \;\; \hat{\boldsymbol{\beta}}_m]$ and $\boldsymbol{x}_f$, find $\hat{\boldsymbol{y}}_f = (\hat{y}_1, ..., \hat{y}_m)^T$ where $\hat{y}_i = \hat{\boldsymbol{\beta}}_i^T \boldsymbol{x}_f$.

15) $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \dfrac{\hat{\boldsymbol{E}}^T \hat{\boldsymbol{E}}}{n-p} = \dfrac{1}{n-p} \sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^T$ while the sample covariance matrix of the residuals is $\boldsymbol{S}_r = \dfrac{n-p}{n-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \dfrac{\hat{\boldsymbol{E}}^T \hat{\boldsymbol{E}}}{n-1}$. Both $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ and $\boldsymbol{S}_r$ are $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ for a large class of error distributions for $\boldsymbol{\epsilon}_i$.

16) The $100(1-\alpha)\%$ nonparametric prediction region for $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ is the nonparametric prediction region from $\oint$ 5.2 applied to $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i = \hat{\boldsymbol{B}}^T \boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$. This takes the data cloud of the n residual vectors $\hat{\boldsymbol{\epsilon}}_i$ and centers the cloud at $\hat{\boldsymbol{y}}_f$. Let

$$D_i^2(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T \boldsymbol{S}_r^{-1} (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$$

for $i = 1, ..., n$. Let $q_n = \min(1 - \alpha + 0.05, 1 - \alpha + m/n)$ for $\alpha > 0.1$ and

$$q_n = \min(1 - \alpha/2, 1 - \alpha + 10\alpha m/n), \quad \text{otherwise.}$$

If $q_n < 1 - \alpha + 0.001$, set $q_n = 1 - \alpha$. Let $0 < \alpha < 1$ and $h = D_{(U_n)}$ where $D_{(U_n)}$ is the q_nth sample quantile of the D_i. The $100(1-\alpha)\%$ nonparametric

prediction region for $\boldsymbol{y}_f$ is

$$\{\boldsymbol{z} : (\boldsymbol{z} - \hat{\boldsymbol{y}}_f)^T \boldsymbol{S}_r^{-1} (\boldsymbol{z} - \hat{\boldsymbol{y}}_f) \leq D^2_{(U_n)}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}\}.$$

a) Consider the n prediction regions for the data where $(\boldsymbol{y}_{f,i}, \boldsymbol{x}_{f,i}) = (\boldsymbol{y}_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. If the order statistic $D_{(U_n)}$ is unique, then U_n of the n prediction regions contain $\boldsymbol{y}_i$ where $U_n/n \to 1 - \alpha$ as $n \to \infty$.

b) If $(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ then the nonparamtric prediction region is a large sample $100(1 - \alpha)\%$ prediction region for $\boldsymbol{y}_f$.

c) If $(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and the $\boldsymbol{\epsilon}_i$ come from an elliptically contoured distribution such that the highest density region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\alpha}\}$, then the nonparametric prediction region is asymptotically optimal.

17) On the DD plot for the residuals, the cases to the left of the vertical line correspond to cases that would have $\boldsymbol{y}_f = \boldsymbol{y}_i$ in the nonparametric prediction region if $\boldsymbol{x}_f = \boldsymbol{x}_i$ while the cases to the right of the line would not have $\boldsymbol{y}_f = \boldsymbol{y}_i$ in the nonparametric prediction region.

18) The DD plot for the residuals is interpreted almost exactly as a DD plot for iid multivariate data is interpreted. Plotted points clustering about the identity line suggests that the $\boldsymbol{\epsilon}_i$ may be iid from a multivariate normal distribution while plotted points that lie above the identity line but cluster about a line through the origin with slope greater than 1 suggests that the $\boldsymbol{\epsilon}_i$ may be iid from an elliptically contoured distribution that is not MVN. The semiparametric and parametric MVN prediction regions correspond to horizontal lines on the DD plot. Robust distances have not been shown to be consistent estimators of the population distances, but are useful for a graphical diagnostic.

19) A robust multivariate linear regression method replaces least squares with the hbreg estimator. The probability that the robust estimator equals the least squares estimator goes to 1 as $n \to \infty$ for a large class of error distributions. Hence the hypothesis tests and nonparametric prediction regions for the classical method can be applied to the robust method. The entries of $\hat{\boldsymbol{B}}$ are hard to drive to $\pm\infty$ for the robust estimator, and the residuals corresponding to outliers are often large. Since the residuals are used to compute $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$, the tests of hypothesis based on the robust estimator are not robust to the presence of outliers. But the robust estimator and classical estimator

tend to give different response and residual plots and test statistics when outliers are present.

12.9 Complements

The least squares estimator $\hat{\boldsymbol{\beta}}$ is a good estimator of $\boldsymbol{\beta}$ under very mild conditions by Theorem 12.3; however, Theorem 12.3 assumes that the model is known before gathering data. If variable selection and response transformation are performed to build a model, then the estimators are biased and results for inference fail to hold in that pvalues and coverage of confidence and prediction intervals will be wrong. See, for example, Berk (1978), Copas (1983), Miller (1984) and Rencher and Pun (1980). Hence it is a good idea to do a pilot study to suggest which transformations and variables to use. Then do a larger study without using variable selection and response transformations.

Cook and Olive (2001) and Olive (2004b, 2013) discuss response plots and transformation plots. Cook and Setodji (2003) use the FF plot while Wilcox (2009) has a robust method for multivariate regression. Su and Cook (2012) give an interesting alternative to least squares. Prediction regions for this method could be made following Section 12.4.3.

Khattree and Naik (1999, p. 91-98) discuss testing $H_0 : \boldsymbol{LBM} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{LBM} \neq \boldsymbol{0}$ where $\boldsymbol{M} = \boldsymbol{I}$ gives a linear test of hypotheses.

12.10 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

12.1*. Refer to the alternative form of the Hotelling Lawley test statistic. Let

$$T(\boldsymbol{W}) = n \ \ [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{LWL}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})].$$

Let

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} = \hat{\boldsymbol{W}}^{-1}.$$

Show $T(\hat{\boldsymbol{W}}) = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$.

12.2. Refer to the alternative form of the Hotelling Lawley test statistic. Let $T = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$. Let $\boldsymbol{L} = \boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$ have a 1 in the jth position. Let $\hat{\boldsymbol{b}}_j^T = \boldsymbol{L}\hat{\boldsymbol{B}}$ be the jth row of $\hat{\boldsymbol{B}}$. Let $d_j = \boldsymbol{L}_j(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}_j^T = (\boldsymbol{X}^T\boldsymbol{X})_{jj}^{-1}$, the jth diagonal entry of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Then $T_j = \frac{1}{d_j}\hat{\boldsymbol{b}}_j^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{b}}_j$. The Hotelling Lawley statistic $U = tr([(n-p)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}]^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}])$. Hence if $\boldsymbol{L} = \boldsymbol{L}_j$, then $U_j = \frac{1}{d_j(n-p)}tr(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{b}}_j\hat{\boldsymbol{b}}_j^T)$.

Using $tr(\boldsymbol{ABC}) = tr(\boldsymbol{CAB})$ and $tr(a) = a$ for scalar a, show the $(n-p)U_j = T_j$.

12.3. Refer to the alternative form of the Hotelling Lawley test statistic. Using the Searle (1982, p. 333) identity
$tr(\boldsymbol{AG}^T\boldsymbol{DGC}) = [vec(\boldsymbol{G})]^T[\boldsymbol{CA} \otimes \boldsymbol{D}^T][vec(\boldsymbol{G})]$, show
$(n-p)U(\boldsymbol{L}) = tr[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}]$
$= [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$ by identifying $\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{D}$, and $\boldsymbol{C}$.

```
$Ftable
            Fj        pvals
[1,] 82.147221 0.000000e+00
[2,] 58.448961 0.000000e+00
[3,] 15.700326 4.258563e-09
[4,]  9.072358 1.281220e-05
[5,] 45.364862 0.000000e+00

$MANOVA
      MANOVAF pval
[1,] 67.80145    0
```

12.4. The above output is for the R Seatbelts data set where $Y_1 = drivers$ = number of drivers killed or seriously injured, $Y_2 = front$ = number of front seat passengers killed or seriously injured, and $Y_3 = back$ = number of back seat passengers killed or seriously injured. The predictors were $x_2 = kms$ = distance driven, $x_3 = price$ = petrol price, $x_4 = van$ = number of van drivers killed, and $x_5 = law = 0$ if the law was in effect that month and 1 otherwise. The data consists of 192 monthly totals in Great Britain from

January 1969 to December 1984, and the compulsory wearing of seat belts law was introduced in February 1983.

a) Do the MANOVA F test.

b) Do the F_4 test.

12.5. a) Sketch a DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ for the multivariate linear regression model if the error vectors $\boldsymbol{\epsilon}_i$ are iid from a multivariate normal distribution. b) Does the DD plot change if the one way MANOVA model is used instead of the multivariate linear regression model?

```
y<-USJudgeRatings[,c(9,10,12)]
x<-USJudgeRatings[,-c(9,10,12)]
mltreg(x,y,indices=c(2,5,6,7,8))
$partial
     partialF      Pval
[1,] 1.649415 0.1855314

$MANOVA
      MANOVAF         pval
[1,] 340.1018 1.121325e-14
```

12.6. The above output is for the R judge ratings data set consisting of lawyer ratings for $n = 43$ judges. $Y_1 = oral$ = sound oral rulings, $Y_2 = writ$ = sound written rulings, and $Y_3 = rten$ = worthy of retention. The predictors were $x_2 = cont$ = number of contacts of lawyer with judge, $x_3 = intg$ = judicial integrity, $x_4 = dmnr$ = demeanor, $x_5 = dilg$ = diligence, $x_6 = cfmg$ = case flow managing, $x_7 = deci$ = prompt decisions, $x_8 = prep$ = preparation for trial, $x_9 = fami$ = familiarity with law, and $x_{10} = phys$ = physical ability.

a) Do the MANOVA F test.

b) Do the MANOVA partial F test for the reduced model that deletes x_2, x_5, x_6, x_7 and x_8.

12.7. Let $\boldsymbol{\beta}_i$ be $p \times 1$ and suppose

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \end{pmatrix} \sim N_{2p}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} \sigma_{11}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{12}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \sigma_{21}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{22}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{bmatrix} \right).$$

Find the distribution of

$$[\boldsymbol{L} \;\; \mathbf{0}] \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \end{pmatrix} = \boldsymbol{L}\hat{\boldsymbol{\beta}}_1$$

where $\boldsymbol{L}\boldsymbol{\beta}_1 = \mathbf{0}$ and $\boldsymbol{L}$ is $r \times p$ with $r \leq p$. Simplify.

R/Splus Problems

Warning: Use the command *source("G:/mpack.txt")* **to download the programs. See Preface or Section 15.2.** Typing the name of the `mpack` function, eg *ddplot*, will display the code for the function. Use the `args` command, eg *args(ddplot)*, to display the needed arguments for the function.

12.8. This problem examines multivariate linear regression on the Cook and Weisberg (1999a) mussels data with $Y_1 = \log(S)$ and $Y_2 = \log(M)$ where S is the shell mass and M is the muscle mass. The predictors are $X_2 = L$, $X_3 = \log(W)$ and $X_4 = H$: the shell length, log(width) and height.

a) The R command for this part make the response and residual plots for each of the three variables. Click the rightmost mouse button and highlight *Stop* to advance the plot. When you have the response and residual plots for one variable on the screen, copy and paste the two plots into *Word*. Do this two times, once for each response variable. The plotted points fall in roughly evenly populated bands about the identity or $r = 0$ line.

b) Copy and paste the output produced from the R command for this part from \$partial on. This gives the output needed to do the MANOVA F test, MANOVA partial F test and the F_j tests.

c) The R command for this plot makes a DD plot of the residuals and adds the lines corresponding to the three prediction regions of Section 5.2. The robust cutoff is larger than the semiparametric cutoff. Place the plot in *Word*. Do the residuals appear to follow a multivariate normal distribution?

d) Do the MANOVA partial F test where the reduced model deletes X_3 and X_4.

e) Do the F_2 test.

f) Do the MANOVA F test.

12.9. This problem examines multivariate linear regression on SAS Institute (1985, p. 146) Fitness Club Data data with $Y_1 = chinups$, $Y_2 = situps$ and $Y_3 = jumps$. The predictors are $X_2 = weight$, $X_3 = waist$ and $X_4 = pulse$.

a) The *R* command for this part make the response and residual plots for each of the three variables. Click the rightmost mouse button and highlight *Stop* to advance the plot. When you have the response and residual plots for one variable on the screen, copy and paste the three plots into *Word*. Do this three times, once for each response variable. Are there any outliers?

b) The *R* command for this plot makes a DD plot of the residuals and adds the lines corresponding to the three prediction regions of Section 5.2. The robust cutoff is larger than the semiparametric cutoff. Place the plot in *Word*. Are there any outliers?

12.6. This problem uses the *mpack* function `mregsim` to simulate the Wilk's Lambda test, Pillai's trace test, Hotelling Lawley trace test, and Roy's largest root test for the F_j tests and the MANOVA F test for multivariate linear regression. When `mnull = T` the first row of $\boldsymbol{B}$ is $\mathbf{1}^T$ while the remaining rows are equal to $\mathbf{0}$. Hence the null hypothesis for the MANOVA F test is true. When `mnull = F` the null hypothesis is true for $p = 2$, but false for $p > 2$. Now the first row of $\boldsymbol{B}$ is $\mathbf{1}^T$ and the last row of $\boldsymbol{B}$ is $\mathbf{0}$. If $p > 2$, then the second to last row of $\boldsymbol{B}$ is (1, 0, ..., 0), the third to last row is (1, 1, 0, ..., 0) etcetera as long as the first row is not changed from $\mathbf{1}^T$. First m iid errors $\boldsymbol{z}_i$ are generated such that the m errors are iid with variance σ^2. Then $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{z}_i$ so that $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T = ((\sigma_{ij}))$ where the diagonal entries $\sigma_{ii} = \sigma^2[1 + (m-1)\rho^2]$ and the off diagonal entries $\sigma_{ij} = \sigma^2[2\rho + (m-2)\rho^2]$ where $\rho = 0.10$. Terms like *Wilkcov* give the percentage of times the Wilk's test rejected the $F_1, F_2, ..., F_p$ tests. The \$mancv wcv pcv hlcv rcv fcv output gives the percentage of times the 4 test statistics reject the MANOVA F test. Here hlcov and fcov both correspond to the Hotelling Lawley test using the formulas in problem A).

5000 runs will be used so the simulation will take several minutes. Sample sizes $n = 10\min(m, p), n = 10\max(m, p)$ and $n = 10mp$ were interesting. Want coverage near 0.05 when H_0 is true and coverage close to 1 for good power when H_0 is false. Multivariate normal errors were used in a) and b) below.

a) Copy the coverage parts of the output produced by the *R* commands for this part. Used $n = 20, m = 2, p = 4$. Here H_0 is true except for the F_1 test. Wilk's and Pillai's tests had low coverage < 0.05 when H_0 was false. Roy's test was good for the F_j tests but why was Roy's test bad for the MANOVA F test?

b) Copy the coverage parts of the output produced by the *R* commands

for this part. Used $n = 20, m = 2, p = 4$. Here H_0 is false except for the F_4 test. Which two tests seem to be the best for this part?

12.11 This problem uses the *mpack* function `mpredsim` to simulate the prediction regions for $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ for multivariate regression. With 5000 runs this simulation takes several minutes. The R command for this problem generate iid lognormal errors then subtract the mean producing $\boldsymbol{z}_i$. Then the $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{z}_i$ are generated as in problem D). Used n=100, m=2, and p=4. The nominal coverage of the prediction region is 90%, and 92% of the training data is covered. The ncvr output gives the coverage of the nonparametric region. What was ncvr?

Chapter 13

Clustering

13.1 Introduction

Clustering is used to classify the n cases into k groups. Discriminant analysis is a type of supervised classification while clustering is a type of unsupervised classification.

For k-means clustering, there are 4 steps.

1) Partition the n cases into k initial groups and find the means of each group. Alternatively, choose k initial seed points. These are groups of size 1 so the mean is equal to the seed point.

2) Compute distances between each case and each mean. Assign case to the cluster whose mean is the nearest.

3) Recalculate the mean of each cluster.

4) Go to 2) and repeat until no more reassignments occur.

Two problems with k-means clustering are i) there could be more or less than k clusters, and ii) two initial means could belong to the same cluster. Then the resulting clusters may be poorly differentiated.

Hierarchical clustering also has several steps. A distance is needed. Single linkage (or nearest neighbor) is the minimum distance between cases in cluster i and cases in cluster j. Complete linkage is the maximum distance between cases in cluster i and cases in cluster j. The average distance between clusters is also sometimes used.

1) Start with m = n clusters. Each case forms a cluster. Compute the distance matrix for the n clusters. Let $d_{U,V}$ be the smallest distance. Combine clusters U and V into a single cluster and set $m = n - 1$.

2) Repeat step 1) with the new m. Continue until there is a single cluster.

3) Plot the resulting clusters as a dendogram. Use the dendogram to select k reasonable clusters of cases.

13.2 Complements

Atkinson, Riani and Cerioli (2004, ch. 7) has some interesting ideas.

13.3 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

13.1*.

R/Splus Problems

Warning: Use the command *source("G:/mpack.txt")* **to download the programs. See Preface or Section 15.2.** Typing the name of the `mpack` function, eg *ddplot*, will display the code for the function. Use the `args` command, eg *args(ddplot)*, to display the needed arguments for the function.

Chapter 14

Other Techniques

14.1 Resistant Regression

Ellipsoidal trimming can be used to create resistant multiple linear regression (MLR) estimators. To perform ellipsoidal trimming, an estimator $(T, \boldsymbol{C})$ is computed and used to create the squared Mahalanobis distances D_i^2 for each vector of observed predictors $\boldsymbol{x}_i$. If the ordered distance $D_{(j)}$ is unique, then j of the $\boldsymbol{x}_i$'s are in the ellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq D_{(j)}^2\}. \tag{14.1}$$

The ith case $(Y_i, \boldsymbol{x}_i^T)^T$ is trimmed if $D_i > D_{(j)}$. Then an estimator of $\boldsymbol{\beta}$ is computed from the remaining cases. For example, if $j \approx 0.9n$, then about 10% of the cases are trimmed, and OLS or L_1 could be used on the cases that remain.

Recall that a response plot is a plot of the fitted values $\hat{Y}_i$ versus the response Y_i and is very useful for detecting outliers. If the MLR model holds and the MLR estimator is good, then the plotted points will scatter about the identity line that has unit slope and zero intercept. The identity line is added to the plot as a visual aid, and the vertical deviations from the identity line are equal to the residuals since $Y_i - \hat{Y}_i = r_i$.

The resistant trimmed views estimator combines ellipsoidal trimming and the response plot. First compute $(T, \boldsymbol{C})$, perhaps using the RFCH estimator or the *R/Splus* function `cov.mcd`. Trim the $M\%$ of the cases with the largest Mahalanobis distances, and then compute the MLR estimator $\hat{\boldsymbol{\beta}}_M$ from the

remaining cases. Use $M = 0$, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten response plots of the fitted values $\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_i$ versus y_i using all n cases. (Fewer plots are used for small data sets if $\hat{\boldsymbol{\beta}}_M$ can not be computed for large M.) These plots are called "trimmed views."

Definition 14.1. The trimmed views (TV) estimator $\hat{\boldsymbol{\beta}}_{T,n}$ corresponds to the trimmed view where the bulk of the plotted points follow the identity line with smallest variance function, ignoring any outliers.

Example 14.1. For the Buxton (1920) data, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet! OLS was used on the cases remaining after trimming, and Figure 14.1 shows four trimmed views corresponding to 90%, 70%, 40% and 0% trimming. The OLS TV estimator used 70% trimming since this trimmed view was best. Since the vertical distance from a plotted point to the identity line is equal to the case's residual, the outliers had massive residuals for 90%, 70% and 40% trimming. Notice that the OLS trimmed view with 0% trimming "passed through the outliers" since the cluster of outliers is scattered about the identity line.

The TV estimator $\hat{\boldsymbol{\beta}}_{T,n}$ has good statistical properties if an estimator with good statistical properties is applied to the cases $(\boldsymbol{X}_{M,n}, \boldsymbol{Y}_{M,n})$ that remain after trimming. Candidates include OLS, L_1, Huber's M–estimator, Mallows' GM–estimator or the Wilcoxon rank estimator. See Rousseeuw and Leroy (1987, p. 12-13, 150). The basic idea is that if an estimator with $O_P(n^{-1/2})$ convergence rate is applied to a set of $n_M \propto n$ cases, then the resulting estimator $\hat{\boldsymbol{\beta}}_{M,n}$ also has $O_P(n^{-1/2})$ rate provided that the response Y was not used to select the n_M cases in the set. If $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ for $M = 0, ..., 90$ then $\|\hat{\boldsymbol{\beta}}_{T,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ by Pratt (1959).

Let $\boldsymbol{X}_n = \boldsymbol{X}_{0,n}$ denote the full design matrix. Often when proving asymptotic normality of an MLR estimator $\hat{\boldsymbol{\beta}}_{0,n}$, it is assumed that

$$\frac{\boldsymbol{X}_n^T \boldsymbol{X}_n}{n} \to \boldsymbol{W}^{-1}.$$

If $\hat{\boldsymbol{\beta}}_{0,n}$ has $O_P(n^{-1/2})$ rate and if for big enough n all of the diagonal elements

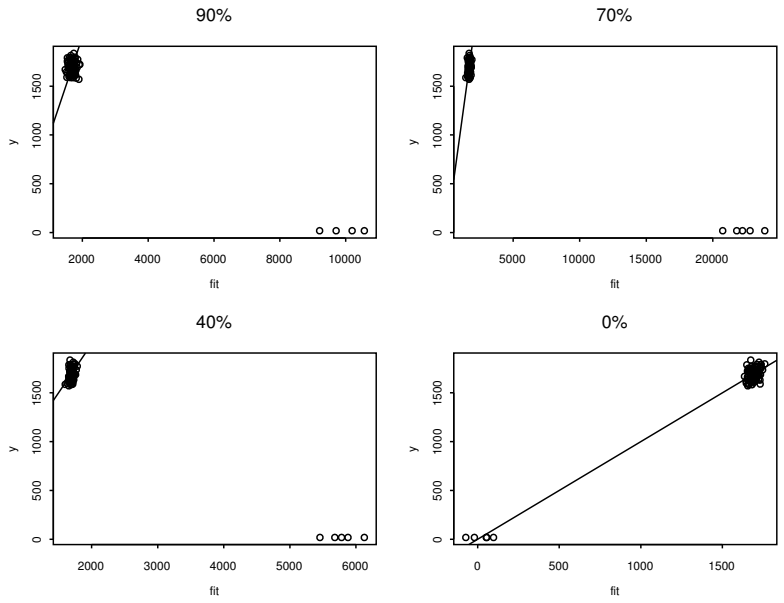

Figure 14.1: 4 Trimmed Views for the Buxton Data

of

$$\left(\frac{\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n}}{n} \right)^{-1}$$

are all contained in an interval $[0, B)$ for some $B > 0$, then $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$.

The distribution of the estimator $\hat{\boldsymbol{\beta}}_{M,n}$ is especially simple when OLS is used and the errors are iid $N(0, \sigma^2)$. Then

$$\hat{\boldsymbol{\beta}}_{M,n} = (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1} \boldsymbol{X}_{M,n}^T \boldsymbol{Y}_{M,n} \sim N_p(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1})$$

and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}) \sim N_p(\boldsymbol{0}, \sigma^2 (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n}/n)^{-1})$. Notice that this result does not imply that the distribution of $\hat{\boldsymbol{\beta}}_{T,n}$ is normal.

Table 14.1 compares the TV, MBA (for MLR), `lmsreg`, `ltsreg`, L_1 and OLS estimators on 7 data sets available from the text's website. The column headers give the file name while the remaining rows of the table give the sample size n, the number of predictors p, the amount of trimming M used by the TV estimator, the correlation of the residuals from the TV estimator with

Table 14.1: Summaries for Seven Data Sets, the Correlations of the Residuals from TV(M) and the Alternative Method are Given in the 1st 5 Rows

Method	Buxton	Gladstone	glado	hbk	major	nasty	wood
MBA	0.997	1.0	0.455	0.960	1.0	-0.004	0.9997
LMSREG	-0.114	0.671	0.938	0.977	0.981	0.9999	0.9995
LTSREG	-0.048	0.973	0.468	0.272	0.941	0.028	0.214
L1	-0.016	0.983	0.459	0.316	0.979	0.007	0.178
OLS	0.011	1.0	0.459	0.780	1.0	0.009	0.227
outliers	61-65	none	119	1-10	3,44	2,6,...,30	4,6,8,19
n	87	274	274	75	112	32	20
p	5	7	7	4	6	5	6
M	70	0	30	90	0	90	20

the corresponding alternative estimator, and the cases that were outliers. If the correlation was greater than 0.9, then the method was effective in detecting the outliers, and the method failed, otherwise. Sometimes the trimming percentage M for the TV estimator was picked after fitting the bulk of the data in order to find the good leverage points and outliers.

Notice that the TV, MBA and OLS estimators were the same for the Gladstone data and for the *major* data (Tremearne 1911) which had two small Y–outliers. For the Gladstone data, there is a cluster of infants that are good leverage points, and we attempt to predict *brain weight* with the head measurements *height, length, breadth, size* and *cephalic index*. Originally, the variable *length* was incorrectly entered as 109 instead of 199 for case 119, and the *glado* data contains this outlier. In 1997, `lmsreg` was not able to detect the outlier while `ltsreg` did. Due to changes in the *Splus* 2000 code, `lmsreg` now detects the outlier but `ltsreg` does not.

The TV estimator can be modified to create a resistant weighted MLR estimator. To see this, recall that the weighted least squares (WLS) estimator using weights W_i can be found using the ordinary least squares (OLS) regression (without intercept) of $\sqrt{W_i}Y_i$ on $\sqrt{W_i}\boldsymbol{x}_i$. This idea can be used for categorical data analysis since the minimum chi-square estimator is often computed using WLS. Let $\boldsymbol{x}_i = (1, x_{i,2}, ..., x_{i,p})^T$, let $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ and let

$\tilde{\boldsymbol{\beta}}$ be an estimator of $\boldsymbol{\beta}$.

Definition 14.2. For a multiple linear regression model with weights W_i, a **weighted response plot** is a plot of $\sqrt{W_i}\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}$ versus $\sqrt{W_i}Y_i$. The **weighted residual plot** is a plot of $\sqrt{W_i}\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}$ versus the WMLR residuals $r_{Wi} = \sqrt{W_i}Y_i - \sqrt{W_i}\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}$.

Application 14.1. For resistant weighted MLR, use the WTV estimator which is selected from ten weighted response plots.

14.2 1D Regression

Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response Y given the $k \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. The scalar Y is a random variable and $\boldsymbol{x}$ is a random vector. A special case of regression was the multiple linear regression model $Y = \alpha + x_1\beta_1 + \cdots + x_k\beta_k + e = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$ where $k = p-1$ and the nontrivial predictors are collected in the $k \times 1$ vector $\boldsymbol{x}$.

Definition 14.3: Cook and Weisberg (1999a, p. 414). In a *1D regression model*, the response Y is conditionally independent of $\boldsymbol{x}$ given a single linear combination $\boldsymbol{\beta}^T\boldsymbol{x}$ of the predictors, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T\boldsymbol{x} \text{ or } Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T\boldsymbol{x}). \tag{14.2}$$

An important 1D regression model, introduced by Li and Duan (1989), has the form

$$Y = g(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}, e) \tag{14.3}$$

where g is a bivariate (inverse link) function and e is a zero mean error that is independent of $\boldsymbol{x}$. The constant term α may be absorbed by g if desired.

Special cases of the 1D regression model (14.2) include many important *generalized linear models* (GLMs) and the additive error *single index model*

$$Y = m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) + e. \tag{14.4}$$

Typically m is the conditional mean or median function. For example if all of the expectations exist, then

$$E[Y|\boldsymbol{x}] = E[m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})|\boldsymbol{x}] + E[e|\boldsymbol{x}] = m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}).$$

The *multiple linear regression model* is an important special case where m is the identity function: $m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. Another important special case of 1D regression is the *response transformation model* where

$$g(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}, e) = t^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e) \tag{14.5}$$

and t^{-1} is a one to one (typically monotone) function. Hence

$$t(Y) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e.$$

Definition 14.4. *Regression* is the study of the conditional distribution of $Y|\boldsymbol{x}$. Focus is often on the *mean function* $E(Y|\boldsymbol{x})$ and/or the *variance function* $\mathrm{VAR}(Y|\boldsymbol{x})$. There is a distribution for each value of $\boldsymbol{x} = \boldsymbol{x}_o$ such that $Y|\boldsymbol{x} = \boldsymbol{x}_o$ is defined. For a 1D regression,

$$E(Y|\boldsymbol{x} = \boldsymbol{x}_o) = E(Y|\boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{\beta}^T\boldsymbol{x}_o) \equiv M(\boldsymbol{\beta}^T\boldsymbol{x}_o)$$

and

$$\mathrm{VAR}(Y|\boldsymbol{x} = \boldsymbol{x}_o) = \mathrm{VAR}(Y|\boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{\beta}^T\boldsymbol{x}_o) \equiv V(\boldsymbol{\beta}^T\boldsymbol{x}_o)$$

where M is the *kernel mean function* and V is the *kernel variance function*.

Notice that the mean and variance functions depend on the *same* linear combination if the 1D regression model is valid. This dependence is typical of GLMs where M and V are known kernel mean and variance functions that depend on the family of GLMs. See Cook and Weisberg (1999a, section 23.1). A *heteroscedastic regression model*

$$Y = M(\boldsymbol{\beta}_1^T\boldsymbol{x}) + \sqrt{V(\boldsymbol{\beta}_2^T\boldsymbol{x})}\ \ e \tag{14.6}$$

is a 1D regression model if $\boldsymbol{\beta}_2 = c\boldsymbol{\beta}_1$ for some scalar c.

Dimension reduction can greatly simplify our understanding of the conditional distribution $Y|\boldsymbol{x}$. If a 1D regression model is appropriate, then the k–dimensional vector $\boldsymbol{x}$ can be replaced by the 1–dimensional scalar $\boldsymbol{\beta}^T\boldsymbol{x}$ *with "no loss of information about the conditional distribution."* Cook and Weisberg (1999a, p. 411) define a *sufficient summary plot* (SSP) to be a plot that contains all the sample regression information about the conditional distribution $Y|\boldsymbol{x}$ of the response given the predictors.

Definition 14.5: If the 1D regression model holds, then $Y \perp\!\!\!\perp \boldsymbol{x} | (a + c\boldsymbol{\beta}^T\boldsymbol{x})$ for any constants a and $c \neq 0$. The quantity $a + c\boldsymbol{\beta}^T\boldsymbol{x}$ is called a *sufficient predictor* (SP), and a *sufficient summary plot* is a plot of any SP versus Y. An *estimated sufficient predictor* (ESP) is $\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T\boldsymbol{x}$ where $\tilde{\boldsymbol{\beta}}$ is an estimator of $c\boldsymbol{\beta}$ for some nonzero constant c. A *response plot* or *estimated sufficient summary plot* (ESSP) is a plot of any ESP versus Y.

If there is only one predictor x, then the plot of x versus Y is both a sufficient summary plot and a response plot, but generally only a response plot can be made. Since a can be any constant, $a = 0$ is often used. The following section shows how to use the OLS regression of Y on $\boldsymbol{x}$ to obtain an ESP. If we plot the fitted values and the ESP versus Y, the plots are called fit–response and ESP-response plots. For multiple linear regression, these two plots are the same.

14.3 Visualizing 1D Regression

Consider the OLS estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$. Li and Duan (1989, p. 1031) show that under regularity conditions, $\hat{\boldsymbol{\beta}}$ is a $\sqrt{n}$ consistent estimator of $c\boldsymbol{\beta}$ for some constant c. If $\hat{\boldsymbol{\beta}} \approx c\boldsymbol{\beta}$ when model (14.2) holds, then the response plot of

$$\hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x} \text{ versus } Y$$

can be used to visualize the conditional distribution $Y|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$ provided that $c \neq 0$. **Often if no strong nonlinearities are present among the predictors,** the bias vector is small enough so that $\hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ is a useful ESP.

Suppose $Y = m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) + e$ and the errors e are small. Suppose $\hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ is a good estimator of $c\boldsymbol{\beta}^T\boldsymbol{x}$. Then m can be visualized with both a plot of $ESP = a + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ versus Y if $c \neq 0$. If $c > 0$ then the plot of ESP versus Y is similar to the plot of $\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ versus Y: except the labels of the horizontal axis change. (The two plots are usually not exactly identical since plotting controls to "fill space" depend on several factors and will change slightly.) If $c < 0$, then the plot appears to be flipped about the vertical axis. OLS often provides a useful estimator of $c\boldsymbol{\beta}$ where $c \neq 0$, but OLS can result in $c = 0$ if g is symmetric about the population median of $\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$.

Definition 14.6. If the 1D regression model (14.2) holds, and OLS is

used, then the ESP may be called the *OLS ESP* and the response plot may be called the *OLS response plot*. Other estimators, such as SIR, may have similar labels.

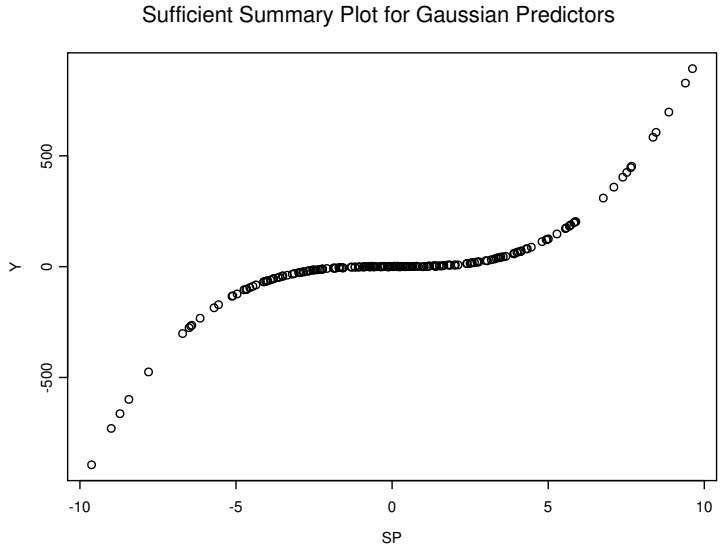

Figure 14.2: SSP for $m(u) = u^3$

Example 14.2. Suppose that $\boldsymbol{x}_i \sim N_3(\boldsymbol{0}, \boldsymbol{I}_3)$ and that

$$Y = m(\boldsymbol{\beta}^T \boldsymbol{x}) + e = (x_1 + 2x_2 + 3x_3)^3 + e.$$

Then a 1D regression model holds with $\boldsymbol{\beta} = (1, 2, 3)^T$. Figure 14.2 shows the sufficient summary plot of $\boldsymbol{\beta}^T \boldsymbol{x}$ versus Y, and Figure 14.3 shows the sufficient summary plot of $-\boldsymbol{\beta}^T \boldsymbol{x}$ versus Y. Notice that the functional form m appears to be cubic in both plots and that both plots can be smoothed by eye or with a scatterplot smoother such as *lowess.* The two figures were generated with the following *R/Splus* commands.

```
 X <- matrix(rnorm(300),nrow=100,ncol=3)
SP <- X%*%1:3
 Y <- (SP)^3 + rnorm(100)
      plot(SP,Y)
      plot(-SP,Y)
```

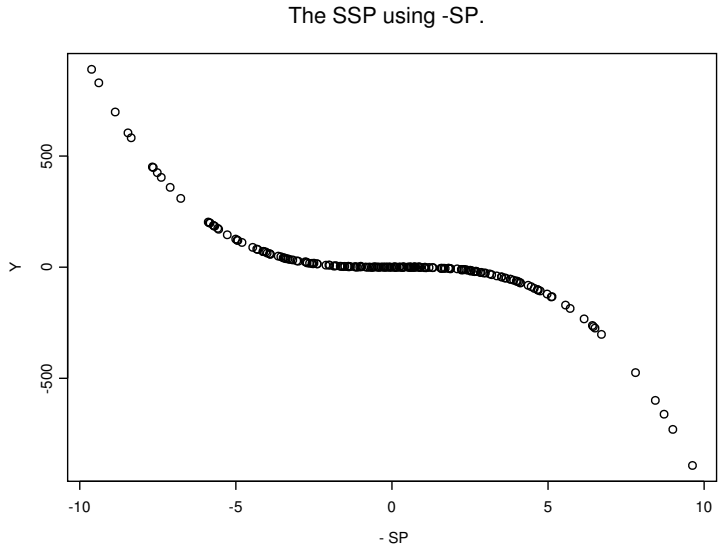

Figure 14.3: Another SSP for $m(u) = u^3$

We particularly want to use the OLS estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ to produce an estimated sufficient summary plot. This estimator is obtained from the usual multiple linear regression of Y_i on $\boldsymbol{x}_i$, but *we are not assuming that the multiple linear regression model holds*; however, we are hoping that the 1D regression model $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T \boldsymbol{x}$ is a useful approximation to the data and that $\hat{\boldsymbol{\beta}} \approx c\boldsymbol{\beta}$ for some nonzero constant c. Nice results exist if the single index model is appropriate. Recall that

$$\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{Y}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))((\boldsymbol{Y} - E(\boldsymbol{Y}))^T].$$

Definition 14.7. Suppose that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid observations and that the positive definite $k \times k$ matrix $\mathrm{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_X$ and the $k \times 1$ vector $\mathrm{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{X,Y}$. Let the OLS estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ be computed from the multiple linear regression of Y on $\boldsymbol{x}$ plus a constant. Then $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ estimates the population quantity $(\alpha_{OLS}, \boldsymbol{\beta}_{OLS})$ where

$$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_{X,Y}. \tag{14.7}$$

The following notation will be useful for studying the OLS estimator. Let the sufficient predictor $\boldsymbol{z} = \boldsymbol{\beta}^T\boldsymbol{x}$ and let $\boldsymbol{w} = \boldsymbol{x} - E(\boldsymbol{x})$. Let $\boldsymbol{r} = \boldsymbol{w} - (\boldsymbol{\Sigma}_X\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w}$.

Theorem 14.1. In addition to the conditions of Definition 14.7, also assume that $Y_i = m(\boldsymbol{\beta}^T\boldsymbol{x}_i) + e_i$ where the zero mean constant variance iid errors e_i are independent of the predictors $\boldsymbol{x}_i$. Then

$$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{X,Y} = c_{m,X}\boldsymbol{\beta} + \boldsymbol{u}_{m,X} \tag{14.8}$$

where the scalar

$$c_{m,X} = E[\boldsymbol{\beta}^T(\boldsymbol{x} - E(\boldsymbol{x}))\ m(\boldsymbol{\beta}^T\boldsymbol{x})] \tag{14.9}$$

and the bias vector

$$\boldsymbol{u}_{m,X} = \boldsymbol{\Sigma}_X^{-1}E[m(\boldsymbol{\beta}^T\boldsymbol{x})\boldsymbol{r}]. \tag{14.10}$$

Moreover, $\boldsymbol{u}_{m,X} = \boldsymbol{0}$ if $\boldsymbol{x}$ is from an EC distribution with nonsingular $\boldsymbol{\Sigma}_X$, and $c_{m,X} \neq 0$ unless $\mathrm{Cov}(\boldsymbol{x}, Y) = \boldsymbol{0}$. If the multiple linear regression model holds, then $c_{m,X} = 1$, and $\boldsymbol{u}_{m,X} = \boldsymbol{0}$.

The proof of the above result is outlined in Problem 14.1 using an argument due to Aldrin, Bϕlviken, and Schweder (1993). See related results in Cook, Hawkins, and Weisberg (1992). If the 1D regression model is appropriate, then typically $\mathrm{Cov}(\boldsymbol{x}, Y) \neq \boldsymbol{0}$ unless $\boldsymbol{\beta}^T\boldsymbol{x}$ follows a symmetric distribution and m is symmetric about the median of $\boldsymbol{\beta}^T\boldsymbol{x}$.

Definition 14.8. Let $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ denote the OLS estimate obtained from the OLS multiple linear regression of Y on $\boldsymbol{x}$. The *OLS view* is a response plot of $a + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ versus Y. Typically $a = 0$ or $a = \hat{\alpha}$.

Remark 14.1. All of this awkward notation and theory leads to a remarkable result, perhaps first noted by Brillinger (1977, 1983) and called the *1D Estimation Result* by Cook and Weisberg (1999a, p. 432). The result is that if the 1D regression model is appropriate, then *the OLS view will frequently be a useful estimated sufficient summary plot* (ESSP). Hence the OLS predictor $\hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ is a useful *estimated sufficient predictor* (ESP).

Although the OLS view is frequently a good ESSP if no strong nonlinearities are present in the predictors and if $c_{m,X} \neq 0$ (eg the sufficient summary plot of $\boldsymbol{\beta}^T\boldsymbol{x}$ versus Y is not approximately symmetric), even better estimated sufficient summary plots can be obtained by using ellipsoidal trimming. This topic is discussed in the following section and follows Olive (2002) closely.

To perform ellipsoidal trimming, an estimator $(T, \boldsymbol{C})$ is computed where T is a $k \times 1$ multivariate location estimator and $\boldsymbol{C}$ is a $k \times k$ symmetric positive definite dispersion estimator. Then the ith squared Mahalanobis distance is the random variable

$$D_i^2 = (\boldsymbol{x}_i - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x}_i - T) \tag{14.11}$$

for each vector of observed predictors $\boldsymbol{x}_i$. If the ordered distances $D_{(j)}$ are unique, then j of the $\boldsymbol{x}_i$ are in the hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T C^{-1} (\boldsymbol{x} - T) \leq D_{(j)}^2\}. \tag{14.12}$$

The ith case $(Y_i, \boldsymbol{x}_i^T)^T$ is trimmed if $D_i > D_{(j)}$. Thus if $j \approx 0.9n$, then about 10% of the cases are trimmed.

We suggest that the estimator $(T, \boldsymbol{C})$ should be the classical sample mean and covariance matrix $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ or a robust multivariate location and dispersion estimator such as RFCH. See Section 4.4. When $j \approx n/2$, the RFCH estimator attempts to make the volume of the hyperellipsoid given by Equation (14.12) small.

Ellipsoidal trimming seems to work for at least three reasons. The trimming divides the data into two groups: the *trimmed cases* and the *remaining cases* $(\boldsymbol{x}_M, Y_M)$ where M% is the amount of trimming, eg $M = 10$ for 10% trimming. If the distribution of the predictors $\boldsymbol{x}$ is EC then the distribution of $\boldsymbol{x}_M$ still retains enough symmetry so that the bias vector is approximately zero. If the distribution of $\boldsymbol{x}$ is not EC, then the distribution of $\boldsymbol{x}_M$ will often have enough symmetry so that the bias vector is small. In particular, trimming often removes strong nonlinearities from the predictors and the weighted predictor distribution is more nearly elliptically symmetric than the predictor distribution of the entire data set (recall Winsor's principle: "all data are roughly Gaussian in the middle"). Secondly, under heavy trimming, the mean function of the remaining cases may be more linear than the mean function of the entire data set. Thirdly, if $|c|$ is very large, then the bias vector may be small relative to $c\boldsymbol{\beta}$. Trimming sometimes inflates $|c|$. From Theorem 14.1, any of these three reasons should produce a better estimated sufficient predictor.

For example, examine Figure 5.4. The data are not EC, but the data within the resistant covering ellipsoid are approximately EC.

Example 14.3. Cook and Weisberg (1999a, p. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. The variables

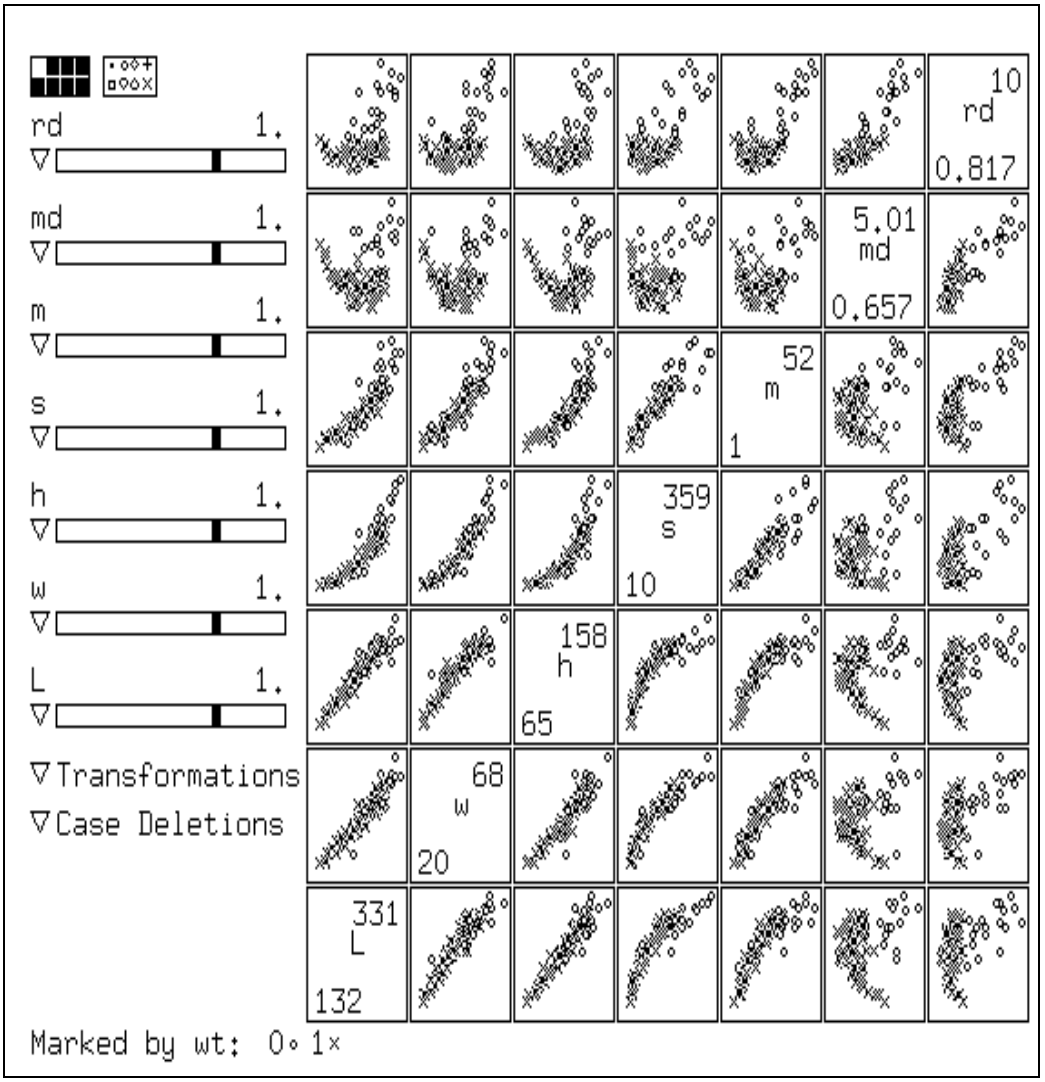

Figure 14.4: Scatterplot for Mussel Data, o Corresponds to Trimmed Cases

are the *muscle mass* M in grams, the *length* L and *height* H of the shell in mm, the *shell width* W and the *shell mass* S. The robust and classical Mahalanobis distances were calculated, and Figure 14.4 shows a scatterplot matrix of the mussel data, the RD_i's, and the MD_i's. Notice that many of the subplots are nonlinear. The cases marked by open circles were given weight zero by the FMCD algorithm, and the linearity of the retained cases has increased. Note that only one trimming proportion is shown and that a heavier trimming proportion would increase the linearity of the cases that were not trimmed.

The two ideas of using ellipsoidal trimming to reduce the bias and choosing a view with a smooth mean function and smallest variance function can be combined into a graphical method for finding the estimated sufficient summary plot and the estimated sufficient predictor. Trim the $M\%$ of the cases with the largest Mahalanobis distances, and then compute the OLS estimator $(\hat{\alpha}_M, \hat{\boldsymbol{\beta}}_M)$ from the cases that remain. Use $M = 0$, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten plots of $\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}$ versus Y using all n cases. In analogy with the Cook and Weisberg procedure for visualizing 1D structure with two predictors, the plots will be called "trimmed views." Notice that $M = 0$ corresponds to the OLS view.

Definition 14.9. The *best trimmed view* is the trimmed view with a smooth mean function and the smallest variance function and is the estimated sufficient summary plot. If $M^* = E$ is the percentage of cases trimmed that corresponds to the best trimmed view, then $\hat{\boldsymbol{\beta}}_E^T \boldsymbol{x}$ is the estimated sufficient predictor.

The following examples illustrate the *R/Splus* function `trviews` that is used to produce the ESSP. If R is used instead of *Splus*, the command

```
library(MASS)
```

needs to be entered to access the function `cov.mcd` called by `trviews`. The function `trviews` is used in Problem 14.2. Also notice the `trviews` estimator is basically the same as the `tvreg` estimator described in Section 14.1. The `tvreg` estimator can be used to simultaneously detect whether the data is following a multiple linear regression model or some other single index model. Plot $\hat{\alpha}_E + \hat{\boldsymbol{\beta}}_E^T \boldsymbol{x}$ versus Y and add the identity line. If the plotted points follow the identity line then the MLR model is reasonable, but if the plotted points

follow a nonlinear mean function, then a nonlinear single index model may be reasonable.

Example 14.2 continued. The command

```
trviews(X, Y)
```

produced the following output.

```
Intercept       X1       X2       X3
0.6701255 3.133926 4.031048 7.593501
Intercept       X1       X2       X3
 1.101398 8.873677 12.99655 18.29054
Intercept       X1       X2       X3
0.9702788 10.71646 15.40126 23.35055
Intercept       X1       X2       X3
0.5937255 13.44889 23.47785 32.74164
Intercept       X1       X2       X3
 1.086138 12.60514 25.06613 37.25504
Intercept       X1       X2       X3
 4.621724 19.54774 34.87627 48.79709
Intercept       X1       X2       X3
 3.165427 22.85721 36.09381 53.15153
Intercept       X1       X2       X3
 5.829141 31.63738 56.56191 82.94031
Intercept       X1       X2       X3
 4.241797 36.24316 70.94507 105.3816
Intercept       X1       X2       X3
 6.485165 41.67623 87.39663 120.8251
```

The function generates 10 trimmed views. The first plot trims 90% of the cases while the last plot does not trim any of the cases and is the OLS view. To advance a plot, press the right button on the mouse (in *R*, highlight `stop` rather than `continue`). After all of the trimmed views have been generated, the output is presented. For example, the 5th line of numbers in the output corresponds to $\hat{\alpha}_{50} = 1.086138$ and $\hat{\boldsymbol{\beta}}_{50}^T$ where 50% trimming was used. The second line of numbers corresponds to 80% trimming while the last line corresponds to 0% trimming and gives the OLS estimate $(\hat{\alpha}_0, \hat{\boldsymbol{\beta}}_0^T) = (\hat{a}, \hat{\boldsymbol{b}})$. The trimmed views with 50% and 90% trimming were very good.

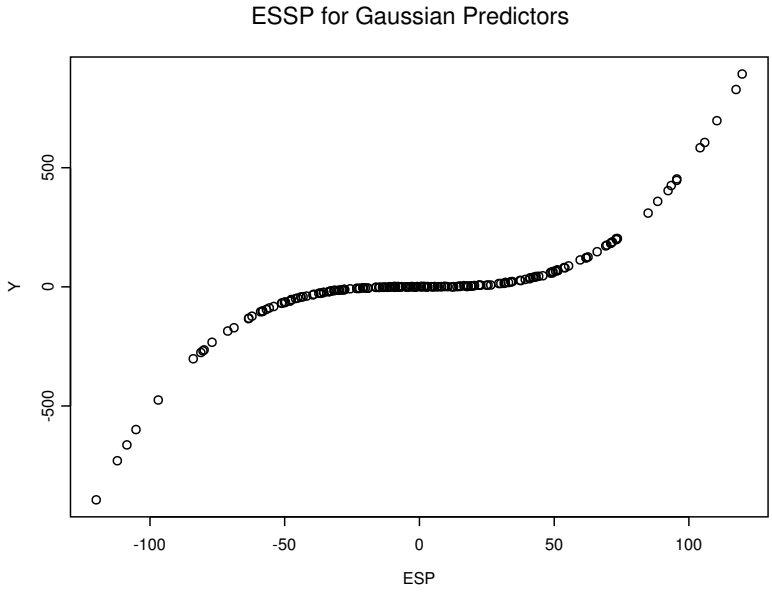

Figure 14.5: Best View for Estimating $m(u) = u^3$

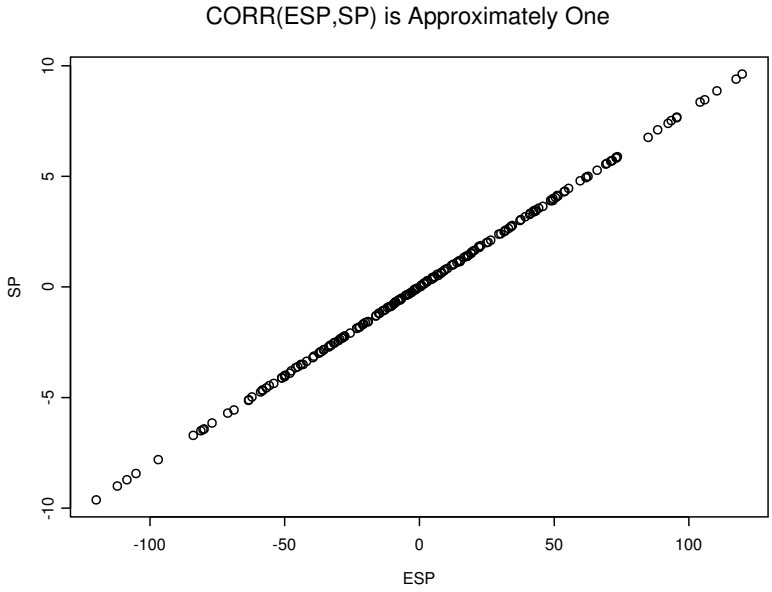

Figure 14.6: The angle between the SP and the ESP is nearly zero.

We decided that the view with 50% trimming was the best. Hence $\hat{\boldsymbol{\beta}}_E = (12.60514, 25.06613, 37.25504)^T \approx 12.5\boldsymbol{\beta}$. The best view is shown in Figure 14.5 and is nearly identical to the sufficient summary plot shown in Figure 14.2. Notice that the OLS estimate $= (41.68, 87.40, 120.83)^T \approx 42\boldsymbol{\beta}$.

The plot of the estimated sufficient predictor versus the sufficient predictor is also informative. Of course this plot can usually only be generated for simulated data since $\boldsymbol{\beta}$ is generally unknown. If the plotted points are highly correlated (with |corr(ESP,SP)| > 0.95) and follow a line through the origin, then the estimated sufficient summary plot is nearly as good as the sufficient summary plot. The simulated data used $\boldsymbol{\beta} = (1, 2, 3)^T$, and the commands

```
SP <- X %*% 1:3
ESP <- X %*% c(12.60514, 25.06613, 37.25504)
plot(ESP,SP)
```

generated the plot shown in Figure 14.6.

Example 14.5. An artificial data set with 200 trivariate vectors $\boldsymbol{x}_i$ was generated. The marginal distributions of $x_{i,j}$ are iid lognormal for $j = 1, 2,$ and 3. Since the response $Y_i = \sin(\boldsymbol{\beta}^T\boldsymbol{x}_i)/\boldsymbol{\beta}^T\boldsymbol{x}_i$ where $\boldsymbol{\beta} = (1, 2, 3)^T$, the random vector $\boldsymbol{x}_i$ is not elliptically contoured and the function m is strongly nonlinear. Figure 14.7d shows the OLS view and Figure 14.8d shows the best trimmed view. Notice that it is difficult to visualize the mean function with the OLS view, and notice that the correlation between Y and the ESP is very low. By focusing on a part of the data where the correlation is high, it may be possible to improve the estimated sufficient summary plot. For example, in Figure 14.8d, temporarily omit cases that have ESP less than 0.3 and greater than 0.75. From the untrimmed cases, obtained the ten trimmed estimates $\hat{\boldsymbol{\beta}}_{90}, ..., \hat{\boldsymbol{\beta}}_0$. Then using *all of the data*, obtain the ten views. The best view could be used as the ESSP.

Application 14.2. Suppose that a 1D regression analysis is desired on a data set, use the trimmed views as an exploratory data analysis technique to visualize the conditional distribution $Y|\boldsymbol{\beta}^T\boldsymbol{x}$. The best trimmed view is an estimated sufficient summary plot. If the single index model (14.4) holds, the function m can be estimated from this plot using parametric models or scatterplot smoothers such as `lowess`. Notice that Y can be predicted visually using *up and over lines*.

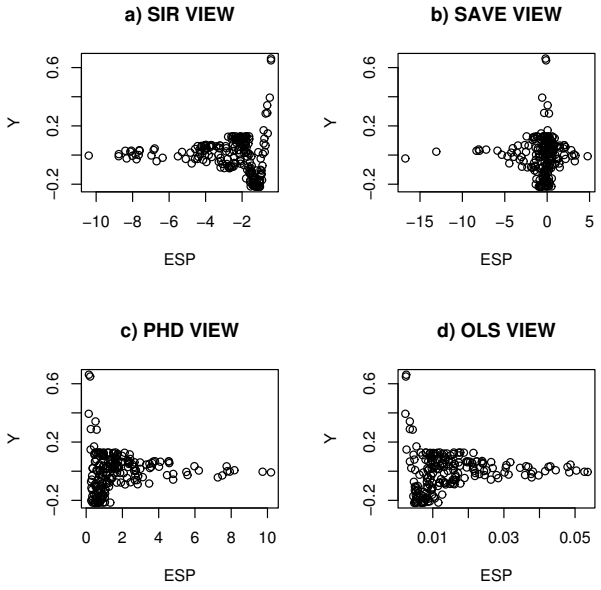

Figure 14.7: Estimated Sufficient Summary Plots Without Trimming

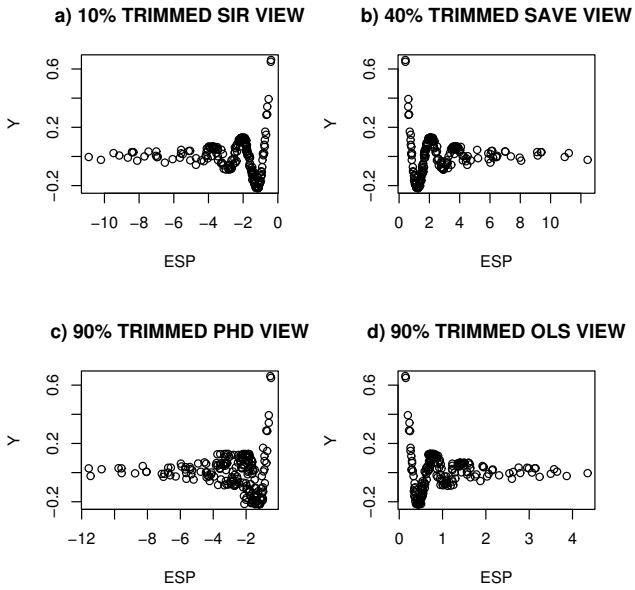

Figure 14.8: 1D Regression with Trimmed Views

Table 14.2: Estimated Sufficient Predictors Coefficients Estimating $c(1, 2, 3)^T$

method	b_1	b_2	b_3
OLS View	0.0032	0.0011	0.0047
90% Trimmed OLS View	0.086	0.182	0.338
SIR View	−0.394	−0.361	−0.845
10% Trimmed SIR VIEW	−0.284	−0.473	−0.834
SAVE View	−1.09	0.870	-0.480
40% Trimmed SAVE VIEW	0.256	0.591	0.765
PHD View	−0.072	−0.029	−0.0097
90% Trimmed PHD VIEW	−0.558	−0.499	−0.664
LMSREG VIEW	−0.003	−0.005	−0.059
70% Trimmed LMSREG VIEW	0.143	0.287	0.428

Application 14.4. The best trimmed view can also be used as a diagnostic for linearity and monotonicity.

For example in Figure 14.5, if ESP = 0, then $\hat{Y} = 0$ and if ESP = 100, then $\hat{Y} = 500$. Figure 14.5 suggests that the mean function is monotone but not linear, and Figure 14.8 suggests that the mean function is neither linear nor monotone.

Application 14.4. Assume that a known 1D regression model is assumed for the data. Then the best trimmed view is a model checking plot and can be used as a diagnostic for whether the assumed model is appropriate.

The trimmed views are sometimes useful even when the assumption of linearly related predictors fails. Cook and Li (2002) summarize when competing methods such as the OLS view, sliced inverse regression (SIR), principal Hessian directions (PHD), and sliced average variance estimation (SAVE) can fail. All four methods frequently perform well if there are no strong nonlinearities present in the predictors.

Example 14.5 (continued). Figure 14.7 shows that the response plots for SIR, PHD, SAVE, and OLS are not very good while Figure 14.8 shows that trimming improved the SIR, SAVE and OLS methods.

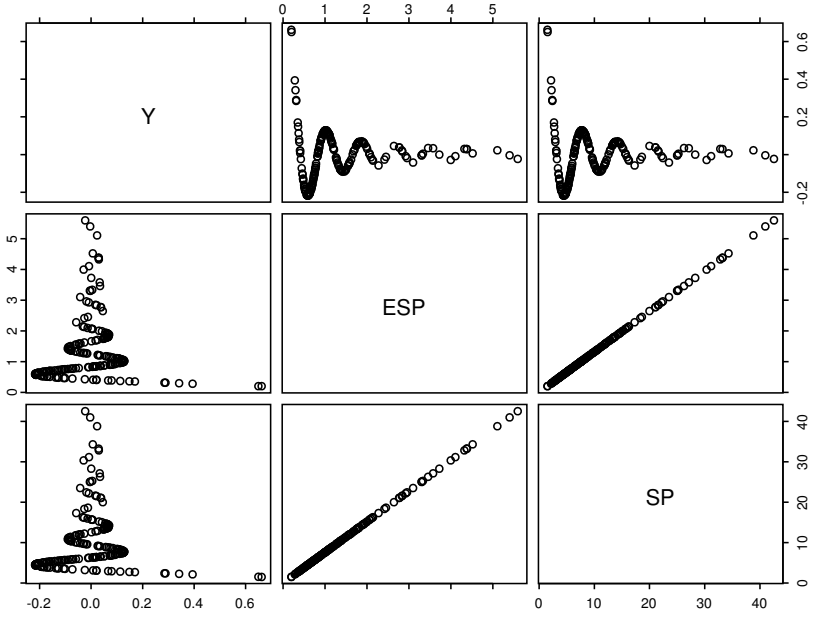

Figure 14.9: 1D Regression with `lmsreg`

One goal for future research is to develop better methods for visualizing 1D regression. Trimmed views seem to become less effective as the number of predictors $k = p - 1$ increases. Consider the sufficient predictor SP $= x_1 + \cdots + x_k$. With the sin(SP)/SP data, several trimming proportions gave good views with $k = 3$, but only one of the ten trimming proportions gave a good view with $k = 10$. In addition to problems with dimension, it is not clear which covariance estimator and which regression estimator should be used. We suggest using the RFCH estimator with OLS, and preliminary investigations suggest that the classical covariance estimator gives better estimates than `cov.mcd`. But among the many *Splus* regression estimators, `lmsreg` often worked well. There is OLS theory, but there is no theory for the robust regression estimators.

Example 14.5 continued. Replacing the OLS trimmed views by alternative MLR estimators often produced good response plots, and for single index models, the `lmsreg` estimator often worked the best. Figure 14.9 shows a scatterplot matrix of Y, ESP and SP where the sufficient predictor SP $= \boldsymbol{\beta}^T \boldsymbol{x}$. The ESP used ellipsoidal trimming with `cov.mcd` and with `lmsreg`

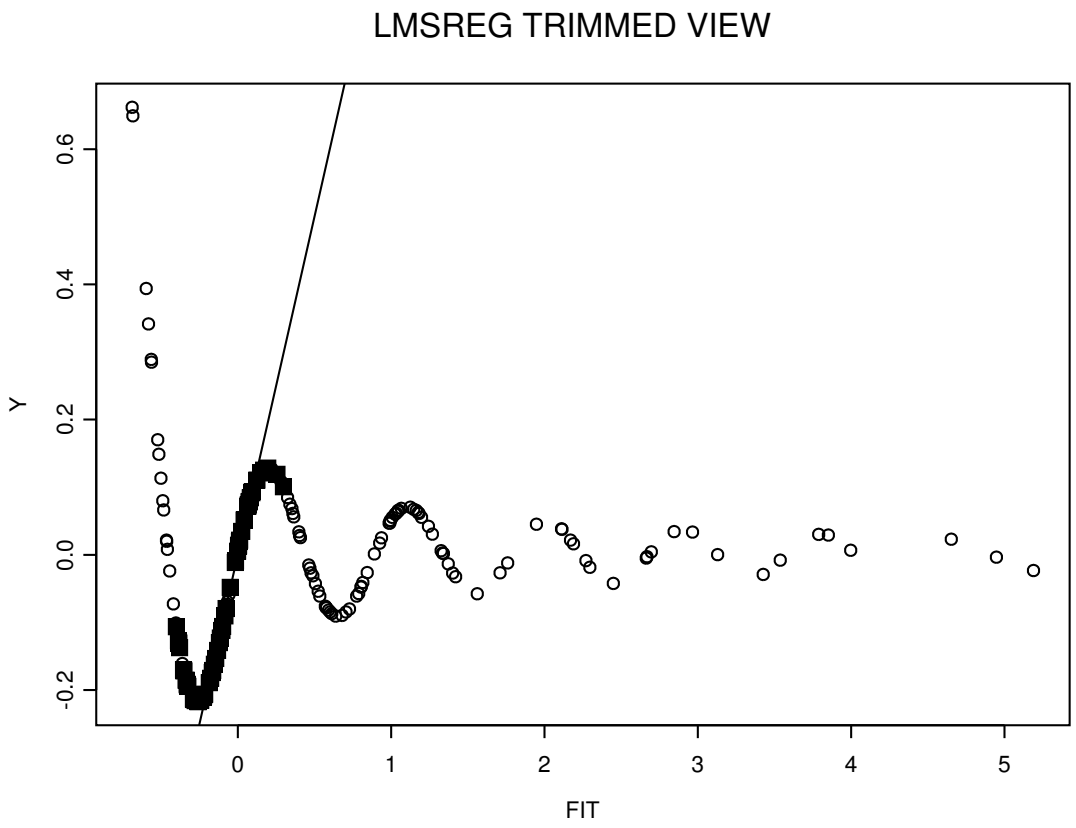

Figure 14.10: The Weighted `lmsreg` Fitted Values Versus Y

instead of OLS. The top row of Figure 14.9 shows that the estimated sufficient summary plot and the sufficient summary plot are nearly identical. Also the correlation of the ESP and the SP is nearly one. Table 14.2 shows the estimated sufficient predictor coefficients $\boldsymbol{b}$ when the sufficient predictor coefficients are $c(1, 2, 3)^T$. Only the SIR, SAVE, OLS and `lmsreg` trimmed views produce estimated sufficient predictors that are highly correlated with the sufficient predictor.

Figure 14.10 helps illustrate why ellipsoidal trimming works. This view used 70% trimming and the open circles denote cases that were trimmed. The highlighted squares correspond to the cases $(\boldsymbol{x}_{70}, Y_{70})$ that were not trimmed. Note that the highlighted cases are far more linear than the data set as a

whole. Also `lmsreg` will give half of the highlighted cases zero weight, further linearizing the function. In Figure 14.10, the `lmsreg` constant $\hat{\alpha}_{70}$ is included, and the plot is simply the response plot of the weighted `lmsreg` fitted values versus Y. The vertical deviations from the line through the origin are the "residuals" $Y_i - \hat{\alpha}_{70} - \hat{\boldsymbol{\beta}}_{70}^T \boldsymbol{x}$ and at least half of the highlighted cases have small residuals.

Example 14.6. This insulation data was contributed by Ms. Spector. A box with insulation was heated for 20 minutes then allowed to cool down. The response variable Y = *temperature* in middle of box was taken at *time* 0, 5, ..., 40. The *type* of insulation was a factor with type 1 = no insulation, 2 = corn pith, 3 = fiberglass, 4 = styrofoam and 5 = bubbles. There were 45 temperature measurements, one for each time type combination. The measurements were averages of ten trials and starting temperatures were close but not exactly equal.

The model using time, (time)^2, type, and the interactions type:time and type:(time)^2 had $E(Y|\boldsymbol{x}) \approx (\boldsymbol{x}^T\boldsymbol{\beta})^2$. A second model used time, (time)^2 and type, and rather awkward R code for producing the response plot in Figure 14.11 is shown below. The solid curve corresponds to $(\boldsymbol{x}^T\hat{\boldsymbol{\beta}}, (\boldsymbol{x}^T\hat{\boldsymbol{\beta}})^2) = (FIT, (FIT)^3)$ where $\hat{\boldsymbol{\beta}}$ is the OLS estimator from regressing Y on $\boldsymbol{x}^T$ = (1, time, (time)^2, type). The thin curve corresponds to lowess. Since the two lines correspond, $E(Y|\boldsymbol{x}) \approx (\boldsymbol{x}^T\boldsymbol{\beta})^3$ or $Y = m(\boldsymbol{x}^T\boldsymbol{\beta}) + e$ where $m(w) = w^3$. See Problem 14.7 for producing the response plot in *Arc*.

```
#assume the insulation data is loaded
ftype <- as.factor(insulation[,2])
zi <- as.data.frame(insulation)
iout <- lm(ytemp~time+I(time^2)+ftype,data=zi)
FIT <- iout$fit
Y <- insulation[,1]
plot(FIT,Y)
lines(lowess(FIT,Y))  #get (FIT,(FIT)^3) curve
zx <- FIT
z <- lsfit(cbind(zx,zx^2,zx^3),Y)
zfit <- Y-z$resid
lines(FIT,zfit)
```

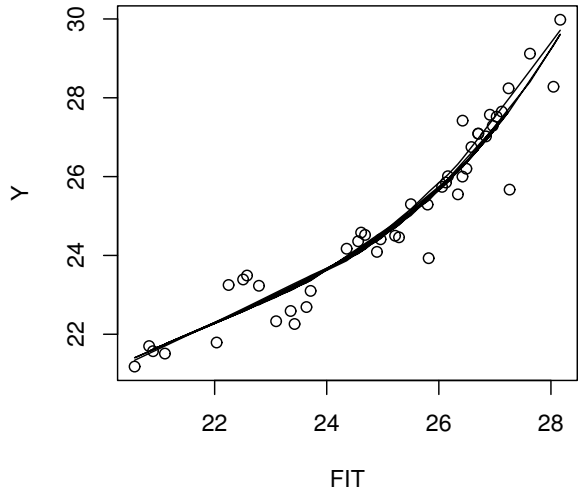

Figure 14.11: Response Plot for Insulation Data

14.4 Complements

The TV estimator was proposed by Olive (2002, 2005) and is similar to an estimator proposed by Rousseeuw and van Zomeren (1992). Although both the TV and MBA estimators have the good $O_P(n^{-1/2})$ convergence rate, their efficiency under normality may be very low. Chang and Olive (2007, 2010) suggest a method of adaptive trimming such that the resulting estimator is asymptotically equivalent to the OLS estimator.

Introduction to 1D regression and regression graphics are Cook and Weisberg (1999a, ch. 18, 19, and 20) and Cook and Weisberg (1999b), while Olive (2010) considers 1D regression. Also see Olive (2013, ch. 12).

14.5 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

14.1*. (Aldrin, Bølviken, and Schweder 1993). Suppose

$$Y = m(\boldsymbol{\beta}^T \boldsymbol{x}) + e \tag{14.13}$$

where m is a possibly unknown function and the zero mean errors e are independent of the predictors. Let $z = \boldsymbol{\beta}^T \boldsymbol{x}$ and let $\boldsymbol{w} = \boldsymbol{x} - E(\boldsymbol{x})$. Let $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \text{Cov}(\boldsymbol{x}, Y)$, and let $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ =Cov$(\boldsymbol{x})$ = Cov$(\boldsymbol{w})$. Let $\boldsymbol{r} = \boldsymbol{w} - (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w}$.

a) Recall that Cov$(\boldsymbol{x}, \boldsymbol{Y}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$ and show that $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = E(\boldsymbol{w}Y)$.

b) Show that $E(\boldsymbol{w}Y) = \boldsymbol{\Sigma}_{\boldsymbol{x},Y} = E[(\boldsymbol{r} + (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})\ m(z)] =$

$$E[m(z)\boldsymbol{r}] + E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}.$$

c) Using $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$, show that $\boldsymbol{\beta}_{OLS} = c(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{u}(\boldsymbol{x})$ where the constant

$$c(\boldsymbol{x}) = E[\boldsymbol{\beta}^T(\boldsymbol{x} - E(\boldsymbol{x}))m(\boldsymbol{\beta}^T\boldsymbol{x})]$$

and the bias vector $\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}E[m(\boldsymbol{\beta}^T\boldsymbol{x})\boldsymbol{r}]$.

d) Show that $E(\boldsymbol{w}z) = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}$. (Hint: Use $E(\boldsymbol{w}z) = E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{x}^T\boldsymbol{\beta}] = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x}^T - E(\boldsymbol{x}^T) + E(\boldsymbol{x}^T))\boldsymbol{\beta}]$.)

e) Assume $m(z) = z$. Using d), show that $c(\boldsymbol{x}) = 1$ if $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$.

f) Assume that $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$. Show that $E(z\boldsymbol{r}) = E(\boldsymbol{r}z) = \boldsymbol{0}$. (Hint: Find $E(\boldsymbol{r}z)$ and use d).)

g) Suppose that $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$ and that the distribution of $\boldsymbol{x}$ is multivariate normal. Then the joint distribution of z and $\boldsymbol{r}$ is multivariate normal. Using the fact that $E(z\boldsymbol{r}) = \boldsymbol{0}$, show Cov$(\boldsymbol{r}, z) = 0$ so that z and $\boldsymbol{r}$ are independent. Then show that $\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{0}$.

(Note: the assumption $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$ can be made without loss of generality since if $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = d^2 > 0$ (assuming $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is positive definite), then $y = m(d(\boldsymbol{\beta}/d)^T\boldsymbol{x}) + e \equiv m_d(\boldsymbol{\eta}^T\boldsymbol{x}) + e$ where $m_d(u) = m(du)$, $\boldsymbol{\eta} = \boldsymbol{\beta}/d$ and $\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\eta} = 1$.)

R/Splus Problems

Warning: Use the command *source("G:/mpack.txt")* **to download the programs. See Preface or Section 15.2.** Typing the name of the

`mpack` function, eg *trviews*, will display the code for the function. Use the `args` command, eg *args(trviews)*, to display the needed arguments for the function.

14.2. Use the following *R/Splus* commands to make 100 $N_3(\mathbf{0}, I_3)$ cases and 100 trivariate non-EC cases.

```
n3x <- matrix(rnorm(300),nrow=100,ncol=3)
ln3x <- exp(n3x)
```

In *R*, type the command *library(MASS).*

a) Using the commands *pairs(n3x)* and *pairs(ln3x)* and include both scatterplot matrices in *Word.* (Click on the plot and hit *Ctrl* and *c* at the same time. Then go to *file* in the *Word* menu and select *paste.*) Are strong nonlinearities present among the MVN predictors? How about the non-EC predictors? (Hint: a box or ball shaped plot is linear.)

b) Make a single index model and the sufficient summary plot with the following commands

```
 ncy <- (n3x%*%1:3)^3 + 0.1*rnorm(100)
plot(n3x%*%(1:3),ncy)
```

and include the plot in *Word.*

c) The command *trviews(n3x, ncy)* will produce ten plots. To advance the plots, click on the *rightmost mouse button* (and in *R* select *stop*) to advance to the next plot. The last plot is the OLS view. Include this plot in *Word.*

d) After all 10 plots have been looked at the output will show 10 estimated predictors. The last estimate is the OLS (least squares) view and might look like

```
Intercept        X1        X2        X3
 4.417988 22.468779 61.242178 75.284664
```

If the OLS view is a good estimated sufficient summary plot, then the plot created from the command (leave out the intercept)

```
plot(n3x%*%c(22.469,61.242,75.285),n3x%*%1:3)
```

should cluster tightly about some line. Your linear combination will be different than the one used above. Using your OLS view, include the plot using the command above (but with your linear combination) in *Word.* Was this plot linear? Did some of the other trimmed views seem to be better than the OLS view, that is, did one of the trimmed views seem to have a smooth mean function with a smaller variance function than the OLS view?

e) Now type the *R/Splus* command

```
lncy <- (ln3x%*%1:3)^3 + 0.1*rnorm(100).
```

Use the command *trviews(ln3x,lncy)* to find the best view with a smooth mean function and the smallest variance function. This view should not be the OLS view. Include your best view in *Word.*

f) Get the linear combination from your view, say $(94.848, 216.719, 328.444)^T$, and obtain a plot with the command

```
plot(ln3x%*%c(94.848,216.719,328.444),ln3x%*%1:3).
```

Include the plot in *Word.* If the plot is linear with high correlation, then your response plot in e) should be good.

14.3. (At the beginning of your *R/Splus* session, use *source("G:/rpack.txt")* command (and *library(MASS)* in *R.*))

a) Perform the commands

```
> nx <- matrix(rnorm(300),nrow=100,ncol=3)
> lnx <- exp(nx)
> SP <- lnx%*%1:3
> lnsincy <- sin(SP)/SP + 0.01*rnorm(100)
```

For parts b), c) and d) below, to make the best trimmed view with `trviews`, `ctrviews` or `lmsviews`, you may need to use the function twice. The first view trims 90% of the data, the next view trims 80%, etc. The last view trims 0% and is the OLS view (or `lmsreg` view). Remember to advance the view with the rightmost mouse button (and in *R*, highlight "stop"). Then click on the plot and next simultaneously hit *Ctrl* and *c*. This makes a copy of the plot. Then in *Word*, use the menu commands "Copy>paste."

b) Find the best trimmed view with `OLS` and `covfch` with the following commands and include the view in *Word.*

```
> trviews(lnx,lnsincy)
```

(With `trviews`, suppose that 40% trimming gave the best view. Then instead of using the procedure above b), you can use the command

```
> essp(lnx,lnsincy,M=40)
```

to make the best trimmed view. Then click on the plot and next simultaneously hit *Ctrl* and *c*. This makes a copy of the plot. Then in *Word*, use the menu commands "Copy>paste". Click the rightmost mouse button (and in *R*, highlight "stop") to return the command prompt.)

c) Find the best trimmed view with `OLS` and $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ using the following commands and include the view in *Word.* See the paragraph above b).

```
> ctrviews(lnx,lnsincy)
```

d) Find the best trimmed view with `lmsreg` and `cov.mcd` using the following commands and include the view in *Word.* See the paragraph above b).

```
> lmsviews(lnx,lnsincy)
```

e) Which method or methods gave the best response plot? Explain briefly.

14.4. Warning: this problem may take too much time. This problem is like Problem 14.3 but uses many more single index models.
a) Make some prototype functions with the following commands.

```
> nx <- matrix(rnorm(300),nrow=100,ncol=3)
> SP <- nx%*%1:3
> ncuby <- SP^3 + rnorm(100)
> nexpy <- exp(SP) + rnorm(100)
> nlinsy <- SP + 4*sin(SP) + 0.1*rnorm(100)
> nsincy <- sin(SP)/SP + 0.01*rnorm(100)
> nsiny <- sin(SP) + 0.1*rnorm(100)
> nsqrty <- sqrt(abs(SP)) + 0.1*rnorm(100)
> nsqy <- SP^2 + rnorm(100)
```

b) Make sufficient summary plots similar to Figures 14.2 and 14.3 with the following commands and include both plots in *Word*.

```
> plot(SP,ncuby)
> plot(-SP,ncuby)
```

c) Find the best trimmed view with the following commands (first type `library(MASS)` if you are using *R*). Include the view in *Word.*

```
> trviews(nx,ncuby)
```

You may need to use the function twice. The first view trims 90% of the data, the next view trims 80%, etc. The last view trims 0% and is the OLS view. Remember to advance the view with the rightmost mouse button (and in *R*, highlight "stop"). Suppose that 40% trimming gave the best view. Then use the command

```
> essp(nx,ncuby, M=40)
```

to make the best trimmed view. Then click on the plot and next simultaneously hit *Ctrl* and *c*. This makes a copy of the plot. Then in *Word*, use the menu commands "Copy>paste".

d) To make a plot like Figure 14.6, use the following commands. Let *tem* $= \hat{\boldsymbol{\beta}}$ obtained from the *trviews* output. In Example 14.2 (continued), *tem* can be obtained with the following command.

```
> tem <- c(12.60514, 25.06613, 37.25504)
```

Include the plot in *Word.*

```
> ESP <- nx%*%tem
> plot(ESP,SP)
```

e) Repeat b), c) and d) with the following data sets.
i) nx and nexpy
ii) nx and nlinsy
iii) nx and nsincy
iv) nx and nsiny
v) nx and nsqrty
vi) nx and nsqy
Enter the following commands to do parts vii) to x).

```
> lnx <- exp(nx)
> SP <- lnx%*%1:3
> lncuby <- (SP/3)^3 + rnorm(100)
> lnlinsy <- SP + 10*sin(SP) + 0.1*rnorm(100)
> lnsincy <- sin(SP)/SP + 0.01*rnorm(100)
> lnsiny <- sin(SP/3) + 0.1*rnorm(100)
> ESP <- lnx%*%tem
```

vii) lnx and lncuby
viii) lnx and lnlinsy
ix) lnx and lnsincy
x) lnx and lnsiny

14.5. Warning: this problem may take too much time. Repeat Problem 14.4 but replace `trviews` with a) `lmsviews`, b) `symviews` (that creates views that sometimes work even when symmetry is present), c) `ctrviews` and d) `sirviews`.

Except for part a), the *essp* command will not work. Instead, for the best trimmed view, click on the plot and next simultaneously hit *Ctrl* and *c*. This makes a copy of the plot. Then in *Word*, use the menu commands "Copy>paste".

14.6. a) In addition to the *source("G:/mpack.txt")* command, also use the *source("G:/mrobdata.txt")* command (and in *R*, type the *library(MASS)* command).

b) Type the command *tvreg(buxx,buxy,ii=1)*. Click the rightmost mouse button (and in *R*, highlight *Stop*). The response plot should appear. Repeat 10 times and remember which plot percentage M (say M = 0) had the best response plot. Then type the command *tvreg2(buxx,buxy, M = 0)* (except use your value of M, not 0). Again, click the rightmost mouse button (and in *R*, highlight *Stop*). The response plot should appear. Hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

c) The estimated coefficients $\hat{\boldsymbol{\beta}}_{TV}$ from the best plot should have appeared on the screen. Copy and paste these coefficients into *Word*.

Problem using ARC

14.7. a) Activate the *insulation.lsp* dataset of Example 14.6 with the menu commands "File > Load > Removable Disk (G:) > insulation.lsp." Scroll up the screen to read the data description.

b) From the insulation menu select *Transform*, click on *time*, change the number in the *p box* to 2 and click on OK to add time^2 to the variable list. From the insulation menu select *Make factors*, click on *type* and click on OK to make the factor {F}type. From the insulation menu select *Make interactions*, click on {F}type and time, then click on OK. Again from the insulation menu select *Make interactions*, click on {F}type and time^2, then click on OK.

c) From the Graph&Fit menu select *Fit linear LS*, place y in the *response*

box and time, time^2 and {F}type in the *Terms/Predictors box*. Click on OK and copy and paste the output into *Word*.

d) To make a response plot use the menu commands "Graph&Fit >Plot of". Select y for the V-box and L1:Fit-Values for the H-box. Click on *OK*. When the graph appears, move the OLS slider bar to 3 and the lowess slider bar to 0.5. Since the lowess curve and the OLS cubic fit to $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ nearly coincide, the approximation $E(Y|\boldsymbol{x}) \approx (\boldsymbol{x}^T\boldsymbol{\beta})^3$ seems to be good. Copy the plot into *Word*.

e) From the Graph&Fit menu select *Fit linear LS*, place y in the *response box* and time, time^2, {F}type and From the Graph&Fit menu select *Fit linear LS*, place y in the *response box* and time, time^2, {F}type, {F}type$*$time and {F}type$*$time^2 in the *Terms/Predictors box*. Click on OK and copy and paste the output into *Word*.

f) To make a response plot for a second 1D regression model use the menu commands "Graph&Fit >Plot of". Select y for the V-box and L2:Fit-Values for the H-box. Click on *OK*. When the graph appears, move the OLS slider bar to 2 and the lowess slider bar to 0.5. Since the lowess curve and the OLS quadratic fit to $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ nearly coincide, the approximation $E(Y|\boldsymbol{x}) \approx (\boldsymbol{x}^T\boldsymbol{\beta})^2$ seems to be good. Copy the plot into *Word*.

Chapter 15

Stuff for Students

15.1 Tips for Doing Research

As a student or new researcher, you will probably encounter researchers who think that their method of doing research is the only correct way of doing research, but there are dozens of methods that have proven effective.

Familiarity with the literature is important since your research should be original. The field of high breakdown (HB) robust statistics has perhaps produced more literature in the past 40 years than any other field in statistics.

This text presents the author's applied research in multivariate analysis from 1997–2012, and a summary of the ideas that most influenced the development of this text follows. Gnanadesikan and Kettenring (1972) suggested an algorithm similar to concentration and suggested that robust covariance estimators could be formed by estimating the elements of the covariance matrix with robust scale estimators. Devlin, Gnanadesikan and Kettenring (1975, 1981) introduced the concentration technique. Rousseeuw (1984) extended the MCD location estimator to the MCD estimator of multivariate location and dispersion. Cook and Nachtsheim (1994) showed that robust Mahalanobis distances could be used to reduce the bias of 1D regression estimators. Rousseeuw and Van Driessen (1999) introduced the DD plot.

Much of the HB literature is not applied or consists of ad hoc methods. In far too many papers, the estimator actually used is an ad hoc inconsistent zero breakdown approximation of an estimator for which there is theory. The MCD, depth and MVE estimators are impractical to compute. The S estimators and projection estimators are currently impossible to compute for

$p > 2$. Unless there is a computational breakthrough, these estimators can rarely be used in practical problems. Similarly, two stage estimators need a good initial HB estimator, but no good initial HB estimator was available until Olive (2004a) and Olive and Hawkins (2007, 2008, 2010).

There are hundreds of papers on outlier detection. Most of these compare their method with an existing method on one or two outlier configurations where their method does better. However, the new method rarely outperforms the existing method (such as `lmsreg` or `cov.mcd`) if a broad class of outlier configurations is examined. In such a paper, check whether the new estimator is consistent and if the author has shown types of outlier configurations where the method fails. **Try to figure out how the method would perform for the cases of one and two predictors**.

Dozens of papers suggest that a classical method can be made robust by replacing a classical estimator with a robust estimator. Again inconsistent robust estimators are usually used. These methods can be very useful, but rely on perfect classification of the data into outliers and clean cases. Check whether these methods can find outliers that can not be found by the response plot, FCH DD plot and FMCD DD plot.

For example consider making a robust Hotelling's t–test. If the paper uses the FMCD `cov.mcd` algorithm, then the procedure is relying on the perfect classification paradigm. On the other hand, Srivastava and Mudholkar (2001) present an estimator that has large sample theory.

Beginners can have a hard time determining whether a robust algorithm estimator is consistent or not. As a rule of thumb, assume that the approximations (including those for depth, MCD, MVE, S, projection estimators and two stage estimators) are inconsistent unless the authors show that they understand Hawkins and Olive (2002) and Olive and Hawkins (2007, 2008, 2010). In particular, the elemental or basic resampling algorithms, concentration algorithms and algorithms based on random projections should be considered inconsistent until you can prove otherwise.

After finding a research topic, **paper trailing** is an important technique for finding related literature. To use this technique, find a paper on the topic, go to the bibliography of the paper, find one or more related papers and repeat. Often your university's library will have useful internet resources for finding literature. Usually a research university will subscribe to either *The Web of Knowledge* with a link to ISI Web of Science or to the *Current Index to Statistics*. Both of these resources allow you to search for literature by author,

eg Olive, or by topic, eg robust statistics. Both of these methods search for recent papers. With Web of Knowledge, find an article with *General Search*, click on the article and then click on the *Find Related Articles* icon to get a list of related articles. For papers before 1997, use the free *Current Index to Statistics* website (http://query.statindex.org/CIS/OldRecords/queryOld).

The search engines (www.google.com), (www.ask.com), (www.msn.com), (www.yahoo.com), (www.info.com) and (www.scirus.com) are also useful. The google search engine also has a useful link to "Google Scholar." When searching, enter a topic and the word *robust* or *outliers*. For example, enter the keywords *robust factor analysis* or *factor analysis and outliers*.

The STATLIB site (http://lib.stat.cmu.edu/) is useful for finding statistics departments, data sets and software. Statistical journals often have websites that make abstracts and preprints available. Two useful websites are given below.

```
(www.stat.ucla.edu/journals/ProbStatJournals/)
(www.statsci.org/jourlist.html)
```

Websites for researchers or research groups can be very useful. Below are websites for Dr. Rousseeuw's group, Dr. Rocke, Dr. Croux, Dr. Hubert's group and for the University of Minnesota.

```
(www.agoras.ua.ac.be/)
(http://handel.cipic.ucdavis.edu/~dmrocke/preprints.html)
(www.econ.kuleuven.ac.be/public/NDBAE06/)
(http://wis.kuleuven.be/stat/robust.html)
(www.stat.umn.edu)
```

The latter website has useful links to software. *Arc* and *R* can be downloaded from these links. **Familiarity with a high level programming language** such as FORTRAN or *R/Splus* is essential. A very useful *R* link is (www.r-project.org/#doc). See *R Development Core Team* (2011).

Finally, a Ph.D. student needs an advisor or **mentor** and most researchers will find collaboration valuable. Attending conferences and making your research available over the internet can lead to contacts.

Some references on research, including technical writing and presentations, include American Society of Civil Engineers (1950), Becker and Keller-McNulty (1996), Ehrenberg (1982), Freeman, Gonzalez, Hoaglin and Kilss (1983), Hamada and Sitter (2004), Rubin (2004) and Smith (1997).

15.2 R/Splus and Arc

R is the free version of *Splus*. The website (www.stat.umn.edu) has useful links for *Arc* which is the software developed by Cook and Weisberg (1999a). The website (www.stat.umn.edu) also has a link to **Cran** which gives *R* support. As of April 2012, the author's personal computer has Version 2.13.1 (July 8, 2011) of *R*, Splus–2000 (see Mathsoft 1999ab) and Version 1.06 (July 2004) of *Arc*. Many of the text *R/Splus* functions and figures were made in the 1990's using *Splus* on a workstation.

Downloading the book's data.lsp files into Arc

Many homework problems use data files for *Arc* contained in the book's website (www.math.siu.edu/olive/mbook.htm). As an example, open the *cbrain.lsp* file with *Notepad*. Then use the menu commands "File>Save As". A window appears. On the top "Save in" box change what is in the box to "Removable Disk (G:)" in order to save the file on flash drive G. Then in *Arc* activate the *cbrain.lsp* file with the menu commands "File > Load > Removable Disk (G:) > cbrain.lsp."

Alternatively, open *cbrain.lsp* file with *Notepad*. Then use the menu commands "File>Save As". A window appears. On the top "Save in" box change what is in the box to "My Documents". Then go to *Arc* and use the menu commands "File>Load". A window appears. Change "Arc" to "My Documents" and open *cbrain.lsp*.

Downloading the book's R/Splus functions *mpack.txt* into *R* or *Splus*:

Many of the homework problems use *R/Splus* functions contained in the book's website (www.math.siu.edu/olive/mbook.htm) under the file name *mpack.txt*. Suppose that you download *mpack.txt* onto flashdrive G. Enter *R* and wait for the cursor to appear. Then go to the *File* menu and drag down *Source R Code*. A window should appear. Navigate the *Look in* box until it says *Removable Disk (G:)*. In the *Files of type* box choose *All files(*.*)* and then select *mpack.txt*. The following line should appear in the main *R* window.

```
> source("G:/mpack.txt")
```

Type *ls()*. About 70 *R/Splus* functions from *mpack.txt* should appear.

When you finish your *R/Splus* session, enter the command *q()*. A window asking "*Save workspace image?*" will appear. Click on *No* if you do not want to save the programs in *R*. (If you do want to save the programs then click on *Yes*.)

If you use *Splus*, the command

```
> source("G:/mpack.txt")
```

will enter the functions into *Splus*. Creating a special workspace for the functions may be useful.

This section gives tips on using *R/Splus*, but is no replacement for books such as Becker, Chambers, and Wilks (1988), Braun and Murdoch (2007), Crawley (2005, 2007), or Venables and Ripley (2003). Also see Mathsoft (1999ab) and use the website (www.google.com) to search for useful websites. For example enter the search words *R documentation*.

The command *q()* gets you out of *R* or *Splus*.

Least squares regression is done with the function *lsfit*.

The commands *help(fn)* and *args(fn)* give information about the function fn, eg if fn = lsfit.

Type the following commands.

```
x <- matrix(rnorm(300),nrow=100,ncol=3)
y <- x%*%1:3 + rnorm(100)
out<- lsfit(x,y)
out$coef
ls.print(out)
```

The first line makes a 100 by 3 matrix x with N(0,1) entries. The second line makes $y[i] = 0 + 1 * x[i,1] + 2 * x[i,2] + 3 * x[i,2] + e$ where e is N(0,1). The term 1:3 creates the vector $(1,2,3)^T$ and the matrix multiplication operator is $\%*\%$. The function `lsfit` will automatically add the constant to the model. Typing "out" will give you a lot of irrelevant information, but *out$coef* and *out$resid* give the OLS coefficients and residuals respectively.

To make a residual plot, type the following commands.

```
fit <- y - out$resid
plot(fit,out$resid)
title("residual plot")
```

The first term in the plot command is always the horizontal axis while the second is on the vertical axis.

To put a graph in *Word,* hold down the *Ctrl* and *c* buttons simultaneously. Then select "paste" from the *Word* Edit menu.

To enter data, open a data set in *Notepad* or *Word*. You need to know the number of rows and the number of columns. Assume that each case is entered in a row. For example, assuming that the file *cyp.lsp* has been saved on your disk from the webpage for this book, open *cyp.lsp* in *Word.* It has 76 rows and 8 columns. In *R* or *Splus*, write the following command.

```
cyp <- matrix(scan(),nrow=76,ncol=8,byrow=T)
```

Then copy the data lines from *Word* and paste them in *R/Splus.* If a cursor does not appear, hit *enter.* The command *dim(cyp)* will show if you have entered the data correctly.

Enter the following commands

```
cypy <- cyp[,2]
cypx<- cyp[,-c(1,2)]
lsfit(cypx,cypy)$coef
```

to produce the output below.

```
   Intercept           X1           X2           X3           X4
205.40825985   0.94653718   0.17514405   0.23415181   0.75927197
 X5           X6
 -0.05318671  -0.30944144
```

To check that the data is entered correctly, fit LS in *Arc* with the response variable *height* and the predictors *sternal height, finger to ground, head length, nasal length, bigonal breadth,* and *cephalic index* (entered in that order). You should get the same coefficients given by *R* or *Splus*.

Making functions in R and Splus is easy.

For example, type the following commands.

```
mysquare <- function(x){
# this function squares x
r <- x^2
r }
```

The second line in the function shows how to put comments into functions.

Modifying your function is easy.

Use the fix command.

fix(mysquare)

This will open an editor such as *Notepad* and allow you to make changes.

In *Splus*, the command *Edit(mysquare)* may also be used to modify the function *mysquare*.

To save data or a function in *R*, when you exit, click on *Yes* when the "*Save worksheet image?*" window appears. When you reenter *R*, type *ls().* This will show you what is saved. You should rarely need to save anything for the material in the first thirteen chapters of this book. In *Splus*, data and functions are automatically saved. To remove unwanted items from the worksheet, eg x, type *rm(x)*,
pairs(x) makes a scatterplot matrix of the columns of x,
hist(y) makes a histogram of y,
boxplot(y) makes a boxplot of y,
stem(y) makes a stem and leaf plot of y,
scan(), source(), and *sink()* are useful on a *Unix* workstation.
To type a simple list, use $y <-$ *c(1,2,3.5).*
The commands *mean(y), median(y), var(y)* are self explanatory.

The following commands are useful for a scatterplot created by the command *plot(x,y).*
lines(x,y), lines(lowess(x,y,f=.2))
identify(x,y)
abline(out$coef), abline(0,1)

The usual arithmetic operators are $2 + 4$, $3 - 7$, $8 * 4$, $8/4$, and

```
2^{10}.
```

The ith element of vector y is $y[i]$ while the ij element of matrix x is $x[i, j]$. The second row of x is $x[2,]$ while the 4th column of x is $x[, 4]$. The transpose of x is *t(x).*

The command *apply(x,1,fn)* will compute the row means if fn = mean. The command *apply(x,2,fn)* will compute the column variances if fn = var.

The commands *cbind* and *rbind* combine column vectors or row vectors with an existing matrix or vector of the appropriate dimension.

Downloading the book's R/Splus data sets *robdata.txt* into *R* or *Splus* is done in the same way for downloading *rpack.txt*. Use the following command.

```
> source("G:/mrobdata.txt")
```

For example the command

```
> lsfit(belx,bely)
```

will perform the least squares regression for the Belgian telephone data.

Transferring Data to and from *Arc* and *R* or *Splus*.
For example, suppose that the Belgium telephone data (Rousseeuw and Leroy 1987, p. 26) has the predictor *year* stored in x and the response *number of calls* stored in y in *R* or *Splus*. Combine the data into a matrix z and then use the *write.table* command to display the data set as shown below. The

```
 sep=' '
```

separates the columns by two spaces.

```
> z <- cbind(x,y)
> write.table(data.frame(z),sep='   ')
row.names   z.1   y
 1   50   0.44
 2   51   0.47
 3   52   0.47
 4   53   0.59
 5   54   0.66
 6   55   0.73
 7   56   0.81
 8   57   0.88
 9   58   1.06
 10   59   1.2
 11   60   1.35
 12   61   1.49
 13   62   1.61
```

```
 14   63   2.12
 15   64   11.9
 16   65   12.4
 17   66   14.2
 18   67   15.9
 19   68   18.2
 20   69   21.2
 21   70   4.3
 22   71   2.4
 23   72   2.7073
 24   73   2.9
```

To enter a data set into *Arc*, use the following template *new.lsp*.

```
dataset=new
begin description
Artificial data.
Contributed by David Olive.
end description
begin variables
col 0 = x1
col 1 = x2
col 2 = x3
col 3 = y
end variables
begin data
```

Next open *new.lsp* in *Notepad*. (Or use the *vi* editor in Unix. Sophisticated editors like *Word* will often work, but they sometimes add things like page breaks that do not allow the statistics software to use the file.) Then copy the data lines from *R/Splus* and paste them below *new.lsp*. Then modify the file *new.lsp* and save it on a disk as the file *belg.lsp*. (Or save it in *mdata* where *mdata* is a data folder added within the *Arc data* folder.) The header of the new file *belg.lsp* is shown on the next page.

```
dataset=belgium
begin description
Belgium telephone data from
```

```
Rousseeuw and Leroy (1987, p. 26)
end description
begin variables
col 0 = case
col 1 = x = year
col 2 = y = number of calls in tens of millions
end variables
begin data
1 50 0.44
.  .  .
.  .  .
.  .  .
24 73 2.9
```

The file above also shows the first and last lines of data. The header file needs a data set name, description, variable list and a *begin data* command. Often the description can be copied and pasted from source of the data, eg from the STATLIB website. Note that the first variable starts with *Col 0*.

To transfer a data set from Arc to R or Splus, select the item "Display data" from the dataset's menu. Select the variables you want to save, and then push the button for "Save in R/Splus format." You will be prompted to give a file name. If you select *bodfat*, then two files *bodfat.txt* and *bodfat.Rd* will be created. The file *bodfat.txt* can be read into either *R* or *Splus* using the *read.table* command. The file *bodfat.Rd* saves the documentation about the data set in a standard format for *R*.

As an example, the following command was used to enter the body fat data into *Splus*. (The *mdata folder* does not come with *Arc*. The folder needs to be created and filled with files from the book's website. Then the file *bodfat.txt* can be stored in the mdata folder.)

```
bodfat <- read.table("C:\\ARC\\DATA\\MDATA\\BODFAT.TXT",header=T)
bodfat[,16] <- bodfat[,16]+1
```

The last column of the body fat data consists of the case numbers which start with 0 in *Arc*. The second line adds one to each case number.

As another example, use the menu commands
"File>Load>Data>Arcg>forbes.lsp" to activate the forbes data set. From the *Forbes* menu, select *Display Data*. A window will appear. Double click

on *Temp* and *Pressure*. Click on *Save Data in R/Splus Format* and save as *forbes.txt* in the folder *mdata*.

Enter *Splus* and type the following command.

```
forbes<-read.table("C:\\ARC\\DATA\\ARCG\\FORBES.TXT",header=T)
```

The command *forbes* will display the data set.

Getting information about a library in R

In *R*, a *library* is an add–on package of *R* code. The command *library()* lists all available libraries, and information about a specific library, such as `MASS` for robust estimators like `cov.mcd` or `ts` for time series estimation, can be found, eg, with the command *library(help=MASS)*.

Downloading a library into R

Many researchers have contributed a *library* of *R* code that can be downloaded for use. To see what is available, go to the website (http://cran.us.r-project.org/) and click on the Packages icon. Suppose you are interested the Weisberg (2002) dimension reduction library *dr*. Scroll down the screen and click on *dr*. Then click on the file corresponding to your type of computer, eg *dr 2.0.0.zip* for *Windows*. My unzipped files are stored in my directory

```
C:\unzipped.
```

The file

```
C:\unzipped\dr
```

contains a folder *dr* which is the *R library*. Cut this folder and paste it into the *R* library folder. (On my computer, I store the folder *rw1011* in the file

```
C:\R-Gui.
```

The folder

```
C:\R-Gui\rw1011\library
```

contains the library packages that came with *R*.) Open *R* and type the following command.

library(dr)

Next type *help(dr)* to make sure that the library is available for use.

Warning: R is free but not fool proof. If you have an old version of R and want to download a library, you may need to update your version of R. The libraries for robust statistics may be useful for outlier detection, but the methods have not been shown to be consistent or high breakdown. All software has some bugs. For example, Version 1.1.1 (August 15, 2000) of R had a random generator for the Poisson distribution that produced variates with too small of a mean θ for $\theta \geq 10$. Hence simulated 95% confidence intervals might contain θ 0% of the time. This bug seems to have been fixed in Version 2.4.1.

15.3 Projects

Straightforward Projects

- Read Bentler and Yuan (1998) and Cattell (1966). These papers use scree plots to determine how many eigenvalues of the covariance matrix are nonzero. This topic is very important for dimension reduction methods such as principal components.

- Remark 4.1 estimates the percentage of outliers that the FMCD algorithm can tolerate. In Section 4.5, data is generated such that the FMCD estimator works well for $p = 4$ but fails for $p = 8$. Generate similar data sets for $p = 8, 9, 10, 12, 15, 20, 25, 30, 35, 40, 45,$ and 50. For each value of p find the smallest integer valued percentage of outliers needed to cause the FMCD and FCH estimators to fail. Use the `mpack` function `concsim`. If `concsim` is too slow for large p, use `covsim2` which will only give counts for the fast FCH estimator. As a criterion, a count ≥ 16 is good. Compare these observed FMCD percentages with Remark 4.1 (use the `gamper2` function). Do not forget the *library(MASS)* command if you use R.

- DD plots: compare classical–FCH vs classical–cov.mcd DD plots on real and simulated data. Do problems 4.4, 5.2 and 5.3 but with a wider variety of data sets, n, p and gamma.

- Many papers substitute the latest MCD algorithm for the classical estimator and have titles like "Fast and Robust Factor Analysis." Find such a paper that analyzes a data set on

i) factor analysis,

ii) discriminant analysis,

iii) principal components,

iv) canonical correlation analysis,

v) Hotelling's t test, or

vi) principal component regression.

For the data, make a scatterplot matrix of the classical, RFCH and FMCD Mahalanobis distances. Delete any outliers and run the classical procedure on the undeleted data. Did the paper's procedure perform as well as this procedure?

- Examine the DD plot as a diagnostic for multivariate normality and elliptically contoured distributions. Use real and simulated data.

- Resistant regression: modify `tvreg` by using OLS–covfch instead of OLS–cov.mcd. (L_1–cov.mcd and L_1–covfch are also interesting.) Compare your function with `tvreg`. The `tvreg` and `covfch` functions are in *rpack.txt*.

- *Using ESP to Search for the Missing Link*: Compare `trimmed views` which uses OLS and `cov.mcd` with another regression–MLD combo. There are 8 possible projects: i) OLS–FCH, ii) OLS–Classical (use `ctrviews`), iii) SIR–cov.mcd (`sirviews`), iv) SIR–FCH, v) SIR–classical, vi) lmsreg–cov.mcd (`lmsviews`), vii) lmsreg–FCH, and viii) lmsreg –classical. Do Problem 14.3ac (but just copy and paste the best view instead of using the essp(nx,ncuby,M=40) command) with both your estimator and `trimmed views`. Try to see what types of functions work for both estimators, when `trimmed views` is better and when the procedure i)–viii) in better. If you can invent interesting 1D functions, do so. See Problem 14.4.

- Investigate using trimmed views to make various procedures such as sliced inverse regression resistant against the presence of nonlinearities. The functions *sirviews*, *drsim5*, *drsim6* and *drsim7* in *rpack.txt* may be useful.

- The DGK estimator with 66% coverage should be able to tolerate a cluster of about 30% extremely distant outliers. Compare the DGK estimators with 50% and 66% coverage for various outlier configurations.

Harder Projects

- Which estimator is better FCH, RFCH, CMBA or RCMBA?
- For large data sets, make the DD plot of the DGK estimator vs MB estimator and the DD plot of the classical estimator versus the MB estimator. Which DD plot is more useful? Does your answer depend on n and p? These two plots are among the fastest outlier diagnostics for multivariate data.
- *The Super Duper Outlier Scooper for Multivariate Location and Dispersion:* Consider the modified MBA estimator for multivariate location and dispersion given in Problem 4.7. This MBA estimator uses 8 starts using 0%, 50%, 60%, 70%, 80%, 90%, 95% and 98% trimming of the cases closest to the coordinatewise median in Euclidean distance. The estimator is $\sqrt{n}$ consistent on elliptically contoured distributions with nonsingular covariance matrix. For small data sets the *cmba2* function can fail because the covariance estimator applied to the closest 2% cases to the coordinatewise median is singular. Modify the function so that it works well on small data sets. Then consider the following proposal that may make the estimator asymptotically equivalent to the classical estimator when the data are from a multivariate normal (MVN) distribution. The attractor corresponding to 0% trimming is the DGK estimator $(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$. Let $(\hat{\boldsymbol{\mu}}_T, \hat{\boldsymbol{\Sigma}}_T) = (\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$ if $det(\hat{\boldsymbol{\Sigma}}_0) \leq det(\hat{\boldsymbol{\Sigma}}_M)$ and $(\hat{\boldsymbol{\mu}}_T, \hat{\boldsymbol{\Sigma}}_T) = (\hat{\boldsymbol{\mu}}_M, \hat{\boldsymbol{\Sigma}}_M)$ otherwise where $(\hat{\boldsymbol{\mu}}_M, \hat{\boldsymbol{\Sigma}}_M)$ is the attractor corresponding to M% trimming. Then make the DD plot of the classical Mahalanobis distances versus the distances corresponding to $(\hat{\mu}_T, \hat{\boldsymbol{\Sigma}}_T)$ for $M = 50, 60, 70, 80, 90, 95$ and 98. If all seven DD plots "look good" then use the classical estimator. The resulting estimator will be asymptotically equivalent to the classical estimator if P(all seven DD plots "look good") goes to one as $n \to \infty$. We conjecture that all seven plots will look good because if n is large and the trimmed attractor "beats" the DGK estimator, then the plot will look good. Also if the data is MVN but not spherical, then the DGK estimator will almost always "beat" the trimmed estimator, so all 7 plots will be identical.

- The TV estimator for MLR has a good combination of resistance and theory. Consider the following modification to make the method asymptotically equivalent to OLS when the Gaussian model holds: if each trimmed view "looks good," use OLS. The method is asymptotically equivalent to OLS if the probability P(all 10 trimmed views look good) goes to one as $n \to \infty$. Rousseeuw and Leroy (1987, p. 128) shows that if the predictors are bounded, then the ith residual r_i converges in probability to the ith error e_i for $i = 1, ..., n$. Hence all 10 trimmed views will look like the OLS view with high probability if n is large.

- Compare outliers and missing values, especially missing and outlying at random. See Little and Rubin (2002).

- Suppose that the data set contains missing values. Code the missing value as $\pm 99999 +$ rnorm(1). Run a robust procedure on the data. The idea is that the case with the missing value will be given weight zero if the variable is important, and the variable will be given weight zero if the case is important. See Hawkins and Olive (1999b).

- Download the `dr` function for R, (contributed by Sanford Weisberg), and make PHD and SAVE trimmed views.

- Implement the Carroll and Pederson (1993) robust logistic regression estimator using the robust MLD estimator RFCH or RMVN and see how well the estimator works.

Research Ideas that have Confounded the Author

- If the attractor of a randomly selected elemental start is (in)consistent, then FMCD is (in)consistent. Hawkins and Olive (2002) showed that the attractor is inconsistent if k concentration steps are used. Suppose K elemental starts are used for an MCD concentration estimator and that the starts are iterated until convergence instead of for k steps. Prove or disprove the conjecture that the resulting estimator is inconsistent. (Intuitively, the elemental starts are inconsistent and hence are tilted away from the parameter of interest. Concentration may reduce but probably does not eliminate the tilt.)

- Prove or disprove Conjectures 4.1, 4.2, and 4.3.

- Prove or disprove Conjecture 5.1. Do elemental set and concentration algorithms for multivariate location and dispersion (MLD) give consistent estimators if the number of starts increases to ∞ with the sample size n? (Algorithms that use a fixed number of elemental sets along with the classical estimator and a biased but easily computed high breakdown estimator will be easier to compute and have better statistical properties. See Theorem 4.9 and Olive and Hawkins, 2007, 2008.)

  It is easy to create consistent algorithm estimators that use $O(n)$ randomly chosen elemental sets. He and Wang (1997) show that the all elemental subset approximation to S estimators for MLD is consistent for $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$. Hence an algorithm that randomly draws $g(n)$ elemental sets and searches all $C(g(n), p+1)$ elemental sets is also consistent if $g(n) \to \infty$ as $n \to \infty$. For example, $O(n)$ elemental sets are used if $g(n) \propto n^{1/(p+1)}$.

  When a fixed number of K elemental starts are used, the best attractor is inconsistent but gets close to $(\boldsymbol{\mu}, c_{MCD}\boldsymbol{\Sigma})$ if the data distribution is EC. (The estimator may be unbiased but the variability of the component estimators does not go to 0 as $n \to \infty$.) If $K \to \infty$, then the best attractor should approximate the highest density region arbitrarily closely and the algorithm should be consistent. However, the time for the algorithm greatly increases, the convergence rate is very poor (possibly between $K^{1/2p}$ and $K^{1/p}$), and the elemental concentration algorithm can not guarantee that the determinant is bounded when outliers are present.

- A promising two stage estimator is the "cross checking estimator" that uses a standard consistent estimator and an alternative consistent estimator with desirable properties such as a high breakdown value. The final estimator uses the standard estimator if it is "close" to the alternative estimator, and hence is asymptotically equivalent to the standard estimator for clean data. One important area of research for robust statistics is finding good computable consistent robust estimators to be used in plots and in the cross checking algorithm. The estimators given in Theorems 4.8 and 4.9 (see Olive 2004a and Olive and Hawkins 2007, 2008) finally make the cross checking estimator practical, but better estimators are surely possible. He and Wang (1996) suggested

the cross checking idea for multivariate location and dispersion.

15.4 Hints for Selected Problems

Chapter 1

1.1 a) $8.25 \pm 0.7007 = (6.020, 10.480)$

b) $8.75 \pm 1.1645 = (7.586, 9.914)$.

1.2 a) $\overline{Y} = 24/5 = 4.8$.

b)

$$S^2 = \frac{138 - 5(4.8)^2}{4} = 5.7$$

so $S = \sqrt{5.7} = 2.3875$.

c) The ordered data are 2,3,5,6,8 and MED$(n) = 5$.

d) The ordered $|Y_i - \text{MED}(n)|$ are 0,1,2,2,3 and MAD$(n) = 2$.

1.2 a) $\overline{Y} = 15.8/10 = 1.58$.

b)

$$S^2 = \frac{38.58 - 10(1.58)^2}{9} = 1.5129$$

so $S = \sqrt{1.5129} = 1.230$.

c) The ordered data set is 0.0,0.8,1.0,1.2,1.3,1.3,1.4,1.8,2.4,4.6 and MED$(n) = 1.3$.

d) The ordered $|Y_i - \text{MED}(n)|$ are 0,0,0.1,0.1,0.3,0.5,0.5,1.1,1.3,3.3 and MAD$(n) = 0.4$.

e) 4.6 is unusually large.

Chapter 2

Chapter 3

3.1 a) $X_2 \sim N(100, 6)$.

b)

$$\begin{pmatrix} X_1 \\ X_3 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 49 \\ 17 \end{pmatrix}, \begin{pmatrix} 3 & -1 \\ -1 & 4 \end{pmatrix} \right).$$

c) $X_1 \perp\!\!\!\perp X_4$ and $X_3 \perp\!\!\!\perp X_4$.

d)

$$\rho(X_1, X_2) = \frac{Cov(X_1, X_3)}{\sqrt{\text{VAR}(X_1)\text{VAR}(X_3)}} = \frac{-1}{\sqrt{3}\sqrt{4}} = -0.2887.$$

3.2 a) $Y|X \sim N(49, 16)$ since $Y \perp\!\!\!\perp X$. (Or use $E(Y|X) = \mu_Y + \Sigma_{12}\Sigma_{22}^{-1}(X - \mu_x) = 49 + 0(1/25)(X - 100) = 49$ and $\text{VAR}(Y|X) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 16 - 0(1/25)0 = 16$.)

b) $E(Y|X) = \mu_Y + \Sigma_{12}\Sigma_{22}^{-1}(X - \mu_x) = 49 + 10(1/25)(X - 100) = 9 + 0.4X$.

c) $\text{VAR}(Y|X) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 16 - 10(1/25)10 = 16 - 4 = 12$.

3.4 The proof is identical to that given in Example 3.2. (In addition, it is fairly simple to show that $M_1 = M_2 \equiv M$. That is, M depends on Σ but not on c or g.)

3.6 a) Sort each column, then find the median of each column. Then $\text{MED}(\boldsymbol{W}) = (1430, 180, 120)^T$.

b) The sample mean of $(X_1, X_2, X_3)^T$ is found by finding the sample mean of each column. Hence $\overline{\boldsymbol{x}} = (1232.8571, 168.00, 112.00)^T$.

3.11 $\boldsymbol{\Sigma B} = E[E(\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X})\boldsymbol{X}^T\boldsymbol{B})] = E(\boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{B}) = \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{\Sigma B}$. Hence $\boldsymbol{M}_B = \boldsymbol{\Sigma B}(\boldsymbol{B}^T\boldsymbol{\Sigma B})^{-1}$.

Chapter 4

4.4 The 4 plots should look nearly identical with the five cases 61–65 appearing as outliers.

4.5 Not only should none of the outliers be highlighted, but the highlighted cases should be ellipsoidal.

4.6 Answers will vary since this is simulated data, but should get gam near 0.4, 0.3, 0.2 and 0.1 as p increases from 2 to 20.

Chapter 5

5.2 b Ideally the answer to this problem and Problem 5.3b would be nearly the same, but students seem to want correlations to be very high and use n too high. Values of n around 20, 40 and 50 for $p = 2, 3$ and 4 should be enough.

5.3 b Values of n should be near 20, 40 and 50 for $p = 2, 3$ and 4.

5.4 This is simulated data, but for most plots the slope is near 2.

Chapter 6

6.1 Note that $o_P(1)O_P(1) = [(\hat{\boldsymbol{\Sigma}} - \hat{\lambda}_i) - c(\boldsymbol{\Sigma} - \lambda_i)]\hat{\boldsymbol{e}}_i = c(\boldsymbol{\Sigma} - \lambda_i)\hat{\boldsymbol{e}}_i \xrightarrow{P} \mathbf{0}$.

Chapter 7

Chapter 8

Chapter 9

Chapter 10

Chapter 11

Chapter 12

Chapter 13

Chapter 14

14.6 The identity line should NOT PASS through the cluster of outliers with Y near 0. The amount of trimming seems to vary some with the computer (which should not happen unless there is a bug in the `tvreg2` function or if the computers are using different versions of `cov.mcd`), but most students liked 70% or 80% trimming.

15.5 F Table

Tabled values are F(0.95,k,d) where $P(F < F(0.95, k, d)) = 0.95$.
00 stands for ∞. Entries produced with the `qf(.95,k,d)` command in *R*. The numerator degrees of freedom are k while the denominator degrees of freedom are d.

k	1	2	3	4	5	6	7	8	9	00
d										
1	161	200	216	225	230	234	237	239	241	254
2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.41
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	1.62
00	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.00

Agulló, J. (1996), "Exact Iterative Computation of the Multivariate Minimum Volume Ellipsoid Estimator with a Branch and Bound Algorithm," In *Proceedings in Computational Statistics,* Prat, A., Physica-Verlag, Heidelberg, 175-180.

Agulló, J., (1998), "Computing the Minimum Covariance Determinant Estimator," unpublished manuscript, Universidad de Alicante.

Aldrin, M., Bϕlviken, E., and Schweder, T. (1993), "Projection Pursuit Regression for Moderate Non-linearities," *Computational Statistics and Data Analysis,* 16, 379-403.

Alkenani, A., and Yu, K. (2012), "A Comparative Study for Robust Canonical Correlation Methods" *Journal of Statistical Computation and Simulation,* to appear.

Alqallaf, F.A. Konis, K.P., Martin, R.D., and Zamar, R.H. (2002), "Scalable Robust Covariance and Correlation Estimates for Data Mining," In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Edmonton.

American Society of Civil Engineers (1950), "So You're Going to Present a Paper," *The American Statistician* 4, 6-8.

Anderson, M.J. (2001), "A New Method for Non-parametric Multivariate Analysis of Variance," *Austral Ecology*, 26, 32-46.

Anderson, T.W. (1984, 2003), *An Introduction to Multivariate Statistical Analysis,* 2nd and 3rd ed., Wiley, New York, NY.

Arcones, M.A. (1995), "Asymptotic Normality of Multivariate Trimmed Means," *Statistics and Probability Letters,* 25, 43-53.

Ash, R.B. (1972), *Real Analysis and Probability,* Academic Press, San Diego, CA.

Atkinson, A., Riani, R., and Cerioli, A. (2004), *Exploring Multivariate Data with the Forward Search*, Springer-Verlag, New York, NY.

Bali, J.L., Boente, G., Tyler, D.E. and Wang, J.L. (2011), "Robust Functional Principal Components: a Projection-Pursuit Approach," *The Annals of Statistics*, 39, 2852-2882.

Barndorff-Nielsen, O. (1982), "Exponential Families," in *Encyclopedia of Statistical Sciences,* Vol. 2, eds. Kotz, S. and Johnson, N.L., Wiley, New York, NY, 587-596.

Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988), *The New S Language A Programming Environment for Data Analysis and Graphics,* Wadsworth and Brooks/Cole, Pacific Grove, CA.

Becker, R.A., and Keller-McNulty, S. (1996), "Presentation Myths," *The American Statistician*, 50, 112-115.

Bentler, P.M., and Yuan, K.H. (1998), "Tests for Linear Trend in the Smallest Eigenvalues of the Correlation Matrix," *Psychometrika,* 63, 131-144.

Berk, K.N. (1978), "Comparing Subset Regression Procedures," *Technometrics,* 20, 1-6.

Bernholt, T., and Fischer, P. (2004), "The Complexity of Computing the MCD-Estimator," *Theoretical Computer Science*, 326, 383-398.

Bhatia, R., Elsner, L., and Krause, G. (1990), "Bounds for the Variation of the Roots of a Polynomial and the Eigenvalues of a Matrix," *Linear Algebra and Its Applications*, 142, 195-209.

Billor, N., Hadi, A., and Velleman, P. (2000), "Bacon: Blocked Adaptive Computationally Efficient Outlier Nominators," *Computational Statistics & Data Analysis*, 34, 279-298.

Bishop, C.M. (2006), *Pattern Recognition and Machine Learning*, Springer Science, New York, NY.

Boente, G. (1987), "Asymptotic Theory for Robust Principal Components," *Journal of Multivariate Analysis*, 21, 67-78.

Boente, G., and Fraiman, R. (1999), "Discussion of 'Robust Principal Component Analysis for Functional Data' by Locantore et al," *Test*, 8, 28-35.

Bogdan, M. (1999), "Data Driven Smooth Tests for Bivariate Normality," *Journal of Multivariate Analysis,* 68, 26-53.

Box, G.E.P. (1990), "Commentary on 'Communications between Statisticians and Engineers/Physical Scientists' by H.B. Hoadley and J.R. Kettenring," *Technometrics,* 32, 251-252.

Box, G.E.P., and Cox, D.R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, B,*26, 211-246.

Box, G.E.P, Hunter, J.S., and Hunter, W.G. (2005), *Statistics for Experimenters*, 2nd ed., Wiley, New York, NY.

Braun, W.J., and Murdoch, D.J. (2007), *A First Course in Statistical Programming with R*, Cambridge University Press, New York, NY.

Brillinger, D.R. (1977), "The Identification of a Particular Nonlinear Time Series," *Biometrika,* 64, 509-515.

Brillinger, D.R. (1983), "A Generalized Linear Model with "Gaussian" Regressor Variables," in *A Festschrift for Erich L. Lehmann,* eds. Bickel,

P.J., Doksum, K.A., and Hodges, J.L., Wadsworth, Pacific Grove, CA, 97-114.

Brown, M.B., and Forsythe, A.B. (1974a), "The ANOVA and Multiple Comparisons for Data with Heterogeneous Variances," *Biometrics*, 30, 719-724.

Brown, M.B., and Forsythe, A.B. (1974b), "The Small Sample Behavior of Some Statistics Which Test the Equality of Several Means," *Technometrics*, 16, 129-132.

Butler, R.W., Davies, P.L., and Jhun, M. (1993), "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics,* 21, 1385-1400.

Buxton, L.H.D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

Cai, T., and Liu, W. (2011), "A Direct Approach to Sparse Linear Discriminant Analysis," *Journal of the American Statistical Association*, 106, 1566-1577.

Cambanis, S., Huang, S., and Simons, G. (1981), "On the Theory of Elliptically Contoured Distributions," *Journal of Multivariate Analysis,* 11, 368-385.

Carroll, R.J., and Pederson, S. (1993), "On Robustness in the Logistic Regression Model," *Journal of the Royal Statistical Society, B,* 55, 693-706.

Casella, G., and Berger, R.L. (2002), *Statistical Inference,* 2nd ed., Duxbury, Belmont, CA.

Cator, E.A., and Lopuhaä, H.P. (2009), "Central Limit Theorem and Influence Function for the MCD Estimators at General Multivariate Distributions," preprint. See (http://arxiv.org/abs/0907.0079).

Cator, E.A., and Lopuhaä, H.P. (2010), "Asymptotic Expansion of the Minimum Covariance Determinant Estimators," *Journal of Multivariate Analysis,* 101, 2372-2388.

Cattell, R.B. (1966), "The Scree Test for the Number of Factors," *Multivariate Behavioral Research,* 1, 245-276.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P. (1983), *Graphical Methods for Data Analysis,* Duxbury Press, Boston, MA.

Chang, J., and Olive, D.J. (2007), *Resistant Dimension Reduction*, Preprint, see (www.math.siu.edu/olive/preprints.htm).

Chang, J., and Olive, D.J. (2010), "OLS for 1D Regression Models," *Communications in Statistics: Theory and Methods*, 39, 1869-1882.

Chmielewski, M.A. (1981), "Elliptically Symmetric Distributions: a Review and Bibliography," *International Statistical Review,* 49, 67-74.

Cobb, G.W. (1998), *Introduction to Design and Analysis of Experiments*, Key College Publishing, Emeryville, CA.

Cook, R.D. (1998), *Regression Graphics: Ideas for Studying Regression Through Graphics,* Wiley, New York, NY.

Cook, R.D., and Hawkins, D.M. (1990), "Comment on 'Unmasking Multivariate Outliers and Leverage Points' by P.J. Rousseeuw and B.C. van Zomeren," *Journal of the American Statistical Association,* 85, 640-644.

Cook, R.D., Hawkins, D.M., and Weisberg, S. (1992), "Comparison of Model Misspecification Diagnostics Using Residuals from Least Mean of Squares and Least Median of Squares," *Journal of the American Statistical Association,* 87, 419-424.

Cook, R.D., Hawkins, D.M., and Weisberg, S. (1993), "Exact Iterative Computation of the Robust Multivariate Minimum Volume Ellipsoid Estimator," *Statistics and Probability Letters,* 16, 213-218.

Cook, R.D., and Li, B. (2002), "Dimension Reduction for Conditional Mean in Regression," *The Annals of Statistics,* 30, 455-474.

Cook, R.D., and Nachtsheim, C.J. (1994), "Reweighting to Achieve Elliptically Contoured Covariates in Regression," *Journal of the American Statistical Association,* 89, 592-599.

Cook, R.D., and Olive, D.J. (2001), "A Note on Visualizing Response Transformations in Regression," *Technometrics,* 43, 443-449.

Cook, R.D., and Setodji, C.M. (2003), "A Model-Free Test for Reduced Rank in Multivariate Regression," *Journal of the American Statistical Association*, 98, 340351.

Cook, R.D., and Weisberg, S. (1999a), *Applied Regression Including Computing and Graphics,* Wiley, New York, NY.

Cook, R.D., and Weisberg, S. (1999b), "Graphs in Statistical Analysis: is the Medium the Message?" *The American Statistician,* 53, 29-37.

Copas, J.B. (1983), "Regression Prediction and Shrinkage," (with discussion), *Journal of the Royal Statistical Society, B*, 45, 311-354.

Cornish, E.A. (1954), "The Multivariate t-Distribution Associated with a Set of Normal Sample Deviates," *Australian Journal of Physics*, 7, 531-542.

Cramér, H. (1946), *Mathematical Methods of Statistics,* Princeton University Press, Princeton, NJ.

Crawley, M.J. (2005), *Statistics an Introduction Using R,* Wiley, Hoboken, NJ.

Crawley, M.J. (2007), *The R Book,* Wiley, Hoboken, NJ.

Croux, C., Dehon, C., Rousseeuw, P.J., and Van Aelst, S. (2001), "Robust Estimation of the Conditional Median Function at Elliptical Models," *Statistics and Probability Letters,* 51, 361-368.

Croux, C., Dehon, C., and Yadine, A. (2010), "The k-step Spatial Sign Covariance Matrix," *Advances in Data Analysis and Classification*, 4, 137-150.

Croux, C., and Van Aelst, S. (2002), "Comment on 'Nearest-Neighbor Variance Estimation (NNVE): Robust Covariance Estimation via Nearest-Neighbor Cleaning' by N. Wang and A.E. Raftery," *Journal of the American Statistical Association,* 97, 1006-1009.

Czörgö, S. (1986), "Testing for Normality in Arbitrary Dimension," *The Annals of Statistics,* 14, 708-723.

DasGupta, A. (2008), *Asymptotic Theory of Statistics and Probability*, Springer-Verlag, New York, NY.

Datta, B.N. (1995), *Numerical Linear Algebra and Applications,* Brooks/Cole Publishing Company, Pacific Grove, CA.

Davidson, J. (1994), *Stochastic Limit Theory*, Oxford University Press, Oxford, UK.

Davies, P.L. (1992), "Asymptotics of Rousseeuw's Minimum Volume Ellipsoid Estimator," *The Annals of Statistics,* 20, 1828-1843.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1975), "Robust Estimation and Outlier Detection with Correlation Coefficients," *Biometrika,* 62, 531-545.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1981), "Robust Estimation of Dispersion Matrices and Principal Components," *Journal of the American Statistical Association,* 76, 354-362.

Duda, R.O., Hart, P.E., and Stork, D.G. (2000), *Pattern Classification,* 2nd ed., Wiley, New York, NY.

Easton, G.S., and McCulloch, R.E. (1990), "A Multivariate Generalization of Quantile Quantile Plots," *Journal of the American Statistical Association,* 85, 376-386.

Eaton, M.L. (1986), "A Characterization of Spherical Distributions," *Journal of Multivariate Analysis,* 20, 272-276.

Eaton, M.L., and Tyler, D.E. (1991), "On Wielands's Inequality and its Application to the Asymptotic Distribution of the Eigenvalues of a Random Symmetric Matrix," *The Annals of Statistics*, 19, 260-271.

Ehrenberg, A.S.C. (1982), "Writing Technical Papers or Reports," *The American Statistician*, 36, 326-329.

Ernst, M.D. (2009), "Teaching Inference for Randomized Experiments," *Journal of Statistical Education*, 17, (online).

Fang, K.T., and Anderson, T.W. (editors) (1990), *Statistical Inference in Elliptically Contoured and Related Distributions,* Allerton Press, New York, NY.

Fang, K.T., Kotz, S., and Ng, K.W. (1990), *Symmetric Multivariate and Related Distributions,* Chapman & Hall, New York, NY.

Ferguson, T.S. (1996), *A Course in Large Sample Theory*, Chapman & Hall, New York, NY.

Flury, B., and Riedwyl, H. (1988), *Multivariate Statistics: a Practical Approach*, Chapman & Hall, London, UK.

Fox, J. (1991), *Regression Diagnostics,* Sage Publications, Newbury Park, CA.

Freeman, D.H., Gonzalez, M.E., Hoaglin, D.C., and Kilss, B.A. (1983), "Presenting Statistical Papers," *The American Statistician,* 37, 106-110.

Furnival, G., and Wilson, R. (1974), "Regression by Leaps and Bounds," *Technometrics,* 16, 499-511.

García-Escudero, L.A., and Gordaliza, A. (2005), "Generalized Radius Processes for Elliptically Contoured Distributions," *Journal of the American Statistical Association*, 100, 1036-1045.

Gladstone, R.J. (1905), "A Study of the Relations of the Brain to the Size of the Head," *Biometrika,* 4, 105-123.

Gnanadesikan, R. (1977, 1997), *Methods for Statistical Data Analysis of Multivariate Observations*, 1st and 2nd ed., Wiley, New York, NY.

Gnanadesikan, R., and Kettenring, J.R. (1972), "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data," *Biometrics,* 28, 81-124.

Good, P.I. (2012), *A Practitioner's Guide to Resampling for Data Analysis, Data Mining, and Modeling*, Chapman & Hall/CRC, Boca Raton, FL.

Grimm, L.G., and Yarnold, P.R. (Eds.) (1995), *Reading and Understanding Multivariate Statistics*, American Psychological Association, Washington, D.C.

Grimm, L.G., and Yarnold, P.R. (Eds.) (2000), *Reading and Understanding More Multivariate Statistics*, American Psychological Association, Washington, D.C.

Grofman, B. (1981), "Fair Apportionment and the Banzhaf Index," *The American Mathematical Monthly,*" 88, 1-5.

Gupta, A.K., and Varga, T. (1993), *Elliptically Contoured Models in Statistics,* Kluwar Academic Publishers, Dordrecht, The Netherlands.

Hair, J.F., Black, B., Anderson, R.E., and Tatham, R.L. (2005), *Multivariate Data Analysis*, 6th ed., Prentice Hall, Upper Saddle River, NJ.

Hamada, M., and Sitter, R. (2004), "Statistical Research: Some Advice for Beginners," *The American Statistician,* 58, 93-101.

Hand, D.J. (2006), "Classifier Technology and the Illusion of Progress," (with discussion), *Statistical Science*, 21, 1-34.

Hawkins, D.M. (1993), "A Feasible Solution Algorithm for the Minimum Volume Ellipsoid Estimator in Multivariate Data," *Computational Statistics,* 9, 95-107.

Hawkins, D.M. (1994), "The Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator in Multivariate Data, *Computational Statistics and Data Analysis*, 17, 197-210.

Hawkins, D.M., and Olive, D.J. (1999a), "Improved Feasible Solution Algorithms for High Breakdown Estimation," *Computational Statistics and Data Analysis,* 30, 1-11.

Hawkins, D.M., and Olive, D. (1999b), "Applications and Algorithms for Least Trimmed Sum of Absolute Deviations Regression," *Computational Statistics and Data Analysis,* 32, 119-134.

Hawkins, D.M., and Olive, D.J. (2002), "Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm," (with discussion), *Journal of the American Statistical Association,* 97, 136-159.

Hawkins, D.M., and Simonoff, J.S. (1993), "High Breakdown Regression and Multivariate Estimation," *Applied Statistics,* 42, 423-432.

He, X., and Wang, G. (1996), "Cross-Checking Using the Minimum Volume Ellipsoid Estimator," *Statistica Sinica,* 6, 367-374.

He, X., and Wang, G. (1997), "Qualitative Robustness of S*- Estimators of Multivariate Location and Dispersion," *Statistica Neerlandica*, 51, 257-268.

Henderson, H.V., and Searle, S.R. (1977), "Vec and Vech Operators for Matrices, with Some Uses in Jacobians and Multivariate Statistics," *The Canadian Journal of Statistics,* 7, 65-81.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (eds.) (1991), *Fundamentals of Exploratory Analysis of Variance*, Wiley, New York, NY.

Hoeffding, W. (1952), "The Large Sample Power of Tests Based on Permutations of Observations," *The Annals of Mathematical Statistics,* 23, 169-192.

Hosmer, D.W., and Lemeshow, S. (2000), *Applied Logistic Regression*, 2nd ed., Wiley, New York, NY.

Huber, P.J. (1981), *Robust Statistics,* Wiley, New York, NY.

Huber, P.J., and Ronchetti, E.M. (2009), *Robust Statistics,* 2nd ed., Wiley, Hoboken, NJ.

Hubert, M. (2001), "Discussion of 'Multivariate Outlier Detection and Robust Covariance Matrix Estimation' by D. Peña and F.J. Prieto," *Technometrics,* 43, 303-306.

Hubert, M., Rousseeuw, P.J., and Van Aelst, S. (2002), "Comment on 'Inconsistency of Resampling Algorithms for High Breakdown Regression and a New Algorithm' by D.M. Hawkins and D.J. Olive," *Journal of the American Statistical Association,* 97, 151-153.

Hubert, M., Rousseeuw, P.J., and Van Aelst, S. (2008), "High Breakdown Multivariate Methods," *Statistical Science*, 23, 92-119.

Hubert, M., Rousseeuw, P.J., and Verdonck, T. (2012), "A Deterministic Algorithm for Robust Location and Scatter," *Journal of Computational and Graphical Statistics,* 21, 618-637.

Huberty, C.J., and Olejnik, S. (2006), *Applied MANOVA and Discriminant Analysis*, 2nd ed., Wiley, Hoboken, NJ.

Jiang, J. (2010) *Large Sample Techniques for Statistics*, Springer, New York, NY.

Johnson, M.E. (1987), *Multivariate Statistical Simulation,* Wiley, New York, NY.

Johnson, N.L., and Kotz, S. (1972), *Distributions in Statistics: Continuous Multivariate Distributions,* Wiley, New York, NY.

Johnson, R.A., and Wichern, D.W. (1988, 2007), *Applied Multivariate Statistical Analysis,* 2nd and 6th ed., Prentice Hall, Englewood Cliffs, NJ.

Jolliffe, I.T. (2010), *Principal Component Analysis*, 2nd ed., Springer, New York, NY.

Jones, H.L., (1946), "Linear Regression Functions with Neglected Variables," *Journal of the American Statistical Association,* 41, 356-369.

Kachigan, S.K. (1991), *Multivariate Statistical Analysis: A Conceptual Introduction*, 2nd ed., Radius Press, New York, NY.

Kelker, D. (1970), "Distribution Theory of Spherical Distributions and a Location Scale Parameter Generalization," *Sankhya, A*, 32, 419-430.

Kendall, M. (1980), *Multivariate Analysis*, 2nd ed., Macmillan Publishing, New York, NY.

Khattree, R., and Naik, D.N. (1999), *Applied Multivariate Statistics with SAS Software,* 2nd ed., SAS Institute, Cary, NC.

Kim, J. (2000), "Rate of Convergence of Depth Contours: with Application to a Multivariate Metrically Trimmed Mean," *Statistics and Probability Letters,* 49, 393-400.

Kirk, R.E. (1982), *Experimental Design: Procedures for the Behavioral Sciences*, 2nd ed., Brooks/Cole Publishing Company, Belmont, CA.

Koltchinskii, V.I., and Li, L. (1998), "Testing for Spherical Symmetry of a Multivariate Distribution," *Journal of Multivariate Analysis,* 65, 228-244.

Kosfeld, R. (1996), "Robust Exploratory Factor Analysis," *Statistical Papers*, 37 (2): 105-122.

Kowalski, C.J. (1973), "Non-normal Bivariate Distributions with Normal Marginals," *The American Statistician,* 27, 103-106.

Kuehl, R.O. (1994), *Statistical Principles of Research Design and Analysis*, Duxbury Press, Belmont, CA.

Ledolter, J., and Swersey, A.J. (2007), *Testing 1-2-3 Experimental Design with Applications in Marketing and Service Operations,* Stanford University Press, Stanford, CA.

Lehmann, E.L. (1983), *Theory of Point Estimation,* Wiley, New York, NY.

Lehmann, E.L. (1999), *Elements of Large–Sample Theory*, Springer-Verlag, New York, NY.

Leon, S.J. (1986), *Linear Algebra with Applications*, 2nd ed., Macmillan Publishing Company, New York, NY.

Li, K.C., and Duan, N. (1989), "Regression Analysis Under Link Violation," *The Annals of Statistics,* 17, 1009-1052.

Li, R., Fang, K., and Zhu, L. (1997), "Some Q-Q Probability Plots to Test Spherical and Elliptical Symmetry," *Journal of Computational and Graphical Statistics,* 6, 435-450.

Little, R.J.A., and Rubin, D.B. (2002), *Statistical Analysis with Missing Data,* 2nd ed., Wiley, New York, NY.

Liu, R.Y., Parelius, J.M., and Singh, K. (1999), "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics, and Inference," *The Annals of Statistics,* 27, 783-858.

Locantore, N., Marron, J.S., Simpson, D.G., Tripoli, N., Zhang, J.T., and Cohen, K.L. (1999), "Robust Principal Component Analysis for Functional

Data," (with discussion), *Test*, 8, 1-73.

Lopuhaä, H.P. (1999), "Asymptotics of Reweighted Estimators of Multivariate Location and Scatter," *The Annals of Statistics,* 27, 1638-1665.

Mai, Q., Zou, H., and Yuan, M. (2012), "A Direct Approach to Sparse Discriminant Analysis in Ultra-High Dimensions," *Biometrika*, 99, 29-42.

Mallows, C. (1973), "Some Comments on C_p," *Technometrics,* 15, 661-676.

Manzotti, A., Pérez, F.J., and Quiroz, A.J. (2002), "A Statistic for Testing the Null Hypothesis of Elliptical Symmetry," *Journal of Multivariate Analysis,* 81, 274-285.

Mardia, K.V. (1971), "The Effect of Nonnormality on Some Multivariate Tests of Robustness to Nonnormality in the Linear Model," *Biometrika*, 58, 105-121.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis,* Academic Press, London.

Maronna, R.A., Martin, R.D., and Yohai, V.J. (2006), *Robust Statistics: Theory and Methods*, Wiley, Hoboken, NJ.

Maronna, R.A., and Yohai, V.J. (2002), "Comment on 'Inconsistency of Resampling Algorithms for High Breakdown Regression and a New Algorithm' by D.M. Hawkins and D.J. Olive," *Journal of the American Statistical Association,* 97, 154-155.

Maronna, R.A., and Zamar, R.H. (2002), "Robust Estimates of Location and Dispersion for High-Dimensional Datasets," *Technometrics,* 44, 307-317.

MathSoft (1999a), *S-Plus 2000 User's Guide,* Data Analysis Products Division, MathSoft, Seattle, WA. (Mathsoft is now Insightful.)

MathSoft (1999b), *S-Plus 2000 Guide to Statistics,* Volume 2, Data Analysis Products Division, MathSoft, Seattle, WA. (Mathsoft is now Insightful.)

McDonald, G.C., and Schwing, R.C. (1973), "Instabilities of Regression Estimates Relating Air Pollution to Mortality," *Technometrics*, 15, 463-482.

McLachlan, G.J. (2004), *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, Hoboken, NJ.

Mehrotra, D.V. (1995), "Robust Elementwise Estimation of a Dispersion Matrix," *Biometrics*, 51, 1344-1351.

Miller, D.M. (1984), "Reducing Transformation Bias in Curve Fitting," *The American Statistician*, 38, 124-126.

Mϕller, S.F., von Frese, J., and Bro, R. (2005), "Robust Methods for Multivariate Data Analysis," *Journal of Chemometrics,* 19, 549-563.

Montgomery, D.C. (1984), *Design and Analysis of Experiments,* 2nd ed., Wiley, New York, NY.

Moore, D.S. (2000), *The Basic Practice of Statistics,* 2nd ed., W.H. Freeman, New York, NY.

Mosteller, F., and Tukey, J.W. (1977), *Data Analysis and Regression,* Addison-Wesley, Reading, MA.

Muirhead, R.J. (1982), *Aspects of Multivariate Statistical Theory,* Wiley, New York, NY.

Muirhead, R.J., and Waternaux, C.M. (1980), "Asymptotic Distribution in Canonical Correlation Analysis and Other Multivariate Procedures for Nonnormal Populations," *Biometrika*, 67, 31-43.

Olive, D.J. (2002), "Applications of Robust Distances for Regression," *Technometrics,* 44, 64-71.

Olive, D.J. (2004a), "A Resistant Estimator of Multivariate Location and Dispersion," *Computational Statistics and Data Analysis*, 46, 99-102.

Olive, D.J. (2004b), "Visualizing 1D Regression," in *Theory and Applications of Recent Robust Methods*, eds. Hubert, M., Pison, G., Struyf, A., and Van Aelst, S., Series: Statistics for Industry and Technology, Birkhäuser, Basel, Switzerland, 221-233.

Olive, D.J. (2005), "Two Simple Resistant Regression Estimators," *Computational Statistics and Data Analysis*, 49, 809-819.

Olive, D.J. (2007), "Prediction Intervals for Regression Models," *Computational Statistics and Data Analysis,* 51, 3115-3122.

Olive, D.J. (2010), *Multiple Linear and 1D Regression Models*, Unpublished Online Text available from (www.math.siu.edu/olive/regbk.htm).

Olive, D.J. (2011b), *The Number of Samples for Resampling Algorithms*, Preprint, see (www.math.siu.edu/olive/).

Olive, D.J. (2012a), "Why the Rousseeuw Yohai Paradigm is One of the Largest and Longest Running Scientific Hoaxes in History," unpublished manuscript available from (www.math.siu.edu/olive/).

Olive, D.J. (2012b), *A Course in Statistical Theory*, Unpublished manuscript available from (www.math.siu.edu/olive/).

Olive, D.J. (2013a), "Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data," *International Journal of Statistics and Probability*, 2, 90-100.

Olive, D.J. (2013b), "Plots for Generalized Additive Models," *Communications in Statistics: Theory and Methods*, 41, to appear.

Olive, D.J. (2014), *Robust Statistics*, To Appear.

Olive, D.J., and Hawkins, D.M. (2003), "Robust Regression with High Coverage," *Statistics and Probability Letters,* 63, 259-266.

Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Olive, D.J., and Hawkins, D.M. (2007), "Robustifying Robust Estimators," Preprint, see (www.math.siu.edu/olive/preprints.htm).

Olive, D.J., and Hawkins, D.M. (2008), "High Breakdown Multivariate Estimators," Preprint, see (www.math.siu.edu/olive/preprints.htm).

Olive, D.J., and Hawkins, D.M. (2010), "Robust Multivariate Location and Dispersion," Preprint, see (www.math.siu.edu/olive/preprints.htm).

Olive, D.J., and Hawkins, D.M. (2011), "Practical High Breakdown Regression," Preprint, see (www.math.siu.edu/olive/preprints.htm).

Peña, D., and Prieto, F.J. (2001), "Multivariate Outlier Detection and Robust Covariance Matrix Estimation," *Technometrics,* 286-299.

Pesch, C. (1999), "Computation of the Minimum Covariance Determinant Estimator," in *Classification in the Information Age, Proceedings of the 22nd Annual GfKl Conference, Dresden 1998,* eds. Gaul W., and Locarek-Junge, H., Springer, Berlin, 225232.

Polansky, A.M. (2011), *Introduction to Statistical Limit Theory*, CRC Press, Boca Rotan, FL.

Poor, H.V. (1988), *An Introduction to Signal Detection and Estimation,* Springer-Verlag, New York, NY.

Pratt, J.W. (1959), "On a General Concept of 'in Probability'," *The Annals of Mathematical Statistics,* 30, 549-558.

Press, S.J. (2005), *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*, 2nd ed., Dover Publications, Mineola, NY.

R Development Core Team (2011), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rao, C.R. (1965, 1973), *Linear Statistical Inference and Its Applications,* 1st and 2nd ed., Wiley, New York, NY.

Raveh, A. (1989), "A Nonparametric Approach to Linear Discriminant Analysis," *Journal of the American Statistical Association*, 84, 176-183.

Rencher, A., and Pun, F. (1980), "Inflation of R^2 in Best Subset Regression," *Technometrics*, 22, 49-53.

Rencher, A.C. (2002), *Methods of Multivariate Analysis*, 2nd ed., Wiley, New York, NY.

Reyen, S.S., Miller, J.J., and Wegman, E.J. (2009), "Separating a Mixture of Two Normals with Proportional Covariances," *Metrika*, 70, 297-314.

Riani, M., Atkinson, A.C., and Cerioli, A. (2009), "Finding an Unknown Number of Outliers," *Journal of the Royal Statistical Society, B*, 71, 447-466.

Rocke, D.M., and Woodruff, D.L. (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association,* 91, 1047-1061.

Rocke, D.M., and Woodruff, D.L. (2001), "Discussion of 'Multivariate Outlier Detection and Robust Covariance Matrix Estimation' by D. Peña and F.J. Prieto," *Technometrics,* 43, 300-303.

Rohatgi, V.K. (1976), *An Introduction to Probability Theory and Mathematical Statistics,* Wiley, New York, NY.

Rohatgi, V.K. (1984), *Statistical Inference,* Wiley, New York, NY.

Ross, S.M. (1989), *Introduction to Probability Models*, 4th ed., Academic Press, San Diego, CA.

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871-880.

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* Wiley, New York, NY.

Rousseeuw, P.J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

Rousseeuw, P.J., and van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," (with discussion), *Journal of the American Statistical Association,* 85, 633-651.

Rousseeuw, P.J., and van Zomeren, B.C. (1992), "A Comparison of Some Quick Algorithms for Robust Regression," *Computational Statistics and Data Analysis,* 14, 107-116.

Rubin, D.B. (2004), "On Advice for Beginners in Statistical Research," *The American Statistician,* 58, 196-197.

Ruppert, D. (1992), "Computing S-Estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics,* 1, 253-270.

SAS Institute (1985), *SAS User's Guide: Statistics*, Version 5, SAS Institute, Cary, NC.

Schaaffhausen, H. (1878), "Die Anthropologische Sammlung Des Anatomischen Der Universitat Bonn," *Archiv fur Anthropologie,* 10, 1-65, Appendix.

Searle, S.R. (1982), *Matrix Algebra Useful for Statistics*, New York, NY.

Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis,* 2nd ed., Wiley, New York, NY.

Sen, P.K., and Singer, J.M. (1993), *Large Sample Methods in Statistics: An Introduction with Applications,* Chapman & Hall, New York, NY.

Sen, P.K., Singer, J.M., and Pedrosa De Lima, A.C. (2010), *From Finite Sample to Asymptotic Methods in Statistics,* Cambridge University Press, New York, NY.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics,* Wiley, New York, NY.

Severini, T.A. (2005), *Elements of Distribution Theory*, Cambridge University Press, New York, NY.

Silverman, B.A. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, NY.

Smith, W.B. (1997), "Publication is as Easy as C-C-C," *Communications in Statistics Theory and Methods,* 26, vii-xii.

Snedecor, G.W., and Cochran, W.G. (1967), *Statistical Methods*, 6th ed., Iowa State College Press, Ames, Iowa.

Srivastava, D.K., and Mudholkar, G.S. (2001), "Trimmed $\tilde{T}^2$: A Robust Analog of Hotelling's T^2," *Journal of Statistical Planning and Inference,* 97, 343-358.

Staudte, R.G., and Sheather, S.J. (1990), *Robust Estimation and Testing,* Wiley, New York, NY.

Stavig, V. (2004), "Bread and Peace," *Minnesota*, 35-42. See (www.minnesotaalumni.org/s/1118/content.aspx?sid=1118&gid=1&pgid=1664).

Stewart, G.M. (1969), "On the Continuity of the Generalized Inverse," *SIAM Journal on Applied Mathematics*, 17, 33-45.

Stigler, S.M (2010), "The Changing History of Robustness," *The American Statistician,* 64, 271-281.

Su, Z., and Cook, R.D. (2012), "Inner Envelopes: Efficient Estimation in Multivariate Linear Regression," *Biometrika*, 99, 687-702.

Tabachnick, B.G., and Fidell, L.S. (2006), *Using Multivariate Statistics*, 5th ed., Pearson Education, Boston, MA.

Tallis, G.M. (1963), "Elliptical and Radial Truncation in Normal Populations," *The Annals of Mathematical Statistics*, 34, 940-944.

Taskinen, S., Koch, I., and Oja, H. (2012), "Robustifying Principal Component Analysis with Spatial Sign Vectors," *Statistics & Probability Letters*, 82, 765-774.

Thode, H.C. (2002), *Testing for Normality,* Marcel Dekker, New York, NY.

Tremearne, A.J.N. (1911), "Notes on Some Nigerian Tribal Marks," *Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 41, 162-178.

Tukey, J.W. (1957), "Comparative Anatomy of Transformations," *Annals of Mathematical Statistics,* 28, 602-632.

Tukey, J.W. (1977), *Exploratory Data Analysis,* Addison-Wesley Publishing Company, Reading, MA.

Tyler, D.E. (1983), "The Asymptotic Distribution of Principal Component Roots Under Local Alternatives to Multiple Roots," *The Annals of Statistics*, 11, 1232-1242.

van der Vaart, A.W. (1998), *Asymptotic Statistics*, Cambridge University Press, Cambridge, UK.

Velilla, S. (1993), "A Note on the Multivariate Box-Cox Transformation to Normality," *Statistics and Probability Letters,* 17, 259-263.

Venables, W.N., and Ripley, B.D. (2003), *Modern Applied Statistics with S,* 4th ed., Springer-Verlag, New York, NY.

Wackerly, D.D., Mendenhall, W., and Scheaffer, R.L., (2008), *Mathematical Statistics with Applications,* 7th ed., Thomson Brooks/Cole, Belmont, CA.

Waternaux, C.M. (1976), "Asymptotic Distribution of the Sample Roots for a Nonnormal Population," *Biometrika*, 63, 639-645.

Weisberg, S. (2002), "Dimension Reduction Regression in R," *Journal of Statistical Software*, 7, webpage (www.jstatsoft.org).

Welch, B.L. (1947), "The Generalization of Student's Problem When Several Different Population Variances are Involved," *Biometrika*, 34, 28-35.

Welch, B.L. (1951), "On the Comparison of Several Mean Values: an Alternative Approach," *Biometrika*, 38, 330-336.

Welsh, A.H. (1986), "Bahadur Representations for Robust Scale Estimators Based on Regression Residuals," *The Annals of Statistics,* 14, 1246-1251.

White, H. (1984), *Asymptotic Theory for Econometricians,* Academic Press, San Diego, CA.

Wilcox, R.R. (2009), "Robust Multivariate Regression When There is Heteroscedasticity," *Communications in Statistics-Simulation and Computation*, 38, 1-13.

Wilcox, R.R. (2012), *Introduction to Robust Estimation and Hypothesis Testing,* 3rd ed., Academic Press Elsevier, New York, NY.

Wilk, M.B., Gnanadesikan, R., Huyett, M.J., and Lauh, E. (1962), "A Study of Alternative Compounding Matrices Used in a Graphical Internal Comparisons Procedure," Bell Laboratories Memorandum.

Willems, G., Pison, G., Rousseeuw, P.J., and Van Aelst, S. (2002), "A Robust Hotelling Test," *Metrika*, 55, 125-138.

Wisnowski, J.W., Simpson J.R., and Montgomery D.C. (2002), "A Performance Study for Multivariate Location and Shape Estimators," *Quality and Reliability Engineering International,* 18, 117-129.

Wisseman, S.U., Hopke, P.K., and Schindler-Kaudelka, E. (1987), "Multielemental and Multivariate Analysis of Italian Terra Sigillata in the World Heritage Museum, University of Illinois at Urbana-Champaign," *Archeomaterials*, 1, 101-107.

Woodruff, D.L., and Rocke, D.M. (1993), "Heuristic Search Algorithms for the Minimum Volume Ellipsoid," *Journal of Computational and Graphical Statistics,* 2, 69-95.

Woodruff, D.L., and Rocke, D.M. (1994), "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators," *Journal of the American Statistical Association,* 89, 888-896.

Yao, Y. (1965), "An Approximate Degrees of Freedom Solution to the Multivariate Behrens Fisher Problem," *Biometrika,* 52, 139-147.

Yeo, I.K., and Johnson, R. (2000), "A New Family of Power Transformations to Improve Normality or Symmetry," *Biometrika,* 87, 954-959.

Zhang, J. (2011), *Applications of a Robust Dispersion Estimator*, Ph.D. Thesis, Southern Illinois University, online at
(www.math.siu.edu/olive/szhang.pdf).

Zhang, Z., and Olive, D.J. (2008), "Robust Covariance Matrix Estimation," online at (www.math.siu.edu/olive/pprcovm.pdf).

Zhang, J., Olive, D.J., and Ye, P. (2012), "Robust Covariance Matrix Estimation With Canonical Correlation Analysis," *International Journal of Statistics and Probability*, 1, 119-136.

Index