David J. Olive

Robust Multivariate Analysis

Springer

Robust Multivariate Analysis

David J. Olive

Robust Multivariate Analysis

Springer

David J. Olive
Department of Mathematics
Southern Illinois University
Carbondale, IL
USA

ISBN 978-3-319-68251-8 ISBN 978-3-319-68253-2 (eBook)
https://doi.org/10.1007/978-3-319-68253-2

Library of Congress Control Number: 2017953809

Mathematics Subject Classification (2010): 62Hxx, 62G35

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

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 $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})^T$. There are n cases.

This book could be the primary text for at least two courses: a course in multivariate statistical analysis at the level of Johnson and Wichern (2007) or a course in Robust Statistics. I) For a course on multivariate statistical analysis, cover Chapters 1–13, omitting much of Chapter 4. The emphasis for this course is on multivariate statistical methods that work well for a large class of underlying distributions. Exact theory for the multivariate normal distribution is usually omitted, but is often replaced by simpler and more applicable large sample theory. II) For a course on Robust Statistics, cover Chapters 1–14, where Chapters 4 and 14 are the most important chapters. This course emphasizes methods that work well for a large class of distributions as well as multivariate statistical methods that are robust to certain types of *outliers*: observations that lie far from the bulk of the data. Outliers can ruin a classical analysis. The text tends to cover the classical method that is not robust to outliers, and then gives a practical outlier robust analog of the classical method that has some large sample theory, and often the robust method can be used in tandem with the classical method. This course on Robust Statistics covers the univariate location model very briefly compared to texts like Huber and Ronchetti (2009) and Wilcox (2017).

I have taught topic courses on I) Robust Statistics using Olive (2008) where I cover a lot of material on the univariate location model, robust multivariate location and dispersion (Chapter 4), and robust regression (Chapter 14), and II) Robust Multivariate Analysis using an earlier version of this text where I cover Chapters 1–13 but omit Chapter 14.

There are many texts on multivariate statistical analysis that are based on rather difficult multivariate normal theory. This text uses simpler and more applicable large sample theory for classical methods of multivariate statistical analysis and provides good practical outlier resistant methods that are backed by theory.

Prediction regions are developed for the multivariate location and dispersion model as well as the multivariate linear regression model. A relationship between the new prediction regions and confidence regions provides a simple way to bootstrap confidence regions. These confidence regions often provide a practical method for testing hypotheses. See Chapter 5.

This book covers robust multivariate analysis. There are two uses of the word “robust.” First, a method is robust to the assumption of multivariate normality if the method gives good results for a large class of underlying distributions. Such methods have good large sample theory. Some of the methods in this text work well, asymptotically, if the data are independent and identically distributed from a population that has a nonsingular covariance matrix. Other methods have large sample theory for a large class of elliptically contoured distributions. Second, the text develops methods that are robust to certain types of outliers.

This book presents classical methods that are robust to the assumption of multivariate normality, and often uses an outlier robust estimator of multivariate location and dispersion to develop an outlier robust method that can be used in tandem with the classical method. The new technique for bootstrapping confidence regions can often be used to perform inference for the outlier robust method. These techniques are illustrated for methods such as principal component analysis, canonical correlation analysis, and factor analysis. More importantly, the technique for making a good robust version of a classical method can be extended to many classical methods. Prediction regions are developed that have good large sample theory, recent large sample theory for multivariate linear regression is presented, and plots for detecting outliers and for checking the model are presented.

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 semiparametric in that the procedures also perform well on a much larger class of elliptically contoured (EC) distributions. This book uses many **acronyms**. See Table 1.1.

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 outlier 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 independent and identically distributed (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.

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. I) 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. II) Practical asymptotically optimal prediction regions are developed. One of the prediction regions can be applied to a bootstrap sample to make a confidence region.

Chapter 1 provides an introduction and some results that will be used later in the text. Some univariate location model results are also given. The material on truncated distributions will be useful for simplifying the large sample theory of robust regression estimators in Chapter 14. 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, and shows how to bootstrap hypothesis tests by making a confidence region using a prediction region applied to the bootstrap sample of the test statistic. 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, including robust regression, 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, statistical learning, and pattern recognition. See Duda et al. (2000), James et al. (2013), and Bishop (2006). The text can also be used to present many statistical methods to students running a statistical consulting laboratory.

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 provides 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 are competitors for parametric prediction regions, which tend to be far too small when the parametric distribution is misspecified, and competitors for bootstrap intervals, especially if the bootstrap intervals take too long to compute. These prediction regions are extended to multivariate regression in Section 12.3.

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 or RMVN estimators from Section 4.4. The resistant trimmed views methods for visualizing 1D regression models graphically are discussed in Section 14.6.

5) Applying a prediction region to a bootstrap sample results in a confidence region that can be used for hypothesis tests based on classical or robust estimators. For example, the bootstrap prediction region method may be useful for testing statistical hypotheses after variable selection. See Section 5.3.

Much of the research on robust multivariate analysis in this book is being published for the first time and will not appear in a journal. Some of the research is also quite recent, and further research and development is likely. See, for example, Olive (2017a, b) and Rupasinghe Arachchige Don and Pelawa Watagoda (2017).

The website (http://lagrange.math.siu.edu/Olive/multbk.htm) for this book provides over 130 *R* programs in the file *mpack.txt* and several *R* data sets in the file *mrobdata.txt*. 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 functions *mpack.txt* and data files *mrobdata.txt* into *R*: The commands

```
source("http://lagrange.math.siu.edu/Olive/mpack.txt")
source("http://lagrange.math.siu.edu/Olive/mrobdata.txt")
```

can be used to download the *R* functions and data sets into *R*. (*Copy and paste these two commands* into *R* from near the top of the file (http://lagrange.math.siu.edu/Olive/mrsashw.txt), which contains commands that are useful for doing many of the *R* homework problems.) Type *ls()*. Over 130 *R* 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 and data are easily obtained with the source commands).

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.

There are **two target audiences** for a **Master's level course in a Statistics department** if students have had only one calculus-based course in statistics (e.g., Wackerly et al. 2008). The text can be used for a course in I) Robust Statistics or for II) a course in multivariate analysis at a level similar to that of Johnson and Wichern (2007), Mardia et al. (1979), Press (2005), and Rencher and Christensen (2012). Anderson (2003) is at a much higher level.

The text is higher than Master's level for students in an applied field like quantitative psychology. Lower level texts on multivariate analysis include Flury and Riedwyl (1988), Grimm and Yarnold (1995, 2000), Hair et al. (2009), Kachigan (1991), Lattin et al. (2003), and Tabachnick and Fidell (2012).

For the two Master's level courses, consider skipping the proofs of the theorems. Chapter 2, Sections 3.1–3.3, and Chapter 5 are important. Then topics from the remaining chapters can be chosen. For a course in Robust Statistics, Chapter 4 and robust regression from Chapter 14 are important. 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 (2014), Poor (1988), and White (1984) meet this requirement.

A *third target audience* consists of those who want to do research in robust statistics or multivariate analysis. The text could be used as a reference or the primary text in a reading course for Ph.D. students.

For robust multivariate analysis, see Atkinson et al. (2004), Farcomeni and Greco (2015), Oja (2010), Shevlyakov and Oja (2016), and Wilcox (2017). Also see Aggarwal (2017). Most work on robust multivariate analysis follows the dominant robust statistics paradigm, described after the next paragraph. See Maronna et al. (2006).

Need for the book:

As a book on robust multivariate analysis, this book is an alternative to the dominant robust statistics 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 non-normality and work well on large classes of distributions. A new bootstrap method is used for hypothesis tests based on classical and robust estimators.

The *dominant robust statistics 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, call 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 dominant robust statistics 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 et al. (2006, ch. 2, 6) and Hubert et al. (2008) provided 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 both consistent and 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 an 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 $(\bar{\boldsymbol{x}}, \boldsymbol{S})$ where $\bar{\boldsymbol{x}}$ 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 resulting FMCD estimator is the classical estimator, which is $\sqrt{n}$ consistent on a large class of distributions. Now suppose there are two trial fits $(\bar{\boldsymbol{x}}, \boldsymbol{S})$ 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 $det(\boldsymbol{I}_p) = 1$, the fit with the smallest determinant will not be the classical estimator if $det(\boldsymbol{S}) > 1$. 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, e.g., $K = 500$. To produce an elemental fit, randomly select h cases and compute the classical estimator $(T_i, \boldsymbol{C}_i)$ (or $T_i = \hat{\beta}_i$ 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 Chapter 4, Section 14.4, and 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 cannot 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 $K = K_n \to \infty$, 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. □

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, FLMS, FMCD, or FMVE. See Hawkins and Olive (2002, p. 157). You can get consistent estimators if $K = K_n \to \infty$ or $h = h_n \to \infty$ as $n \to \infty$. You can get high breakdown estimators and avoid singular starts if all $K = K_n = C(n, h)$ elemental sets are used, but such an estimator is impractical.

Researchers are starting to use intelligently chosen trial fits. Maronna and Yohai (2015) used 500 elemental sets plus the classical estimator to produce an FS estimator used as the initial estimator for an FMM estimator. However, choosing from a fixed number of elemental sets and the classical estimator results in a zero breakdown initial FS estimator, and the FMM estimator has the same breakdown as the initial estimator. Hence the FMM estimator is zero breakdown. For regression, they show that the FS estimator is consistent on a class of symmetric error distributions, so the FMM estimator is asymptotically equivalent to the MM estimator that has the smallest criterion value. This FMM estimator is asymptotically equivalent to the FMM estimator that uses OLS as the initial estimator. See Section 14.7.1 for more on this regression estimator. For multivariate location and dispersion, suppose the algorithm uses elemental sets and the sample covariance matrix: These trial fits are unbiased estimators of the population covariance estimator $Cov(\boldsymbol{x}) = c_x \boldsymbol{\Sigma}$ for elliptically contoured distributions. But for S estimators, the global minimizer is estimating $d_x \boldsymbol{\Sigma}$ asymptotically, where the constant $c_x \neq d_x$. Hence the probability that the initial estimator is an elemental set is likely bounded away from 0, and the zero breakdown FMM estimator is likely inconsistent.

Acknowledgements

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 (including R Core Team (2016)). A 1997 preprint of Rousseeuw and Van Driessen (1999) was the starting point for much of my work in multivariate analysis. Working with students was also valuable, and an earlier version of this text was used in a robust multivariate analysis course in 2012.

Thanks also go to Springer, Springer's Associate Editor Donna Chernyk, and to several reviewers.

Carbondale, USA David J. Olive

Contents

Chapter 1
Introduction

This chapter gives a brief introduction to multivariate analysis, including some matrix optimization results, mixture distributions, and the special case of the location model. Section 1.2 gives an overview of the book along with a table of abbreviations. Truncated distributions, covered in Section 1.7, will be useful for large sample theory for the location model and for the regression model. See Chapter 14.

1.1 Introduction

Multivariate analysis is a set of statistical techniques used to analyze possibly 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 $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})^T = (x_{i1}, ..., x_{ip})^T$.

Definition 1.1. A **case** or **observation** consists of p random variables measured for one person or thing. The ith case $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$.

Notation: Typically lowercase boldface letters such as $\boldsymbol{x}$ denote column vectors, while uppercase boldface letters such as $\boldsymbol{S}$ denote matrices with two 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 the 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 uppercase letter such as Y will usually be a random variable. A lowercase letter such as x_1 will also often be a random variable. An exception to this notation is

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_1

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.

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 independent and identically distributed (iid): 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.

Some techniques have a vector of response variables $(Y_1, ..., Y_m)^T$ that is predicted with a vector of predictor variables $(x_1, ..., x_p)^T$. See Chapters 10 and 12. 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. Robust regression is briefly covered in Chapter 14.

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. See Section 1.2 for a table of acronyms.

Robust multivariate analysis consists of i) techniques that are robust to non-normality or ii) techniques that are robust to outliers. Techniques that are robust to outliers tend to have some robustness to non-normality. The classical sample mean $\overline{\boldsymbol{x}}$ and covariance matrix $\boldsymbol{S}$, defined in Section 2.2, are very robust to non-normality, but are not robust to outliers. Large sample theory is useful for both robust techniques. See Section 3.4.

Statistical Learning could be defined as the statistical analysis of multivariate data. Machine learning, data mining, Big Data, analytics, business analytics, data analytics, and predictive analytics are synonymous terms. The techniques are useful for Data Science and Statistics, the science of extracting information from data. Often Statistical Learning methods are useful when n is large, but n/p is not large. Most of the methods in this text are for n/p large, but Section 4.7 shows how to detect outliers when n/p is not large. The outlier detection method gives a *covmb2 set B* of at least $n/2$ cases. If x is a matrix of predictor variables and Y is a vector of response variables, the following R commands produce cleaned data that can be used in Statistical Learning techniques, such as lasso, even if $p > n$. Section 8.9 uses a similar technique for discriminant analysis.

```
tem <- getB(x)
Yc <- Y[tem$indx]
xc <- x[tem$indx,]
```

Statistical Learning problems are supervised or unsupervised. For supervised learning, the goal is to predict a response variable given predictors. Discriminant analysis and regression are important examples. See Chapters 8 and 14. For unsupervised learning, the goal is to describe associations and patterns among the p variables. Clustering, described in Chapter 13, is an important example. Excellent texts for Statistical Learning are Efron and Hastie (2016), Hair et al. (2009), Hastie et al. (2015), and James et al. (2013). Also see Olive (2017c).

1.2 Overview and Acronyms

Chapters 1, 2, and 3 present some results useful for multivariate analysis, including matrix optimization results, the sample mean and covariance matrix, Mahalanobis distances, the multivariate normal distribution, and elliptically contoured distributions. This material is essential for any first course in multivariate analysis. Some of the sections of Chapter 1 are useful for robust regression which is covered in Chapter 14.

Chapters 4 and 5 are the most important chapters in the book and are needed for the following chapters except Chapter 13 on clustering. Chapter 4 discusses classical and outlier resistant methods of multivariate location and dispersion (MLD). Chapter 5 shows how to use the DD plot to detect outliers and gives a prediction region for multivariate data. Applying this prediction region on bootstrapped data gives a confidence region that can be used for hypothesis testing. This "prediction region method" will be used to perform inference on outlier resistant methods of multivariate analysis. There is a subset U that is used to compute the robust MLD estimator. Often applying a standard method, such as principal components, on the subset U results in a robust version of the standard method.

Most of the remaining chapters focus on standard methods of multivariate analysis such as principal component analysis, canonical correlation analysis,

Table 1.1 Acronyms

Acronym	Description
CCA	canonical correlation analysis
cdf	cumulative distribution function
cf	characteristic function
CI	confidence interval
CLT	central limit theorem
DA	discriminant analysis
Det-MCD	practical approximate MCD estimator not backed by theory
DGK	an MLD estimator (DGK are the initials of the paper's authors)
EC	elliptically contoured
ESP	estimated sufficient predictor
ESSP	estimated sufficient summary plot = response plot
Fast-MCD	a slow FMCD estimator
FCH	name of a fast, consistent, highly outlier resistant MLD estimator
FDA	Fisher's discriminant analysis
FLTS	practical approximate LTS estimator not backed by theory
FMCD	practical approximate MCD estimator not backed by theory
GAM	generalized additive model
GLM	generalized linear model
HB	high breakdown
hbreg	practical high breakdown regression estimator backed by theory
iid	independent and identically distributed
KNN	K-nearest neighbors discriminant analysis
LDA	linear discriminant analysis
LMS	least median of squares (robust regression)
LR	logistic regression
LTA	least trimmed sum of absolute deviations (robust regression)
LTS	least trimmed sum of squares (robust regression)
MAD	median absolute deviation
MANOVA	multivariate analysis of variance
MB	median ball estimator
MBA	an MLD estimator made obsolete by FCH
MBA	or the median ball algorithm is the mbareg estimator
mbareg	a resistant regression estimator backed by theory
MCD	the impractical minimum covariance determinant estimator
MCLT	multivariate central limit theorem
MED	the median
mgf	moment generating function
MLD	multivariate location and dispersion
MLR	multiple linear regression
MVE	the impractical minimum volume ellipsoid estimator
MVN	multivariate normal

(continued)

Table 1.1 (continued)

Acronym	Description
OGK	an MLD estimator not backed by theory
OLS	ordinary least squares
PCA	principal component analysis
pdf	probability density function
PI	prediction interval
pmf	probability mass function
QDA	quadratic discriminant analysis
RFCH	the reweighted FCH estimator
RMVN	a reweighted FCH estimator that works well for MVN data
SE	standard error
SSP	sufficient summary plot
SUR	seemingly unrelated regressions
TVREG	a resistant "trimmed views" regression estimator

discriminant analysis, MANOVA, factor analysis, and multivariate linear regression. Emphasis is on methods that are robust to normality: the methods have large sample theory that shows that the methods work on a large class of distributions. Of secondary importance is how to make outlier resistant methods that are backed by large sample theory. Chapter 14 considers other techniques, including robust regression.

Acronyms are widely used in robust statistics and multivariate analysis, and some of the more important acronyms are in Table 1.1 Also see the text's index. The letter "R" tends to stand for "robust" (RPCA) or "reweighted" (RFCH). The letter "F" before a brand-name robust estimator (FMCD) tends to mean a practical estimator that used a fixed number of trial fits, where the criterion of the brand-name estimator was used to select the trial fit used in the final estimator. The letter "C" before a brand-name estimator (CLTS) tends to mean a concentration algorithm that was used for the F-brand-name estimator. The letter "A," standing for "algorithm," was also used for concentration algorithms (ALTS). These acronyms (with A, C, F, or R) are often omitted from Table 1.1.

1.3 Some 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 test 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 test 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 laboratory for analysis. It is often the case that even if the shards are sent to the same laboratory twice, the two sets of laboratory 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 test 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) noted 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 test 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 test 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 techniques. For supervised classification, Hand (2006) noted 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) suggested 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.

1.4 Some Matrix Optimization Results

The following results will be used throughout the text and are useful for principal component analysis, canonical correlation analysis, Fisher's discriminant analysis, and the Hotelling's T^2 test. Let $\boldsymbol{B} > 0$ denote that $\boldsymbol{B}$ is a positive definite matrix. The *generalized eigenvalue problem* finds eigenvalue eigenvector pairs $(\lambda, \boldsymbol{g})$ such that $\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{g} = \lambda\boldsymbol{g}$ which are also solutions to the equation $\boldsymbol{A}\boldsymbol{g} = \lambda\boldsymbol{C}\boldsymbol{g}$. Then the pairs are used to maximize or minimize the *Rayleigh quotient* $\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}$. Results from linear algebra show that if $\boldsymbol{C} > 0$ and $\boldsymbol{A}$ are both symmetric, then the p eigenvalues of $\boldsymbol{C}^{-1}\boldsymbol{A}$ are real, and the number of nonzero eigenvalues of $\boldsymbol{C}^{-1}\boldsymbol{A}$ is equal to $\text{rank}(\boldsymbol{C}^{-1}\boldsymbol{A}) = \text{rank}(\boldsymbol{A})$. Note that if $\boldsymbol{a}_1 = c_1\boldsymbol{g}_1$ is the maximizer and $\boldsymbol{a}_p = c_p\boldsymbol{g}_p$ is the minimizer of the Rayleigh quotient for any nonzero constants c_1 and c_p, then there is a vector $\boldsymbol{\beta}$ that is the maximizer or minimizer such that $\|\boldsymbol{\beta}\| = 1$.

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, pp. 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 $\max_{\boldsymbol{a}\neq\boldsymbol{0}} \frac{n\boldsymbol{a}^T(\overline{\boldsymbol{x}}-\boldsymbol{\mu})(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{S}\boldsymbol{a}} = n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}-\boldsymbol{\mu}) = T^2$ where the max is attained for $\boldsymbol{a} = c\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}-\boldsymbol{\mu})$ for any constant $c \neq 0$.

f) Let $\boldsymbol{A}$ be a $p \times p$ symmetric matrix. Let $\boldsymbol{C} > 0$ be a $p \times p$ symmetric matrix. Then $\max_{\boldsymbol{a}\neq\boldsymbol{0}} \frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}} = \lambda_1(\boldsymbol{C}^{-1}\boldsymbol{A})$, the largest eigenvalue of $\boldsymbol{C}^{-1}\boldsymbol{A}$. The value of $\boldsymbol{a}$ that achieves the max is the eigenvector $\boldsymbol{g}_1$ of $\boldsymbol{C}^{-1}\boldsymbol{A}$ corresponding to $\lambda_1(\boldsymbol{C}^{-1}\boldsymbol{A})$. Similarly, $\min_{\boldsymbol{a}\neq\boldsymbol{0}} \frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}} = \lambda_p(\boldsymbol{C}^{-1}\boldsymbol{A})$, the smallest eigenvalue of $\boldsymbol{C}^{-1}\boldsymbol{A}$. The value of $\boldsymbol{a}$ that achieves the min is the eigenvector $\boldsymbol{g}_p$ of $\boldsymbol{C}^{-1}\boldsymbol{A}$ corresponding to $\lambda_p(\boldsymbol{C}^{-1}\boldsymbol{A})$.

Proof Sketch. For a), note that $\text{rank}(\boldsymbol{C}^{-1}\boldsymbol{A}) = 1$, where $\boldsymbol{C} = \boldsymbol{B}$ and $\boldsymbol{A} = \boldsymbol{d}\boldsymbol{d}^T$, since $\text{rank}(\boldsymbol{C}^{-1}\boldsymbol{A}) = \text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{d}) = 1$. Hence $\boldsymbol{C}^{-1}\boldsymbol{A}$ has one nonzero eigenvalue eigenvector pair $(\lambda_1, \boldsymbol{g}_1)$. Since

$$(\lambda_1 = \boldsymbol{d}^T\boldsymbol{B}^{-1}\boldsymbol{d}, \boldsymbol{g}_1 = \boldsymbol{B}^{-1}\boldsymbol{d})$$

is a nonzero eigenvalue eigenvector pair for $\boldsymbol{C}^{-1}\boldsymbol{A}$ and $\lambda_1 > 0$, the result follows by f).

Note that b) and c) are special cases of f) with $\boldsymbol{A} = \boldsymbol{B}$ and $\boldsymbol{C} = \boldsymbol{I}$.

Note that e) is a special case of a) with $\boldsymbol{d} = (\overline{\boldsymbol{x}} - \boldsymbol{\mu})$ and $\boldsymbol{B} = \boldsymbol{S}$.

(Also note that $(\lambda_1 = (\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}), \boldsymbol{g}_1 = \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}))$ is a nonzero eigenvalue eigenvector pair for the rank 1 matrix $\boldsymbol{C}^{-1}\boldsymbol{A}$ where $\boldsymbol{C} = \boldsymbol{S}$ and $\boldsymbol{A} = (\overline{\boldsymbol{x}} - \boldsymbol{\mu})(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T$.)

For f), see Mardia et al. (1979, p. 480). □

Suppose $\boldsymbol{A} > 0$ and $\boldsymbol{C} > 0$ are $p \times p$ symmetric matrices, and let $\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{a} = \lambda \boldsymbol{a}$. Then $\boldsymbol{A}\boldsymbol{a} = \lambda \boldsymbol{C}\boldsymbol{a}$, or $\boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{a} = \frac{1}{\lambda}\boldsymbol{a}$. Hence if $(\lambda_i(\boldsymbol{C}^{-1}\boldsymbol{A}), \boldsymbol{a})$ are eigenvalue eigenvector pairs of $\boldsymbol{C}^{-1}\boldsymbol{A}$, then $\left(\lambda_i(\boldsymbol{A}^{-1}\boldsymbol{C}) = \frac{1}{\lambda_i(\boldsymbol{C}^{-1}\boldsymbol{A})}, \boldsymbol{a}\right)$ are eigenvalue eigenvector pairs of $\boldsymbol{A}^{-1}\boldsymbol{C}$. Thus we can maximize $\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}$ with the eigenvector $\boldsymbol{a}$ corresponding to the smallest eigenvalue of $\boldsymbol{A}^{-1}\boldsymbol{C}$ and minimize $\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}$ with the eigenvector $\boldsymbol{a}$ corresponding to the largest eigenvalue of $\boldsymbol{A}^{-1}\boldsymbol{C}$.

Remark 1.1. Suppose $\boldsymbol{A}$ and $\boldsymbol{C}$ are symmetric $p \times p$ matrices, $\boldsymbol{A} > 0$, $\boldsymbol{C}$ is singular, and it is desired to make $\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}$ large but finite. Hence $\frac{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}$ should be made small but nonzero. The above result suggests that the eigenvector $\boldsymbol{a}$ corresponding to the smallest nonzero eigenvalue of $\boldsymbol{A}^{-1}\boldsymbol{C}$ may be useful. Similarly, suppose it is desired to make $\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}$ small but nonzero. Hence $\frac{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}$ should be made large but finite. Then the eigenvector $\boldsymbol{a}$ corresponding to the largest eigenvalue of $\boldsymbol{A}^{-1}\boldsymbol{C}$ may be useful.

1.5 The Location Model

The location model is used when there is one variable Y, such as height, of interest. The location model is a special case of the multivariate location and dispersion model, where there are p variables $x_1, ..., x_p$ of interest, such as height and weight if $p = 2$. See Chapter 2.

The *location model* is

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

where $e_1, ..., e_n$ are error random variables, often independent and identically distributed (iid) with zero mean. For example, if the Y_i are iid from a normal distribution with mean μ and variance σ^2, written $Y_i \sim N(\mu, \sigma^2)$, then the e_i are iid with $e_i \sim N(0, \sigma^2)$. 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 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; i.e., assume that $Y_1, ..., Y_n$ are iid.

Definition 1.2. 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 $Y_i \le m$, 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 $Y_i \le m$ 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 $Y_i \le \overline{Y}$ is equal to the weight of the rods corresponding to $Y_i > \overline{Y}$. The sample mean is drawn toward an outlier since the absolute residual corresponding to a single outlier is large.

If the data set $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \le \cdots \le Y_{(n)}$, 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 MED$_c(n) = 3$ where the sample size $n = 5$.

Definition 1.3. 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 1.4. 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 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 determines 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 1.5. 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.

The following confidence interval provides considerable resistance to gross outliers while being very simple to compute. See Olive (2008, pp. 238, 261–262). The standard error SE(MED(n)) is due to Bloch and Gastwirth (1968), but the degrees of freedom p is motivated by the confidence interval for the trimmed mean. Let $\lfloor x \rfloor$ denote the "greatest integer function" (e.g., $\lfloor 7.7 \rfloor = 7$). Let $\lceil x \rceil$ denote the smallest integer greater than or equal to x (e.g., $\lceil 7.7 \rceil = 8$).

Application 1.1: inference with the sample median. Let $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4}\ \rceil$ and use

$$SE(\text{MED}(n)) = 0.5(Y_{(U_n)} - Y_{(L_n+1)}).$$

Let $p = U_n - L_n - 1$ (so $p \approx \lceil\ \sqrt{n}\ \ \rceil$). Then a $100(1-\alpha)\%$ confidence interval for the population median is

$$\text{MED}(n) \pm t_{p,1-\alpha/2} SE(\text{MED}(n)). \tag{1.6}$$

Example 1.2. Let the data be 6, 9, 9, 7, 8, 9, 9, 7. Assume the data came from a symmetric distribution with mean μ, and find a 95% CI for μ.

Solution. When computing small examples by hand, the steps are to sort the data from smallest to largest value and find n, L_n, U_n, $Y_{(L_n+1)}$, $Y_{(U_n)}$, p, MED(n), and $SE(\text{MED}(n))$. After finding $t_{p,1-\alpha/2}$, plug the relevant quantities into the formula for the CI. The sorted data are 6, 7, 7, 8, 9, 9, 9, 9. Thus $\text{MED}(n) = (8+9)/2 = 8.5$. Since $n = 8$, $L_n = \lfloor 4 \rfloor - \lceil \sqrt{2} \rceil = 4 - \lceil 1.414 \rceil = 4 - 2 = 2$ and $U_n = n - L_n = 8 - 2 = 6$. Hence $SE(\text{MED}(n)) = 0.5(Y_{(6)} - Y_{(3)}) = 0.5 * (9-7) = 1$. The degrees of freedom $p = U_n - L_n - 1 = 6 - 2 - 1 = 3$. The cutoff $t_{3,0.975} = 3.182$. Thus the 95% CI for MED(Y) is

$$\text{MED}(n) \pm t_{3,0.975} SE(\text{MED}(n))$$

$= 8.5 \pm 3.182(1) = [5.318, 11.682]$. The classical t-interval uses $\overline{Y} = (6+7+7+8+9+9+9+9)/8$ and $S_n^2 = (1/7)[(\sum_{i=1}^n Y_i^2) - 8(8^2)] = (1/7)[(522 - 8(64)] = 10/7 \approx 1.4286$, and $t_{7,0.975} \approx 2.365$. Hence the 95% CI for μ is $8 \pm 2.365(\sqrt{1.4286/8}) = [7.001, 8.999]$. Notice that the t-cutoff = 2.365 for the classical interval is less than the t-cutoff = 3.182 for the median interval and that $SE(\overline{Y}) < SE(\text{MED}(n))$. The parameter μ is between 1 and 9 since the test scores are integers between 1 and 9. Hence for this example, the t-interval is considerably superior to the overly long median interval.

Example 1.3. In the last example, what happens if the 6 becomes 66 and a 9 becomes 99?

Solution. Then the ordered data are 7, 7, 8, 9, 9, 9, 66, 99. Hence $\text{MED}(n) = 9$. Since L_n and U_n only depend on the sample size, they take the same values as in the previous example and $SE(\text{MED}(n)) = 0.5(Y_{(6)} - Y_{(3)}) = 0.5 * (9 - 8) = 0.5$. Hence the 95% CI for MED(Y) is $\text{MED}(n) \pm t_{3,0.975} SE(\text{MED}(n)) = 9 \pm 3.182(0.5) = [7.409, 10.591]$. Notice that with discrete data, it is possible to drive $SE(\text{MED}(n))$ to 0 with a few outliers if n is small. The classical confidence interval $\overline{Y} \pm t_{7,0.975} S/\sqrt{n}$ blows up and is equal to $[-2.955, 56.455]$.

Example 1.4. The Buxton (1920) data contains 87 heights of men, but five of the men were recorded to be about 0.75 inches tall! The mean height is $\overline{Y} = 1598.862$, and the classical 95% CI is [1514.206, 1683.518]. MED$(n) = 1693.0$, and the resistant 95% CI based on the median is [1678.517, 1707.483].

The heights for the five men were recorded under their head lengths, so the outliers can be corrected. Then $\overline{Y} = 1692.356$, and the classical 95% CI is [1678.595, 1706.118]. Now MED$(n) = 1694.0$, and the 95% CI based on the median is [1678.403, 1709.597]. Notice that when the outliers are corrected, the two intervals are very similar although the classical interval length is slightly shorter. Also notice that the outliers roughly shifted the median confidence interval by about 1 mm, while the outliers greatly increased the length of the classical t-interval. See Problem 1.3 for *mpack* software.

1.6 Mixture Distributions

Mixture distributions are often used as outlier models, and certain mixtures of elliptically contoured distributions have an elliptically contoured distribution. See Problem 3.4. 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 1.7 is the same as the usual definition for expectation if Y is a discrete or continuous random variable. The two definitions and proposition can be extended to random vectors.

Definition 1.6. The distribution of a random variable Y is a *mixture distribution* if the cdf of Y has the form

$$F_Y(y) = \sum_{i=1}^{k} \alpha_i F_{W_i}(y) \tag{1.7}$$

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

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y) dF(y). \tag{1.8}$$

Proposition 1.2. a) If Y is a discrete random variable that has a probability mass function (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 probability distribution function (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 1.5. 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 VAR$(Y) = E[Y^2] - (E[Y])^2 = 1 - \epsilon + \epsilon k^2$. If $\epsilon = 0.1$ and $k = 10$, then $EY = 0$, and 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 1.2. 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.

1.7 Truncated Distributions

Truncated and Winsorized random variables are important because they simplify the asymptotic theory of robust estimators. See Section 14.7. Let Y be a random variable with continuous cdf F, and let $\alpha = F(a) < F(b) = \beta$. Then α is the *left trimming proportion* and $1-\beta$ is the *right trimming proportion.* Let $F(a-) = P(Y < a)$. (Refer to Proposition 1.2 for the notation used below.)

Definition 1.8. The *truncated random variable* $Y_T \equiv Y_T(a,b)$ with *truncation points* a and b has cdf

$$F_{Y_T}(y|a,b) = G(y) = \frac{F(y) - F(a-)}{F(b) - F(a-)} \tag{1.9}$$

for $a \le y \le b$. Also G is 0 for $y < a$, and G is 1 for $y > b$. The mean and variance of Y_T are

$$\mu_T = \mu_T(a,b) = \int_{-\infty}^{\infty} y dG(y) = \frac{\int_a^b y dF(y)}{\beta - \alpha} \tag{1.10}$$

and

$$\sigma_T^2 = \sigma_T^2(a,b) = \int_{-\infty}^{\infty} (y-\mu_T)^2 dG(y) = \frac{\int_a^b y^2 dF(y)}{\beta-\alpha} - \mu_T^2.$$

See Cramér (1946, p. 247).

Definition 1.9. The *Winsorized random variable*

$$Y_W = Y_W(a,b) = \begin{cases} a, & Y \le a \\ Y, & a \le Y \le b \\ b, & Y \ge b. \end{cases}$$

If the cdf of $Y_W(a,b) = Y_W$ is F_W, then

$$F_W(y) = \begin{cases} 0, & y < a \\ F(a), & y = a \\ F(y), & a < y < b \\ 1, & y \ge b. \end{cases}$$

Since Y_W is a mixture distribution with a point mass at a and at b, the mean and variance of Y_W are

$$\mu_W = \mu_W(a,b) = \alpha a + (1-\beta)b + \int_a^b y dF(y)$$

and

$$\sigma^2_W = \sigma^2_W(a,b) = \alpha a^2 + (1-\beta)b^2 + \int_a^b y^2 dF(y) - \mu^2_W.$$

Truncated distributions can be used to simplify the asymptotic theory of robust estimators of location and regression. The following four subsections will be useful when the underlying distribution is exponential, double exponential, normal, or Cauchy. If Y has an exponential distribution, $Y \sim$ EXP(λ), then the pdf of Y is

$$f(y) = \frac{1}{\lambda} \exp\left(\frac{-y}{\lambda}\right) I(y \ge 0)$$

where $\lambda > 0$ and the indicator $I(y \ge 0)$ is one if $y \ge 0$ and zero otherwise. If Y has a double exponential distribution (or Laplace distribution), $Y \sim$ DE(θ, λ), then the pdf of Y is

$$f(y) = \frac{1}{2\lambda} \exp\left(\frac{-|y-\theta|}{\lambda}\right)$$

where y is real and $\lambda > 0$. If Y has a normal distribution (or Gaussian distribution), $Y \sim N(\mu, \sigma^2)$, then the pdf of Y is

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and μ and y are real. If Y has a Cauchy distribution, $Y \sim C(\mu, \sigma)$, then the pdf of Y is

$$f(y) = \frac{\sigma}{\pi} \frac{1}{\sigma^2 + (y-\mu)^2} = \frac{1}{\pi\sigma[1 + (\frac{y-\mu}{\sigma})^2]}$$

where y and μ are real numbers and $\sigma > 0$.

Definitions 1.8 and 1.9 defined the truncated random variable $Y_T(a,b)$ and the Winsorized random variable $Y_W(a,b)$. Let Y have cdf F, and let the truncated random variable $Y_T(a,b)$ have the cdf $F_{T(a,b)}$. The following lemma illustrates the relationship between the means and variances of $Y_T(a,b)$ and $Y_W(a,b)$. Note that $Y_W(a,b)$ is a mixture of $Y_T(a,b)$ and two point masses at a and b. Let $c = \mu_T(a,b) - a$ and $d = b - \mu_T(a,b)$.

Lemma 1.3. Let $a = \mu_T(a,b) - c$ and $b = \mu_T(a,b) + d$. Then
a) $\mu_W(a,b) = \mu_T(a,b) - \alpha c + (1-\beta)d,$ and
b) $\sigma^2_W(a,b) = (\beta - \alpha)\sigma^2_T(a,b) + (\alpha - \alpha^2)c^2 + [(1-\beta) - (1-\beta)^2]d^2 + 2\alpha(1-\beta)cd.$
c) If $\alpha = 1 - \beta$ then

$$\sigma^2_W(a,b) = (1-2\alpha)\sigma^2_T(a,b) + (\alpha - \alpha^2)(c^2 + d^2) + 2\alpha^2 cd.$$

d) If $c = d$ then

$$\sigma^2_W(a,b) = (\beta - \alpha)\sigma^2_T(a,b) + [\alpha - \alpha^2 + 1 - \beta - (1-\beta)^2 + 2\alpha(1-\beta)]d^2.$$

e) If $\alpha = 1 - \beta$ and $c = d$, then $\mu_W(a,b) = \mu_T(a,b)$ and

$$\sigma^2_W(a,b) = (1-2\alpha)\sigma^2_T(a,b) + 2\alpha d^2.$$

Proof. We will prove b) since its proof contains the most algebra. Now

$$\sigma^2_W = \alpha(\mu_T - c)^2 + (\beta - \alpha)(\sigma^2_T + \mu^2_T) + (1-\beta)(\mu_T + d)^2 - \mu^2_W.$$

Collecting terms shows that

$$\sigma^2_W = (\beta - \alpha)\sigma^2_T + (\beta - \alpha + \alpha + 1 - \beta)\mu^2_T + 2[(1-\beta)d - \alpha c]\mu_T$$

$$+\alpha c^2 + (1-\beta)d^2 - \mu^2_W.$$

From a),

$$\mu^2_W = \mu^2_T + 2[(1-\beta)d - \alpha c]\mu_T + \alpha^2 c^2 + (1-\beta)^2 d^2 - 2\alpha(1-\beta)cd,$$

and we find that

$$\sigma^2_W = (\beta - \alpha)\sigma^2_T + (\alpha - \alpha^2)c^2 + [(1-\beta) - (1-\beta)^2]d^2 + 2\alpha(1-\beta)cd. \quad \square$$

1.7.1 The Truncated Exponential Distribution

Let Y be a (one sided) truncated exponential $TEXP(\lambda, b)$ random variable. Then the pdf of Y is

$$f_Y(y|\lambda, b) = \frac{\frac{1}{\lambda}e^{-y/\lambda}}{1 - \exp(-\frac{b}{\lambda})}$$

for $0 < y \le b$ where $\lambda > 0$. Let $b = k\lambda$, and let

$$c_k = \int_0^{k\lambda} \frac{1}{\lambda} e^{-y/\lambda} dy = 1 - e^{-k}.$$

Next, we will find the first two moments of $Y \sim TEXP(\lambda, b = k\lambda)$ for $k > 0$.

Lemma 1.4. If Y is $TEXP(\lambda, b = k\lambda)$ for $k > 0$, then

$$a) \ E(Y) = \lambda \left[\frac{1 - (k+1)e^{-k}}{1 - e^{-k}}\right],$$

and

$$b) \ E(Y^2) = 2\lambda^2 \left[\frac{1 - \frac{1}{2}(k^2 + 2k + 2)e^{-k}}{1 - e^{-k}}\right].$$

Proof. a) Note that

$$c_k E(Y) = \int_0^{k\lambda} \frac{y}{\lambda} e^{-y/\lambda} dy = -ye^{-y/\lambda}|_0^{k\lambda} + \int_0^{k\lambda} e^{-y/\lambda} dy$$

(use integration by parts). So

$$c_k E(Y) = -k\lambda e^{-k} + (-\lambda e^{-y/\lambda})|_0^{k\lambda} = -k\lambda e^{-k} + \lambda(1 - e^{-k}).$$

Hence

$$E(Y) = \lambda \left[\frac{1 - (k+1)e^{-k}}{1 - e^{-k}}\right].$$

b) Note that

$$c_k E(Y^2) = \int_0^{k\lambda} \frac{y^2}{\lambda} e^{-y/\lambda} dy.$$

Since

$$\frac{d}{dy}[-(y^2 + 2\lambda y + 2\lambda^2)e^{-y/\lambda}] = \frac{1}{\lambda} e^{-y/\lambda}(y^2 + 2\lambda y + 2\lambda^2) - e^{-y/\lambda}(2y + 2\lambda)$$

$$= y^2 \frac{1}{\lambda} e^{-y/\lambda},$$

we have $c_k E(Y^2) = [-(y^2 + 2\lambda y + 2\lambda^2)e^{-y/\lambda}]_0^{k\lambda} =$
$-(k^2\lambda^2 + 2\lambda^2 k + 2\lambda^2)e^{-k} + 2\lambda^2$. So the result follows. □

Since as $k \to \infty$, $E(Y) \to \lambda$, and $E(Y^2) \to 2\lambda^2$, we have $\text{VAR}(Y) \to \lambda^2$. If $k = 9\log(2) \approx 6.24$, then $E(Y) \approx .998\lambda$, and $E(Y^2) \approx 0.95(2\lambda^2)$.

1.7.2 The Truncated Double Exponential Distribution

Suppose that X is a double exponential $DE(\mu, \lambda)$ random variable. Then $\text{MED}(X) = \mu$ and $\text{MAD}(X) = \log(2)\lambda$. Let $c = k\log(2)$, and let the truncation points $a = \mu - k\text{MAD}(X) = \mu - c\lambda$ and $b = \mu + kMAD(X) = \mu + c\lambda$. Let $X_T(a, b) \equiv Y$ be the truncated double exponential $TDE(\mu, \lambda, a, b)$ random variable. Then for $a \le y \le b$, the pdf of Y is

$$f_Y(y|\mu, \lambda, a, b) = \frac{1}{2\lambda(1 - \exp(-c))} \exp(-|y - \mu|/\lambda).$$

Lemma 1.5. a) $E(Y) = \mu$.

$$b)\ \text{VAR}(Y) = 2\lambda^2 \left[\frac{1 - \frac{1}{2}(c^2 + 2c + 2)e^{-c}}{1 - e^{-c}}\right].$$

Proof. a) follows by symmetry and b) follows from Lemma 1.4 b) since $\text{VAR}(Y) = E[(Y - \mu)^2] = E(W_T^2)$ where W_T is $TEXP(\lambda, b = c\lambda)$. □

As $c \to \infty$, $\text{VAR}(Y) \to 2\lambda^2$. If $k = 9$, then $c = 9\log(2) \approx 6.24$ and $\text{VAR}(Y) \approx 0.95(2\lambda^2)$.

1.7.3 The Truncated Normal Distribution

Now if X is $N(\mu, \sigma^2)$, then let Y be a truncated normal $TN(\mu, \sigma^2, a, b)$ random variable. Then

$$f_Y(y) = \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} I_{[a,b]}(y)$$

where Φ is the standard normal cdf. The indicator function

$$I_{[a,b]}(y) = 1 \text{ if } \mathrm{a \le y \le b}$$

and is zero otherwise. Let ϕ be the standard normal pdf.

Lemma 1.6. $E(Y) = \mu + \left[\frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}\right] \sigma$, and

$$V(Y) = \sigma^2 \left[1 + \frac{(\frac{a-\mu}{\sigma})\phi(\frac{a-\mu}{\sigma}) - (\frac{b-\mu}{\sigma})\phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}\right] - \sigma^2 \left[\frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}\right]^2.$$

(See Johnson and Kotz 1970a, p. 83.)

Proof. Let $c =$

$$\frac{1}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}.$$

Then $E(Y) = \int_a^b y f_Y(y) dy$. Hence

$$\frac{1}{c} E(Y) = \int_a^b \frac{y}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$= \int_a^b \left(\frac{y-\mu}{\sigma}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy + \frac{\mu}{\sigma} \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$= \int_a^b \left(\frac{y-\mu}{\sigma}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy + \mu \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy.$$

Note that the integrand of the last integral is the pdf of a $N(\mu, \sigma^2)$ distribution. Let $z = (y-\mu)/\sigma$. Thus $dz = dy/\sigma$, and $E(Y)/c =$

$$\int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \sigma \frac{z}{\sqrt{2\pi}} e^{-z^2/2} dz + \frac{\mu}{c} = \frac{\sigma}{\sqrt{2\pi}} (-e^{-z^2/2})\Big|_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} + \frac{\mu}{c}.$$

Multiplying both sides by c gives the expectation result.

$$E(Y^2) = \int_a^b y^2 f_Y(y)dy.$$

Hence

$$\frac{1}{c}E(Y^2) = \int_a^b \frac{y^2}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy$$

$$= \sigma \int_a^b \left(\frac{y^2}{\sigma^2} - \frac{2\mu y}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy$$

$$+\sigma \int_a^b \frac{2y\mu - \mu^2}{\sigma^2} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy$$

$$= \sigma \int_a^b \left(\frac{y-\mu}{\sigma}\right)^2 \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy + 2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c}.$$

Let $z = (y-\mu)/\sigma$. Then $dz = dy/\sigma$, $dy = \sigma dz$, and $y = \sigma z + \mu$. Hence

$$\frac{E(Y^2)}{c} = 2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} + \sigma \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \sigma \frac{z^2}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

Next integrate by parts with $w = z$ and $dv = ze^{-z^2/2}dz$. Then $E(Y^2)/c =$

$$2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} + \frac{\sigma^2}{\sqrt{2\pi}}[(-ze^{-z^2/2})|_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} + \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-z^2/2}dz]$$

$$= 2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} + \sigma^2\left[\left(\frac{a-\mu}{\sigma}\right)\phi\left(\frac{a-\mu}{\sigma}\right) - \left(\frac{b-\mu}{\sigma}\right)\phi\left(\frac{b-\mu}{\sigma}\right) + \frac{1}{c}\right].$$

Using

$$\text{VAR}(Y) = c\frac{1}{c}E(Y^2) - (E(Y))^2$$

gives the result. □

Corollary 1.7. Let Y be $TN(\mu, \sigma^2, a = \mu - k\sigma, b = \mu + k\sigma)$. Then $E(Y) = \mu$ and $V(Y) = \sigma^2 \left[1 - \frac{2k\phi(k)}{2\Phi(k) - 1}\right]$.

Proof. Use the symmetry of ϕ, the fact that $\Phi(-x) = 1 - \Phi(x)$, and the above lemma to get the result. □

Examining $V(Y)$ for several values of k shows that the $TN(\mu, \sigma^2, a = \mu - k\sigma, b = \mu + k\sigma)$ distribution does not change much for $k > 3.0$. See Table 1.2.

1.7.4 The Truncated Cauchy Distribution

If X is a Cauchy $C(\mu, \sigma)$ random variable, then $\text{MED}(X) = \mu$ and $\text{MAD}(X) = \sigma$. If Y is a truncated Cauchy $TC(\mu, \sigma, \mu - a\sigma, \mu + b\sigma)$ random variable, then

Table 1.2 Variances for several truncated normal distributions

k	$V(Y)$
2.0	$0.774\sigma^2$
2.5	$0.911\sigma^2$
3.0	$0.973\sigma^2$
3.5	$0.994\sigma^2$
4.0	$0.999\sigma^2$

$$f_Y(y) = \frac{1}{\tan^{-1}(b) + \tan^{-1}(a)} \frac{1}{\sigma[1 + (\frac{y-\mu}{\sigma})^2]}$$

for $\mu - a\sigma < y < \mu + b\sigma$. Moreover,

$$E(Y) = \mu + \sigma \left(\frac{\log(1 + b^2) - \log(1 + a^2)}{2[\tan^{-1}(b) + \tan^{-1}(a)]} \right), \quad \text{and}$$

$$V(Y) = \sigma^2 \left[\frac{b + a - \tan^{-1}(b) - \tan^{-1}(a)}{\tan^{-1}(b) + \tan^{-1}(a)} - \left(\frac{\log(1 + b^2) - \log(1 + a^2)}{\tan^{-1}(b) + \tan^{-1}(a)} \right)^2 \right].$$

Lemma 1.8. If $a = b$, then $E(Y) = \mu$, and $V(Y) = \sigma^2 \left[\frac{b - \tan^{-1}(b)}{\tan^{-1}(b)} \right]$.

See Johnson and Kotz (1970a, p. 162) and Dahiya et al. (2001).

1.8 Summary

1) Given a small data set, find $\overline{Y}$, S, $\text{MED}(n)$, and $\text{MAD}(n)$. 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 $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ 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 $\text{MAD}(n)$, find $D_i = |Y_i - \text{MED}(n)|$, and then find the sample median of the D_i by ordering them from smallest to largest and finding the middle value or values.

1.9 Complements

Olive (2008, chapters 2–4) covered robust estimators for the location model, including alternatives to the Olive (2005b) confidence interval for the median given in Application 1.1.

Riani et al. (2009) found the population mean and covariance matrix of an elliptically trimmed (truncated) multivariate normal distribution, using Tallis (1963).

1.10 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?

R Problem

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, e.g., *medci*, will display the code for the function. Use the `args` command, e.g., *args(medci)*, to display the needed arguments for the function.

1.3. a) Use the commands

```
height <- rnorm(87, mean=1692, sd = 65)
height[61:65] <- 19.0
```

to simulate data similar to the Buxton heights.

Download the *mpack* functions, `cci` and `medci`, which produce a classical CI, and a CI using the median and the Bloch and Gastwirth (1968) SE. The default is a 95% CI.

b) Compute a 95% CI for the artificial height data set with the command *cci(height)*.

c) Compute a 95% CI for the artificial height data set with the command *medci(height)*.

Chapter 2
Multivariate Distributions

This chapter describes the multivariate location and dispersion (MLD) model, random vectors, the population mean, the population covariance matrix, and the classical MLD estimators: the sample mean and the sample covariance matrix. Some important results on Mahalanobis distances and the volume of a hyperellipsoid are given. Often methods of multivariate analysis work best when the variables $x_1, ..., x_p$ are linearly related. Section 2.4 discusses power transformations to remove gross linearities from the variables.

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, e.g., $f(\boldsymbol{z}) = c \exp(\boldsymbol{z}^T \boldsymbol{z})$, $\boldsymbol{z}$ is a column vector.

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_2

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.* Some texts, for example Johnson and Wichern (1988, pp. 7, 53), use $\boldsymbol{X}$ to denote the $n \times p$ data matrix and an $n \times 1$ 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* 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

The population location vector $\boldsymbol{\mu}$ need not be the population mean, but often the population mean is denoted by $\boldsymbol{\mu}$. For elliptically contoured distributions, such as the multivariate normal distribution, $\boldsymbol{\mu}$ is usually the point of symmetry for the population distribution. See Chapter 3.

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*

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

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

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

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

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

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

Then

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

and

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

Let the population standardized random variables

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

for $i = 1, ..., p$. Then $\text{Cor}(\boldsymbol{x}) = \boldsymbol{\rho}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{z})$ 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

$$E(\boldsymbol{x}) = \boldsymbol{\mu} \text{ and } \operatorname{Cov}(\boldsymbol{x}) = \mathrm{c_x}\boldsymbol{\Sigma} = \boldsymbol{\Sigma_x}$$

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

Definition 2.5. Let $x_{1j}, ..., x_{nj}$ be measurements on the jth 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 $\operatorname{Cov}(X_i, X_j) = \sigma_{ij}$, 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 $\operatorname{Cor}(X_i, X_j) = \rho_{ij}$, 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 $(n-1)\boldsymbol{S} = \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\ \overline{\boldsymbol{x}}^T =$

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

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

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

$$Z_j = \frac{x_j - \overline{x}_j}{\sqrt{S_{jj}}}$$

for $j = 1, ..., p$. Then the sample correlation matrix $\boldsymbol{R}$ is the sample covariance matrix of the $\boldsymbol{z}_i = (Z_{i1}, ..., Z_{ip})^T$ where $i = 1, ..., n$.

Often it is useful to standardize variables with a robust location estimator and a robust scale estimator. The *R* function `scale` is useful. The *R* code below shows how to standardize using

$$Z_j = \frac{x_j - \text{MED}(x_j)}{\text{MAD}(x_j)}$$

for $j = 1, ..., p$. Here $\text{MED}(x_j) = \text{MED}(x_{1j}, ..., x_{nj})$ and $\text{MAD}(x_j) = \text{MAD}(x_{1j}, ..., x_{nj})$ are the sample median and sample median absolute deviation of the data for the jth variable: $x_{1j}, ..., x_{nj}$. See Definitions 1.3 and 1.5. Some of these results are illustrated with the following *R* code.

```
x <- buxx[,1:3]; cov(x)
                len        nasal     bigonal
len     118299.9257 -191.084603 -104.718925
nasal     -191.0846   18.793905   -1.967121
bigonal   -104.7189   -1.967121   36.796311

cor(x)
                len        nasal     bigonal
len      1.00000000 -0.12815187 -0.05019157
nasal   -0.12815187  1.00000000 -0.07480324
bigonal -0.05019157 -0.07480324  1.00000000
z <- scale(x)
cov(z)
                len        nasal     bigonal
len      1.00000000 -0.12815187 -0.05019157
nasal   -0.12815187  1.00000000 -0.07480324
bigonal -0.05019157 -0.07480324  1.00000000

medd <- apply(x,2,median)
madd <- apply(x,2,mad)/1.4826
z <- scale(x,center=medd,scale=madd)
ddplot4(z)#scaled data still has 5 outliers
cov(z)    #in the length variable
                len       nasal   bigonal
len     4731.997028 -12.738974 -6.981262
nasal    -12.738974   2.088212 -0.218569
```

```
bigonal   -6.981262  -0.218569  4.088479

cor(z)
                len       nasal     bigonal
len      1.00000000 -0.12815187 -0.05019157
nasal   -0.12815187  1.00000000 -0.07480324
bigonal -0.05019157 -0.07480324  1.00000000

apply(z,2,median)
len   nasal bigonal
0       0       0
#scaled data has coord. median = (0,0,0)^T
apply(z,2,mad)/1.4826
len   nasal bigonal
1       1       1 #scaled data has unit MAD
```

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

Rule of thumb 2.1. Multivariate procedures start to give good results for $n \geq 10p$, especially if the distribution is close to multivariate normal. In particular, we want $n \geq 10p$ for the sample covariance and correlation matrices. For procedures with large sample theory on a large class of distributions, for any value of n, there are always distributions where the results will be poor, but will eventually be good for larger sample sizes. Norman and Streiner (1986, pp. 122, 130, 157) gave this rule of thumb and note that some authors recommend $n \geq 30p$. This rule of thumb is much like the rule of thumb that says the central limit theorem normal approximation for $\overline{Y}$ starts to be good for many distributions for $n \geq 30$. See the paragraph below Theorem 3.7.

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

$$\boldsymbol{S} = \boldsymbol{DRD}, \tag{2.4}$$

and

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

The inverse covariance matrix or inverse correlation matrix can be used to find the partial correlation $r_{ij,\boldsymbol{x}(ij)}$ between x_i and x_j where $\boldsymbol{x}(ij)$ is the vector of predictors with x_i and x_j deleted where $i \neq j$. This partial correlation is the correlation of x_i and x_j after eliminating the linear effects of $\boldsymbol{x}(ij)$ from both variables: regress x_i and x_j on $\boldsymbol{x}(ij)$ and get the two sets of residuals, and then find the correlation of the two sets of residuals. If $p \geq 3$ and $\boldsymbol{S}^{-1} = (S^{ij})$, then

$$r_{ij,\boldsymbol{x}(ij)} = \frac{-S^{ij}}{(S^{ii}S^{jj})^{1/2}} = \frac{-r^{ij}}{(r^{ii}r^{jj})^{1/2}}.$$

Srivastava and Khatri(1979, p. 53) proved this result. The second equality holds since $\boldsymbol{R}^{-1} = \boldsymbol{D}\boldsymbol{S}^{-1}\boldsymbol{D} = (r^{ij}) = (S^{ij}\sqrt{S_{ii}}\sqrt{S_{jj}})$.

Some *R* code illustrating this result is shown below. The function `lsfit` is used to regress x_1 on x_3 and then regress x_2 on x_3. Note that $\boldsymbol{x}(i = 1, j = 2) = x_3$ once x_1 and x_2 have been deleted since $p = 3$.

```
x <- buxx[,1:3]; z<-solve(cor(x))
z #inverse correlation matrix

               len      nasal    bigonal
len     1.02042523 0.13535798 0.06134196
nasal   0.13535798 1.02358206 0.08336109
bigonal 0.06134196 0.08336109 1.00931453

out1 <- lsfit(x[,3],x[,1])$resid
out2 <- lsfit(x[,3],x[,2])$resid
cor(out1,out2)
[1] -0.1324439

-z[1,2]/sqrt(z[1,1]*z[2,2])
[1] -0.1324439

zz <- solve(var(x)) #inverse covariance matrix
-zz[1,2]/sqrt(zz[1,1]*zz[2,2])
[1] -0.1324439
```

2.3 Mahalanobis 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 $\boldsymbol{e}_i^T\boldsymbol{e}_j = 0$ if $i \neq j$ 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, pp. 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 $\boldsymbol{A}$ is a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^p \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$, 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, pp. 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, e.g., 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 Definitions 3.2 and 3.3) if g is continuous and decreasing. Let $\boldsymbol{w} = \boldsymbol{x} - \boldsymbol{\mu}$. Then points at a 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 a 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, pp. 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 $\boldsymbol{Z}$ 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 (1999a). The 10%, 30%, 50%, 70%, 90%, and 98% highest density regions are shown for two multivariate normal (MVN) distributions. Both distributions have $\boldsymbol{\mu} = \boldsymbol{0}$. 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.

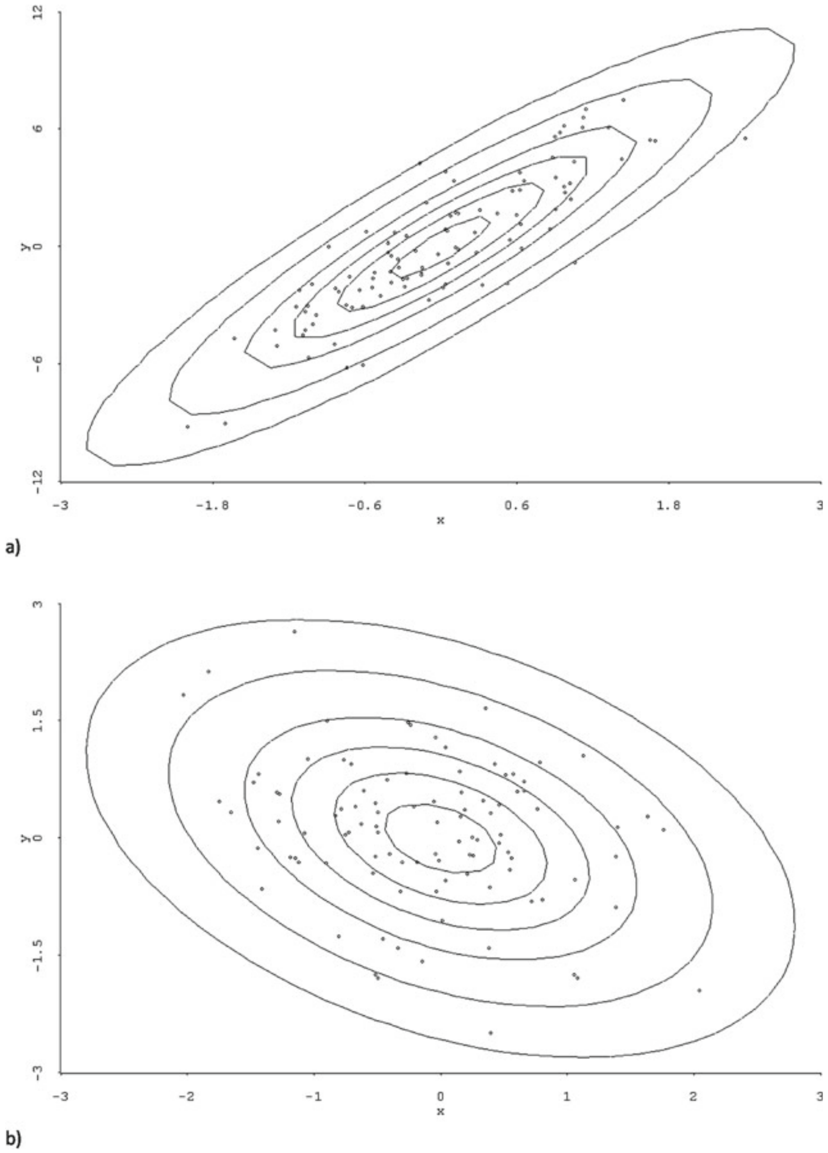

Fig. 2.1 Highest Density Regions for 2 MVN Distributions

2.4 Predictor Transformations

In regression, there is a response variable $w_1 = Y$ of interest, and predictor variables $w_2, ..., w_p$ are used to predict Y. In multivariate analysis, all p random variables $x_1, ..., x_p$ are of interest.

Predictor transformations are used to remove gross nonlinearities in the predictors w_i or the random variables x_i, 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 $\lambda = 0$. The *modified power transformation* also has $x = t_0(w) = \log(w)$, but for $\lambda \neq 0$,

$$x = t_\lambda(w) = \frac{w^\lambda - 1}{\lambda}.$$

For both the power and modified power transformations, often $\lambda \in \varLambda_L$ where

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

is called the *ladder of powers*. Often when a power transformation is needed, a transformation that goes "down the ladder," e.g., from $\lambda = 1$ to $\lambda = 0$ will be useful. If the transformation goes too far down the ladder, e.g., 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 random variables.

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, along with the name of the variable. The software *R* 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 the 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 random variables.

Rule of thumb 2.2. a) If strong nonlinearities are apparent in the scatterplot matrix of the random variables $x_1, ..., x_p$, it is often useful to remove the nonlinearities by transforming the random variables 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 toward large values of X_1, increase the power of X_1. If the curve bulges toward small values of X_1, decrease the power of X_1. See Tukey (1977, pp. 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 the 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. 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 toward 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 toward 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 toward 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 val-

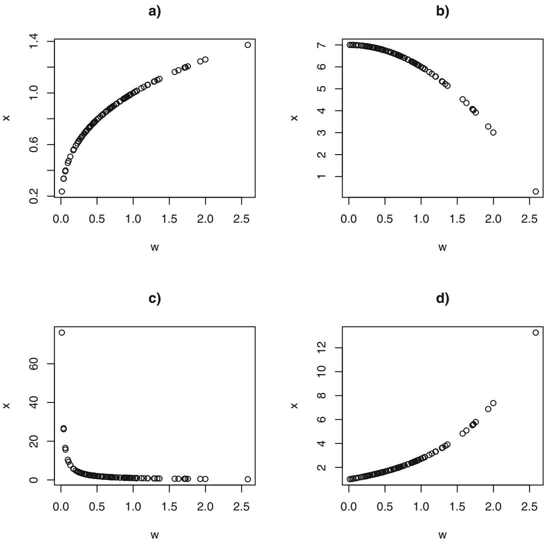

Fig. 2.2 Plots to Illustrate the Bulging and Ladder Rules

ues of x need spreading. Notice that the plotted points bulge down toward large values of the horizontal variable.

Example 2.2. Mussel Data. Cook and Weisberg (1999a, pp. 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, *height* H, and the *width* W of the shell in mm, 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 labeled from 150 to 300 on the bottom of the 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 $\log(W)$.

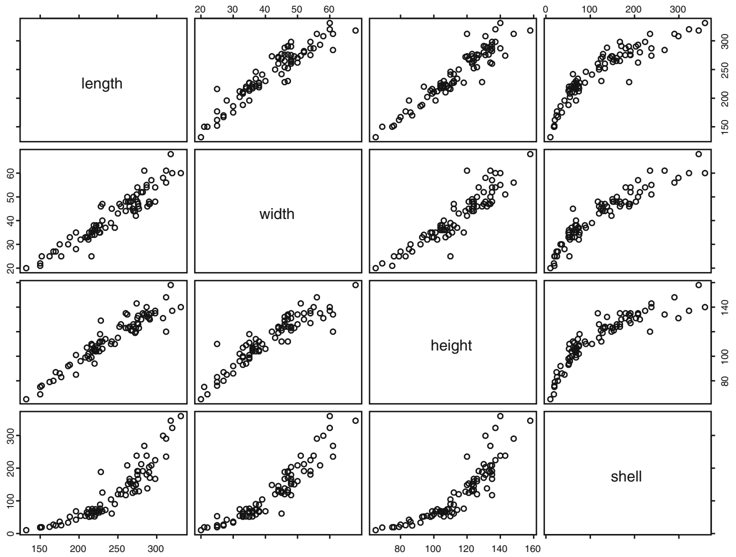

Fig. 2.3 Scatterplot Matrix for Original Mussel Data Predictors

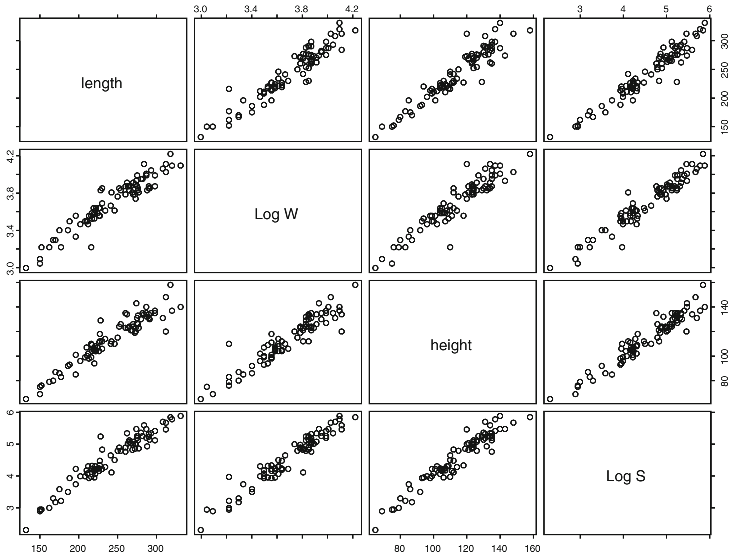

Fig. 2.4 Scatterplot Matrix for Transformed Mussel Data Predictors

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* $\text{Cov}(\boldsymbol{x}) = E(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T = (\sigma_{ij}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$.

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

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

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

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 $\boldsymbol{S} = \boldsymbol{DRD}$, and $\boldsymbol{R} = \boldsymbol{D}^{-1}\boldsymbol{S}\boldsymbol{D}^{-1}$.

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

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 $\{\boldsymbol{x}|D^2_{\boldsymbol{x}} \le h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T\boldsymbol{S}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{x}}) \le h^2\}$ is centered at $\overline{\boldsymbol{x}}$ and has volume equal to

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

Let $\boldsymbol{S}$ have eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ where $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_p$. 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 $i \ne j$ 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.

22) For multivariate procedures with p variables, we want $n \geq 10p$. This rule of thumb will be used for the sample covariance matrix $\boldsymbol{S}$, the sample correlation matrix $\boldsymbol{R}$, and procedures that use these matrices such as principal component analysis, factor analysis, canonical correlation analysis, Hotelling's T^2, discriminant analysis for each group, and one way MANOVA for each group.

2.6 Complements

Section 2.3 will be useful for principal component analysis and for prediction regions. Fan (2017) gave a useful one-number summary of the correlation matrix that acts like a squared correlation.

2.7 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

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

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

2.3. 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.

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 the 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 $\boldsymbol{A}\boldsymbol{e} = \lambda\boldsymbol{e}$ so $\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{e} = \boldsymbol{A}^{-1}\lambda\boldsymbol{e}$. Find an eigenvalue eigenvector pair for $\boldsymbol{A}^{-1}$.

Problems using ARC

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 2.7a." 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 HW2d7.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 Problem

Note: For the following problem, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/mrsashw.txt) into *R*.

2.10. Use the following *R* 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

This chapter considers elliptically contoured distributions, including the multivariate normal distribution. These distributions are important models for multivariate data. Sample Mahalanobis distances and a brief review of large sample theory are also covered.

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

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_3

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

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).

From Definition 2.3, recall that 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*

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

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

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

Thus

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

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, pp. 127-132) or Severini (2005, ch. 8).

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

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

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 $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)^T$ and $\boldsymbol{\Sigma} = diag(\sigma_1^2, ..., \sigma_p^2)$ (so the off-diagonal entries $\sigma_{ij} = 0$ while the diagonal entries of $\boldsymbol{\Sigma}$ are $\sigma_{ii} = \sigma_i^2$).

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 and b is a constant, then $\boldsymbol{a} + b\boldsymbol{X} \sim N_p(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$. (Note that $b\boldsymbol{X} = b\boldsymbol{I}_p\boldsymbol{X}$ with $\boldsymbol{A} = b\boldsymbol{I}_p$.)

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

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

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

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

and the conditional variance

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

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

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

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

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

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 $\text{VAR}(X) = \text{VAR}(Y) = 1$, but $\text{Cov}(X,Y) = \pm\rho$. Hence $f(x,y) =$

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

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

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 $Y|\boldsymbol{X}_2 = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e$ follows the multiple linear regression model.

3.2 Elliptically Contoured Distributions

Definition 3.2: Johnson (1987, pp. 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 3.3. 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 (3.6). See Eaton (1986) and Cook (1998, pp. 57, 130).

Lemma 3.4. Let $\boldsymbol{X}$ be a $p \times 1$ random vector with 1st moments; i.e., $E(\boldsymbol{X})$ 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 second moments, $\text{Cov}(\boldsymbol{X})$ 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

$$\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \mu_Y \\ \boldsymbol{\mu}_X \end{pmatrix}, \ \text{and } \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{pmatrix}.$$

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 $\text{Cov}(\boldsymbol{X})$ is nonsingular,

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

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

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{0} \end{pmatrix}.$$

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

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

$$\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{M}_{1B} \\ \boldsymbol{M}_{2B} \end{pmatrix} \begin{pmatrix} \boldsymbol{A}^T & \boldsymbol{0}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{X}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{X}_2 - \boldsymbol{\mu}_2 \end{pmatrix}$$

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

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

$$= \begin{pmatrix} \boldsymbol{M}_{1B} \\ \boldsymbol{M}_{2B} \end{pmatrix}.$$

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 $\boldsymbol{\Sigma}_{11}$. □

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

$$E\left[\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \mid \boldsymbol{X}\right] = E\left[\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \mid \boldsymbol{B}^T \begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix}\right]$$

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

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 et al. (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. See Problem 3.4 for a related result.

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) showed 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

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

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

The following $EC(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution for a $p \times 1$ random vector $\boldsymbol{x}$ is the uniform distribution on a hyperellipsoid where $f(\boldsymbol{z}) = c$ for $\boldsymbol{z}$ in the hyperellipsoid where c is the reciprocal of the volume of the hyperellipsoid. The pdf of the distribution is

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

See Theorem 2.4 or Equation (5.16) where $h^2 = p + 2$. Then $E(\boldsymbol{x}) = \boldsymbol{\mu}$ by symmetry and is can be shown that $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$.

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. See Problem 3.24.

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 3.4. 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

$$D^2_{\boldsymbol{X}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \tag{3.13}$$

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 hyperellipsoid boundaries of the form $(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = a^2$. 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

$$\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \le \chi^2_p(\alpha)\} = \{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \le \chi^2_p(\alpha)\}$$

where $P(W \le \chi^2_p(\alpha)) = \alpha$ if $W \sim \chi^2_p$. 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 $k \ne j$. 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. Often the bootstrap can be used to compute the SE. See Section 5.3.

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 distributions "close" to the normal distribution, the central limit theorem provides a good approximation if the sample size $n \geq 30$. Hesterberg (2014, pp. 41, 66) suggests $n \geq 5000$ is needed for moderately skewed distributions. A special case of the CLT is proven after Theorem 3.20.

Notation. The notation $X \sim Y$ and $X \overset{D}{=} Y$ 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. See Definition 3.5. 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 $\text{VAR}(X) = \sigma_X^2$.

Example 3.4. a) Let $Y_1, ..., Y_n$ be iid Ber(ρ). Then $E(Y) = \rho$ and $\text{VAR}(Y) = \rho(1-\rho)$. (The Bernoulli ($\rho$) distribution is the binomial (1,ρ) distribution.) 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 $Y_n \overset{D}{=} \sum_{i=1}^n X_i$ where $X_1, ..., X_n$ are iid Ber(ρ). Hence

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

since

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

by a).

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

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

or

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

Theorem 3.8: the Delta Method. If g does not depend on n, $g'(\theta) \neq 0$, 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 $\text{VAR}(Y) = \sigma^2$. 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 $X \sim \text{Binomial}(n, p)$ 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,

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

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

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

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

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

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

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

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

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 3.5. 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 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 $X_n \overset{D}{\to} X$, then $P(a < X_n \le b) = F_n(b) - F_n(a) \to F(b) - F(a) = P(a < X \le b)$ 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 3.9. Suppose that $X_n \sim U(-1/n, 1/n)$. Then the cdf $F_n(x)$ of X_n is

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

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 $n \ge 1$. 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

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

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 $F_X(x)$ (i.e., for $x \ne 0$),

$$X_n \overset{D}{\to} X.$$

Example 3.10. Suppose $Y_n \sim U(0, n)$. Then $F_n(t) = t/n$ for $0 < t \le n$ and $F_n(t) = 0$ for $t \le 0$. Hence $\lim_{n\to\infty} F_n(t) = 0$ for $t \le 0$. 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 3.6. A sequence of random variables X_n *converges in distribution to a constant* $\tau(\theta)$, written

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

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

Definition 3.7. 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 $\epsilon > 0$,

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

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 $\epsilon > 0$,

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

Definition 3.8. Let the *parameter space* Θ be the set of possible values of θ. 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 3.9. 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

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

if

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

as $n \to \infty$.

Lemma 3.9: Generalized Chebyshev's Inequality. Let $u : \mathbb{R} \to [0, \infty)$ 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 $\sigma^2 = \text{VAR}(Y)$ exists, then we obtain
Chebyshev's Inequality:

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

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

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

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

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

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

The following proposition gives sufficient conditions for T_n to be a consistent estimator of $\tau(\theta)$. Notice that $E_\theta[(T_n - \tau(\theta))^2] = MSE_{\tau(\theta)}(T_n) \to 0$ for all $\theta \in \Theta$ is equivalent to $T_n \overset{qm}{\to} \tau(\theta)$ 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

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

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

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

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

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_θ be a random variable with distribution depending on θ, and $0 < \delta \leq 1$. If

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

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 3.10. 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 $n \to \infty$. □

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 3.11. Lehmann (1999, pp. 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

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

for all $n \geq N_\epsilon$.

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

Definition 3.12. 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.

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 $W_n \asymp_P X_n$,

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

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

b) Since $W_n \asymp_P X_n$,

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

for all $n \ge N_\epsilon$. 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

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

for all $n \ge N_1$, and

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

for all $n \ge N_2$. Hence

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

and

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

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

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

for all $n \ge N$. Hence $W_n \asymp_P X_n$. $\square$

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

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

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 $\|T^*_n - \boldsymbol{\beta}\| = O_P(n^{-\delta})$. $\square$

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. We are assuming that the function τ does not depend on n.

Example 3.11. 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 Theorems 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. Assume that the function g does not depend on n.
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 3.12. (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 3.13. 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: $Y_n \overset{D}{=} Z/\sqrt{V_n/n}$ 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 $V_n \overset{D}{=} \sum_{i=1}^n X_i$ 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 3.18e.

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 characteristic function (cf) $\phi(t)$.

a)

$$Y_n \xrightarrow{D} Y \ \text{ iff } \ \phi_{\text{n}}(\text{t}) \to \phi(\text{t}) \ \forall \text{t} \in \mathbb{R}.$$

b) Also assume that Y_n has moment generating function (mgf) m_n and Y has mgf m. Assume that all of the mgfs m_n and m are defined on $|t| \le d$ 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, pp. 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 $|t| < \sigma t_o$. We 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

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

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

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,

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

3.4.4 Multivariate Limit Theorems

Many of the univariate results of the previous three 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 3.13. 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** r**th 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 $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$, 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 $\boldsymbol{g}$ does not depend on n and

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

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

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

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

Definition 3.14. 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 functions $\phi_n(\boldsymbol{t})$, and let $\boldsymbol{X}$ be a $k \times 1$ random vector with cf $\phi(\boldsymbol{t})$. Then

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

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

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

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

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

Application: Proof of the MCLT Theorem 3.21. Note that for fixed $\boldsymbol{t}$, the $\boldsymbol{t}^T\boldsymbol{X}_i$ are iid random variables with mean $\boldsymbol{t}^T\boldsymbol{\mu}$ and variance $\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}$. Hence by the CLT, $\boldsymbol{t}^T\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N(0, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$. The right-hand side has distribution $\boldsymbol{t}^T\boldsymbol{X}$ where $\boldsymbol{X} \sim N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. Hence by the Cramér Wold Device, $\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. □

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

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

Let $g(n) \geq 1$ be an increasing function of the sample size n: $g(n) \uparrow \infty$, e.g., $g(n) = \sqrt{n}$. 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 3.15. 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 $g(n)$.

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

The following two theorems are taken from Severini (2005, pp. 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 $\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{C}$, 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, e.g., $Y_n = n^{-1/2}$.)

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) Let $\boldsymbol{\Sigma} > 0$. 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 $s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

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

(Note that $D^2_{\boldsymbol{x}}(T,\boldsymbol{C}) = s^{-1}D^2_{\boldsymbol{x}}(\boldsymbol{\mu},\boldsymbol{\Sigma}) + O_P(n^{-\delta})$ if $(T,\boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^δ where $0 < \delta \le 0.5$ if $[\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}] = O_P(n^{-\delta})$.)

Alternatively, $D^2_{\boldsymbol{x}}(T,\boldsymbol{C})$ is a continuous function of $(T,\boldsymbol{C})$ if $\boldsymbol{C} > 0$ for $n > 10p$. Hence $D^2_{\boldsymbol{x}}(T,\boldsymbol{C}) \xrightarrow{P} D^2_{\boldsymbol{x}}(\mu, s\boldsymbol{\Sigma})$.

iii) 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

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

Also

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

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

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

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 $\boldsymbol{X}+\boldsymbol{a} \sim N_p(\boldsymbol{\mu}+\boldsymbol{a},\boldsymbol{\Sigma})$.

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

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)

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

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 joint pdf

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

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

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

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 if $n >> p$.

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 $F_n(\boldsymbol{x}) \rightarrow F(\boldsymbol{x})$ as $n \rightarrow \infty$ 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 $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \rightarrow 0$ as $n \rightarrow \infty$.

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 $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$, 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 et al. (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 et al. (1990), Fang and Anderson (1990), and Gupta et al. (2013). Fang et al. (1990) sketched the history of elliptically contoured distributions while Gupta et al. (2013) discussed matrix valued elliptically contoured distributions. Cambanis et al. (1981), Chmielewski (1981), and Eaton (1986) are also important references. Also see Muirhead (1982, pp. 30–42).

Section 3.4 followed Olive (2014, ch. 8) closely, which is a good master's level treatment of large sample theory. 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) considered asymptotic theory for econometric applications. Also see DasGupta (2008), Davidson (1994), Jiang (2010), Polansky (2011), Sen et al. (2010), and van der Vaart (1998).

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.4). 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 $\text{Cov}(\boldsymbol{X})$.

3.6*. The table ($\boldsymbol{W}$) below represents three head measurements on six 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 $\text{MED}(\boldsymbol{W})$.

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

```
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
```

3.7. 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 $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$. 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

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

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 $1 \le r \le p$. 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 depends 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 $\text{Cov}(\boldsymbol{x})$. Let $\boldsymbol{A}$ be an $r \times p$ constant matrix and let $\boldsymbol{B}$ be a $q \times p$ constant matrix. Find $\text{Cov}(\boldsymbol{Ax}, \boldsymbol{Bx})$ in terms of $\boldsymbol{A}, \boldsymbol{B}$, and $\text{Cov}(\boldsymbol{x})$.

3.13. The table $\boldsymbol{W}$ shown below represents four measurements on five 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 $\text{MED}(\boldsymbol{W})$.

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 $\text{Cov}(\boldsymbol{x}) = \dfrac{d}{d-2}\boldsymbol{\Sigma}$ 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 $\text{Cov}(\boldsymbol{x}_i) = [1 + \gamma(c-1)]\boldsymbol{\Sigma}$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{d})$ for appropriate vector $\boldsymbol{d}$.

3.17. Let $\boldsymbol{X}$ be an $n \times p$ constant matrix and let $\boldsymbol{\beta}$ be a $p \times 1$ constant vector. Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$. 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.

3.18. 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 $\text{Var}(Y|X)$.

3.19. 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 $\rho(X_1, X_4)$.

3.20. 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 $\rho(X_1, X_3)$.

3.21. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors where $E(\boldsymbol{x}_i) = e^{0.5}\boldsymbol{1}$ 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 $\boldsymbol{c}$.

3.22. 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 $\rho(X_1, X_3)$.

3.23. 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 $\text{Var}(Y|X)$.

3.24. 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, pp. 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
$\text{Cov}(X_{i1}, X_{i2}) = \exp[\mu_1 + \mu_2 + 0.5(\sigma_1^2 + \sigma_2^2) + \sigma_{12}][\exp(\sigma_{12}) - 1]$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

3.25. Following Srivastava and Khatri (2001, p. 47), let

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N_p\left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right].$$

a) Show that the nonsingular linear transformation

$$\begin{pmatrix} \boldsymbol{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2 \\ \boldsymbol{X}_2 \end{pmatrix} \sim$$

$$N_p\left[\begin{pmatrix} \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right].$$

b) Then $\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2 \perp\!\!\!\perp \boldsymbol{X}_2$, and

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

By independence, $\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2$ has the same distribution as $(\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2)|\boldsymbol{X}_2$, and the term $-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2$ is a constant, given $\boldsymbol{X}_2$. Use this result to show that

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

Chapter 4
MLD Estimators

This chapter is the most important chapter for outlier robust statistics and covers robust estimators of multivariate location and dispersion. The practical, highly outlier resistant, $\sqrt{n}$ consistent FCH, RFCH, and RMVN estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ are developed along with proofs. The RFCH and RMVN estimators are reweighted versions of the FCH estimator. It is shown why competing "robust estimators" fail to work, are impractical, or are not yet backed by theory. The RMVN and RFCH sets are defined and will be used to create practical robust methods of principal component analysis, canonical correlation analysis, discriminant analysis, factor analysis, and multivariate linear regression in the following chapters.

Warning: This chapter contains many acronyms, abbreviations, and estimator names such as FCH, RFCH, and RMVN. See Section 1.2 and Table 1.1 for many of the acronyms and for the acronyms that start with the added letter A, C, F, or R: A stands for *algorithm*, C for *concentration*, F for estimators that use a *fixed* number of trial fits, and R for *reweighted*.

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

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_4

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 $(\overline{\boldsymbol{z}} = \boldsymbol{0}, \boldsymbol{R})$, we want $n \geq 10p$. We want $n \geq 20p$ 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 $\boldsymbol{B} = \boldsymbol{1}\boldsymbol{b}^T$ where $\boldsymbol{1}$ 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 $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}$ where $\boldsymbol{A}$ is any nonsingular $p \times p$ matrix. An affine transformation changes $\boldsymbol{x}_i$ to $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b}$ for $i = 1, ..., n$, and affine equivariant multivariate location and dispersion estimators change in natural ways.

Definition 4.1. The multivariate location and dispersion estimator $(T, \boldsymbol{C})$ is *affine equivariant* if

$$T(\boldsymbol{Z}) = T(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}T(\boldsymbol{W}) + \boldsymbol{b}, \tag{4.1}$$

$$\text{and } \boldsymbol{C}(\boldsymbol{Z}) = \boldsymbol{C}(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}\boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}^T. \tag{4.2}$$

The following proposition shows that the Mahalanobis distances are invariant under affine transformations. See Rousseeuw and Leroy (1987, pp. 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^2_i.

Proposition 4.1. If $(T, \boldsymbol{C})$ is affine equivariant, then

$$D^2_i(\boldsymbol{W}) \equiv D^2_i(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = D^2_i(T(\boldsymbol{Z}), \boldsymbol{C}(\boldsymbol{Z})) \equiv D^2_i(\boldsymbol{Z}). \tag{4.3}$$

Proof. Since $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}$ has ith row $\boldsymbol{z}_i^T = \boldsymbol{x}_i^T\boldsymbol{A}^T + \boldsymbol{b}^T$,

$$D^2_i(\boldsymbol{Z}) = [\boldsymbol{z}_i - T(\boldsymbol{Z})]^T \boldsymbol{C}^{-1}(\boldsymbol{Z})[\boldsymbol{z}_i - T(\boldsymbol{Z})]$$

$$= [\boldsymbol{A}(\boldsymbol{x}_i - T(\boldsymbol{W}))]^T[\boldsymbol{A}\boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}^T]^{-1}[\boldsymbol{A}(\boldsymbol{x}_i - T(\boldsymbol{W}))]$$

$$= [\boldsymbol{x}_i - T(\boldsymbol{W})]^T \boldsymbol{C}^{-1}(\boldsymbol{W})[\boldsymbol{x}_i - T(\boldsymbol{W})] = D^2_i(\boldsymbol{W}). \quad \square$$

Warning: Estimators that use randomly chosen elemental sets or projections are not affine equivariant since these estimators often change when they are computed several times (corresponding to the identity transformation $\boldsymbol{A} = \boldsymbol{I}_p$). Such estimators can sometimes be made pseudo-affine equivariant by using the same fixed random number seed and random number generator each time the estimator is used. Then the pseudo-affine equivariance of the estimator depends on the random number seed and the random number generator, and such estimators are not as attractive as affine equivariant estimators that do not depend on a fixed random number seed and random number generator.

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

$$B(T, \boldsymbol{W}) = \min \left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty \right\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$ and $1 \le d_n \le n$. Let $\lambda_1(\boldsymbol{C}(\boldsymbol{W})) \ge \cdots \ge \lambda_p(\boldsymbol{C}(\boldsymbol{W})) \ge 0$ 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

$$B(\boldsymbol{C}, \boldsymbol{W}) = \min \left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \max \left[\frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{W}_d^n))}, \lambda_1(\boldsymbol{C}(\boldsymbol{W}_d^n)) \right] = \infty \right\}.$$

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. For an affine equivariant estimator, the largest possible breakdown value is $n/2$ or $(n+1)/2$ for n even or odd, respectively. Hence in the proof of the following result, we could replace $d_n < d_T$ by $d_n < \min(n/2, d_T)$.

Proposition 4.2. Fix n. 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 $d_T = d_{T,n}$ 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. Suppose the multivariate location estimator T satisfies $\|T(\boldsymbol{W}_d^n)\| \le M$ for some constant M if $d_n < d_T$. Note that for a fixed data set $\boldsymbol{W}_d^n$ with ith row $\boldsymbol{w}_i$, the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) \le \max_{i=1,\ldots,n} \|\boldsymbol{x}_i - T(\boldsymbol{W}_d^n)\| \le \max_{i=1,\ldots,n} \|\boldsymbol{x}_i\| + M$ if $d_n < d_T$. Similarly, suppose $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) \le M$ for some constant M if $d_n < d_T$, then $\|T(\boldsymbol{W}_d^n)\|$ is bounded if $d_n < d_T$. □

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 cannot 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 $0 < a < \lambda_p \le \lambda_1 < b$ 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 $0 \le d_n < n\gamma_n < n/2$ 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 $p+1 \le c_n \le n$, 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

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \le D^2_{(c_n)}\} \tag{4.4}$$

where $D^2_{(c_n)}$ is the c_nth smallest squared Mahalanobis distance based on $(T, \boldsymbol{C})$. This hyperellipsoid contains the c_n cases with the smallest D_i^2. 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

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \le h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \le h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \le h\}$$

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, pp. 103–104).

4.3 The Concentration Algorithm

Concentration algorithms are widely used since impractical brand name estimators, such as the MCD estimator given in Definition 4.4, take too long to compute. The concentration algorithm, defined in Definition 4.5, use K starts and attractors. A *start* is an initial estimator, and an *attractor* is an estimator obtained by refining the start. For example, let the start be the classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$. Then the attractor could be the classical estimator $(T_1, \boldsymbol{C}_1)$ applied to the half set of cases with the smallest Mahalanobis distances. This concentration algorithm uses one concentration step, but the process could be iterated for k concentration steps, producing an estimator $(T_k, \boldsymbol{C}_k)$.

If more than one attractor is used, then some criterion is needed to select which of the K attractors is to be used in the final estimator. If each attractor $(T_{k,j}, \boldsymbol{C}_{k,j})$ is the classical estimator applied to $c_n \approx n/2$ cases, then the minimum covariance determinant (MCD) criterion is often used: choose the attractor that has the minimum value of $det(\boldsymbol{C}_{k,j})$ where $j = 1, ..., K$.

The remainder of this chapter will explain the concentration algorithm, explain why the MCD criterion is useful but can be improved, provide some theory for practical robust multivariate location and dispersion estimators, and show how the set of cases used to compute the recommended RMVN or RFCH estimator can be used to create robust multivariate analogs of methods such as principal component analysis and canonical correlation analysis. The RMVN and RFCH estimators are reweighted versions of the practical FCH estimator, given in Definition 4.8.

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 $(T_{MCD}(\boldsymbol{W}), \boldsymbol{C}_{MCD}(\boldsymbol{W}))$.

Here
$$C(n, i) = \binom{n}{i} = \frac{n!}{i! \ \ (n-i)!}$$
is the binomial coefficient.

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. See Chapter 14. The population analog of the MCD estimator is closely related to the hyperellipsoid of highest concentration that contains $c_n/n \approx$ half of the mass. The MCD estimator is a $\sqrt{n}$ consistent HB asymptotically normal estimator for $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ where a_{MCD} is some positive constant when the data $\boldsymbol{x}_i$ are iid from a large class of distributions. See Cator and Lopuhaä (2010, 2012) who extended some results of Butler et al. (1993).

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) noted 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 *concentration 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 (1999a). Using $k = 10$ concentration steps often works well. The following proposition is useful and shows that $det(\boldsymbol{S}_{0,j})$ tends to be greater than the determinant of the attractor $det(\boldsymbol{S}_{k,j})$.

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 of 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, e.g., $K = 500$, 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 $(\boldsymbol{0}, \boldsymbol{I}_p)$ and $(\boldsymbol{1}, diag(1, 3, ..., p))$ be the high breakdown attractors where $\boldsymbol{0}$ and $\boldsymbol{1}$ are the $p \times 1$ vectors of zeroes and ones. If the minimum determinant criterion is used, then the final estimator is $(\boldsymbol{0}, \boldsymbol{I}_p)$. 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 et al. (2006, pp. 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, e.g., 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: $\gamma_n \geq \min(\gamma_{n,1}, ..., \gamma_{n,K}) \rightarrow 0.5$ as $n \rightarrow \infty$.

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 $E(\boldsymbol{S}_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$, $E[\overline{\boldsymbol{x}}_{-1,j}] = E(\boldsymbol{x}) = \boldsymbol{\mu}$, and $\text{Cov}(\overline{\boldsymbol{x}}_{-1,j}) = \text{Cov}(\boldsymbol{x})/(p{+}1) = \boldsymbol{\Sigma}_{\boldsymbol{x}}/(p{+}1)$ 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 $P(\min_j \|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = P(\text{all } |\overline{\boldsymbol{x}}_{-1,\text{j}} - \boldsymbol{\mu}\| > \epsilon) \rightarrow \delta_\epsilon^{\text{K}} > 0$ as $n \rightarrow \infty$ 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 $\gamma_n \leq K/n \rightarrow 0$ as $n \rightarrow \infty$. □

Since the FMCD estimator is a zero breakdown elemental concentration algorithm, the Hubert et al. (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,

$$\gamma_o \approx \min\left(\frac{n - c_n}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h}\right) 100\% \tag{4.5}$$

if n is large, $c_n \geq n/2$ and $h = p + 1$.

Proof. Suppose that the data set contains n cases with d outliers and $n - d$ clean cases. Suppose K elemental sets are chosen with replacement. If W_i is the number of outliers in the ith elemental set, then the W_i are iid hypergeometric$(d, n - d, h)$ random variables. Suppose that it is desired to find K such that the probability P(that at least one of the elemental sets is clean) $\equiv P_1 \approx 1 - \alpha$ where $0 < \alpha < 1$. Then $P_1 = 1-$ P(none of the K elemental sets is clean) $\approx 1 - [1 - (1 - \gamma)^h]^K$ by independence. If the contamination proportion γ is fixed, then the probability of obtaining at least one clean subset of size h with high probability (say $1 - \alpha = 0.8$) is given by

$0.8 = 1 - [1 - (1-\gamma)^h]^K$. Fix the number of starts K and solve this equation for γ. $\square$

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* $(T_{-1,j}, \boldsymbol{C}_{-1,j})$. After finding $(T_{0,j}, \boldsymbol{C}_{0,j})$, 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 et al. 1975, 1981) defined below is one example. Gnanadesikan and Kettenring (1972, pp. 94–95) provided 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 (2004a) MB estimator is a 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) provided 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 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 MED$(\boldsymbol{W})$ in Euclidean distance.

The proof of the following theorem implies that a high breakdown estimator $(T, \boldsymbol{C})$ has MED$(D_i^2) \leq V$ and that the hyperellipsoid $\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \leq D_{(c_n)}^2\}$ that contains $c_n \approx n/2$ 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, e.g., $k = 10$. 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 $\tau_1 \geq \cdots \geq \tau_p$, then for any nonzero vector $\boldsymbol{x}$,

$$0 < \|\boldsymbol{x}\|^2 \ \tau_p \leq \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \leq \|\boldsymbol{x}\|^2 \ \tau_1. \tag{4.6}$$

Let $\lambda_1 \geq \cdots \geq \lambda_p$ be the eigenvalues of $\boldsymbol{C}$. By (4.6),

$$\frac{1}{\lambda_1}\|\boldsymbol{x} - T\|^2 \leq (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) \leq \frac{1}{\lambda_p}\|\boldsymbol{x} - T\|^2. \tag{4.7}$$

By (4.7), if the $D_{(i)}^2$ are the order statistics of the $D_i^2(T, \boldsymbol{C})$, then $D_{(i)}^2 < V$ 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 $1/\lambda_p$ and 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, pp. 50, 103), the boundary of the set $\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \leq h^2\} = \{\boldsymbol{x} | (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) \leq h^2\}$ is a hyperellipsoid centered at T with axes of length $2h\sqrt{\lambda_i}$. Hence $\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \leq D_{(c_n)}^2\}$ 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 $\lambda_{p,0} > \inf det(\boldsymbol{C}_{MCD})/\lambda_{1,0}^{p-1} > 0$ where the infimum 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 Definitions 3.14 and 3.15.

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 (2010, 2012) showed 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 $\text{Cov}(\boldsymbol{x}_i)$ where g is continuously differentiable with finite fourth moment: $\int (\boldsymbol{x}^T\boldsymbol{x})^2 g(\boldsymbol{x}^T\boldsymbol{x}) d\boldsymbol{x} < \infty$.

Lopuhaä (1999) showed 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 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). Assume the number of concentration steps k is fixed. 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 100ξth 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 $D_i^2(T, \boldsymbol{C}) \leq D^2_{(c_n)}(T, \boldsymbol{C})$. Let $c_n \approx n/2$ and

$$b = D^2_{0.5}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

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

$(T, \tilde{\boldsymbol{C}}) \equiv (T, D^2_{(c_n)}(T, \boldsymbol{C})\ \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$, and $D^2_i(T, \tilde{\boldsymbol{C}}) \leq 1$ is equivalent to $D^2_i(T, \boldsymbol{C}) \leq D^2_{(c_n)}(T, \boldsymbol{C})$). 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, $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi^2_p$.

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 4.2, 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 affine equivariant 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 Cator and Lopuhaä (2010, 2012). 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 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$(\boldsymbol{W})$ 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

$$\boldsymbol{C}_{FCH} = \frac{\text{MED}(D^2_i(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{4.10}$$

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 $\boldsymbol{C}_{MBA}$ are 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 $det(\boldsymbol{C}_{FCH})$ 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 $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_A) \leq det(\boldsymbol{C}_{0,M}) < \infty$ 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 Proposition 4.4 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 criterion 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 $\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,j})^T \boldsymbol{S}_{k,j}^{-1}(\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,j}) \le h^2\}$ 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, pp. 103–104). The "MVE(c_n)" criterion is $h^p\sqrt{\det(\boldsymbol{S}_{k,j})}$ 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 $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH}) \le \chi^2_{p,0.975}$, 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 $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1) \le \chi^2_{p,0.975}$, 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

$$\boldsymbol{C}_{RMVN} = \frac{\text{MED}(D_i^2(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2))}{\chi^2_{p,q_2}} \tilde{\boldsymbol{\Sigma}}_2.$$

Definition 4.11. Let the n_2 cases in Definition 4.10 be known as the *RMVN set U*. Hence $(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2) = (\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ is the classical estimator applied to the RMVN set U, which can be regarded as the untrimmed data (the data not trimmed by ellipsoidal trimming) or the cleaned data. Also $\boldsymbol{S}_U$ is the unscaled estimated dispersion matrix while $\boldsymbol{C}_{RMVN}$ is the scaled estimated dispersion matrix.

Remark 4.4. Classical methods will be applied to the RMVN subset U to make robust methods throughout this text. Under (E1), $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c_U \boldsymbol{\Sigma})$ for some constant $c_U > 0$ that depends on the underlying distribution of the iid $\boldsymbol{x}_i$. For a general estimator of multivariate location and dispersion $(T_A, \boldsymbol{C}_A)$, typically a reweight for efficiency step is performed, resulting in a set U such that the classical estimator $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ is the classical estimator applied to a set U. For example, use $U = \{\boldsymbol{x}_i | D_i^2(T_A, \boldsymbol{C}_A) \leq \chi^2_{p,0.975}\}$. Then the final estimator is $(T_F, \boldsymbol{C}_F) = (\overline{\boldsymbol{x}}_U, a\boldsymbol{S}_U)$ where scaling is done as in Equation (4.10) in an attempt to make $\boldsymbol{C}_F$ a good estimator of $\boldsymbol{\Sigma}$ if the iid data are from a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Then $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ can be shown to be a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c_U \boldsymbol{\Sigma})$ for a large class of distributions for the RMVN set, for the RFCH set, or if $(T_A, \boldsymbol{C}_A)$ is an affine equivariant $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c_A \boldsymbol{\Sigma})$ on a large class of distributions. The necessary theory is not yet available for other practical robust reweighted estimators such as OGK and Det-MCD.

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 $0.5 \le q < 1$ 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 $0 \le \gamma < 0.5$ 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 $N_2(\boldsymbol{0}, \boldsymbol{\Sigma})$ 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 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)
```

Remark 4.5. The RFCH and RMVN estimators are recommended. If these estimators are too slow and outlier detection is of interest, try the RMB estimator, the reweighted MB estimator. If RMB is too slow or if $n < 2(p+1)$, the Euclidean distances $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I})$ of $\boldsymbol{x}_i$ from the coordinatewise median MED($\boldsymbol{W}$) may be useful. A DD plot of $D_i(\overline{\boldsymbol{x}}, \boldsymbol{I})$ versus $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I})$ is also useful for outlier detection and for whether $\overline{\boldsymbol{x}}$ and MED($\boldsymbol{W}$) are giving similar estimates of multivariate location. See Section 4.7.

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 $\text{MED}(\boldsymbol{W}) = (1680, 1240)^T$, but if the distances correspond to the contours of a covering ellipsoid, then case 44 has the largest distance. For $k = 0$, $(T_{0,M}, \boldsymbol{C}_{0,M})$ 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.

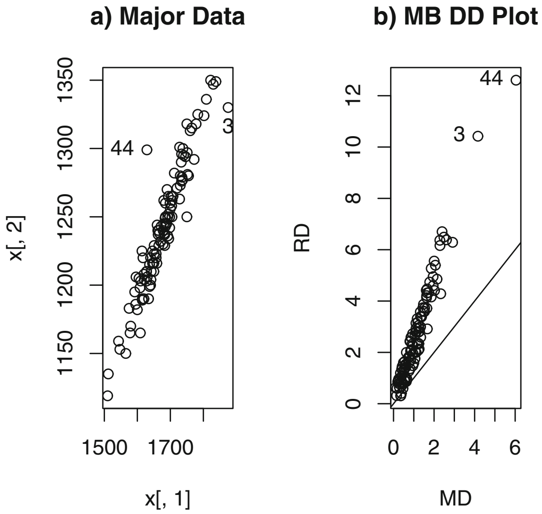

Fig. 4.1 Plots for Major Data

As the dimension p gets larger, outliers that cannot 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 or stick 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 $p = 2$.

Example 4.2. The estimators can be useful when the data is not elliptically contoured. The Gladstone (1905) 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,

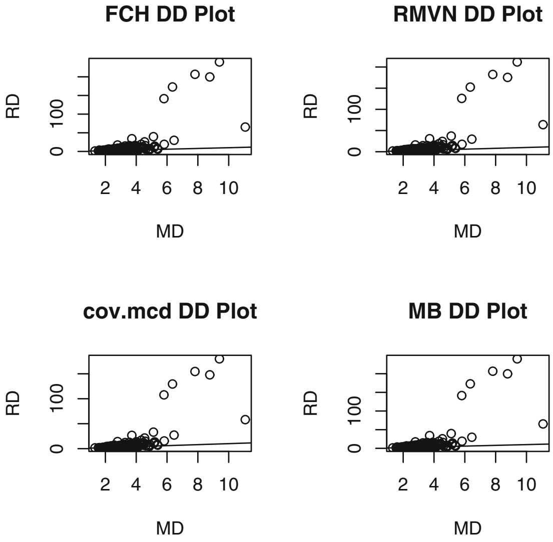

Fig. 4.2 DD Plots for Gladstone Data

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

```
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
```

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 above are the averages, using 20 runs and $n = 1000$, of the dispersion matrices when the bulk of the data are iid $N_4(\boldsymbol{0}, \boldsymbol{\Sigma})$ 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 $0.94\boldsymbol{\Sigma}$.

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 0.013
-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 55.65
```

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_4(15\ \boldsymbol{1}, \boldsymbol{\Sigma})$, 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^2_W is the sample variance of the W_{in}, then $E(nS^2_W) = \tau^2$ and $V(nS^2_W) = 2\tau^4/(r-1)$. So $nS^2_W \pm \sqrt{5}SE(nS^2_W) \approx \tau^2 \pm \sqrt{10}\tau^2/\sqrt{r-1}$. So for $r = 1000$ runs, we expect nS^2_W 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, e.g., $n^{1/3}$ rate, then the scaled sample variance $nS^2_W \to \infty$ as $n \to \infty$.

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

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 $\boldsymbol{x} \sim N_p(\boldsymbol{0}, diag(1, ..., p))$. 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 a hyperellipsoidal 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 three data sets the robust distances did not achieve complete separation. In Spring 2015, Det-MCD simulated much like FMCD, but was more likely to cause an error in R.

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 $(0, ..., 0, pm/\sqrt{p})^T$.

For near point mass outliers, a hyperellipsoid with very small volume can cover half of the data if the outliers are at one end of the hyperellipsoid 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

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

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).

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

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, and 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.

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 $c\boldsymbol{\Sigma} = c\ diag(1, 2)$.

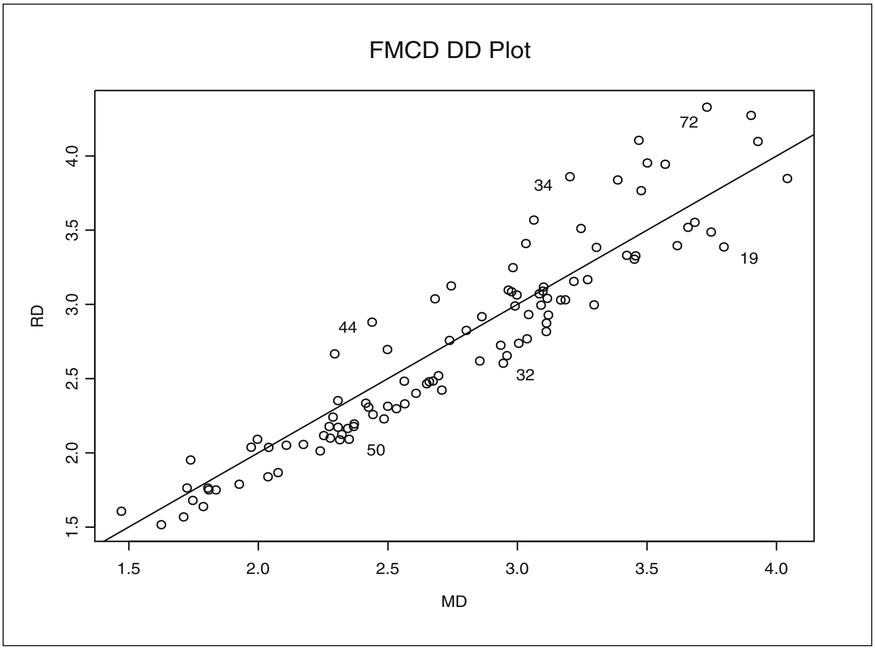

Fig. 4.3 The FMCD Estimator Failed

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 $\boldsymbol{C}_{MBA} \approx 2\boldsymbol{C}_{FMCD} \approx 2\boldsymbol{\Sigma}$.

In an older mean shift simulation, when p was 8 or larger, 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 hyperellipsoid while the outliers fall in another hugely distant compact hyperellipsoid, then a concentration algorithm using a single start can sometimes tolerate nearly 25% outliers.

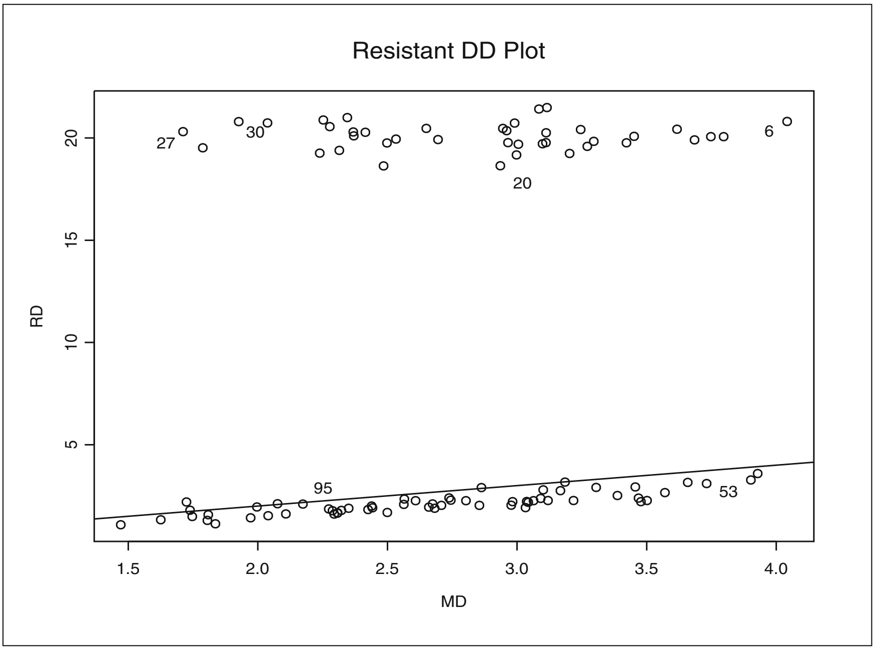

Fig. 4.4 The Outliers are Large in the MBA DD Plot

For example, suppose 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 cannot 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 $0 \leq \det(\boldsymbol{C}_{MCD}) \leq \det(\boldsymbol{C}_A)$. □

Software

The `robustbase` library was downloaded from (www.r-project.org/#doc). $\oint$ 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. To use Det-MCD instead of FMCD, change

```
out <- covMcd(x)  to out <- covMcd(x,nsamp=``deterministic''),
```

but in Spring 2015, this change was more likely to cause errors.

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, MB estimators, and the `covmb2` estimator of Section 4.7. 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.

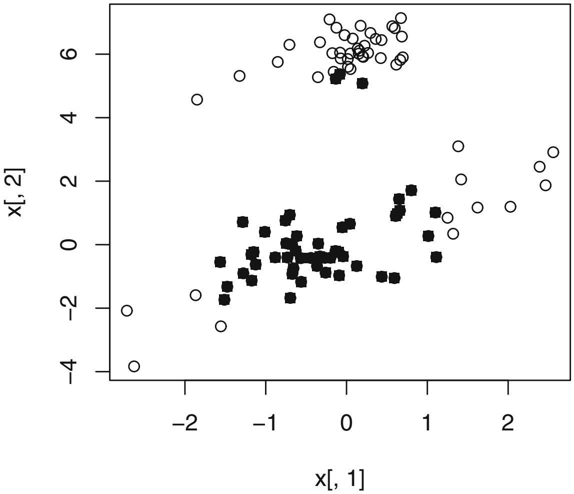

Fig. 4.5 highlighted cases = half set with smallest RD = $(T_0, \boldsymbol{C}_0)$

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. **Better programs would use MB if DGK causes an error.**

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 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

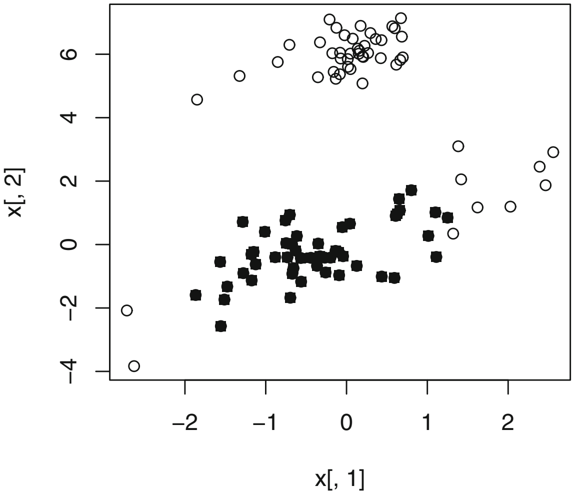

Fig. 4.6 highlighted cases = half set with smallest RD = $(T_1, \boldsymbol{C}_1)$

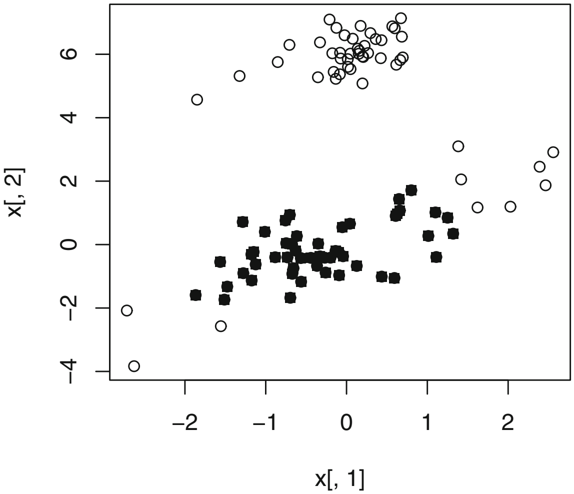

Fig. 4.7 highlighted cases = half set with smallest RD = $(T_2, \boldsymbol{C}_2)$

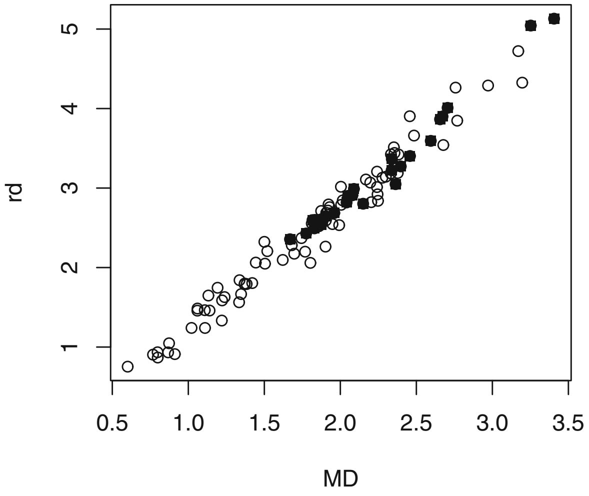

Fig. 4.8 highlighted cases = outliers, RD = $(T_{0,D}, \boldsymbol{C}_{0,D})$

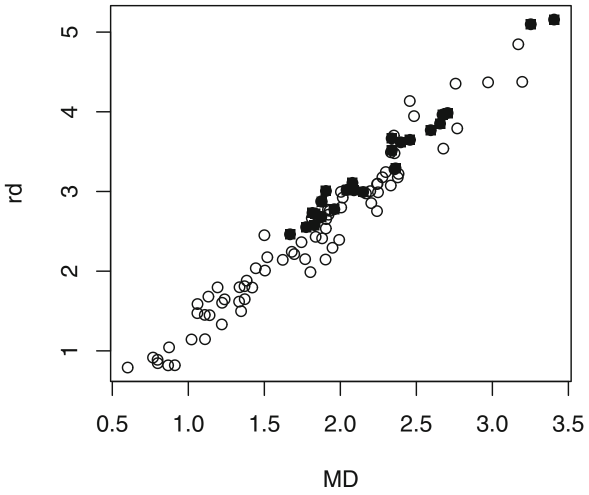

Fig. 4.9 highlighted cases = outliers, RD = $(T_{1,D}, \boldsymbol{C}_{1,D})$

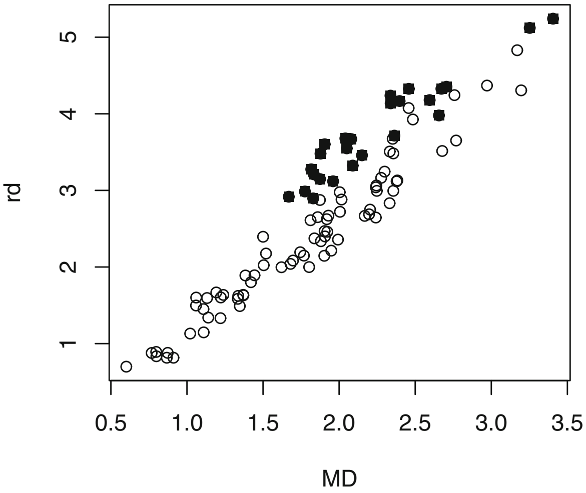

Fig. 4.10 highlighted cases = outliers, RD = $(T_{2,D}, \boldsymbol{C}_{2,D})$

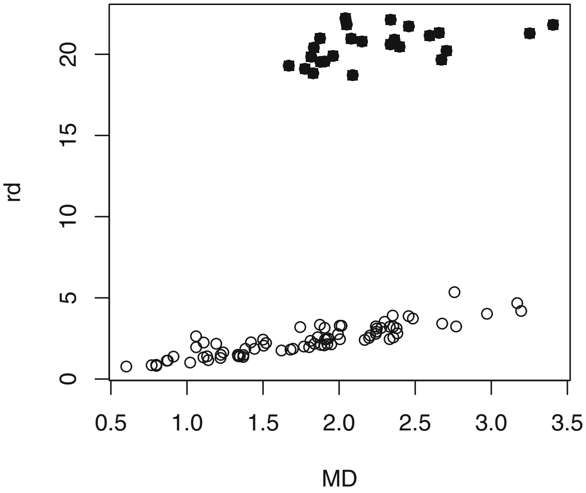

Fig. 4.11 highlighted cases = outliers, RD = $(T_{3,D}, \boldsymbol{C}_{3,D})$

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 The RMVN and RFCH Sets

Both the RMVN and RFCH estimators compute the classical estimator $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ on some set U containing $n_U \geq n/2$ of the cases. Referring to Definition 4.9, for the RFCH estimator, $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U) = (T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2)$, and then $\boldsymbol{S}_U$ is scaled to form $\boldsymbol{C}_{RFCH}$. Referring to Definition 4.10, for the RMVN estimator, $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U) = (T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2)$, and then $\boldsymbol{S}_U$ is scaled to form $\boldsymbol{C}_{RMVN}$. See Definition 4.11.

The two main ways to handle outliers are i) apply the multivariate method to the cleaned data and ii) plug in robust estimators for classical estimators. Subjectively cleaned data may work well for a single data set, but we can't get large sample theory since sometimes too many cases are deleted (delete outliers and some nonoutliers) and sometimes too few (do not get all of the outliers). Practical plug in robust estimators have rarely been shown to be $\sqrt{n}$ consistent and highly outlier resistant.

Using the RMVN or RFCH, set U is simultaneously a plug in method and an objective way to clean the data such that the resulting robust method is often backed by theory. This result is extremely useful computationally: find the RMVN set or RFCH set U, then apply the classical method to the cases in the set U. This procedure is often equivalent to using $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ as plug in estimators. The method can be applied if $n > 2(p+1)$ but may not work well unless $n > 20p$. The *mpack* function `getu` gets the RMVN set U as well as the case numbers corresponding to the cases in U.

The set U is a small volume hyperellipsoid containing at least half of the cases since concentration is used. The set U can also be regarded as the "untrimmed data": the data that was not trimmed by ellipsoidal trimming. Theory has been proved for a large class of elliptically contoured distributions, but it is conjectured that theory holds for a much wider class of distributions. See Conjectures 4.2 and 4.3 in Section 4.9. In simulations, RFCH and RMVN 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 $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{x}) = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T$. 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 is affine equivariant, and RFCH and RMVN are asymptotically equivalent to a scaled DGK estimator. But in the simulations the results also held for skewed distributions.

In this text, usually the RMVN set U will be used. Several illustrative applications are given next, where the theory usually assumes that the cases are iid from a large class of elliptically contoured distributions. There are many other "robust methods" in the literature that use plug in estimators like FMCD. Replacing the plug in estimator by RMVN or RFCH will often greatly improve the robust method. See Chapter 14, and in Section 5.2, note that the prediction regions using RMVN performed much like the prediction regions using $(\overline{\boldsymbol{x}}, \boldsymbol{S})$.

i) The classical estimator of multivariate location and dispersion applied to the cases in U gives $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$, a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ for some constant $c > 0$. See Remark 4.4.

ii) The classical estimator of the correlation matrix applied to the cases in U gives $\boldsymbol{R}_U$, a consistent estimator of the population correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}}$.

iii) For principal component analysis (PCA), RPCA is the classical PCA method applied to the set U. See Theorems 6.1, 6.2, 6.3, and Section 6.2.

iv) For canonical correlation analysis (CCA), RCCA is the classical CCA method applied to the set U. See Theorem 7.1 and Section 7.2.

v) Let U_i be the RMVN or RFCH subset applied to the n_i cases from group i for $i = 1, ..., G$. Let $(\overline{\boldsymbol{x}}_{U_i}, \boldsymbol{S}_{U_i})$ be the sample mean and covariance applied to the cases in U_i. Let $Y = i$ for cases in U_i which are from group i. Let $U_{big} = U_1 \cup U_2 \cup \cdots \cup U_G$ be the combined sample. Then apply the discriminant analysis method to U_{big} with the corresponding labels Y. For example, RFDA consists of applying classical FDA on U_{big}. See Section 8.9.

vi) For factor analysis, apply the factor analysis method to the set U. This method can be used as a diagnostic for methods such as the maximum likelihood method of factor analysis, but is backed by theory for principal component factor analysis. See Section 11.2.

vii) For multiple linear regression, let Y be the response variable, $x_1 = 1$ and $x_2, ..., x_p$ be the predictor variables. Let $\boldsymbol{z}_i = (Y_i, x_{i2}, ..., x_{ip})^T$. Let U be the RMVN or RFCH set formed using the $\boldsymbol{z}_i$. Then a classical regression estimator applied to the set U results in a robust regression estimator. For least squares, this is implemented with the *mpack* function `rmreg2`.

viii) For multivariate linear regression, let $Y_1, ..., Y_m$ be the response variables, $x_1 = 1$ and $x_2, ..., x_p$ be the predictor variables. Let

$$\boldsymbol{z}_i = (Y_{i1}, ...Y_{im}, x_{i2}, ..., x_{ip})^T.$$

Let U be the RMVN or RFCH set formed using the $\boldsymbol{z}_i$. Then a classical least squares multivariate linear regression estimator applied to the set U results in a robust multivariate linear regression estimator. For least squares, this is implemented with the *mpack* function `rmreg2`. The method for multiple linear regression in vii) corresponds to $m = 1$. See Section 12.6.2.

There are also several variants on the method. Suppose there are tentative predictors $Z_1, ..., Z_J$. After transformations assume that predictors $X_1, ..., X_k$ are linearly related. Assume the set U used cases $i_1, i_2, ..., i_{n_U}$. To add variables like $X_{k+1} = X_1^2$, $X_{k+2} = X_3 X_4$, $X_{k+3} = gender$, ..., X_p, augment U with the variables $X_{k+1}, ..., X_p$ corresponding to cases $i_1, ..., i_{n_U}$. Adding variables results in cleaned data that is more likely to contain outliers.

If there are g groups ($g = G$ for discriminant analysis, $g = 2$ for binary regression, and $g = p$ for one way MANOVA), the function `getubig` gets the RMVN set U_i for each group and combines the g RMVN sets into one large set $U_{big} = U_1 \cup U_2 \cup \cdots \cup U_g$.

4.7 What if $p > n$?

Most of the methods in this text work best if $n \geq 20p$, but often data sets have $p > n$, and outliers are a major problem. Usually the population covariance matrix is singular when $p > n$, but the sample covariance matrix, MB estimator with no concentration steps, and the sign covariance matrix can be computed. Weighted versions of the last two estimators are useful since concentration steps need nonsingular dispersion matrices. Compute the squared Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median $D_i^2 = D_i^2(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median MED_j: the coordinatewise median computed from the cases $\boldsymbol{x}_i$ with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \boldsymbol{I}_p))$ where $\text{MED}_0 = \text{MED}(\boldsymbol{W})$. We often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, ..., D_n) + k\text{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise. This weighting corresponds to the weighting that would be used in a one sided metrically trimmed mean (Huber type skipped mean) of the distances. Just as classical methods can be applied to the set RMVN set U to create a robust estimator, classical methods for the $p > n$ case, such as PLS (partial least squares), can be applied to the `covmb2` set B defined below.

Definition 4.12. Let the *covmb2 set* B of at least $n/2$ cases correspond to the cases with weight 1. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set B. Hence

$$T = \frac{\sum_{i=1}^n W_i \boldsymbol{x}_i}{\sum_{i=1}^n W_i} \quad \text{and } \boldsymbol{C} = \frac{\sum_{i=1}^n W_i (\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^n W_i \ - \ 1}.$$

The `covmb2` estimator attempts to give a robust dispersion estimator that reduces the bias by using a big ball about the coordinatewise median instead of a ball that contains half of the cases. The weighting is the default method, but you can also plot the squared Euclidean distances and estimate the number $m \geq n/2$ of cases with the smallest distances to be used. The *mpack* function `medout` makes the plot, and the *mpack* function `getB` gives the set B of cases that got weight 1 along with the index `indx` of the case numbers that got weight 1. The function `vecw` stacks the columns of the dispersion matrix $\boldsymbol{C}$ into a vector. Then the elements of the matrix can be plotted.

The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically, the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers. An alternative for outlier detection is to replace $\boldsymbol{C}$ by $\boldsymbol{C}_d = diag(\hat{\sigma}_{11}, ..., \hat{\sigma}_{pp})$. For example, use $\hat{\sigma}_{ii} = \boldsymbol{C}_{ii}$. See Ro et al. (2015) and Tarr et al. (2016) for references.

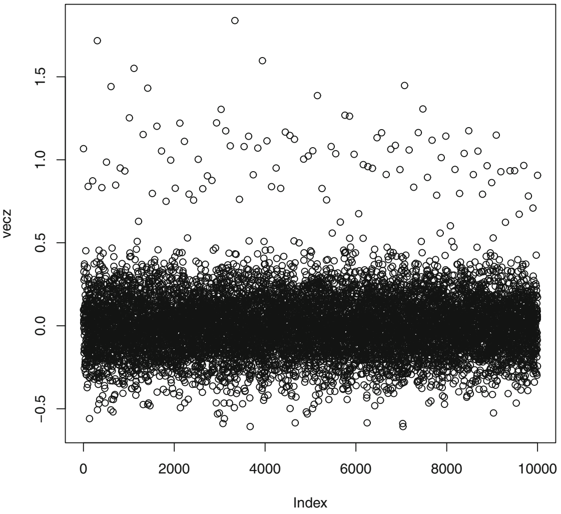

Fig. 4.12 Elements of $\boldsymbol{C}$ for outlier data

Example 4.3. This example helps illustrate the effect of outliers on classical methods. The artificial data set had $n = 50, p = 100$, and the clean data was iid $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Hence the diagonal elements of the population covariance matrix are 0, and the diagonal elements are 1. Plots of the elements of the sample covariance matrix S and the `covmb2` estimator $\boldsymbol{C}$ are not shown, but were similar to Figure 4.12. Then the first ten cases were contaminated: $\boldsymbol{x}_i \sim N_p(\boldsymbol{\mu}, 100\boldsymbol{I}_p)$ where $\boldsymbol{\mu} = (10, 0, ..., 0)^T$. Figure 4.12 shows that the `covmb2` dispersion matrix $\boldsymbol{C}$ was not much effected by the outliers. The diagonal elements are near 1 and the off-diagonal elements are near 0. Figure 4.13 shows that the sample covariance matrix $\boldsymbol{S}$ was greatly effected by the outliers. Several sample covariances are less than -20, and several sample variances are over 40.

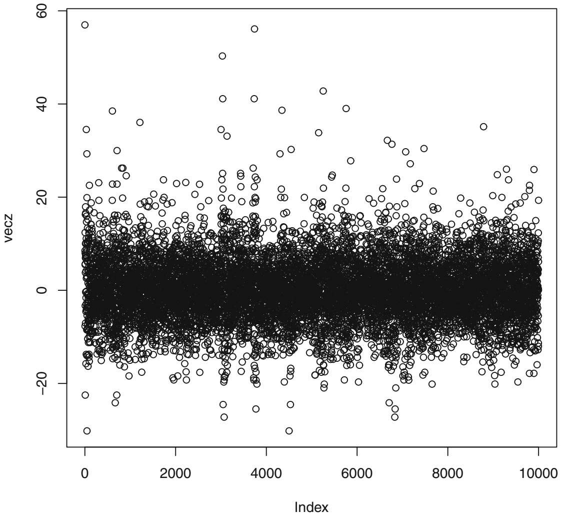

Fig. 4.13 Elements of the classical covariance matrix $\boldsymbol{S}$ for outlier data

R code to used to produce Figures 4.12 and 4.13 is shown below.

```
#n = 50, p = 100
x<-matrix(rnorm(5000),nrow=50,ncol=100)
out<-medout(x) #no outliers, try ddplot5(x)
out <- covmb2(x,msteps=0)
z<-out$cov
plot(diag(z))  #plot the diagonal elements of C
plot(out$center) #plot the elements of  T
vecz <- vecw(z)$vecz
plot(vecz)

out<-covmb2(x,m=45)
plot(out$center)
plot(diag(out$cov))

#outliers
x[1:10,] <- 10*x[1:10,]
x[1:10,1] <- x[1:10]+10
medout(x) #The 10 outliers are easily detected in
#the plot of the distances from the MED(X).
ddplot5(x) #two widely separated clusters of data
tem <- getB(x,msteps=0)
tem$indx #all 40 clean cases were used
dim(tem$B) #40 by 100
out<-covmb2(x,msteps=0)
```

```
z<-out$cov
plot(diag(z))
plot(out$center)
vecz <- vecw(z)$vecz
plot(vecz) #plot the elements of C
#Figure 4.12

#examine the sample covariance matrix and mean
plot(diag(var(x)))
plot(apply(x,2,mean)) #plot elements of xbar
zc <- var(x)
vecz <- vecw(zc)$vecz
plot(vecz) #plot the elements of S
#Figure 4.13

out<-medout(x) #10 outliers
out<-covmb2(x,m=40)
plot(out$center)
plot(diag(out$cov))
```

The `covmb2` estimator can also be used for $n > p$. The *mpack* function `mldsim6` suggests that for 40% outliers, the outliers need to be further away from the bulk of the data (`covmb2(k=5)` needs a larger value of *pm*) than for the other six estimators if $n \geq 20p$. With outlier types like those in Tables 4.4 and 4.5, `covmb2(k=5)` was often near best. Try the following commands. The other estimators need $n > 2p$, and as n gets close to $2p$, `covmb2` may outperform the other estimators.

```
#near point mass on major axis
mldsim6(n=100,p=10,outliers=1,gam=0.25,pm=25)
mldsim6(n=100,p=10,outliers=1,gam=0.4,pm=25) #bad
mldsim6(n=100,p=40,outliers=1,gam=0.1,pm=100)
mldsim6(n=200,p=60,outliers=1,gam=0.1,pm=100)
#mean shift outliers
mldsim6(n=100,p=40,outliers=3,gam=0.1,pm=10)
mldsim6(n=100,p=40,outliers=3,gam=0.25,pm=20)
mldsim6(n=200,p=60,outliers=3,gam=0.1,pm=10)
#concentration steps can help
mldsim6(n=100,p=10,outliers=3,gam=0.4,pm=10,osteps=0)
mldsim6(n=100,p=10,outliers=3,gam=0.4,pm=10,osteps=9)
```

The following estimator can also be used if $p > n$, but improvements are needed, and suggested in the following paragraphs. For other estimators that can be used when $p > n$, see Boudt et al. (2017).

Definition 4.13. The Sign Covariance Matrix

$$\hat{\boldsymbol{\Sigma}}_S = \frac{1}{n}\sum_{i=1}^{n} \frac{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n)^T}{\|\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n\|^2} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^T.$$

This estimator is similar to the classical covariance estimator computed from $\boldsymbol{z}_i = \dfrac{\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n}{\|\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n\|}$, assuming $\overline{\boldsymbol{z}}_i \approx \boldsymbol{0}$. Note that $\|\boldsymbol{z}_i\| = 1$, so the $\boldsymbol{z}_i$ lie on the unit hypersphere centered at $\hat{\boldsymbol{\mu}}_n$. Here $\hat{\boldsymbol{\mu}}_n$ is the L_1-median or spatial median, defined as

$$\hat{\boldsymbol{\mu}}_n = \text{argmin}_{\boldsymbol{\mu}} \quad \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|,$$

and $\hat{\boldsymbol{\mu}}_n$ is a fairly practical high breakdown estimator of multivariate location. An argument similar to the proof of Lemma 4.3 can be used to show that the maximum eigenvalue of $\hat{\boldsymbol{\Sigma}}_S$ is bounded.

Draw a circle and then ellipsoidal data centered at the center of the circle. Project the data on the circle. Unless the data is spherical, the projection severely distorts the shape of the data. Hence $\hat{\boldsymbol{\Sigma}}_S$ is not a consistent estimator of $c\boldsymbol{\Sigma}$ for nonspherical elliptically contoured data. Suppose $p = 2$ and the highest density region is a ellipse with the x-axis as the major axis and the y-axis as the minor axis. Assume the data projected on the x-axis has standard deviation 3, and the data projected on the y-axis has standard deviation 0.3. Then we expect that $\hat{\boldsymbol{\Sigma}}_S$ will underestimate the variability about the major axis and overestimate the variability about the minor axis.

Suppose p is fixed and $n \to \infty$. Croux et al. (2010) showed that the Sign Covariance Matrix is a high breakdown estimator, and the minimum eigenvalue can be driven to zero if more than half of the cases are outliers. They also claim, that under regularity conditions, that for clean data, the Sign Covariance Matrix consistently estimates the orientation of the dispersion matrix: for a class of elliptically contoured distributions, the eigenvectors $\hat{\boldsymbol{e}}_i$ estimate the population eigenvectors $\boldsymbol{e}_i$.

Remark 4.6. Suppose the $\hat{\boldsymbol{e}}_i$ estimate the dispersion matrix orientation, but the eigenvalues are severely biased. Let $L_i = \text{MAD}(\hat{\boldsymbol{e}}_i^T \boldsymbol{x}_1, ..., \hat{\boldsymbol{e}}_i^T \boldsymbol{x}_n)$ for $i = 1, ..., p$. Note that $\hat{\boldsymbol{e}}_i^T \boldsymbol{x}_1, ..., \hat{\boldsymbol{e}}_i^T \boldsymbol{x}_n$ are the data projected onto the line in the direction of the ith orthonormal eigenvector $\hat{\boldsymbol{e}}_i$. Let $\hat{\lambda}_1 = L_{(p)} \geq \hat{\lambda}_2 = L_{(p-1)} \geq \cdots \geq \hat{\lambda}_{p-1} = L_{(2)} \geq \hat{\lambda}_p = L_{(1)}$. If necessary, relabel the $\hat{\boldsymbol{e}}_i$ so that $\hat{\lambda}_i$ corresponds to eigenvector $\hat{\boldsymbol{e}}_i$. Then let $\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^p \hat{\lambda}_i \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^T$. (Compare Theorem 2.1.) Then on a class of elliptically contoured distributions, $\hat{\boldsymbol{\Sigma}}$ may be a better estimator of $c\boldsymbol{\Sigma}$ for some $c > 0$.

Outliers can still severely bias the estimator since they have the same weight as the clean data. For example, draw a circle centered at the origin. Suppose the data $\boldsymbol{x}_i$ is tightly clustered about the horizontal axis. Then the projected data $\boldsymbol{z}_i$ tightly clusters about the points $(-1, 0)$ and (1,0). Placing outliers with $\boldsymbol{z}_i$ at (0,1) or $(0, -1)$ will inflate the minor axis.

To examine breakdown, suppose that more than half of the $\boldsymbol{x}_i = \boldsymbol{x}_o$, so there is a point mass at $\boldsymbol{x}_o$. Because of the exact fit property, the spatial median is equal to $\boldsymbol{x}_o$. Draw a cloud for the clean data and put $\boldsymbol{x}_o$ far to the right of the cloud. Draw a unit circle around $\boldsymbol{x}_o$ and a left opening cone with vertex at $\boldsymbol{x}_o$ that just contains the data cloud. Then the clean data gets projected on an arc where the cone intersects the left of the circle. As $\boldsymbol{x}_o$ is moved further to the right, the arc length goes to zero, so the projected data form a near point mass. Hence the Sign Covariance Matrix is roughly the covariance matrix of a point mass (it is not clear what $\boldsymbol{z}_i$ is if $\boldsymbol{x}_i = \hat{\boldsymbol{\mu}}_n$) and a near point mass, which will have smallest eigenvalue → 0 as $\boldsymbol{x}_0$ gets further away from the clean data cloud.

To examine the effect of outliers on the eigenvectors, assume the proportion of outliers $\boldsymbol{x}_o$ is 40%. Then the spatial median gets dragged to the right of the data cloud. Draw a unit circle around the spherical median and project the data onto the unit circle. Most of the data is projected onto the left half of the circle, so the eigenvectors $\hat{\boldsymbol{e}}_i$ no longer estimate the orientation of the dispersion matrix.

A simple solution is to compute the spatial median on the `covmb2` set B, perhaps using the spatial median instead of the coordinatewise median to form the D_i, but it seems that applying the classical estimator on the data in set B makes more sense.

The median ball estimator and `covmb2` put a hypersphere, centered at the coordinatewise median, about the data and gives zero weight to data outside the hypersphere. The Sign Covariance Matrix projects the data onto a unit hypersphere centered at the spherical median. We conjecture that the median ball and `covbm2` estimators also estimate the orientation of the dispersion matrix for a class of elliptically contoured distributions. The reweighted median ball estimator with five concentration steps seems to estimate the $c\boldsymbol{\Sigma}$ with very small bias for many elliptically contoured distributions when $n >> p$. When $p > n$ and `covmb2` is used, using Remark 4.6 to find the L_i and $\hat{\lambda}_i$ may be useful.

4.8 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 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.

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 (e.g., some plotted points are far from the bulk of the data or the plotted points follow two lines).

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 $E(\overline{\boldsymbol{x}}_j) = E(\boldsymbol{x}_i)$, $\text{Cov}(\overline{\boldsymbol{x}}_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}}/(p + 1)$, and $E(\boldsymbol{S}_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$.

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 $(\overline{\boldsymbol{z}}, \boldsymbol{R})$, we want $n \geq 10p$. We want $n \geq 20p$ for FCH, RFCH, or RMVN.

10) Brand-name robust MLD estimators from the dominant robust statistics 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.

4.9 Complements

Most of this chapter focused on robust estimators where $n \geq 10p$. Dispersion estimators for $n < 10p$ are discussed in Section 4.7, Pourahmadi (2013), and Yao et al. (2015). Tiao and Tsay (1983) gave an interesting bound on the determinant.

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 $\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu}_{t-1})^T \boldsymbol{\Sigma}^{-1}_{t-1}(\boldsymbol{x} - \boldsymbol{\mu}_{t-1}) \leq D^2_{0.5}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})\}$ 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 $(\overline{\boldsymbol{x}}, \boldsymbol{S}) = (T_{-1,D}, \boldsymbol{C}_{-1,D})$ as the only start. Thus $(\boldsymbol{\mu}_{-1,D}, \boldsymbol{\Sigma}_{-1,D})$ is the population mean and covariance matrix. For a large class of elliptically contoured distributions 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 et al. (2009) found 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 $(T_{-1,M}, \boldsymbol{C}_{-1,M}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ 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 MED($\boldsymbol{W}$) 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, $\boldsymbol{\mu}_{k,M} = \boldsymbol{\mu}$.

The covmb2 estimator is also a natural estimator of a population mean and covariance matrix $(\boldsymbol{\mu}_{j,B}, \boldsymbol{\Sigma}_{j,B})$ corresponding to a truncated distribution, and we conjecture that covmb2 is $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}_{j,B}, \boldsymbol{\Sigma}_{j,B})$ under mild conditions.

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 $(\overline{\boldsymbol{x}}_{0,M}, \boldsymbol{S}_{0,M}) = (T_{0,M}, \boldsymbol{C}_{0,M})$ is a high breakdown estimator. If the data distribution is EC but not spherical about $\boldsymbol{\mu}$, then for $k \geq 0$, $\boldsymbol{S}_{k,M} = \boldsymbol{C}_{MB}$ underestimates the major axis and overestimates 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 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 MED($\boldsymbol{W}$) is the sample median of the jth predictor. Hence with $m-1$ outliers, 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

$$B^2 = \max_{i \in \mathcal{C}} \|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|^2 \leq \sum_{i=1}^{p} R_i^2 \leq p(\max R_i^2).$$

One of the most effective methods for detecting outliers for large data sets or if $p > n$ is to use $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. See the *mpack* function `medout`. Section 4.7 suggests more useful techniques.

The MB estimator has outlier resistance similar to $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ for distant outliers but, as shown in Example 4.1, can be much more effective for detecting certain types of outliers that cannot 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 with continuous decreasing g) 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 hyperellipsoid will have approximately the same center as the hypersphere, and the hypersphere will be drawn toward 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) suggested an estimator similar to the MB estimator, also see Croux and Van Aelst (2002). Another estimator similar to MB was suggested by Wilk et al. (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 if g is continuous and decreasing, then 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 Olive (2004a). The computational and theoretical simplicity of the FCH estimator makes it interesting. 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.

The recommended estimators are the RFCH and RMVN estimators. These two estimators are about an order of magnitude faster than most alternative robust estimators, but take slightly longer to compute than the FCH estimator, and may have slightly less resistance to outliers. RFCH and RMVN appear to have much less variability than FCH when there are no outliers. These three estimators are in Zhang et al. (2012). Also see Zhang and Olive (2009).

In addition to concentration and randomly selecting elemental sets, three other algorithm techniques are important. He and Wang (1996) suggested 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 $\mathrm{VAR(X)} = \sigma^2(\mathrm{X})$ suggests that a robust estimator of dispersion can be created by replacing the sample standard deviation $\hat{\sigma}$ by a robust estimator of scale. (This idea seems best if the outlying cases have outliers that can be detected if projected on two of the p coordinate axes. It is possible to have outliers that do not appear in one- or two-dimensional projections. Hence the method may not perform well compared to methods that use Mahalanobis or Euclidean distances like FCH and `covmb2`.) Maronna and Zamar (2002) modified this idea to create a fairly fast (possibly high breakdown consistent) OGK estimator of multivariate location and dispersion. Also see Alqallaf et al. (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 et al. (2000) proposed 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.

There certainly exist types of outlier configurations where the FMCD estimator outperforms the robust RFCH estimator. The RFCH estimator is vulnerable to certain types of outliers that lie inside the hypersphere based on the median Euclidean distance from the coordinatewise median. 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, RFCH, 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 et al. (2002) 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).

Rocke and Woodruff (1996, p. 1050) claimed 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 et al. (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 (2007b, 2008), Rousseeuw and Van Driessen (1999), Rousseeuw and van Zomeren (1990), Ruppert (1992), and Woodruff and Rocke (1993). Rousseeuw and Leroy (1987, $\oint$ 7.1) also described many methods.

The discussion by Rocke and Woodruff (2001) and by Hubert (2001) of Peña and Prieto (2001) stressed the fact that no one estimator can dominate all others for every outlier configuration. These papers and Wisnowski et al. (2002) gave 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, pp. 214, 233) noted 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 $(T_{FMCD}, \boldsymbol{C}_{FMCD})$ 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) pointed out that the FMVE is not the minimum volume ellipsoid (MVE) estimator.

Where the Dominant Robust Statistics 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, and constrained M estimators. For some depth estimators, like the Stahel–Donoho estimator, the exact algorithm of Liu and Zuo (2014) appears to take too long if $p \geq 6$ and $n \geq 100$, and simulations may need $p \leq 3$.

It is possible to compute the MCD and MVE estimators for $p = 4$ and $n = 100$ in a few hours using branch and bound algorithms (like estimators with $O(100^4)$ complexity). See Agulló (1996, 1998) and Pesch (1999). These algorithms take too long if both $p \geq 5$ and $n \geq 100$. Simulations may need $p \leq 2$. Two-stage estimators such as the MM estimator, that need an initial high breakdown consistent estimator, take longer to compute than the initial estimator. See Maronna et al. (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 et al. (2009) simulated the OGK and the Olive (2004a) median ball algorithm (MBA) estimators for $p = 100$ and n up to 50000 and noted 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 is roughly 100 to 200 times faster.

ii) No practical useful "high breakdown" estimator of multivariate location and dispersion (MLD) from this paradigm has been shown to be both consistent and 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 MLD estimator has not been shown to be both high breakdown and consistent.** Also Hawkins and Olive (2002) showed that elemental concentration estimators using K starts are zero breakdown estimators, and these estimators are inconsistent if they use k concentration steps where k is fixed.

The **main competitors** for the Olive and Hawkins (2010) multivariate location and dispersion FCH, RFCH, and RMVN estimators are the Maronna and Zamar (2002) *OGK estimator*, the Hubert et al. (2012) *Det-MCD estimator* which have not been proven to be consistent or positive breakdown, and the *Sign Covariance Matrix* shown to be high breakdown by Croux et al. (2010). Also see Taskinen et al. (2012). Croux et al. (2010) showed that the practical Sign Covariance Matrix and k-step Spatial Sign Covariance Matrix are high breakdown. They claimed that under regularity conditions, these two estimators consistently estimate the orientation of the scatter matrix. Sections 4.5 and 4.7 suggest that a tight cluster of outliers severely bias these competing estimators, unless the outlier proportion is small.

Papers with titles like Rousseeuw and Van Driessen (1999) "A Fast Algorithm for the Minimum Covariance Determinant Estimator" and Hubert et al. (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.

iii) Many papers give theory for an impractical estimator such as MCD, then replace the estimator by a zero breakdown practical estimator such as FMCD.

It may be reasonable to call an estimator a brand-name high breakdown estimator (e.g., S estimator or MCD estimator) if a) the estimator is high breakdown and b) the estimator has the same asymptotic distribution as the brand-name estimator.

Since the brand-name estimators are impractical to compute, practical algorithm estimators that use a fixed number of trial fits are used. If a) and b) have not been proven for the practical estimator, call the practical estimator an F-brand-name estimator (e.g., FS or FMCD), where F denotes that the criterion of the brand-name estimator was used to select a trial fit to be used in the F-brand-name estimator from a fixed number of trial fits.

The F-brand-name estimators can have a wide variety of theory. Suppose the final estimator is one of the trial fits. a) If K elemental sets are used as the trial fits then the final estimator is inconsistent and zero breakdown by Theorem P.1 in the preface. b) If there is only one trial fit, computing the criterion does not change the trial fit. Then the final estimator is the trial fit. So if the classical estimator is the trial fit, then the classical estimator is the final estimator. c) If K trial fits are used and each trial fit is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$, then the final estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$.

A necessary condition for a practical estimator to be the global minimizer of a brand-name criterion is that the practical estimator has a criterion value at least as small as any other estimator used. So if the FMCD estimator is the MCD estimator, $det(\boldsymbol{C}_{FMCD}) \leq det(\boldsymbol{C}_A)$ for any other estimator $\boldsymbol{C}_A$ that is the sample covariance matrix applied to the same number of cases c as the FMCD estimator. Hubert et al. (2012) claimed the Fast-MCD is the MCD estimator, but sometimes Fast-MCD has the smallest determinant and sometimes Det-MCD has the smallest determinant. This result proves that neither Fast-MCD nor Det-MCD is computing MCD. Similarly, permuting the data should not change the criterion if the global minimizer is being computed. See Problem 4.9.

To get theory for Det-MCD, we need breakdown results when the concentration steps are iterated to convergence, and we need a result like Theorem 4.9 when the initial estimator is consistent but not necessarily affine equivariant, and the initial estimator is iterated to convergence.

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, pp. 200, 283) noted 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 warning at 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.

The major algorithm producers are now using intelligently chosen starts or trial fits, such as the classical estimator, in their F-brand-name estimators. These estimators are typically not affine equivariant. See Hawkins and Olive (2002), Olive (2004a), Olive and Hawkins (2010), Hubert et al. (2012), and Maronna and Yohai (2015).

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 maliciously placed cases. In particular, the breakdown value of an estimator tells nothing about more important properties such as consistency or asymptotic normality.

4.10 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

R 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, e.g., *covmba*, will display the code for the function. Use the args command, e.g., *args(covmba)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/mrsashw.txt) into *R*.

4.1. a) Download the maha function that creates the classical Mahalanobis distances.

b) Copy and paste the commands for this problem and check whether observations 1–40 look like outliers.

4.2. Download the rmaha function that creates the robust Mahalanobis distances using cov.mcd (FMCD). Obtain outx2 as in Problem 4.1b). Enter the *R* 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 the *R* command *library(MASS)*), 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) 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 robust estimators cannot 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 four 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 four 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. You need to right click *Stop* on the plot 7 times.

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 $\sim N_p(0, diag(1, ..., p))$.

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 an 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) 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

This chapter examines the DD plot of classical versus robust Mahalanobis distances, and develops practical prediction regions for a future test observation $\boldsymbol{x}_f$ that work even if the iid training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ come from an unknown distribution. The prediction regions can be visualized with the DD plot. The classical prediction region assumes that the data are iid from a multivariate normal distribution, and the region tends to have too small of a volume if the MVN assumption is violated. The undercoverage of the volume of the classical region becomes worse as the number of variables p increases since the volume of the region $\{\boldsymbol{x} : D_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq h\} \propto h^p$. The classical region uses $h_c = \sqrt{\chi^2_{p,1-\delta}}$, which tends to be much smaller than the value of h that gives correct coverage.

A relationship between confidence regions and prediction regions is used to derive bootstrap tests for $H_0 : \boldsymbol{\mu} = \boldsymbol{c}$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{c}$ where $\boldsymbol{c}$ is some constant vector. Using $\boldsymbol{\mu} = \boldsymbol{A\beta}$ may be useful for testing after variable selection.

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.

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_5

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 $(\boldsymbol{\mu}, c_{\boldsymbol{x}}\boldsymbol{\Sigma}) = (E(\boldsymbol{x}), \text{Cov}(\boldsymbol{x}))$. 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

$$\text{RD}_i = \text{RD}_i(T_R, \boldsymbol{C}_R) = \sqrt{(\boldsymbol{x}_i - T_R(\boldsymbol{W}))^T [\boldsymbol{C}_R(\boldsymbol{W})]^{-1} (\boldsymbol{x}_i - T_R(\boldsymbol{W}))}$$

$= \tau\ D_i(T_A, \boldsymbol{C}_A)$ 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 $(\text{MD}_{n,\alpha}, \text{RD}_{n,\alpha})$ where $0 < \alpha < 1$ and $\text{MD}_{n,\alpha}$ is the 100αth sample percentile of the MD_i. Nevertheless, the variability in the DD plot may increase with the distances. Let $K > 0$ be a constant, e.g., 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): $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) =$

$$(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j)^T\left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \ - \ \frac{\boldsymbol{\Sigma}^{-1}}{a_j}+\hat{\boldsymbol{\Sigma}}_j^{-1}\right)(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j)$$

$$=(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j)^T\left(\frac{-\boldsymbol{\Sigma}^{-1}}{a_j}+\hat{\boldsymbol{\Sigma}}_j^{-1}\right)(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j)+(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j)^T\left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j}\right)(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j)$$

$$=\frac{1}{a_j}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j)^T(-\boldsymbol{\Sigma}^{-1}+a_j\ \hat{\boldsymbol{\Sigma}}_j^{-1})(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j) \ +$$

$$(\boldsymbol{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_j)^T\left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j}\right)(\boldsymbol{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_j)$$

$$=\frac{1}{a_j}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

$$+\frac{2}{a_j}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_j)+\frac{1}{a_j}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_j)$$

$$+\frac{1}{a_j}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j)^T[a_j\hat{\boldsymbol{\Sigma}}_j^{-1}-\boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j) \tag{5.1}$$

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. □

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 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, $(\text{med}(\text{MD}_i), \text{med}(D_i(A)))$. This line segment has slope

$$\text{med}(D_i(A))/\text{med}(\text{MD}_i)$$

which is generally not one. By taking $\tau = \text{med}(\text{MD}_i)/\text{med}(D_i(A))$, 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 $\text{med}(\text{MD}_i)$ 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 nonunit 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 et al. (1997), and Liu et al. (1999) for references.

Example 5.1. Rousseeuw and Van Driessen (1999) chose the multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution as the target. If the data are indeed iid MVN vectors, then the $(\text{MD}_i)^2$ are asymptotically χ^2_p random variables, and $\text{MED} = \sqrt{\chi^2_{p,0.5}}$ where $\chi^2_{p,0.5}$ is the median of the χ^2_p distribution. Since the target distribution is Gaussian, let

$$\text{RD}_i = \frac{\sqrt{\chi^2_{p,0.5}}}{\text{med}(D_i(A))} D_i(A) \quad \text{so that} \quad \tau = \frac{\sqrt{\chi^2_{\text{p},0.5}}}{\text{med}(\text{D}_\text{i}(\text{A}))}. \tag{5.2}$$

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 $\text{med}(\text{MD}_i) \approx \text{MED}$ where MED is the target population analog (obtained, e.g., via simulation, or from the actual target distribution as in Equation 3.10), then use

$$\text{RD}_i = \tau\ D_i(A) = \frac{\text{MED}}{\text{med}(D_i(A))} D_i(A). \tag{5.3}$$

We recommend using RFCH or RMVN as the robust estimators in DD plots. 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 hyperellipsoids that are nearly concentric. We suspect that the situation worsens as p increases. The `cov.mcd` estimator 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 under mild regularity conditions 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 $n \to \infty$.

Notice that if this conjecture is true, and if the data is EC with 2nd moments, then

$$\left[\frac{\text{med}(D_i(A))}{\text{med}(\text{MD}_i)}\right]^2 \boldsymbol{C}_A \tag{5.4}$$

Table 5.1 Corr(RD_i, MD_i) for $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ 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

estimates 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.

In a simulation study, $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ 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 nonsingular estimator of multivariate location and dispersion defines a hyperellipsoid, the DD plot can be used to examine which points are in the robust hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T_R)^T \boldsymbol{C}_R^{-1} (\boldsymbol{x} - T_R) \leq RD^2_{(h)}\} \tag{5.5}$$

where $RD^2_{(h)}$ is the hth smallest squared robust Mahalanobis distance, and which points are in a classical hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \leq MD^2_{(h)}\}. \tag{5.6}$$

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, e.g., if the data is EC. The DD plot will follow the identity line closely if $\text{med}(\text{MD}_i) \approx \text{MED}$, and $\text{RD}_i^2 =$

$$(\boldsymbol{x}_i - T_A)^T \left[\left(\frac{\text{MED}}{\text{med}(D_i(A))} \right)^2 \boldsymbol{C}_A^{-1} \right] (\boldsymbol{x}_i - T_A) \approx (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) = \text{MD}_i^2$$

for $i = 1, ..., n$. When the distribution is not EC, the RMVN (or RFCH or FMCD) estimator 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 toward elliptical symmetry. This application is important since many statistical methods assume that the underlying data distribution is MVN or EC.

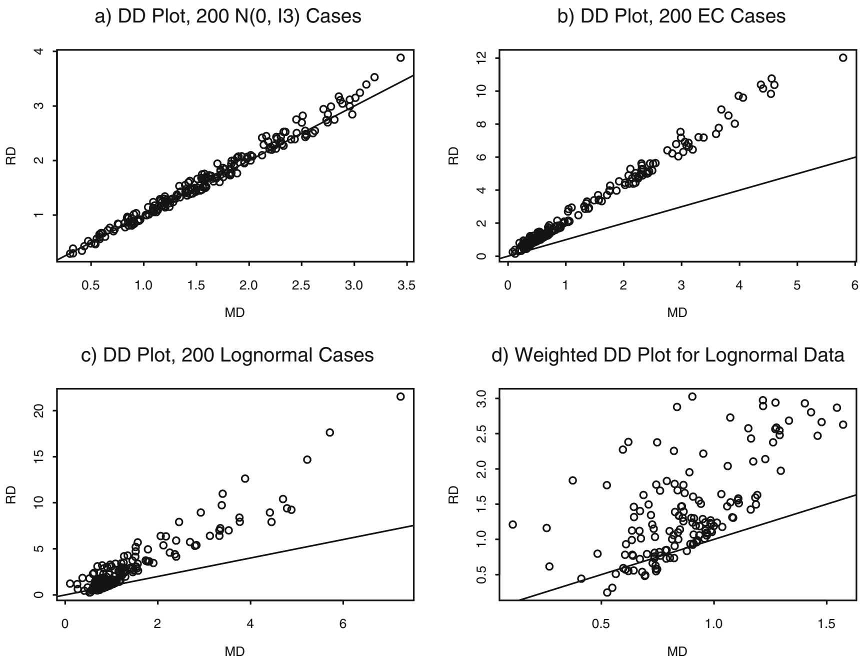

Fig. 5.1 4 DD Plots

For this application, the RFCH and RMVN estimators 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 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 $\text{RD}_i \geq \sqrt{\chi^2_{p,.975}}$. 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.

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; i.e., 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 $\text{RD}_i \geq \sqrt{\chi^2_{3,.975}} \approx 3.06$ 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 toward a target EC distribution have been developed. Generalizations of the Box–Cox transformation toward a multivariate normal distribution are described in Velilla (1993). Alternatively, Cook and Nachtsheim (1994) gave a two-step numerical procedure for transforming data toward 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.

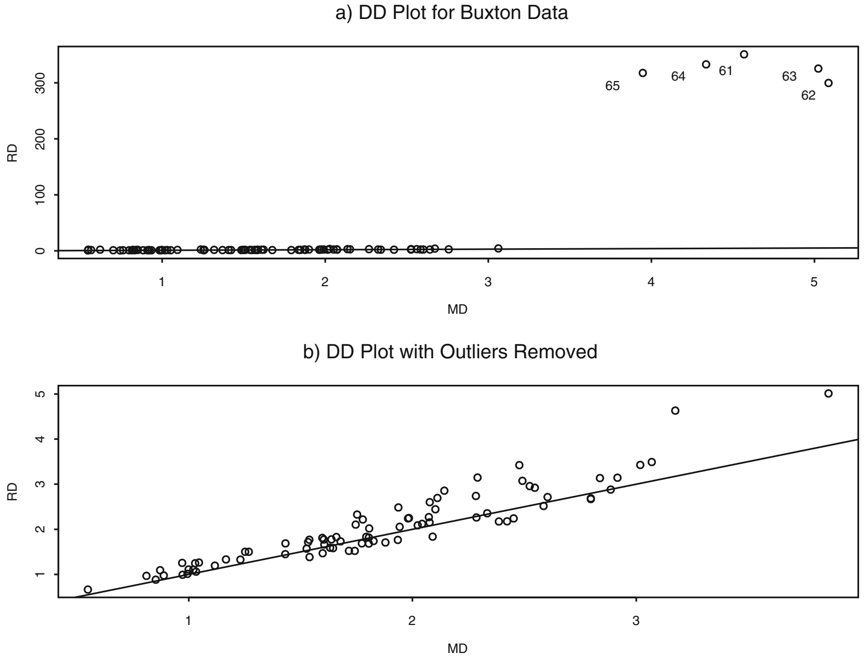

Fig. 5.2 DD Plots for the Buxton Data

Example 5.2. Buxton (1920, pp. 232–5) gave 20 measurements of 88 men. We will examine whether the multivariate normal distribution is a reasonable 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 multivariate normal distribution is reasonable. (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.

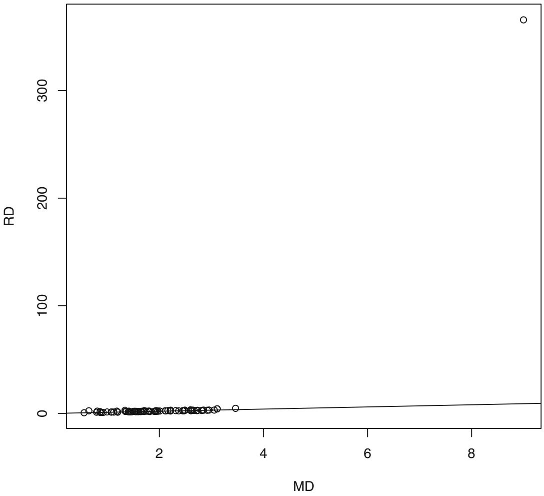

Fig. 5.3 DD Plot With One Outlier in the Upper Right Corner

Application 5.2. The DD plot can be used to detect multivariate outliers. See Figures 4.2, 4.4, 5.2a, and 5.3.

Warning: It is important to know that plots fill space. If there is a single outlier, then often it will appear in the upper left or upper right corner of the DD plot, where RD is large, since the plot has to cover the outlier. The rest of the data will often appear to be tightly clustered about the identity line. Beginners sometimes fail to spot the single outlier because they do not know that the plot will fill space. There is a lot of blank space because of the outlier. If the outlier was not present, then the box would not extend much above the identity line in the upper right corner of the plot. For example, suppose all of the outliers except point 63 were deleted from the Buxton data. Then compare the DD plot in Figure 5.2 b) where all of the outliers have been deleted, with the DD plot in Figure 5.3 where the single outlier is in the upper right corner. *R* commands to produce Figures 5.2 and 5.3 are shown below.

```
library(MASS)
x <- cbind(buxy,buxx)
ddplot(x,type=3) #Figure 5.2a), right click Stop

zx <- x[-c(61:65),]
ddplot(zx,type=3) #Figure 5.2b), right click Stop

zz <- x[-c(61,62,64,65),]
ddplot(zz,type=3) #Figure 5.3, right click Stop
```

5.2 Prediction Regions

Consider predicting a future test value $\boldsymbol{x}_f$, given past training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid. Much as confidence regions and intervals give a measure of precision for the point estimator $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$, prediction regions and intervals give a measure of precision of the point estimator $T = \hat{\boldsymbol{x}}_f$ of the future random vector $\boldsymbol{x}_f$. The material in this section will be useful for Section 5.3 and for developing practical prediction regions for multivariate linear regression. See Section 12.3.

Definition 5.2. A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n) \to 1-\delta$ as $n \to \infty$. A prediction region is *asymptotically optimal* if its volume converges in probability to the volume of the minimum volume covering region or the highest density region of the distribution of $\boldsymbol{x}_f$.

Some researchers define a large sample prediction region $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n) \geq 1-\delta$, asymptotically. As an example for $p = 1$, a large sample $100(1-\delta)\%$ *prediction interval* (PI) has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1-\delta$ as the sample size $n \to \infty$. Open intervals are often used. If the highest density region is an interval, then a PI is asymptotically optimal if it has the shortest asymptotic length that gives the desired asymptotic coverage. The following definition makes sense when the highest density region is unique. If $p = 1$, nonzero flat spots in the pdf can cause the region to have higher than nominal coverage. For example, the highest density region of a uniform(θ_1, θ_2) random variable is not unique.

Definition 5.3. When unique, the $100(1-\delta)\%$ *highest density region* $R(f_{1-\delta}) = \{\boldsymbol{z} : f(\boldsymbol{z}) \geq f_\delta\}$ where f_δ is the largest constant such that $P[\boldsymbol{x} \in R(f_{1-\delta})] \geq 1-\delta$ and $f(\boldsymbol{z})$ is the probability density function (pdf) of $\boldsymbol{x}$.

For elliptically contoured distributions with continuous decreasing g, the highest density region is the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) \leq u_{1-\delta}\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq u_{1-\delta}\} \tag{5.7}$$

where $P(U \leq u_{1-\delta}) = 1-\delta$, and U is given by (3.9). If $HDR_Y(1-\delta)$ is the $100(1-\delta)\%$ highest density region for a random variable Y, and $X \sim U(0,\theta) \perp\!\!\!\perp Y$ (meaning X is independent of Y), then the $100(1-\delta)\%$ highest density region for (X_f, Y_f) is

$$\{(x,y) : x \in (0,\theta), y \in HDR_Y(1-\delta)\}.$$

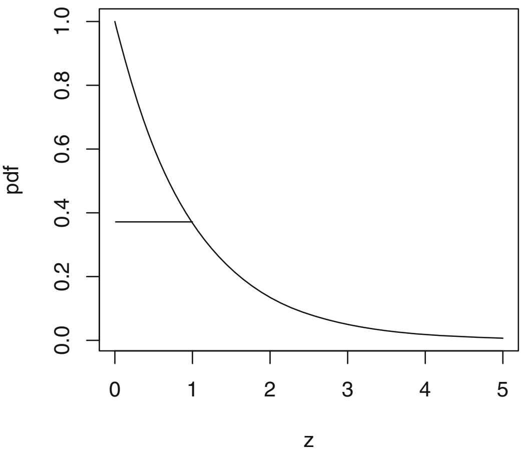

Fig. 5.4 Highest 36.8% Density Region is (0,1)

To illustrate the highest density region, first let $p = 1$. Suppose $X_1, ..., X_n$ are iid from a unimodal pdf that has interval support, and that the pdf $f(z)$ decreases rapidly as z moves away from the mode. Let (a, b) be the shortest interval such that $F(b) - F(a) = 1 - \delta$ where the cumulative distribution function $F(x) = P(X \leq x)$. Then the interval is the highest density region containing $1-\delta$ of the mass. To find the $(1-\delta)100\%$ highest density region of a pdf, move a horizontal line down from the top of the pdf. The line will intersect the pdf or the boundaries of the support of the pdf at $(a_1, b_1), ..., (a_k, b_k)$ for some $k \geq 1$. Stop moving the line when the area under the pdf corresponding to the intervals is equal to $1 - \delta$. See Figure 5.4 where the area under the pdf from 0 to 1 gives the 36.8% highest density region. The region will often have $f(a) = f(b)$, e.g., if the support where $f(z) > 0$ is $(-\infty, \infty)$. For $p = 2$, a horizontal plane is moved up and down the joint pdf until the area under the "intersection" of the plane and the joint pdf (or the boundaries of the support of the pdf) equals $1 - \delta$. (For a joint pmf $f(\boldsymbol{z}) = P(\boldsymbol{x} = \boldsymbol{z})$, the sum of the $f(\boldsymbol{z})$ such that $f(\boldsymbol{z}) \geq f_\delta$ is $\geq 1-\delta$.) Figure 2.1 shows the highest density regions for two bivariate normal distributions.

There is a moderate amount of literature for prediction regions that may perform well for small p. Let $\hat{f}_{(1)}, ..., \hat{f}_{(n)}$ be the order statistics of $\hat{f}(\boldsymbol{x}_1), ..., \hat{f}(\boldsymbol{x}_n)$. Hyndman (1996) used the estimated highest density region

$$\hat{R}(f_{1-\delta}) = \{\boldsymbol{z} : d\hat{f}(\boldsymbol{z}) \geq d\hat{f}_{(h)}\} \tag{5.8}$$

where $d > 0$ can be any constant, $h = max(1, \lfloor n\delta \rfloor)$, and $\lfloor x \rfloor$ is the integer part of x. Also see Lei et al. (2013), who estimated $f(\boldsymbol{z})$ with a kernel density estimator, for references. See Section 8.6 for kernel density estimators.

For $p = 1$ and positive integer c, the shorth(c) estimator is a useful estimator of the highest density region when the region is an interval.

Definition 5.4. Let $Z_{(1)}, ..., Z_{(n)}$ be the order statistics of $Z_1, ..., Z_n$. Consider intervals that contain c cases: $[Z_{(1)}, Z_{(c)}], [Z_{(2)}, Z_{(c+1)}], ..., [Z_{(n-c+1)}, Z_{(n)}]$. Compute $Z_{(c)} - Z_{(1)}, Z_{(c+1)} - Z_{(2)}, ..., Z_{(n)} - Z_{(n-c+1)}$. Then let the shortest closed interval containing at least c of the Z_i be

$$\mathrm{shorth(c)} = [\mathrm{Z_{(s)}}, \mathrm{Z_{(s+c-1)}}]. \tag{5.9}$$

Let

$$k_n = \lceil n(1 - \delta) \rceil \tag{5.10}$$

where $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. Frey (2013) showed that for large $n\delta$ and iid data, the shorth(k_n) PI has undercoverage that depends on the distribution of the data with maximum undercoverage $\approx 1.12\sqrt{\delta/n}$, and used the shorth($c$) estimator as the large sample $100(1-\delta)\%$ PI where

$$c = \min(n, \lceil n[1 - \delta + 1.12\sqrt{\delta/n}\ ]\ \rceil). \tag{5.11}$$

Example 5.3. Given below were votes for preseason 1A basketball poll from Nov. 22, 2011 WSIL News where the 778 was a typo: the actual value was 78. As shown below, finding shorth(3) from the ordered data is simple. If the outlier was corrected, shorth(3) = [76,78].

```
111   89    778   78    76

order data: 76 78 89 111 778

                   13 = 89 - 76

                       33 = 111 - 78

                          689 = 778 - 89
shorth(3) = [76,89]
```

Let $D^2_{(c)}$ be the cth order statistic of $D^2_1, ..., D^2_n$, and consider the hyperellipsoid

$$\mathcal{A}_n = \{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(c)}\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(c)}\}. \tag{5.12}$$

If n is large, we can use $c = k_n = \lceil n(1-\delta) \rceil$. If n is not large, using $c = U_n$ where U_n decreases to k_n can improve small sample performance. U_n will be defined in the paragraph below Equation (5.16). Olive (2013a) showed that (5.12) is a large sample $100(1-\delta)\%$ prediction region under mild conditions, although regions with smaller volumes may exist. Note that the result follows since if $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{S}$ are nonsingular, then the Mahalanobis distance is a continuous function of $(\overline{\boldsymbol{x}}, \boldsymbol{S})$. Let $D = D(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Then $D_i \overset{D}{\rightarrow} D$ and $D_i^2 \overset{D}{\rightarrow} D^2$. Hence the sample percentiles of the D_i are consistent estimators of the population percentiles of D at continuity points of the cumulative distribution function (cdf) of D. The prediction region (5.12) estimates the highest density region for a large class of elliptically contoured distributions.

A problem with the prediction regions that cover $\approx 100(1-\delta)\%$ of the training data cases $\boldsymbol{x}_i$ (such as (5.8), (5.9), and (5.12) for appropriate h or c) is that they have coverage lower than the nominal coverage of $1-\delta$ for moderate n. This result is not surprising since empirically *statistical methods perform worse on test data than on training data.* Increasing c will improve the coverage for moderate samples. Then empirically for many distributions, for $n \approx 20p$, the two prediction regions (5.9) and (5.12) applied to iid data or pseudodata using $k_n = \lceil n(1-\delta) \rceil$ tend to have undercoverage as high as 5%. The undercoverage decreases rapidly as n increases. Let
$q_n = \min(1-\delta+0.05, 1-\delta+p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1-\delta/2, 1-\delta+10\delta p/n), \quad \text{otherwise.} \tag{5.13}$$

If $1-\delta < 0.999$ and $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. Using

$$c = \lceil n q_n \rceil \tag{5.14}$$

in (5.12) decreased the undercoverage.

If $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\ \boldsymbol{\Sigma})$ for some constant $d > 0$ where $\boldsymbol{\Sigma}$ is nonsingular, then $D^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - d^{-1}\boldsymbol{\Sigma}^{-1} + d^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$$

$$= d^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_p(1).$$

Thus the sample percentiles of $D_i^2(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $d^{-1}D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (at continuity points $D_{1-\delta}$ of the cdf of $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$). If $\boldsymbol{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi^2_m$.

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

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \le h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \le h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \le h\} \tag{5.15}$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)}. \tag{5.16}$$

A future observation (random vector) $\boldsymbol{x}_f$ is in the region (5.15) if $D_{\boldsymbol{x}_f} \leq h$.

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ for some constant $d > 0$ where $\boldsymbol{\Sigma}$ is nonsingular, then (5.15) is a large sample $100(1-\delta)\%$ prediction region if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the $100q_n$th sample quantile of the D_i where q_n is defined above (5.14). For example, use $U_n = c = \lceil nq_n \rceil$. If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ and $\boldsymbol{x}_f$ are iid, then region (5.15) is asymptotically optimal on a large class of elliptically contoured distributions in that the region's volume converges in probability to the volume of the highest density region (5.7).

The Olive (2013a) nonparametric prediction region uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. For the classical prediction region, see Chew (1966) and Johnson and Wichern (1988, pp. 134, 151). Refer to the above paragraph for $D_{(U_n)}$.

Definition 5.5. The large sample $100(1-\delta)\%$ *nonparametric prediction region* for a future value $\boldsymbol{x}_f$ given iid data $\boldsymbol{x}_1, ..., , \boldsymbol{x}_n$ is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}, \tag{5.17}$$

while the large sample $100(1-\delta)\%$ *classical prediction region* is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi^2_{p,1-\delta}\}. \tag{5.18}$$

If p is small, Mahalanobis distances tend to be right skewed with a population shorth that discards the right tail. For $p = 1$ and $n \geq 20$, the finite sample correction factors c/n for c given by (5.11) and (5.14) do not differ by much more than 3% for $0.01 \leq \delta \leq 0.5$. See Figure 5.5 where ol = (eq. 5.14)$/n$ is plotted versus fr = (eq. 5.11)$/n$ for $n = 20, 21, ..., 500$. The top plot is for $\delta = 0.01$, while the bottom plot is for $\delta = 0.3$. The identity line is added to each plot as a visual aid. The value of n increases from 20 to 500 from the right of the plot to the left of the plot. Examining the axes of each plot shows that the correction factors do not differ greatly. *R* code to create Figure 5.5 is shown below.

```
cmar <- par("mar"); par(mfrow = c(2, 1))
par(mar=c(4.0,4.0,2.0,0.5))
frey(0.01); frey(0.3)
par(mfrow = c(1, 1)); par(mar=cmar)
```

Remark 5.1. The nonparametric prediction region (5.17) is useful if $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from a distribution with a nonsingular covariance matrix, and the sample size n is large enough. The distribution could be continuous, discrete, or a mixture. The asymptotic coverage is $1-\delta$ if the $100(1-\delta)$th percentile $D_{1-\delta}$ of D is a continuity point of the distribution of D, although prediction regions with smaller volume may exist. If $D_{1-\delta}$ is not a continuity point of the distribution of D, then the asymptotic coverage tends to be $\geq 1-\delta$ since a sample percentile with cutoff q_n that decreases to $1-\delta$ is used

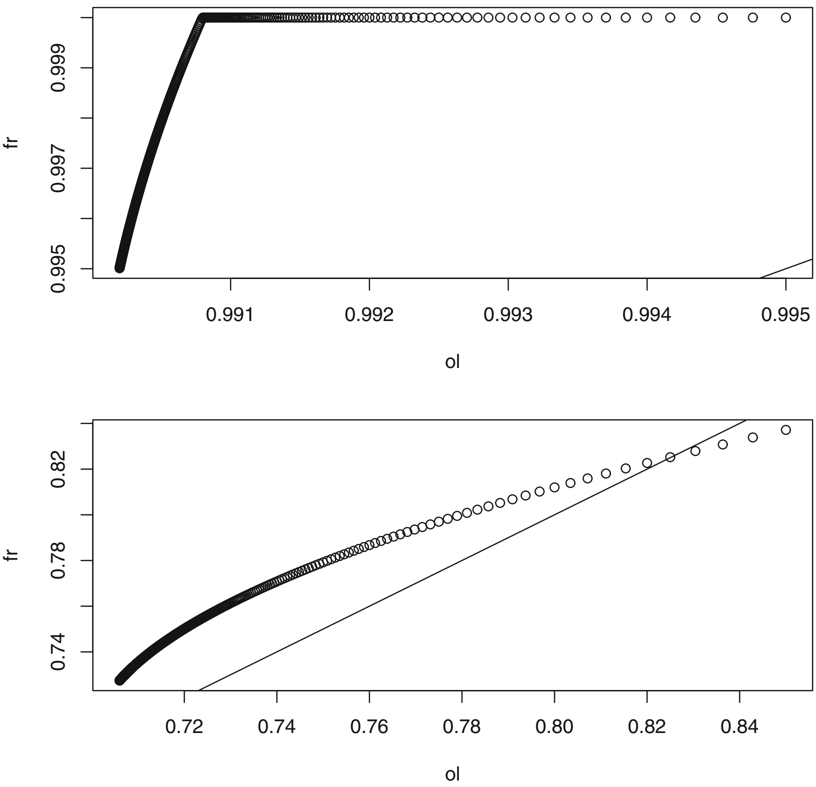

Fig. 5.5 Correction Factor Comparison when $\delta = 0.01$ (Top Plot) and $\delta = 0.3$ (Bottom Plot)

and a closed region is used. Often D has a continuous distribution and hence has no discontinuity points for $0 < \delta < 1$. (If there is a jump in the distribution from 0.9 to 0.96 at discontinuity point a, and the nominal coverage is 0.95, we want 0.96 coverage instead of 0.9. So we want the sample percentile to decrease to a.) The nonparametric prediction region (5.17) contains U_n of the training data cases $\boldsymbol{x}_i$ provided that $\boldsymbol{S}$ is nonsingular, even if the model is wrong. For many distributions, the coverage started to be close to $1 - \delta$ for $n \geq 10p$ where the coverage is the simulated percentage of times that the prediction region contained $\boldsymbol{x}_f$.

Remark 5.2. The most used prediction regions assume that the error vectors are iid from a multivariate normal distribution. The ratio of the volumes of regions (5.18) and (5.17) is

$$\left(\frac{\chi^2_{p,1-\delta}}{D^2_{(U_n)}}\right)^{p/2},$$

which can become close to zero rapidly as p gets large if the $\boldsymbol{x}_i$ are not from the light-tailed multivariate normal distribution. For example, suppose $\chi^2_{4,0.5} \approx 3.33$ and $D^2_{(U_n)} \approx D^2_{\boldsymbol{x},0.5} = 6$. Then the ratio is $(3.33/6)^2 \approx 0.308$. Hence if the data is not multivariate normal, severe undercoverage can occur if the classical prediction region is used, and the undercoverage tends to get worse as the dimension p increases. The coverage need not to go to 0, since by the multivariate Chebyshev's inequality, $P(D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq \gamma) \geq 1 - p/\gamma > 0$ for $\gamma > p$ where the population covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{x})$. See Budny (2014), Chen (2011), and Navarro (2014, 2016). Using $\gamma = h^2 = p/\delta$ in (5.15) usually results in prediction regions with volume and coverage that is too large.

Remark 5.3. There may not yet be any practical competing prediction regions that do not have the form of (5.15) if p is much larger than two and the distribution of the $\boldsymbol{x}_i$ is unknown. Remark 5.1 suggests that the nonparametric prediction region (5.17) starts to have good coverage for $n \geq 10p$ for a large class of distributions. Of course for any n, there are error distributions that will have severe undercoverage. Prediction regions that estimate the pdf $f(\boldsymbol{z})$ with a kernel density estimator quickly become impractical as p increases since large sample sizes are needed for good estimates. A similar problem occurs when using a kernel density estimator for discriminant analysis. See Silverman (1986, p. 129).

For example, the Hyndman nominal 95% prediction region (5.8) was computed for iid $N_p(\boldsymbol{0}, \boldsymbol{I})$ data with 1000 runs. Let the coverage be the observed proportion of prediction regions that contained the future value $\boldsymbol{x}_f$. For $p = 1$, the coverage was 0.933 for $n = 40$. For $p = 2$, the coverage was 0.911 for $n = 50$ and 0.930 for $n = 150$. For $p = 4$, the coverage was 0.920 for $n = 250$. For $p = 5$, the coverage was 0.866 for $n = 200$ and 0.934 for $n = 2000$. For $p = 8$, the coverage was 0.735 for $n = 125$. For the multivariate lognormal distribution with $n = 20p$, the Olive (2013a) large sample nonparametric 95% prediction region (5.17) had coverages 0.970, 0.959, and 0.964 for $p = 100, 200$, and 500. Some *R* code is below.

```
nruns=1000 #p = 1
count<-0
for(i in 1:nruns){
x <- rnorm(40)
xff <- rnorm(1)
count <- count + hdr2(x,xf=xff)$inr}
count #933/1000
```

```
count<-0 #p = 5
for(i in 1:nruns){
x <- matrix(rnorm(1000),ncol=5,nrow=200)
xff <- as.vector(rnorm(5))
count <- count + hdr2(x,xf=xff)$inr}
count #886/1000

#lognormal, p = 100
count<-0
for(i in 1:nruns){
x <- exp(matrix(rnorm(200000),ncol=100,nrow=2000))
xff <- exp(as.vector(rnorm(100)))
count <- count + predrgn(x,xf=xff)$inr}
count #970/1000
```

Olive (2013a) used three prediction regions (5.15) that can be displayed with the DD plot. The nonparametric prediction region (5.17) uses the classical estimator $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ and $h = D_{(U_n)}$. The other two prediction regions are defined below.

Definition 5.6. The *semiparametric prediction region* uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h = D_{(U_n)}$. The *parametric MVN prediction region* uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h^2 = \chi^2_{p,q_n}$ where $P(W \le \chi^2_{p,\delta}) = \delta$ if $W \sim \chi^2_p$.

All three prediction regions are asymptotically optimal for MVN distributions with nonsingular $\boldsymbol{\Sigma}$. The first two prediction regions are asymptotically optimal for a large class of $EC(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distributions given by Assumption (E1) used in Theorem 4.8, provided g is continuous and decreasing. For distributions with nonsingular covariance matrix $c_X \boldsymbol{\Sigma}$, the nonparametric region is a large sample $(1-\delta)100\%$ prediction region (provided $D_{1-\delta}$ is a continuity point of the cdf of D), but regions with smaller volume may exist.

Notice that for the training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, if $\boldsymbol{C}^{-1}$ exists, then $c \approx 100 q_n\%$ of the n cases are in the prediction regions for $\boldsymbol{x}_f = \boldsymbol{x}_i$, and $q_n \to 1-\delta$ even if $(T, \boldsymbol{C})$ is not a good estimator. Hence the coverage q_n of the training 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 $q_n = 1 - \delta/2$ or $q_n = 1 - \delta + 0.05$ for $n \le 20p$ and $q_n \to 1 - \delta$ as $n \to \infty$. If $q_n \equiv 1 - \delta$ and $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ where $d > 0$ and $\boldsymbol{\Sigma}$ is nonsingular, then (5.15) is a large sample prediction region, but taking q_n given by (5.13) improves the finite sample performance of the prediction region. Taking $q_n \equiv 1 - \delta$ 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 $\min(0.05, \delta/2)$. Using (5.13) helped reduce undercoverage for small n due to the unknown variability of $(T, \boldsymbol{C})$.

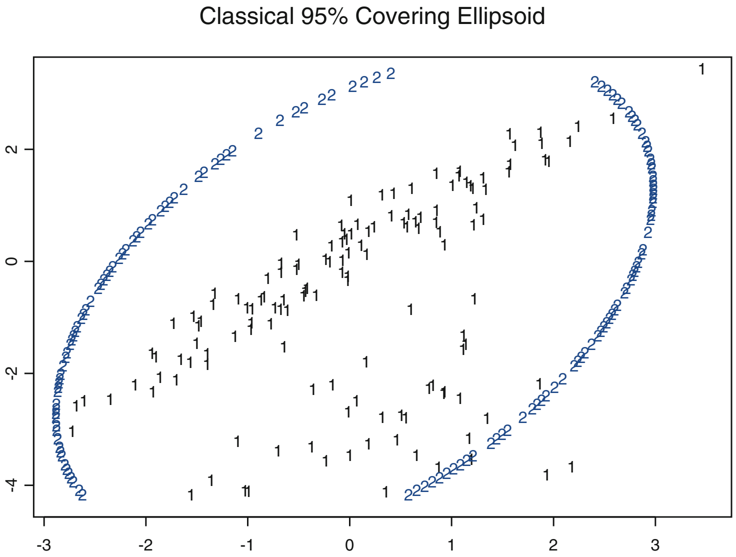

Fig. 5.6 Artificial Bivariate Data

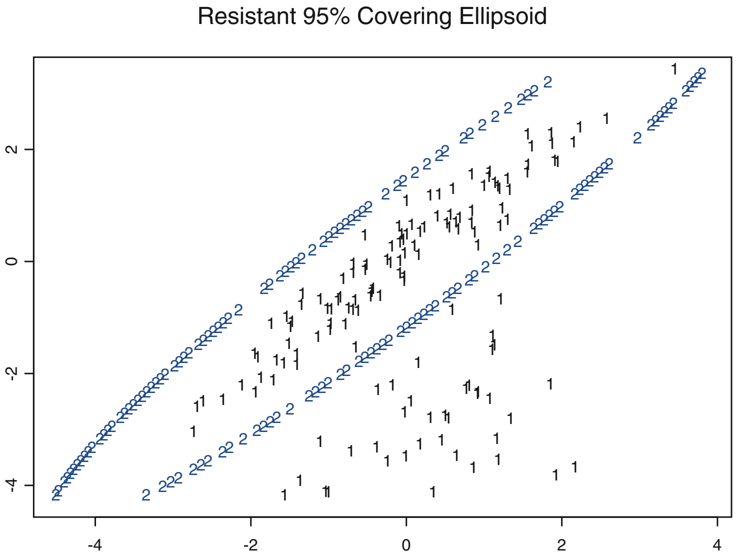

Fig. 5.7 Artificial Data

Example 5.4. An artificial data set consisting of 100 iid cases from a

$$N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1.49 & 1.4\\1.4 & 1.49\end{pmatrix}\right)$$

distribution and 40 iid cases from a bivariate normal distribution with mean $(0,-3)^T$ and covariance $\boldsymbol{I}_2$. Figure 5.6 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. There is an *R* package that makes an ellipse. Notice that the classical parametric ellipsoid covers almost all of the data. Figure 5.7 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 $q_n = 0.95$.

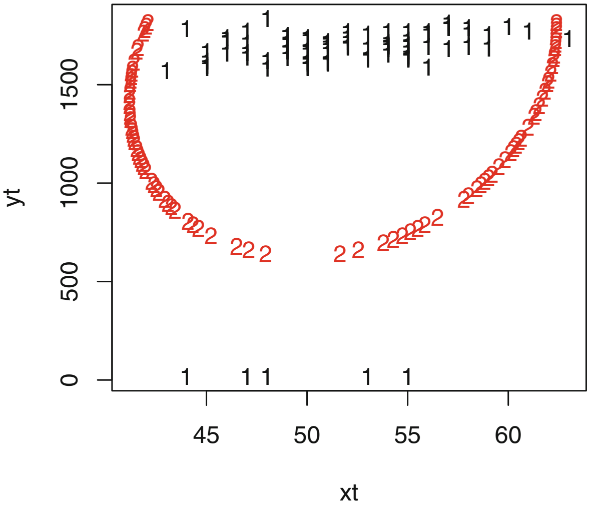

Fig. 5.8 Ellipsoid is Inflated by Outliers

Example 5.5. Buxton (1920) gave 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 the actual heights for these cases were recorded as the head lengths) and are massive outliers. First *height* and *nasal height* were used with $q_n = 0.95$. Figure 5.8 shows that the classical

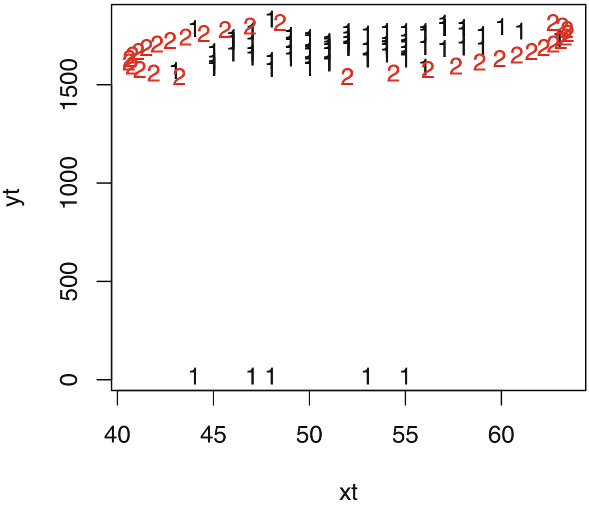

Fig. 5.9 Ellipsoid Ignores Outliers

parametric prediction region (using $MD \leq \sqrt{\chi^2_{2,.95}}$) is quite large but does not include any of the outliers. Figure 5.9 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 5.10 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 nonparametric and semiparametric 90% regions blow up unless the outlier proportion is small.

Figure 5.10 can be made with the following *R* commands, assuming source commands for `mpack` and `mrobdata` have been performed. See the Preface or Section 15.2. Right click `Stop` to get the cursor.

```
x <- cbind(buxy,buxx)
ddplot4(x)  #right click Stop
```

Figure 5.11 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 again at 3.33, slightly below the semiparametric region cutoff of 3.44. Cases to the left of the vertical line $MD = 3.33$ (not shown since you can mentally drop down a vertical line where the horizontal line ends at the identity line) correspond to a (modified) classical prediction region.

Figure 5.11 can be made with the following *R* commands. Right click `Stop` to get the cursor and the output following the two commands.

```
zx <- x[-c(61:65),]
ddplot4(zx) #right click Stop
$cuplim
     95%
3.086005
$ruplim
     95%
3.438821
$mvnlim
[1] 3.327236
```

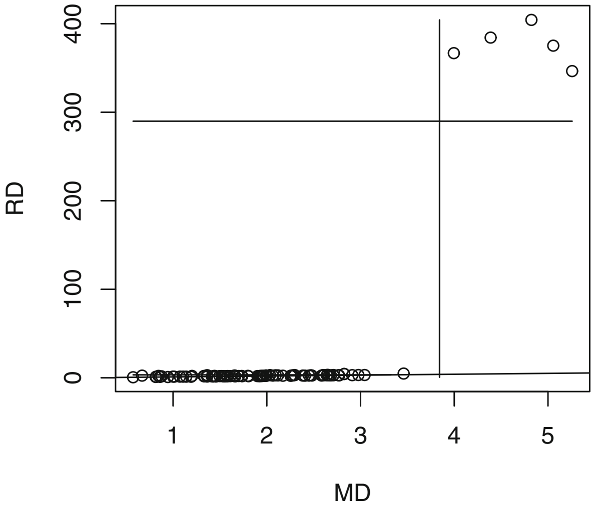

Fig. 5.10 Prediction Regions for Buxton Data

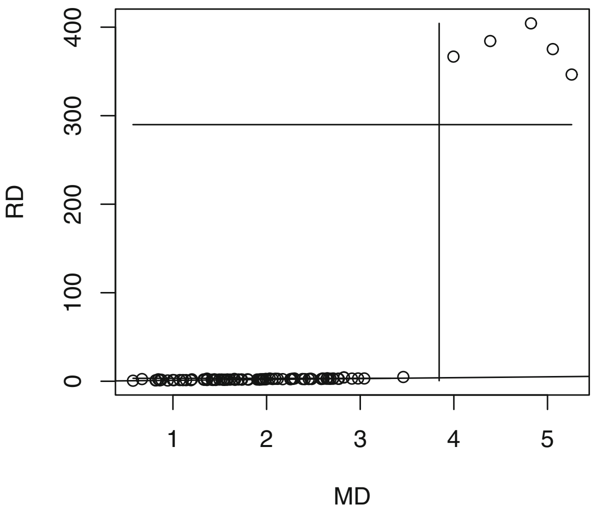

Fig. 5.11 Prediction Regions for Buxton Data without Outliers

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 $\delta = 0.1$.

Often the coverage for the semiparametric region was better than that of the nonparametric region for n near $10p$. The nonparametric covering region $\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1}(\boldsymbol{z} - \overline{\boldsymbol{x}}) \leq D^2_{(n)}(\overline{\boldsymbol{x}}, \boldsymbol{S})\}$ uses all of the data, but for small n, data is sparse, and the covering region overfits and hence the volume is too small. The nonparametric prediction region is a hyperellipsoid that is concentric with the covering region (that replaces $D^2_{(U_n)}$ with $D^2_{(n)}$). The semiparametric region is based on the RMVN half set of data. This region is not a good estimator of the population 50% covering region for small n. Hence when it is blown up to cover 95% of the training data, the region is quite large, so it is likely that a future $\boldsymbol{x}_f$ is in the region.

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 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 $10p \le n \le 20p$, the nonparametric region often had good coverage (and volume ratio near 0.5 for MVN data).

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 5000 runs, this result held for $2 \le p \le 80$. For the nonelliptically 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 Bootstrapping Hypothesis Tests and Confidence Regions

This section shows that, under regularity conditions, applying the nonparametric prediction region of Section 5.2 to a bootstrap sample results in a confidence region. When teaching confidence intervals, it is often noted that by the central limit theorem, the probability that $\overline{Y}_n$ is within two standard deviations ($2SD(\overline{Y}_n) = 2\sigma/\sqrt{n}$) of μ is about 95%. Hence the probability that μ is within two standard deviations of $\overline{Y}_n$ is about 95%. Thus the interval $[\mu - 1.96S/\sqrt{n}, \mu + 1.96S/\sqrt{n}\ ]$ is a large sample 95% prediction interval for a future value of the sample mean $\overline{Y}_{n,f}$ if μ is known, while $[\overline{Y}_n - 1.96S/\sqrt{n}, \overline{Y}_n + 1.96S/\sqrt{n}\ ]$ is a large sample 95% confidence interval for the population mean μ. Note that the lengths of the two intervals are the same. Where the interval is centered determines whether the interval is a confidence or a prediction interval.

Definition 5.7. A *large sample* $100(1-\delta)\%$ *confidence region* for a vector of parameters $\boldsymbol{\mu}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\mu} \in \mathcal{A}_n) \to 1 - \delta$ as $n \to \infty$.

Some researchers define a large sample confidence region $\mathcal{A}_n$ such that $P(\boldsymbol{\mu} \in \mathcal{A}_n) \ge 1 - \delta$, asymptotically. The following theorem shows that the hyperellipsoid R_c centered at the statistic T_n is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$, but the hyperellipsoid R_p centered at known $\boldsymbol{\mu}$ is a large sample $100(1-\delta)\%$ prediction region for a future value of the statistic $T_{f,n}$. The result uses the fact that the squared distance of $\boldsymbol{\mu}$ from a statistic T_n in a hyperellipsoid R_c centered at T_n is equal to the squared distance of T_n from a parameter $\boldsymbol{\mu}$ in a hyperellipsoid R_p centered at $\boldsymbol{\mu}$:

$$D^2_{\boldsymbol{\mu}}(T_n, \hat{\boldsymbol{\Sigma}}_T) = (\boldsymbol{\mu} - T_n)^T \hat{\boldsymbol{\Sigma}}_T^{-1} (\boldsymbol{\mu} - T_n) =$$

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

As in Remark 5.1, if the $100(1-\delta)$th percentile of D^2 is not a continuity point of the distribution of D^2, then the asymptotic coverage tends to be $\geq 1-\delta$ if a sample percentile with cutoff q_n that decreases to $1-\delta$ is used, since a closed region is used. Often D^2 has a continuous distribution and hence has no discontinuity points for $0 < \delta < 1$.

Theorem 5.2. Let the $100(1-\delta)$th percentile $D^2_{1-\delta}$ be a continuity point of the distribution of D^2. Assume that $D^2_{\boldsymbol{\mu}}(T_n, \boldsymbol{\Sigma}_T) \xrightarrow{D} D^2$, $D^2_{\boldsymbol{\mu}}(T_n, \hat{\boldsymbol{\Sigma}}_T) \xrightarrow{D} D^2$, and $\hat{D}^2_{1-\delta} \xrightarrow{P} D^2_{1-\delta}$ where $P(D^2 \leq D^2_{1-\delta}) = 1-\delta$. i) Then $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}^2_{1-\delta}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$, and if $\boldsymbol{\mu}$ is known, then $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}^2_{1-\delta}\}$ is a large sample $100(1-\delta)\%$ prediction region for a future value of the statistic $T_{f,n}$. ii) Region R_c contains $\boldsymbol{\mu}$ iff region R_p contains T_n.

Proof. i) From the discussion above, $D^2_{\boldsymbol{\mu}}(T_n, \hat{\boldsymbol{\Sigma}}_T) = D^2_{T_n}(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_T)$. Thus the probability that R_c contains $\boldsymbol{\mu}$ is $P(D^2_{\boldsymbol{\mu}}(T_n, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}^2_{1-\delta}) \to 1-\delta$, and the probability that R_p contains $T_{f,n}$ is $P(D^2_{\boldsymbol{\mu}}(T_{f,n}, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}^2_{1-\delta}) \to 1-\delta$, as $n \to \infty$.

ii) $D^2_{\boldsymbol{\mu}}(T_n, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}^2_{1-\delta}$ iff $D^2_{T_n}(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}^2_{1-\delta}$ since $D^2_{\boldsymbol{\mu}}(T_n, \hat{\boldsymbol{\Sigma}}_T) = D^2_{T_n}(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_T)$. □

Hence if there was an iid sample $T_{1,n}, ..., T_{B,n}$ of the statistic, the large sample $100(1-\delta)\%$ nonparametric prediction region $\{\boldsymbol{w} : D^2(\overline{T}, \boldsymbol{S}_T) \leq D^2_{(c)}\}$ for $T_{f,n}$ contains $E(T_n) = \boldsymbol{\mu}$ with asymptotic coverage $\geq 1-\delta$. To make the asymptotic coverage equal to $1-\delta$, use the large sample $100(1-\delta)\%$ confidence region $\{\boldsymbol{w} : D^2(T_{1,n}, \boldsymbol{S}_T) \leq D^2_{(c)}\}$. The prediction region method bootstraps this procedure by using a bootstrap sample of the statistic $T^*_{1,n}, ..., T^*_{B,n}$. Centering the region at $T^*_{1,n}$ instead of $\overline{T^*}$ is not needed since the bootstrap sample is centered near T_n: the distribution of $\sqrt{n}(T_n - \boldsymbol{\mu})$ is approximated by the distribution of $\sqrt{n}(T^* - T_n)$ or by the distribution of $\sqrt{n}(T^* - \overline{T^*})$. See Equations (5.20), (5.25), and (5.26). Also note that if $T_n = t(\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$, $T^*_{in} = t(\boldsymbol{w}^*_{i1}, ..., \boldsymbol{w}^*_{in})$, and $\boldsymbol{w}^*_{i1}, ..., \boldsymbol{w}^*_{in}$ is permutation of $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$, then $T^*_{in} = T_n$ if t is permutation invariant.

A bootstrap sample is a sample from the empirical distribution: there are n cases, and a sample of size n of the cases is drawn with replacement, and each case is equally likely to be drawn. The residual bootstrap draws a bootstrap sample from the residuals. The following subsection will help clarify ideas.

5.3.1 The Bootstrap

Definition 5.8. Suppose that data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ has been collected and observed. Often the data is a random sample (iid) from a distribution with cdf F. The *empirical distribution* is a discrete distribution where the $\boldsymbol{x}_i$ are the possible values, and each value is equally likely. If $\boldsymbol{w}$ is a random variable having the empirical distribution, then $p_i = P(\boldsymbol{w} = \boldsymbol{x}_i) = 1/n$ for $i = 1, ..., n$. The *cdf of the empirical distribution* is denoted by F_n.

Example 5.6. Let $\boldsymbol{w}$ be a random variable having the empirical distribution given by Definition 5.8. Show that $E(\boldsymbol{w}) = \overline{\boldsymbol{x}} \equiv \overline{\boldsymbol{x}}_n$ and $\text{Cov}(\boldsymbol{w}) = \frac{n-1}{n}\boldsymbol{S} \equiv \frac{n-1}{n}\boldsymbol{S}_n$.

Solution: Recall that for a discrete random vector, the population expected value $E(\boldsymbol{w}) = \sum \boldsymbol{x}_i p_i$ where $\boldsymbol{x}_i$ are the values that $\boldsymbol{w}$ takes with positive probability p_i. Similarly, the population covariance matrix

$$\text{Cov}(\boldsymbol{w}) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(\boldsymbol{w} - E(\boldsymbol{w}))^T] = \sum (\boldsymbol{x}_i - E(\boldsymbol{w}))(\boldsymbol{x}_i - E(\boldsymbol{w}))^T p_i.$$

Hence

$$E(\boldsymbol{w}) = \sum_{i=1}^n \boldsymbol{x}_i \frac{1}{n} = \overline{\boldsymbol{x}},$$

and

$$\text{Cov}(\boldsymbol{w}) = \sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \frac{1}{n} = \frac{n-1}{n}\boldsymbol{S}. \quad \square$$

Example 5.7. If $W_1, ..., W_n$ are iid from a distribution with cdf F_W, then the empirical cdf F_n corresponding to F_W is given by

$$F_n(y) = \frac{1}{n}\sum_{i=1}^n I(W_i \leq y)$$

where the indicator $I(W_i \leq y) = 1$ if $W_i \leq y$ and $I(W_i \leq y) = 0$ if $W_i > y$. Fix n and y. Then $nF_n(y) \sim$ binomial $(n, F_W(y))$. Thus $E[F_n(y)] = F_W(y)$ and $V[F_n(y)] = F_W(y)[1 - F_W(y)]/n$. By the central limit theorem,

$$\sqrt{n}(F_n(y) - F_W(y)) \xrightarrow{D} N(0, F_W(y)[1 - F_W(y)]).$$

Thus $F_n(y) - F_W(y) = O_P(n^{-1/2})$, and F_n is a reasonable estimator of F_W if the sample size n is large.

Suppose there is data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ collected into an $n \times p$ matrix $\boldsymbol{W}$. Let the statistic $T_n = t(\boldsymbol{W}) = T(F_n)$ be computed from the data. Suppose the statistic estimates $\boldsymbol{\mu} = T(F)$, and let $t(\boldsymbol{W}^*) = t(F_n^*) = T_n^*$ indicate that t was computed from an iid sample from the empirical distribution F_n: a sample $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ of size n was drawn with replacement from the observed sample $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. This notation is used for von Mises differentiable statistical functions in large sample theory. See Serfling (1980, ch. 6). The empirical distribution is also important for the influence function (widely used in robust statistics). The *empirical bootstrap* or *nonparametric bootstrap* or *naive bootstrap* draws B samples of size n from the rows of $\boldsymbol{W}$, e.g., from the empirical distribution of $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. Then T_{jn}^* is computed from the jth bootstrap sample for $j = 1, ..., B$.

Example 5.8. Suppose the data is 1, 2, 3, 4, 5, 6, 7. Then $n = 7$ and the sample median T_n is 4. Using R, we drew $B = 2$ bootstrap samples (samples of size n drawn with replacement from the original data) and computed the sample median $T_{1,n}^* = 3$ and $T_{2,n}^* = 4$.

```
b1 <- sample(1:7,replace=T)
b1
[1] 3 2 3 2 5 2 6
median(b1)
[1] 3
b2 <- sample(1:7,replace=T)
b2
[1] 3 5 3 4 3 5 7
median(b2)
[1] 4
```

The bootstrap has been widely used to estimate the population covariance matrix of the statistic $\text{Cov}(T_n)$, for testing hypotheses, and for obtaining confidence regions (often confidence intervals). An iid sample $T_{1n}, ..., T_{Bn}$ of size B of the statistic would be very useful for inference, but typically we only have one sample of data and one value $T_n = T_{1n}$ of the statistic. Often $T_n = t(\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$, and the bootstrap sample $T_{1n}^*, ..., T_{Bn}^*$ is formed where $T_{jn}^* = t(\boldsymbol{w}_{j1}^*, ..., \boldsymbol{w}_{jn}^*)$. Section 5.3.3 will show that $T_{1n}^* - T_n, ..., T_{Bn}^* - T_n$ is pseudodata for $T_{1n} - \boldsymbol{\mu}, ..., T_{Bn} - \boldsymbol{\mu}$ when n is large.

The *residual bootstrap* is often useful for additive error regression models of the form $Y_i = m(\boldsymbol{x}_i) + e_i = \hat{m}(\boldsymbol{x}_i) + r_i = \hat{Y}_i + r_i$ for $i = 1, ..., n$ where the ith residual $r_i = Y_i - \hat{Y}_i$. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, $\boldsymbol{r} = (r_1, ..., r_n)^T$, and let $\boldsymbol{X}$ be an $n \times p$ matrix with ith row $\boldsymbol{x}_i^T$. Then the fitted values $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i)$ and the residuals are obtained by regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$. Here the errors e_i are iid, and it would be useful to be able to generate B iid samples $e_{1j}, ..., e_{nj}$ from the distribution of e_i where $j = 1, ..., B$. If the $m(\boldsymbol{x}_i)$ were known, then we could form a vector $\boldsymbol{Y}_j$ where the ith element $Y_{ij} = m(\boldsymbol{x}_i) + e_{ij}$ for $i = 1, ..., n$. Then regress $\boldsymbol{Y}_j$ on $\boldsymbol{X}$. Instead, draw samples $r^*_{1j}, ..., r^*_{nj}$ with replacement from the residuals, then form a vector $\boldsymbol{Y}^*_j$ where the ith element $Y^*_{ij} = \hat{m}(\boldsymbol{x}_i) + r^*_{ij}$ for $i = 1, ..., n$. Then regress $\boldsymbol{Y}^*_j$ on $\boldsymbol{X}$.

Example 5.9. For multiple linear regression, $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i$ is written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. (This model is a special case of the multivariate linear regression model in Chapter 12 where there is $m = 1$ response variable Y_i. Outlier resistant methods are discussed in Chapter 14.) Regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ to obtain $\hat{\boldsymbol{\beta}}$, $\boldsymbol{r}$, and $\hat{\boldsymbol{Y}}$ with ith element $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i) = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$. For $j = 1, ..., B$, regress $\boldsymbol{Y}^*_j$ on $\boldsymbol{X}$ to form $\hat{\boldsymbol{\beta}}^*_{1,n}, ..., \hat{\boldsymbol{\beta}}^*_{B,n}$ using the residual bootstrap.

Consider the residual bootstrap, and let $\boldsymbol{r}^W$ denote an $n \times 1$ random vector of elements selected with replacement from the n residuals $r_1, ..., r_n$. Then there are $K = n^n$ possible values for $\boldsymbol{r}^W$. Let $\boldsymbol{r}^W_1, ..., \boldsymbol{r}^W_K$ be the possible values of $\boldsymbol{r}^W$. These values are equally likely, so are selected with probability $= 1/K$. Note that the random vector $\boldsymbol{r}^W$ has a discrete distribution. Then

$$E(\boldsymbol{r}^W_j) = \begin{pmatrix} E(r^*_{1j}) \\ \vdots \\ E(r^*_{nj}) \end{pmatrix}.$$

Now the marginal distribution of r^*_{ij} takes on the n values $r_1, ..., r_n$ with the same probability $1/n$. So each of the n marginal distributions is the empirical distribution of the residuals. Hence $E(r^*_{ij}) = \sum_{i=1}^n r_i/n = \overline{r}$, and $\overline{r} = 0$ for least squares residuals for multiple linear regression when there is a constant in the model. So for least squares, $E(\boldsymbol{r}^W_j) = \boldsymbol{0}$, and $E(\hat{\boldsymbol{\beta}}^*_j) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\hat{\boldsymbol{Y}} + \boldsymbol{r}^W_j) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \hat{\boldsymbol{Y}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \boldsymbol{H}\boldsymbol{Y} =$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_n$$

since $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{H} = \boldsymbol{X}^T$. Here, $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, and $j = 1, ..., B$. Also, the expectation is with respect to the bootstrap distribution where $\hat{\boldsymbol{Y}}$ acts as a constant.

For the (ordinary) least squares estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$, the estimated covariance matrix of $\hat{\boldsymbol{\beta}}_{OLS}$ is $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_{OLS}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Efron (1982, p. 36) noted that for the residual bootstrap, $E(\hat{\boldsymbol{\beta}}^*) = \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_n$ and $\text{Cov}(\hat{\boldsymbol{\beta}}^*) = \dfrac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, where expectations are taken with respect to the bootstrap distribution. The sample covariance matrix of the $\hat{\boldsymbol{\beta}}^*_j$ is estimating $\text{Cov}(\hat{\boldsymbol{\beta}}^*)$ as $B \to \infty$. Hence the residual bootstrap standard error $SE(\hat{\beta}^*_i) \approx \sqrt{\frac{n-p}{n}}\ SE(\hat{\beta}_i)$ for $i = 1, ..., p$ where $\hat{\boldsymbol{\beta}}_{OLS} = \hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$.

Example 5.10. Suppose there is training data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ for the model $\boldsymbol{y}_i = m(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$, and it is desired to predict a future test value $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ and the training data. The model can be fit, and the residual vectors are formed. One method for obtaining a prediction region for $\boldsymbol{y}_f$ is to form the pseudodata $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$ and apply the nonparametric prediction region (5.17) to the pseudodata. See Section 12.3. The residual bootstrap could also be used to make a bootstrap sample $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}^*_1, ..., \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}^*_B$ where the $\hat{\boldsymbol{\epsilon}}^*_j$ are selected with replacement from the residual vectors for $j = 1, ..., B$. As $B \to \infty$, the bootstrap sample will take on the n values $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ (the pseudodata) with probabilities converging to $1/n$ for $i = 1, ..., n$.

Suppose there is a statistic T_n that is an $r \times 1$ vector. Let

$$\overline{T^*} = \frac{1}{B}\sum_{i=1}^{B} T^*_i \quad \text{and} \quad \boldsymbol{S}^*_T = \frac{1}{B-1}\sum_{i=1}^{B}(T^*_i - \overline{T^*})(T^*_i - \overline{T^*})^T \tag{5.19}$$

be the sample mean and sample covariance matrix of the bootstrap sample $T^*_1, ..., T^*_B$ where $T^*_i = T^*_{i,n}$. Fix n, and let $E(T^*_{i,n}) = \boldsymbol{\mu}_n$ and $\text{Cov}(T^*_{i,n}) = \boldsymbol{\Sigma}_n$. For example, using least squares and the residual bootstrap for the multiple linear regression model, $\boldsymbol{\mu}_n = \hat{\boldsymbol{\beta}}$, $\boldsymbol{\Sigma}_n = \frac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Suppose the $T^*_i = T^*_{i,n}$ are iid from some distribution with cdf $\tilde{F}_n$. For example, if $T^*_{i,n} = t(F^*_n)$ where iid samples from F_n are used, then $\tilde{F}_n$ is the cdf of $t(F^*_n)$. With respect to $\tilde{F}_n$, both $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ are parameters, but with respect to F, $\boldsymbol{\mu}_n$ is a random vector and $\boldsymbol{\Sigma}_n$ is a random matrix. For fixed n, by the multivariate central limit theorem,

$$\sqrt{B}(\overline{T^*} - \boldsymbol{\mu}_n) \xrightarrow{D} N_r(\boldsymbol{0}, \boldsymbol{\Sigma}_n) \quad \text{and} \quad \mathrm{B}(\overline{\mathrm{T}^*} - \boldsymbol{\mu}_\mathrm{n})^\mathrm{T}[\boldsymbol{S}^*_\mathrm{T}]^{-1}(\overline{\mathrm{T}^*} - \boldsymbol{\mu}_\mathrm{n}) \xrightarrow{\mathrm{D}} \chi^2_\mathrm{r}$$

as $B \to \infty$.

Remark 5.4. For Examples 5.6, 5.9, and 5.10, the bootstrap works but is expensive compared to alternative methods. For Example 5.6, fix n, then $\overline{T^*} \xrightarrow{P} \boldsymbol{\mu}_n = \overline{\boldsymbol{x}}$ and $\boldsymbol{S}^*_T \xrightarrow{P} (n-1)\boldsymbol{S}/n$ as $B \to \infty$, but using $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ makes more sense. For Example 5.9, using $\hat{\boldsymbol{\beta}}$ and the classical estimated covariance matrix $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ makes more sense than using the bootstrap. For Example 5.10, use the pseudodata instead of the residual bootstrap. For these three examples, it is known how the bootstrap sample behaves as $B \to \infty$. The bootstrap can be very useful when $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_r(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, but it not known how to estimate $\boldsymbol{\Sigma}_A$ without using a resampling method like the bootstrap. The bootstrap may be useful when $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{U}$, but the limiting distribution (the distribution of $\boldsymbol{U}$) is unknown.

Remark 5.5. From Example 5.9, $\text{Cov}(\hat{\boldsymbol{\beta}}^*) = \frac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{n-p}{n}\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ where $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ starts to give good estimates of $\text{Cov}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Sigma}_T$ for many error distributions if $n \geq 10p$ and $T = \hat{\boldsymbol{\beta}}$. For the residual bootstrap with large B, note that $\boldsymbol{S}^*_T \approx 0.95\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ for $n = 20p$ and $\boldsymbol{S}^*_T \approx 0.99\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ for $n = 100p$. Hence we may need $n >> p$ before the $\boldsymbol{S}^*_T$ is a good estimator of $\text{Cov}(T) = \boldsymbol{\Sigma}_T$. The distribution of $\sqrt{n}(T_n - \boldsymbol{\mu})$ is approximated by the distribution of $\sqrt{n}(T^* - T_n)$ or by the distribution of $\sqrt{n}(T^* - \overline{T}^*)$, but n may need to be large before the approximation is good.

Suppose the bootstrap sample mean $\overline{T^*}$ estimates $\boldsymbol{\mu}$, and the bootstrap sample covariance matrix $\boldsymbol{S}^*_T$ estimates $c_n\widehat{\text{Cov}}(T_n) \approx c_n\boldsymbol{\Sigma}_T$ where c_n increases to 1 as $n \to \infty$. Then $\boldsymbol{S}^*_T$ is not a good estimator of $\widehat{\text{Cov}}(T_n)$ until $c_n \approx 1$ ($n \geq 100p$ for OLS $\hat{\boldsymbol{\beta}}$), but the squared Mahalanobis distance $D^{2*}_{\boldsymbol{w}}(\overline{T^*}, \boldsymbol{S}^*_T) \approx D^2_{\boldsymbol{w}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_T)/c_n$ and $D^{2*}_{(U_B)} \approx D^2_{1-\delta}/c_n$. Hence the prediction region method, described below, has a cutoff $D^{2*}_{(U_B)}$ that estimates the cutoff $D^2_{1-\delta}/c_n$. Thus the prediction region method may give good results for much smaller n than a bootstrap method that uses a $\chi^2_{r,1-\delta}$ cutoff when a cutoff $\chi^2_{r,1-\delta}/c_n$ should be used for moderate n.

5.3.2 The Prediction Region Method for Hypothesis Testing

Consider testing $H_0 : \boldsymbol{\mu} = \boldsymbol{c}$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{c}$ where $\boldsymbol{c}$ is a known $r \times 1$ vector. If a confidence region can be constructed for $\boldsymbol{\mu} - \boldsymbol{c}$, then fail to reject H_0 if $\boldsymbol{0}$ is in the confidence region, and reject H_0 if $\boldsymbol{0}$ is not in the confidence region. For example, let $\boldsymbol{\mu} = \boldsymbol{A\beta}$ where $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters, and $\boldsymbol{A}$ is a known full rank $r \times p$ matrix with $1 \leq r \leq p$.

The **prediction region method** makes a bootstrap sample $\boldsymbol{w}_i = \hat{\boldsymbol{\mu}}^*_i - \boldsymbol{c}$ for $i = 1, ..., B$. Make the nonparametric prediction region (5.17) for the $\boldsymbol{w}_i$, and reject H_0 if $\boldsymbol{0}$ is not in the prediction region. As shown below, the prediction region method is a special case of the percentile method, and a special case of bootstrapping a test statistic.

For $r = 1$, the percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta)\rceil$ of the $T^*_{i,n}$ from a bootstrap sample $T^*_{1,n}, ..., T^*_{B,n}$ where the statistic T_n is an estimator of μ based on a sample of size n. Often the n is suppressed in the double subscripts. Here $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.8 \rceil = 8$. Let $T^*_{(1)}, T^*_{(2)}, ..., T^*_{(B)}$ be the order statistics of the bootstrap sample. Then one version of the percentile method discards the largest and smallest $\lceil B\delta/2 \rceil$ order statistics, resulting in an interval $[\hat{L}_B, \hat{R}_B]$ that is a large sample $100(1-\delta)\%$ confidence interval (CI) for μ, and also a large sample $100(1-\delta)\%$ prediction interval (PI) for a future bootstrap value $T^*_{f,n}$.

Olive (2017b, d, 2014: p. 283) recommend using the shorth(c) estimator for the percentile method. The shorth interval tends to be shorter than the interval that deletes the smallest and largest $\lceil B\delta/2 \rceil$ observations $W_i = T^*_{i,n}$ when the W_i do not come from a symmetric distribution. Frey (2013) showed that for large $B\delta$ and iid data, the shorth(k_B) PI has maximum undercoverage $\approx 1.12\sqrt{\delta/B}$.

Consider testing $H_0 : \boldsymbol{\mu} = \boldsymbol{c}$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{c}$, and the statistic $T_i = \hat{\boldsymbol{\mu}} - \boldsymbol{c}$. If $E(T_i) = \boldsymbol{\theta}$ and $\text{Cov}(T_i) = \boldsymbol{\Sigma}_T$ were known, then the squared Mahalanobis distance $D_i^2(\boldsymbol{\theta}, \boldsymbol{\Sigma}_T) = (T_i - \boldsymbol{\theta})^T \boldsymbol{\Sigma}_T^{-1}(T_i - \boldsymbol{\theta})$ would be a natural statistic to use if the percentile $D^2_{1-\delta}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_T)$ was known. The prediction region method bootstraps the squared Mahalanobis distances, forming the bootstrap sample $\boldsymbol{w}_i = T_i^* = \hat{\boldsymbol{\mu}}_i^* - \boldsymbol{c}$ and the squared Mahalanobis distances $D_i^2 = D_i^2(\overline{T^*}, \boldsymbol{S}_T^*) = (T_i^* - \overline{T^*})^T [\boldsymbol{S}_T^*]^{-1}(T_i^* - \overline{T^*})$ where $(\overline{T^*}, \boldsymbol{S}_T^*)$ are the sample mean and sample covariance matrix of $T_1^*, ..., T_B^*$. See (5.19). Then the percentile method that contains the smallest U_B distances is used to get the closed interval $[0, D_{(U_B)}]$. If H_0 is true and $E[\hat{\boldsymbol{\mu}}] = \boldsymbol{c}$, then $\boldsymbol{\theta} = \boldsymbol{0}$. Let $D^2_{\boldsymbol{0}} = \overline{T^*}^T [\boldsymbol{S}_T^*]^{-1} \overline{T^*}$ and fail to reject H_0 if $D_{\boldsymbol{0}} \leq D_{(U_B)}$ and reject H_0 if $D_{\boldsymbol{0}} > D_{(U_B)}$. This percentile method is equivalent to computing the prediction region (5.17) on the $\boldsymbol{w}_i = T_i^*$ and checking whether $\boldsymbol{0}$ is in the prediction region.

Remark 5.6. For $r = 1$, we will use the shorth(c) intervals with $c = \min(B, \lceil 1 - \delta + 1.12\sqrt{\delta/B} \rceil)$. Let $q_B = \min(1 - \delta + 0.05, 1 - \delta + r/B)$ for $\delta > 0.1$. Let $q_B = \min(1 - \delta/2, 1 - \delta + 10\delta r/B)$ for $\delta \leq 0.1$. If $1 - \delta < 0.999$ and $q_B < 1 - \delta + 0.001$, set $q_B = 1 - \delta$. Let

$$c = \lceil B q_B \rceil.$$

Let $D_{(U_B)}$ be the $100 q_B$th percentile of the D_i. We may need $n \geq 50r$ and $B \geq \max(100, n, 50r)$. If d is the model degrees of freedom, we also want $n \geq 20d$. Sometimes much larger n is needed to avoid undercoverage.

Note that the percentile method makes an interval that contains U_B of the scalar-valued T_i^*. The prediction region method makes a hyperellipsoid that contains U_B of the $r \times 1$ vectors $T_i^* = \boldsymbol{w}_i$, and equivalently, makes an interval $[0, D_{(U_B)}]$ that contains U_B of the D_i.

When $r = 1$, a hyperellipsoid is an interval. Suppose the parameter of interest is μ, and there is a bootstrap sample $T_1^*, ..., T_B^*$. Let $a_i = |T_i^* - \overline{T^*}|$. Let $\overline{T^*}$ and S_T^{2*} be the sample mean and variance of the T_i^*. Then the squared Mahalanobis distance $D_\mu^2 = (\mu - \overline{T^*})^2 / S_T^{2*} \leq D^2_{(U_B)}$ is equivalent to $\mu \in$

$[\overline{T^*} - S^*_T D_{(U_B)}, \overline{T^*} + S^*_T D_{(U_B)}] = [\overline{T^*} - a_{(U_B)}, \overline{T^*} + a_{(U_B)}]$, which is an interval centered at $\overline{T^*}$ just long enough to cover U_B of the T_i^*. Hence the prediction region method is a special case of the percentile method if $r = 1$. Note that when $r = 1$, then S^*_T and $D_{(U_B)}$ do not need to be computed.

Bootstrapping test statistics is well known, and the prediction region method is a special case bootstrapping a test statistic using $D_i^2 = D_i^2(\overline{T^*}, \boldsymbol{S}^*_T)$ as the test statistic. See Bickel and Ren (2001).

Example 5.11, Bootstrapping Multiple Linear Regression with the Prediction Region Method. Following Example 5.9, we suggest using the residual bootstrap. If the $\boldsymbol{z}_i = (Y_i, \boldsymbol{x}_i^T)^T$ are iid observations from some population, then a sample of size n can be drawn with replacement from $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$. Then the response and predictor variables can be formed into vector $\boldsymbol{Y}_1^*$ and design matrix $\boldsymbol{X}_1^*$. Then $\boldsymbol{Y}_1^*$ is regressed on $\boldsymbol{X}_1^*$ resulting in the estimator $\hat{\boldsymbol{\beta}}_1^*$. This process is repeated B times resulting in the estimators $\hat{\boldsymbol{\beta}}_1^*, ..., \hat{\boldsymbol{\beta}}_B^*$. If the $\boldsymbol{z}_i$ are the rows of a matrix $\boldsymbol{Z}$, then this nonparametric bootstrap uses the empirical distribution of the $\boldsymbol{z}_i$.

Consider testing $H_0 : \boldsymbol{A\beta} = \boldsymbol{c}$ where $\boldsymbol{A}$ is an $r \times p$ matrix with full rank r and $\boldsymbol{\mu} = \boldsymbol{A\beta}$. To use the prediction region method to perform the test, suppose a bootstrap sample $\hat{\boldsymbol{\beta}}_1^*, ..., \hat{\boldsymbol{\beta}}_B^*$ has been generated. Form the prediction region (5.17) for $\boldsymbol{w}_1 = \boldsymbol{A}\hat{\boldsymbol{\beta}}_1^* - \boldsymbol{c}, ..., \boldsymbol{w}_B = \boldsymbol{A}\hat{\boldsymbol{\beta}}_B^* - \boldsymbol{c}$. If $\boldsymbol{0}$ is in the prediction region, fail to reject H_0, otherwise reject H_0.

Following Seber and Lee (2003, p. 100), the classical test statistic for testing H_0 is

$$F_R = \frac{(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T [MSE \ \ \boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c})}{r},$$

and when H_0 is true, $rF_R \xrightarrow{D} \chi^2_r$ for a large class of error distributions. The sample covariance matrix $\boldsymbol{S}_{\boldsymbol{w}}$ of the $\boldsymbol{w}_i$ is estimating

$$\frac{n-p}{n} MSE \ \ \boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T,$$

and $\overline{\boldsymbol{w}} \approx \boldsymbol{0}$ when H_0 is true. Thus under H_0, the squared distance $D_i^2 = (\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T \boldsymbol{S}_{\boldsymbol{w}}^{-1} (\boldsymbol{w}_i - \overline{\boldsymbol{w}}) \approx$

$$\frac{n}{n-p}(\boldsymbol{A}\hat{\boldsymbol{\beta}}^* - \boldsymbol{c})^T [MSE \ \ \boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}^* - \boldsymbol{c}),$$

and we expect $D^2_{(U_B)} \approx \frac{n}{n-p}\chi^2_{r,1-\delta}$, for large n and B, and small p.

5.3.3 Theory for the Prediction Region Method

When the bootstrap is used, a large sample $100(1-\delta)\%$ confidence region for an $r \times 1$ parameter vector $\boldsymbol{\mu}$ is a set $\mathcal{A}_{n,B}$ such that $P(\boldsymbol{\mu} \in \mathcal{A}_{n,B}) \to 1-\delta$ as $n, B \to \infty$. Assume $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A$ as $n, B \to \infty$ where $\boldsymbol{\Sigma}_A$ and $\boldsymbol{S}_T^*$ are nonsingular $r \times r$ matrices, and T_n is an estimator of $\boldsymbol{\mu}$ such that

$$\sqrt{n}\ \ (T_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{U} \tag{5.20}$$

as $n \to \infty$. Then

$$\sqrt{n}\ \boldsymbol{\Sigma}_A^{-1/2}\ (T_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{\Sigma}_A^{-1/2}\boldsymbol{U} = \boldsymbol{Z},$$

$$n\ (T_n - \boldsymbol{\mu})^T\ \hat{\boldsymbol{\Sigma}}_A^{-1}\ (T_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{Z}^T\boldsymbol{Z} = D^2$$

as $n \to \infty$ where $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of $\boldsymbol{\Sigma}_A$, and

$$(T_n - \boldsymbol{\mu})^T\ [\boldsymbol{S}_T^*]^{-1}\ (T_n - \boldsymbol{\mu}) \xrightarrow{D} D^2 \tag{5.21}$$

as $n, B \to \infty$. Assume the cumulative distribution function (cdf) of D^2 is continuous and increasing in a neighborhood of $D^2_{1-\delta}$ where $P(D^2 \le D^2_{1-\delta}) = 1-\delta$. If the distribution of D^2 is known, then a common bootstrap large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$ is $\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \le D^2_{1-\delta}\}$

$$= \{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \le D^2_{1-\delta}\}. \tag{5.22}$$

Often by a central limit theorem or the multivariate delta method, $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_r(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, and $D^2 \sim \chi^2_r$. Note that $[\boldsymbol{S}_T^*]^{-1}$ could be replaced by $n\hat{\boldsymbol{\Sigma}}_A^{-1}$. Machado and Parente (2005) gave sufficient conditions and references for when $n\boldsymbol{S}_T^*$ is a consistent estimator of $\boldsymbol{\Sigma}_A$.

Bickel and Ren (2001) used $n\hat{\boldsymbol{\Sigma}}_A^{-1}$ instead of $[\boldsymbol{S}_T^*]^{-1}$ and replaced the D^2 cutoff in (5.22) by $D^2_{(k_B)}$ where $D^2_{(k_B)}$ is computed from $D_i^2 = n(T_i^* - T_n)^T\hat{\boldsymbol{\Sigma}}_A^{-1}(T_i^* - T_n)$ for $i = 1, ..., B$. If $n\boldsymbol{S}_T^* = \hat{\boldsymbol{\Sigma}}_A$, the (modified) large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$ is $\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \le D^2_{(U_B)}\}$

$$= \{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \le D^2_{(U_B)}\} \tag{5.23}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - T_n)^T[\boldsymbol{S}_T^*]^{-1}(T_i^* - T_n)$ for $i = 1, ..., B$.

The prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$ is $\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \le D^2_{(U_B)}\}$

$$= \{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - \overline{T}^*) \le D^2_{(U_B)}\} \tag{5.24}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\mu}_0)^T [\boldsymbol{S}_T^*]^{-1} (\overline{T}^* - \boldsymbol{\mu}_0) > D^2_{(U_B)}$. This procedure is basically the one sample Hotelling's T^2 test applied to the T_i^* using $\boldsymbol{S}_T^*$ as the estimated covariance matrix and replacing the $\chi^2_{p,1-\delta}$ cutoff by $D^2_{(U_B)}$. See Section 9.1.

Given (5.20) and (5.21), a sufficient condition for (5.23) to be a confidence region is

$$\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{U}, \tag{5.25}$$

while sufficient conditions for (5.24) to be a confidence region are

$$\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{U}, \tag{5.26}$$

and

$$\sqrt{n}(\overline{T}^* - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{U}. \tag{5.27}$$

(We could replace $\boldsymbol{U}$ by $\boldsymbol{W}$ in (5.26) and (5.27), but $\boldsymbol{W} \sim \boldsymbol{U}$ works.) Note (5.26) and (5.27) follow from (5.25) and (5.20) if $\sqrt{n}(T_n - \overline{T}^*) \xrightarrow{P} \boldsymbol{0}$, so $T_n - \overline{T}^* = o_P(n^{-1/2})$.

Following Bickel and Ren (2001), let $\boldsymbol{\mu} = T(F)$, $T_n = T(F_n)$, and $T^* = T(F_n^*)$ where F_n^* is the empirical cdf of $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$, a sample from F_n using the nonparametric bootstrap. If $\sqrt{n}(F_n - F) \xrightarrow{D} \boldsymbol{z}_F$, a Gaussian random process, and if T is sufficiently smooth (Hadamard differentiable with a well behaved Hadamard derivative $\dot{T}(F)$), then (5.20) and (5.25) hold with $\boldsymbol{U} = \dot{T}(F)\boldsymbol{z}_F$. Note that F_n is a perfectly good cdf "F" and F_n^* is a perfectly good empirical cdf from F_n = "F." Thus if n is fixed, and a sample of size m is drawn with replacement from the empirical distribution, then $\sqrt{m}(T(F_m^*) - T_n) \xrightarrow{D} \dot{T}(F_n)\boldsymbol{z}_{F_n}$. Now let $n \to \infty$ with $m = n$. Then bootstrap theory gives $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \lim_{n\to\infty} \dot{T}(F_n)\boldsymbol{z}_{F_n} = \dot{T}(F)\boldsymbol{z}_F \sim \boldsymbol{U}$.

To justify the prediction region method, assume that (5.20) and (5.25) hold where $\boldsymbol{U} \sim N_r(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$. Use $\boldsymbol{Z}_n \sim AN_r(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ to indicate that a normal approximation is used: $\boldsymbol{Z}_n \approx N_r(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$. Let $T_i^* = T_{i,n}^*$. Then $T_i^* \sim AN_r\left(T_n, \dfrac{\boldsymbol{\Sigma}_A}{n}\right)$. Fix n temporarily and let $\boldsymbol{W}_i = \sqrt{n}(T_i^* - T_n)$. Then with respect to the bootstrap distribution (so conditional on the data), $\boldsymbol{W}_1, ..., \boldsymbol{W}_B$ are iid, and $\sqrt{n}(\overline{T}^* - T_n) = \dfrac{1}{B}\sum_{i=1}^B \boldsymbol{W}_i \sim AN_r\left(\boldsymbol{0}, \dfrac{\boldsymbol{\Sigma}_A}{B}\right)$ is a normal approximation. Hence $\sqrt{nB}(\overline{T}^* - T_n) \sim AN_r(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$. Now unfix n. Since the same normal approximation holds for n and B large (and $AN_r(\mathbf{0}, \boldsymbol{\Sigma}_A)$ does not depend on n or B), it follows that $\overline{T}^* - T_n = o_P(n^{-1/2})$.

The prediction region method should often work if $E(\overline{T}^*) - T_n = o_P(n^{-1/2})$ and the asymptotic covariance matrix of $\sqrt{nB}(\overline{T}^* - T_n)$ is $\boldsymbol{\Sigma}_A$ as $n, B \to \infty$. Following Efron (2014), $\overline{T^*}$ is the bagging or smoothed bootstrap estimator of $\boldsymbol{\mu}$, which often outperforms T_n for inference. See Bühlmann and Yu (2002) and Friedman and Hall (2007) for theory and references for the bagging estimator.

These results suggest that under reasonable conditions, (5.20), (5.25), (5.26), and (5.27) hold: $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{U}$, $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{U}$, $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{U}$, and $\sqrt{n}(\overline{T}^* - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{U}$. Stronger conditions are needed for $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A$. The regularity conditions for the prediction region method are weaker when $r = 1$, since $\boldsymbol{S}_T^*$ does not need to be computed.

The following result is also informative. Let $T_i = T_{i,n}$, and assume $T_1, ..., T_B$ are iid where

$$\frac{n}{B} \sum_{i=1}^{B} (T_i - \boldsymbol{\mu})(T_i - \boldsymbol{\mu})^T \xrightarrow{P} \boldsymbol{\Sigma}_A \quad \text{and} \quad \frac{\mathrm{n}}{\mathrm{B}} \sum_{\mathrm{i=1}}^{\mathrm{B}} (\mathrm{T}_\mathrm{i}^* - \overline{\mathrm{T}}^*)(\mathrm{T}_\mathrm{i}^* - \overline{\mathrm{T}}^*)^\mathrm{T} \xrightarrow{\mathrm{P}} \boldsymbol{\Sigma}_\mathrm{A}.$$

Then

$$\frac{n}{B} \sum_{i=1}^{B} (T_i - \boldsymbol{\mu})(T_i - \boldsymbol{\mu})^T - \frac{n}{B} \sum_{i=1}^{B} (T_i^* - \overline{T}^*)(T_i^* - \overline{T}^*)^T \xrightarrow{P} \mathbf{0}, \tag{5.28}$$

the $r \times r$ matrix of zeroes. The trace is a continuous linear function. Post multiply both sides of (5.28) by $[\boldsymbol{S}_T^*]^{-1}$, and take the trace of both sides to get

$$\frac{n}{B} \sum_{i=1}^{B} (T_i - \boldsymbol{\mu})^T [\boldsymbol{S}_T^*]^{-1} (T_i - \boldsymbol{\mu}) - \frac{n}{B} \sum_{i=1}^{B} (T_i^* - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - \overline{T}^*) \xrightarrow{P} 0. \tag{5.29}$$

Now $(T_i - \boldsymbol{\mu})^T [\boldsymbol{S}_T^*]^{-1} (T_i - \boldsymbol{\mu}) - n(T_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_A^{-1} (T_i - \boldsymbol{\mu}) \xrightarrow{P} 0$. Hence the first sum in (5.29) behaves like a sum of iid nonnegative terms that each converge in distribution to D^2. If n is fixed, then the T_i^* are iid with respect to the bootstrap distribution where $\overline{T}^* \approx E(T_i^*) = \boldsymbol{\mu}_n$ and $\boldsymbol{S}_T^* \approx Cov(T_i^*) = \boldsymbol{\Sigma}_n$ with respect to the bootstrap distribution. Hence the second sum in (5.29) behaves like a sum of iid nonnegative terms with respect to the bootstrap distribution.

The prediction region method will often simulate well even if B is rather small. Figure 2.1 shows 10%, 30%, 50%, 70%, 90%, and 98% prediction regions for a future value of T_f for two multivariate normal statistics. The plotted points are iid $T_1, ..., T_B$. If the T_i^* are iid from the bootstrap distribution, then $Cov(\overline{T}^*) \approx Cov(T)/B \approx \boldsymbol{\Sigma}_A/(nB)$. Consider the 90% region. Suppose many iid samples are generated to produce $\overline{T}^*$. By Theorem 5.2, if $\overline{T}^*$ is in the 90% prediction region with probability near 90%, then the confidence region should give simulated coverage near 90% and the volume of the confidence region should be near that of the 90% prediction region. If $B = 100$, then $\overline{T}^*$ falls in a covering region of the same shape as the prediction region, but centered near T_n and the lengths of the axes are divided by $\sqrt{B}$. Hence if $B = 100$, then the axes lengths are about one tenth of those in Figure 2.1. Hence when T_n falls within the 70% prediction region, the probability that $\overline{T}^*$ falls in the 90% prediction region is near one. If T_n is just within or just without the boundary of the 90% prediction region, $\overline{T}^*$ tends to be just within or just without of the 90% prediction region. Hence the coverage and volume of prediction region confidence region is near that of the nominal coverage 90% and near the volume of the 90% prediction region.

Hence B does not need to be large provided that n and B are large enough so that $S^*_T \approx Cov(T^*) \approx \boldsymbol{\Sigma}_A/n$. If n is large, the sample covariance matrix starts to be a good estimator of the population covariance matrix when $B \geq Jr$ where $J = 20$ or 50. For small r, using $B = 1000$ often led to good simulations, but $B = \max(50r, 100)$ may work well.

Often D^2 is unknown, and we use $D^2_{(U_B)}$ to estimate $D^2_{1-\delta}$ instead of assuming $D^2 \sim \chi^2_r$. For $r = 1$, Efron (2014) used confidence intervals $\overline{T^*} \pm z_{1-\delta} SE(\overline{T^*})$ where $P(Z \leq z_{1-\delta}) = 1 - \delta$ if $Z \sim N(0,1)$. Efron used a delta method estimate of $SE(\overline{T^*})$ to avoid using the computationally expensive double bootstrap. The prediction region method, $\overline{T}^* \pm S^*_T D_{(U_B)}$, avoids assuming a normal limiting distribution and estimates the cutoff using quantiles of the Mahalanobis distances of the $T^*_{i,n}$ from $\overline{T^*}$. The shorth(c) estimator is recommended since it can be much shorter.

The following theorem will provide some intuition for why the percentile method works if T_n has a central limit theorem. Let Z_δ be the 100δth percentile of Z: $P(Z \leq Z_\delta) = \delta$, and let $P(Z_{\delta_L} \leq Z \leq Z_{\delta_U}) = 1 - \delta$. Let $T_{n,\delta}$ be the 100δth percentile of (the sampling distribution of) T_n. Then a population prediction interval for $T_{f,n}$ is $[T_{n,\delta_L}, T_{n,\delta_U}]$ which can be estimated by the sample percentiles $[T_{(c_L)}, T_{(c_U)}]$ when there is an iid sample $T_1, ..., T_n$. The shortest such interval can be estimated by the shorth.

Theorem 5.3. Suppose $r = 1$ and $\sqrt{n}(T_n - \mu) \xrightarrow{D} X$ and $\dfrac{\sqrt{n}(T_n - \mu)}{\sigma} \xrightarrow{D} \dfrac{1}{\sigma} X = W$. If the percentiles are continuity points of the distribution of W, then for each large sample $100(1-\delta)$% PI $[T_{(c_L)}, T_{(c_U)}]$ for $T_{f,n}$, there is a large sample $100(1-\delta)$% CI $\left[T_n - W_{\delta_U}\dfrac{\hat{\sigma}}{\sqrt{n}}, T_n - W_{\delta_L}\dfrac{\hat{\sigma}}{\sqrt{n}}\right]$ for μ with approximately the same length.

Proof. Note that $1 - \delta = P(T_{n,\delta_L} \leq T_n \leq T_{n,\delta_U}] \approx P(T_{(c_L)} \leq T_n \leq T_{(c_U)}) \approx P(W_{\delta_L} \leq \frac{\sqrt{n}(T_n - \mu)}{\sigma} \leq W_{\delta_U}) = P(T_n - W_{\delta_U} \frac{\sigma}{\sqrt{n}} \leq \mu \leq T_n - W_{\delta_L} \frac{\sigma}{\sqrt{n}}) = P(W_{\delta_L} \frac{\sigma}{\sqrt{n}} + \mu \leq T_n \leq W_{\delta_U} \frac{\sigma}{\sqrt{n}} + \mu)$. Hence $T_{n,\delta_L} \approx W_{\delta_L} \frac{\sigma}{\sqrt{n}} + \mu$ and $T_{n,\delta_U} \approx W_{\delta_U} \frac{\sigma}{\sqrt{n}} + \mu$. Thus $T_{(c_U)} - T_{(c_L)} \approx \frac{\sigma}{\sqrt{n}}(W_{\delta_U} - W_{\delta_L}) \approx T_{n,\delta_U} - T_{n,\delta_L}$. □

Theorem 5.3 suggests that the Frey (2013) shorth(c) interval applied to the bootstrap sample estimates the shortest large sample $100(1-\delta)\%$ CI $\left[T_n - W_{\delta_U} \frac{\hat{\sigma}}{\sqrt{n}}, T_n - W_{\delta_L} \frac{\hat{\sigma}}{\sqrt{n}}\right]$ based on the asymptotic pivot. Note that if $Z_i = T_n + \mu - T_i$ for $i = 1, ..., n$, then $P(Z_{(c_L)} \leq \mu \leq Z_{(c_U)}) \approx P(T_n + \mu - T_{n,\delta_U} \leq \mu \leq T_n + \mu - T_{n,\delta_L}) = P(T_{n,\delta_L} \leq T_n \leq T_{n,\delta_U}) = 1 - \delta$. Then the Z_i are centered at T_n with deviations equal to $\mu - T_i$. Note that the distribution of $T_n - \mu$ is the same as the distribution of $T_i - \mu$: $T_i - \mu \overset{D}{=} T_n - \mu$. Now the bootstrap approximation says that the distribution of $T_n - \mu$ can be approximated by the distribution of $T_i^* - T_n$. Thus $T_i - \mu \overset{D}{=} T_n - \mu \approx T_i^* - T_n$, or $T_i^* \approx T_i + T_n - \mu$. If the distribution of $T_n - \mu$ is approximately the same as the distribution of $\mu - T_n$ (asymptotic symmetry), then the percentile method should work. Since $\sqrt{n}(T_n - \mu) \xrightarrow{D} X$, we have $n^\gamma(T_n - \mu) \xrightarrow{D} 0$ if $0 < \gamma < 0.5$. The point mass at 0 is a symmetric distribution, and $n^\gamma(T_i + T_n - \mu) \approx n^\gamma \mu$ for large n.

Remark 5.7. Remark 5.5 suggests that even if the statistic T_n is asymptotically normal so the Mahalanobis distances are asymptotically χ^2_r, the prediction region method can give better results for moderate n by using the cutoff $D^2_{(U_B)}$ instead of the cutoff $\chi^2_{r,1-\delta}$. Theorem 5.2 says that the hyperellipsoidal prediction and confidence regions have exactly the same volume. We compensate for the prediction region undercoverage when n is moderate by using $D^2_{(U_n)}$. If n is large, by using $D^2_{(U_B)}$, the prediction region method confidence region compensates for undercoverage when B is moderate, say $B \geq Jr$ where $J = 20$ or 50. See Remark 5.6. This result can be useful if a simulation with $B = 1000$ or $B = 10000$ is much slower than a simulation with $B = Jr$. The price to pay is that the prediction region method confidence region is inflated to have better coverage, so the power of the hypothesis test is decreased if moderate B is used instead of larger B.

Software. The prediction region method will be used several times in the text, sometimes as an "exploratory test." The *mpack* functions `corboot` and `corbootsim` are used to bootstrap the correlation matrix as described in Section 5.3.5. The function `predrgn` makes the nonparametric prediction region and determines whether $\boldsymbol{x}_f$ is in the region. The function `predreg` also makes the nonparametric prediction region and determines if $\mathbf{0}$ is in the region. For multiple linear regression, the function `regboot` does the residual bootstrap for multiple linear regression, `regbootsim` simulates the residual bootstrap for regression, and the function `rowboot` does the empirical nonparametric bootstrap. The function `vsbootsim` simulates the bootstrap for all subsets variable selection, so needs p small, while `vsbootsim2` simulates the prediction region method for forward selection. The functions `fselboot` and `vselboot` bootstrap the forward selection and all subsets variable selection estimators that minimize C_p. See Examples 5.12 and 5.13.

The functions `rhotboot` and `rhotsim2` are used to bootstrap and simulate a Hotelling's type T^2 test based on the RMVN estimator. See Section 9.1. The functions `rmregboot` and `rmregbootsim` are used to bootstrap and simulate the robust regression estimator `rmreg2`. See Chapter 14. The `shorth3` function computes the shorth(c) intervals with the Frey (2013) correction used when $r = 1$. See Remark 5.6.

5.3.4 Bootstrapping 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. A *model for variable selection* in multiple linear regression can be described by

$$Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = \boldsymbol{\beta}^T\boldsymbol{x} + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + e \tag{5.30}$$

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, $\boldsymbol{\beta}_E = \boldsymbol{0}$ 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

$$Y = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + \boldsymbol{x}_O^T\boldsymbol{\beta}_O + e. \tag{5.31}$$

Suppose that S is a subset of I and that model (5.30) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I \tag{5.32}$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S.

The model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ that uses all of the predictors is called the *full model.* A model $Y = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + e$ that only uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel.* Criteria such as $C_p(I)$ and $AIC(I)$ are often used to select a subset. See Olive and Hawkins (2005) and Olive (2017a, Section 3.4).

Suppose model I is selected after variable selection. Then least squares output for the model $\boldsymbol{Y} = \boldsymbol{X}_I \boldsymbol{\beta}_I + \boldsymbol{e}$ can be obtained, but the least squares output is not correct for inference. In particular, $MSE(I)(\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1}$ is not the correct estimated covariance matrix of $\hat{\boldsymbol{\beta}}_I$. The selected model tends to fit the data too well, so $SE(\hat{\beta}_i)$ from the incorrect estimated covariance matrix tends to be too small. Hence the confidence intervals for β_i are too short, and hypothesis tests reject $H_0 : \beta_i = 0$ too often.

Hastie et al. (2009, p. 57) noted that variable selection is a shrinkage estimator: the coefficients are shrunk to 0 for the omitted variables. Suppose $n \geq 10p$. If $\hat{\boldsymbol{\beta}}_I$ is $k \times 1$, form $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. Then $\hat{\boldsymbol{\beta}}_{I,0}$ is a nonlinear estimator of $\boldsymbol{\beta}$, and the residual bootstrap method can be applied. For example, suppose $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ is formed from model I_{min} that minimizes C_p from some variable selection method such as forward selection, backward elimination, stepwise selection, or all subsets variable selection. Instead of computing the least squares estimator from regressing $\boldsymbol{Y}_i^*$ on $\boldsymbol{X}$, perform variable selection on $\boldsymbol{Y}_i^*$ and $\boldsymbol{X}$, fit the model that minimizes the criterion, and add 0s corresponding to the omitted variables, resulting in estimators $\hat{\boldsymbol{\beta}}_1^*, ..., \hat{\boldsymbol{\beta}}_B^*$ where $\hat{\boldsymbol{\beta}}_i^* = \hat{\boldsymbol{\beta}}_{I_{min},0,i}^*$.

Suppose the variable selection method, such as forward selection or all subsets, produces K models. Let model I_{min} be the model that minimizes the criterion, e.g., $C_p(I)$ or $AIC(I)$. Following Seber and Lee (2003, p. 448) and Nishi (1984), the probability that model I_{min} from C_p or AIC underfits goes to zero as $n \to \infty$. Since there are a finite number of regression models I that contain the true model, and each model gives a consistent estimator $\hat{\boldsymbol{\beta}}_{I,0}$ of $\boldsymbol{\beta}$, the probability that I_{min} picks one of these models goes to one as $n \to \infty$. Hence $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a consistent estimator of $\boldsymbol{\beta}$ under model (5.30).

Note that if $S \subseteq I$, and $\boldsymbol{Y} = \boldsymbol{X}_I \boldsymbol{\beta}_I + \boldsymbol{e}_I$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_k(\boldsymbol{0}, \sigma_I^2 \boldsymbol{W}_I)$ under mild regularity conditions where $n(\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \to \boldsymbol{W}_I$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma_I^2 \boldsymbol{W}_{I,0})$ where the $\boldsymbol{W}_{I,0}$ has a column and row of zeroes added for each variable not in I. Note that $\boldsymbol{W}_{I,0}$ is singular unless I corresponds to the full model. For example, if $p = 3$ and model I uses a constant $x_1 \equiv 1$ and x_3 with

$$\boldsymbol{W}_I = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}, \quad \text{then} \quad \boldsymbol{W}_{\mathrm{I},0} = \begin{bmatrix} W_{11} & 0 & W_{12} \\ 0 & 0 & 0 \\ W_{21} & 0 & W_{22} \end{bmatrix}.$$

Hence it is reasonable to conjecture that $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{U}$ where

$$\boldsymbol{U} = \sum_{i=1}^{K} \pi_i N_p(\boldsymbol{0}, \sigma_{I_i}^2 \boldsymbol{W}_{I_i,0}),$$

$0 \leq \pi_i \leq 1$, $\sum_{i=1}^{K} \pi_i = 1$, and K is the number of subsets I_i that contain S.

Inference techniques for the variable selection model have not had much success. Efron (2014) let $t(\boldsymbol{Z})$ be a scalar-valued statistic, based on all of the data $\boldsymbol{Z}$, that estimates a parameter of interest μ. Form a bootstrap sample $\boldsymbol{Z}_i^*$ and $t(\boldsymbol{Z}_i^*)$ for $i = 1, ..., B$. Then $\tilde{\mu} = s(\boldsymbol{Z}) = \dfrac{1}{B}\sum_{i=1}^{B} t(\boldsymbol{Z}_i^*)$, a "bootstrap smoothing" or "bagging" estimator. In the regression setting with variable selection, $\boldsymbol{Z}_i^*$ can be formed with the nonparametric or residual bootstrap using the full model. The prediction region method can also be applied to $t(\boldsymbol{Z})$. For example, when $\boldsymbol{A}$ is $1 \times p$, the prediction region method uses $\mu = \boldsymbol{A\beta} - c$, $t(\boldsymbol{Z}) = \boldsymbol{A}\hat{\boldsymbol{\beta}} - c$ and $\overline{T^*} = \tilde{\mu}$. Efron (2014) used the confidence interval $\overline{T^*} \pm z_{1-\delta} SE(\overline{T^*})$ which is symmetric about $\overline{T^*}$. The prediction region method uses $\overline{T^*} \pm S_T^* D_{(U_B)}$ which is also a symmetric interval centered at $\overline{T^*}$. If both the prediction region method and Efron's method are large sample confidence intervals for μ, then they have the same asymptotic length (scaled by multiplying by $\sqrt{n}$), since otherwise the shorter interval will have lower asymptotic coverage. Since the prediction region interval is a percentile interval, the shorth(c) interval could have much shorter length than the Efron interval and the prediction region interval if the bootstrap distribution is not symmetric.

The prediction region method can be used for vector-valued statistics and parameters and may not need the statistic to be asymptotically normal. These features are likely useful for variable selection models. Prediction intervals and regions can have higher than the nominal coverage $1 - \delta$ if the distribution is discrete or a mixture of a discrete distribution and some other distribution. In particular, coverage can be high if the $\boldsymbol{w}_i$ distribution is a mixture of a point mass at $\boldsymbol{0}$, and the method checks whether $\boldsymbol{0}$ is in the prediction region. Such a mixture often occurs for variable selection methods. The bootstrap sample for the $W_i = \hat{\boldsymbol{\beta}}_{ij}^*$ can contain many zeroes and be highly skewed if the jth predictor is weak. Then the computer program may fail because $\boldsymbol{S_w}$ is singular, but if all or nearly all of the $\hat{\boldsymbol{\beta}}_{ij}^* = 0$, then there may be strong evidence that the jth predictor is not needed given that the other predictors are in the variable selection method if n and B are large.

As an extreme simulation case, suppose $\hat{\boldsymbol{\beta}}_{ij}^* = 0$ for $i = 1, ..., B$ and for each run in the simulation. Consider testing $H_0 : \beta_j = 0$. Then regardless of the nominal coverage $1 - \delta$, the closed interval [0,0] will contain 0 for each run and the observed coverage will be $1 > 1 - \delta$. Using the open interval (0,0) would give observed coverage 0. Also intervals $[0, b]$ and $[a, 0]$ correctly suggest failing to reject $\beta_j = 0$, while intervals $(0, b)$ and $(a, 0)$ incorrectly suggest rejecting $H_0 : \beta_j = 0$. Hence closed regions and intervals make sense.

Warning: The bootstrap tests for variable selection are *exploratory tests*: **for variable selection**, the prediction region method has not yet been proven to be a large sample test, and it is only conjectured that the shorth intervals are large sample CIs for β_i. The sufficient conditions in Section 5.3.3 needed asymptotic normality. The prediction region method fails if $\boldsymbol{S}_{\boldsymbol{w}}$ is singular. Singularity will occur if a $\hat{\beta}^*_{ij} = 0$ for $j = 1, ..., B$. (Such a result may give strong evidence that the predictor x_i is not needed in the model given the other predictors are in the model.) Singularity is likely if many $\beta_i = 0$.

Example 5.12. Cook and Weisberg (1999a, pp. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. Let the response variable be the logarithm $\log(M)$ of the *muscle mass*, 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. Inference for the full model is shown below along with the shorth(c) nominal 95% confidence intervals for β_i computed using the nonparametric and residual bootstraps. As expected, the residual bootstrap intervals are close to the classical least squares confidence intervals $\approx \hat{\beta}_i \pm 2SE(\hat{\beta}_i)$.

The minimum C_p model from all subsets variable selection uses a constant, H, and $\log(S)$. The shorth(c) nominal 95% confidence intervals for β_i using the residual bootstrap are shown. Note that the interval for H is right skewed and contains 0 when closed intervals are used instead of open intervals. The least squares output is also shown, but should only be used for inference if the model was selected before looking at the data.

```
    large sample full model inference
   Est.    SE    t    Pr(>|t|)  nparboot   95% shorth CI
i -1.249 0.838 -1.49 0.14 [-2.93,-0.048][-3.138,0.194]
L -0.001 0.002 -0.28 0.78 [-0.005,0.003][-0.005,0.004]
W  0.130 0.374  0.35 0.73 [-0.384,0.827][-0.555,0.971]
H  0.008 0.005  1.50 0.14 [-0.002,0.018][-0.003,0.017]
S  0.640 0.169  3.80 0.00 [ 0.188,1.001][ 0.276,0.955]
output and shorth intervals for the min Cp submodel
      Est.      SE       t      Pr(>|t|)   95% shorth CI
int   -0.9573   0.1519 -6.3018 0.0000   [-2.769, 0.460]
L      0                                [-0.004, 0.004]
W      0                                [-0.595, 0.869]
H      0.0072   0.0047  1.5490 0.1254   [ 0.000, 0.016]
S      0.6530   0.1160  5.6297 0.0000   [ 0.324, 0.913]
```

It was expected that $\log(S)$ may be the only predictor needed, along with a constant, since $\log(S)$ and $\log(M)$ are both log(mass) measurements and likely highly correlated. Hence we want to test $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ with the I_{min} model selected by all subsets variable selection. (Of course, this test would be easy to do with the full model using least squares theory.) Then $H_0 : \boldsymbol{A\beta} = (\beta_2, \beta_3, \beta_4)^T = \boldsymbol{0}$. Using the prediction region method with the full model gave an interval [0,2.930] with $D_{\mathbf{0}} = 1.641$. Note that $\sqrt{\chi^2_{3,0.95}} = 2.795$. So fail to reject H_0. Using the prediction region method with the I_{min} variable selection model had $[0, D_{(U_B)}] = [0, 3.293]$ while $D_{\mathbf{0}} = 1.134$. So fail to reject H_0. The R code used to produce the above output is shown below.

```
library(leaps)
y <- log(mussels[,5]); x <- mussels[,1:4]
x[,4] <- log(x[,4]); x[,2] <- log(x[,2])
out <- regboot(x,y,B=1000)
tem <- rowboot(x,y,B=1000)
outvs <- vselboot(x,y,B=1000) #get bootstrap CIs,
apply(out$betas,2,shorth3);
apply(tem$betas,2,shorth3);
apply(outvs$betas,2,shorth3)
ls.print(outvs$full)
ls.print(outvs$sub)
#test if beta_2 = beta_3 = beta_4 = 0
Abeta <- out$betas[,2:4]
#prediction region method with residual bootstrap
predreg(Abeta)
Abeta <- outvs$betas[,2:4]
#prediction region method with Imin
predreg(Abeta)
```

Example 5.13. Consider the Gladstone (1905) data set where the response variable is *brain weight* and the predictor variables are as in Example 4.2. Output is shown below for the full model and the bootstrapped minimum C_p forward selection estimator. Note that the shorth intervals for *length* and *sex* are quite long. These variables are often in and often deleted from the bootstrap forward selection. Model I_I is the model with the fewest predictors such that $C_P(I_I) \leq C_P(I_{min}) + 1$. For this data set, $I_I = I_{min}$. The bootstrap CIs differ due to different random seeds.

```
large sample full model inference for Ex. 5.12
       Estimate   SE      t    Pr(>|t|)  95% shorth CI
Int   -3021.255 1701.070 -1.77 0.077 [-6549.8,322.79]
age      -1.656    0.314 -5.27 0.000 [ -2.304,-1.050]
breadth  -8.717   12.025 -0.72 0.469 [-34.229,14.458]
cephalic 21.876   22.029  0.99 0.322 [-20.911,67.705]
circum    0.852    0.529  1.61 0.109 [ -0.065, 1.879]
headht    7.385    1.225  6.03 0.000 [  5.138, 9.794]
height   -0.407    0.942 -0.43 0.666 [ -2.211, 1.565]
len      13.475    9.422  1.43 0.154 [ -5.519,32.605]
sex      25.130   10.015  2.51 0.013 [  6.717,44.19]
output and shorth intervals for the min Cp submodel
```

```
       Estimate   SE      t    Pr(>|t|) 95% shorth CI
Int   -1764.516  186.046 -9.48 0.000 [-6151.6,-415.4]
age      -1.708    0.285 -5.99 0.000 [ -2.299,-1.068]
breadth   0                          [-32.992, 8.148]
cephalic  5.958    2.089  2.85 0.005 [-10.859,62.679]
circum    0.757    0.512  1.48 0.140 [  0.000, 1.817]
headht    7.424    1.161  6.39 0.000 [  5.028, 9.732]
height    0                          [ -2.859, 0.000]
len       6.716    1.466  4.58 0.000 [  0.000,30.508]
sex      25.313    9.920  2.55 0.011 [  0.000,42.144]
output and shorth for I_I model
       Estimate   SE      t    Pr(>|t|) 95% shorth CI
Int   -1764.516  186.046 -9.48 0.000 [-6104.9,-778.2]
age      -1.708    0.285 -5.99 0.000 [ -2.259,-1.003]
breadth   0                          [-31.012, 6.567]
cephalic  5.958    2.089  2.85 0.005 [ -6.700,61.265]
circum    0.757    0.512  1.48 0.140 [  0.000, 1.866]
headht    7.424    1.161  6.39 0.000 [  5.221,10.090]
height    0                          [ -2.173, 0.000]
len       6.716    1.466  4.58 0.000 [  0.000,28.819]
sex      25.313    9.920  2.55 0.011 [  0.000,42.847]
```

The *R* code used to produce the above output is shown below. The last four commands are useful for examining the variable selection output.

```
x<-cbrainx[,c(1,3,5,6,7,8,9,10)]
y<-cbrainy
library(leaps)
out <- regboot(x,y,B=1000)
outvs <- fselboot(x,cbrainy) #get bootstrap CIs,
apply(out$betas,2,shorth3)
apply(outvs$betas,2,shorth3)
ls.print(outvs$full)
ls.print(outvs$sub)
outvs <- modIboot(x,cbrainy) #get bootstrap CIs,
apply(outvs$betas,2,shorth3)
ls.print(outvs$sub)
tem<-regsubsets(x,y,method="forward")
tem2<-summary(tem)
tem2$which
tem2$cp
```

A small simulation study was done in *R* using $B = \max(1000, n, 20p)$ and 5000 runs. The regression model used $\boldsymbol{\beta} = (1, 1, 0, 0)^T$ with $n = 100$, $p = 4$, and various zero mean iid error distributions. The design matrix $\boldsymbol{X}$ consisted of iid N(0,1) random variables. Hence the full model least squares confidence intervals for β_i should have length near $2t_{96,0.975}\sigma/\sqrt{n} \approx 2(1.96)\sigma/10 = 0.392\sigma$ when the iid zero mean errors have variance σ^2. The simulation computed the shorth(c) interval for each β_i and used the prediction region method to test $H_0 : \beta_3 = \beta_4 = 0$. See Remark 5.6. The nominal coverage was 0.95 with $\delta = 0.05$. Observed coverage between 0.94 and 0.96 would suggest coverage is close to the nominal value.

The regression models used the residual bootstrap on the full model least squares estimator and on the all subsets variable selection estimator for the model I_{min}. The residuals were from least squares applied to the full model in both cases. Results are shown for when the iid errors $e_i \sim N(0, 1)$. Table 5.3 shows two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The term "reg" is for the full model regression, and the term "vs" is for the all subsets variable selection. The column for the "test" gives the length and coverage = P(fail to reject H_0) for the interval $[0, D_{(U_B)}]$ where $D_{(U_B)}$ is the cutoff for the confidence region. The volume of the confidence region will decrease to 0 as $n \to \infty$. The cutoff will often be near $\sqrt{\chi^2_{r,0.95}}$ if the statistic T is asymptotically normal. Note that $\sqrt{\chi^2_{2,0.95}} = 2.448$ is very close to 2.449 for the full model regression bootstrap test. The coverages were near 0.95 for the regression bootstrap on the full model. For I_{min}, the coverages were near 0.95 for β_1 and β_2, but higher for the other 3 tests since zeroes often occurred for $\hat{\beta}^*_j$ for $j = 3, 4$. The average lengths and coverages were similar for the full model and all subsets variable selection I_{min} for β_1 and β_2, but the lengths were shorter for I_{min} for β_3 and β_4. Since the predictor variables are iid, they are nearly orthogonal. Hence the active variables with nonzero coefficients should have $\hat{\beta}_i$ that are similar for the least squares models that contain the active predictors. The full model contains the active predictors and the probability that I_{min} contains the active predictors goes to 1 as $n \to \infty$.

Table 5.3 Bootstrapping Regression and Variable Selection

model	cov/len	β_1	β_2	β_3	β_4	test
reg	cov	0.9496	0.9430	0.9440	0.9454	0.9414
	len	0.3967	0.3996	0.3997	0.3997	2.4493
vs	cov	0.9482	0.9486	0.9974	0.9974	0.9896
	len	0.3965	0.3990	0.3241	0.3257	2.6901

The *R* code for the simulation is shown below.

```
regbootsim(nruns=5000) #takes a while
library(leaps)
vsbootsim(nruns=5000)  #takes a long while
vsbootsim2(nruns=5000) #bootstraps forwards selection
```

5.3.5 Bootstrapping the Correlation Matrix

Table 5.4 Bootstrapping the Correlation Matrix

n	ψ	cov/len	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	test
100	0	cov	0.943	0.939	0.942	0.937	0.940	0.941	0.848
		len	0.391	0.391	0.391	0.391	0.392	0.392	3.549
400	0	cov	0.944	0.948	0.943	0.946	0.950	0.952	0.923
		len	0.199	0.199	0.199	0.199	0.199	0.199	3.559
400	0.03	cov	0.950	0.950	0.948	0.949	0.948	0.951	0.441
		len	0.198	0.198	0.198	0.198	0.198	0.198	3.558
400	0.1	cov	0.947	0.949	0.952	0.949	0.952	0.951	0.000
		len	0.190	0.190	0.189	0.190	0.189	0.189	3.561

Larger sample sizes n are needed as r increases. Section 5.2 suggested that for iid elliptically contoured data $\boldsymbol{x}_i$ where $\boldsymbol{x}_i$ is $p \times 1$, the nonparametric prediction region (5.17) coverage for a future value $\boldsymbol{x}_f$ started to get close to the nominal coverage when $n \geq 20p$, but volume ratios needed $n \geq 50p$. Hence we may need $B \geq 50r$ for the volume of the confidence region to be good. Remark 5.6 suggests that the bootstrap may need $n >> r$ for many problems: if d is the model degrees of freedom, we may need $n \geq \max(50r, 20d)$ and $B \geq \max(100, 50r)$ if the test statistic has an approximate multivariate normal distribution. Sample sizes may need to be much larger for other limiting distributions.

Consider testing whether correlations in a correlation matrix are 0. There are $r = p(p-1)/2$ correlations $\rho_{i,j} = cor(X_i, X_j)$ where $i < j$. There are better ways to do this test than the prediction region method, but large sample sizes tend to be needed when the raw correlations are used. (A graphical technique is given in Remark 5.8 after Problem 5.10.)

The simulation simulated iid data $\boldsymbol{w}$ with $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{w}$ and $\boldsymbol{A}_{ij} = \psi$ for $i \neq j$ and $\boldsymbol{A}_{ii} = 1$. Hence $\text{cor}(X_i, X_j) = [2\psi + (p-2)\psi^2]/[1 + (p-1)\psi^2]$. Let $\boldsymbol{\mu} = (\rho_{12}, ..., \rho_{1p}, \rho_{23}, ..., \rho_{2p}, ..., \rho_{p-1,p})^T$.

Table 5.4 shows the results for multivariate normal data with $p = 4$ so $r = 6$ for testing $H_0 : \boldsymbol{\mu} = \boldsymbol{0}$. The nominal coverage was 0.95. For $n = 100$ and $\psi = 0$, the test failed to reject H_0 85% of the time, but 92% of the time for $n = 400$. Note that $\sqrt{\chi^2_{6,0.95}} = 3.548$. With $n = 400$ and $\psi > 0$, for the test the coverage $= 1 -$ power. For $\psi = 0.03$, the simulated power was 0.56, but 1.0 for $\psi = 0.1$. Some *R* code is shown below.

```
corbootsim(type=1,n=100,nruns=5000)
corbootsim(type=1,n=400,nruns=5000) #takes a while
corbootsim(type=1,n=400,psi=0.03,nruns=5000)
corbootsim(type=1,n=400,psi=0.1,nruns=5000)
```

5.4 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 $100(1-\delta)\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n) \xrightarrow{P} 1-\delta$ where $0 < \delta < 1$.

2) The classical $100(1-\delta)\%$ large sample prediction region is $\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi^2_{p,1-\delta}\}$ and works well if n is large and the data are iid MVN.

3) Let $q_n = \min(1-\delta+0.05, 1-\delta+p/n)$ for $\delta > 0.1$ and $q_n = \min(1-\delta/2, 1-\delta+10\delta p/n)$, otherwise. If $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. 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 $100(1-\delta)\%$ prediction regions if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the $100q_n$th 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 $D_{(U_n)}(\overline{\boldsymbol{x}}, \boldsymbol{S})$. Cases in the semiparametric region correspond to points below the horizontal line corresponding to $D_{(U_n)}(T_{RMVN}, \boldsymbol{C}_{RMVN})$ 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 $100(1-\delta)\%$ prediction region (if $D_{1-\delta}$ is a continuity point of the cdf of D) for distributions with a nonsingular covariance matrix, although large sample prediction regions with smaller volume may exist.

5) Suppose m independent large sample $100(1-\delta)\%$ 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 $(m, 1-\delta_n)$ where $1-\delta_n$ is the true coverage and $1-\delta_n \to 1-\delta$ as $n \to \infty$. Simulation can be used to see if the true or actual coverage $1-\delta_n$ is close to the nominal coverage $1-\delta$. A prediction region with $1-\delta_n < 1-\delta$ is liberal, and a region with $1-\delta_n > 1-\delta$ 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, we want $n \geq 10p$ for the nonparametric prediction region and $n \geq 20p$ for the semiparametric prediction region.

7) Consider testing $H_0 : \boldsymbol{\mu} = \boldsymbol{c}$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{c}$ where $\boldsymbol{c}$ is a known $r \times 1$ vector. The **prediction region method** makes a bootstrap sample $\boldsymbol{w}_i = \hat{\boldsymbol{\mu}}_i^* - \boldsymbol{c}$ for $i = 1, ..., B$. Make the nonparametric prediction region (5.17) for the $\boldsymbol{w}_i$, and reject H_0 if $\boldsymbol{0}$ is not in the prediction region.

5.5 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 toward 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) provided references for tests for multivariate normality while Bianco et al. (2017), Koltchinskii and Li (1998), and Manzotti et al. (2002) gave references for tests for elliptically contoured distributions.

There are few practical competitors for the Olive (2013a) prediction regions in Section 5.2 if p is larger than 2. The *mpack* function `ddplot4` can be used to make plots similar to Figures 5.10 and 5.11. Another use of the DD plot is to display the nonparametric and semiparametric prediction regions. 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 $p > 2$. Lei et al. (2013) estimated highest density prediction regions using nonparametric kernel density estimators, and the method may work well for very small p. Similar methods are used for discriminant analysis. See Silverman (1986, pp. 120–130). The multivariate Chebyshev's inequality is due to Chen (2011).

Section 5.3 followed Olive (2017b, d) closely. Good references for the bootstrap include Efron (1982) and Efron and Tibshirani (1993). Janssen and Pauls (2003) and Mammen (1992) suggested that the bootstrap works if there is a central limit theorem for the statistic T_n. Also see Beran (1988), Bickel and Freedman (1981), Horowitz (2001), Machado and Parente (2005), and MacKinnon (2009). The shorth interval given by Remark 5.6 is a practical implementation of the Hall (1988) shortest bootstrap interval based on all bootstrap samples.

We want the bootstrap to produce pseudodata that resembles the data actually collected. For bootstrapping multiple linear regression ($\boldsymbol{Y} = \boldsymbol{X\beta}+\boldsymbol{e}$) methods with $n \geq 10p$, we suggest using the residual bootstrap using full model OLS residuals since OLS is well behaved for a large class of error distributions (see, e.g., Theorem 12.7). When $n >> p$, we conjecture that the prediction region method, using the residual bootstrap with the full model OLS residuals, can be used to bootstrap several methods, including forward selection. Inference (such as prediction intervals and bootstrap hypothesis tests) for other variable selection methods such as lasso, partial least squares, principal component regression, and ridge regression, is given in Pelawa Watagoda (2017), Olive (2017c), and Pelawa Watagoda and Olive (2017).

For survival regression, often the response and predictors $\boldsymbol{x}_i$ are measurements taken on the ith person. Hence the nonparametric bootstrap combined with the prediction region method may be useful to examine forward selection or backward elimination variable selection that minimizes AIC.

Let $T_n = T_{1,n}$ where $\sqrt{n}(T_n - \boldsymbol{\mu}) \overset{D}{\rightarrow} \boldsymbol{U}$, and suppose there was an iid sample $T_{1,n}, ..., T_{B,n}$. Then standard inference techniques could be used to examine how the statistic T_n behaves. Usually there is only one sample and one value of the statistic T_n, but if the empirical distribution is well behaved, and if the statistic T_n is sufficiently smooth, then a bootstrap sample of the statistic $T_1^*, ..., T_B^*$ is useful under regularity conditions: $T_1^* - T_n, ..., T_B^* - T_n$ is pseudodata for $T_{1,n} - \boldsymbol{\mu}, ..., T_{B,n} - \boldsymbol{\mu}$, and applying the Olive (2013a) large sample $100(1-\delta)\%$ prediction region to the $T_1^*, ..., T_B^*$ results in a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$. If T_n is asymptotically normal, then under regularity conditions, the large sample confidence region and equivalent hypothesis test are closely related to applying the Hotelling's T^2 test statistic and confidence region to the $T_1^*, ..., T_B^*$. See Section 9.1.

A technique similar to the prediction region method can be used to estimate the $100(1-\delta)\%$ Bayesian credible region for $\boldsymbol{\theta}$. Generate $B = \max(100000, n)$ values of $\boldsymbol{\theta}_i$ from the posterior distribution, and compute the prediction region (5.17). See Olive (2017b). Olive (2014, p. 364) used the shorth estimator to estimate Bayesian credible intervals.

5.6 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

5.1*. If X and Y are random variables, show that

$$\mathrm{Cov(X,Y) = [Var(X+Y) - Var(X-Y)]/4.}$$

R 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, e.g., *ddplot*, will display the code for the function. Use the `args` command, e.g., *args(ddplot)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/mrsashw.txt) into *R*.

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? (Right click *Stop* once on the plot.)

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.* (Right click *Stop* once on the plot.)

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 region uses the $100(1 - \delta + p/n) =$ 84.6th percentile. 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. (Right click *Stop* once on the plot.)

5.9. Type the *R* command predsim() and paste the output into *Word.*

This program 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, and mcvr count 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 tend to be 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 $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_0$ for known $\boldsymbol{\Sigma}_0$ is to plot $D_i(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i(\overline{\boldsymbol{x}}, \boldsymbol{\Sigma}_0)$ for $i = 1, ..., n$. If $n \geq 10p$ and H_0 is true, then the plotted points in the DD plot should start to cluster tightly about the identity line.

a) A test for sphericity is a test of $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = \sigma^2 \boldsymbol{I}_p$ for some unknown constant $\sigma^2 > 0$. Make a "D^2 plot" of $D_i^2(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i^2(\overline{\boldsymbol{x}}, \boldsymbol{I}_p)$. If $n \geq 10p$ and H_0 is true, then the plotted points in the D^2 plot should cluster tightly about the line through the origin with slope σ^2. 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 we 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, consider 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 $n_i \geq 10p$ and H_0 is true, what line will the plotted points cluster about in each of the k DD plots? (See Equation (8.2) for $\boldsymbol{S}_{pool}$.)

Remark 5.8. Lots of other diagnostic DD plots can be made. Suppose known parts of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ are hypothesized to be $\mathbf{0}$. Let $\boldsymbol{S}_Z$ be the sample covariance matrix with the known parts set to $\mathbf{0}$. Then plot $D_i(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i(\overline{\boldsymbol{x}}, \boldsymbol{S}_Z)$. For example, a diagnostic for $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = diag(\boldsymbol{\Sigma}_{11}, ..., \boldsymbol{\Sigma}_{kk})$ where the $\boldsymbol{\Sigma}_{ii}$ are unknown block matrices is the above plot with $\boldsymbol{S}_Z = diag(\boldsymbol{S}_{11}, ..., \boldsymbol{S}_{kk})$. A diagnostic for $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = diag(\sigma_{11}, ..., \sigma_{pp})$ where the σ_{ii} are unknown would use $\boldsymbol{S}_Z = diag(s_{11}, ..., s_{pp})$ if $\boldsymbol{S} = (s_{ij})$. Another diagnostic would check whether the population correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}} = \boldsymbol{I}_p$. See the following paragraph.

Similar diagnostic DD plots can be made for the population correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}}$ where scaled data $\boldsymbol{z}_i$ is used in the D_i such that the sample mean of the scaled data is $\overline{\boldsymbol{z}} = \mathbf{0}$ and the sample covariance matrix of the scaled data is $\boldsymbol{S}_{\boldsymbol{z}} = \boldsymbol{R} = (r_{ij})$. If the data matrix is x with rows $\boldsymbol{x}_i^T$, then the R command

```
z <- scale(x)
```

will make a data matrix z with rows $\boldsymbol{z}_i^T$. For example, consider $H_0 : \boldsymbol{\rho}_{\boldsymbol{x}} = \boldsymbol{\rho}_0 = (\rho_{ij})$ where $\rho_{ij} = \rho$ for $i \neq j$ where $-1 < \rho < 1$ is unknown, and $\rho_{ii} = 1$ for $i = 1, ..., p$. Let $\hat{\rho}$ be the average of the r_{ij} where $i < j$. Let $\boldsymbol{R}_r = (p_{ij})$ where $p_{ij} = \hat{\rho}$ for $i \neq j$ and $p_{ii} = 1$ for $i = 1, ..., p$. Then make a DD plot of $D_i(\mathbf{0}, \boldsymbol{R})$ versus $D_i(\mathbf{0}, \boldsymbol{R}_r)$.

The RMVN matrix $\boldsymbol{C}_{RMVN}$ could be used in place of $\boldsymbol{S}$ in some of the plots if $\boldsymbol{C}_{RMVN} \xrightarrow{P} c\boldsymbol{\Sigma}_{\boldsymbol{x}}$ for some constant $c > 0$. Then for some of the plots, the plotted points might scatter about some line through the origin instead of the identity line.

Chapter 6
Principal Component Analysis

This chapter considers classical and robust principal component analysis (PCA). Principal component analysis 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 $\boldsymbol{\rho}$ defined below is the correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}}$ 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* $\boldsymbol{\rho} = \big(\rho_{ij}\big)$ where

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

6.1 Introduction

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 positive definite dispersion matrix $\boldsymbol{\Sigma}$ has eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_6

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 $\boldsymbol{e}_i = \boldsymbol{g}_i/\|\boldsymbol{g}_i\|$ is also an eigenvector corresponding to λ_i where $\boldsymbol{g}_i = a_j\boldsymbol{e}_j + a_k\boldsymbol{e}_k$ 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. We want $n \geq 10p$ for classical PCA and $n \geq 20p$ 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 square $p \times p$ matrix $\boldsymbol{A}$ is the sum of the diagonal elements of $\boldsymbol{A}$, and the trace is also the sum of the eigenvalues of $\boldsymbol{A}$: $\text{trace}(\boldsymbol{A}) = tr(\boldsymbol{A}) = \sum_{i=1}^p \boldsymbol{A}_{ii} = \sum_{i=1}^p \lambda_i$. 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 $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \cdots \ge \hat{\lambda}_p$. 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 k 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, this quantity is called the "*proportion of the variance explained*" by the first k principal components. 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 jth case are $Y_{ji} = \boldsymbol{e}_i^T \boldsymbol{x}_j$ for $i = 1, ..., p$. Hence $Z_{ji} = \hat{Y}_{ji}$.

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 $\boldsymbol{u}_k$ corresponds to the kth principal component: the n random variables $\hat{\boldsymbol{e}}_k^T \boldsymbol{x}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{e}}_k$ for $i = 1, ..., n$ are the data $\boldsymbol{x}_i$ projected onto a line in the direction of the kth eigenvector $\hat{\boldsymbol{e}}_k$.

Definition 6.3. A *biplot* is a plot of the jth principal component versus the kth principal component, especially the first versus the second principal component where the plotted points are $(Z_{i1}, Z_{i2}) = (\hat{\boldsymbol{e}}_1^T \boldsymbol{x}_i, \hat{\boldsymbol{e}}_2^T \boldsymbol{x}_i)$. The classical biplot uses $i = 1, ..., n$; or $\boldsymbol{u}_j$ versus $\boldsymbol{u}_k$; while the robust biplot uses cases in some set U. Let $\hat{\boldsymbol{e}}_j = (\hat{e}_{1j}, \hat{e}_{2j}, ..., \hat{e}_{pj})^T$. Then $\hat{e}_{mj}$ is called the *loading* of the mth variable on the jth principal component. In a biplot, an arrow with the mth variable name is the vector from the origin $(0,0)^T$ to the loadings $(\hat{e}_{mj}, \hat{e}_{mk})^T$. 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 jth principal component is large and positive, while the loading with the kth principal component is small and negative.

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 $\hat{\boldsymbol{\Sigma}}$. See Theorem 2.4. These axes are a rotation of the usual axes about the origin.

If $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}$, then the definition of the proportion of the variance explained may make little sense if the variables are measured on different scales. Assume the population covariance matrix is $\boldsymbol{I}_2$. 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 $\{\boldsymbol{x} | D^2_{\boldsymbol{x}} \leq h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leq h^2\}$, 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 distribution with continuous decreasing g. 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 principal 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 from the projection on the major axis, the second principal component is from the projection on the next longest axis, ..., the pth principal component is from 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 et al. (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 $\|\boldsymbol{r}_i(k)\|^2 = \boldsymbol{r}_i(k)^T \boldsymbol{r}_i(k) = Z_{i,k+1}^2 + \cdots + Z_{i,p}^2$.

Suppose $\boldsymbol{S}$ or $\boldsymbol{R}$ is used as the as the dispersion matrix and that $T = \mathbf{0}$ so the hyperellipsoid 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.

Consider the population PCA. When the covariance matrix is used, the first principal component $\boldsymbol{e}_1^T \boldsymbol{x}$ is the linear combination $\boldsymbol{g}_1^T \boldsymbol{x}$ that maximizes $\text{Var}(\boldsymbol{g}_1^T \boldsymbol{x})$ 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 $\text{Var}(\boldsymbol{g}_j^T \boldsymbol{x})$ subject to $\boldsymbol{g}_j^T \boldsymbol{g}_j = 1$ and $\text{Cov}(\boldsymbol{g}_j^T \boldsymbol{x}, \boldsymbol{g}_k^T \boldsymbol{x}) = 0$ for $k < j$. This result can be proved using Theorem 1.1. Hence PCA is a special case of the generalized eigenvalue problem with $\boldsymbol{A} = \boldsymbol{B} = \boldsymbol{\Sigma}$ and $\boldsymbol{C} = \boldsymbol{I}_p$. The classical PCA uses $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{x})$ or the correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}}$.

Dimension reduction involves using the first k principal components to approximate the data matrix without losing much important information. We want the proportion of the trace explained by the first k principal components to be higher than 0.8, 0.9, or 0.95. The scree plot is useful for estimating k.

Definition 6.4. A *scree plot* is a plot of component number versus eigenvalue.

Rule of thumb 6.2. The value of k should be such that

$$\frac{\sum_{i=1}^k \hat{\lambda}_i}{\sum_{i=1}^p \hat{\lambda}_i} \geq c$$

where $c = 0.9$ to explain the structure of the dispersion matrix and $c = 0.95$ if the k principal components $Z_1, ..., Z_k$ are to be used instead of the p variables $X_1, ..., X_p$ in a statistical method such as regression or discriminant analysis.

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, pp. 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 $\boldsymbol{\rho}_{\boldsymbol{x}}$ 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, pp. 352–353), some software that uses $\boldsymbol{S}$ centers the data by using $\boldsymbol{u}_j = \boldsymbol{x}_j - \overline{\boldsymbol{x}}$. Centering does not change $\boldsymbol{S}$ (or $\boldsymbol{R}$), but makes the ith principal component equal to $\hat{\boldsymbol{e}}_i^T \boldsymbol{u}_j = \hat{\boldsymbol{e}}_i^T (\boldsymbol{x}_j - \overline{\boldsymbol{x}})$ for observation $\boldsymbol{x}_j$.

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, e.g., $X_p \approx \sum_{i=1}^{p-1} c_i X_i$. 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

Classical PCA is affected by outliers. If the clean data forms a big cluster and the distant outliers form another cluster, then often $\hat{\boldsymbol{e}}_1$ corresponding to the first principal component is on a line passing through the clean data and the cluster of outliers = the major axis of the covering hyperellipsoid based on $(\overline{\boldsymbol{x}}, \boldsymbol{S})$. Good robust methods, like RPCA described below, can give good results in the presence of certain types of outlier configurations.

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 U, using either the sample covariance matrix $\boldsymbol{C}_U = \boldsymbol{S}_U$ or the sample correlation matrix $\boldsymbol{R}_U$ computed from the cases in U. See Definition 4.11 and Section 4.6. 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 $\boldsymbol{D}\text{Cov}(\boldsymbol{x})\boldsymbol{D} = \boldsymbol{\rho}_{\boldsymbol{x}}$, respectively, where $\boldsymbol{D} = \text{diag}(1/\sqrt{\sigma}_{11}, ..., 1/\sqrt{\sigma}_{\text{pp}})$ and the σ_{ii} are the diagonal entries of $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma_x} = c_X \boldsymbol{\Sigma}$. 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 $\hat{\lambda}_1(\hat{\boldsymbol{A}}) \geq \hat{\lambda}_2(\hat{\boldsymbol{A}}) \geq \cdots \geq \hat{\lambda}_p(\hat{\boldsymbol{A}})$.

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 et al. (1990). Let $\lambda_i(\boldsymbol{\Sigma}) = \lambda_i$ be the eigenvalues of $\boldsymbol{\Sigma}$ so $c_X \lambda_i$ are the eigenvalues of $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma_x}$. Under (E1), $\lambda_i(\boldsymbol{S}) \xrightarrow{P} c_X \lambda_i$ and $\lambda_i(\boldsymbol{C}_U) \xrightarrow{P} c\lambda_i = \dfrac{c}{c_X} c_X \lambda_i = d\, c_X\, \lambda_i$. Hence the population eigenvalues of $\boldsymbol{\Sigma_x}$ and $d\, \boldsymbol{\Sigma_x}$ 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 $\boldsymbol{\rho_x}$ 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 $\boldsymbol{\Sigma_x}\ \boldsymbol{e} = \lambda \boldsymbol{e}$, then

$$d\ \boldsymbol{\Sigma_x}\ \boldsymbol{e} = d\lambda \boldsymbol{e}.$$

Thus $\hat{\lambda}_i(\boldsymbol{S}) \xrightarrow{P} \lambda_i(\boldsymbol{\Sigma_x})$ and $\hat{\lambda}_i(\boldsymbol{C}_U) \xrightarrow{P} d\lambda_i(\boldsymbol{\Sigma_x})$ 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. Theorems 6.2 and 6.3 imply 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 $\boldsymbol{\Sigma_x}\ \boldsymbol{e} = \lambda \boldsymbol{e}$ implies that $\boldsymbol{\Sigma_x}\ (-\boldsymbol{e}) = \lambda(-\boldsymbol{e})$.

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

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

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.

Remark 6.1. When $\boldsymbol{R}$ is used, the correlation of the ith variable with the jth principal component is proportional to the ith entry of the jth eigenvector $\hat{\boldsymbol{e}}_j$. The same result holds when $\boldsymbol{R}_U$ is used *if the correlation is computed on cases in the RMVN set* U. 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 the principal component is said to *load high* on the variables corresponding to entries that are large in magnitude. i) Sometimes only one entry is large. ii) Sometimes all of the large entries have approximately the same size and sign. Then the principal component is interpreted as an average of these entries. iii) 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. iv) If there are $j \geq 2$ large entries that differ in magnitude, then the principal component is interpreted as a linear combination of the corresponding variables.

Warning: The above interpretations may not be valid if $\boldsymbol{S}$ or $\boldsymbol{S}_U$ is used, although the principal component will be a linear combination of the variables. Let $Y_j = \boldsymbol{e}_j^T \boldsymbol{x}$ be the jth population principal component, where $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$. Then $\text{Cov}(\boldsymbol{x}, Y_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{e}_j = \lambda_j \boldsymbol{e}_j$. Let $\boldsymbol{e}_j = (e_{1j}, ..., e_{ij}, ..., e_{pj})^T$. Let $\boldsymbol{x} = (X_1, ..., X_p)^T$ where X_i is the ith random variable with $V(X_i) = \sigma_{ii}$. Problem 6.3 c) will show that $\mathrm{corr(X_i, Y_j)} = \sqrt{\mathrm{\lambda_j}} \frac{\mathrm{e_{ij}}}{\sqrt{\mathrm{\sigma_{ii}}}}$. Recall that the correlation matrix is the covariance matrix of standardized variables $\boldsymbol{z} = (Z_1, ..., Z_p)$, with $\sigma_{ii} = 1$. Hence if a correlation matrix is used for PCA with $Y_j = \boldsymbol{e}_j^T \boldsymbol{z}$, then $\mathrm{corr(X_i, Y_j) = corr(Z_i, Y_j)} = \sqrt{\mathrm{\lambda_j}}\mathrm{e_{ij}}$, and the constant $\sqrt{\lambda_j}$ is the same for X_i for $i = 1, ..., p$. If the covariance matrix is used, Remark 6.1 applies to $e_{ij}/\sqrt{\sigma_{ii}}$ instead of the loadings e_{ij}.

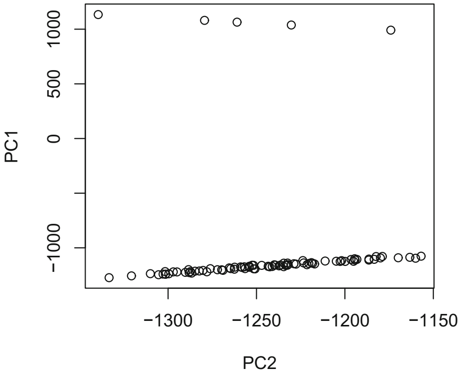

Fig. 6.1 First Two Principal Components for Buxton data

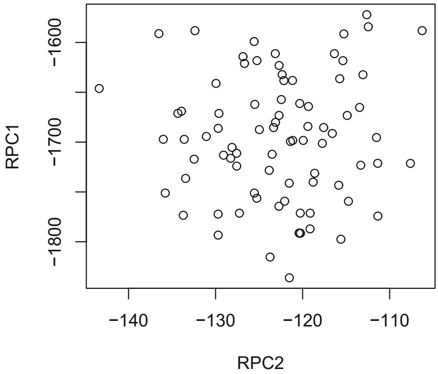

Fig. 6.2 First Two Robust Principal Components with Outliers Omitted From Plot

Example 6.1. Buxton (1920) gave 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 recorded under head length. Performing a classical principal component 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 $\hat{\boldsymbol{e}}_2 \approx (-0.005, -0.848, -0.054, -0.048, 0.525)^T$ that loads high on *head length* and *cephalic*. 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)
```

Using the commands

```
plot(PC1,PC2)
biplot(zz)
```

will give similar plots, except PC1 will be on the x-axis.

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. The default for `rprcomp` is the correlation matrix $\boldsymbol{R}_U$ applied to subset U, while the default for `prcomp` is the covariance matrix $\boldsymbol{S}$.

```
zzcor <- prcomp(z,scale=T)
zrcor <- rprcomp(z,cor=T)
```

An equivalent way to do RPCA is to get the RMVN set U and then perform classical PCA.

```
u <- getu(z)$U
zrcor <- prcomp(u,scale=T)
```

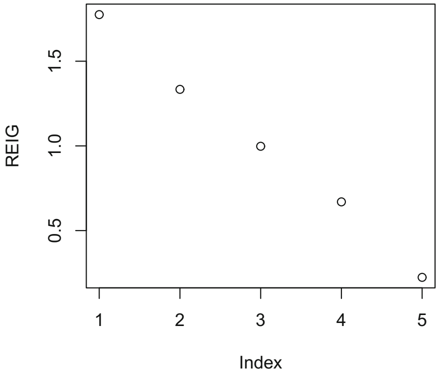

Fig. 6.3 Robust Scree Plot

Then

```
zrcor$out$sd^2
```

gives the eigenvalues, and *zrcor$out$rot* 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.

```
EIG <- zzcor$sd^2
plot(EIG)
#robust scree plot
REIG <- zrcor$out$sd^2
plot(REIG)
```

The summary command can be used to find the proportion of variance explained. The output shown below suggests that the 5th principal component can be omitted. For this example, the outliers did not affect the variance explained much in that the cumulative proportions for PCA and RPCA are similar. The biplots and eigenvectors do differ greatly.

```
summary(zzcor) #classical PCA
Importance of components:
                         PC1   PC2   PC3   PC4   PC5
Standard deviation     1.431 1.074 0.964 0.926 0.106
Proportion of Variance 0.410 0.231 0.186 0.172 0.002
Cumulative Proportion  0.410 0.640 0.826 0.998 1.000
summary(zrcor$out) #RPCA Importance of components:
                         PC1   PC2   PC3   PC4   PC5
Standard deviation     1.332 1.155 0.999 0.818 0.473
Proportion of Variance 0.355 0.267 0.200 0.134 0.045
Cumulative Proportion  0.355 0.622 0.821 0.955 1.000
```

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.

From the following output, 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 eigenvectors sometimes differ in sign.

```
zw <-z[-c(61,62,63,64,65),]
zzcorc <- prcomp(zw,scale=T)
# classical PCA on 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 #RPCA on full data with outliers
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 #RPCA on cleaned data
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
```

The robust biplot is shown in Figure 6.4 and uses cases from the RMVN subset U since outliers can obscure details in the plot. Note that the biplot is similar to Figure 6.2 if the axes were interchanged. The first principal component (eigenvector) from RPCA can be interpreted as a difference *head length* $-$ *cephalic*. In the biplot, the arrows corresponding to these two variables are nearly horizontal and of the same length, but in the opposite direction. The second principal component can be interpreted as an average of *height* (buxy) and *nasal height* (nasal). In the biplot, the arrows corresponding to these two variables are nearly vertical with the same length and direction. All other arrows in the biplot have very short length. The third principal component is highly correlated with *bigonal*, the fourth principal component is roughly proportional to *height* $-$ *nasal*, and the fifth principal component is roughly an average of *length* and *cephalic*. The robust biplot was made with the following command.

```
biplot(zrcor$out) #use biplot(zzcor) for a PCA biplot
```

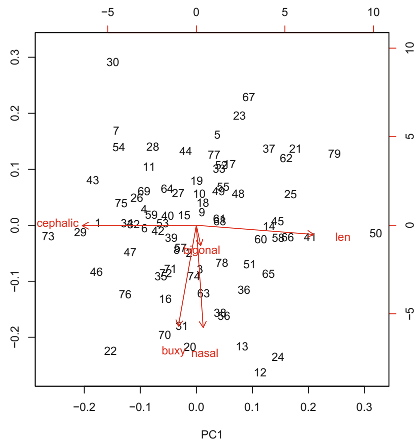

Fig. 6.4 Robust Biplot

In simulations for principal component analysis, FCH, RMVN, OGK, and FMCD 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 $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{x}) = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T$ if second moments exist. 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. See Rocke and Woodruff (1996, p. 1050). DGK is affine equivariant, and FMCD is pseudo-affine equivariant (see the Warning at the end of Section 4.1). 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 $\boldsymbol{x} = \boldsymbol{Az}$ 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 $\text{Cov}(\boldsymbol{x}) = \sigma^2 diag(1, ..., p)$ when second moments exist. 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 $1/p$. Use the command `pcasim(n=100,q=4,nruns=1000,corr=F,rot=F,xtype=1)` for the third and fourth lines of Table 6.2. Also see Problem 6.9.

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 6.2 shows the mean "variance explained" (M) along with the standard deviations (S) 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 nonequivariant estimators, Maronna and Zamar (2002) suggested using $\boldsymbol{A}_{i,i} = 1$ and $\boldsymbol{A}_{i,j} = \psi$ for $i \neq j$ and $\psi = 0, 0.5, 0.7, 0.9$, and 0.99. Then $\text{Cov}(\boldsymbol{x}) = \sigma^2 \boldsymbol{A}^2$ if second moments exist. If ψ is high, or if p is high and $\psi \geq 0.5$, then the data are concentrated about the line with direction $\mathbf{1} = (1, ..., 1)^T$. For $p = 50$ and $\psi = 0.99$, 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 ψ are small.

Table 6.3 Variance Explained by PCA and RPCA, SSD $= 10^7$ SD, $p = 50$, $\psi = 0.99$

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

Table 6.3 shows the mean "variance explained" along with the standard deviations multiplied by 10^7 for the first principal component, denoted by "SSD." The a_1 value is given but p_1 was always 1.0 to many decimal places even with Cauchy data. 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. RPCA had a much lower SSD than PCA for the Cauchy data.

6.3 Eigenvalue Inference

We would like to test hypotheses $H_0 : (\lambda_{p-k}, \lambda_{p-k-1}, ..., \lambda_{p-1}, \lambda_p)^T = \mathbf{0}$ and $H_0 : \lambda_i = 0$. Waternaux (1976) and Tyler (1983) gave some large sample theory for PCA. In particular, if the $\boldsymbol{x}_j = (X_{1j}, ..., X_{pj})^T$ are iid from a multivariate distribution with fourth moments and a covariance matrix $\boldsymbol{\Sigma_x}$ such that the eigenvalues are distinct and positive, then Waternaux (1976) claims $\sqrt{n}(\hat{\lambda}_i - \lambda_i) \xrightarrow{D} N(0, \kappa_i + 2\lambda_i^2)$ where κ_i is the kurtosis of the marginal distribution of X_i, for $i = 1, ..., p$. (Probably also need $\kappa_i \equiv \kappa$ for $i = 1, ..., p$ or $X_1, ..., X_p$ independent with $V(X_1) > V(X_2) > \cdots > V(X_p)$. For example, assume $X_1, ..., X_p$ are independent with $V(X_i) = i$ or $V(X_i) = p + 1 - i$.) For a MVN distribution, $\kappa_i = 0$. The limiting distribution depends on the distribution of $\boldsymbol{x}$, so several tests and confidence intervals are not robust to the assumption of normality. In particular, the $100(1-\delta)\%$ confidence interval

(CI)

$$\hat{\lambda}_i \pm z_{1-\delta}\sqrt{2}\sqrt{\hat{\lambda}_i/n}$$

should not be used unless n is large and the plotted points in the DD plot cluster tightly about the identity line. Here $P(Z \leq z_{1-\delta}) = 1 - \delta$ if $Z \sim N(0, 1)$.

Also, if $\lambda_i = 0$, since the dispersion matrix is positive semidefinite, $\hat{\lambda}_i \geq 0$. Hence for $j = 1, ..., B$, bootstrap values $\hat{\lambda}^*_{ij} \geq 0$ and are often all positive. So the shorth $100(1-\delta)\%$ CI may not contain 0. Similarly, when testing $H_0 : (\lambda_{p-k}, \lambda_{p-k-1}, ..., \lambda_{p-1}, \lambda_p)^T = \mathbf{0}$ with the prediction region method, the prediction region may not contain $\mathbf{0}$ when H_0 is true. Beran and Srivastava (1985) also suggest methods for bootstrapping PCA.

As in the simulation that produced Table 6.2, the simulation for the shorth CIs often used $\boldsymbol{z} = (z_1, ..., z_p)^T$ where the z_i are iid from a continuous distribution and variance σ^2 if second moments exist. Then $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z}$ where $\boldsymbol{A} = diag(\sqrt{1}, ..., \sqrt{p})$ and $\text{Cov}(\boldsymbol{x}) = \sigma^2 diag(1, ..., p)$ when second moments exist. The rows of the data matrix were sampled with replacement to bootstrap the sample covariance matrix. Then the classical or robust PCA was fit to find $\hat{\lambda}^*_{1j}, ..., \hat{\lambda}^*_{pj}$ for $j = 1, ..., B = 1000$. In addition to the shorth 95% CIs, let $T_i = \overline{\hat{\lambda}^*_{ij}}$ and let $S^*_i = \sqrt{\boldsymbol{S}^*_{ii}}$ be the sample standard deviation of the $\hat{\lambda}^*_{ij}$ where $\boldsymbol{S}^*$ is the sample covariance matrix of vectors $(\hat{\lambda}^*_{1j}, ..., \hat{\lambda}^*_{pj})^T$ and $j = 1, ..., B$. We tried the nominal large sample 95% bootstrap CIs $T_i \pm 2S^*_i$ and $\hat{\lambda}_i \pm 2S^*_i$. Note that the standard deviation S^*_i is not divided by $\sqrt{n}$ since the statistic variability is proportional to $1/\sqrt{n}$. (If the statistic was $\overline{X}$, then the bootstrap variance would estimate σ^2/n, not σ^2.) The first choice seemed to perform a bit better and was used in the simulation.

The *mpack* function `pcabootsim` determined the proportion of times the shorth CI and bootstrap CI contained $\lambda_i = i\ \sigma^2$. We used the multivariate normal distribution with $\sigma^2 = 1$, the $MVT_p(d)$ distribution with $\sigma^2 = d/(d-2)$ for $d \geq 3$, the mixture distribution with the z_i iid $(1-\epsilon)N(0,1)+\epsilon N(0,25)$ with $\sigma^2 = 1 + 24\epsilon$ where $0 < \epsilon < 1$ (see Section 1.6), and the multivariate lognormal distribution with $\sigma^2 = e(e-1) \approx 4.6708$.

The *mpack* function `pcaboot` produces the shorth and bootstrap CIs for λ_i for PCA or RPCA. For RPCA, the eigenvalues equal $\lambda_i = ci\ \sigma^2$ in the simulation, where $c = 1$ for MVN data but is unknown otherwise. So the simulation for RPCA could only be done for MVN data. The bootstrap CIs and shorth CIs for λ_i did seem to work when the conditions for the above PCA large sample theory hold, but large samples were needed: $n \geq 100p$ for the MVN distribution and $n \geq 400p$ for some of the distributions in the simulation. Fourth moments seemed important, the $MVT_p(3)$ distribution only has third moments, and the bootstrap and shorth CIs had coverage well under 95%. The bootstrap CI were denoted by lsci.

Some output is shown below for the classical PCA using $p = 4$. The default for `pcabootsim` is to use the sample covariance matrix $\boldsymbol{S}$ (coverages will be incorrect if $\boldsymbol{R}$ is used since then $\lambda_i = 1$ when second moments exist), while the default for pcaboot is to use the sample correlation matrix $\boldsymbol{R}$.

```
pcabootsim(n=400,xtype=1,nruns=100) #MVN data
$lscicv #nominal 95% coverage
[1] 0.95 0.94 0.92 0.92
$shcicv #observed coverages > 91%
[1] 0.98 0.95 0.92 0.92
pcabootsim(n=400,xtype=2,nruns=100)
$lscicv
[1] 0.97 0.92 0.91 0.91
$shcicv
[1] 0.97 0.92 0.92 0.91
pcabootsim(n=400,xtype=3,nruns=100)
$lscicv
[1] 0.98 0.96 0.94 0.92
$shcicv
[1] 0.98 0.98 0.92 0.96
pcabootsim(n=1600,xtype=4,nruns=100)
$lscicv
[1] 1.00 0.95 0.93 0.92
$shcicv
[1] 1.00 0.96 0.91 0.93
pcabootsim(n=1000,xtype=5,nruns=100)
$lscicv
[1] 0.96 0.97 0.92 0.92
$shcicv
[1] 0.95 0.97 0.92 0.93
pcabootsim(n=1600,xtype=6,nruns=100)
#third moments, might need fourth moments
$lscicv
[1] 0.87 0.67 0.71 0.64
$shcicv
[1] 0.89 0.66 0.74 0.65
```

```
pcaboot(buxx,rob=T) #Buxton data RPCA with
#(generalized) correlation matrix
#n = 87 is probably to small to have coverage
#near the nominal 95%
$shorci[[1]]$shorth
[1] 1.6409 2.2055
$shorci[[2]]$shorth
[1] 0.9424 1.3176
$shorci[[3]]$shorth
[1] 0.4523 0.9918
$shorci[[4]]$shorth
```

```
[1] 0.1162 0.3710
$lscis
     lsciL  lsciU
[1,] 1.6002 2.1644
[2,] 0.8960 1.2967
[3,] 0.4860 1.0967
[4,] 0.1046 0.3555
```

6.4 Summary

1) Let $\boldsymbol{\Sigma} = (\sigma_{ij})$ be a positive definite symmetric $p \times p$ dispersion matrix. A *generalized correlation matrix* $\boldsymbol{\rho} = \left(\rho_{ij}\right)$ where

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

The generalized correlation matrix is the correlation matrix when second moments exist if $\boldsymbol{\Sigma} = c \operatorname{Cov}(\boldsymbol{x})$ 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 U. The robust PCA does classical PCA on the cases in the RMVN set U, which is equivalent to doing the PCA using $\boldsymbol{S}_U$ or $\boldsymbol{R}_U$. 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 $\boldsymbol{C}$.

4) We want $n \geq 10p$ for the classical PCA and $n \geq 20p$ 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

$$\sqrt{\hat{\lambda}_1}, \sqrt{\hat{\lambda}_2}, ..., \sqrt{\hat{\lambda}_p}$$

under the label "standard deviations," while *SAS* output gives the eigenvalues $\hat{\lambda}_i$. Typical *R* output is shown below.

```
Standard deviations:
[1] 1.3369152 1.1466891 1.0016463 0.8123854 0.4842482
```

```
Rotation: PC1      PC2      PC3      PC4      PC5
len        0.67273 -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
buxy      -0.21306 -0.67557 -0.01727  0.68852  0.15447
```

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 95%, 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 $\boldsymbol{R}_U$ with cases restricted to 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 entries. If exactly two entries are of similar large magnitude but of different sign, the principal component is interpreted as a difference of the two entries. If there are $j \geq 2$ large entries 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: $\hat{e}_{ij}/\sqrt{S_{ii}}$.

Hence PCA based on $\boldsymbol{S}$ is harder to interpret if the 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) Let $\hat{\boldsymbol{\Sigma}}$ be a consistent estimator of $\boldsymbol{\Sigma} > 0$. The following theorems show that asymptotically, the eigenvalues and eigenvectors of $\hat{\boldsymbol{\Sigma}}$ act as those of $\boldsymbol{\Sigma}$ and vice versa. 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 positive definite dispersion matrix.

i) **Theorem 6.1.** Suppose the dispersion matrix $\boldsymbol{\Sigma} > 0$ 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.

12) 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 $\hat{\boldsymbol{e}}_j^T \boldsymbol{w}_i = \hat{\boldsymbol{e}}_j^T(\boldsymbol{x}_i - T)$.

13) 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

14) For PCA, the most important *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.

6.5 Complements

Jolliffe (2010) is an authoritative text on PCA. See Cook (2007) for a good discussion on dimension reduction. Cattell (1966) and Bentler and Yuan (1998) are good references for scree plots. Møller et al. (2005) discussed PCA, principal component regression, and drawbacks of M estimators.

```
x<-cbind(buxx,buxy) # data matrix
mn <- apply(x,2,mean) #sample mean
J <- 0*1:87 + 1  # vector of n ones, n = 87
J <- J%*%t(J)/87 #J%*%x has rows = mn
zc <- x-J%*%x #centered x
yc <- zc/sqrt(87-1) #t(yc) %*% yc = cov(x)
svd(yc)$v              #right eigenvectors of Yc
          [,1]     [,2]     [,3]     [,4]    [,5]
[1,]  0.653883  0.75596 -0.01173  0.00988  0.0268
[2,] -0.001366  0.03980  0.06800 -0.42534 -0.9016
[3,] -0.000489 -0.01276 -0.99161 -0.12775 -0.0151
[4,] -0.000714  0.00251 -0.10890  0.89588 -0.4308
[5,] -0.756594  0.65327 -0.00952  0.00854  0.0252
> svd(t(yc))$u        #left eigenvectors of Yc^T
          [,1]     [,2]     [,3]     [,4]    [,5]
[1,] -0.653883 -0.75596  0.01173 -0.00988 -0.0268
[2,]  0.001366 -0.03980 -0.06800  0.42534  0.9016
[3,]  0.000489  0.01276  0.99161  0.12775  0.0151
[4,]  0.000714 -0.00251  0.10890 -0.89588  0.4308
[5,]  0.756594 -0.65327  0.00952 -0.00854 -0.0252
> prcomp(x)
Standard deviations:
[1] 523.70760  42.50435   6.06073   4.39067   3.80398
Rotation:
               PC1      PC2      PC3      PC4     PC5
len       0.653883  0.75596 -0.01173  0.00988  0.0268
nasal    -0.001366  0.03980  0.06800 -0.42534 -0.9016
bigonal  -0.000489 -0.01276 -0.99161 -0.12775 -0.0151
cephalic -0.000714  0.00251 -0.10890  0.89588 -0.4308
buxy     -0.756594  0.65327 -0.00952  0.00854  0.0252
svd(yc)$d        #singular values = sqrt(eigenvalues)
[1] 523.70760  42.50435   6.06073   4.39067   3.80398
svd(t(yc))$d     #singular values = sqrt(eigenvalues)
[1] 523.70760  42.50435   6.06073   4.39067   3.80398
```

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 σ_i 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 $\boldsymbol{Y}^T = \boldsymbol{V}\boldsymbol{\Lambda}^T\boldsymbol{U}^T$, and

$$\boldsymbol{Y}\boldsymbol{Y}^T = \frac{1}{n-1}\begin{bmatrix} \boldsymbol{z}_1^T\boldsymbol{z}_1 & \boldsymbol{z}_1^T\boldsymbol{z}_2 & \dots & \boldsymbol{z}_1^T\boldsymbol{z}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{z}_n^T\boldsymbol{z}_1 & \boldsymbol{z}_n^T\boldsymbol{z}_2 & \dots & \boldsymbol{z}_n^T\boldsymbol{z}_n \end{bmatrix}$$

which is the matrix of scalar products divided by $(n-1)$. Similarly, if $\boldsymbol{Z}_c$ is the centered data matrix (subtract the means), then $\boldsymbol{Y}_c = \boldsymbol{Z}_c/\sqrt{n-1}$, and the covariance matrix $\boldsymbol{S} = \boldsymbol{Y}_c^T\boldsymbol{Y}_c = \boldsymbol{Z}_c^T\boldsymbol{Z}_c/(n-1)$. For more information about the SVD, see Datta (1995, pp. 552–556) and Fogel et al. (2013).

The output on the previous page shows how to do classical PCA with $\boldsymbol{S}$ on a data set using the SVD. The eigenvectors agree up to sign.

The literature for robust PCA is large, but the "high breakdown" competitors for RPCA are impractical or not backed by theory. See Hubert et al. (2008; 2012) and Wilcox (2008) for references. 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 component is a mildly outlier resistant-bounded influence approach suggested by Locantore et al. (1999). Boente and Fraiman (1999) claimed that the basis of the eigenvectors is consistently estimated by spherical principal components for elliptically contoured distributions. Also see the end of Section 4.7, Maronna et al. (2006, pp. 212–213), and Taskinen et al. (2012). A potentially useful method for robust principal components uses alternating robust regressions. See Croux et al. (2007), Chen et al. (2008), and Liu et al. (2003).

It may be possible to do robust PCA when $n < p$ by standardizing the data with the MED(X_i) and MAD(X_i). Let the standardized data be in the matrix $\boldsymbol{Z}$. Then plot the Euclidean distances of the standardized data from the coordinatewise median MED$(\boldsymbol{Z})$ and delete outliers, leaving m cases in an $m \times p$ matrix $\boldsymbol{Y}$ (use *mpack* functions `medout` and `ddplot5`). Alternatively, find the `covrmb2` subset $B = \boldsymbol{Y}$. Then use the SVD of $\boldsymbol{Y}$ to perform a "robust" PCA. Also see Feng and He (2014).

Sparse PCA attempts to increase the interpretability of PCA by making many of the loading entries equal to 0. See Zou et al. (1993). Apply sparse PCA on the RMVN subset U if $n \geq 20p$, and on the `covrmb2` subset B if n is not much larger than p. Also see Croux et al. (2013). Xu et al. (2011) suggests that sparse algorithms are not stable.

Bali et al. (2011) gave possibly impressive theory for infinite complexity impractical robust projection estimators, but should have given theory for the practical F-projection estimator actually used. This error occurs far too often in multivariate "robust statistics" papers. Hubert et al. (2008; 2012) gave many references for methods, including PCA, where the practical plug-in estimator is not yet backed by theory and should be replaced by RFCH or RMVN. Also see Croux et al. (2007; 2013) for comparison of projection estimators with other methods.

To estimate the first principal direction for principal component analysis, the F-projection (CR) estimator uses n projections $\boldsymbol{z}_i = \boldsymbol{w}_i/\|\boldsymbol{w}_i\|$ where $\boldsymbol{w}_i = \boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n$. 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 F-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 an $\boldsymbol{x}_i$ 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 $\boldsymbol{x}_i$ 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 $\boldsymbol{x}_i$ falls in "hypercone") $\to 1$, and $n \to \infty$. If these heuristics are correct, we 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 F-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 $p > 1$.

In general, some criterion is needed to pick the estimated first principal component from the n candidates $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$. A possibility is to compute MAD$(\boldsymbol{z}_i^T \boldsymbol{w}_1, ..., \boldsymbol{z}_i^T \boldsymbol{w}_n)$ for $i = 1, ..., n$ and take the $\boldsymbol{z}_j$ that maximizes the MAD.

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 first three simulations and have more outlier configurations than the two configurations 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 $\hat{\boldsymbol{\Sigma}}$ is a consistent estimator of $\boldsymbol{\Sigma}$, then the inverse, determinant, and eigenvalues of $\hat{\boldsymbol{\Sigma}}$ are consistent estimators of the inverse, determinant, and eigenvalues of $\boldsymbol{\Sigma} > 0$. See, for example, Bhatia et al. (1990), Stewart (1969), and Severini (2005, pp. 348–349).

6.6 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

6.1*. Assume the $p \times p$ dispersion matrix $\boldsymbol{\Sigma}$ is positive definite. If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ for some constant $c > 0$, prove that $\boldsymbol{\Sigma}\hat{\boldsymbol{e}}_i - \lambda_i \hat{\boldsymbol{e}}_i \xrightarrow{P} \boldsymbol{0}$.

```
rprcomp(z)  #robust PCA for Problem 6.2
Standard deviations:
[1] 1.3369152 1.1466891 1.0016463 0.8123854 0.4842482
Rotation: PC1       PC2       PC3       PC4       PC5
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
buxy      -0.21306 -0.67557 -0.01727  0.68852  0.15446

prcomp(z,scale=T)  #classical PCA
Standard deviations:
[1] 1.3184358 1.1723991 1.0155266 0.7867349 0.4867867

Rotation:      PC1      PC2      PC3      PC4      PC5
len       -0.70308 -0.06778 0.07744  0.16901  0.6830
nasal     -0.15038  0.68868 0.02042 -0.70385  0.0854
bigonal   -0.11646 -0.04882 0.96504 -0.02261 -0.2285
cephalic   0.68502  0.08950 0.24854  0.03071  0.6782
buxy      -0.01551  0.71466 0.02247  0.68890 -0.1181
```

6.2. Shown above 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 (to make the variability of buxy similar to that of the other variables). The "standard deviations" line corresponds to the square roots of the eigenvalues. The Rotation matrix gives the five principal components.

a) For the robust `rprcomp` output make a scree plot. What proportion of the trace is explained by the first four principal components?

b) Which principal component corresponds to i) bigonal, ii) nasal + buxy, iii) length + cephalic, iv) length − cephalic, and v) nasal − buxy?

6.3*. Let $Y_j = \boldsymbol{e}_j^T \boldsymbol{x}$ be the jth 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 five cases have massive outliers in the height and length variables. Interpret PC1 and PC2.

```
prcomp(z,scale=T)  #Problem 6.4
[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.

```
Output for Problem 6.5.
              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
```

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, shown below, is based on the (robust) correlation matrix.

a) How many principal components are needed?

b) Interpret the first principal component.

```
output for Problem 6.6
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
```

6.7. 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.

a) Interpret the first principal component.

b) Interpret the second principal component.

```
> rprcomp(USJudgeRatings) #Problem 6.7
Standard deviations:
[1] 3.22195 1.03833 0.51050 0.41049 0.22798 0.16243
[7] 0.11156 0.09407 0.07441 0.05595 0.04492 0.03806
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
```

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   154.3106   145.147647       0.9439        0.9439
2     9.1630                    0.0561        1.0000
                             Eigenvectors
                                  Prin1         Prin2
                July           0.343532      0.939141
                January        0.939141      -.343532
```

R 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, e.g., *ddplot*, will display the code for the function. Use the `args` command, e.g., *args(pcasim)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/mrsashw.txt) into *R*.

6.9. a) Type the *R* command `pcasim()` and paste the output into *Word*.

This command computes the first three eigenvalues and eigenvectors for the classical and robust PCA using $\boldsymbol{R}$ 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 three eigenvalues. The terms abscoreigvi give the absolute correlation between the ith classical and robust eigenvectors 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) to 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 four 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 the proportion of variance explained and the 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?

b) 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. Consider the *biplot* of the first principal component versus the second principal component. See Definition 6.3.

The Buxton (1920) data has a cluster of five massive outliers. The first classical principal component tends to go right through a cluster of large outliers.

a) The *R* commands for this part make the classical scree plot and biplot. Paste the plots into *Word.*

b) The *R* commands for this part 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 five massive outliers?

e) From the robust biplot, which two variables loaded highest with the first principal component?

6.13. The Johnson (1996) STATLIB *bodyfat* data has $n = 252$ and 15 variables x_1 = *density* determined by underwater weighing, x_2 = *bfat* = the person's body fat percentage, x_3 = *age*, x_4 = *weight*, x_5 = *height*, and measurements x_6 = *neck*, x_7 = *chest*, x_8 = *abdomen*, x_9 = *hip*, x_{10} = *thigh*, x_{11} = *knee*, x_{12} = *ankle*, x_{13} = *biceps*, x_{14} = *forearm*, and x_{15} = *wrist.*

a) The *R* commands for this part make the classical scree plot and biplot. Paste the biplot into *Word.*

b) The *R* commands for this part make the robust scree plot and biplot. Paste the biplot into *Word.*

c) From the robust biplot, what is the relationship between *bfat* and *density*?

Chapter 7
Canonical Correlation Analysis

This chapter covers classical and robust canonical correlation analysis (CCA). 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 $q \times 1$ with $m = p - q \leq q$ and $m, q \geq 1$. If $m = 1$ and $q = 1$, then the canonical correlation is the usual correlation. Hence usually $q > 1$ and $m > 1$. The population canonical correlation analysis 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*.

7.1 Introduction

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 $q \times 1$.

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_7

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. We want $n \geq 10p$ for classical CCA and $n \geq 20p$ for robust CCA that uses FCH, RFCH, or RMVN. Also make the DD plot for the $\boldsymbol{w}$ variables and the DD plot for the $\boldsymbol{y}$ 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

$$U_k = \boldsymbol{a}_k^T \boldsymbol{w} = \boldsymbol{e}_k^T \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{w} \quad \text{and} \quad V_k = \boldsymbol{b}_k^T \boldsymbol{y} = \boldsymbol{g}_k^T \boldsymbol{\Sigma}_{22}^{-1/2} \boldsymbol{y}$$

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, pp. 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 is the linear combinations (U, V) with unit variances that maximize $\text{corr}(U, V)$ among all choices uncorrelated with the previous $i - 1$ 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 $\boldsymbol{\Sigma} = \boldsymbol{\rho}_{\boldsymbol{x}}$. Hence $\boldsymbol{\Sigma}_{ii} = \boldsymbol{\rho}_{\boldsymbol{x},ii}$ 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

$$U_k = \boldsymbol{a}_k^T \boldsymbol{w} = \boldsymbol{e}_k^T \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{w} \quad \text{and} \quad V_k = \boldsymbol{b}_k^T \boldsymbol{y} = \boldsymbol{g}_k^T \boldsymbol{\Sigma}_{22}^{-1/2} \boldsymbol{y}$$

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 it can be shown that 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

$$\hat{U}_k = \hat{\boldsymbol{a}}_k^T \boldsymbol{w} = \hat{\boldsymbol{e}}_k^T \hat{\boldsymbol{\Sigma}}_{11}^{-1/2} \boldsymbol{w} \quad \text{and} \quad \hat{V}_k = \hat{\boldsymbol{b}}_k^T \boldsymbol{y} = \hat{\boldsymbol{g}}_k^T \hat{\boldsymbol{\Sigma}}_{22}^{-1/2} \boldsymbol{y}$$

for $k = 1, ..., m$. Then the kth *sample canonical correlation* $\hat{\rho}_k = corr(\hat{U}_k, \hat{V}_k) = \sqrt{\hat{\lambda}_k}$ 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 $(\hat{U}_i, \hat{V}_i)$ is the pair of 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 $i-1$ canonical variable pairs and $\text{corr}(\hat{U}_i, \hat{V}_i) = \hat{\rho}_i$.

Note that $\boldsymbol{x} = (\boldsymbol{w}^T, \boldsymbol{y}^T)^T$ are labels. The labels $\boldsymbol{w} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T$ are also often used. The *R* function `cancor` is used to perform the classical CCA and produces output \$cor, \$xcoef, and \$ycoef. These are the canonical correlations $\hat{\rho}_k$, the $\hat{\boldsymbol{a}}_i$, and the $\hat{\boldsymbol{b}}_i$. *R* output is shown in symbols for the following table, then output is given for Example 7.1.

corr					
$\hat{\rho}_1$	$\cdots$	$\hat{\rho}_m$			
wcoef					
$\boldsymbol{w}$	$\hat{\boldsymbol{a}}_1$	$\cdots$	$\hat{\boldsymbol{a}}_m$		
ycoef					
$\boldsymbol{y}$	$\hat{\boldsymbol{b}}_1$	$\cdots$	$\hat{\boldsymbol{b}}_m$	$\cdots$	$\hat{\boldsymbol{b}}_q$

Example 7.1. This example will be continued in more detail in the following section. The output is for the mussel data described in Example 2.2, where the log transformation was used on the five variables. For this data set, $m = 2$ canonical correlations are found. The 1st canonical correlation $\hat{\rho}_1 = 0.982$, while the 2nd canonical correlation can be ignored. Note that $\hat{\boldsymbol{a}}_1 = (0.127, 0.019)^T$ which puts most of the weight on $\log(S)$. Note that $\hat{\boldsymbol{b}}_1 = (0.157, 0.161, 0.214)^T$.

```
zm <- log(mussels); x <- zm[,c(4,5)];
y <- zm[,-c(4,5)]; out<-cancor(x,y)
out$cor
[1] 0.9818605 0.1555381    out$ycoef
out$xcoef                       [,1]     [,2]     [,3]
      [,1]      [,2]    L 0.15675  0.72779  2.19359
S 0.126505  0.40778    W 0.16051  0.86505 -1.06764
M 0.018973 -0.48725    H 0.21438 -2.06346 -0.83039
```

Rule of thumb 7.2. Interpret the $\hat{\boldsymbol{a}}_i$ and $\hat{\boldsymbol{b}}_i$ much as $\hat{\boldsymbol{e}}_j$ is interpreted for PCA. The first pair $(\hat{\boldsymbol{a}}_1, \hat{\boldsymbol{b}}_1)$ corresponding to $\hat{\rho}_1$ is the most important. Pairs with low $\hat{\rho}_k$ can be ignored.

Interpretation of CCA output is often hard. i) If $\boldsymbol{x} = (\boldsymbol{w}^T, \boldsymbol{y}^T)^T$ and $\boldsymbol{w}$ and $\boldsymbol{y}$ are independent: $\boldsymbol{w} \perp\!\!\!\perp \boldsymbol{y}$, then often $\hat{\rho}_1$ is not close to 0 until the sample size n is quite large.

ii) Let $\boldsymbol{w} = (W_1, ..., W_m)^T$ and $\boldsymbol{y} = (Y_1, ..., Y_q)^T$. Unlike PCA with the correlation matrix, $\text{cor}(W_i, \hat{\boldsymbol{a}}_j)$ and $\text{cor}(Y_i, \hat{\boldsymbol{b}}_j)$ are not proportional to $\hat{a}_{ij}$ and $\hat{b}_{ij}$. This means the first sentence of Rule of thumb 7.2 is not very good. Computing $\text{cor}(W_i, \hat{\boldsymbol{a}}_j)$ and $\text{cor}(W_i, \hat{\boldsymbol{b}}_j)$ for $i = 1, ..., m$ can help. Similarly, compute $\text{cor}(Y_i, \hat{\boldsymbol{a}}_j)$ and $\text{cor}(Y_i, \hat{\boldsymbol{b}}_j)$ for $i = 1, ..., q$.

iii) Multicollinearity occurs if some of the W_i are highly correlated and/or some of the Y_i are highly correlated. Then the canonical variates $\hat{\boldsymbol{a}}_i$ and $\hat{\boldsymbol{b}}_j$ can be hard to interpret.

iv) The W_i's should have similar variances S_i^2, and the Y_j's should have similar variances S_j^2. Otherwise, the components of $\hat{\boldsymbol{a}}_k$ and $\hat{\boldsymbol{b}}_k$ are hard to interpret.

To interpret CCA, i) We want to know which W_i variables are most important for $\hat{\boldsymbol{a}}_1$, and so which W_i variables most explain $\hat{\boldsymbol{b}}_1^T \boldsymbol{y}$ (log(S), likely due to multicollinearity).

ii) We want to know which Y_i variables are most important for $\hat{\boldsymbol{b}}_1$, and so which Y_i variables most explain $\hat{\boldsymbol{a}}_1^T \boldsymbol{w}$ (all three are important, but there is multicollinearity).

iii) Are $\hat{\boldsymbol{a}}_1^T \boldsymbol{w}$ and $\hat{\boldsymbol{b}}_1^T \boldsymbol{y}$ meaningful? Sometimes the output has a "wow factor": the client says "wow that makes sense," but often interpretation is difficult.

7.2 Robust CCA

The *R* function `cancor` does classical CCA and the *mpack* function `rcancor` does robust CCA (RCCA) by applying `cancor` on the RMVN set U: the subset of the data used to compute RMVN. See Definition 4.11 and Section 4.6. Recall that the subset U is found using ellipsoidal trimming and can be regarded as cleaned data where the cleaned data is such that the classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ applied to the data U results in the estimator $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ which is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma}_{\boldsymbol{x}})$ for a large class of elliptically contoured distributions. Also, $(T_{RMVN}, \boldsymbol{C}_{RMVN}) = (\overline{\boldsymbol{x}}_U, a\boldsymbol{S}_U)$. Hence $\boldsymbol{S}_U$ is the unscaled RMVN dispersion estimator.

Some theory is simple: the FCH, RFCH, RMVN, and RCCA methods of robust CCA produce consistent estimators of the kth canonical correlation ρ_k on a large class of elliptically contoured distributions.

To see this, suppose $\text{Cov}(\boldsymbol{x}) = c_x \boldsymbol{\Sigma}$ 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_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 positive definite 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 robust CCA methods and RCCA produce consistent estimators the kth canonical correlation ρ_k on that class of distributions. These remarks prove the following theorem.

Theorem 7.4. For RCCA, Theorem 7.1 holds if the $\boldsymbol{x}_i$ are iid from a large class of elliptically contoured distributions.

Example 7.1, continued. 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 was taken on all five variables, and y contains $\log(L), \log(H)$, and $\log(W)$.

The data set *zm* contains x and y, and the DD plot (not shown) showed case 48 was separated from the bulk of the data, but near the identity line. The DD plot for x (not shown) 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 (not shown) 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.

From the output shown below, 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 we 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.

```
> zm <- log(mussels); x <- zm[,c(4,5)]
> y <- zm[,-c(4,5)]

> 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
```

The RCCA output can also be obtained by performing classical CCA on the RMVN subset U. Try this with the following R commands.

```
u <- getu(zm)$U
ux <- u[,c(4,5)]
uy <- u[,-c(4,5)]
cancor(ux,uy)
```

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 $q \times 1$ with $m = p - q \leq q$ and $m, q \geq 1$. 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					
$\hat{\rho}_1$	$\cdots$	$\hat{\rho}_m$			
wcoef					
$\boldsymbol{w}$	$\hat{\boldsymbol{a}}_1$	$\cdots$	$\hat{\boldsymbol{a}}_m$		
ycoef					
$\boldsymbol{y}$	$\hat{\boldsymbol{b}}_1$	$\cdots$	$\hat{\boldsymbol{b}}_m$	$\cdots$	$\hat{\boldsymbol{b}}_q$

```
$out$cor
[1] 0.98596703 0.06797587   $out$ycoef
$out$xcoef                         [,1]      [,2]      [,3]
       [,1]       [,2]   L 0.162545  0.423752 -2.849268
S 0.149662  0.646012   W 0.236969  1.537968  0.935650
M 0.032363 -0.854339   H 0.253032 -2.680646  1.778593
```

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

Koch (2014, p. 115) showed that CCA can be cast as a generalized eigenvector problem. Muirhead and Waternaux (1980) showed 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 $k = 1, ..., m$. If the data are iid from an elliptically contoured distribution with standardized 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 $k = 1, ..., m$ is $N_m(\mathbf{0}, (\kappa + 1)\boldsymbol{I}_p)$. Note that $\kappa = 0$ for multivariate normal data. The prediction region method can be used to create shorth confidence intervals for the ρ_k if the ρ_k are distinct, positive, and if the underlying population distribution has finite fourth moments. See how the prediction region method was used for eigenvalue inference for PCA in Section 6.3.

The literature for robust CCA is large, but the "high breakdown" competitors for RCCA are impractical or not yet backed by theory. More work is needed to show that Theorem 7.1 holds for other practical robust methods. Some of these methods may be useful as outlier diagnostics. Alkenani and Yu (2013), Zhang (2011), and Zhang et al. (2012) gave references for practical robust CCA that is not yet backed by theory and developed robust CCA based on FCH, RFCH, and RMVN. Alternating regressions may be a useful method if $p > n$, but the practical robust methods are not yet backed by theory. See Dehon et al. (2000).

7.5 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

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 = $20/\sqrt{(\text{estimated performance} + 1)}$. These six variables had a linear scatterplot matrix and DD plot, and similar variances. We 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) #Problem 7.2 output
$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  -0.0136133  0.057004  0.0897574 -0.0114237
mmin   0.0374853 -0.018749  0.0844425  0.0058597
chmin  0.0069323  0.098436 -0.0217826  0.0907567
chmax  0.0199989  0.011597  0.0078556 -0.0941986
```

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 were 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  #Problem 7.3 output
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 correlation 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 $\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T$. Then from the DD plot, the $\boldsymbol{z}_i$ appeared to follow a multivariate normal distribution. Sketch the DD plot.

```
> rcancor(x,y)  #Problem 7.4 output
$out
$out$cor
[1] 0.8116953 0.5064619 0.1376399

$out$xcoef
                     [,1]        [,2]        [,3]
x.kms         -2.0802e-05 -0.00002339 -2.2597e-06
x.PetrolPrice -1.8480e+00  3.71737158  5.2920e+00
x.VanKilled    1.5976e-03 -0.01684508  1.6737e-02

$out$ycoef
                 [,1]        [,2]        [,3]
y.drivers  1.6788e-06 -2.4873e-05  0.00047179
y.front    5.5947e-04 -7.7970e-05 -0.00081576
y.rear    -9.9650e-04 -7.5216e-04  0.00050458
```

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]  #Problem 7.5 output
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
```

R Problem

Warning: For the following problem, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/mrsashw.txt) into *R*.

7.6. Copy and paste the *R* commands for this problem into *R*. These commands make $\boldsymbol{w}$ 3×1 and $\boldsymbol{y}$ 5×1 where there are $n_1 = n_2 \equiv n_i$ cases for both $\boldsymbol{w}$ and $\boldsymbol{y}$. The eight variables $w_1, w_2, w_3, y_1, ..., y_5$ are iid $N(0,1)$. Hence the population canonical correlations are 0. The output starts with $n_i = 500$ and gives the first sample classical and robust correlation $\hat{\rho}_1$, then increases n_i by 500 and repeats. How large does n_i need to be before $\hat{\rho}_1 < 0.05$ for the classical estimator and for the robust estimator?

Chapter 8
Discriminant Analysis

This chapter considers discriminant analysis: given p measurements $\boldsymbol{w}$, we want to correctly classify $\boldsymbol{w}$ into one of G groups or populations. The maximum likelihood, Bayesian, and Fisher's discriminant rules are used to show why methods like linear and quadratic discriminant analysis can work well for a wide variety of group distributions.

8.1 Introduction

Definition 8.1. In *supervised classification*, there are G known groups and m cases. Each case is assigned to exactly one group based on its measurements $\boldsymbol{w}_i$.

Suppose there are G populations or groups or classes where $G \geq 2$. 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 $j = 1, ..., G$. 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 $p \times 1$ (observed) random vector from one of the G groups, but the group is unknown. Usually there are many $\boldsymbol{w}_i$, and *discriminant analysis* (DA) attempts to allocate the $\boldsymbol{w}_i$ to the correct groups. The $\boldsymbol{w}_1, ..., \boldsymbol{w}_m$ are known as the *test data.* Let π_k = the (prior) probability that a randomly selected case $\boldsymbol{w}_i$ belongs to the kth group. If $\boldsymbol{x}_{1,1} ..., \boldsymbol{x}_{n_G,G}$ are a random sample of cases from the collection of G populations, then $\hat{\pi}_k = n_k/n$ where $n = \sum_{i=1}^{G} n_i$. Often the *training data* $\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{n_G,G}$ is not collected in this manner. Often the n_k are fixed numbers such that n_k/n does not estimate π_k. For example, could have $G = 2$ where $n_1 = 100$ and $n_2 = 100$ where patients in group 1 have a deadly disease and patients in group 2 are healthy, but an attempt has been made to match the sick patients with healthy patients on p variables

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_8

such as age, weight, height, an indicator for smoker or nonsmoker, and gender. Then using $\hat{\pi}_j = 0.5$ does not make sense because π_1 is much smaller than π_2. Here the indicator variable is qualitative, so the p variables do not have a pdf.

Let $\boldsymbol{W}_i$ be the random vector and $\boldsymbol{w}_i$ be the observed random vector. Let $Y = j$ if $\boldsymbol{w}_i$ comes from the jth group for $j = 1, ..., G$. Then $\pi_j = P(Y = j)$ and the *posterior probability* that $Y = k$ or that $\boldsymbol{w}_i$ belongs to group k is

$$p_k(\boldsymbol{w}_i) = P(Y = k|\boldsymbol{W}_i = \boldsymbol{w}_i) = \frac{\pi_k f_k(\boldsymbol{w}_i)}{\sum_{j=1}^G \pi_j f_j(\boldsymbol{w}_i)}. \tag{8.1}$$

Definition 8.2. a) The *maximum likelihood discriminant rule* allocates case $\boldsymbol{w}_i$ to group a if $\hat{f}_a(\boldsymbol{w}_i)$ maximizes $\hat{f}_j(\boldsymbol{w}_i)$ for $j = 1, ..., G$.

b) The *Bayesian discriminant rule* allocates case $\boldsymbol{w}_i$ to group a if $\hat{p}_a(\boldsymbol{w}_i)$ maximizes

$$\hat{p}_k(\boldsymbol{w}_i) = \frac{\hat{\pi}_k \hat{f}_k(\boldsymbol{w}_i)}{\sum_{j=1}^G \hat{\pi}_j \hat{f}_j(\boldsymbol{w}_i)}$$

for $k = 1, ..., G$.

c) The (population) *Bayes classifier* allocates case $\boldsymbol{w}_i$ to group a if $p_a(\boldsymbol{w}_i)$ maximizes $p_k(\boldsymbol{w}_i)$ for $k = 1, ..., G$.

Note that the above rules are robust to nonnormality of the G groups. Following James et al. (2013, pp. 38–39, 139), the Bayes classifier has the lowest possible expected test error rate out of all classifiers using the same p predictor variables $\boldsymbol{w}$. Of course, typically, the π_j and f_j are unknown. Note that the maximum likelihood rule and the Bayesian discriminant rule are equivalent if $\hat{\pi}_j \equiv 1/G$ for $j = 1, ..., G$. If p is large, or if there is multicollinearity among the predictors, or if some of the predictor variables are noise variables (useless for prediction), then there is likely a subset $\boldsymbol{z}$ of d of the p variables $\boldsymbol{w}$ such that the Bayes classifier using $\boldsymbol{z}$ has lower error rate than the Bayes classifier using $\boldsymbol{w}$.

Several of the discriminant rules in this chapter can be modified to incorporate π_j and costs of correct and incorrect allocation. See Johnson and Wichern (1988, ch. 11). We will assume that costs of correct allocation are unknown or equal to 0, and that costs of incorrect allocation are unknown or equal. Unless stated otherwise, assume that the probabilities π_j that $\boldsymbol{w}_i$ is in group j are unknown or equal: $\pi_j = 1/G$ for $j = 1, ..., G$. Some rules can handle discrete predictors.

8.2 LDA and QDA

Often it is assumed that the G groups have the same covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. Then the pooled covariance matrix estimator is

$$\boldsymbol{S}_{pool} = \frac{1}{n-G}\sum_{j=1}^{G}(n_j - 1)\boldsymbol{S}_j \tag{8.2}$$

where $n = \sum_{j=1}^{G} n_j$. The pooled estimator $\boldsymbol{S}_{pool}$ can also be useful if some of the n_i are small so that the $\boldsymbol{S}_j$ are not good estimators. Let $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$ be the estimator of multivariate location and dispersion for the jth group, e.g., the sample mean and sample covariance matrix $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$. Then a pooled estimator of dispersion is

$$\hat{\boldsymbol{\Sigma}}_{pool} = \frac{1}{k-G}\sum_{j=1}^{G}(k_j - 1)\hat{\boldsymbol{\Sigma}}_j \tag{8.3}$$

where often $k = \sum_{j=1}^{G} k_j$ and often k_j is the number of cases used to compute $\hat{\boldsymbol{\Sigma}}_j$.

LDA is especially useful if the population dispersion matrices are equal: $\boldsymbol{\Sigma}_j \equiv \boldsymbol{\Sigma}$ for $j = 1, ..., G$. Then $\hat{\boldsymbol{\Sigma}}_{pool}$ is an estimator of $c\boldsymbol{\Sigma}$ for some constant $c > 0$ if each $\hat{\boldsymbol{\Sigma}}_j$ is a consistent estimator of $c_j\boldsymbol{\Sigma}$ where $c_j > 0$ for $j = 1, ..., G$. If LDA does not work well with predictors $\boldsymbol{x} = (X_1, ..., X_p)$, try adding squared terms X_i^2 and possibly two way interaction terms X_iX_j. If all squared terms and two way interactions are added, LDA will often perform like QDA.

Definition 8.3. Let $\hat{\boldsymbol{\Sigma}}_{pool}$ be a pooled estimator of dispersion. 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 $j = 1, ..., G$. *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 $j = 1, ..., G$. *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 $j = 1, ..., G$.

Examining some of the rules for $G = 2$ 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 hyperellipsoids from the two groups have little overlap, then the distance rule does well. If $\boldsymbol{\mu} = \boldsymbol{0}$, then expect nearly all of the $\boldsymbol{w}$ 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 hyperellipsoid 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 a hyperellipsoidal region.

Now suppose the G groups come from the same family of elliptically contoured $EC(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g)$ distributions where g is a continuous decreasing function that does not depend on j for $j = 1, ..., G$. For example, the jth distribution 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 where $g(z) = \exp(-z/2)$, 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 result is true since then maximizing $f_j(\boldsymbol{w})$ is equivalent to minimizing $D^2_{\boldsymbol{w}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Plugging in estimators leads to the distance rule. The same determinant assumption is a much weaker assumption than 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 $|\text{Cov}(\boldsymbol{x})| = |\text{Cov}(\boldsymbol{y})|$.

Note that if the G 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. The constant c_X is given below Equation (3.8).

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 $j = 1, ..., G$, then we want to minimize $\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1}(-2\boldsymbol{w} + \boldsymbol{\mu}_j)$ or maximize $\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1}(2\boldsymbol{w} - \boldsymbol{\mu}_j)$. Plugging in estimators 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 $p > 2$. The linear discriminant rule and distance rule are robust to nonnormality, as is the logistic regression discriminant rule if $G = 2$. Expect the distance rule to be best when the ellipsoidal covering regions of the G groups have little overlap. The distance rule can be very poor if the groups overlap and have very different variability.

Rule of thumb 8.1. It is often useful to use predictor transformations from Section 2.4 to remove nonlinearities from the predictors. The log rule is especially useful for highly skewed predictors. After making transformations, assume that there are $1 \le k \le p$ continuous predictors $X_1, ..., X_k$ where no terms like $X_2 = X_1^2$ or $X_3 = X_1 X_2$ are included. If $n_j \ge 10k$ for $j = 1, ..., G$, then make the G DD plots using the k predictors from each group to check for outliers, which could be cases that were incorrectly classified. Then use p predictors which could include squared terms, interactions, and categorical predictors. Try several discriminant rules. For a given rule, the error rates computed using the training data $\boldsymbol{x}_{i,j}$ with known groups give a lower bound on the error rates for the test data $\boldsymbol{w}_i$. 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 for the test data $\boldsymbol{w}_i$ 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 G groups, where the $\boldsymbol{x}_{i,j}$ are used for $i = 1, ..., n_j$. If all of the n_j are large, say $n_j \ge 30p$, then the plotted points should cluster tightly about the identity line in each of the G plots if the assumption of equal covariance matrices is reasonable. The linear discriminant rule has some robustness against the assumption of equal covariance matrices. See Remark 8.2.

8.3 LR

Definition 8.6. Assume that $G = 2$ 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 $\rho(\boldsymbol{w})$. 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 $\hat{\rho}(\boldsymbol{w}) < 0.5$. The training data for logistic regression are cases $(\boldsymbol{x}_i, Y_i)$ where $Y_i = j$ if the ith case is in group j for $j = 0, 1$ and $i = 1, ..., n$. Logistic regression produces an *estimated sufficient predictor* $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$. Then

$$\hat{\rho}(\boldsymbol{x}) = \frac{e^{ESP}}{1 + e^{ESP}} = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}.$$

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 $n_i \geq 5p$, then the logistic regression model was useful if the step function of observed slice proportions scatter fairly closely about the logistic curve $\hat{\rho}(ESP)$.

An extension of the above binary logistic regression model uses

$$\hat{\rho}(\boldsymbol{w}) = \frac{e^{\hat{h}(\boldsymbol{w})}}{1 + e^{\hat{h}(\boldsymbol{w})}},$$

and will be discussed below after some notation. Note that $\hat{h}(\boldsymbol{w}) > 0$ corresponds to $\hat{\rho}(\boldsymbol{w}) > 0.5$ while $\hat{h}(\boldsymbol{w}) < 0$ corresponds to $\hat{\rho}(\boldsymbol{w}) < 0.5$. LR uses $\hat{h}(\boldsymbol{w}) = ESP$, and the binary logistic GAM defined in Definition 8.9 uses $\hat{h}(\boldsymbol{w}) = ESP = EAP$. These two methods are robust to nonnormality.

Definition 8.7. In a *1D regression*, Y is independent of $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$ where $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$ for a *generalized linear model* (GLM). In a *generalized additive model* (GAM), Y is independent of $\boldsymbol{x} = (x_1, ..., x_p)^T$ given the *additive predictor* $SP = AP = \alpha + \sum_{j=1}^p S_j(x_j)$ for some (usually unknown) functions S_j. The *estimated sufficient predictor* ESP $= \hat{h}(\boldsymbol{x})$. For a GAM, the *estimated additive predictor* ESP = EAP = $\hat{\alpha} + \sum_{j=1}^p \hat{S}_j(\boldsymbol{x}_j)$. A *response plot* is a plot of ESP versus Y.

Note that a GLM is a special case of the GAM using $S_j(x_j) = \beta_j x_j$ for $j = 1, ..., p$. A GLM with SP $= \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ is a special case of a GAM with $x_3 \equiv x_1 x_2$. A GLM with SP $= \alpha + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$ is a special case of a GAM with $S_1(x_1) = \beta_1 x_1 + \beta_2 x_1^2$ and $S_2(x_2) = \beta_3 x_2$. A GLM with p terms may be equivalent to a GAM with k terms $w_1, ..., w_k$ where $k < p$.

The plotted points in the EE plot defined below should scatter tightly about the identity line if the GLM is appropriate and if the sample size is large enough so that the ESP is a good estimator of the SP and the EAP is a good estimator of the AP. If the clustering is not tight but the GAM gives a reasonable approximation to the data, as judged by the EAP–response plot, then examine the $\hat{S}_j$ of the GAM to see if some simple terms such as x_i^2 can be added to the GLM so that the modified GLM has a good ESP–response plot. (This technique is easiest if the GLM and GAM have the same p terms $x_1, ..., x_p$. The technique is more difficult, for example, if the GLM has terms x_1, x_1^2, and x_2 while the GAM has terms x_1 and x_2.)

Definition 8.8. An *EE plot* is a plot of EAP versus ESP.

Definition 8.9. Let $\rho(w) = \exp(w)/[1 + \exp(w)]$.

a) For the *binary logistic GLM*, $Y_1, ..., Y_n$ are independent with $Y|SP \sim$ *binomial*$(1, \rho(SP))$ where $\rho(SP) = P(Y = 1|SP)$. This model has $E(Y|SP)$ $= \rho(SP)$ and $V(Y|SP) = \rho(SP)(1 - \rho(SP))$.

b) For the *binary logistic GAM*, $Y_1, ..., Y_n$ are independent with $Y|AP \sim$ binomial$(1, \rho(\text{AP}))$ where $\rho(AP) = P(Y = 1|AP)$. This model has $E(Y|AP)$ $= \rho(AP)$ and $V(Y|AP) = \rho(AP)(1 - \rho(AP))$. The response plot and discriminant rule are similar to those of Definition 8.6, and the EAP–response plot adds the estimated mean function $\rho(EAP)$ and a step function to the plot. The *logistic GAM discriminant rule* allocates $\boldsymbol{w}$ to group 1 if $\hat{\rho}(\boldsymbol{w}) \geq 0.5$ and allocates $\boldsymbol{w}$ to group 0 if $\hat{\rho}(\boldsymbol{w}) < 0.5$ where

$$\hat{\rho}(\boldsymbol{w}) = \frac{e^{EAP}}{1 + e^{EAP}}$$

and $EAP = \hat{\alpha} + \sum_{j=1}^{p} \hat{S}_j(\boldsymbol{w}_j)$.

Rule of thumb 8.2. For binary data, Kay and Little (1987) suggest examining the two distributions $x|Y = 0$ and $x|Y = 1$. Use predictor x if the two distributions are roughly symmetric with similar spread. Use x and x^2 if the distributions are roughly symmetric with different spread. Use x and $\log(x)$ if one or both of the distributions are skewed. From Section 2.4, recall that the log rule says add $\log(x)$ to the model if $\min(x) > 0$ and $\max(x)/\min(x) > 10$. The rules shown in Table 8.1 are used if x is an indicator variable or if x is a continuous variable. Replace normality by "symmetric with similar spreads" and "symmetric with different spreads" in the second and third lines of the table.

Example 8.1. The ICU data is available from the text's website and from STATLIB (http://lib.stat.cmu.edu/DASL/Datafiles/ICU.html). Also

Table 8.1 Building the Logistic Regression Model

distribution of $x\|y = i$	variables to include in the model
$x\|y = i$ is an indicator	x
$x\|y = i \sim N(\mu_i, \sigma^2)$	x
$x\|y = i \sim N(\mu_i, \sigma_i^2)$	x and x^2
$x\|y = i$ has a skewed distribution	x and $\log(x)$
$x\|y = i$ has support on (0,1)	$\log(x)$ and $\log(1 - x)$

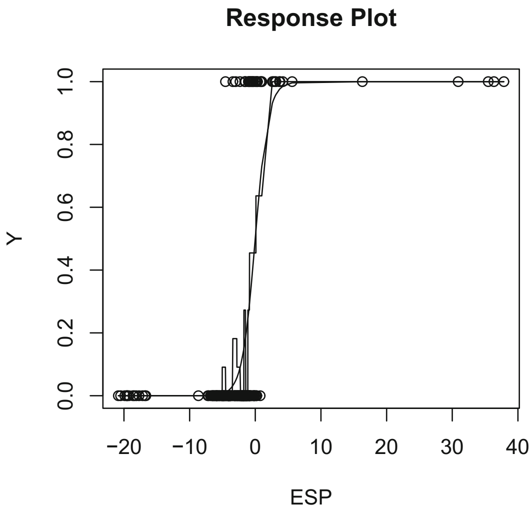

Fig. 8.1 Visualizing the ICU Data With a GLM

see, Hosmer and Lemeshow (2000, pp. 23–25) and Olive (2013b). The survival of 200 patients following admission to an intensive care unit was studied with logistic regression. The response variable was STA (0 = Lived, 1 = Died). Predictors were AGE, SEX (0 = Male, 1 = Female), RACE (1 = White, 2 = Black, 3 = Other), SER= Service at ICU admission (0 = Medical, 1 = Surgical), CAN= Is cancer part of the present problem? (0 = No, 1 = Yes), CRN= History of chronic renal failure (0 = No, 1 = Yes), INF= Infection probable at ICU admission (0 = No, 1 = Yes), CPR= CPR prior to ICU admission (0 = No, 1 = Yes), SYS= Systolic blood pressure at ICU admission (in mm Hg), HRA= Heart rate at ICU admission (beats/min), PRE= Previous admission to an ICU within 6 months (0 = No, 1 = Yes), TYP= Type of admission (0 = Elective, 1 = Emergency), FRA= Long bone, multiple, neck, single area, or hip fracture (0 = No, 1 = Yes), PO2= PO2 from initial blood gases (0 if $>$ 60, 1 if $\leq$ 60), PH= PH from initial blood gases (0 if $\geq$ 7.25, 1 if $<$ 7.25), PCO= PCO2 from initial blood gases (0 if $\leq$ 45, 1 if $>$ 45), Bic= Bicarbonate from initial blood gases

(0 if $\geq$ 18, 1 if $<$ 18), CRE= Creatinine from initial blood gases (0 if $\leq$ 2.0, 1 if $>$ 2.0), and LOC= Level of consciousness at admission (0 = no coma or stupor, 1= deep stupor, 2 = coma).

Factors LOC and RACE had two indicator variables to model the three levels. The response plot in Figure 8.1 shows that the logistic regression model using the 19 predictors is useful for predicting survival, although the output has $\hat{\rho}(\boldsymbol{x}) = 1$ or $\hat{\rho}(\boldsymbol{x}) = 0$ exactly for some cases. Note that the step function of slice proportions tracks the model logistic curve fairly well. Also note that cases with ESP values greater that 0 get classified in group 1 (died), and in group 0 (survived), otherwise.

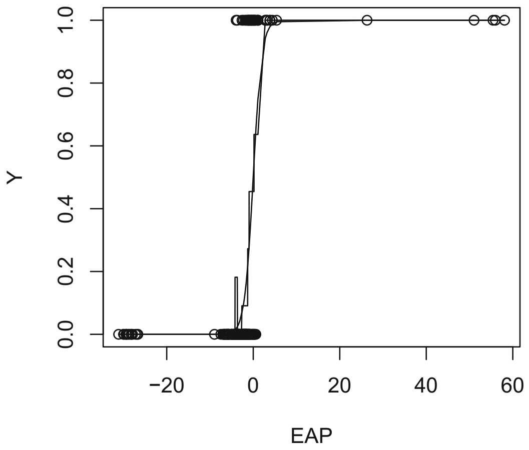

Fig. 8.2 Visualizing the ICU GAM

Next, a binary generalized additive model was fit with unspecified functions for AGE, SYS, and HRA, and linear functions for the remaining sixteen variables. Output suggested that functions for SYS and HRA are linear but the function for AGE may be slightly curved. Several cases had $\hat{\rho}(AP)$ equal to zero or one, but the response plot in Figure 8.2 suggests that the full model is useful for predicting survival. Note that the ten slice step function closely tracks the logistic curve. Figure 8.3 shows the plot of EAP versus ESP from the binary logistic regression. The plot shows that the near zero and near one probabilities are handled differently by the GAM and GLM, but the estimated success probabilities for the two models are similar: $\hat{\rho}(ESP) \approx \hat{\rho}(EAP)$. Some R commands for producing the above figures are shown below. The R library `mgcv` for fitting GAMs is described in Wood (2006). The `mpack`

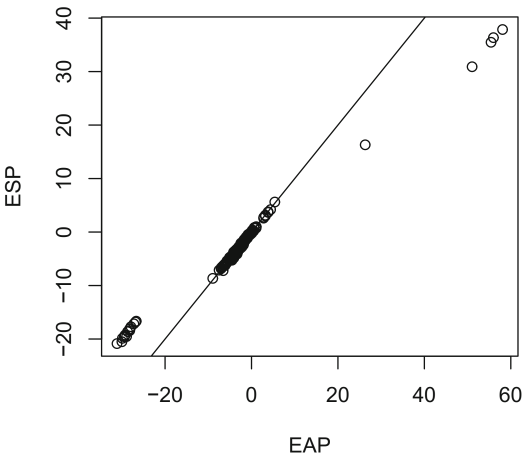

Fig. 8.3 GAM and GLM give Similar Success Probabilities

function `lrplot3` is needed to make the response plot, and the plots may look different due to *R* changes for fitting GAMs.

```
##ICU data from Statlib or URL
#http://lagrange.math.siu.edu/Olive/ICU.lsp
#delete header of ICU.lsp and delete last parentheses
#at the end of the file. Save the file on F drive as
#icu.txt.

icu <- read.table("F:\\icu.txt")

names(icu) <- c("ID", "STA", "AGE", "SEX", "RACE",
"SER", "CAN", "CRN", "INF", "CPR", "SYS", "HRA",
"PRE", "TYP", "FRA", "PO2", "PH", "PCO", "Bic",
"CRE", "LOC")

icu[,5] <- as.factor(icu[,5])
icu[,21] <- as.factor(icu[,21])
icu2<-icu[,-1]
outf <- glm(formula=STA~.,family=binomial,data=icu2)
ESP <- predict(outf)

library(mgcv)
outgam <- gam(STA ~ s(AGE)+SEX+RACE+SER+CAN+CRN+INF+
CPR+s(SYS)+s(HRA)+PRE+TYP+FRA+PO2+PH+PCO+Bic+CRE+LOC,
family=binomial,data=icu2)
EAP <- predict.gam(outgam)
```

```
plot(EAP,ESP)
abline(0,1)
#Figure 8.3

Y <- icu2[,1]
lrplot3(ESP=EAP,Y,slices=18)
#Figure 8.2

lrplot3(ESP,Y,slices=18)
#Figure 8.1
```

8.4 KNN

The K-nearest neighbors (KNN) method identifies the K cases in the training data that are closest to $\boldsymbol{w}$. Suppose m_j of the K cases are from group j. Then the KNN estimate of $p_j(\boldsymbol{w}) = P(Y = j|\boldsymbol{W} = \boldsymbol{w}) = P(\boldsymbol{w}$ is from the jth group) is $\hat{p}_j(\boldsymbol{w}) = m_j/K$. (Actually, $m_j/K \approx c p_j(\boldsymbol{w})$ so $m_j/m_k \approx p_j(\boldsymbol{w})/p_k(\boldsymbol{w})$. See the end of this section.) Applying the Bayesian discriminant rule to the $\hat{p}_j(\boldsymbol{w})$ gives the KNN discriminant rule.

Definition 8.10. The K-nearest neighbors (KNN) discriminant rule allocates $\boldsymbol{w}$ to group a if m_a maximizes m_j for $j = 1, ..., G$.

A couple of examples will be useful. When $K = 1$, find the case in the training data closest to $\boldsymbol{w}$. If that training data case is from group j, then allocate $\boldsymbol{w}$ to group j. Suppose n_j is the largest n_k for $k = 1, ..., G$. Hence group j is the group with the most training data cases. Then if $K = n$, $\boldsymbol{w}$ is always allocated to group j. The $K = n$ rule is bad. The $K = 1$ rule is surprisingly good but tends to have low bias and high variability. Generally, values of $K > 1$ will have smaller test error rates.

For KNN and other discriminant analysis rules, it is often useful to standardize the data so that all variables have a sample mean of 0 and sample standard deviation of 1. The `scale` function in *R* can be used to standardize data.

To see why KNN might be reasonable, let D_ϵ be a hypersphere of radius ϵ centered at $\boldsymbol{w}$. Since the pdf $f_j(\boldsymbol{x})$ is continuous, there exists $\epsilon > 0$ small enough such that $f_j(\boldsymbol{x}) \approx f_j(\boldsymbol{w})$ for all $\boldsymbol{x} \in D_\epsilon$ and for each $j = 1, ..., G$. If $\boldsymbol{z}$ is a random vector from a distribution with pdf $f_j(\boldsymbol{x})$, then $P_j(\boldsymbol{z} \in D_\epsilon) =$

$$\int_{D_\epsilon} f_j(\boldsymbol{x}) d\boldsymbol{x} \approx f_j(\boldsymbol{w}) \int_{D_\epsilon} 1 d\boldsymbol{x} = f_j(\boldsymbol{w}) Vol(D_\epsilon) = f_j(\boldsymbol{w}) \frac{2\pi^{p/2}}{p\Gamma(p/2)} \epsilon^p.$$

Here P_j denotes the probability when the distribution has pdf $f_j(\boldsymbol{x})$.

If for $i = 1, ..., n$, the $\boldsymbol{z}_i$ are iid from a distribution with pdf $f_j(\boldsymbol{x})$, ϵ is fixed, and if $f_j(\boldsymbol{w}) > 0$, then the number of $\boldsymbol{z}_i$ in D_ϵ is proportional to n. Hence if the number of $\boldsymbol{z}_i$ in D_ϵ is proportional to n^δ with $0 < \delta < 1$, then $\epsilon \to 0$. So if $K/n \to 0$ in KNN, then the hypersphere containing the K cases has radius $\epsilon \to 0$ as $n \to \infty$. Hence the above approximations will be valid for large n. Note that if $p = 1$, then D_ϵ is the line segment $(w - \epsilon, w + \epsilon)$ and $Vol(D_\epsilon) = 2\epsilon =$ length of the line segment. If $p = 2$, then D_ϵ is the circle of radius ϵ centered at $\boldsymbol{w}$ and $Vol(D_\epsilon) = \pi\epsilon^2 =$ the area of the circle. If $p = 3$, then D_ϵ is the sphere of radius ϵ centered at $\boldsymbol{w}$ and $Vol(D_\epsilon) = 4\pi\epsilon^3/3 =$ the volume of the sphere.

Now suppose that the training data $\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{n_G,G}$ is a random sample from the G populations so that $n_j/n \xrightarrow{P} \pi_j$ as $n \to \infty$ for $j = 1, ..., G$. Then for ϵ small and K large, $m_j/K \approx$

$$P(\boldsymbol{W} \in D_\epsilon, Y = j) = P(\boldsymbol{W} \in D_\epsilon | Y = j) P(Y = j) \approx \pi_j f_j(\boldsymbol{w}) Vol(D_\epsilon).$$

Now $P(\boldsymbol{W} \in D_\epsilon) = \sum_{j=1}^G P(\boldsymbol{W} \in D_\epsilon, Y = j) =$
$\sum_{j=1}^G P(\boldsymbol{W} \in D_\epsilon | Y = j) P(Y = j)$ since the sets $\{Y = j\}$ form a disjoint partition. Hence

$$P(Y = k | \boldsymbol{W} \in D_\epsilon) = \frac{P(Y = k, \boldsymbol{W} \in D_\epsilon)}{P(\boldsymbol{W} \in D_\epsilon)} = \frac{P(\boldsymbol{W} \in D_\epsilon) | Y = k) P(Y = k)}{P(\boldsymbol{W} \in D_\epsilon)}$$

$$\approx \frac{\pi_k f_k(\boldsymbol{w}) Vol(D_\epsilon)}{\sum_{j=1}^G \pi_j f_j(\boldsymbol{w}) Vol(D_\epsilon)},$$

which is the quantity used by the Bayes classifier since the constant $Vol(D_\epsilon)$ cancels. This argument can also be used to justify Equation (8.1). Since the denominator is a constant, allocating $\boldsymbol{w}$ to group a with the largest m_a/K, or equivalently with the largest m_a, approximates the Bayes classifier if n is very large, K is large, and ϵ is very small.

This approximation likely needs unrealistically large n, especially if p is large and $\boldsymbol{w}$ is in a region where there is a lot of group overlap. However, KNN often works well in practice. Silverman (1986, pp. 96–100) also discusses using KNN to find an estimator $\hat{f}(\boldsymbol{w})$ of $f(\boldsymbol{w})$.

As claimed above Definition 8.10, note, for large K and small ϵ, that

$$m_j/K \approx P(\boldsymbol{W} \in D_\epsilon, Y = j) = P(Y = j | \boldsymbol{W} \in D_\epsilon) P(\boldsymbol{W} \in D_\epsilon) \approx$$

$$cP(Y = j | \boldsymbol{W} = \boldsymbol{w}) = cp_k(\boldsymbol{w})$$

where $c = P(\boldsymbol{W} \in D_\epsilon)$.

8.5 FDA

The FDA method of discriminant analysis, a special case of the generalized eigenvalue problem, finds eigenvalue eigenvector pairs so that the $\hat{\boldsymbol{e}}_1^T \boldsymbol{x}_{ij}$ have low variability in each group, but the variability of the $\hat{\boldsymbol{e}}_1^T \boldsymbol{x}_{ij}$ between groups is large. More precisely, let $\hat{\boldsymbol{W}}$ be a $p \times p$ dispersion matrix used to measure variability within groups and let $\hat{\boldsymbol{B}}$ be a $p \times p$ symmetric matrix used to measure variability between classes. Let the eigenvalue eigenvector pairs of a matrix $\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{B}}$ 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 from Theorem 1.1 f), $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T \hat{\boldsymbol{B}} \boldsymbol{a}}{\boldsymbol{a}^T \hat{\boldsymbol{W}} \boldsymbol{a}} = \hat{\lambda}_1$, the largest eigenvalue of $\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{B}}$. The value of $\boldsymbol{a}$ that achieves the max is the eigenvector $\hat{\boldsymbol{e}}_1$. Then $\hat{\boldsymbol{e}}_2$ will achieve the max among all unit vectors orthogonal to $\hat{\boldsymbol{e}}_1$. Similarly, $\hat{\boldsymbol{e}}_3$ will achieve the max among all unit vectors orthogonal to $\hat{\boldsymbol{e}}_1$ and $\hat{\boldsymbol{e}}_2$, et cetera.

Many choices of $\hat{\boldsymbol{W}}$ have been suggested. Typically assume rank$(\hat{\boldsymbol{W}}) = p$ and rank$(\hat{\boldsymbol{B}}) = \min(p, G-1)$. Let $q \leq \min(p, G-1)$ be the number of nonzero eigenvalues $\hat{\lambda}_i$ of $\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{B}}$. Let $(T_i, \boldsymbol{C}_i)$ be an estimator of multivariate location and dispersion for the ith group. Let $\overline{T} = \dfrac{1}{G}\sum_{i=1}^G T_i$. Let $\hat{\boldsymbol{B}}_T = \sum_{i=1}^G (T_i - \overline{T})(T_i - \overline{T})^T$. Note that $\hat{\boldsymbol{B}}_T/(G-1)$ is the sample covariance matrix of the $T_1, ..., T_G$. Let $\hat{\boldsymbol{W}}_T = \sum_{i=1}^G \boldsymbol{C}_i$. Typically $(T_i, \boldsymbol{C}_i) = (\overline{\boldsymbol{x}}_i, \boldsymbol{S}_i)$ is used where the notation $\overline{T} = \overline{\overline{\boldsymbol{x}}}$ is used. Let $\hat{\boldsymbol{B}}_B = \sum_{i=1}^G \hat{\pi}_i (T_i - \overline{T})(T_i - \overline{T})^T$, and $\hat{\boldsymbol{W}}_B = \sum_{i=1}^G \hat{\pi}_i \boldsymbol{C}_i$. Let $\hat{\boldsymbol{W}}_L = G\hat{\boldsymbol{\Sigma}}_{pool}$. See Equation (8.3). Let $\boldsymbol{A} = (a_{ij})$ be a $p \times p$ matrix, and let $diag(\boldsymbol{A}) = diag(a_{11}, ..., a_{pp})$ be the diagonal matrix with the a_{ii} along the diagonal. Let $\hat{\boldsymbol{W}}_D = diag(\hat{\boldsymbol{W}}_A)$ for any previously defined $\hat{\boldsymbol{W}}_A$, e.g., $A = T$. Then $\hat{\boldsymbol{W}}_D$ is nonsingular if all $w_{ii} > 0$ even if $\hat{\boldsymbol{W}}_A = (w_{ij})$ is singular. Sometimes $\overline{T}_B = \sum_{i=1} \hat{\pi}_i T_i$ is used instead of $\overline{T}$. The rule may also use $\hat{\boldsymbol{B}} = c_1 \hat{\boldsymbol{B}}_A$ and $\hat{\boldsymbol{W}} = c_2 \hat{\boldsymbol{W}}_A$ for positive constants c_1 and c_2, e.g., $c_1 = 1/(G-1)$ and $c_2 = 1/(n-G)$.

The FDA rule finds $\hat{\boldsymbol{e}}_1$ and summarizes the group by the linear combination $\hat{\boldsymbol{e}}_1^T T_i$. Then FDA allocates $\boldsymbol{w}$ to the group a for which $\hat{\boldsymbol{e}}_1^T \boldsymbol{w}$ is closest to $\hat{\boldsymbol{e}}_1^T T_a$. (We can view $\hat{\boldsymbol{e}}_1^T T_i$ as a summary of the n_i linear combinations of the predictors $\hat{\boldsymbol{e}}_1^T \boldsymbol{x}_{ij}$ in the ith group where $j = 1, ..., n_i$.) The FDA method should work well if the within group variability is small and the between group variability is large.

Definition 8.11. For *Fisher's discriminant analysis* (FDA), the *FDA discriminant rule* allocates $\boldsymbol{w}$ to group a that minimizes $|\hat{\boldsymbol{e}}_1^T \boldsymbol{w} - \hat{\boldsymbol{e}}_1^T T_i|$ for $i = 1, ..., G$.

Remark 8.1. a) Often it is suggested to use PCA for DA: find D such that the first D principal components explain at least 95% of the variance. Then use the $D \leq \min(n, p)$ principal components as the variables. The problem with this idea is that principal components are used to explain the structure of the dispersion matrix of the data, not as linear combinations of the data that are good for DA. Using the J linear combinations from FDA such that

$$\sum_{i=1}^{J} \hat{\lambda}_i / \sum_{i=1}^{p} \hat{\lambda}_i \geq 0.95$$

might be a better choice for DA, especially if the number of nonzero eigenvalues q is not too small.

b) Often DA rules from the other FDA eigenvectors simply replace $\hat{\boldsymbol{e}}_1$ with $\hat{\boldsymbol{e}}_j$. It might be better to consider J rules such that $(\hat{\boldsymbol{e}}_1^T \boldsymbol{w}, ..., \hat{\boldsymbol{e}}_k^T \boldsymbol{w})^T$ is closest to $(\hat{\boldsymbol{e}}_1^T T_a, ..., \hat{\boldsymbol{e}}_k^T T_a)^T$ for $k = 1, ..., J$ where $a \in \{1, ..., G\}$ and J is as in Remark 8.1 a). Or let $\hat{\boldsymbol{V}} = [\hat{\boldsymbol{e}}_1 \ \hat{\boldsymbol{e}}_2 \ \cdots \ \hat{\boldsymbol{e}}_q]$. Then allocate $\boldsymbol{w}$ to group a that minimizes $D_j^2(\boldsymbol{w})$ where $D_j^2(\boldsymbol{w}) = (\boldsymbol{w} - T_j)^T \hat{\boldsymbol{V}} \hat{\boldsymbol{V}}^T (\boldsymbol{w} - T_j)^T - 2\log(\hat{\pi}_j)$ where $\hat{\boldsymbol{W}}_B$ and $\hat{\boldsymbol{B}}_B$ are used. See Filzmoser et al. (2006).

c) If $\hat{\boldsymbol{W}}$ is singular and $\hat{\boldsymbol{B}}$ is nonsingular, then the eigenvalue eigenvector pair(s) corresponding to the smallest nonzero eigenvalue(s) of $\hat{\boldsymbol{B}}^{-1}\hat{\boldsymbol{W}}$ may be of interest, as argued below Theorem 1.1.

Following Koch (2014, pp. 120–124) closely, consider the population version of FDA where the ith group has mean and covariance matrix $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\boldsymbol{x}_i})$ for $i = 1, ..., G$ where $\boldsymbol{x}_i$ is a random vector from the population corresponding to the ith group. Let $\overline{\boldsymbol{\mu}} = \frac{1}{G}\sum_{i=1}^G \boldsymbol{\mu}_i$, $\boldsymbol{B} = \sum_{i=1}^G (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^T$, and $\boldsymbol{W} = \sum_{i=1}^G \boldsymbol{\Sigma}_{\boldsymbol{x}_i}$. Then the *between group variability*

$$b(\boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a} = \sum_{i=1}^{G} |\boldsymbol{a}^T(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})|, \tag{8.4}$$

and the *within group variability* =

$$w(\boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{W} \boldsymbol{a} = \sum_{i=1}^{G} \boldsymbol{a}^T \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \boldsymbol{a} = \sum_{i=1}^{G} \mathrm{Var}(\boldsymbol{a}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}) \tag{8.5}$$

since $\mathrm{Var}(\boldsymbol{a}^T \boldsymbol{x}_i) = E[(\boldsymbol{a}^T \boldsymbol{x}_i - E(\boldsymbol{a}^T \boldsymbol{x}_i))^2] = E[\boldsymbol{a}^T(\boldsymbol{x}_i - E(\boldsymbol{x}_i))(\boldsymbol{x}_i - E(\boldsymbol{x}_i))^T \boldsymbol{a}]$ $= \boldsymbol{a}^T \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \boldsymbol{a}$. Then

$$\max_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{b(\boldsymbol{a})}{w(\boldsymbol{a})} = \max_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{W} \boldsymbol{a}}$$

is achieved by $\boldsymbol{a} = \boldsymbol{e}_1$, the eigenvector corresponding to the largest eigenvalue $\lambda_1(\boldsymbol{W}^{-1}\boldsymbol{B})$ of $\boldsymbol{W}^{-1}\boldsymbol{B}$. Hence $b(\boldsymbol{e}_1)$ is large while $w(\boldsymbol{e}_1)$ is small in that the ratio is a max.

FDA approximates Equations (8.4) and (8.5) by using $\hat{\boldsymbol{B}}_T$ and $\hat{\boldsymbol{W}}_T$ with $(T_i, \boldsymbol{C}_i) = (\overline{\boldsymbol{x}}_i, \boldsymbol{S}_i)$. Note that $\boldsymbol{W}/G$ tends not to be a good estimator of dispersion unless the G groups have the same covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}_i} = \boldsymbol{\Sigma}_{\boldsymbol{x}}$ for $i = 1, ..., G$, but $w(\boldsymbol{a})$ is a good measure of within group variability even if the $\boldsymbol{\Sigma}_{\boldsymbol{x}_i}$ are not equal. Also, if $\hat{\boldsymbol{W}}_A$ is such that $\boldsymbol{a}^T \hat{\boldsymbol{W}}_A \boldsymbol{a}$ can be made small, then FDA will likely work well with $\hat{\boldsymbol{B}}_T$ and $\hat{\boldsymbol{W}}_A$ if there are no outliers.

Remark 8.2. If $G = 2$, $(T_i, \boldsymbol{C}_i) = (\overline{\boldsymbol{x}}_i, \boldsymbol{S}_i)$, $\hat{\boldsymbol{B}} = \hat{\boldsymbol{B}}_T$, and $\hat{\boldsymbol{W}} = 2\boldsymbol{S}_{pool}$, then LDA and FDA are equivalent. See Koch (2014, p. 129). This result helps explain why LDA works well on so many data sets.

Two special cases are illustrative. First, let $\hat{\boldsymbol{W}} = \boldsymbol{I}_p$ and use $\hat{\boldsymbol{B}}_T$. Then FDA attempts to find a vector $\hat{\boldsymbol{e}}_1$ such that the $\hat{\boldsymbol{e}}_1^T T_i$ are far from $\hat{\boldsymbol{e}}_1^T \overline{T}$. Then find group a such that $\hat{\boldsymbol{e}}_1^T \boldsymbol{w}$ is closer to $\hat{\boldsymbol{e}}_1^T T_a$ than to $\hat{\boldsymbol{e}}_1^T T_i$ for $i \neq a$. Second, consider $G = 2$. Then $\hat{\boldsymbol{B}}_T = (T_1 - T_2)(T_1 - T_2)^T/2$. Using Theorem 1.1a) with $\boldsymbol{d} = (T_1 - T_2)/\sqrt{2}$ shows that $\hat{\boldsymbol{e}}_1 = \dfrac{\hat{\boldsymbol{W}}^{-1}(T_1 - T_2)}{\|\hat{\boldsymbol{W}}^{-1}(T_1 - T_2)\|}$. If the $\hat{\boldsymbol{W}}^{-1}\boldsymbol{x}_{ij}$ are "standardized data," and the $\hat{\boldsymbol{W}}^{-1}T_i$ are standardized centers for $i = 1, 2$, then FDA projects $\boldsymbol{w}$ on the line between the standardized centers and allocates $\boldsymbol{w}$ to the group with the standardized center closest to $\hat{\boldsymbol{e}}_1^T \boldsymbol{w}$.

```
library(MASS) ##Use ?lda
out <- lda(as.matrix(iris[, 1:4]), iris$Species)
names(out); out; plot(out) #plots LD1 versus LD2
Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333
Group means:
           Sep.Len Sep.Wid Pet.Len Pet.Wid
setosa       5.006   3.428   1.462   0.246
versicolor   5.936   2.770   4.260   1.326
virginica    6.588   2.974   5.552   2.026
Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  0.8293776  0.02410215
Sepal.Width   1.5344731  2.16452123
Petal.Length -2.2012117 -0.93192121
Petal.Width  -2.8104603  2.83918785
Proportion of trace:
   LD1    LD2
0.9912 0.0088

gp <- as.integer(iris$Species)
x <- as.matrix(iris[,1:4])            #AER 0.02
out<- lda(x,gp); 1-mean(predict(out,x)$class==gp)
```

```
plot(out) #Get numbers in Figure 8.4.
```

Example 8.2. The library *MASS* has a function `lda` that does FDA. The famous iris data set has variables x_1 = sepal length, x_2 = sepal width, x_3 = petal length, and x_4 = petal width. There are three groups corresponding to types of iris: setosa, versicolor, and virginica. The above *R* code performs FDA. Figure 8.4 shows the plot of LD1 $= \hat{\boldsymbol{e}}_1$ versus LD2 $= \hat{\boldsymbol{e}}_2$. Since the proportion of trace for LD2 is small, LD2 is not needed. Note that LD1 separates setosa from the other two types of iris, and versicolor and virginica are nearly separated.

Let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{e}}_1 =$ LD1 be the first eigenvector from FDA. The function `FDAboot` bootstraps $\hat{\boldsymbol{\beta}}$ and gives the nominal 95% shorth CIs. Also shown below is the sample mean vector of the bootstrapped $\hat{\boldsymbol{\beta}}^*_i$ where $i = 1, ..., B = 1000$. The bootstrap is performed by taking samples of size n_i with replacement from each group for $i = 1, ..., G$. Perform FDA on the combined sample to get $\hat{\boldsymbol{\beta}}^*_j$. Since $\hat{\boldsymbol{\beta}}$ is an eigenvector, the bootstrapped eigenvector could estimate $\hat{\boldsymbol{\beta}}$ or $-\hat{\boldsymbol{\beta}}$. Pick a $\hat{\beta}_j$ that is large in magnitude, and see how many times the $\hat{\beta}^*_j$ have the same sign as $\hat{\beta}_j$. Multiply the bootstrap vector by -1 if it has opposite sign. In the output below, all $B = 1000$ bootstrap vectors had $\hat{\beta}^*_4 < 0$.

```
#Sample sizes may not be large enough for the
#shorth CI coverage to be close to the nominal 95%
out<-FDAboot(x,gp)
apply(out$betas,2,mean)
[1]  0.8468  1.5807 -2.2558 -2.9180
sum(out$betas[,4]<0) #all betahat^*
[1] 1000  #estimate betahat, not -betahat
ddplot4(out$betas) #right click Stop
#covers the identity line
out$shorci[[1]]$shorth
[1] 0.3148 1.4634
out$shorci[[2]]$shorth
[1] 0.7745 2.3096
out$shorci[[3]]$shorth
[1] -2.9276 -1.6260
out$shorci[[4]]$shorth
[1] -3.8609 -1.8875
```

Next, *R* code is given for robust FDA. The function `getUbig` gets the RMVN set U_i for each group for $i = 1, ..., G$ and combines the sets into one large data set. Then RFDA is the classical FDA applied to this cleaned data set. See Section 8.9 and the output below. Like the robust biplot, Figure 8.5 only uses the cleaned cases since outliers could obscure the plot, and this technique can distort the amount of group overlap.

```
tem<-getubig(x,gp) ##Robust FDA
outr<-lda(tem$Ubig,tem$grp)
1-mean(predict(outr,x)$class==gp) #AER 0.03
plot(outr)
outr
```

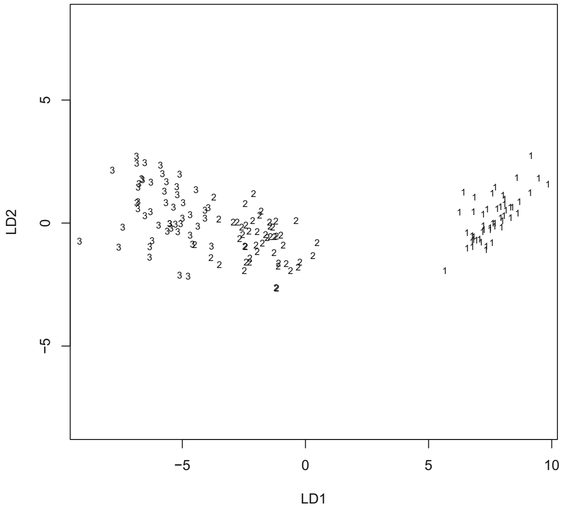

Fig. 8.4 Plot of LD1 versus LD2 for the iris data

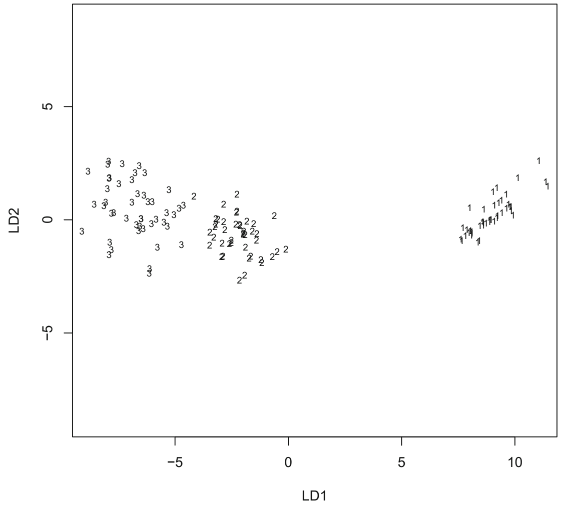

Fig. 8.5 RFDA Plot of LD1 versus LD2 for the iris data

```
Prior probabilities of groups:
        1         2         3
0.3206107 0.3282443 0.3511450
Group means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     5.026190    3.438095     1.464286   0.2309524
2     5.923256    2.813953     4.234884   1.3093023
3     6.486957    2.950000     5.454348   2.0173913
Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  0.4281837 -0.06899442
Sepal.Width   2.5221645  2.01270912
Petal.Length -2.3230167 -1.11944258
Petal.Width  -3.2947263  3.25076179
Proportion of trace:
   LD1    LD2
0.9942 0.0058
```

The covmb2 subset B can be found when $p < n$ or $p \geq n$. See Section 4.7. The function getBbig gets the set B_i for each group for $i = 1, ..., G$ and combines the sets into one large data set. Then a robust FDA is the classical FDA applied to this cleaned data set. For the iris data, using covmb2 did not discard any cases, so the robust FDA and classical FDA had identical output. See Section 8.9 and the *R* code below.

```
#Robust FDA with covmb2 set B from each group.
#This subset of cases can be found when p > n.
tem<-getBbig(x,gp)
outr<-lda(tem$Bbig,tem$grp)      #AER 0.02
plot(outr); 1-mean(predict(outr,x)$class==gp)
outr #Output is same as that for classical FDA.
```

8.6 The Kernel Density Estimator

Definition 8.12. Let $K(\boldsymbol{z})$ be a joint 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

$$K(\boldsymbol{z}) = \frac{p\Gamma(p/2)}{2\pi^{p/2}} I(\boldsymbol{z}^T\boldsymbol{z} \le 1)$$

so

$$\hat{f}(\boldsymbol{z}) = \frac{p\Gamma(p/2)}{2\pi^{p/2}} \ \frac{1}{n} \ \frac{1}{h^p} \ \sum_{i=1}^{n} I(\|\boldsymbol{z} - \boldsymbol{x}_i\|^2 \le h^2).$$

Following Silverman (1986, pp. 84–85), we 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) $= f(\boldsymbol{w})\dfrac{2\pi^{p/2}}{p\Gamma(p/2)}r^p \propto r^p$. 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 $\sqrt{n}$.

Example 8.3 in Section 8.8 will consider two toy methods of DA. To define the kernel density estimator used in Method 2, let $v_j = \lceil 2\sqrt{n_j}\rceil$ and let $r_j^2 = \|\boldsymbol{x}_{i,j} - \overline{\boldsymbol{x}}_j\|^2_{(v_j)} = D^2_{(v_j)}(\overline{\boldsymbol{x}}_j, \boldsymbol{I}_p)$ where the n_j $\boldsymbol{x}_{i,j}$ are in group j. Hence the hypersphere centered at $\overline{\boldsymbol{x}}_j$ with radius r_j contains $\approx 2\sqrt{n_j}$ of the $\boldsymbol{x}_{i,j}$ in group j. Then the kernel density estimator used in Method 2 is

$$\hat{f}_j(\boldsymbol{w}) = \frac{p\Gamma(p/2)}{2\pi^{p/2}} \ \frac{1}{n_j} \ \frac{1}{(r_j)^p} \ \sum_{i=1}^{n_j} I(\|\boldsymbol{w} - \boldsymbol{x}_{i,j}\|^2 \le r_j^2) \qquad (8.6)$$

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. This kernel density estimator was also used in the `hdr` function for the Hyndman (1996) large sample prediction region (5.8).

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 $p > 2$, speed is more important than asymptotic optimality. Also, only a crude estimator is needed since if $f_a(\boldsymbol{w})$ is the pdf that maximizes $f_j(\boldsymbol{w})$, the method only needs $\hat{f}_a(\boldsymbol{w})$ 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).

8.7 Estimating the Test Error

Definition 8.13. The test error rate L_n is the population proportion of misclassification errors made by the DA method.

The Bayes classifier has the smallest expected test error, but the Bayes classifier generally cannot be computed used since the π_k and f_k are unknown. If it was known that $\pi_1 = 0.9$, a simple DA rule would be to always allocate $\boldsymbol{w}$ to group 1. Then the test error of this rule would be $L_n = 0.1$.

Generally the test error L_n needs to be estimated by $\hat{L}_n$. A simple method for estimating the test error is to apply the DA method to the training data and find the proportion of classification errors made. To help see why this method is poor, consider KNN with $K = 1$. Then the training data is perfectly classified with a training error rate of 0, although the test error rate may be quite high.

Definition 8.14. The *training error rate* or *apparent error rate* (AER) is

$$AER = \hat{L}_n = \frac{1}{n} \sum_{i=1}^{n_j} \sum_{j=1}^{G} I[\hat{Y}_{ij} \neq Y_{ij}]$$

where $\hat{Y}_{ij}$ is the DA estimate of Y_{ij} using all n training cases $\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{G,n_G}$. Note that $Y_{ij} = j$ since $\boldsymbol{x}_{ij}$ comes from the jth group. 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 $m_A = \sum_{j=1}^{G} m_j$ of the $n = \sum_{j=1}^{G} n_j$ training cases are correctly classified, then $AER = 1 - m_A/n$.

DA methods fit the training data better than test data, so the AER tends to underestimate the error rate for test data. We want to use a DA method with a low test error rate. Cross validation (CV) divides the training data into a big part and a small part, perhaps J times. For each of the J divisions, the DA rule is computed for the big part and applied to the small part. Hence the small part is used as a validation set. The proportion of errors made for the small part is recorded.

For leave one out or delete one cross validation, $J = n$, the big part uses $n - 1$ cases from the training data while the small part uses the 1 case left out of the big part. This case will either be correctly or incorrectly classified. The leave one out CV rule can sometimes be rapidly computed, but usually requires the DA method to be fit n times.

Definition 8.15. An estimator of the test error rate is the *leave one out cross validation* error rate

$$\hat{L}_n = \frac{1}{n} \sum_{i=1}^{n_j} \sum_{j=1}^{G} I(\hat{Y}_{ij} \neq Y_{ij})$$

where $\hat{Y}_{ij}$ is the estimate of Y_{ij} when $\boldsymbol{x}_{ij}$ is deleted from the n training cases $\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{G,n_G}$. Note that $\hat{L}_n$ is the proportion of training cases that are misclassified by the n leave one out rules. If m_C is the number of cases correctly classified by leave one out classification, then $\hat{L}_n = 1 - m_C/n$.

For KNN, find the K cases in the training data closest to $\boldsymbol{x}_{i,j}$ not including $\boldsymbol{x}_{i,j}$. Then compute the leave one out cross validation error rate as in Definition 8.15.

Assume that the training data $\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{n_G,G}$ is a random sample from the G populations so that $n_j/n \xrightarrow{P} \pi_j$ as $n \to \infty$ for $j = 1, ..., G$. Hence n_j/n is a consistent estimator of π_j. Following Devroye and Wagner (1982), when $K = 1$ the test error rate L_n of KNN method converges in probability to L where $L_B \leq L \leq 2L_B$ and L_B is the test error rate of the Bayes classifier. If $K_n \to \infty$ and $K_n/n \to 0$ as $n \to \infty$, then the KNN method converges to the Bayes classifier in that the KNN test error rate $L_n \xrightarrow{P} L_B$. Then the leave one out cross validation error rate $\hat{L}_n$ is a good estimator of L_n in that $2e^{-2n\epsilon^2}$ was usually an upper bound on $P[|\hat{L}_n - L_n| \geq \epsilon]$ for small $\epsilon > 0$.

For the method below, $J = 1$ and the validation set or hold-out set is the small part of the data. Typically, 10% or 20% of the data is randomly selected to be in the validation set. Note that the DA method is only computed once to compute the error rate.

Definition 8.16. The *validation set* approach has $J = 1$. Let the validation set contain n_v cases $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_{n_v}, Y_{n_v})$, say. Then the *validation set* error rate is

$$\hat{L}_n = \frac{1}{n_v} \sum_{i=1}^{n_v} I(\hat{Y}_i \neq Y_i)$$

where $\hat{Y}_i$ is the estimate of Y_i computed from the DA method applied to the $n - n_v$ cases not in the validation set. If m_L is the number of the n_v cases from the validation set correctly classified, then $\hat{L}_n = 1 - m_L/n_v$.

The k-fold CV has $J = k$ partitions of the data into big and small sets, and the DA method is computed k times. The values $k = 5$ and 10 are common because they have been shown empirically to work well.

Definition 8.17. For *k-fold cross validation* (k-fold CV), randomly divide the training data into k groups or folds of approximately equal size $n_j \approx n/k$

for $j = 1, ..., k$. Leave out the first fold, fit the DA method to the $k-1$ remaining folds, and then find the proportion of errors for the first fold. Repeat for folds 2, ..., k. The k-fold CV error rate is

$$\hat{L}_n = \frac{1}{n} \sum_{i=1}^{n_j} \sum_{j=1}^{G} I(\hat{Y}_{ij} \neq Y_{ij})$$

where $\hat{Y}_{ij}$ is the estimate of Y_{ij} when $\boldsymbol{x}_{ij}$ is in the deleted fold. If m_k is the number of the n training cases correctly classified, then $\hat{L}_n = 1 - m_k/n$.

8.8 Some Examples

Example 8.3. This example derives two toy DA methods.

Assume the G groups come from G 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 $j = 1, ..., G$. That is, there may be G different elliptically contoured distributions with different location vectors and dispersion matrices.

Two toy 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 $j = 1, ..., G$. 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 $j = 1, ..., G$, then both Methods 1 and 2 allocate $\boldsymbol{w}$ to the group a with the smallest value of

$$\frac{D_{\boldsymbol{w}}(j)}{D_{(n_j)}(j)}. \tag{8.7}$$

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 G covering regions, distances are only computed for the groups that have covering regions that contain $\boldsymbol{w}$. Also, Equation (8.7) 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 $\hat{f}_j(\boldsymbol{w})$ using Equation (8.6) for the groups where $D_{\boldsymbol{w}}(j) \leq D_{(n_j)}(j)$ and allocate $\boldsymbol{w}$ to the group a with the largest $\hat{f}_a(\boldsymbol{w})$.

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 row of x is from 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 G 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 8.4. We generated n random $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ 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 ($D_{\boldsymbol{w}}(2) < D_{\boldsymbol{w}}(1)$) so the error rate will be near 1 for group 1 and near 0 for group 2. The *out1* output below with $p = 2$ 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 8.5. 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
```

Example 8.6. The following output illustrates crude variable selection using the *LDA* function. See Problems 8.5 and 8.6. The code deletes predictors as long as the AER does not increase if the predictor is deleted. Using all of the data, the AER = 0.0357. Eventually the AER = 0.

```
library(MASS) #Output for Example 8.6.
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)$class==group)
[1] 0.03571429 #AER using all of the predictors.
out<-lda(x[,-c(1)],group)
1-mean(predict(out,x[,-c(1)])$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 #Can delete predictors 1-5.
out<-lda(x[,-c(1,2,3,4,5,6)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,6)])$class==group)
[1] 0.07142857 #Predictor x6 is important.
out<-lda(x[,-c(1,2,3,4,5,7)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7)])$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 #Predictor x11 is important.
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 #Predictor x14 is important.
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  #Predictor x18 is important.
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
#Predictors x6, x11, x14, x18 seem good for LDA.
```

Example 8.7. This example illustrates that the AER tends to underestimate the test error rate compared to the validation set approach. The validation test error estimates can change greatly when the random number generator seed is changed. See Definitions 8.14 and 8.16. The men's basketball data set `mbb1415` is described in Problem 13.4, which tells how to get the data set into *R*. The KNN method AER is especially poor when K is small ($K < 10$, say). The KNN method also depends on a random number seed, perhaps to handle ties. (If there are three groups and $K = 3$, it is possible that the three nearest neighbors to $\boldsymbol{w}$ come from groups 1, 2, and 3. How does KNN decide which group to allocate $\boldsymbol{w}$?) The *R* commands below standardize the variables to have mean 0 and variance 1, puts guards into group 1, small forwards into group 2, centers and power forwards into group 3, and individuals with unknown position into group 0. Then individuals who do not play much (are in the bottom quartile in playing time) are deleted. Next, players in group 0 are deleted, leaving a data set `z` with 86 cases, 3 groups, and 35 predictor variables. The data set z is also divided into a validation test set ztest of 20 cases and a training set ztrain of 66 cases.

```
set.seed(1)
z <- mbb1415[,-1]
z <- scale(z) #standardize the variables
grp <- mbb1415[,1]
grp[grp==2]<-1
grp[grp==3]<-2
grp[grp==4]<-3
grp[grp==5]<-3
#Put guards in group 1, small forwards in group 2,
#centers and power forwards in group 3,
#unknowns in group 0.
#Get rid of players who did not play much.
z <- z[mbb1415[,3]>182,]
grp <- grp[mbb1415[,3]>182]
#Get rid of group 0, 86 cases left.
z <- z[grp>0,]
grp<-grp[grp>0]
indx<-sample(1:86,replace=F)
train <- indx[21:86]
test <- indx[1:20]
ztest <- z[test,]   #20 test cases
grptest <- grp[test]
ztrain <- z[train,]
grptrain <- grp[train]
```

Since x_1 is used as group, $z_i = x_{i+1}$. Below we use z_7 = turnovers, z_{10} = stl.pos (stolen possessions, a ball handling rating), z_{12} = rebounds, z_{13} = offensive rebounds, z_{28} = three point field goal percentage, and z_{32} = free throw percentage. With 2 nearest neighbors, the AER is 0.151, but (the validation error rate) VER = 0.45. With 1 nearest neighbor, the AER = 0 since each training case is its own nearest neighbor. Hence the training cases are perfectly classified.

```
#see what the variables are
z[1,c(7,10,12,13,28,32)]

library(class)
out <- knn(z[,c(7,10,12,13,28,32)],
z[,c(7,10,12,13,28,32)],grp,k=2)
mean(grp!=out)  #0.151 AER

out<-knn(ztrain[,c(7,10,12,13,28,32)],
ztest[,c(7,10,12,13,28,32)],grptrain,k=2)
```

```
mean(grptest!=out) #0.45 validation ER

out <- knn(z[,c(7,10,12,13,28,32)],
z[,c(7,10,12,13,28,32)],grp,k=1)
mean(grp!=out)  #0.0 AER

out<-knn(ztrain[,c(7,10,12,13,28,32)],
ztest[,c(7,10,12,13,28,32)],grptrain,k=1)
mean(grptest!=out) #0.45 validation ER
```

The output below shows that VER = 0.5 and AER = 0.22 with FDA (LDA), and VER = 0.45 and AER = 0.13 with QDA.

```
library(MASS) #three ways to get VER = 0.5
out <- lda(z[,c(7,10,12,13,28,32)],grp, subset=train)
1-mean(predict(out,z[-train,c(7,10,12,13,28,32)])
$class==grp[-train])
1-mean(predict(out,z[test,c(7,10,12,13,28,32)])
$class==grptest)
1-mean(predict(out,ztest[,c(7,10,12,13,28,32)])
$class==grptest)
out<-lda(z[,c(7,10,12,13,28,32)],grp)
1-mean(predict(out,z[,c(7,10,12,13,28,32)])
$class==grp) #AER =0.22

out <- qda(z[,c(7,10,12,13,28,32)],grp, subset=train)
#VER = 0.45
1-mean(predict(out,ztest[,c(7,10,12,13,28,32)])
$class==grptest)
out<-qda(z[,c(7,10,12,13,28,32)],grp)
1-mean(predict(out,z[,c(7,10,12,13,28,32)])
$class==grp) #AER =0.13
```

8.9 Robust Estimators

The literature on robust DA is fairly large. See Alrawashdeh et al. (2012), Hawkins and McLachlan (1997), Todorov and Pires (2007), and Pires and Branco (2010) for references. Several of the discussed methods could be robustified by using RMVN as the plug in estimator.

The RMVN set gives a method to objectively clean data such that the classical method applied to the cleaned data corresponds to using robust plug in estimators. Assume that there are tentative predictors $Z_1, ..., Z_J$.

After transformations assume that predictors $X_1, ..., X_k$ are linearly related. First, consider $k = p$. Let U_i be the RMVN subset applied to the n_i cases from group i for $i = 1, ..., G$. Let $(\overline{\boldsymbol{x}}_{U_i}, \boldsymbol{S}_{U_i})$ be the sample mean and covariance applied to the cases in U_i. Note that $Y = i$ for cases in U_i which are from group i. Let $U_{big} = U_1 \cup U_2 \cup \cdots \cup U_G$ be the combined sample. Then apply the DA method to U_{big} with the corresponding labels Y_{big}.

For example, RFDA consists of applying classical FDA on U_{big} resulting in finding $\hat{\boldsymbol{e}}_1$ that maximizes $\max\limits_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T \hat{\boldsymbol{B}} \boldsymbol{a}}{\boldsymbol{a}^T \hat{\boldsymbol{W}} \boldsymbol{a}} = \hat{\lambda}_1$ using $\hat{\boldsymbol{W}} = \hat{\boldsymbol{W}}_T = \sum_{i=1}^G \boldsymbol{S}_{U_i}$, and $\hat{\boldsymbol{B}} = \hat{\boldsymbol{B}}_T = \sum_{i=1}^G (\overline{\boldsymbol{x}}_{U_i} - \overline{\overline{\boldsymbol{x}}}_{big})(\overline{\boldsymbol{x}}_{U_i} - \overline{\overline{\boldsymbol{x}}}_{big})^T$ where $\overline{\overline{\boldsymbol{x}}}_{big} = \dfrac{1}{G} \sum\limits_{i=1}^G \overline{\boldsymbol{x}}_{U_i}$.

Remark 8.3. Modifications are simple if $k < p$. We can add variables like $X_{k+1} = X_1^2$, $X_{k+2} = X_3 X_4$, and $X_{k+3} = gender$. Assume the RMVN set U_i used cases $j_{1,i}, ..., j_{d_i,i}$ for $i = 1, ..., G$. Then augment U_i with the variables $X_{k+1}, ..., X_p$ corresponding to these cases. Adding variables results in cleaned data that is more likely to contain outliers.

The *mpack* function `getubig` gets U_{big}. If it can be assumed that the G groups only differ by G location vectors, then `getuc` subtracts the group coordinatewise median from each group, combines the centered data into one data set $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$, gets the case indices corresponding the centered data set, then returns the $\boldsymbol{x}_i$ corresponding to these case indices, resulting in a set U_c. Both functions return *indx*, the indices of the cases in the cleaned data set (U_{big} or U_c), and also return $Y_{big} = grp$ which has the group labels for each case in the cleaned data set.

Similarly, let B_i be the `covmb2` subset (see Section 4.7) applied to the n_i cases from group i for $i = 1, ..., G$. Let $(\overline{\boldsymbol{x}}_{B_i}, \boldsymbol{S}_{B_i})$ be the sample mean and covariance applied to the cases in B_i. Let $B_{big} = B_1 \cup B_2 \cup \cdots \cup B_G$ be the combined sample. Then apply the DA method to B_{big} with the corresponding labels Y_{big}. The function `getBbig` gets B_{big}, the group labels, and *indx*.

```
library(MASS) #need mrobdata
x<-turtle[,1:3]
group<-turtle[,4]+1
cleanb <- getubig(x,group)
cleanc <- getuc(x,group)
cleanB <- getBbig(x,group)
outb <- lda(cleanb$Ubig,cleanb$grp)
outc <- lda(cleanc$Uc,cleanc$grp)
outB <- lda(cleanB$Bbig,cleanB$grp)
#same as lda(x,group)
1-mean(predict(outb,x)$class==group) #AER 0.083
1-mean(predict(outc,x)$class==group) #0.063
1-mean(predict(outB,x)$class==group) #0.083
```

```
x[1:3,] <- x[1:3,] + 100 #Make 3 outliers.
cbo <- getubig(x,group)
cco <- getuc(x,group)
cBo <- getBbig(x,group)
outbo <- lda(cbo$Ubig,cbo$grp)
outco <- lda(cco$Uc,cco$grp)
outBo <- lda(cBo$Bbig,cBo$grp)
1-mean(predict(outbo,x)$class==group) #AER 0.083
1-mean(predict(outco,x)$class==group) #0.083
1-mean(predict(outBo,x)$class==group) #0.083
out<-lda(x,group)
1-mean(predict(out,x)$class==group)# 0.125
#classical LDA AER increases from 0.083 to 0.125
dim(x); dim(cco$Uc); dim(cbo$Ubig); dim(cBo$Bbig)
```

See Example 8.2 for how to do robust FDA using U_{big} or B_{big}. The B_{big} set can be used to make outlier resistant DA methods that work when $p > n$. The above output gives another example where B_{big} used all of the data when there were no outliers. When three outliers were added, B_{big} deleted the three outliers but used all 45 clean cases. Care needs to be taken since we want to know how well the resistant method works on the entire data set, not just on the cleaned data set. More importantly, we want to know how well the resistant method works on test data. Also theory needs to be developed for the resistant methods.

Choosing the outliers to demonstrate that the robust method is useful can be challenging. Consider $G = 2$ groups. If outliers are added to both groups in a similar manner so that both groups are fairly well separated, both classical and robust methods will likely do well. For example suppose that both groups come from an elliptically contoured distribution with the same $\boldsymbol{\Sigma}$ but different means $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. If 20% near point mass outliers are put on the major axis of each hyperellipsoidal highest density region of the clean data, LDA and QDA will likely work well. Consider contamination as in Problem 3.4: $\boldsymbol{x}_i \sim (1-\gamma)EC_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}, g_1) + \gamma EC_p(\boldsymbol{\mu}_i, c\boldsymbol{\Sigma}, g_2)$ where $i = 1, 2$, $c > 0$, and $0 < \gamma < 1$. After contamination, both groups are still elliptically contoured with means $\boldsymbol{\mu}_i$, so again LDA and QDA likely work well.

Outliers can be chosen so that the robust methods are much better than the classical methods. Let $\boldsymbol{\mu}_1 = \boldsymbol{0}$ and suppose no outliers are used for group 1. For group two, a) choose the outliers so that the sample mean $\overline{\boldsymbol{x}}_2 = \boldsymbol{0}$, or b) make the outliers a near point mass at $\boldsymbol{0}$.

Similar problems occur for one way MANOVA models and the Hotelling's T^2 test, so the outlier configurations used by Rupasinghe Arachchige Don and Olive (2017) and Rupasinghe Arachchige Don and Pelawa Watagoda (2017) may be interesting.

8.10 Summary

1) In *supervised classification*, there are G known groups or populations and m test 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 $j = 1, ..., G$. 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. The $n = \sum_{j=1}^G n_j$ cases make up the training data. Let $(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ denote the sample mean and covariance matrix for each group. Let the ith test case $\boldsymbol{w}_i$ be a new $p \times 1$ random vector from one of the G groups, but the group is unknown. *Discriminant analysis* attempts to allocate the $\boldsymbol{w}_i$ to the correct groups for $i = 1, ..., m$.

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 $j = 1, ..., G$. This rule is robust to nonnormality and the assumption of equal population dispersion matrices, but f_j is hard to estimate for $p > 2$.

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 $\rho_j(\boldsymbol{w}_i)$ that $\boldsymbol{w}_i$ is in group j are unknown or equal: $\rho_j(\boldsymbol{w}_i) = 1/G$ for $j = 1, ..., G$. Often it is assumed that the G groups have the same covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. Then the pooled covariance matrix estimator is

$$\boldsymbol{S}_{pool} = \frac{1}{n-G} \sum_{j=1}^{G} (n_j - 1) \boldsymbol{S}_j$$

where $n = \sum_{j=1}^G n_j$. Let $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$ be the estimator of multivariate location and dispersion for the jth group, e.g., 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 $j = 1, ..., G$. 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 $j = 1, ..., G$. *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 $j = 1, ..., G$. *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 $n_j \geq 10p$ for $j = 1, ..., G$.

7) Assume that $G = 2$ 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 $ESP \geq 0$ 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 $\hat{\rho}(\boldsymbol{x}) \equiv \hat{\rho}(ESP)$ 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 $n_i \geq 5p$, 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, $n_0 \geq 5p$ and $n_1 \geq 5p$, 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 $G \geq 2$ 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 $m_A = \sum_{j=1}^G m_j$ of the $n = \sum_{j=1}^G n_j$ cases were correctly classified, then the *apparent error rate* AER $= 1 - m_A/n$.

11) For the ddiscr method, get the apparent error rate for each of the G 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 test 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 use m between $0.1n$ and $0.5n$.

15) Variable selection is the search for a subset of variables that do 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.

8.11 Complements

Discriminant analysis has a massive literature. James et al. (2013) and Hastie et al. (2009) discussed many other important methods such as trees, random forests, boosting, and support vector machines. Koch (2014, pp. 120–124) showed that Fisher's discriminant analysis is a generalized eigenvalue problem. James et al. (2013) gave useful *R* code for fitting KNN. Cook and Zhang (2015) showed that envelope methods have the potential to significantly improve standard methods of linear discriminant analysis.

For $G = 2$, an alternative to the logistic regression model is the discriminant function model. See Hosmer and Lemeshow (2000, pp. 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 8.18. Under the conditions above, the **discriminant function** parameters are given by

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{8.8}$$

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 8.18 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.

Huberty and Olejnik (2006) and McLachlan (2004) are useful references for discriminant analysis. Silverman (1986, $\oint$ 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 et al. (2012). See Friedman (1989) for regularized discriminant analysis.

A DA method for two groups can be extended to G groups by performing the DA method G times where $Y_{ij} = 1$ if $\boldsymbol{x}_{ij}$ is in the jth group and $Y_{ij} = 0$ if $\boldsymbol{x}_{ij}$ is not in the jth group for $j = 1, ..., G$. Then compute $\hat{\rho}_j = \hat{P}(\boldsymbol{w}$ is in the jth) group and assign $\boldsymbol{w}$ to group a where $\hat{\rho}_a$ is a max.

There are variable selection methods for DA, and some implementations are needed in *R*, especially forward selection for when $p > n$. Witten and Tibshirani (2011) gave a LASSO type FDA method useful for $p > n$. See the *R* package *penalizedLDA*. An outlier resistant version can be made using *getBbig* to find B_{big}. See Sections 4.7 and 8.9.

Olive and Hawkins (2005) suggested that fast variable selection methods originally meant for multiple linear regression are also often effective for logistic regression when the C_p criterion is used. Also see Todorov (2007). See Olive (2010: ch. 10, 2013b, 2017a: ch. 13) for more information about variable selection and response plots for logistic regression.

Hand (2006) noted that supervised classification is a research area in statistics, machine learning, pattern recognition, computational learning theory, and data mining. Hand (2006) argued 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.

8.12 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

8.1*. Assume the cases in each of the G groups are iid from a population with covariance matrix $\boldsymbol{\Sigma_x}(j)$ Find $E(\boldsymbol{S}_{pool})$ assuming that the k groups have the same covariance matrix $\boldsymbol{\Sigma_x}(j) \equiv \boldsymbol{\Sigma_x}$ for $j = 1, ..., G$.

```
Logistic Regression Output for Problem 8.2
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 $x3 = 0$. Refer to the above output.

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) #Problem 8.4
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 above PCA and LDA output 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) #Problem 8.5
> 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, and 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 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, e.g., *ddplot*, will display the code for the function. Use the `args` command, e.g., *args(ddplot)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/mrsashw.txt) into *R*.

8.6. The Wisseman et al. (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. Nonetheless, variable selection using QDA will be done. Copy and paste the *R* commands for this part into *R*. The first nine variables result in no misclassification errors.

c) Now use commands like those shown in Example 8.6 to delete variables whose deletion does not result in a classification error. You should get four variables are needed for perfect classification. What are they (e.g., X1, X2, X3, and X4)?

8.7. Variable selection for LDA used the pottery data described in Problem 8.6, and suggested that variables X6, X11, X14, and X18 are good. Use the *R* commands for this problem to get the apparent error rate AER.

8.8. 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. See Example 8.3.

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

The Hotelling's T^2 test is used to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ when there is one sample, and $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ when there are two samples. Other applications include the multivariate matched pairs test and a test in the repeated measurements setting. These tests are robust to nonnormality.

The one-sample Hotelling's T^2 test, multivariate matched pairs test, and two-sample Hotelling's T^2 test are analogs of the univariate one-sample t test, matched pairs t test, and two-sample t test, respectively. For the multivariate Hotelling's T^2 tests, there are $p > 1$ variables and their correlations are important.

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

$$T_C^2 = n(T - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{D}}^{-1} (T - \boldsymbol{\mu}_0) > \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$$

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_9

where $\hat{\boldsymbol{D}}$ 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) $\to \alpha$ as $n \to \infty$. We want $n \geq 10p$ 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_p^2 approximation to T_H^2 is poor.

Let pval be an estimate of the pvalue. We typically use $T_C^2 = T_H^2$ in the following four-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 $T_C^2 = n(T - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{D}}^{-1} (T - \boldsymbol{\mu}_0)$.
iii) Find pval =

$$P\left(T_C^2 < \frac{(n-1)p}{n-p} F_{p,n-p}\right) = P\left(\frac{n-p}{(n-1)p} \ \ T_C^2 < F_{p,n-p}\right).$$

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 $\leq \alpha$ and fail to reject H_0 if pval $> \alpha$. 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 $P[\boldsymbol{\mu} \in R(\boldsymbol{W})] \to 1 - \alpha$ as $n \to \infty$. 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

$$R(\boldsymbol{W}) = \{\boldsymbol{w} | n(\overline{\boldsymbol{x}} - \boldsymbol{w})^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{w}) \leq \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}\}$$

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 $n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) = D^2_{\boldsymbol{\mu}}(\overline{\boldsymbol{x}}, \boldsymbol{S}/n)$. If $\boldsymbol{\mu}$ is close to $\overline{\boldsymbol{x}}$ with respect to the Mahalanobis distance based on dispersion matrix $\boldsymbol{S}/n$, then $\boldsymbol{\mu}$ will be 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 $\boldsymbol{a} \neq \boldsymbol{0}$, $P(L_{\boldsymbol{a}} \leq \boldsymbol{a}^T\boldsymbol{\mu} \leq U_{\boldsymbol{a}}) \to 1-\alpha$ as $n \to \infty$ 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, pp. 184–5), the p confidence intervals (CIs) for μ_i and the $p(p-1)/2$ CIs for $\mu_i - \mu_k$ can be made such that for each of the two types of CI, they all hold simultaneously with confidence $\to 1-\alpha$. Hence if $\alpha = 0.05$, then in 100 samples, we expect all p CIs to contain μ_i about 95 times, and we expect all $p(p-1)/2$ CIs to contain $\mu_i - \mu_k$ about 95 times. For each of the two types of CI, about 5 times at least one of the CIs will fail to contain its parameter (μ_i or $\mu_i - \mu_k$). The simultaneous 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 simultaneous CIs for $\mu_i - \mu_k$ are

$$[L, U] = \overline{x}_i - \overline{x}_k \pm \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p,1-\alpha}} \sqrt{\frac{S_{ii} - 2S_{ik} + S_{kk}}{n}}.$$

Example 9.1. Following Mardia et al. (1979, p. 126), data for first and second adult sons had $n = 25$ and variables X_1 = head length of first son and X_2 = head length of second son. Suppose $\boldsymbol{\mu}_0 = (182, 182)^T$ and $T_C^2 = 1.28$. Perform the one-sample Hotelling's T^2 test.

Solution: i) $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$

ii) $T_C^2 = 1.28$

iii) $\dfrac{n-p}{(n-1)p} T_C^2 = \dfrac{25-2}{(24(2)} 1.28 = 0.6133$, and pval $= P(0.613 < F_{2,23}) >$ 0.05

iv) Fail to reject H_0, so $\boldsymbol{\mu} = (182, 182)^T$.

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, pp. 1651–1652) showed that if (E1) holds, then the classical estimator applied to cases with $D_i(\overline{\boldsymbol{x}}, S) \leq h$ is asymptotically normal with

$$\sqrt{n}(T_{0,D} - \boldsymbol{\mu}) \xrightarrow{D} N_p(\mathbf{0}, \kappa_p \boldsymbol{\Sigma}).$$

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:

$$\sqrt{n}(T_{RMVN} - \boldsymbol{\mu}) \xrightarrow{D} N_p(\mathbf{0}, \tau_p \boldsymbol{\Sigma})$$

for a wide variety of elliptically contoured distributions where τ_p depends on both p and the underlying distribution. 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

$$\frac{T_R^2}{f_{n,p}} = n(T_{RMVN} - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{C}}^{-1}_{RMVN}(T_{RMVN} - \boldsymbol{\mu}_0)/f_{n,p} > \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$$

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 $2 \leq p \leq 100$.

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

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.

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

```
n<-4000; p <- 30 #May take a few minutes.
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)
```

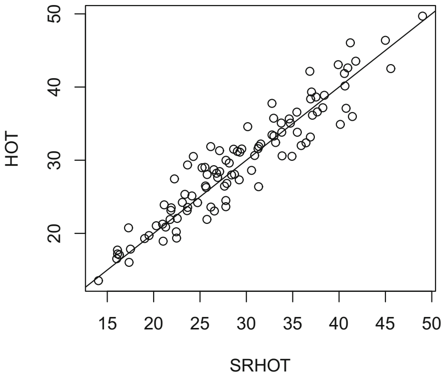

Fig. 9.1 Scaled "Robust" Statistic Versus T^2_H Statistic

For the Hotelling's T^2_H 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^2_H 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 (the proportion 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^2_H has more power than the robust test, but suggests that the power of both tests rapidly increases to one as δ increases.

9.1.2 Bootstrapping Hotelling's T^2 Type Tests

The prediction region method of Section 5.3 is useful for bootstrapping the test $H_0 : \boldsymbol{\mu}_T = \boldsymbol{\mu}_0$ versus $H_A : \boldsymbol{\mu}_T \neq \boldsymbol{\mu}_0$ where the test statistic T estimates the parameter $\boldsymbol{\mu}_T$. Take a sample of size n with replacement from the cases $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ to make the bootstrap statistic T^*_1. Repeat to get the bootstrap sample $T^*_1, ..., T^*_B$. Apply the nonparametric prediction region to the bootstrap sample and see if $\boldsymbol{\mu}_0$ is in the region. Equivalently, apply the nonparametric prediction region to $\boldsymbol{w}_i = T^*_i - \boldsymbol{\mu}_0$, $i = 1, ..., B$, and fail to reject H_0 if $\mathbf{0}$ is in the region, otherwise reject H_0.

The *mpack* function `rhotboot` bootstraps T where T is the coordinatewise median or T is the RMVN location estimator. The function `medhotsim` simulates the test with $\boldsymbol{\mu}_0 = \mathbf{0}$ when T is the coordinatewise median. The simulated data are as in Section 6.3, with $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z}$, except that $\boldsymbol{z} = \boldsymbol{u} - \mathbf{1}$ was used for the multivariate lognormal distribution with $u_i = \exp(w_i)$ and

$w_i \sim N(0,1)$, so that the population coordinatewise median of $\boldsymbol{x}$ and $\boldsymbol{z}$ was $\boldsymbol{0}$ when H_0 is true. When H_0 was false, $\boldsymbol{\mu}_0 = \delta \boldsymbol{1}$ with $\delta > 0$.

The term *hotcov* was the proportion of times the bootstrap test rejected H_0 with a nominal level of 0.05. With $n = 100$ and $p = 2$, *hotcov* was near 0.05 when H_0 was true. The test usually had good power if $\boldsymbol{\mu} = (0.5, 0.5)^T$. See output below where 1000 runs were used.

```
medhotsim(xtype=1,nruns=1000)
0.046              #MVN((0,0)^T, diag(1,2)) data
medhotsim(xtype=1,nruns=1000,delta=0.5)
0.995              #MVN((0.5,0.5)^T, diag(1,2)) data
```

9.2 Matched Pairs

Assume that there are $k = 2$ treatments, and both treatments are given to the same n cases or units. Then p measurements are taken for both treatments. For example, systolic and diastolic blood pressure could be compared before and after the patient (case) receives blood pressure medication. Then $p = 2$. Alternatively use n 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 $\boldsymbol{y}_i = (Y_{i1}, Y_{i2}, ..., Y_{ip})^T$ denotes the p measurements from the 1st treatment, and $\boldsymbol{z}_i = (Z_{i1}, Z_{i2}, ..., Z_{ip})^T$ denotes 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 $\boldsymbol{\mu}_0 = \boldsymbol{0}$. If the p random variables are continuous, make three 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 four steps.
i) State the hypotheses $H_0 : \boldsymbol{\mu} = \boldsymbol{0} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{0}$.
ii) Find the test statistic $T^2_M = n\overline{\boldsymbol{x}}^T \boldsymbol{S}^{-1} \overline{\boldsymbol{x}}$.
iii) Find pval =

$$P\left(T^2_M < \frac{(n-1)p}{n-p} F_{p,n-p}\right) = P\left(\frac{n-p}{(n-1)p}\, T^2_M < F_{p,n-p}\right).$$

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 $\leq \alpha$ and fail to reject H_0 if pval $> \alpha$. 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

$$\{\boldsymbol{w} |\ n(\overline{\boldsymbol{x}} - \boldsymbol{w})^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{w}) \leq \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}\},$$

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 $\boldsymbol{S}$.

Example 9.2. Following Johnson and Wichern (1988, pp. 213–214), wastewater from a sewage treatment plant was sent to two labs for measurements of biochemical demand (BOD) and suspended solids (SS). Suppose $n = 11$, $p = 2$, and $T_M^2 = 13.6$. Perform the appropriate test.

Solution: i) $H_0 : \boldsymbol{\mu} = \boldsymbol{0} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{0}$

ii) $T_M^2 = 13.6$

iii) $\dfrac{n-p}{(n-1)p} T_M^2 = \dfrac{11-2}{(11-1)2} 13.6 = 6.12$, and pval $= P(6.12 < F_{2,9}) < 0.05$

iv) Reject H_0. Hence $\boldsymbol{\mu} \neq (0,0)^T$, and the two labs are giving different mean measurements for $(\mu_{BOD}, \mu_{SS})^T$.

To get a bootstrap analog of this test, bootstrap the $\boldsymbol{d}_i = \boldsymbol{x}_i$ as in Section 9.1.2 where usually $H_0 : \boldsymbol{\mu} \equiv \boldsymbol{\mu}_T = \boldsymbol{0}$. Again robust location estimators, such as the coordinatewise median or RMVN location estimator T_{RMVN}, could be used on the $\boldsymbol{x}_i$.

9.3 Repeated Measurements

Repeated measurements = longitudinal data analysis. Take p measurements on the same unit, often the same measurement, e.g., blood pressure, at several time periods. Hence each unit or individual is measured repeatedly over time. The variables are X_1, ..., X_p where often X_k is the measurement at the kth time period. Then $E(\boldsymbol{x}) = (\mu_1, ..., \mu_p)^T = (\mu + \tau_1, ..., \mu + \tau_p)^T$. Let the $(p-1) \times 1$ vector $\boldsymbol{y}_j = (x_{1j} - x_{2j}, x_{2j} - x_{3j}, ..., x_{p-1,j} - x_{pj})^T$ for $j = 1, ..., n$. Hence $y_{ij} = x_{ij} - x_{i+1,j}$ for $j = 1, ..., n$ and $i = 1, ..., p-1$. 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 $\boldsymbol{\mu}_{\boldsymbol{y}} = E(\boldsymbol{y}_i)$, then $\boldsymbol{\mu}_{\boldsymbol{y}} = \boldsymbol{0}$ is equivalent to $\mu_1 = \cdots = \mu_p$ where $E(X_k) = \mu_k$. Let $\boldsymbol{S}_{\boldsymbol{y}}$ be the sample covariance matrix of the $\boldsymbol{y}_i$.

The **large sample repeated measurements test** has four steps.
i) State the hypotheses $H_0 : \boldsymbol{\mu}_{\boldsymbol{y}} = \boldsymbol{0} \quad H_1 : \boldsymbol{\mu}_{\boldsymbol{y}} \neq \boldsymbol{0}$.
ii) Find the test statistic $T_R^2 = n\overline{\boldsymbol{y}}^T \boldsymbol{S}_{\boldsymbol{y}}^{-1} \overline{\boldsymbol{y}}$.
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 $\boldsymbol{\mu}_{\boldsymbol{y}} \neq \boldsymbol{0}$ so not all p of the μ_i are equal, while if you fail to reject H_0 conclude that the population mean $\boldsymbol{\mu}_{\boldsymbol{y}} = \boldsymbol{0}$ or that there is not enough evidence to conclude that $\boldsymbol{\mu}_{\boldsymbol{y}} \neq \boldsymbol{0}$. Reject H_0 if pval $\leq \alpha$ and fail to reject H_0 if pval $> \alpha$. Give a nontechnical sentence, if possible.

Example 9.3. Following Morrison (1967, pp. 139–141), reaction times to visual stimuli were obtained for $n = 20$ normal young men under conditions A, B, and C of stimulus display. Let $\overline{x}_A = 21.05, \overline{x}_B = 21.65$, and $\overline{x}_C = 28.95$. Test whether $\mu_A = \mu_B = \mu_C$ if $T_R^2 = 882.8$.

Solution: i) $H_0 : \boldsymbol{\mu}_{\boldsymbol{y}} = \boldsymbol{0} \quad H_1 : \boldsymbol{\mu}_{\boldsymbol{y}} \neq \boldsymbol{0}$
ii) $T_R^2 = 882.8$
iii) $\dfrac{n-p+1}{(n-1)(p-1)} T_R^2 = \dfrac{20-3+1}{(20-1)(3-1)} 882.8 = 418.168$, and
pval $= P(418.168 < F_{2,18}) \approx 0$
iv) Reject H_0. The three mean reaction times are different.

An alternative test would use a statistic T, such as the coordinatewise median or RMVN location estimator, on the $\boldsymbol{y}_j$, and the bootstrap method of Section 9.1.2 can be applied with $\boldsymbol{\mu}_{\boldsymbol{y}} = \boldsymbol{0}$. This test is equivalent to $H_0 : \mu_1 = \cdots = \mu_p$ where μ_k is a population location parameter for the kth measurement. Hence if the coordiatewise median is being used, then μ_k is the population median of the kth measurement.

9.4 Two Samples

Suppose there are two independent random samples $\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{n_1,1}$ and $\boldsymbol{x}_{1,2}, ..., \boldsymbol{x}_{n_2,2}$ 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 $n_1 = k n_2$ for some positive real number k.

By the multivariate central limit theorem,

$$\begin{pmatrix} \sqrt{n_1}\ (\overline{\boldsymbol{x}}_1 - \boldsymbol{\mu}_1) \\ \sqrt{n_2}\ (\overline{\boldsymbol{x}}_2 - \boldsymbol{\mu}_2) \end{pmatrix} \xrightarrow{D} N_{2p}\left[\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x}_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{\boldsymbol{x}_2} \end{pmatrix}\right],$$

or

$$\begin{pmatrix} \sqrt{n_2}\ (\overline{\boldsymbol{x}}_1 - \boldsymbol{\mu}_1) \\ \sqrt{n_2}\ (\overline{\boldsymbol{x}}_2 - \boldsymbol{\mu}_2) \end{pmatrix} \xrightarrow{D} N_{2p}\left[\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}_1}}{k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{\boldsymbol{x}_2} \end{pmatrix}\right].$$

Hence

$$\sqrt{n_2}\ [(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \xrightarrow{D} N_p(\boldsymbol{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}_1}}{k} + \boldsymbol{\Sigma}_{\boldsymbol{x}_2}).$$

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

$$T_0^2 = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \left(\frac{\boldsymbol{S}_1}{n_1} + \frac{\boldsymbol{S}_2}{n_2}\right)^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) \xrightarrow{D} \chi^2_p.$$

The above result is easily generalized to other statistics. See Rupasinghe Arachchige Don and Pelawa Watagoda (2017). If the sequence of positive integers $d_n \to \infty$ 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 $\dfrac{pF_{p,d_n,1-\alpha}}{\chi^2_{p,1-\alpha}}$ 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 $P(Y_n \le \chi^2_{p,\alpha}) = \alpha$ if Y_n has a χ^2_p distribution, and $P(Y_n \le F_{p,d_n,\alpha}) = \alpha$ if Y_n has an F_{p,d_n} distribution.

Let pval denote the estimated pvalue. The four-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 $\leq \alpha$ and fail to reject H_0 if pval $> \alpha$. Give a nontechnical sentence if possible. As a benchmark for this text, use $\alpha = 0.05$ if α is not given.

Example 9.4. Following Mardia et al. (1979, p. 153), cranial length and breadth (X_1 and X_2) were measured on $n_1 = 35$ female frogs and $n_2 = 14$ male frogs with $\overline{\boldsymbol{x}}_1 = (22.86, 24.397)^T$ and $\overline{\boldsymbol{x}}_2 = (21.821, 22.442)^T$. Test $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ if $T_0^2 = 2.550$.

Solution: i) $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$
ii) $t_0 = T_0^2/p = 2.550/2 = 1.275$
iii) pval $= P(1.275 < F_{2,14-2}) > 0.05$
iv) Fail to reject H_0. There is not enough evidence to conclude that the mean lengths and breadths differ for the male and female frogs.

The plots for the one way MANOVA model in Section 10.2 are also useful for the two-sample Hotelling's T^2 test. An alternative to the above test is to used the pooled covariance matrix. This Hotelling's T^2 test is a special case of the one way MANOVA model with two groups covered in Section 10.3.

9.4.1 Bootstrapping Two-Sample Tests

Bootstrapping the two-sample test is similar to bootstrapping discriminant analysis and one way MANOVA models. Take a sample of size n_i with replacement from random sample i for $i = 1, 2$, and compute $T_{11}^* - T_{21}^*$. Repeat B times to get the bootstrap sample $\boldsymbol{w}_1 = T_{11}^* - T_{21}^*, ..., \boldsymbol{w}_B = T_{1B}^* - T_{2B}^*$. Apply the nonparametric prediction region on the $\boldsymbol{w}_i$, and fail to reject $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ if $\boldsymbol{0}$ is in the prediction region, and reject H_0, otherwise. See Rupasinghe Arachchige Don and Pelawa Watagoda (2017).

Some *R* output is below for the Gladstone (1905) data where several infants are outliers. We first tested the first 133 cases versus the last 134 cases. It turned out that the first group was younger and had all of the infants, so H_0 was rejected. Then a random sample of 133 was used as the first group and the remaining 134 as the second group. Then the test failed to reject H_0. Using the nominal level $\alpha = 0.05$ of the large sample bootstrap test, reject H_0 if the test statistic is larger than the cutoff, where 4.102 was the cutoff for the first test which used RMVN.

```
zz <- cbrainx[,c(1,3,5,6,7,8,9,11)]
#get rid of qualitative variables
zx <- zz[1:133,]
zy <- zz[134:267,]
out<-rhot2boot(zx,zy,med=F) #RMVN takes a while.
tem<-predreg(out$mus)
> tem$cuplim
   95.4%
4.101788
> tem$D0
[1] 7.529998 #> 4.102 so reject Ho
out<-rhot2boot(zx,zy,med=T) #coord. median is fast
tem<-predreg(out$mus)
> tem$cuplim
   95.4%
4.046958
> tem$D0
[1] 12.87506 #> 4.05 so reject Ho
plot(zx[,1],zy[-134,1])
#zx people tend to be older, infants are in zy
indx <- sample(1:267,133)#random sample for zx and zy
zx <- zz[indx,]
zy <- zz[-indx,]
out<-rhot2boot(zx,zy,med=F)
tem<-predreg(out$mus) #RMVN
> tem$cuplim
   95.4%
4.065357
> tem$D0
[1] 2.94968 #< 4.07 so fail to reject Ho
out<-rhot2boot(zx,zy,med=T)
tem<-predreg(out$mus) #coord. median
> tem$cuplim
   95.4%
3.915687
> tem$D0
[1] 2.802046 #< 3.92 so fail to reject Ho
```

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 $T_H^2 = n(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0) > \dfrac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$ 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 $T_C^2 = n(T - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{D}}^{-1}(T - \boldsymbol{\mu}_0) > \dfrac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$ where $\hat{\boldsymbol{D}}$ 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 $\dfrac{(n-1)p}{n-p} F_{p,n-p,1-\alpha} \rightarrow \chi^2_{p,1-\alpha}$ as $n \rightarrow \infty$.

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 four-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 $\leq \alpha$ and fail to reject H_0 if pval $> \alpha$.

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 $\boldsymbol{\mu} = E(\boldsymbol{x}) = E(\boldsymbol{y}) - E(\boldsymbol{z})$. We want to test if $\boldsymbol{\mu} = \boldsymbol{0}$, so $E(\boldsymbol{y}) = E(\boldsymbol{z})$. The test can also be used if $(\boldsymbol{y}_i, \boldsymbol{z}_i)$ are matched (highly dependent) in some way. For example, if identical twins are in the study, $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$ could be the measurements on each twin. Let $(\overline{\boldsymbol{x}}, \boldsymbol{S}_{\boldsymbol{x}})$ be the sample mean and covariance matrix of the $\boldsymbol{x}_i$.

5) The **large sample multivariate matched pairs test** has four steps.
i) State the hypotheses $H_0 : \boldsymbol{\mu} = \boldsymbol{0} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{0}$.
ii) Find the test statistic $T_M^2 = n\overline{\boldsymbol{x}}^T \boldsymbol{S}_{\boldsymbol{x}}^{-1} \overline{\boldsymbol{x}}$.
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 $\leq \alpha$ and fail to reject H_0 if pval $> \alpha$. Give a nontechnical sentence if possible.

6) Repeated measurements = longitudinal data analysis. Take p measurements on the same unit, often the same measurement, e.g., 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. Then $E(\boldsymbol{x}) = (\mu_1, ..., \mu_p)^T = (\mu + \tau_1, ..., \mu + \tau_p)^T$. Let $\boldsymbol{y}_j = (x_{1j} - x_{2j}, x_{2j} - x_{3j}, ..., x_{p-1,j} - x_{pj})^T$ for $j = 1, ..., n$. 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 $\boldsymbol{\mu}_{\boldsymbol{y}} = E(\boldsymbol{y}_i)$, then $\boldsymbol{\mu}_Y = \boldsymbol{0}$ is equivalent to $\mu_1 = \cdots = \mu_p$ where $E(X_k) = \mu_k$. Let $\boldsymbol{S}_{\boldsymbol{y}}$ be the sample covariance matrix of the $\boldsymbol{y}_i$.

7) The **large sample repeated measurements test** has four 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 $\leq \alpha$ and fail to reject H_0 if pval $> \alpha$. Give a nontechnical sentence, if possible.

8) The F tables give left tail area and the pval is a right tail area. The Section 15.5 table 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 $\dfrac{n-p}{(n-1)p} \; T_C^2 \geq F_{k,d,0.95}$, 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 $n \geq 10p$, the tests in 3), 5), and 7) 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 $\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{n_1,1}$ and $\boldsymbol{x}_{1,2}, ..., \boldsymbol{x}_{n_2,2}$ 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 four-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 $\leq \alpha$ and fail to reject H_0 if pval $> \alpha$. Give a nontechnical sentence if possible.

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 et al. (2002) used similar reasoning to present a diagnostic based on the FMCD estimator.

Yao (1965) suggested 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, pp. 55–57), which provides randomization tests as competitors for the two-sample Hotelling's T^2 test. Bootstrapping the tests with robust estimators seems to be effective. For bootstrapping the two-sample Hotelling's T^2 test, see Rupasinghe Arachchige Don and Pelawa Watagoda (2017). Gregory et al. (2015) and Feng and Sun (2015) considered the two-sample test when $p \geq n$.

9.7 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

9.1. Following Morrison (1967, pp. 122–123), the Wechsler Adult Intelligence Scale scores of $n = 101$ subjects aged 60 to 64 were recorded, giving a verbal score (X_1) and performance score (X_2) for each subject. Suppose $\boldsymbol{\mu}_0 = (60, 50)^T$ and $T^2_C = 357.43$. Perform the one-sample Hotelling's T^2 test.

9.2. Following Morrison (1967, pp. 137–138), the levels of free fatty acid (FFA) in the blood were measured in $n = 15$ hypnotized normal volunteers who had been asked to experience fear, depression, and anger effects while in the hypnotic state. The mean FFA changes were $\overline{x}_1 = 2.669$, $\overline{x}_2 = 2.178$, and $\overline{x}_3 = 2.558$. Let $\mu_F = \mu + \tau_1$, $\mu_D = \mu + \tau_2$, and $\mu_A = \mu + \tau_3$. We want to know if the mean stress FFA changes were equal. So test whether $\mu_F = \mu_D = \mu_F$ if $T^2_R = 2.68$.

9.3. Data is taken or modified from Johnson and Wichern (1988, pp. 185, 224).

a) Suppose $S^2_2 = S_{22} = 126.05$, $\overline{x}_2 = 54.69$, $n = 87$, and $p = 3$. Find a large sample simultaneous 95% CI for μ_2.

b) Suppose a random sample of 50 bars of soap from method 1 and a random sample of 50 bars of soap from method 2 are obtained. Let $X_1 =$ lather and $X_2 =$ mildness with $\overline{\boldsymbol{x}}_1 = (8.4, 4.1)^T$ and $\overline{\boldsymbol{x}}_2 = (10.2, 3.9)^T$. Test $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ if $T_0^2 = 52.4722$.

R 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, e.g., *rhotsim*, will display the code for the function. Use the `args` command, e.g., *args(rhotsim)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/mrsashw.txt) into *R*.

9.4*. Use the *R* commands in Subsection 9.1.1 to make a plot similar to Figure 9.1. The program may take a minute to run.

9.5. Conjecture:

$$\sqrt{n}(T_{RMVN} - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \tau_p \boldsymbol{\Sigma})$$

for a wide variety of elliptically contoured distributions where τ_p depends on both p and the underlying distribution. 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

$$\frac{T_R^2}{f_{n,p}} = n(T_{RMVN} - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{C}}_{RMVN}^{-1}(T_{RMVN} - \boldsymbol{\mu}_0)/f_{n,p} > \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$$

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

$$\boldsymbol{x}_i \sim N_p(\boldsymbol{0}, diag(1, 2, ..., p))$$

where $H_0 : \boldsymbol{\mu} = \boldsymbol{0}$ 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. We want hcv and rhcv near 0.05. THIS SIMULATION MAY TAKE A FEW MINUTES. Record hcv and rhcv. Were hcv and rhcv near 0.05?

b) The *R* commands for this part use simulated data

$$\boldsymbol{x}_i \sim N_p(\delta \boldsymbol{1}, diag(1, 2, ..., p))$$

where $H_0 : \boldsymbol{\mu} = \boldsymbol{0}$ 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. We want hcv and rhcv near 1 so that the power is high. Paste the output into *Word*. THIS SIMULATION MAY TAKE A FEW MINUTES. Record hcv and rhcv. Were hcv and rhcv near 1?

Chapter 10
MANOVA

This chapter gives the multivariate linear model which includes the following two models. i) The multivariate linear regression model of Chapter 12 has at least one quantitative predictor variable. ii) For the MANOVA model, the predictors are indicator variables. Often observations $(Y_1, ..., Y_m, x_1, x_2, ..., x_p)$ are collected on the same person or thing and hence are correlated. If transformations can be found such that the m response plots and residual plots of Section 10.2 look good, and $n \geq (m+p)^2$ (and $n_i \geq 10m$ if there are p treatment groups and $n = \sum_{i=1}^m n_i$), then the MANOVA model can often be used to efficiently analyze the data. These two plots and the DD plot of the residuals are useful for checking the model and for outlier detection.

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. A multivariate linear model has $m \geq 2$ response variables. A *multiple linear model* = *univariate linear model* has $m = 1$ response variable, but at least two nontrivial predictors, and usually a constant (so $p \geq 3$). A simple linear model has $m = 1$, one nontrivial predictor, and usually a constant (so $p = 2$). Multiple linear regression models and ANOVA models are special cases of multiple linear models.

Definition 10.2. The **multivariate linear model**

$$\boldsymbol{y}_i = \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i$$

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_10

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 $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$, where the matrices are defined below. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\sigma_{ij})$ for $k = 1, ..., n$. Then the $p \times m$ coefficient matrix $\boldsymbol{B} = \left[\boldsymbol{\beta}_1 \; \boldsymbol{\beta}_2 \ldots \boldsymbol{\beta}_m\right]$ and the $m \times m$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j$. The $\boldsymbol{\epsilon}_i$ are assumed to be iid. The univariate linear model corresponds to $m = 1$ response variable and is written in matrix form as $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$. Subscripts are needed for the m univariate linear models $\boldsymbol{Y}_j = \boldsymbol{X\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$, where $E(\boldsymbol{e}_j) = \boldsymbol{0}$. For the multivariate linear model, $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij} \;\; \boldsymbol{I}_n$ for $i, j = 1, ..., m$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

Definition 10.3. The multivariate analysis of variance (**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 p predictor variables $X_1, X_2, ..., X_p$. The MANOVA model is a special case of the multivariate linear model. For the MANOVA model, the predictors are not quantitative variables, so the predictors are indicator variables. Sometimes, the trivial predictor $\boldsymbol{1}$ is also in the model. In matrix form, the MANOVA model is $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$. 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 $E(Y_{ij}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j$.

The data matrix $\boldsymbol{W} = [\boldsymbol{X} \;\; \boldsymbol{Z}]$. If the model contains a constant, then usually the first column of ones $\boldsymbol{1}$ of $\boldsymbol{X}$ is omitted from the data matrix for software such as *R* and *SAS*. The $n \times m$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & \ldots & Y_{1,m} \\ Y_{2,1} & Y_{2,2} & \ldots & Y_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n,1} & Y_{n,2} & \ldots & Y_{n,m} \end{bmatrix} = \left[\boldsymbol{Y}_1 \; \boldsymbol{Y}_2 \ldots \boldsymbol{Y}_m\right] = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix}.$$

The $n \times p$ design matrix $\boldsymbol{X}$ of predictor variables is not necessarily of full rank p, and

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} = \left[\boldsymbol{v}_1 \; \boldsymbol{v}_2 \ldots \boldsymbol{v}_p\right] = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix},$$

where often $\boldsymbol{v}_1 = \boldsymbol{1}$.

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

$$\boldsymbol{E} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \dots & \epsilon_{1,m} \\ \epsilon_{2,1} & \epsilon_{2,2} & \dots & \epsilon_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \dots & \epsilon_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \dots & \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix}.$$

Considering the ith row of $\boldsymbol{Z}, \boldsymbol{X}$, and $\boldsymbol{E}$ shows that $\boldsymbol{y}_i^T = \boldsymbol{x}_i^T \boldsymbol{B} + \boldsymbol{\epsilon}_i^T$.

Warning: The $\boldsymbol{e}_i$ are error vectors, not orthonormal eigenvectors.

Definition 10.4. 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,

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i \tag{10.1}$$

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 $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors, and $n \geq p$. Equivalently,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{10.3}$$

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, e.g., 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 depends 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

$$\begin{bmatrix} Y_{i1} = \beta_{11}x_{i1} + \cdots + \beta_{p1}x_{ip} + \epsilon_{i1} = \boldsymbol{x}_i^T\boldsymbol{\beta}_1 + \epsilon_{i1} \\ Y_{i2} = \beta_{12}x_{i1} + \cdots + \beta_{p2}x_{ip} + \epsilon_{i2} = \boldsymbol{x}_i^T\boldsymbol{\beta}_2 + \epsilon_{i2} \\ \vdots \\ Y_{im} = \beta_{1m}x_{i1} + \cdots + \beta_{pm}x_{ip} + \epsilon_{im} = \boldsymbol{x}_i^T\boldsymbol{\beta}_m + \epsilon_{im} \end{bmatrix}$$

or $\boldsymbol{y}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i$, 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 $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ 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.

Let $\hat{\boldsymbol{B}}$ be the MANOVA estimator of $\boldsymbol{B}$. MANOVA models are often fit by least squares. Then the **least squares estimators** are

$$\hat{\boldsymbol{B}} = \hat{\boldsymbol{B}}_g = (\boldsymbol{X}^T\boldsymbol{X})^{-}\boldsymbol{X}^T\boldsymbol{Z} = \left[\hat{\boldsymbol{\beta}}_1 \; \hat{\boldsymbol{\beta}}_2 \; \ldots \; \hat{\boldsymbol{\beta}}_m\right]$$

where $(\boldsymbol{X}^T\boldsymbol{X})^{-}$ is a generalized inverse of $\boldsymbol{X}^T\boldsymbol{X}$. Here $\hat{\boldsymbol{B}}_g$ depends on the generalized inverse. If $\boldsymbol{X}$ has full rank p, then $(\boldsymbol{X}^T\boldsymbol{X})^{-} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\hat{\boldsymbol{B}}$ is unique.

Definition 10.5. The *predicted values* or *fitted values*

$$\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \left[\hat{\boldsymbol{Y}}_1 \; \hat{\boldsymbol{Y}}_2 \; \dots \; \hat{\boldsymbol{Y}}_m\right] = \begin{bmatrix} \hat{Y}_{1,1} & \hat{Y}_{1,2} & \dots & \hat{Y}_{1,m} \\ \hat{Y}_{2,1} & \hat{Y}_{2,2} & \dots & \hat{Y}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{Y}_{n,1} & \hat{Y}_{n,2} & \dots & \hat{Y}_{n,m} \end{bmatrix}.$$

The *residuals* $\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} =$

$$\begin{bmatrix} \hat{\boldsymbol{\epsilon}}_1^T \\ \hat{\boldsymbol{\epsilon}}_2^T \\ \vdots \\ \hat{\boldsymbol{\epsilon}}_n^T \end{bmatrix} = \left[\hat{\boldsymbol{r}}_1 \; \hat{\boldsymbol{r}}_2 \; \dots \; \hat{\boldsymbol{r}}_m\right] = \begin{bmatrix} \hat{\epsilon}_{1,1} & \hat{\epsilon}_{1,2} & \dots & \hat{\epsilon}_{1,m} \\ \hat{\epsilon}_{2,1} & \hat{\epsilon}_{2,2} & \dots & \hat{\epsilon}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\epsilon}_{n,1} & \hat{\epsilon}_{n,2} & \dots & \hat{\epsilon}_{n,m} \end{bmatrix}.$$

These quantities can be found by fitting m ANOVA models $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ to get $\widehat{\boldsymbol{\beta}}_j$, $\widehat{\boldsymbol{Y}}_j = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j$, 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. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ be the usual estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ for the MANOVA model. If least squares is used with a full rank $\boldsymbol{X}$, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=p}$.

10.2 Plots for MANOVA Models

As in Chapter 12, this section suggests using residual plots, response plots, and the DD plot to examine the multivariate linear model. The residual plots are often used to check for lack of fit of the multivariate linear model. The response plots are used to check linearity (and to detect influential cases and outliers for linearity). The response and residual plots are used exactly as in the $m = 1$ case corresponding to multiple linear regression and experimental design models. See Olive (2010, 2017a), Olive and Hawkins (2005), and Cook and Weisberg (1999a, p. 432; 1999b).

Definition 10.6. 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 10.1. Make the m response and residual plots for any MANOVA model. In a response plot, the vertical deviations from the identity line are the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij}$. Suppose the model is good, the error distribution is not highly skewed, and $n \geq 10p$. 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 transformed 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.

For some MANOVA models that do not use replication, the response and residual plots look much like those for multivariate linear regression in Section 12.2. The response and residual plots for the one-way MANOVA model need some notation, and it is useful to use three subscripts. Suppose there are independent random samples of size n_i from p different populations (treatments), or n_i cases are randomly assigned to p treatment groups with $n = \sum_{i=1}^p n_i$. 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 $E(\boldsymbol{y}_{ij}) = \boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{im})^T$ and $\text{Cov}(\boldsymbol{y}_{ij}) = \boldsymbol{\Sigma_\epsilon}$. Hence the p treatments have possibly different mean vectors $\boldsymbol{\mu}_i$, but common covariance matrix $\boldsymbol{\Sigma_\epsilon}$.

Then for the kth response variable, the *response plot* is a plot of $\hat{Y}_{ijk} \equiv \hat{\mu}_{ik}$ versus Y_{ijk} and the *residual plot* is a plot of $\hat{Y}_{ijk} \equiv \hat{\mu}_{ik}$ versus r_{ijk}, where $\hat{\mu}_{ik}$ is the sample mean of the n_i responses Y_{ijk} corresponding to the ith treatment for the kth response variable. 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 $Z_1, ..., Z_n$ consists of an axis and n points each corresponding to the value of Z_i. The response plot for the kth response variable consists of p dot plots, one for each value of $\hat{\mu}_{ik}$. The dot plot corresponding to $\hat{\mu}_{ik}$ is the dot plot of $Y_{i,1,k}, ..., Y_{i,n_i,k}$. Similarly, the residual plot for the kth response variable consists of p dot plots, and the plot corresponding to $\hat{\mu}_{ik}$ is the dot plot of $r_{i,1,k}, ..., r_{i,n_i,k}$. Assuming the $n_i \geq 10$, the p dot plots for the kth response variable should have roughly the same shape and spread in both the response and residual plots. Note that $\hat{\mu}_{ik} = \overline{Y}_{iok} = \dfrac{1}{n_i} \displaystyle\sum_{j=1}^{n_i} Y_{ijk}$.

Assume that each $n_i \geq 10$. It is easier to check shape and spread in 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 may not be much difference in the μ_{ik}. In the response plot, if the identity line fits the plotted points better than any horizontal line, then conclude that at least some of the means μ_{ik} differ.

Definition 10.7. 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 two lines parallel to the identity line and two 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 two 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_i cases corresponding to treatment i with mean μ_{ik} that is far from the bulk of the data. This dot plot is probably not a cluster of "bad outliers" if $n_i \geq 4$ and $n \geq 5p$. If $n_i = 1$, 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.

Rule of thumb 10.3. Suppose all $n_i \geq 5$, and consider the spreads of the p dot plots. If the maximum spread is no more that twice the minimum spread, then the one-way MANOVA model may be useful for tests of hypotheses. This rule of thumb is used for the one-way ANOVA model since 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), where standard deviations replace the dot plot spreads. 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 10.2. The above rules are mainly for linearity and tend to use marginal models. The marginal models are useful for checking linearity, but are not very useful for checking other model violations such as outliers in the error vector distribution. The RMVN DD plot of the residual vectors is a global method (takes into account the correlations of $Y_1, ..., Y_m$) for checking the error vector distribution, but is not real effective for detecting outliers since OLS is used to find the residual vectors. A DD plot of residual vectors from a robust MANOVA method might be more effective for detecting outliers. This remark also applies to the plots used in Section 12.2 for multivariate linear regression.

The RMVN DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ is used to check the error vector distribution, to detect outliers, and to display the nonparametric prediction region developed in Section 12.3. The DD plot suggests that the error vector distribution is elliptically contoured if the plotted points cluster tightly about a line through the origin as $n \to \infty$. The plot suggests that the error vector distribution is multivariate normal if the line is the identity line. If n is large and the plotted points do not cluster tightly about a line through the origin, then the error vector distribution may not be elliptically contoured. Make a DD plot of the continuous predictor variables to check for $\boldsymbol{x}$-outliers. These applications of the DD plot for iid multivariate data are discussed in Olive (2002, 2008, 2013a) and Chapter 5. The RMVN estimator has not yet been proven to be a consistent estimator for residual vectors, but simulations suggest that the RMVN DD plot of the residual vectors is a useful diagnostic plot.

Response transformations can also be made as in Section 2.4, but also make the response plot of $\hat{\boldsymbol{Y}}_j$ versus $\boldsymbol{Y}_j$ and use the rules of Section 2.4 on Y_j to linearize the response plot for each of the m response variables $Y_1, ..., Y_m$.

Example 10.1. Consider the one-way MANOVA model on the famous iris data set with $n = 150$ and $p = 3$ species of iris: setosa, versicolor, and virginica. The $m = 4$ variables are $Y_1 =$ *sepal length*, $Y_2 =$ *sepal width*, $Y_3 =$ *petal length*, and $Y_4 =$ *petal width*. See Becker et al. (1988). The plots for the $m = 4$ response variables look similar, and Figure 10.1 shows the response and residual plots for Y_4. Note that the spread of the three dot plots is similar. The dot plot intersects the identity line at the sample mean of the cases in the dot plot. The setosa cases in lowest dot plot have a sample mean of 0.246, and the horizontal line $Y_4 = 0.246$ is below the dot plots for versicolor and virginica which have means of 1.326 and 2.026. Hence the mean petal widths differ for the three species, and it is easier to see this difference in the response plot than the residual plot. The plots for the other three variables are similar. Figure 10.2 shows that the DD plot of the residual vectors suggests that the error vector distribution is elliptically contoured but not multivariate normal.

The DD plot also shows the prediction regions of Section 5.2 computed using the residual vectors $\hat{\boldsymbol{\epsilon}}_i$. From Section 12.3, if $\{\hat{\boldsymbol{\epsilon}} | D_{\hat{\boldsymbol{\epsilon}}}(\boldsymbol{0}, \boldsymbol{S}_r) \leq h\}$ is a prediction region for the residual vectors, then $\{\boldsymbol{y} | D_{\boldsymbol{y}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq h\}$ is a prediction region for $\boldsymbol{y}_f$. For the one-way MANOVA model, a prediction region for $\boldsymbol{y}_f$ would only be valid for an $\boldsymbol{x}_f$ which was observed, i.e., for $\boldsymbol{x}_f = \boldsymbol{x}_j$, since only observed values of the categorical predictor variables make sense. The 90% nonparametric prediction region corresponds to $\boldsymbol{y}$ with distances to the left of the vertical line $MD = 3.2$.

R commands for these two figures are shown below and will also show the plots for Y_1, Y_2, and Y_3. The *mpack* function `manova1w` makes the response and residual plots, while `ddplot4` makes the DD plot. The last command shows that the pvalue = 0 for the one-way MANOVA test discussed in the following section.

```
library(MASS)
y <- iris[,1:4] #m = 4 = number of response variables
group <- iris[,5]
#p = number of groups = number of dot plots
out<- manova1w(y,p=3,group=group) #right click
```

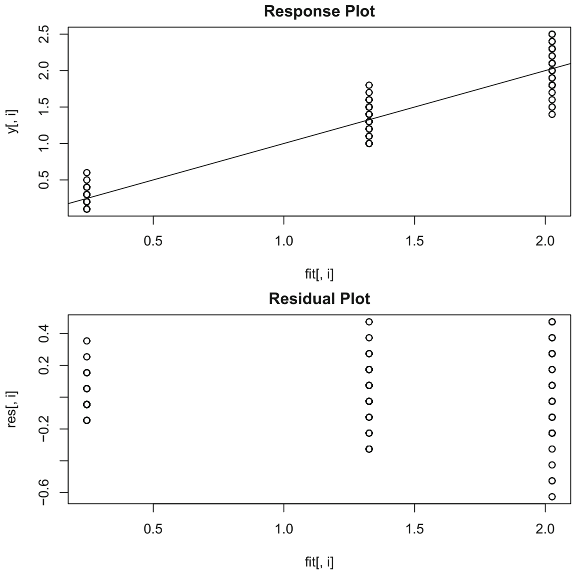

Fig. 10.1 Plots for Y_4 = Petal Width

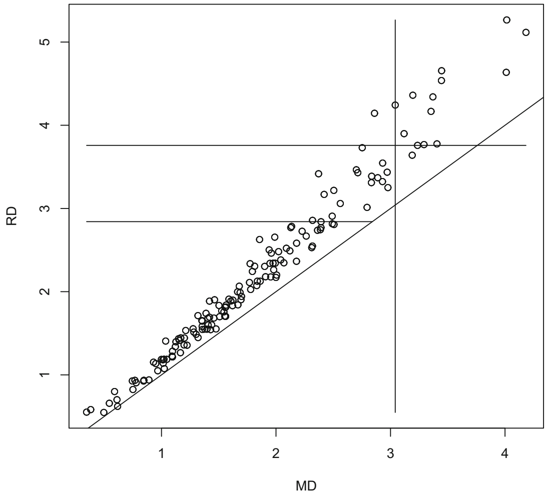

Fig. 10.2 DD Plot of the Residual Vectors for Iris Data

```
#Stop 8 times
ddplot4(out$res) #right click Stop
summary(out$out) #default is Pillai's test
```

10.3 One-Way MANOVA

Using double subscripts will be useful for describing the one-way MANOVA model. Suppose there are independent random samples of size n_i from p different populations (treatments), or n_i cases are randomly assigned to p treatment groups. Then $n = \sum_{i=1}^p n_i$ and 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 group and the jth case (often an individual or thing) in the group. 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 $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. (The common covariance matrix 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 $p = 2$. (In Chapter 9, the notation was different since a case was treated as if it was from a multivariate location and dispersion model with p measurements for each of the $k = 2$ treatments. Now the measurements are treated as m response variables for a multivariate linear model and there are p treatments. Hence k and p from Chapter 9 correspond to p to m, respectively, in this chapter.)

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 $\boldsymbol{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_j + \boldsymbol{\epsilon}_{ij}$, where $\boldsymbol{\epsilon}_{ij}$ 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 vector $\hat{\boldsymbol{\epsilon}}_{ij} = \boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i = \boldsymbol{y}_{ij} - \hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\tau}}_i$. Then $\boldsymbol{y}_{ij} = \overline{\boldsymbol{y}} + (\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}}) + (\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i) = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\tau}}_i + \hat{\boldsymbol{\epsilon}}_{ij}$.

Several $m \times m$ matrices will be useful. 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

$$\boldsymbol{B}_T = \sum_{i=1}^p n_i (\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}})(\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}})^T.$$

The total corrected (for the mean) sum of squares and cross products matrix is $\boldsymbol{T} = \boldsymbol{B}_T + \boldsymbol{W} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}})(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}})^T$. Note that $\boldsymbol{S} = \boldsymbol{T}/(n-1)$ is the usual sample covariance matrix of the $\boldsymbol{y}_{ij}$ 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	$\boldsymbol{B}_T$	$p-1$
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 $\boldsymbol{A} = \boldsymbol{W}, \boldsymbol{B}_T$, or $\boldsymbol{T}$, and *df* is the corresponding degrees of freedom. Let t_0 be the test statistic. Often Pillai's trace statistic, the Hotelling Lawley trace statistic, or Wilks' lambda are used. Wilks' lambda

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{B}_T + \boldsymbol{W}|} = \frac{|\boldsymbol{W}|}{|\boldsymbol{T}|} = \frac{|\sum_{i=1}^{p}(n_i - 1)\boldsymbol{S}_i|}{|(n-1)\boldsymbol{S}|} =$$

$$\frac{|\sum_{i=1}^{p}\sum_{j=1}^{n_i}(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i)(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i)^T|}{|\sum_{i=1}^{p}\sum_{j=1}^{n_i}(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}})(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}})^T|}.$$

Then $t_o = -[n - 0.5(m + p - 2)]\log(\Lambda)$ 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.

Another way to perform the one-way MANOVA test is to get *R* output. The default test is Pillai's test, but other tests can be obtained with the *R* output shown below.

```
summary(out$out) #default is Pillai's test
summary(out$out, test = "Wilks")
summary(out$out, test = "Hotelling-Lawley")
summary(out$out, test = "Roy")
```

Example 10.1, continued. The R output for the iris data gives a Pillai's F statistic of 53.466 and pval = 0.
i) $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_4 \quad H_1$: not H_0
ii) $F = 53.466$
iii) pval = 0
iv) Reject H_0. The means for the three varieties of iris do differ.

Following Mardia et al. (1979, p. 335), let $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_m$ be the eigenvalues of $\boldsymbol{W}^{-1}\boldsymbol{B}_T$. Then $1 + \lambda_i$ for $i = 1, ..., m$ are the eigenvalues of $\boldsymbol{W}^{-1}\boldsymbol{T}$ and $\Lambda = \prod_{i=1}^m (1 + \lambda_i)^{-1}$.

Following Fujikoshi (2002), let the Hotelling Lawley trace statistic $U = tr(\boldsymbol{B}_T \boldsymbol{W}^{-1}) = tr(\boldsymbol{W}^{-1}\boldsymbol{B}_T) = \sum_{i=1}^m \lambda_i$, and let Pillai's trace statistic $V = tr(\boldsymbol{B}_T \boldsymbol{T}^{-1}) = tr(\boldsymbol{T}^{-1}\boldsymbol{B}_T) = \displaystyle\sum_{i=1}^m \frac{\lambda_i}{1 + \lambda_i}$. If the $\boldsymbol{y}_{ij} - \boldsymbol{\mu}_j$ are iid with common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, and if H_0 is true, then under regularity conditions $-[n - 0.5(m + p - 2)] \log(\Lambda) \xrightarrow{D} \chi^2_{m(p-1)}$, $(n - m - p - 1)U \xrightarrow{D} \chi^2_{m(p-1)}$, and $(n - 1)V \xrightarrow{D} \chi^2_{m(p-1)}$. Note that the common covariance matrix assumption implies that each of the p treatment groups or populations has the same covariance matrix $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ for $i = 1, ..., p$, an extremely strong assumption.

A possible alternative method for one-way MANOVA is to use the model $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$ or

$$\begin{bmatrix} Y_{111} & Y_{112} & \cdots & Y_{11m} \\ \vdots & \vdots & \ldots & \vdots \\ Y_{1,n_1,1} & Y_{1,n_1,2} & \cdots & Y_{1,n_1,m} \\ Y_{211} & Y_{211} & \cdots & Y_{21m} \\ \vdots & \vdots & \ldots & \vdots \\ Y_{2,n_2,1} & Y_{2,n_2,2} & \cdots & Y_{2,n_2,m} \\ \vdots & \vdots & \ldots & \vdots \\ Y_{p,11} & Y_{p,1m} & \cdots & Y_{p,1m} \\ \vdots & \vdots & \ldots & \vdots \\ Y_{p,n_p,1} & Y_{p,n_p,2} & \cdots & Y_{p,n_p,m} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \ldots & 1 \\ 1 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \ldots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \ldots & \beta_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p,1} & \beta_{p,2} & \ldots & \beta_{p,m} \end{bmatrix} + \boldsymbol{E}.$$

Then $\boldsymbol{X}$ is full rank, where the ith column of $\boldsymbol{X}$ is an indicator for group $i - 1$ for $i = 2, ..., p$, $\hat{\beta}_{1k} = \overline{Y}_{pok} = \hat{\mu}_{pk}$ for $k = 1, ..., m$, and

$$\hat{\beta}_{ik} = \overline{Y}_{i-1,ok} - \overline{Y}_{pok} = \hat{\mu}_{i-1,k} - \hat{\mu}_{pk}$$

for $k = 1, ..., m$ and $i = 2, ..., p$. Thus testing $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ is equivalent to testing $H_0 : \boldsymbol{LB} = \boldsymbol{0}$, where $\boldsymbol{L} = [\boldsymbol{0} \ \boldsymbol{I}_{p-1}]$. Such tests are discussed in Section 12.4. Press (2005, p. 262) used the above model.

Then $\boldsymbol{y}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_{ij}$ and

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\mu}_p^T \\ \boldsymbol{\mu}_1^T - \boldsymbol{\mu}_p^T \\ \boldsymbol{\mu}_2^T - \boldsymbol{\mu}_p^T \\ \vdots \\ \boldsymbol{\mu}_{p-2}^T - \boldsymbol{\mu}_p^T \\ \boldsymbol{\mu}_{p-1}^T - \boldsymbol{\mu}_p^T \end{bmatrix}.$$

Remark 10.3. Since the common covariance matrix assumption $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_\epsilon$ for $k = 1, ..., n$ is extremely strong, using the bootstrap prediction region method to test $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ may be a useful alternative. Take a sample of size n_i with replacement from the n_i cases for each group $i = 1, 2, ..., p$. Let the $(p-1)m \times 1$ vector $\boldsymbol{w}_i = vec(\boldsymbol{L}\hat{\boldsymbol{B}}_i^*) = ((\hat{\boldsymbol{\mu}}_1^* - \hat{\boldsymbol{\mu}}_p^*)^T, ..., (\hat{\boldsymbol{\mu}}_{p-1}^* - \hat{\boldsymbol{\mu}}_p^*)^T)^T$ for $i = 1, ..., B$, where $vec(\boldsymbol{A})$ is defined below Theorem 12.6. For a robust test, use $\boldsymbol{w}_i = ((T_1^* - T_p^*)^T, ..., (T_{p-1}^* - T_p^*)^T)^T$ where T_i is a robust location estimator, such as the coordinatewise median or RMVN location estimator, applied to the cases in the ith treatment group. Likely need $n \geq 40mp$, $n \geq (m+p)^2$, and $n_i \geq 40m$. See Rupasinghe Arachchige Don (2017) and Rupasinghe Arachchige Don and Olive (2017).

Large sample theory can be also be used to derive a better test. Let $\boldsymbol{\Sigma}_i$ be the nonsingular population covariance matrix of the ith treatment group or population. To simplify the large sample theory, assume $n_i = \pi_i n$, where $0 < \pi_i < 1$ and $\sum_{i=1}^p \pi_i = 1$. Assume H_0 is true, and let $\boldsymbol{\mu}_i = \boldsymbol{\mu}$ for $i = 1, ..., p$. Then by the multivariate central limit theorem, $\sqrt{n_i}(\overline{\boldsymbol{y}}_i - \boldsymbol{\mu}) \xrightarrow{D} N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$, and $\sqrt{n}(\overline{\boldsymbol{y}}_i - \boldsymbol{\mu}) \xrightarrow{D} N_m\left(\boldsymbol{0}, \dfrac{\boldsymbol{\Sigma}_i}{\pi_i}\right)$. Let

$$\boldsymbol{w} = \begin{bmatrix} \overline{\boldsymbol{y}}_1 - \overline{\boldsymbol{y}}_p \\ \overline{\boldsymbol{y}}_2 - \overline{\boldsymbol{y}}_p \\ \vdots \\ \overline{\boldsymbol{y}}_{p-2} - \overline{\boldsymbol{y}}_p \\ \overline{\boldsymbol{y}}_{p-1} - \overline{\boldsymbol{y}}_p \end{bmatrix}.$$

Then $\sqrt{n}\boldsymbol{w} \xrightarrow{D} N_{m(p-1)}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ with $\boldsymbol{\Sigma}_{\boldsymbol{w}} = (\boldsymbol{\Sigma}_{ij})$, where $\boldsymbol{\Sigma}_{ij} = \text{Cov}(\sqrt{n}(\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}}_p), \sqrt{n}(\overline{\boldsymbol{y}}_j - \overline{\boldsymbol{y}}_p)) = \frac{\boldsymbol{\Sigma}_p}{\pi_p}$ for $i \neq j$, and $\boldsymbol{\Sigma}_{ii} = \text{Cov}(\sqrt{n}(\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}}_p)) = \frac{\boldsymbol{\Sigma}_i}{\pi_i} + \frac{\boldsymbol{\Sigma}_p}{\pi_p}$ for $i = j$. Hence

$$t_0 = n\boldsymbol{w}^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1} \boldsymbol{w} = \boldsymbol{w}^T \left(\frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}}{n} \right)^{-1} \boldsymbol{w} \xrightarrow{D} \chi^2_{m(p-1)}$$

as the $n_i \to \infty$ if H_0 is true. Here

$$\frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}}{n} = \begin{bmatrix} \frac{\boldsymbol{S}_1}{n_1} + \frac{\boldsymbol{S}_p}{n_p} & \frac{\boldsymbol{S}_p}{n_p} & \frac{\boldsymbol{S}_p}{n_p} & \cdots & \frac{\boldsymbol{S}_p}{n_p} \\ \frac{\boldsymbol{S}_p}{n_p} & \frac{\boldsymbol{S}_2}{n_2} + \frac{\boldsymbol{S}_p}{n_p} & \frac{\boldsymbol{S}_p}{n_p} & \cdots & \frac{\boldsymbol{S}_p}{n_p} \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{\boldsymbol{S}_p}{n_p} & \frac{\boldsymbol{S}_p}{n_p} & \frac{\boldsymbol{S}_p}{n_p} & \cdots & \frac{\boldsymbol{S}_{p-1}}{n_{p-1}} + \frac{\boldsymbol{S}_p}{n_p} \end{bmatrix}$$

is a block matrix, where the off-diagonal block entries equal $\boldsymbol{S}_p/n_p$ and the ith diagonal block entry is $\dfrac{\boldsymbol{S}_i}{n_i} + \dfrac{\boldsymbol{S}_p}{n_p}$ for $i = 1, ..., (p-1)$. Reject H_0 if $t_0 > m(p-1)F_{m(p-1),d_n}(1-\alpha)$, where $d_n = \min(n_1, ..., n_p)$. It may make sense to relabel the groups so that n_p is the largest n_i or $\boldsymbol{S}_p/n_p$ has the smallest generalized variance of the $\boldsymbol{S}_i/n_i$. This test may start to outperform the one-way MANOVA test if $n \geq (m+p)^2$ and $n_i \geq 20m$ for $i = 1, ..., p$.

10.4 Two-Way MANOVA

The two-way MANOVA model is the multivariate generalization of the two-way ANOVA model. There are m response variables $Y_1, ..., Y_m$. There are two factors A and B. Factor A has a levels, and factor B has b levels. Example 10.2 will illustrate the R output that can be generated for this model.

Definition 10.8. The **main effects** are A and B. The AB interaction is not a main effect.

Remark 10.4. If A and B are factors, then there are five possible models.
i) The two-way MANOVA model has terms A, B, and AB.
ii) The additive model or main effects model has terms A and B.
iii) The one-way MANOVA model that uses factor A.
iv) The one-way MANOVA model that uses factor B.
v) The null model does not use any of the three terms A, B, or AB. If the null model holds, then the factors have no effect on the m response variables $Y_1, ..., Y_m$, or each combination of the ab factor levels has the same effect on the m response variables.

Example 10.2. This example on producing plastic film is taken from the R help files and uses data from Krzanowski (1988, p. 381). There are

$m = 3$ response variables on the plastic: *tear*, *gloss*, and *opacity*. There are two explanatory variables *rate* and *additive* that have two levels: low and high. First, the *R* commands below fit a one-way MANOVA model using *rate* as the explanatory variable. The means for the two values of *rate* appear to differ, both from the plots and from the small pval of the one-way MANOVA test.

The output for the two-way MANOVA tests suggest that the interaction is not needed since pval = 0.3. The two-way MANOVA model is refitted with the interaction deleted. For this competing model, both factors were significant.

```
tear <- c(6.5, 6.2, 5.8, 6.5, 6.5, 6.9, 7.2, 6.9, 6.1,
6.3, 6.7, 6.6, 7.2, 7.1, 6.8, 7.1, 7.0, 7.2, 7.5, 7.6)
gloss <- c(9.5, 9.9, 9.6, 9.6, 9.2, 9.1, 10.0, 9.9,
9.5, 9.4, 9.1, 9.3, 8.3, 8.4, 8.5, 9.2, 8.8, 9.7,
10.1, 9.2)
opacity <- c(4.4, 6.4, 3.0, 4.1, 0.8, 5.7, 2.0, 3.9,
1.9, 5.7, 2.8, 4.1, 3.8, 1.6, 3.4, 8.4, 5.2, 6.9,
2.7, 1.9)
Y <- cbind(tear, gloss, opacity)
rate <- factor(gl(2,10), labels=c("Low", "High"))
additive <- factor(gl(2, 5, length=20),
labels=c("Low", "High"))

#one way MANOVA model using rate as a predictor
fit <- manova(Y ~ rate )
summary.aov(fit)           #univariate ANOVA tables
summary(fit)               #MANOVA table with Pillai
summary(fit, test="Wilks") #Wilks’ lambda
summary(fit,test = "Hotelling-Lawley")
summary(fit,test = "Roy")
#for one way MANOVA with df = 1 (p = 2 groups),
#the 4 tests are the same =
#two sample Hotelling’s T^2 test
grp <- as.integer(rate)
out<-manova1w(y=Y,p=2,group=grp)
summary(out$out,test="Hotelling-Lawley")
#for two way MANOVA, the 4 tests seem to be the same
#for df = 1 or p = 2 groups
#two way MANOVA model
fit <- manova(Y ~ rate * additive)
summary.aov(fit)   # univariate two way ANOVA tables
summary(fit, test="Wilks") # Wilks’ lambda
summary(fit)             # Pillai’s test: the default
```

```
                Df  Pillai approx F numDf denDf  Pr(>F)
rate             1 0.61814  7.5543      3    14  0.00303
additive         1 0.47697  4.2556      3    14  0.02475
rate:additive 1 0.22289  1.3385      3    14  0.30178

#delete the interaction to get the additive model
fit <- manova(Y ~ rate + additive)
summary.aov(fit)   # univariate two way ANOVA tables
summary(fit, test="Wilks")#MANOVA with Wilks' lambda
summary(fit)                #Pillai
          Df  Pillai approx F num Df den Df  Pr(>F)
rate       1 0.61316   7.9253      3     15   0.00212
additive   1 0.44616   4.0279      3     15   0.02753
```

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 $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$. 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 $E(Y_{ij}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j$.

The data matrix $\boldsymbol{W} = [\boldsymbol{X} \;\; \boldsymbol{Z}]$ except usually the first column $\boldsymbol{1}$ of $\boldsymbol{X}$ is omitted if $x_{i,1} \equiv 1$. 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 $\boldsymbol{v}_1 = \boldsymbol{1}$.

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 $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, 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 $\boldsymbol{Y}_j = \boldsymbol{X\beta}_j + \boldsymbol{e}_j$ 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) 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 $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$.

5) 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 residual vectors $\hat{\boldsymbol{\epsilon}}_i$.

6) 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 $\boldsymbol{\Sigma_\epsilon}$. 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 $i = 1, ..., m$.

7) For the one-way MANOVA model, make a DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ 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. We want $n \geq (m + p)^2$ and $n_i \geq 10m$.

8) 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 $\hat{Y}_{ijk} = \hat{\mu}_{ik} = \overline{Y}_{iok}$ for $i = 1, ..., p$. So for the kth variable, the means $\mu_{1k}, ..., \mu_{pk}$ are of interest. The residuals are $r_{ijk} = Y_{ijk} - \hat{Y}_{ijk}$. For each variable Y_k make a response plot of $\overline{Y}_{iok}$ versus Y_{ijk} and a residual plot of $\overline{Y}_{iok}$ versus r_{ijk}. Both plots will consist of p dot plots of n_i cases located at the $\overline{Y}_{iok}$. 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 $n_i \geq 5$, we want $\max(R_{1k}, ..., R_{pk}) \leq 2 \min(R_{1k}, ..., R_{pk})$. The one-way MANOVA model may be reasonable for the test in point 9) if the m response and residual plots satisfy the above graphical checks.

9) 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. As a textbook convention, use $\alpha = 0.05$ if α is not given.

10) The one-way MANOVA test assumes that the p treatment groups or populations have the same covariance matrix: $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_p$, but the test has some resistance to this assumption. See points 6) and 8).

10.6 Complements

Many other MANOVA models can be made that are multivariate generalizations of the corresponding univariate ANOVA models, and the R function `manova` can likely fit several of these models. Functions to produce the response and residual plots as well as the DD plot of the residual vectors should be made. The large sample theory from the literature should be examined to see which tests are good and robust to nonnormality. See Olive (2010; 2017a; ch. 5-9) for references on univariate $m = 1$ ANOVA models.

Fujikoshi (2002) showed that the one-way MANOVA test statistics have an asymptotic chi-square distribution for a large class of iid error distributions. See Wakaki et al. (2002) for more results including some for the two-way MANOVA model. Kakizawa (2009) showed that the Hotelling Lawley, Pillai's trace, and Wilks' Λ tests for some MANOVA models are large sample tests for a large class of iid error distributions. Similar tests are developed in Chapter 12 for multivariate linear regression.

Hand and Taylor (1987), Huberty and Olejnik (2006), and Khattree and Naik (1999, ch. 4) are useful reference for MANOVA. Mardia (1971) noted 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). Konietschke et al. (2015) proposed bootstrap tests that appear to perform better than the Wilks' Λ test.

Aelst and Willems (2011) gave references for robust one-way MANOVA tests. The FS and FMM estimators used are not yet backed by theory.

Section 9.1 gives two robust methods when $p = 2$ where the one-way MANOVA model reduces to a Hotelling's T^2 test. Two diagnostics for the one-way MANOVA model are very similar to the robust discriminant analysis methods of Section 8.9. As a diagnostic, run the classical one-way MANOVA model on U_{big} given in Sections 4.6 and 8.9, where the G groups for discriminant analysis are replaced by the p groups for one-way MANOVA.

Alternatively, let T_j be the coordinatewise median for the jth group. Let $\boldsymbol{z}_{ij} = \boldsymbol{y}_{ij} - T_j$ for $i = 1, ..., n_j$ and $j = 1, ..., p$. Then find the RMVN set U_c for all n $\boldsymbol{z}_{ij}$ and the cases in the set. Then run the classical one-way MANOVA model on the $\boldsymbol{y}_{ij}$ corresponding to the cases in the RMVN set. It is unlikely that the test statistics run on the cleaned data have a limiting χ^2 distribution. The output below demonstrates the diagnostics based on U_{big} and U_c.

```
y<-turtle[,1:3] #need mrobdata
group<-turtle[,4]+1
cleanb <- getubig(y,group)
Yb <- cleanb$Ubig
grpb <- cleanb$grp #m = 3, p = 2 groups
outb <- manova1w(y=Yb,p=2,group=grpb)
#right click Stop 6 times
cleanc <- getuc(y,group)
Yc <- cleanc$Uc
grpc <- cleanc$grp
outc <- manova1w(y=Yc,p=2,group=grpc)
#right click Stop 6 times
```

Bootstrapping analogs of the one-way MANOVA test is useful. Consider testing $H_0 : \boldsymbol{LB} = \boldsymbol{0}$. Take a sample of size n_i with replacement from each group. From the combined sample, find $\boldsymbol{w}_1 = vec(\boldsymbol{L}\hat{\boldsymbol{B}}^*_1)$. Repeat B times to get a bootstrap sample $\boldsymbol{w}_1, ..., \boldsymbol{w}_B$ where $\boldsymbol{w}_i = vec(\boldsymbol{L}\hat{\boldsymbol{B}}^*_i)$. (The vec operator is described under Theorem 12.6.) Apply the nonparametric prediction region to the $\boldsymbol{w}_i$ and see if $\boldsymbol{0}$ is in the region. If $\boldsymbol{L}$ is $(p-1)\times p$, then $\boldsymbol{w}$ is $m(p-1)\times 1$, and we likely need $n \geq 40mp$, $n \geq (m+p)^2$, and $n_i \geq 40m$. See Remark 10.3, Rupasinghe Arachchige Don (2017), and Rupasinghe Arachchige Don and Olive (2017).

Some other tests that assume the different groups have different covariance matrices are given by Zhang and Liu (2013), and Zhang et al. (2016).

10.7 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

10.1*. If $\boldsymbol{X}$ is of full rank and least squares is used to fit the MANOVA model, then $\hat{\boldsymbol{\beta}}_i = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_i$, and $\boldsymbol{Y}_i = \boldsymbol{X}\boldsymbol{\beta}_i + \boldsymbol{e}_i$. Treating $\boldsymbol{X}\boldsymbol{\beta}_i$ as a constant, $\text{Cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = \text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$. Using this information, show $\text{Cov}(\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\beta}}_j) = \sigma_{ij}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

10.2. SAS Institute (1985), pp. 498 - 501 described a one-way MANOVA model. There are two groups for gender: female and male. There were $p = 4$ (skull measurements) variables $X_1 = \textit{length}$, $X_2 = \textit{basilar}$, $X_3 = \textit{zygomat}$, and $X_4 = \textit{postorb}$. 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 Wilks' lambda, but the other three test statistics gave the same pvalue. Do a four-step one-way MANOVA test.

R 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, e.g., *ddplot*, will display the code for the function. Use the `args` command, e.g., *args(ddplot)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/mrsashw.txt) into *R*.

10.3. 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 makes 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 tend to be smaller than the female turtles.

b) The *R* command for this plot makes a DD plot of the residual vectors 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 residual vectors appear to follow a multivariate normal distribution? (Right click *Stop* once on the plot.)

Problem 10.4. Use the *R* commands for this problem to obtain output for Example 10.2.

Chapter 11
Factor Analysis

Factor analysis gives an approximation of the dispersion matrix

$$\hat{\boldsymbol{\Sigma}} \approx \hat{\boldsymbol{L}}^T \hat{\boldsymbol{L}} + \hat{\boldsymbol{\Psi}},$$

so $\hat{\boldsymbol{\Sigma}} \approx \hat{\boldsymbol{L}}^T \hat{\boldsymbol{L}}$ if $\hat{\boldsymbol{\Psi}}$ is small. Factor analysis clusters variables into groups called factors and suggests that the factors can explain the dispersion more simply than $X_1, ..., X_p$.

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.

Sometimes candidate groups can be spotted using the correlation matrix $\boldsymbol{R} = (r_{ij})$: join the two variables with the highest absolute correlation $|r_{ij}|$ into a group, provided $|r_{ij}| \geq c$. The other $p-2$ variables form groups of size 1. Then add the variable with the highest $|r_{ij}| \geq c$ to a group. Continue the process. If all $|r_{ij}| < c$, then there will be p groups each with one variable. Use $c = 0.9, 0.8, 0.7$, etc.

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_11

	1	2	3	4	5	6
1 Classics		0.83	0.78	0.70	0.66	0.63
2 French			0.67	0.67	0.65	0.57
3 English				0.64	0.54	0.51
4 Math					0.45	0.51
5 Discrimination						0.40
6 Music						

Example 11.1. Spearman (1904), the starting point of factor analysis, gave the correlation matrix for examination scores in six subjects for 33 students, shown in the above table. See Kendall (1980, p. 47). If $0 < c \leq 0.4$, put all six variables in one group. If $c = 0.83$, put Classics and French in group 1 while all other variables have their own group. If $0.7 < c \leq 0.78$, let group 1 = Classics, French, and English = "language factor," group 2 = Math, group 3 = Discrimination, and group 4 = music. If $0.67 \leq c \leq 0.7$, let group 1 = Classics, French, English, and Math, group 2 = Discrimination, and group 3 = music. If $c = 0.66$, let group 1 be all variables except Music and group 2 = Music.

Some notation is needed before presenting factor analysis models. 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 $\boldsymbol{e}_i \boldsymbol{e}_i^T = (-\boldsymbol{e}_i)(-\boldsymbol{e}_i)^T$.

Definition 11.1. Suppose there are p random variables, let the $p \times 1$ random vector $\boldsymbol{x} = (X_1, ..., X_p)^T$ and assume that m factors are used.

a) A population factor analysis approximation of the dispersion matrix is $\boldsymbol{\Sigma} \approx \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi} \equiv \boldsymbol{\Sigma}_F$, where the $p \times m$ *matrix of factor loadings* $\boldsymbol{L} = (l_{ij})$, and $\boldsymbol{\Psi} = diag(\psi_1, ..., \psi_p)$ is a diagonal matrix so that the approximation is exact for the diagonal elements: $\boldsymbol{\Sigma}_{ii} = \boldsymbol{\Sigma}_{F,ii}$.

b) A sample factor analysis approximation of the dispersion matrix is $\hat{\boldsymbol{\Sigma}} \approx \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}} \equiv \hat{\boldsymbol{\Sigma}}_F$, where the $p \times m$ *matrix of factor loadings* $\hat{\boldsymbol{L}} = (l_{ij})$, and $\hat{\boldsymbol{\Psi}} = diag(\hat{\psi}_1, ..., \hat{\psi}_p)$ is a diagonal matrix so that the approximation is exact for the diagonal elements: $\hat{\boldsymbol{\Sigma}}_{ii} = \hat{\boldsymbol{\Sigma}}_{F,ii}$. Hence $(\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T)_{ii} + \hat{\psi}_i = \hat{\boldsymbol{\Sigma}}_{ii}$. 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 $\hat{\psi}_i$ are called *uniquenesses*. 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 estimated communalities are unaffected by the choice of $\boldsymbol{\Gamma}$ since $\hat{h}_i^2 = (\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T)_{ii} = [\hat{\boldsymbol{L}}^*(\hat{\boldsymbol{L}}^*)^T]_{ii}$.

Several methods of factor analysis have been proposed. The principal component factor analysis and maximum likelihood factor analysis models are special cases of the orthogonal factor analysis model.

Definition 11.2. For *principal component factor analysis*, the ith column of the $p \times m$ matrix $\hat{\boldsymbol{L}}$ is $\sqrt{\hat{\lambda}_i}\hat{\boldsymbol{e}}_i$ where $m < p$. Then

$$\hat{\boldsymbol{L}} = \left[\sqrt{\hat{\lambda}_1}\hat{\boldsymbol{e}}_1 \ \sqrt{\hat{\lambda}_2}\hat{\boldsymbol{e}}_2 \ \dots \ \sqrt{\hat{\lambda}_m}\hat{\boldsymbol{e}}_m \right].$$

Then $\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^m \hat{\lambda}_i \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^T + \sum_{i=m+1}^p \hat{\lambda}_i \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^T = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \sum_{i=m+1}^p \hat{\lambda}_i \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^T \approx \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}} \equiv \hat{\boldsymbol{\Sigma}}_F$ where $\hat{\boldsymbol{\Psi}} = diag(\hat{\psi}_1, ..., \hat{\psi}_p)$ and $\hat{\boldsymbol{\Sigma}}_{ii} = \hat{\boldsymbol{\Sigma}}_{F,ii}$. Hence $(\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T)_{ii} + \hat{\psi}_i = \hat{\boldsymbol{\Sigma}}_{ii}$. The ith column $\sqrt{\hat{\lambda}_i}\hat{\boldsymbol{e}}_i$ of $\hat{\boldsymbol{L}}$ gives the estimated factor loadings for factor F_i. These estimated factor loadings do not change as m is increased.

Definition 11.3. The *orthogonal factor analysis model* is $\boldsymbol{x} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\epsilon}$ where the $p \times 1$ random vector $\boldsymbol{x} = (X_1, ..., X_p)^T$, 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 $\boldsymbol{\Sigma} \approx \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi} = \boldsymbol{\Sigma}_F$ 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 *loading* of the ith variable on the jth factor $= l_{ij}$. Note that

$$X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \epsilon_1$$

$$X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \epsilon_2$$

$$\vdots \qquad\qquad \vdots$$

$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \epsilon_p.$$

Data often does not have this structure, so an important question is 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 $E(\boldsymbol{F}) = \boldsymbol{0}$, $\text{Cov}(\boldsymbol{F}) = \boldsymbol{I}_m$, $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$, $\text{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi}$, and that $\boldsymbol{F}$ and $\boldsymbol{\epsilon}$ are independent. Then $\text{Cov}(\boldsymbol{x}, \boldsymbol{F}) = \boldsymbol{L}$ or $\text{Cov}(X_i, F_j) = l_{ij}$, and $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi} = \boldsymbol{\Sigma}_F$.

Rule of thumb 11.1. The fact that $\text{Cov}(X_i, F_j) = l_{ij}$ and that different methods of factor analysis tend to give similar results means that the factor analysis output is interpreted much like PCA output, especially when the principal component factor analysis method is used.

a) 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
$\hat{\boldsymbol{L}}_1$	$\hat{\boldsymbol{L}}_2$	$\cdots$	$\hat{\boldsymbol{L}}_m$

b) 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 entries. 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.

c) 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%.

If $\boldsymbol{\Gamma}$ is an orthogonal matrix, then $\hat{\boldsymbol{L}}^* = \hat{\boldsymbol{L}}\boldsymbol{\Gamma}$ is also a matrix of estimated factor loadings. This multiplication corresponds to a rotation. The varimax and promax rotations seek $\hat{\boldsymbol{\Gamma}}$ such that $\hat{\boldsymbol{L}}^* = \hat{\boldsymbol{L}}\hat{\boldsymbol{\Gamma}}$ 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. Then variables with nonzero loadings are "important." Hence Rule of thumb 11.1 is often easier to apply after a varimax or promax rotation. The varimax rotation is orthogonal, but the promax rotation is oblique (nonorthogonal), and thus the approximation $\hat{\boldsymbol{\Sigma}} \approx \hat{\boldsymbol{L}}^T\hat{\boldsymbol{L}} + \hat{\boldsymbol{\Psi}}$ is often not good if $\hat{\boldsymbol{L}}$ is from the promax rotation.

Rule of thumb 11.2. To use factor analysis, assume the DD plot and subplots of the scatterplot matrix are linear. Typically, $\hat{\boldsymbol{\Sigma}} = \boldsymbol{R}$ and standardized data are used. We want $n \geq 10p$ for classical factor analysis and $n \geq 20p$ 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 $\boldsymbol{R}$. We want the *proportion of the trace explained* by the first m factors $= \sum_{i=1}^m \hat{\lambda}_i / \sum_{j=1}^p \hat{\lambda}_j = \sum_{i=1}^m \hat{\lambda}_i / tr(\hat{\boldsymbol{\Sigma}}) > 0.7$. We want $m < \min(10, p)$. Suppose $(T, \hat{\boldsymbol{\Sigma}})$ is the estimator of multivariate location and dispersion. Make a plot of $D_i(T, \hat{\boldsymbol{\Sigma}}_F)$ versus $D_i(T, \hat{\boldsymbol{\Sigma}})$ with the identity line that has unit slope and zero intercept added as a visual aid. If $\hat{\boldsymbol{\Sigma}}_F$ is an adequate approximation of $\hat{\boldsymbol{\Sigma}}$, then the plotted points should cluster tightly about the identity line. See Figure 11.1.

Definition 11.4. *Principal axis factor analysis* or *principal factor factor analysis* replaces the correlation matrix $\boldsymbol{R} = (r_{ij})$ by $\boldsymbol{R}_P = \boldsymbol{R} + diag(\hat{c}_1 -$

$r_{11}, ..., \hat{c}_p - r_{pp})$ where $\hat{c}_i$ is an estimated communality. Hence the 1s on the diagonal of $\boldsymbol{R}$ are replaced by the $\hat{c}_i$ on the diagonal of $\boldsymbol{R}_P$. There are three common methods. a) Use $\hat{c}_i = \hat{h}_i^2$ from the principal component factor analysis. b) Use the "squared multiple correlation coefficients" $\hat{c}_i = R_i^2 = 1 - 1/r^{ii}$ where r^{ii} is the ith diagonal element of $\boldsymbol{R}^{-1}$ (the r^{ii} are partial correlations: see the second paragraph after Rule of thumb 2.1). c) The *iterated principal factor method*: use a) as the starting point: the $\hat{c}_{i1}$ are the estimated communalities from the principal component factor analysis. The factorization will produce estimated communalities $\hat{c}_{i2}$. Iterate until the estimated communalities converge, usually after a few iterations.

Since $\boldsymbol{R}$ and the generalized correlation matrix based on the FCH, RFCH, and RMVN estimator converge in distribution to the population correlation matrix $\boldsymbol{\rho}$, the robust and classical principal axis factor analyses should give similar results if n is large and the data is iid from a large class of elliptically contoured distributions for methods a) and b) above. This result is conjectured to hold for method c).

Definition 11.5. *Maximum likelihood factor analysis* uses

$$\hat{\boldsymbol{\Sigma}} \approx \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}} \equiv \hat{\boldsymbol{\Sigma}}_F$$

where $\hat{\boldsymbol{L}}$ and $\hat{\boldsymbol{\Psi}}$ are maximum likelihood estimators of $\boldsymbol{L}$ and $\boldsymbol{\Psi}$ assuming a multivariate normal likelihood and m factors, subject to $\hat{\boldsymbol{L}}^T\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{L}}$ being diagonal.

Johnson and Wichern (1988, p. 395) suggested that if $\hat{\boldsymbol{\Sigma}}$ is the correlation matrix $\boldsymbol{R}$, then the elements of $\boldsymbol{R} - \hat{\boldsymbol{\Sigma}}_F$ tend to be smaller for the maximum likelihood method than for the principal component method. If this result is true, then the maximum likelihood method is robust to normality. In general, factor analysis methods are used to approximate $\boldsymbol{R}$ or $\boldsymbol{S}$, and if the factor method was not robust to nonnormality, then the method would not be popular since multivariate normal data sets are rather rare.

Remark 11.1. A k factor model makes sense if the degrees of freedom $d \geq 0$ where $d = 0.5(p-k)^2 - 0.5(p+k)$.

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 $\hat{\boldsymbol{\Sigma}} \overset{P}{\to} c\boldsymbol{\Sigma} = d\text{Cov}(\boldsymbol{x})$ while $\boldsymbol{S} \overset{P}{\to} c_X\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{x})$. If the generalized correlation matrix is used as $\hat{\boldsymbol{\Sigma}}$, then the classical and robust methods both satisfy $\hat{\boldsymbol{\Sigma}} \overset{P}{\longrightarrow} \boldsymbol{\rho}$. The RMVN method is easy to program since it is a classical factor analysis applied to the RMVN subset U. See Definition 4.11 and Section 4.6.

As a diagnostic, run the factor analysis method on the RMVN set U. This method is backed by theory for the principal component factor analysis method, and for at least two of the principal axis factor analysis methods, (a) and b) in Definition 11.4), for a large class of elliptically contoured distributions. For the maximum likelihood factor analysis method, using the RMVN set U is currently only backed by theory for multivariate normal data. We need theory proving the conjecture that the maximum likelihood method works for a large class of elliptically contoured distributions.

Example 11.2. The Venables and Ripley (2003) *MASS* library function `factanal` computes maximum likelihood factor analysis. The default appears to use the correlation matrix. The *R* help files make the following artificial data set. The plot command shows that factor 1 loads high on variables 1 and 2 as does the column under "Factor 1." The "uniquenesses" are the $\hat{\psi}_i$.

```
v1 <- c(1,1,1,1,1,1,1,1,1,1,3,3,3,3,3,4,5,6)
v2 <- c(1,2,1,1,1,1,2,1,2,1,3,4,3,3,3,4,6,5)
v3 <- c(3,3,3,3,3,1,1,1,1,1,1,1,1,1,1,5,4,6)
v4 <- c(3,3,4,3,3,1,1,2,1,1,1,1,2,1,1,5,6,4)
v5 <- c(1,1,1,1,1,3,3,3,3,3,1,1,1,1,1,6,4,5)
v6 <- c(1,1,1,2,1,3,3,3,4,3,1,1,1,2,1,6,5,4)
x <- cbind(v1,v2,v3,v4,v5,v6); cor(x)
   v1     v2     v3     v4     v5     v6
v1 1.0000 0.9393 0.5129 0.4320 0.4665 0.4086
v2 0.9393 1.0000 0.4124 0.4084 0.4364 0.4326
v3 0.5129 0.4124 1.0000 0.8771 0.5129 0.4320
v4 0.4320 0.4084 0.8771 1.0000 0.4320 0.4323
v5 0.4665 0.4364 0.5129 0.4320 1.0000 0.9473
v6 0.4086 0.4326 0.4320 0.4323 0.9473 1.0000
out1 <- factanal(x, factors = 3)
plot(out1$loadings[,1]); out1
Uniquenesses:v1   v2    v3    v4    v5    v6
          0.005 0.101 0.005 0.224 0.084 0.005
Loadings:
   Factor1 Factor2 Factor3
v1 0.944   0.182   0.267
v2 0.905   0.235   0.159
v3 0.236   0.210   0.946
v4 0.180   0.242   0.828
v5 0.242   0.881   0.286
v6 0.193   0.959   0.196
               Factor1 Factor2 Factor3
SS loadings      1.893   1.886   1.797
Proportion Var   0.316   0.314   0.300
```

```
Cumulative Var   0.316   0.630   0.929
#used varimax rotation
#factor 1 is almost an average of v1 and v2
#factor 2 is almost an average of v5 and v6
#factor 3 is almost an average of v3 and v4
out2 <- factanal(x, factors = 3, rotation = "promax")
out2
Uniquenesses:   v1    v2    v3    v4    v5    v6
             0.005 0.101 0.005 0.224 0.084 0.005
Loadings:   Factor1   Factor2   Factor3
            v1           0.985
            v2           0.951
            v3                     1.003
            v4                     0.867
            v5  0.910
            v6  1.033
               Factor1 Factor2 Factor3
SS loadings      1.903   1.876   1.772
Proportion Var   0.317   0.313   0.295
Cumulative Var   0.317   0.630   0.925
##promax rotation tries to give 0 loadings to lots of
##variables in the factor
```

Example 11.3. Factor analysis can also be performed by supplying a covariance matrix or a correlation matrix. As a diagnostic, supply the RMVN dispersion matrix or the RMVN generalized correlation matrix. The following *R* covariance matrix was used. The program can be run with 1 factor, then 2, ..., k. *R* gives a test for whether the j factors are significant. Sometimes pvalue $< \alpha$ seems to suggest that the j factors are inadequate, as in this example, but sometimes pvalue $< \alpha$ seems to suggest that j factors are adequate. See Example 11.5. Use $\alpha = 0.05$ or 0.01. See output below.

```
out1 <- factanal(factors = 1, covmat=ability.cov)
out1
Uniquenesses:
general picture  blocks    maze reading   vocab
  0.535   0.853   0.748   0.910   0.232   0.280

Loadings: Factor1
general    0.682
picture    0.384
blocks     0.502
maze       0.300
reading    0.877
vocab      0.849
```

```
               Factor1
SS loadings      2.443
Proportion Var   0.407

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 75.18 on 9 degrees of
freedom. The p-value is 1.46e-12

out2 <- factanal(factors = 2, covmat=ability.cov)
out2
Uniquenesses:
general picture  blocks    maze reading   vocab
  0.455   0.589   0.218   0.769   0.052   0.334
Loadings: Factor1 Factor2
  general 0.499   0.543
  picture 0.156   0.622
  blocks  0.206   0.860
  maze    0.109   0.468
  reading 0.956   0.182
  vocab   0.785   0.225

               Factor1 Factor2
SS loadings      1.858   1.724
Proportion Var   0.310   0.287
Cumulative Var   0.310   0.597

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 6.11 on 4 degrees of
freedom. The p-value is 0.191.
##Seem to want pvalue > 0.01 to suggest that there
##are enough factors.
```

Example 11.4. The Buxton (1920) data has five massive outliers for the variables len = head length and buxy = height. Supplying the RMVN dispersion matrix will still use all n cases. Using the RMVN set U does not use all n cases, so the test statistic for the number of factors changes.

```
z <- cbind(buxx,buxy) #Outliers ruin the FA.
covhat <- var(z)
out2 <- factanal(factors = 2, covmat=covhat)
out2 #out2 <- factanal(z,factors = 2) gives
     #the same result
Uniquenesses:
     len    nasal  bigonal cephalic     buxy
   0.018    0.005    0.992    0.982    0.005
```

```
Loadings:
         Factor1 Factor2
len      -0.983   0.123
nasal     0.251   0.965
bigonal
cephalic         -0.106
buxy      0.996
               Factor1 Factor2
SS loadings      2.029   0.969
Proportion Var   0.406   0.194
Cumulative Var   0.406   0.600

The degrees of freedom for the model is 1 and the fit
was 0.0125.

#The following command lets you examine a different
#rotation.
update(out2,rotation="promax")

rcovhat <- covrmvn(z)$cov #Outlier Resistant Method.
rout2 <- factanal(factors = 2, covmat=rcovhat)
rout2 #The program can make a correlation matrix
      #given a scaled covariance matrix.
Uniquenesses:
     len    nasal  bigonal cephalic     buxy
   0.412    0.884    0.999    0.005    0.005
Loadings:
         Factor1 Factor2
len      -0.760   0.102
nasal             0.338
bigonal
cephalic  0.997
buxy      0.154   0.986
               Factor1 Factor2
SS loadings      1.598   1.097
Proportion Var   0.320   0.219
Cumulative Var   0.320   0.539

The degrees of freedom for the model is 1 and the fit
was 0.0197.

#Robust Factor Analysis with RMVN Subset U
u <- getu(z)$U
rout3 <- factanal(u,factors = 2)
rout3
```

```
#The change in the output is the test statistic.

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 1.46 on 1 degree of
freedom. The p-value is 0.227.
```

Next the outliers are deleted, and the robust and classical maximum likelihood methods gave similar results. Note that the low loadings that differ for the two methods can likely be set to 0 when interpreting the factors. So Factor 1 loadings are roughly a difference of *length* and *cephalic*, while Factor 2 loadings are roughly *buxy* + 0.37 *nasal*. The robust method has about the same Factor 2 loadings for the clean data and the outlier data. For Factor 1 loadings, *length* and *cephalic* both have high loadings with opposite signs for the clean data and outlier data.

```
zc <- z[-c(61,62,63,64,65),] #delete outliers
uc <- getu(zc)$U #robust method on cleaned data
rout4 <- factanal(uc,factors=2); rout4
Uniquenesses:
     len    nasal  bigonal cephalic     buxy
   0.005    0.858    0.999    0.437    0.005
Loadings:Factor1 Factor2
len       0.997
nasal    -0.102   0.362
bigonal
cephalic -0.740   0.126
buxy              0.997
               Factor1 Factor2
SS loadings      1.552   1.144
Proportion Var   0.310   0.229
Cumulative Var   0.310   0.539
Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 2.36 on 1 degree of
freedom. The p-value is 0.124.
outc <- factanal(zc,factors=2)
outc #classical method on cleaned data
Uniquenesses:
     len    nasal  bigonal cephalic     buxy
   0.005    0.853    0.966    0.484    0.005
Loadings:Factor1 Factor2
len       0.995
nasal             0.383
bigonal   0.185
cephalic -0.718
buxy              0.995
```

```
                 Factor1 Factor2
SS loadings        1.545   1.143
Proportion Var     0.309   0.229
Cumulative Var     0.309   0.538
Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 8 on 1 degree of freedom.
The p-value is 0.00468.
```

Example 11.5. The output below is for a factor analysis of the Hebbler (1847) data from 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.

The maximum likelihood factor analysis was run using all of the data and the RMVN set U. Both analyses are very similar and suggest that Factor 1 is the average of the first three variables while Factor 2 is the average of the last two variables.

```
library(MASS)
out<-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

Factor Correlations:
        Factor1 Factor2
Factor1   1.000  -0.496
Factor2  -0.496   1.000

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 88.11 on 1 degree of
freedom. The p-value is 6.19e-21.
```

```
u<-getu(marry)$U   #Use RMVN set U.
outr<-factanal(u,factors=2,rotation="promax")
Uniquenesses:
    pop    mmen    mwmn mmilmen  milwmn
  0.011   0.005   0.005   0.005   0.005
Loadings:
        Factor1 Factor2
pop      1.005
mmen     0.994
mwmn     0.993
mmilmen          0.995
milwmn           0.988

               Factor1 Factor2
SS loadings      2.984   1.967
Proportion Var   0.597   0.393
Cumulative Var   0.597   0.990

Factor Correlations:
        Factor1 Factor2
Factor1   1.000  -0.427
Factor2  -0.427   1.000

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 41.17 on 1 degree of
freedom. The p-value is 1.39e-10.
#Now a very small pvalue seems good.
z <- scale(marry)
zu <- scale(u)
biplot(out$loadings[,1:2],z)
biplot(outr$loading[,1:2],zu)
```

Biplots, not shown, can also be useful. The last two R commands above can be used to make biplots. With the varimax rotation, $\hat{\boldsymbol{\Sigma}}_F \approx \boldsymbol{R}$, but this approximation is not good for the promax oblique rotation. See Figure 11.1, the DD plot using $\boldsymbol{R}$ and $\hat{\boldsymbol{\Sigma}}_F$ using varimax, made with the R commands below.

```
out<-factanal(marry,factors=2) #varimax is default
Lhat <- out$loadings[,1:2]
sigf <- Lhat
cor(marry)-sigf
center <- 0*1:dim(marry)[2]
cov <- cor(marry)
```

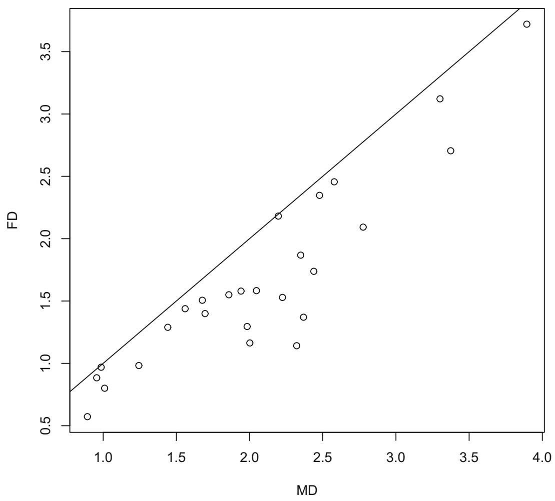

Fig. 11.1 DD Plot of $MD = D_i(\mathbf{0}, \boldsymbol{R})$ Versus $FD = D_i(\mathbf{0}, \hat{\boldsymbol{\Sigma}}_F)$

```
z<-scale(marry)
MD <- sqrt(mahalanobis(z, center, cov))
FD <- sqrt(mahalanobis(z, center, sigf))
plot(MD, FD)
abline(0, 1)
```

11.3 Summary

1) Factor analysis is used to get $\hat{\boldsymbol{\Sigma}} \approx \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}} = \hat{\boldsymbol{\Sigma}}_F$. Factor analysis clusters variables into groups called factors and suggests that the $m < p$ factors explain the dispersion more simply than $X_1, ..., X_p$. $\hat{\boldsymbol{L}} = [\hat{\boldsymbol{L}}_1, ..., \hat{\boldsymbol{L}}_m]$ 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, we 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. Perhaps a k factor model with $pval < 0.05$ is not sufficient: more factors are needed, while a k factor model with $pval > 0.05$ is sufficient.

7) Let $\hat{\boldsymbol{\Gamma}}$ be an orthogonal matrix. Then $\hat{\boldsymbol{L}}_{\Gamma}\hat{\boldsymbol{L}}_{\Gamma}^T = \hat{\boldsymbol{L}}\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Gamma}}^T\hat{\boldsymbol{L}}^T = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T$. The varimax and promax rotations seek $\hat{\boldsymbol{\Gamma}}$ such that $\hat{\boldsymbol{L}}^* \equiv \hat{\boldsymbol{L}}_{\Gamma} = \hat{\boldsymbol{L}}\hat{\boldsymbol{\Gamma}}$ 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

Brown et al. (2012) is a useful reference for factor analysis. Kosfeld (1996) did factor analysis with the DGK estimator. Pison et al. (2003) gave references for robust methods of factor analysis. The practical plug in estimators in the literature are not yet backed by theory and should be replaced by RMVN or RFCH.

Hastie et al. (2009, p. 560) noted that *independent component analysis* is approximately factor analysis with a rotation. Nordhausen and Tyler (2015) suggested that robust plug in estimators tend not to work for independent component analysis.

11.5 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

```
out <- factanal(factors = 2, covmat=Harman23.cor,
rotation="promax")
out                          (Output for 11.1.)
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 an *R* data set with a correlation matrix of eight physical measurements on 305 girls between ages seven and seventeen.

a) What is the cumulative variance explained by the two 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          (Output for 11.2.)
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 two factors?

b) Explain Factor 1.
c) Explain Factor 2.

```
Uniquenesses: (Output for Problem 11.3)
age  breadth cephalic  circum headht height len
0.005  0.005    0.005  0.142   0.005  0.303 0.005
size  cbrainy
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) data. The variables included *log*(*age*) and *height* and seven head measurements breadth, cephalic, circum, headht, len, size, and brain weight.

a) What is the cumulative variance explained by the four factors?

b) Which factor has a nonzero loading for log(age)?

c) Explain Factor 3.

R Problem

Note: For the following problem, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/mrsashw.txt) into *R*.

11.4. The Buxton data has five massive outliers in variables len and buxy = height.

a) The *R* commands for this part do a factor analysis on the Buxton data, likely using the sample correlation matrix obtained from 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, likely using a robust correlation 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

This chapter will show that multivariate linear regression with $m \geq 2$ response variables is nearly as easy to use, at least if m is small, as multiple linear regression which has 1 response variable. *For multivariate linear regression, at least one predictor variable is quantitative.* Plots for checking the model, including outlier detection, are given. Prediction regions that are robust to nonnormality are developed. For hypothesis testing, it is shown that the Wilks' lambda statistic, Hotelling Lawley trace statistic, and Pillai's trace statistic are robust to nonnormality.

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.

Definition 12.2. 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 $x_1, x_2, ..., x_p$ where $x_1 \equiv 1$ 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. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$ where the matrices are defined below. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\sigma_{ij})$ for $k = 1, ..., n$. Then the $p \times m$ coefficient matrix $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \ldots & \boldsymbol{\beta}_m \end{bmatrix}$ and the $m \times m$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{X}\boldsymbol{B}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j$. The $\boldsymbol{\epsilon}_i$ are assumed to be iid. Multiple linear regression corresponds to $m = 1$ response variable and is written in matrix form as

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_12

$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$. Subscripts are needed for the m multiple linear regression models $\boldsymbol{Y}_j = \boldsymbol{X\beta}_j + \boldsymbol{e}_j$ for $j = 1,...,m$ where $E(\boldsymbol{e}_j) = \boldsymbol{0}$. For the multivariate linear regression model, $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij} \quad \boldsymbol{I}_n$ for $i, j = 1,...,m$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

Notation. The **multiple linear regression model** uses $m = 1$. The **multivariate linear model** $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1,...,n$ has $m \geq 2$, and multivariate linear regression and MANOVA models are special cases. See Definition 10.2. This chapter will use $x_1 \equiv 1$ for the multivariate linear regression model. The **multivariate location and dispersion model** is the special case where $\boldsymbol{X} = \boldsymbol{1}$ and $p = 1$.

The data matrix $\boldsymbol{W} = [\boldsymbol{X} \quad \boldsymbol{Z}]$ except usually the first column $\boldsymbol{1}$ of $\boldsymbol{X}$ is omitted for software. 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 $n \times p$ design matrix of predictor variables is

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

where $\boldsymbol{v}_1 = \boldsymbol{1}$.

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

$$\boldsymbol{E} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \dots & \epsilon_{1,m} \\ \epsilon_{2,1} & \epsilon_{2,2} & \dots & \epsilon_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \dots & \epsilon_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 \ \boldsymbol{e}_2 \ \dots \ \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix}.$$

Considering the ith row of $\boldsymbol{Z}, \boldsymbol{X}$, and $\boldsymbol{E}$ shows that $\boldsymbol{y}_i^T = \boldsymbol{x}_i^T\boldsymbol{B} + \boldsymbol{\epsilon}_i^T$.

Warning: The $\boldsymbol{e}_i$ are error vectors, not orthonormal eigenvectors.

Definition 12.3. In the *multiple linear regression model*, $m = 1$ and

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i \tag{12.1}$$

for $i = 1, \ldots, n$. 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,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{12.3}$$

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 $\boldsymbol{X}$.

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 (e.g., ancient Egyptians or modern US citizens). The predictor variables might be $x_1 \equiv 1$, $x_2 =$ *femur length*, and $x_3 =$ *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 $x_4 =$ *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

$$\hat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z} = \left[\hat{\boldsymbol{\beta}}_1 \; \hat{\boldsymbol{\beta}}_2 \; \ldots \; \hat{\boldsymbol{\beta}}_m\right].$$

The *predicted values* or *fitted values*

$$\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \left[\hat{\boldsymbol{Y}}_1 \; \hat{\boldsymbol{Y}}_2 \; \ldots \; \hat{\boldsymbol{Y}}_m\right] = \begin{bmatrix} \hat{Y}_{1,1} & \hat{Y}_{1,2} & \ldots & \hat{Y}_{1,m} \\ \hat{Y}_{2,1} & \hat{Y}_{2,2} & \ldots & \hat{Y}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{Y}_{n,1} & \hat{Y}_{n,2} & \ldots & \hat{Y}_{n,m} \end{bmatrix}.$$

The *residuals* $\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} =$

$$\begin{bmatrix} \hat{\boldsymbol{\epsilon}}_1^T \\ \hat{\boldsymbol{\epsilon}}_2^T \\ \vdots \\ \hat{\boldsymbol{\epsilon}}_n^T \end{bmatrix} = \left[\boldsymbol{r}_1 \; \boldsymbol{r}_2 \; \ldots \; \boldsymbol{r}_m\right] = \begin{bmatrix} \hat{\epsilon}_{1,1} & \hat{\epsilon}_{1,2} & \ldots & \hat{\epsilon}_{1,m} \\ \hat{\epsilon}_{2,1} & \hat{\epsilon}_{2,2} & \ldots & \hat{\epsilon}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\epsilon}_{n,1} & \hat{\epsilon}_{n,2} & \ldots & \hat{\epsilon}_{n,m} \end{bmatrix}.$$

These quantities can be found from the m multiple linear regressions of $\boldsymbol{Y}_j$ on the predictors: $\hat{\boldsymbol{\beta}}_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_j$, $\hat{\boldsymbol{Y}}_j = \boldsymbol{X}\hat{\boldsymbol{\beta}}_j$, and $\boldsymbol{r}_j = \boldsymbol{Y}_j - \hat{\boldsymbol{Y}}_j$

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 $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},p}$ 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

$$\hat{\boldsymbol{E}} = [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}]\boldsymbol{Z}.$$

The following two theorems show that the least squares estimators are fairly good. Also see Theorem 12.7 in Section 12.4. Theorem 12.2 can also be used for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} = \dfrac{n-1}{n-d}\boldsymbol{S}_r$.

Theorem 12.1. Johnson and Wichern (1988, p. 304): Suppose $\boldsymbol{X}$ has full rank $p < n$ and the covariance structure of Definition 12.2 holds. Then $E(\hat{\boldsymbol{B}}) = \boldsymbol{B}$ so $E(\hat{\boldsymbol{\beta}}_j) = \boldsymbol{\beta}_j$, $\text{Cov}(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_k) = \sigma_{jk}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ 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. $\boldsymbol{S}_r = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + O_P(n^{-1/2})$ and $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + O_P$ $(n^{-1/2})$ if the following three conditions hold: $\boldsymbol{B} - \hat{\boldsymbol{B}} = O_P(n^{-1/2})$, $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{x}_i^T = O_P(1)$, and $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T = O_P(n^{1/2})$.

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

$$\boldsymbol{S}_r = \frac{n}{n-1} \frac{1}{n} \sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^T. \quad \square$$

$\boldsymbol{S}_r$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ are also $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ by Cook (2012, p. 692). See Theorem 12.7.

12.2 Plots for the Multivariate Linear Regression Model

As in Chapter 10, this section suggests using residual plots, response plots, and the DD plot to examine the multivariate linear model. The DD plot is used to examine the distribution of the iid error vectors. The residual plots are often used to check for lack of fit of the multivariate linear model. The response plots are used to check linearity and to detect influential cases for the linearity assumption. The response and residual plots are used exactly as in the $m = 1$ case corresponding to multiple linear regression and experimental design models. See Olive (2010, 2017a, c), Olive et al. (2015), Olive and Hawkins (2005), and Cook and Weisberg (1999a, p. 432; 1999b). Review Remark 10.2 which also applies to multivariate linear regression.

Notation. Plots will be used to simplify the 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 $r_{ij} = Y_{ij} - \hat{Y}_{ij}$. Suppose the model is good, the jth error distribution is unimodal and not highly skewed for $j = 1, ..., m$, and $n \geq 10p$. 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 transformed or corrected. If the model is good, then 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.

Rule of thumb 12.1. Use multivariate linear regression if

$$n \geq \max((m+p)^2, mp + 30, 10p))$$

provided that the m response and residual plots 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. The *mpack* function `MLRsim` simulates response and residual plots for various distributions when $m = 1$.

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 emphasize *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. Some applications for multivariate linear regression need the m error vectors to be linearly related, and larger sample sizes may be needed if the error vectors are not linearly related. For example, the asymptotic optimality of the prediction regions of Section 12.3 needs the error vectors to be iid from an elliptically contoured distribution. Make the RR plot and a DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ to check that the error vectors are linearly related. Make a DD plot of the continuous predictor variables to check for $\boldsymbol{x}$-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.

The RMVN DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ is used to check the error vector distribution, to detect outliers, and to display the nonparametric prediction region developed in Section 12.3. The DD plot suggests that the error vector distribution is elliptically contoured if the plotted points cluster tightly about a line through the origin as $n \to \infty$. The plot suggests that the error vector distribution is multivariate normal if the line is the identity line. If n is large and the plotted points do not cluster tightly about a line through the origin, then the error vector distribution may not be elliptically contoured. These applications of the DD plot for iid multivariate data are discussed in Olive (2002, 2008, 2013a) and Chapter 5. The RMVN estimator has not yet been proven to be a consistent estimator when computed from residual vectors, but simulations suggest that the RMVN DD plot of the residual vectors is a useful diagnostic plot. The *mpack* function `mregddsim` can be used to simulate the DD plots for various distributions.

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` with $m = 1$, 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 also make the response plot of $\hat{\boldsymbol{Y}}_j$ versus $\boldsymbol{Y}_j$, and use the rules of Section 2.4 on Y_j to linearize the response plot for each of the m response variables $Y_1, ..., Y_m$.

12.3 Asymptotically Optimal Prediction Regions

In this section, we will consider a more general multivariate regression model and then consider the multivariate linear model as a special case. Given n cases of training or past data $(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$ and a vector of predictors $\boldsymbol{x}_f$, suppose it is desired to predict a future test vector $\boldsymbol{y}_f$.

Definition 12.8. A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{y}_f \in \mathcal{A}_n) \to 1-\delta$ as $n \to \infty$ and is *asymptotically optimal* if the volume of the region converges in probability to the volume of the population minimum volume covering region.

The classical large sample $100(1-\delta)\%$ prediction region for a future value $\boldsymbol{x}_f$ given iid data $\boldsymbol{x}_1, ..., , \boldsymbol{x}_n$ is $\{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \le \chi^2_{p,1-\delta}\}$, while for multivariate linear regression, the classical large sample $100(1-\delta)\%$ prediction region for a future value $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ and past data $(\boldsymbol{x}_1, \boldsymbol{y}_i), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$ is $\{\boldsymbol{y} : D^2_{\boldsymbol{y}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le \chi^2_{m,1-\delta}\}$. See Wichern (1988, pp. 134, 151, 312). By Equation (3.10), these regions may work for multivariate normal $\boldsymbol{x}_i$ or $\boldsymbol{\epsilon}_i$, but otherwise tend to have undercoverage. Section 5.2 and Olive (2013a) replaced $\chi^2_{p,1-\delta}$ by the order statistic $D^2_{(U_n)}$ where U_n decreases to $\lceil n(1-\delta) \rceil$. This section will use a similar technique from Olive (2017b) to develop possibly the first practical large sample prediction region for the multivariate linear model with unknown error distribution. The following technical theorem will be needed to prove Theorem 12.4.

Theorem 12.3. Let $a > 0$ and assume that $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$.

a) $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) - \frac{1}{a} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = o_P(1)$.

b) Let $0 < \delta \le 0.5$. If $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) - (\boldsymbol{\mu}, a\boldsymbol{\Sigma}) = O_p(n^{-\delta})$ and $a\hat{\boldsymbol{\Sigma}}^{-1}_n - \boldsymbol{\Sigma}^{-1} = O_P(n^{-\delta})$, then

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) - \frac{1}{a} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = O_P(n^{-\delta}).$$

Proof. Let B_n denote the subset of the sample space on which $\hat{\boldsymbol{\Sigma}}_n$ has an inverse. Then $P(B_n) \to 1$ as $n \to \infty$. Now

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

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a} - \frac{\boldsymbol{\Sigma}^{-1}}{a} + \hat{\boldsymbol{\Sigma}}^{-1}_n \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n) =$$

$$(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_n)^T\left(\frac{-\boldsymbol{\Sigma}^{-1}}{a}+\hat{\boldsymbol{\Sigma}}_n^{-1}\right)(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_n)+(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_n)^T\left(\frac{\boldsymbol{\Sigma}^{-1}}{a}\right)(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_n)=$$

$$\frac{1}{a}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_n)^T(-\boldsymbol{\Sigma}^{-1}+a\ \hat{\boldsymbol{\Sigma}}_n^{-1})(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_n)\ +$$

$$(\boldsymbol{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_n)^T\left(\frac{\boldsymbol{\Sigma}^{-1}}{a}\right)(\boldsymbol{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_n)$$

$$=\frac{1}{a}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})+\frac{2}{a}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_n)+$$

$$\frac{1}{a}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_n)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_n)+\frac{1}{a}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_n)^T[a\hat{\boldsymbol{\Sigma}}_n^{-1}-\boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-\hat{\boldsymbol{\mu}}_n)$$

on B_n, and the last three terms are $o_P(1)$ under a) and $O_P(n^{-\delta})$ under b). □

Now suppose a prediction region for an $m \times 1$ random vector $\boldsymbol{y}_f$ given a vector of predictors $\boldsymbol{x}_f$ is desired for the multivariate linear model. If we had many cases $\boldsymbol{z}_i = \boldsymbol{B}^T\boldsymbol{x}_f + \boldsymbol{\epsilon}_i$, then we could use the multivariate prediction region for m variables from Section 5.2. Instead, Theorem 12.4 will use the prediction region from Section 5.2 on the pseudodata $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{B}}^T\boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ 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 $\hat{\boldsymbol{z}}_i = (\boldsymbol{B} - \boldsymbol{B} + \hat{\boldsymbol{B}})^T\boldsymbol{x}_f + (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i) = \boldsymbol{z}_i + (\hat{\boldsymbol{B}} - \boldsymbol{B})^T\boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i = \boldsymbol{z}_i + (\hat{\boldsymbol{B}} - \boldsymbol{B})^T\boldsymbol{x}_f - (\hat{\boldsymbol{B}} - \boldsymbol{B})^T\boldsymbol{x}_i = \boldsymbol{z}_i + O_P(n^{-1/2})$. Hence the distances based on the $\boldsymbol{z}_i$ and the distances based on the $\hat{\boldsymbol{z}}_i$ have the same quantiles, asymptotically (for quantiles that are continuity points of the distribution of $\boldsymbol{z}_i$).

If the $\boldsymbol{\epsilon}_i$ are iid from an $EC_m(\boldsymbol{0}, \boldsymbol{\Sigma}, g)$ distribution with continuous decreasing g and nonsingular covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = c\boldsymbol{\Sigma}$ for some constant $c > 0$, then the population asymptotically optimal prediction region is $\{\boldsymbol{y} : D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}\}$ where $P(D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}) = 1 - \delta$. For example, if the iid $\boldsymbol{\epsilon}_i \sim N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then $D_{1-\delta} = \sqrt{\chi^2_{m,1-\delta}}$. If the error distribution is not elliptically contoured, then the above region still has $100(1-\delta)\%$ coverage, but prediction regions with smaller volume may exist.

A natural way to make a large sample prediction region is to estimate the target population minimum volume covering region, but for moderate samples and many error distributions, the natural estimator that covers $\lceil n(1-\delta)\rceil$ of the cases tends to have undercoverage as high as $min(0.05, \delta/2)$. This empirical result is not too surprising since it is well known that the performance of a prediction region on the training data is superior to the performance on future test data, due in part to the unknown variability of the estimator. To compensate for the undercoverage, let q_n be as in Theorem 12.4.

Theorem 12.4. Suppose $\boldsymbol{y}_i = E(\boldsymbol{y}_i|\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i = \hat{\boldsymbol{y}}_i + \hat{\boldsymbol{\epsilon}}_i$ where $\text{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, and where the zero mean $\boldsymbol{\epsilon}_f$ and the $\boldsymbol{\epsilon}_i$ are iid for $i = 1, ..., n$. Given $\boldsymbol{x}_f$, 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

$$D_i^2 \equiv D_i^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$$

for $i = 1, ..., n$. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + m/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta m/n), \quad \text{otherwise.}$$

If $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let $0 < \delta < 1$ and $h = D_{(U_n)}$ where $D_{(U_n)}$ is the 100 q_nth sample quantile of the Mahalanobis distances D_i. Let the nominal $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$ be given by

$$\{\boldsymbol{z} : (\boldsymbol{z} - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{z} - \hat{\boldsymbol{y}}_f) \le D^2_{(U_n)}\} =$$

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D^2_{(U_n)}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{(U_n)}\}. \tag{12.4}$$

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 $U_n/n \to 1 - \delta$ 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 (12.4) is a large sample $100(1-\delta)\%$ 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 unique highest density region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}\}$, then the prediction region (12.4) 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 $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{(U_n)}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ iff $\hat{\boldsymbol{\epsilon}}_i$ is in prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{(U_n)}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$, 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 $100(1-\delta)$th percentile of the D_i asymptotically, $U_n/n \to 1 - \delta$.

b) Let $P[D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})] = 1 - \delta$. Since $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, Theorem 12.3 shows that if $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} (E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ then $D(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{D} D_{\boldsymbol{z}}$ $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Hence the percentiles of the distances converge in distribution, and the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ converges to $1 - \delta =$ the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le$

$D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$ at continuity points $D_{1-\delta}$ of the distribution of $D(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$.

c) The asymptotically optimal prediction region is the region with the smallest volume (hence highest density) such that the coverage is $1-\delta$, as $n \to \infty$. This region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$ if the asymptotically optimal region for the $\boldsymbol{\epsilon}_i$ is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$. Hence the result follows by b). □

Notice that if $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}$ exists, then $100q_n\%$ of the n training data $\boldsymbol{y}_i$ are in their corresponding prediction region with $\boldsymbol{x}_f = \boldsymbol{x}_i$, and $q_n \to 1-\delta$ even if $(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is not a good estimator or if the regression model is misspecified. Hence the coverage q_n of the training data is robust to model assumptions. Of course the volume of the prediction region could be large if a poor estimator $(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is used or if the $\boldsymbol{\epsilon}_i$ do not come from an elliptically contoured distribution. The response, residual, and DD plots can be used to check model assumptions. If the plotted points in the RMVN DD plot cluster tightly about some line through the origin and if $n \geq \max[3(m+p)^2, mp+30]$, we expect the volume of the prediction region may be fairly low for the least squares estimators.

If n is too small, then multivariate data is sparse and the covering ellipsoid for the training data may be far too small for future data, resulting in severe undercoverage. Also notice that $q_n = 1-\delta/2$ or $q_n = 1-\delta+0.05$ for $n \leq 20p$. At the training data, the coverage $q_n \geq 1-\delta$, and q_n converges to the nominal coverage $1-\delta$ as $n \to \infty$. Suppose $n \leq 20p$. Then the nominal 95% prediction region uses $q_n = 0.975$ while the nominal 50% prediction region uses $q_n = 0.55$. Prediction distributions depend both on the error distribution and on the variability of the estimator $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$. This variability is typically unknown but converges to 0 as $n \to \infty$. Also, residuals tend to underestimate errors for small n. For moderate n, ignoring estimator variability and using $q_n = 1-\delta$ resulted in undercoverage as high as $\min(0.05, \delta/2)$. Letting the "coverage" q_n decrease to the nominal coverage $1-\delta$ inflates the volume of the prediction region for small n, compensating for the unknown variability of $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$.

Consider the multivariate linear regression model. The semiparametric and parametric regions are only conjectured to be large sample prediction regions, but are useful as diagnostics. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=p}$, $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$, and $D_i^2(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T \boldsymbol{S}_r^{-1} (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$ for $i = 1, ..., n$. Then the large sample nonparametric $100(1-\delta)\%$ prediction region is

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}\}, \tag{12.5}$$

while the (Johnson and Wichern (1988, p. 312) classical large sample $100(1-\delta)\%$ prediction region is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le \chi^2_{m,1-\delta}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le \sqrt{\chi^2_{m,1-\delta}}\}. \qquad (12.6)$$

Theorem 12.5 will show that this prediction region (12.5) can also be found by applying the nonparametric prediction region (5.17) on the $\hat{\boldsymbol{z}}_i$. Recall that $\boldsymbol{S}_r$ defined in Definition 12.4 is the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$. Section 5.2 describes the nonparametric, semiparametric, and parametric MVN prediction regions. Similar regions are used for multivariate linear regression. For the multivariate linear regression model, if $D_{1-\delta}$ is a continuity point of the distribution of D, Assumption D1 above Theorem 12.7 holds, and the $\boldsymbol{\epsilon}_i$ have a nonsingular covariance matrix, then (12.5) is a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$.

Theorem 12.5. For multivariate linear regression, when least squares is used to compute $\hat{\boldsymbol{y}}_f$, $\boldsymbol{S}_r$, and the pseudodata $\hat{\boldsymbol{z}}_i$, prediction region (12.5) is the Section 5.2 nonparametric prediction region applied to the $\hat{\boldsymbol{z}}_i$.

Proof. Multivariate linear regression with least squares satisfies Theorem 12.4 by Su and Cook (2012). (See Theorem 12.7.) Let $(T, \boldsymbol{C})$ be the sample mean and sample covariance matrix (see Definition 2.6) applied to the $\hat{\boldsymbol{z}}_i$. The sample mean and sample covariance matrix of the residual vectors is $(\boldsymbol{0}, \boldsymbol{S}_r)$ 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 $D_{(U_n)}$. □

The RMVN DD plot of the residual vectors will be used to display the prediction regions for multivariate linear regression. See Example 12.3. The nonparametric prediction region for multivariate linear regression of Theorem 12.5 uses $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ in (12.4) and has simple geometry. Let R_r be the nonparametric prediction region (12.5) applied to the residuals $\hat{\boldsymbol{\epsilon}}_i$ with $\hat{\boldsymbol{y}}_f = \boldsymbol{0}$. Then R_r is a hyperellipsoid with center $\boldsymbol{0}$, and the nonparametric prediction region is the hyperellipsoid R_r translated to have center $\hat{\boldsymbol{y}}_f$. Hence in a DD plot, all points to the left of the line $MD = D_{(U_n)}$ correspond to $\boldsymbol{y}_i$ that are in their prediction region, while points to the right of the line are not in their prediction region.

The nonparametric prediction region has some interesting properties. This prediction region is asymptotically optimal if the $\boldsymbol{\epsilon}_i$ are iid for a large class of elliptically contoured $EC_m(\boldsymbol{0}, \boldsymbol{\Sigma}, g)$ distributions. Also, if there are 100 different values $(\boldsymbol{x}_{jf}, \boldsymbol{y}_{jf})$ to be predicted, we only need to update $\hat{\boldsymbol{y}}_{jf}$ for $j = 1, ..., 100$, we do not need to update the covariance matrix $\boldsymbol{S}_r$.

It is common practice to examine how well the prediction regions work on the training 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 $1-\delta$ 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 (12.5) for the data since the prediction region (12.5) 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 $\hat{\boldsymbol{\epsilon}}_i = \boldsymbol{y}_i - \hat{\boldsymbol{y}}_i$. Thus $100q_n\%$ of prediction regions corresponding to the data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ contain $\boldsymbol{y}_i$, and $100q_n\% \to 100(1-\delta)\%$. Hence the prediction regions work well on the training data and should work well on $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ similar to the training data. Of course simulation should be done for $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ that are not equal to training data cases. See Section 12.5.

This training data 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 training data cases $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is 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 for a future value $\boldsymbol{y}_f$ 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.

12.4 Testing Hypotheses

This section considers testing a linear hypothesis $H_0 : \boldsymbol{LB} = \mathbf{0}$ versus
$H_1 : \boldsymbol{LB} \neq \mathbf{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix.

Definition 12.9. Assume rank$(\boldsymbol{X}) = p$. The *total corrected (for the mean) sum of squares and cross products matrix* is

$$\boldsymbol{T} = \boldsymbol{R} + \boldsymbol{W}_e = \boldsymbol{Z}^T \left(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T \right) \boldsymbol{Z}.$$

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, e.g., if $\boldsymbol{B} = \mathbf{0}$. The *regression sum of squares and cross products matrix* is

$$\boldsymbol{R} = \boldsymbol{Z}^T \left[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T \right] \boldsymbol{Z} = \boldsymbol{Z}^T\boldsymbol{X}\hat{\boldsymbol{B}} - \frac{1}{n}\boldsymbol{Z}^T\mathbf{1}\mathbf{1}^T\boldsymbol{Z}.$$

Let $\boldsymbol{H} = \hat{\boldsymbol{B}}^T \boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}$. The *error or residual sum of squares and cross products matrix* is

$$\boldsymbol{W}_e = (\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T(\boldsymbol{Z} - \hat{\boldsymbol{Z}}) = \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{Z}^T[\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Z}.$$

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)	$\boldsymbol{T}$	$n-1$

Definition 12.10. 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 *Roy's maximum root statistic* is $\lambda_{max}(\boldsymbol{L}) = \lambda_1$.

The *Wilks' 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 $U(\boldsymbol{L}) = tr[\boldsymbol{W}_e^{-1}\boldsymbol{H}] = \sum_{i=1}^{m} \lambda_i$.

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 last three test statistics do not lead to the same conclusions (Roy's test may not be trustworthy for $r > 1$). Theory and simulations developed below for the four statistics will provide more information about the sample sizes needed to use the four test statistics. See the following page for notation used in the next theorem.

Theorem 12.6. *The Hotelling-Lawley trace statistic*

$$U(\boldsymbol{L}) = \frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]. \quad (12.7)$$

Proof. Using the Searle (1982, p. 333) identity $tr(\boldsymbol{A}\boldsymbol{G}^T\boldsymbol{D}\boldsymbol{G}\boldsymbol{C}) = [vec(\boldsymbol{G})]^T[\boldsymbol{C}\boldsymbol{A} \otimes \boldsymbol{D}^T][vec(\boldsymbol{G})]$, it follows that $(n-p)U(\boldsymbol{L}) = tr[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}]$ $= [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] = T$ where $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}$, $\boldsymbol{G} = \boldsymbol{L}\hat{\boldsymbol{B}}, \boldsymbol{D} = [\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}$, and $\boldsymbol{C} = \boldsymbol{I}$. Hence (12.7) holds. □

Some notation is useful to show (12.7) and to show that $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$ under mild conditions if H_0 is true. Following Henderson and Searle (1979), let matrix $\boldsymbol{A} = [\boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \ldots \quad \boldsymbol{a}_p]$. 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

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2n}\boldsymbol{B} \\ \vdots & \vdots & \ldots & \vdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix}.$$

An important fact is that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonsingular square matrices, then $[\boldsymbol{A} \otimes \boldsymbol{B}]^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$. The following assumption is important.

Assumption D1: Let h_i be the ith diagonal element of $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Assume $\max_{1 \leq i \leq n} h_i \xrightarrow{P} 0$ as $n \to \infty$, assume that the zero mean iid error vectors have finite fourth moments, and assume that $\dfrac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \xrightarrow{P} \boldsymbol{W}^{-1}$.

Su and Cook (2012) proved a central limit type theorem for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{B}}$ for the partial envelopes estimator, and the least squares estimator is a special case. These results prove the following theorem. Their theorem also shows that for multiple linear regression $(m = 1)$, $\hat{\sigma}^2 = MSE$ is a $\sqrt{n}$ consistent estimator of σ^2.

Theorem 12.7. Multivariate Least Squares Central Limit Theorem (MLS CLT). For the least squares estimator, if assumption D1 holds, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ and

$$\sqrt{n} \;\; vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W}).$$

Theorem 12.8. If assumption D1 holds and if H_0 is true, then $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

Proof. By Theorem 12.7, $\sqrt{n} \;\; vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$. Then under H_0, $\sqrt{n} \;\; vec(\boldsymbol{L}\hat{\boldsymbol{B}}) \xrightarrow{D} N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)$, and $n \;\; [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes$

$(\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}$. This result also holds if $\boldsymbol{W}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are replaced by $\hat{\boldsymbol{W}} = n(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Hence under H_0 and using the proof of Theorem 12.6,

$$T = (n-p)U(\boldsymbol{L}) = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}.$$

□

Some more details on the above results may be useful. Consider testing a linear hypothesis $H_0 : \boldsymbol{L}\boldsymbol{B} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{L}\boldsymbol{B} \neq \boldsymbol{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix. For now assume the error distribution is multivariate normal $N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Then

$$vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m \end{pmatrix} \sim N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

where

$$\boldsymbol{C} = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} \sigma_{11}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{12}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{1m}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \sigma_{21}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{22}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{2m}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{m1}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{m2}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{mm}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{bmatrix}.$$

Now let $\boldsymbol{A}$ be an $rm \times pm$ block diagonal matrix: $\boldsymbol{A} = diag(\boldsymbol{L}, ..., \boldsymbol{L})$. Then $\boldsymbol{A}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = vec(\boldsymbol{L}(\hat{\boldsymbol{B}} - \boldsymbol{B})) =$

$$\begin{pmatrix} \boldsymbol{L}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \\ \boldsymbol{L}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \\ \vdots \\ \boldsymbol{L}(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m) \end{pmatrix} \sim N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)$$

where $\boldsymbol{D} = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T = \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^T =$

$$\begin{bmatrix} \sigma_{11}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{12}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{1m}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \\ \sigma_{21}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{22}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{2m}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{m1}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{m2}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{mm}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \end{bmatrix}.$$

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

$$T = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}. \tag{12.8}$$

A large sample level δ test will reject H_0 if $pval \leq \delta$ where

$$pval = P\left(\frac{T}{rm} < F_{rm,n-mp}\right). \tag{12.9}$$

Since least squares estimators are asymptotically normal, if the $\boldsymbol{\epsilon}_i$ are iid 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

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \xrightarrow{P} \boldsymbol{W}^{-1}.$$

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 (12.8) holds, and (12.9) gives a large sample level δ test if the least squares estimators are asymptotically normal.

Kakizawa (2009) showed, under stronger assumptions than Theorem 12.8, that for a large class of iid error distributions, the following test statistics have the same χ^2_{rm} limiting distribution when H_0 is true, and the same noncentral $\chi^2_{rm}(\omega^2)$ limiting distribution with noncentrality parameter ω^2 when H_0 is false under a local alternative. Hence the three tests are robust to the assumption of normality. The limiting null distribution is well known when the zero mean errors are iid from a multivariate normal distribution. See Khattree and Naik (1999, p. 68): $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, $(n-p)V(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, and $-[n-p-0.5(m-r+3)]\log(\Lambda(\boldsymbol{L})) \xrightarrow{D} \chi^2_{rm}$. Results from Kshirsagar (1972, p. 301) suggest that the third chi-square approximation is very good if $n \geq 3(m+p)^2$ for multivariate normal error vectors.

Theorems 12.6 and 12.8 are useful for relating multivariate tests with the partial F test for multiple linear regression that tests whether a reduced model that omits some of the predictors can be used instead of the full model that uses all p predictors. The partial F test statistic is

$$F_R = \left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F)$$

where the residual sums of squares $SSE(F)$ and $SSE(R)$ and degrees of freedom df_F and df_r are for the full and reduced model while the mean square error $MSE(F)$ is for the full model. Let the null hypothesis for the partial F test be $H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$ where $\boldsymbol{L}$ sets the coefficients of the predictors in the full model but not in the reduced model to 0. Seber and Lee (2003, p. 100) show that

$$F_R = \frac{[\boldsymbol{L}\hat{\boldsymbol{\beta}}]^T(\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}[\boldsymbol{L}\hat{\boldsymbol{\beta}}]}{r\hat{\sigma}^2}$$

is distributed as $F_{r,n-p}$ if H_0 is true and the errors are iid $N(0,\sigma^2)$. Note that for multiple linear regression with $m = 1$, $F_R = (n-p)U(\boldsymbol{L})/r$ since $\hat{\boldsymbol{\Sigma}}_\epsilon^{-1} = 1/\hat{\sigma}^2$. Hence the scaled Hotelling Lawley test statistic is the partial F test statistic extended to $m > 1$ predictor variables by Theorem 12.6.

By Theorem 12.8, for example, $rF_R \xrightarrow{D} \chi^2_r$ for a large class of nonnormal error distributions. If $Z_n \sim F_{k,d_n}$, then $Z_n \xrightarrow{D} \chi^2_k/k$ as $d_n \to \infty$. Hence using the $F_{r,n-p}$ approximation gives a large sample test with correct asymptotic level, and the partial F test is robust to nonnormality.

Similarly, using an $F_{rm,n-pm}$ approximation for the following test statistics gives large sample tests with correct asymptotic level by Kakizawa (2009) and similar power for large n. The large sample test will have correct asymptotic level as long as the denominator degrees of freedom $d_n \to \infty$ as $n \to \infty$, and $d_n = n - pm$ reduces to the partial F test if $m = 1$ and $U(\boldsymbol{L})$ is used. Then the three test statistics are

$$\frac{-[n-p-0.5(m-r+3)]}{rm} \ \log(\Lambda(\boldsymbol{L})), \ \ \frac{n-p}{rm} \ V(\boldsymbol{L}), \text{ and } \frac{\mathrm{n-p}}{\mathrm{rm}} \ \mathrm{U}(\boldsymbol{L}).$$

By Berndt and Savin (1977) and Anderson (1984, pp. 333, 371),

$$V(\boldsymbol{L}) \leq -\log(\Lambda(\boldsymbol{L})) \leq U(\boldsymbol{L}).$$

Hence the Hotelling Lawley test will have the most power, and Pillai's test will have the least power.

Following Khattree and Naik (1999, pp. 67–68), there are several approximations used by the SAS software. For the Roy's largest root test, if $h = \max(r, m)$, use

$$\frac{n-p-h+r}{h} \lambda_{max}(\boldsymbol{L}) \approx F(h, n-p-h+r).$$

The simulations in Section 12.5 suggest that this approximation is good for $r = 1$ but poor for $r > 1$. Anderson (1984, p. 333) stated that Roy's largest root test has the greatest power if $r = 1$ but is an inferior test for $r > 1$. Let $g = n - p - (m - r + 1)/2$, $u = (rm - 2)/4$ and $t = \sqrt{r^2m^2 - 4}/\sqrt{m^2 + r^2 - 5}$ for $m^2 + r^2 - 5 > 0$ and $t = 1$, otherwise. Assume H_0 is true. Thus $U \xrightarrow{P} 0, V \xrightarrow{P} 0$, and $\Lambda \xrightarrow{P} 1$ as $n \to \infty$. Then

$$\frac{gt-2u}{rm} \ \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \approx F(rm, gt-2u) \ \text{ or } \ (n-p)t \ \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \approx \chi^2_{rm}.$$

For large n and $t > 0$, $-\log(\Lambda) = -t\log(\Lambda^{1/t}) = -t\log(1 + \Lambda^{1/t} - 1) \approx t(1 - \Lambda^{1/t}) \approx t(1 - \Lambda^{1/t})/\Lambda^{1/t}$. If it cannot be shown that

$$(n-p)[-\log(\Lambda) - t(1-\Lambda^{1/t})/\Lambda^{1/t}] \xrightarrow{P} 0 \ \text{ as } \ \mathrm{n} \to \infty,$$

then it is possible that the approximate χ^2_{rm} distribution may be the limiting distribution for only a small class of iid error distributions. When the $\boldsymbol{\epsilon}_i$ are iid $N_m(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon})$, there are some exact results. For $r = 1$,

$$\frac{n-p-m+1}{m} \ \frac{1-\Lambda}{\Lambda} \sim F(m, n-p-m+1).$$

For $r = 2$,

$$\frac{2(n-p-m+1)}{2m} \; \frac{1-\Lambda^{1/2}}{\Lambda^{1/2}} \sim F(2m, 2(n-p-m+1)).$$

For $m = 2$,

$$\frac{2(n-p)}{2r} \; \frac{1-\Lambda^{1/2}}{\Lambda^{1/2}} \sim F(2r, 2(n-p)).$$

Let $s = \min(r, m)$, $m_1 = (|r-m|-1)/2$, and $m_2 = (n-p-m-1)/2$. Note that $s(|r-m|+s) = \min(r,m)\max(r,m) = rm$. Then

$$\frac{n-p}{rm} \; \frac{V}{1-V/s} = \frac{n-p}{s(|r-m|+s)} \; \frac{V}{1-V/s} \approx \frac{2m_2+s+1}{2m_1+s+1} \; \frac{V}{s-V} \approx$$

$$F(s(2m_1+s+1), s(2m_2+s+1)) \approx F(s(|r-m|+s), s(n-p)) = F(rm, s(n-p)).$$

This approximation is asymptotically correct by Slutsky's theorem since $1 - V/s \xrightarrow{P} 1$. Finally, $\frac{n-p}{rm}U =$

$$\frac{n-p}{s(|r-m|+s)}U \approx \frac{2(sm_2+1)}{s^2(2m_1+s+1)}U \approx F(s(2m_1+s+1), 2(sm_2+1))$$

$$\approx F(s(|r-m|+s), s(n-p)) = F(rm, s(n-p)).$$

This approximation is asymptotically correct for a wide range of iid error distributions.

Multivariate analogs of tests for multiple linear regression can be derived with appropriate choice of $\boldsymbol{L}$. Assume a constant $x_1 = 1$ is in the model. As a textbook convention, use $\delta = 0.05$ if δ is not given.

The 4 step MANOVA test of linear hypotheses is useful.

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 $\leq \delta$, reject H_0 and conclude that $\boldsymbol{LB} \neq \boldsymbol{0}$. If pval $> \delta$, fail to reject H_0 and conclude that $\boldsymbol{LB} = \boldsymbol{0}$ or that there is not enough evidence to conclude that $\boldsymbol{LB} \neq \boldsymbol{0}$.

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 $\boldsymbol{H} = \hat{\boldsymbol{B}}^T \boldsymbol{X}^T \boldsymbol{X} \hat{\boldsymbol{B}} = \hat{\boldsymbol{Z}}^T \hat{\boldsymbol{Z}}$, but is usually not a test of interest.

The analog of the ANOVA F test for multiple linear regression is the MANOVA F test that uses $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_{p-1}]$ to test whether the nontrivial predictors are needed in the model. This 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. Response and residual plots are often useful for $n \geq 10p$.

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_0 from output.
iii) Find the pval from output.
iv) If pval $\leq \delta$, reject H_0. If pval $> \delta$, 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. (Or there is not enough evidence to conclude that there is a mreg relationship between the response variables and the predictors. Get the variable names from the story problem.)

The F_j test of hypotheses uses $\boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$, where the 1 is in the jth position, to test whether the jth predictor x_j is needed in the model given that the other $p - 1$ predictors are in the model. This test is an analog of the t tests for multiple linear regression. Note that x_j is not needed in the model corresponds to $H_0 : \boldsymbol{B}_j = \boldsymbol{0}$ while x_j needed in the model corresponds to $H_1 : \boldsymbol{B}_j \neq \boldsymbol{0}$ where $\boldsymbol{B}_j^T$ is the jth row of $\boldsymbol{B}$.

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.
i) State the hypotheses $H_0 :\; x_j$ is not needed in the model $H_1 :\; x_j$ is needed.
ii) Find the test statistic F_j from output.
iii) Find pval from output.
iv) If pval $\leq \delta$, reject H_0. If pval $> \delta$, 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. (Or there is not enough evidence to conclude that x_j is needed in the model. Get the variable names from the story problem.)

The Hotelling Lawley statistic

$$F_j = \frac{1}{d_j} \hat{\boldsymbol{B}}_j^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \hat{\boldsymbol{B}}_j = \frac{1}{d_j} (\hat{\beta}_{j1}, \hat{\beta}_{j2}, ..., \hat{\beta}_{jm}) \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \begin{pmatrix} \hat{\beta}_{j1} \\ \hat{\beta}_{j2} \\ \vdots \\ \hat{\beta}_{jm} \end{pmatrix}$$

where $\hat{\boldsymbol{B}}_j^T$ is the jth row of $\hat{\boldsymbol{B}}$ and $d_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}$, the jth diagonal entry of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. The statistic F_j could be used for forward selection and backward elimination in variable selection.

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 $x_2, \ldots, x_p$ corresponds to the MANOVA 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 predictors are in the model.

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 $\leq \delta$, reject H_0 and conclude that the full model should be used. If pval $> \delta$, fail to reject H_0 and conclude that the reduced model is good.

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 $= 0.06$). Right click `Stop` on the plots m times to advance the plots and to get the cursor back on the command line in *R*.

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


#Output for Example 12.2
y<-marry[,c(2,3)]; x<-marry[,-c(2,3)];
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 1843 Prussia census. Sometimes if the wife or husband was not at the household, then she/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, and x_4 = number of women married to military men in the district. The reduced model deletes x_3 and x_4. The constant uses $x_1 = 1$.

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
```

Solution:

a) i) H_0: the nontrivial predictors are not needed in the mreg model
H_1: at least one of the nontrivial predictors is needed
ii) $F_0 = 295.071$
iii) pval = 0
iv) Reject H_0, the nontrivial predictors are needed in the mreg model.
b) i) H_0: x_2 is not needed in the model H_1: x_2 is needed
ii) $F_2 = 600.57$
iii) pval = 0
iv) Reject H_0, *population of the district* is needed in the model.
c) i) H_0: x_4 is not needed in the model H_1: x_4 is needed
ii) $F_4 = 0.065$
iii) pval = 0.937
iv) Fail to reject H_0, *number of women married to military men* is not needed in the model given that the other predictors are in the model.
d) i) H_0: The reduced model is good H_1: Use the full model.
ii) $F_R = 0.200$
iii) pval = 0.935
iv) Fail to reject H_0, so the reduced model is good.

e) i) H_0: The reduced model is good H_1: Use the full model.
ii) $F_R = 569.6$
iii) pval = 0.00
iv) Reject H_0, so use the full model.

12.5 An Example and Simulations

The semiparametric prediction region and parametric MVN prediction region from Section 5.2 applied to the $\hat{\boldsymbol{z}}_i$ are only conjectured to be large sample prediction regions, but are added to the DD plot as visual aids. Cases below the horizontal line that crosses the identity line correspond to the semiparametric region, while cases below the horizontal line that ends at the identity line correspond to the parametric MVN region. A vertical line dropped down from this point of intersection does correspond to a large sample prediction region for multivariate normal error vectors. Note that $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$, and adding a constant $\hat{\boldsymbol{y}}_f$ to all of the residual vectors does not change the Mahalanobis distances, so the DD plot of the residual vectors can be used to display the prediction regions.

Example 12.3. Cook and Weisberg (1999a, pp. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. Let $Y_1 = \log(S)$ 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) First use the multivariate location and dispersion model for this data. Figure 12.1 shows a scatterplot matrix of the data, and Figure 12.2 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.

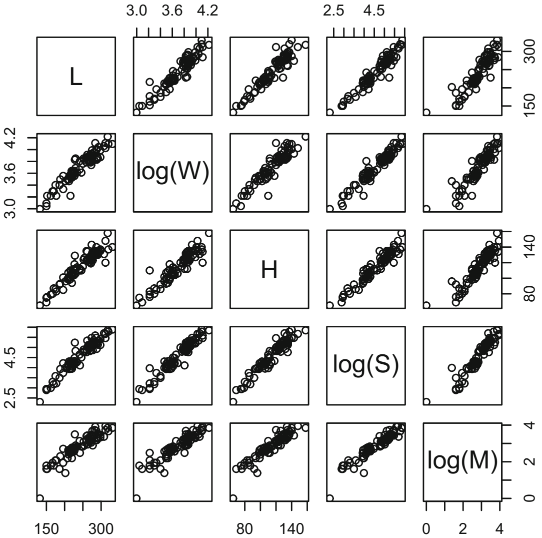

Fig. 12.1 Scatterplot Matrix of the Mussels Data

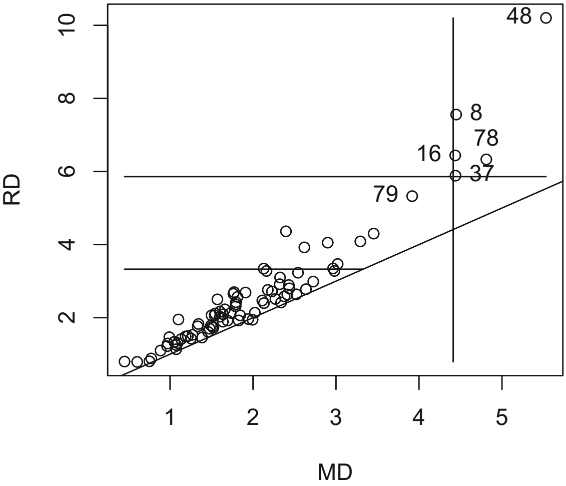

Fig. 12.2 DD Plot of the Mussels Data, MLD Model

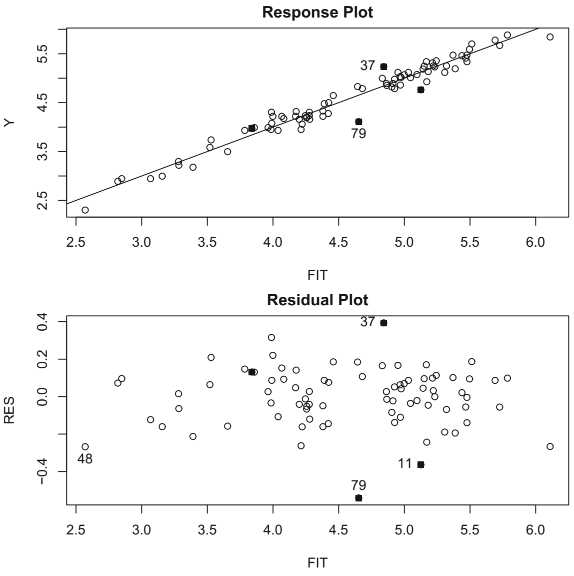

Fig. 12.3 Plots for $Y_1 = \log(S)$

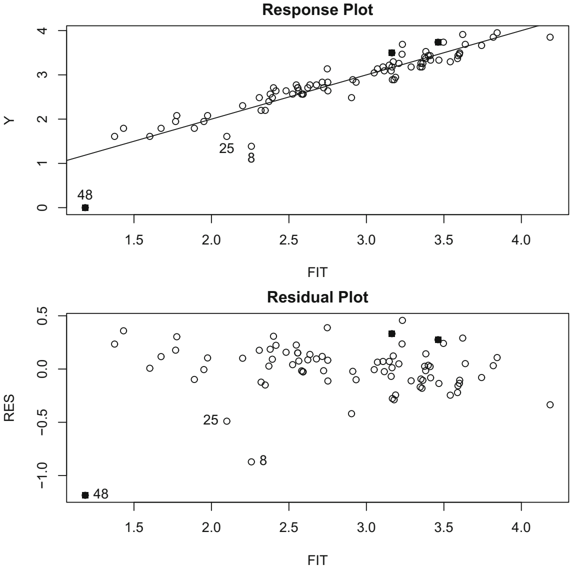

Fig. 12.4 Plots for $Y_2 = \log(M)$

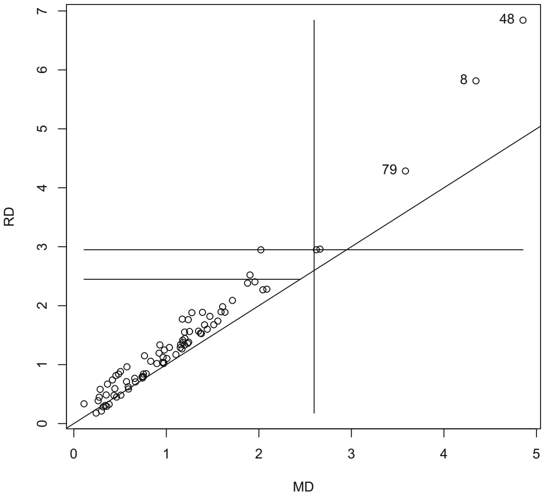

Fig. 12.5 DD Plot of the Residual Vectors for the Mussels Data

b) Now consider the multivariate linear regression model. To check linearity, Figures 12.3 and 12.4 give the response and residual plots for Y_1 and Y_2. The response plots show strong linear relationships. For Y_1, case 79 sticks out while for Y_2, cases 8, 25, and 48 are not fit well. Highlighted cases had Cook's distance $> \min(0.5, 2p/n)$. See Cook (1977).

To check the error vector distribution, the DD plot should be used instead of univariate residual plots, which do not take into account the correlations of the random variables $\epsilon_1, ..., \epsilon_m$ in the error vector $\boldsymbol{\epsilon}$. A residual vector $\hat{\boldsymbol{\epsilon}} = (\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}) + \boldsymbol{\epsilon}$ is a combination of $\boldsymbol{\epsilon}$ and a discrepancy $\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}$ that tends to have an approximate multivariate normal distribution. The $\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}$ term can dominate for small to moderate n when $\boldsymbol{\epsilon}$ is not multivariate normal, incorrectly suggesting that the distribution of the error vector $\boldsymbol{\epsilon}$ is closer to a multivariate normal distribution than is actually the case. Figure 12.5 shows the DD plot of the residual vectors. The plotted points are highly correlated but do not cover the identity line, suggesting an elliptically contoured error distribution that is not multivariate normal. The nonparametric 90% prediction region for the residuals consists of the points to the left of the vertical line $MD = 2.60$. Comparing Figures 12.2 and 12.5, the residual distribution is closer to a multivariate normal distribution. Cases 8, 48, and 79 have especially large distances.

The four Hotelling Lawley F_j statistics were greater than 5.77 with pvalues less than 0.005, and the MANOVA F statistic was 337.8 with pvalue ≈ 0.

The response, residual, and DD plots are effective for finding influential cases, for checking linearity, for checking whether the error distribution is multivariate normal or some other elliptically contoured distribution, and for displaying the nonparametric prediction region. Note that cases to the right of the vertical line correspond to cases with $\boldsymbol{y}_i$ that are not in their prediction region. These are the cases corresponding to residual vectors with large Mahalanobis distances. Adding a constant does not change the distance, so the DD plot for the residual vectors is the same as the DD plot for the $\hat{\boldsymbol{z}}_i$.

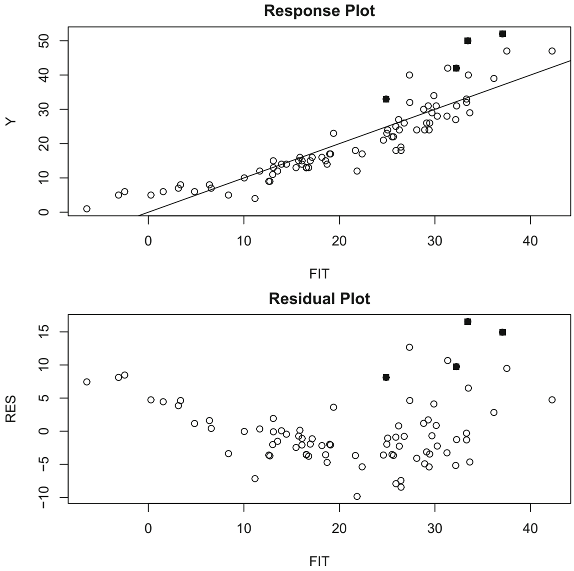

Fig. 12.6 Plots for $Y_2 = M$

c) Now suppose the same model is used except $Y_2 = M$. Then the response and residual plots for Y_1 remain the same, but the plots shown in Figure 12.6 show curvature about the identity and $r = 0$ lines. Hence the linearity condition is violated. Figure 12.7 shows that the plotted points in the DD plot have correlation well less than one, suggesting that the error vector distribution is no longer elliptically contoured. The nonparametric 90% prediction region for the residual vectors consists of the points to the left of the vertical line $MD = 2.52$ and contains 95% of the training data. Note that the plots can be used to quickly assess whether power transformations have resulted in a linear model, and whether influential cases are present. *R* code for producing the seven figures is shown below.

```
y <- log(mussels)[,4:5]
x <- mussels[,1:3]
```

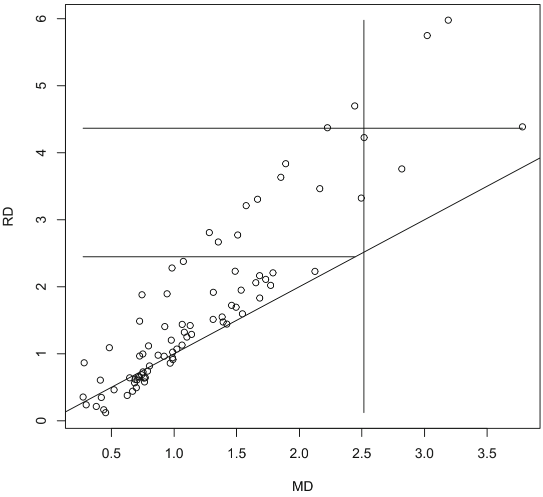

Fig. 12.7 DD Plot When $Y_2 = M$

```
x[,2] <- log(x[,2])
z<-cbind(x,y)
pairs(z, labels=c("L","log(W)","H","log(S)","log(M)"))
ddplot4(z)  #right click Stop
out <- mltreg(x,y) #right click Stop 4 times
ddplot4(out$res) #right click Stop
y[,2] <- mussels[,5]
tem <- mltreg(x,y) #right click Stop 4 times
ddplot4(tem$res) #right click Stop
```

12.5.1 Simulations for Testing

A small simulation was used to study the Wilks' Λ test, the Pillai's trace test, the Hotelling Lawley trace test, and the Roy's largest root test for the F_j tests and the MANOVA F test for multivariate linear regression. The first row of $\boldsymbol{B}$ was always $\mathbf{1}^T$, and the last row of $\boldsymbol{B}$ was always $\mathbf{0}^T$. When the null hypothesis for the MANOVA F test is true, all but the first row corresponding to the constant are equal to $\mathbf{0}^T$. When $p \geq 3$ and the null hypothesis for the MANOVA F test is false, then the second to last row of $\boldsymbol{B}$ is (1, 0, ..., 0), the third to last row is (1, 1, 0, ..., 0) etc., as long as the first row is not changed from $\mathbf{1}^T$. First, $m \times 1$ error vectors $\boldsymbol{w}_i$ were generated such that the m random variables in the vector $\boldsymbol{w}_i$ are iid with variance σ^2. Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and

$a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = \sigma^2[1 + (m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = \sigma^2[2\psi + (m-2)\psi^2]$ where $\psi = 0.10$. Hence the correlations are $(2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$. As ψ gets close to 1, the error vectors cluster about the line in the direction of $(1, ..., 1)^T$. We used $\boldsymbol{w}_i \sim N_m(\boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{w}_i \sim (1-\tau)N_m(\boldsymbol{0}, \boldsymbol{I}) + \tau N_m(\boldsymbol{0}, 25\boldsymbol{I})$ with $0 < \tau < 1$ and $\tau = 0.25$ in the simulation, $\boldsymbol{w}_i \sim$ multivariate t_d with $d = 7$ degrees of freedom, or $\boldsymbol{w}_i \sim$ lognormal - E(lognormal): where the m components of $\boldsymbol{w}_i$ were iid with distribution $e^z - E(e^z)$ where $z \sim N(0,1)$. Only the lognormal distribution is not elliptically contoured.

Table 12.1 Test Coverages: MANOVA F H_0 is True.

$\boldsymbol{w}$ dist	n	test	F_1	F_2	F_{p-1}	F_p	F_M
MVN	300	W	1	0.043	0.042	0.041	0.018
MVN	300	P	1	0.040	0.038	0.038	0.007
MVN	300	HL	1	0.059	0.058	0.057	0.045
MVN	300	R	1	0.051	0.049	0.048	0.993
MVN	600	W	1	0.048	0.043	0.043	0.034
MVN	600	P	1	0.046	0.042	0.041	0.026
MVN	600	HL	1	0.055	0.052	0.050	0.052
MVN	600	R	1	0.052	0.048	0.047	0.994
MIX	300	W	1	0.042	0.043	0.044	0.017
MIX	300	P	1	0.039	0.040	0.042	0.008
MIX	300	HL	1	0.057	0.059	0.058	0.039
MIX	300	R	1	0.050	0.050	0.051	0.993
MVT(7)	300	W	1	0.048	0.036	0.045	0.020
MVT(7)	300	P	1	0.046	0.032	0.042	0.011
MVT(7)	300	HL	1	0.064	0.049	0.058	0.045
MVT(7)	300	R	1	0.055	0.043	0.051	0.993
LN	300	W	1	0.043	0.047	0.040	0.020
LN	300	P	1	0.039	0.045	0.037	0.009
LN	300	HL	1	0.057	0.061	0.058	0.041
LN	300	R	1	0.049	0.055	0.050	0.994

The simulation used 5000 runs, and H_0 was rejected if the F statistic was greater than $F_{d_1,d_2}(0.95)$ where $P(F_{d_1,d_2} < F_{d_1,d_2}(0.95)) = 0.95$ with $d_1 = rm$ and $d_2 = n - mp$ for the test statistics

$$\frac{-[n-p-0.5(m-r+3)]}{rm} \log(\Lambda(\boldsymbol{L})), \quad \frac{n-p}{rm} V(\boldsymbol{L}), \text{ and } \frac{\mathrm{n-p}}{\mathrm{rm}} \mathrm{U}(\boldsymbol{L}),$$

while $d_1 = h = \max(r, m)$ and $d_2 = n - p - h + r$ for the test statistic

$$\frac{n - p - h + r}{h} \lambda_{max}(\boldsymbol{L}).$$

Table 12.2 Test Coverages: MANOVA F H_0 is False.

n	$m = p$	test	F_1	F_2	F_{p-1}	F_p	F_M
30	5	W	0.012	0.222	0.058	0.000	0.006
30	5	P	0.000	0.000	0.000	0.000	0.000
30	5	HL	0.382	0.694	0.322	0.007	0.579
30	5	R	0.799	0.871	0.549	0.047	0.997
50	5	W	0.984	0.955	0.644	0.017	0.963
50	5	P	0.971	0.940	0.598	0.012	0.871
50	5	HL	0.997	0.979	0.756	0.053	0.991
50	5	R	0.996	0.978	0.744	0.049	1
105	10	W	0.650	0.970	0.191	0.000	0.633
105	10	P	0.109	0.812	0.050	0.000	0.000
105	10	HL	0.964	0.997	0.428	0.000	1
105	10	R	1	1	0.892	0.052	1
150	10	W	1	1	0.948	0.032	1
150	10	P	1	1	0.941	0.025	1
150	10	HL	1	1	0.966	0.060	1
150	10	R	1	1	0.965	0.057	1
450	20	W	1	1	0.999	0.020	1
450	20	P	1	1	0.999	0.016	1
450	20	HL	1	1	0.999	0.035	1
450	20	R	1	1	0.999	0.056	1

Denote these statistics by W, P, HL, and R. Let the coverage be the proportion of times that H_0 is rejected. We want coverage near 0.05 when H_0 is true and coverage close to 1 for good power when H_0 is false. With 5000 runs, coverage outside of (0.04,0.06) suggests that the true coverage is not 0.05. Coverages are tabled for the F_1, F_2, F_{p-1}, and F_p test and for the MANOVA F test denoted by F_M. The null hypothesis H_0 was always true for the F_p test and always false for the F_1 test. When the MANOVA F test was true, H_0 was true for the F_j tests with $j \neq 1$. When the MANOVA F test was false, H_0 was false for the F_j tests with $j \neq p$, but the F_{p-1} test should be hardest to reject for $j \neq p$ by construction of $\boldsymbol{B}$ and the error vectors.

When the null hypothesis H_0 was true, simulated values started to get close to nominal levels for $n \geq 0.8(m+p)^2$ and were fairly good for $n \geq 1.5(m+p)^2$. The exception was Roy's test which rejects H_0 far too often if $r > 1$. See Table 12.1 where we want values for the F_1 test to be close to 1 since H_0 is false for the F_1 test, and we want values close to 0.05, otherwise. Roy's test was very good for the F_j tests but very poor for the MANOVA F test. Results are shown for $m = p = 10$. As expected from Berndt and Savin (1977), Pillai's test rejected H_0 less often than Wilks' test which rejected H_0 less often than the Hotelling Lawley test. Based on a much larger simulation study Pelawa Watagoda (2013, pp. 111–112), using the four types of error vector distributions and $m = p$, the tests had approximately correct level if $n \geq 0.83(m+p)^2$ for the Hotelling Lawley test, if $n \geq 2.80(m+p)^2$ for the Wilks' test (agreeing with Kshirsagar (1972) $n \geq 3(m+p)^2$ for multivariate normal data), and if $n \geq 4.2(m+p)^2$ for Pillai's test.

In Table 12.2, H_0 is only true for the F_p test where $p = m$, and we want values in the F_p column near 0.05. We want values near 1 for high power otherwise. If H_0 is false, often H_0 will be rejected for small n. For example, if $n \geq 10p$, then the m residual plots should start to look good, and the MANOVA F test should be rejected. For the simulated data, the test had fair power for n not much larger than mp. Results are shown for the lognormal distribution.

Some *R* output for reproducing the simulation is shown below. The *mpack* function is `mregsim`, and etype = 1 uses data from a MVN distribution. The fcov line computed the Hotelling Lawley statistic using Equation (12.8) while the hotlawcov line used Definition 12.10. The mnull=T part of the command means we want the first value near 1 for high power and the next three numbers near the nominal level 0.05 except for mancv where we want all of the MANOVA F test statistics to be near the nominal level of 0.05. The mnull=F part of the command means we want all values near 1 for high power except for the last column (for the terms other than mancv) corresponding to the F_p test where H_0 is true so we want values near the nominal level of 0.05. The "coverage" is the proportion of times that H_0 is rejected, so "coverage" is short for "power" and "level": we want the coverage near 1 for high power when H_0 is false, and we want the coverage near the nominal level 0.05 when H_0 is true. Also see Problem 12.10.

```
mregsim(nruns=5000,etype=1,mnull=T)
$wilkcov
[1] 1.0000 0.0450 0.0462 0.0430
$pilcov
[1] 1.0000 0.0414 0.0432 0.0400
$hotlawcov
[1] 1.0000 0.0522 0.0516 0.0490
$roycov
[1] 1.0000 0.0512 0.0500 0.0480
```

```
$fcov
[1] 1.0000 0.0522 0.0516 0.0490
$mancv
        wcv    pcv  hlcv    rcv   fcv
[1,] 0.0406 0.0332 0.049 0.1526 0.049

mregsim(nruns=5000,etype=2,mnull=F)

$wilkcov
[1] 0.9834 0.9814 0.9104 0.0408
$pilcov
[1] 0.9824 0.9804 0.9064 0.0372
$hotlawcov
[1] 0.9856 0.9838 0.9162 0.0480
$roycov
[1] 0.9848 0.9834 0.9156 0.0462
$fcov
[1] 0.9856 0.9838 0.9162 0.0480
$mancv
       wcv    pcv   hlcv    rcv    fcv
[1,] 0.993 0.9918 0.9942 0.9978 0.9942
```

12.5.2 Simulations for Prediction Regions

The same type of data and 5000 runs were used to simulate the prediction regions for $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ for multivariate regression. With n=100, m=2, and p=4, the nominal coverage of the prediction region is 90%, and 92% of the training data is covered. Following Olive (2013a), consider the prediction region $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\}$=$\{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\}$=$\{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$. Then the ratio of the prediction region volumes

$$\frac{h_i^m \sqrt{det(\boldsymbol{C}_i)}}{h_2^m \sqrt{det(\boldsymbol{C}_2)}}$$

was recorded where $i = 1$ was the nonparametric region, $i = 2$ was the semiparametric region, and $i = 3$ was the parametric MVN region. Here h_1 and h_2 were the cutoff $D_{(U_n)}(T_i, \boldsymbol{C}_i)$ for $i = 1, 2$, and $h_3 = \sqrt{\chi^2_{m,q_n}}$.

Table 12.3 Coverages for 90% Prediction Regions.

$\boldsymbol{w}$ dist	n	$m = p$	ncov	scov	mcov	voln	volm
MVN	48	2	0.901	0.905	0.888	0.941	0.964
MVN	300	5	0.889	0.887	0.890	1.006	1.015
MVN	1200	10	0.899	0.896	0.896	1.004	1.001
MIX	48	2	0.912	0.927	0.710	0.872	0.097
MIX	300	5	0.906	0.911	0.680	0.882	0.001
MIX	1200	10	0.904	0.911	0.673	0.889	0+
MVT(7)	48	2	0.903	0.910	0.825	0.914	0.646
MVT(7)	300	5	0.899	0.909	0.778	0.916	0.295
MVT(7)	1200	10	0.906	0.911	0.726	0.919	0.061
LN	48	2	0.912	0.926	0.651	0.729	0.090
LN	300	5	0.915	0.917	0.593	0.696	0.009
LN	1200	10	0.912	0.916	0.593	0.679	0+

If, as conjectured, the RMVN estimator is a consistent estimator when applied to the residual vectors instead of iid data, then the volume ratios converge in probability to 1 if the iid zero mean errors $\sim N_m(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and the volume ratio converges to 1 for $i = 1$ for a large class of elliptically contoured distributions. These volume ratios were denoted by voln and volm for the nonparametric and parametric MVN regions. The coverage was the proportion of times the prediction region contained $\boldsymbol{y}_f$ where ncov, scov, and mcov are for the nonparametric, semiparametric, and parametric MVN regions.

In the simulations, we took $n = 3(m + p)^2$ and $m = p$. Table 12.3 shows that the coverage of the nonparametric region was close to 0.9 in all cases. The volume ratio voln was fairly close to 1 for the three elliptically contoured distributions. Since the volume of the prediction region is proportional to h^m, the volume can be very small if h is too small and m is large. Parametric prediction regions usually give poor estimates of h when the parametric distribution is misspecified. Hence the parametric MVN region only performed well for multivariate normal data.

Some *R* output for reproducing the simulation is shown below. The *mpack* function is `mpredsim`, and etype = 1 uses data from a MVN distribution. The term "ncvr" is "ncov" in Table 12.3. Since up = 0.94, 94% of the training data is covered by the nominal 90% nominal prediction region. Also see Problem 12.11.

```
mpredsim(nruns=5000,etype=1)
$ncvr
[1] 0.9162
$scvr
[1] 0.916
$mcvr
```

```
[1] 0.9138
$voln
[1] 0.9892485
$vols
[1] 1
$volm
[1] 1.004964
$up
[1] 0.94
```

12.6 Two Robust Estimators

12.6.1 The rmreg Estimator

The classical multivariate linear regression estimator is found from m least squares multiple linear regressions of Y_j on the predictors. The first way to make a robust multivariate linear regression estimator is to replace least squares by a robust estimator, such as the m `hbreg` multiple linear regressions of Y_j on the predictors. Olive and Hawkins (2011) showed that the probability the `hbreg` estimator is equal to the least squares estimator goes to one as $n \to \infty$ for a large class of (univariate) error distributions. See Section 14.4. The class of (univariate) error distributions contains distributions that are not symmetric; however, for a skewed distribution, the slope estimates are similar, but the intercept $\hat{\beta}_1$ will differ for least squares and `hbreg`. See the Warning in Section 14.2.

The *mpack* function `rmreg` replaces least squares with `hbreg` to make the first robust multivariate linear regression estimator. Then the probability that the `rmreg` estimator is equal to the classical multivariate linear regression estimator also goes to 1 on a large class of distributions for the error vector $\boldsymbol{\epsilon}$, including many elliptically contoured distributions.

Hence the large sample nonparametric prediction region and the large sample Wilks' test, Pillai's test, and Hotelling Lawley test using the robust estimator `rmreg` are asymptotically equivalent to their analogs using the classical estimator for a large class of error vector distributions.

The `rmreg` estimator has some useful theory for clean data, but replacing the least squares multiple linear regression estimator by a highly outlier-resistant multiple linear regression estimator results in a multivariate linear regression estimator with outlier resistance that is still quite low. For `rmreg`, the tests are not valid when outliers are present since $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ uses the outliers.

Example 12.4. Buxton (1920) gave various measurements of 88 men. *Head length* and person's *height* were the response variables, while an intercept, *nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multivariate linear regression model. Observation 9 was deleted since it had missing values. Five individuals, numbers 62–66, were reported to be about 0.75 inches tall with head lengths well over five feet! Figure 12.8 shows the response and residual plots corresponding to Y_1 for the robust estimator `rmreg`. The response plot for the classical estimator, not shown, has the identity line tilted slightly above most of the plotted points in the lower part of the plot, while the plotted points in the lower part of the residual plot follow a line with negative slope instead of the $r = 0$ line. Figure 12.9 shows the response and residual plots corresponding to Y_2 for the robust estimator. The response plot for the classical estimator, not shown, has the identity line tilted slightly below most of the plotted points in the upper part of the plot, while the plotted points in the upper part of the residual plot follow a line with negative slope instead of the $r = 0$ line. Figure 12.10 shows the DD plot. The 90% semiparametric and nonparametric regions use the 95th percentile which is a linear combination of an outlying case with a nonoutlying case. The parametric MVN region contains cases below the RD = 2.448 line, which is obscured by the identity line. The tests of hypotheses for the robust estimator are not robust to outliers because all $n = 87$ residual vectors are

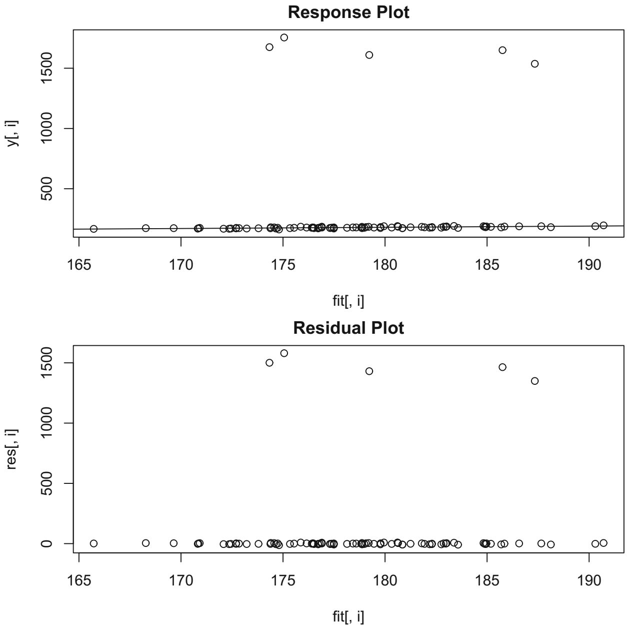

Fig. 12.8 Plots for Y_1 = head length

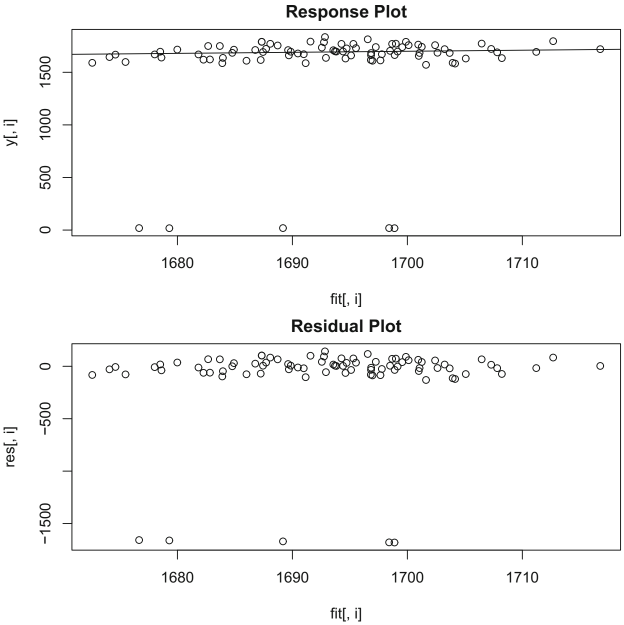

Fig. 12.9 Plots for Y_2 = height

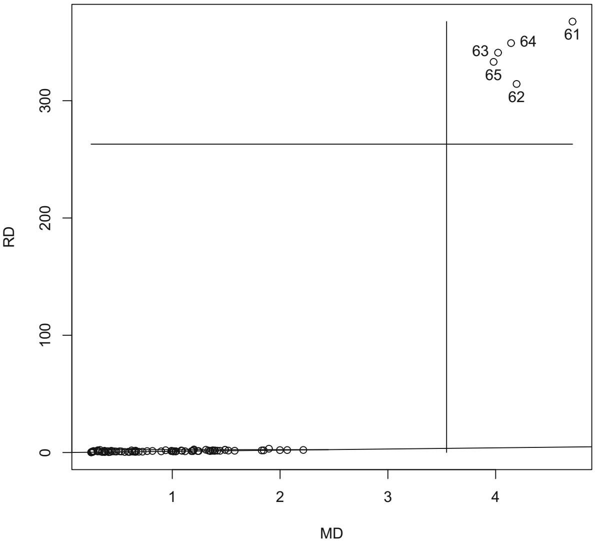

Fig. 12.10 DD Plot of the Residual Vectors for the Buxton Data

used to make $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. As is often the case, outliers can be detected with the plots using the classical or robust estimator.

These three figures can be made with the following *R* commands.

```
ht <- buxy
z <- cbind(buxx,ht)
y <- z[,c(1,5)]
x <- z[,2:4]
# compare mltreg(x,y)
out<-rmltreg(x,y) #right click Stop 4 times
out; ddplot4(out$res) #right click Stop
```

R functions for simulating testing and prediction regions using `rmreg` are `rmregsim` and `rmpredism`. The similar functions for the classical estimator delete the initial "r." The function `rmltreg` produces output similar to the *R* function `mltreg` for the classical estimator, including response and residual plots and output for testing with the `rmreg` estimator. For prediction, the simulation results for the robust estimator were similar to those for the classical estimator. For testing, the robust estimator needed larger values of n, and the tests did not work for the highly skewed lognormal distribution since the robust and classical estimators estimate the m constants (β_{1j} for $j = 1, ..., m$) differently when the data is skewed. Rupasinghe Arachchige Don (2013) did a large simulation study for testing using `rmreg`. Results suggested that for the three elliptically contoured distributions used in Section 12.5.1 and $m = p$, the tests had approximately correct level if $n > 150 + (m + p)^2$ for the Hotelling Lawley test, if $n > 140 + 3(m + p)^2$ for the Wilks' test, and if $n > 90 + 3.6(m + p)^2$ for Pillai's test.

12.6.2 The rmreg2 Estimator

The robust multivariate linear regression estimator `rmreg2` is the classical multivariate linear regression estimator applied to the RMVN set when RMVN is computed from the vectors $\boldsymbol{u}_i = (x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})^T$ for $i = 1, ..., n$. Hence $\boldsymbol{u}_i$ is the ith case with $x_{i1} = 1$ deleted. This regression estimator has considerable outlier resistance and is one of the most outlier resistant practical robust regression estimator for the $m = 1$ multiple linear regression case. See Chapter 14. The `rmreg2` estimator has been shown to be consistent if the $\boldsymbol{u}_i$ are iid from a large class of elliptically contoured distributions, which is a much stronger assumption than having iid error vectors $\boldsymbol{\epsilon}_i$.

First, we will review some results for multiple linear regression. Let $\boldsymbol{x} = (1, \boldsymbol{w}^T)^T$ and let

$$\mathrm{Cov}(\boldsymbol{w}) = \mathrm{E}[(\boldsymbol{w} - \mathrm{E}(\boldsymbol{w}))(\boldsymbol{w} - \mathrm{E}(\boldsymbol{w}))^{\mathrm{T}}] = \boldsymbol{\Sigma}_{\boldsymbol{w}}$$

and $\mathrm{Cov}(\boldsymbol{w}, Y) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(Y - E(Y))] = \boldsymbol{\Sigma}_{\boldsymbol{w}Y}$. Let $\boldsymbol{\beta} = (\alpha, \boldsymbol{\eta}^T)^T$ be the population OLS coefficients from the regression of Y on $\boldsymbol{x}$ ($\boldsymbol{w}$ and a constant), where α is the constant and $\boldsymbol{\eta}$ is the vector of slopes. Let the OLS estimator be $\hat{\boldsymbol{\beta}} = (\hat{\alpha}, \hat{\boldsymbol{\eta}}^T)^T$. Then the population coefficients from an OLS regression of Y on $\boldsymbol{x}$ are

$$\alpha = E(Y) - \boldsymbol{\eta}^T E(\boldsymbol{w}) \ \text{ and } \ \boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{w}\mathrm{Y}}. \tag{12.10}$$

Then the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. The sample covariance matrix of $\boldsymbol{w}$ is

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T \text{ where the sample mean } \overline{\boldsymbol{w}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \boldsymbol{w}_{\mathrm{i}}.$$

Similarly, define the sample covariance vector of $\boldsymbol{w}$ and Y to be

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(Y_i - \overline{Y}).$$

Suppose that $(Y_i, \boldsymbol{w}_i^T)^T$ are iid random vectors such that $\boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{w}Y}$ exist. Then

$$\hat{\alpha} = \overline{Y} - \hat{\boldsymbol{\eta}}^T\overline{\boldsymbol{w}} \xrightarrow{P} \alpha$$

and

$$\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y} \xrightarrow{P} \boldsymbol{\eta} \ \text{ as } \ \mathrm{n} \to \infty.$$

Now for multivariate linear regression, $\hat{\boldsymbol{\beta}}_j = (\hat{\alpha}_j, \hat{\boldsymbol{\eta}}_j^T)^T$ where $\hat{\alpha}_j = \overline{Y}_j - \hat{\boldsymbol{\eta}}_j^T\overline{\boldsymbol{w}}$ and $\hat{\boldsymbol{\eta}}_j = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y_j}$. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}\boldsymbol{y}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^T$ which has jth column $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y_j}$ for $j = 1, ..., m$. Let

$$\boldsymbol{u} = \begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{y} \end{pmatrix}, \quad E(\boldsymbol{u}) = \boldsymbol{\mu}_{\boldsymbol{u}} = \begin{pmatrix} E(\boldsymbol{w}) \\ E(\boldsymbol{y}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{w}} \\ \boldsymbol{\mu}_{\boldsymbol{y}} \end{pmatrix}, \ \text{ and } \ \mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} =$$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{ww}} & \boldsymbol{\Sigma}_{\boldsymbol{wy}} \\ \boldsymbol{\Sigma}_{\boldsymbol{yw}} & \boldsymbol{\Sigma}_{\boldsymbol{yy}} \end{pmatrix}.$$

Let the vector of constants be $\boldsymbol{\alpha}^T = (\alpha_1, ..., \alpha_m)$ and the matrix of slope vectors $\boldsymbol{B}_S = \begin{bmatrix} \boldsymbol{\eta}_1 & \boldsymbol{\eta}_2 & \dots & \boldsymbol{\eta}_m \end{bmatrix}$. Then the population least squares coefficient matrix is

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{\alpha}^T \\ \boldsymbol{B}_S \end{pmatrix}$$

where $\boldsymbol{\alpha} = \boldsymbol{\mu}_{\boldsymbol{y}} - \boldsymbol{B}_S^T \boldsymbol{\mu}_{\boldsymbol{w}}$ and $\boldsymbol{B}_S = \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{wy}}$ where $\boldsymbol{\Sigma}_{\boldsymbol{w}} = \boldsymbol{\Sigma}_{\boldsymbol{ww}}$.

If the $\boldsymbol{u}_i$ are iid with nonsingular covariance matrix $\text{Cov}(\boldsymbol{u})$, the least squares estimator

$$\hat{\boldsymbol{B}} = \begin{pmatrix} \hat{\boldsymbol{\alpha}}^T \\ \hat{\boldsymbol{B}}_S \end{pmatrix}$$

where $\hat{\boldsymbol{\alpha}} = \overline{\boldsymbol{y}} - \hat{\boldsymbol{B}}_S^T \overline{\boldsymbol{w}}$ and $\hat{\boldsymbol{B}}_S = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{wy}}$. The least squares multivariate linear regression estimator can be calculated by computing the classical estimator $(\overline{\boldsymbol{u}}, \boldsymbol{S}_{\boldsymbol{u}}) = (\overline{\boldsymbol{u}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}})$ of multivariate location and dispersion on the $\boldsymbol{u}_i$, and then plug in the results into the formulas for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{B}}_S$.

Let $(T, \boldsymbol{C}) = (\tilde{\boldsymbol{\mu}}_{\boldsymbol{u}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}})$ be a robust estimator of multivariate location and dispersion. If $\tilde{\boldsymbol{\mu}}_{\boldsymbol{u}}$ is a consistent estimator of $\boldsymbol{\mu}_{\boldsymbol{u}}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}$ is a consistent estimator of $c\ \boldsymbol{\Sigma}_{\boldsymbol{u}}$ for some constant $c > 0$, then a robust estimator of multivariate linear regression is the plug-in estimator $\tilde{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}} - \tilde{\boldsymbol{B}}_S^T \tilde{\boldsymbol{\mu}}_{\boldsymbol{w}}$ and $\tilde{\boldsymbol{B}}_S = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{wy}}$.

For the `rmreg2` estimator, $(T, \boldsymbol{C})$ is the classical estimator applied to the RMVN set when RMVN is applied to vectors $\boldsymbol{u}_i$ for $i = 1, ..., n$ (could use $(T, \boldsymbol{C})$ = RMVN estimator since the scaling does not matter for this application). Then $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}_{\boldsymbol{u}}, c\ \boldsymbol{\Sigma}_{\boldsymbol{u}})$ if the $\boldsymbol{u}_i$ are iid from a large class of $EC_d(\boldsymbol{\mu}_{\boldsymbol{u}}, \boldsymbol{\Sigma}_{\boldsymbol{u}}, g)$ distributions where $d = m + p - 1$. Thus the classical and robust estimators of multivariate linear regression are both $\sqrt{n}$ consistent estimators of $\boldsymbol{B}$ if the $\boldsymbol{u}_i$ are iid from a large class of elliptically contoured distributions. This assumption is quite strong, but the robust estimator is useful for detecting outliers. When there are categorical predictors or the joint distribution of $\boldsymbol{u}$ is not elliptically contoured, it is possible that the robust estimator is bad and very different from the good classical least squares estimator.

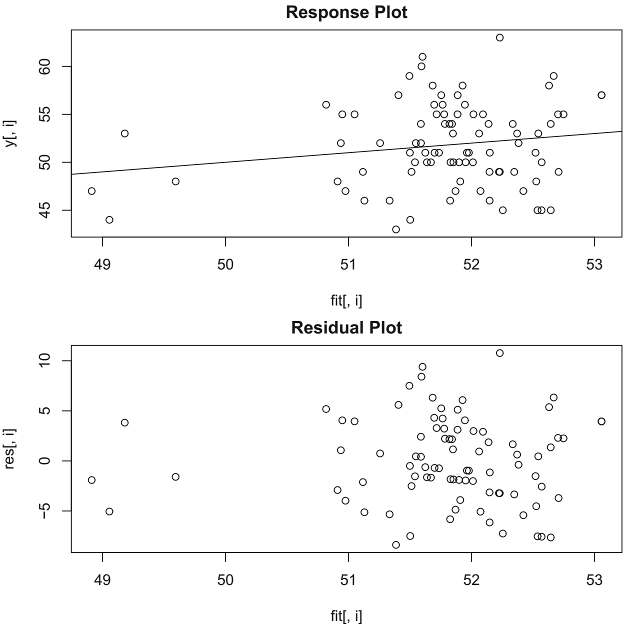

Fig. 12.11 Plots for Y_1 = nasal height using `rmreg`

Example 12.4, continued. The *mpack* function `rmreg2` computes the `rmreg2` estimator and produces the response and residual plots. The plots for the `rmreg2` estimator were very similar to Figures 12.8 and 12.9.

Now let Y_1 = *nasal height* and Y_2 = *height* with x_2 = *head length*, x_3 = *bigonal breadth*, and x_4 = *cephalic index*. Then Y_2 and x_2 have massive outliers. Then the response and residual plots for the classical estimator and the robust estimator `rmreg` using `hbreg` were nearly identical. Figures 12.11 and 12.12 show that the fit using `rmreg` went right through the outliers. Figures 12.13 and 12.14 show that the response and residual plots corresponding to `rmreg2` do not have fits that pass through the outliers.

These figures can be made with the following *R* commands.

```
ht <- buxy; z <- cbind(buxx,ht);
y <- z[,c(2,5)]; x <- z[,c(1,3,4)]
# compare mltreg(x,y) #right click Stop 4 times
rmltreg(x,y) #right click Stop 4 times
out <- rmreg2(x,y) #right click Stop 4 times
# try ddplot4(out$res) #right click Stop
```

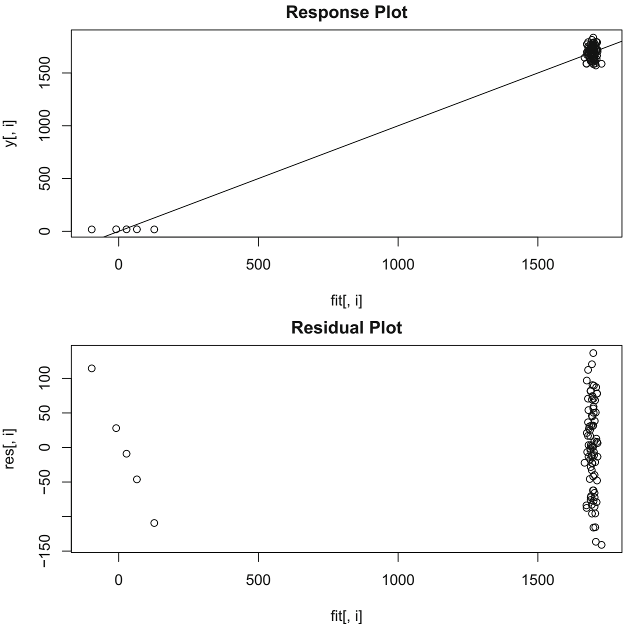

Fig. 12.12 Plots for Y_2 = height using rmreg

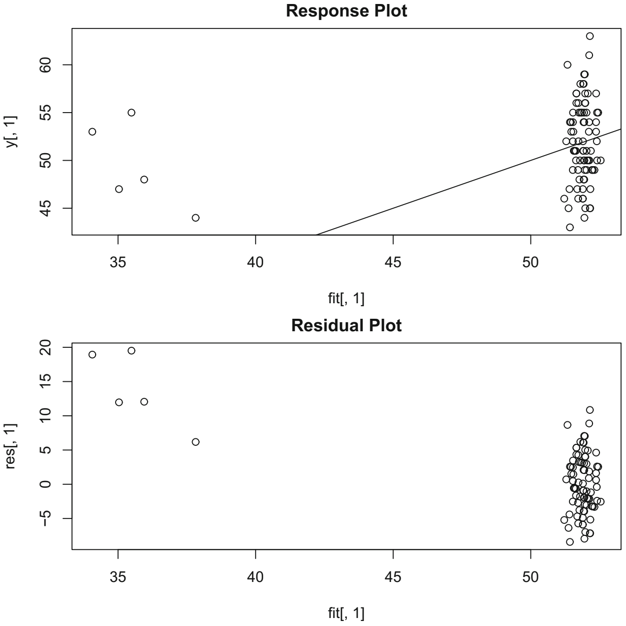

Fig. 12.13 Plots for Y_1 = nasal height using rmreg2

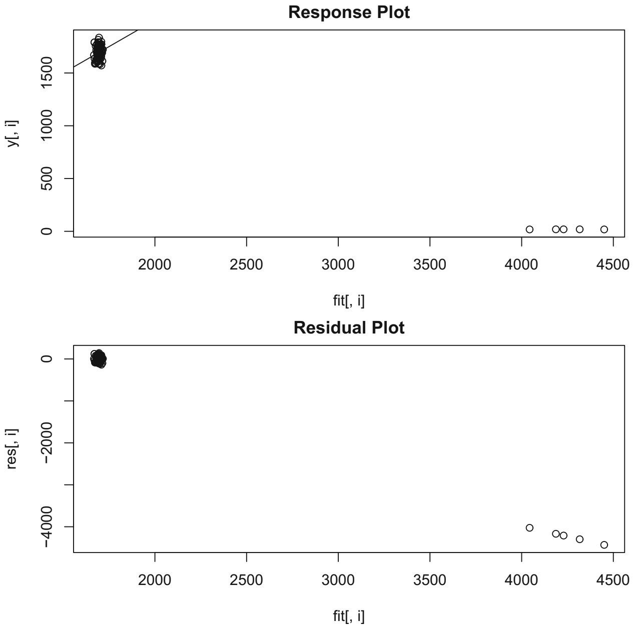

Fig. 12.14 Plots for Y_2 = height using `rmreg2`

The residual bootstrap for the test $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ may be useful. Take a sample of size n with replacement from the residual vectors to form $\boldsymbol{Z}_1^*$ with ith row $\boldsymbol{y}_i^{*T}$ where $\boldsymbol{y}_i^* = \hat{\boldsymbol{y}}_i + \boldsymbol{\epsilon}_i^*$. The function `rmreg3` gets the `rmreg2` estimator without the plots. Using `rmreg3`, regress $\boldsymbol{Z}$ on $\boldsymbol{X}$ to get $vec(\boldsymbol{L}\hat{\boldsymbol{B}}_1^*)$. Repeat B times to get a bootstrap sample $\boldsymbol{w}_1, ..., \boldsymbol{w}_B$ where $\boldsymbol{w}_i = vec(\boldsymbol{L}\hat{\boldsymbol{B}}_i^*)$. The nonparametric bootstrap uses n cases drawn with replacement and may also be useful. Apply the nonparametric prediction region to the $\boldsymbol{w}_i$ and see if $\boldsymbol{0}$ is in the region. If $\boldsymbol{L}$ is $r \times p$, then $\boldsymbol{w}$ is $rp \times 1$, and we likely need $n \geq \max[50rp, 3(m+p)^2]$.

12.7 Seemingly Unrelated Regressions

Each response variable in a multivariate linear regression model follows a multiple linear regression model $\boldsymbol{Y}_j = \boldsymbol{X\beta}_j + \boldsymbol{e}_j$ 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 unrelated 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 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 12.11. 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 are 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 $\boldsymbol{Y}_j = \boldsymbol{X}_j \boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$.
3) A prediction region for $\boldsymbol{y}_f$ is made as in Section 12.3 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.

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 in Chapter 10 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 $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$. 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 $E(Y_{ij}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j$.

3) 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 $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj} \boldsymbol{I}_n$.

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 transformed or corrected. If the model is good, then 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 an $r \times p$ 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 Wilks' 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 $\lambda_{max}(\boldsymbol{L}) = \lambda_1$.

7) **Theorem**: The Hotelling-Lawley trace statistic

$$U(\boldsymbol{L}) = \frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})].$$

8) **Assumption D1**: Let h_i be the ith diagonal element of $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ $\boldsymbol{X}^T$. Assume $\max(h_1, ..., h_n) \xrightarrow{P} 0$ as $n \to \infty$, assume that the zero mean iid error vectors have finite fourth moments, and assume that $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \xrightarrow{P} \boldsymbol{W}^{-1}$.

9) **Multivariate Least Squares Central Limit Theorem (MLS CLT):** For the least squares estimator, if assumption D1 holds, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, and $\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$.

10) **Theorem**: If assumption D1 holds and if H_0 is true, then $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

11) Under regularity conditions, $-[n-p+1-0.5(m-r+3)]$ $\log(\Lambda(\boldsymbol{L})) \xrightarrow{D} \chi^2_{rm}$, $(n-p)V(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, and $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

These statistics are robust against nonnormality.

12) For the Wilks' Lambda test,

$$pval = P\left(\frac{-[n-p+1-0.5(m-r+3)]}{rm}\ \ \log(\Lambda(\boldsymbol{L})) < F_{rm,n-rm}\right).$$

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) $\to \delta$ as $n \to \infty$, under regularity conditions.

13) 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 $\leq \delta$, reject H_0. If pval $> \delta$, 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.)

14) 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 $\boldsymbol{B}$. The hypotheses are equivalent to $H_0: \ \boldsymbol{B}_j^T = \boldsymbol{0} \quad H_1: \ \ \boldsymbol{B}_j^T \neq \boldsymbol{0}$. i) State the hypotheses
H_0: x_j is not needed in the model H_1: x_j is needed in the model.
ii) Find the test statistic F_j from output.
iii) Find pval from output.
iv) If pval $\leq \delta$, reject H_0. If pval $> \delta$, 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 $Y_1, ..., Y_m$. 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.

15) 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 $\leq \delta$, reject H_0 and conclude that the full model should be used.
If pval $> \delta$, fail to reject H_0 and conclude that the reduced model is good.

16) 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.

17) 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 com-

mand `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
```

18) 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 $\hat{y}_i = \hat{\boldsymbol{\beta}}_i^T \boldsymbol{x}_f$.

19) $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \dfrac{\hat{\boldsymbol{E}}^T \hat{\boldsymbol{E}}}{n-p} = \dfrac{1}{n-p} \displaystyle\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^T$ 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 distributions for the error vectors $\boldsymbol{\epsilon}_i$.

20) The $100(1-\delta)\%$ 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 $q_n = \min(1 - \delta + 0.05, 1 - \delta + m/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta m/n), \quad \text{otherwise.}$$

If $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let $0 < \delta < 1$ and $h = D_{(U_n)}$ where $D_{(U_n)}$ is the q_nth sample quantile of the D_i. The $100(1 - \delta)\%$ nonparametric prediction region for $\boldsymbol{y}_f$ is

$$\{\boldsymbol{y} : (\boldsymbol{y} - \hat{\boldsymbol{y}}_f)^T \boldsymbol{S}_r^{-1} (\boldsymbol{y} - \hat{\boldsymbol{y}}_f) \leq D^2_{(U_n)}\} = \{\boldsymbol{y} : D_{\boldsymbol{y}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}\}.$$

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 $U_n/n \to 1 - \delta$ 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 nonparametric prediction region is a large sample $100(1 - \delta)\%$ 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 unique highest density region is $\{\boldsymbol{y} : D_{\boldsymbol{y}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}\}$, then the nonparametric prediction region is asymptotically optimal.

21) On the DD plot for the residual vectors, 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.

22) The DD plot for the residual vectors 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 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.

23) The robust multivariate linear regression method `rmreg` replaces least squares with the `hbreg` estimator. The probability that the `rmreg` estimator equals the classical 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 sometimes 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.

24) The robust multivariate linear regression method `rmreg2` computes the classical estimator on the RMVN set where RMVN is computed from

the n cases $\boldsymbol{u}_i = (x_{i2}, ..., x_{pi}, Y_{i1}, ..., Y_{im})^T$. This estimator has considerable outlier resistance but theory currently needs very strong assumptions. The response and residual plots and DD plot of the residuals from this estimator are useful for outlier detection. The `rmreg2` estimator is superior to the `rmreg` estimator for outlier detection.

12.9 Complements

Multivariate linear regression is a semiparametric method that is nearly as easy to use as multiple linear regression if m is small. The material on plots and testing followed Olive et al. (2015) closely. The m response and residual plots should be made as well as the DD plot, and the response and residual plots are very useful for the $m = 1$ case of multiple linear regression and experimental design. These plots speed up the model building process for multivariate linear models since the success of power transformations achieving linearity can be quickly assessed, and influential cases can be quickly detected. See Cook and Cook and Olive (2001). Work is needed on variable selection and on determining the sample sizes for when the tests and prediction regions start to work well. Response and residual plots can look good for $n \geq 10p$, but for testing and prediction regions, we may need $n \geq a(m+p)^2$ where $0.8 \leq a \leq 5$ even for well behaved elliptically contoured error distributions. Cook and Setodji (2003) used the FF plot.

Often observations $(Y_1, ..., Y_m, x_2, ..., x_p)$ are collected on the same person or thing and hence are correlated. If transformations can be found such that the m response plots and residual plots look good, and n is large ($n \geq \max[(m+p)^2, mp + 30)]$ starts to give good results), then multivariate linear regression can be used to efficiently analyze the data. Examining m multiple linear regressions is an incorrect method for analyzing the data.

In addition to robust estimators and seemingly unrelated regressions, envelope estimators and partial least squares (PLS) are competing methods for multivariate linear regression. See recent work by Cook such as Cook and Su (2013), Cook et al. (2013), and Su and Cook (2012). Methods like ridge regression and lasso can also be extended to multivariate linear regression. See, for example, Obozinski et al. (2011). Prediction regions for alternative methods with $n >> p$ could be made following Section 12.3.

Section 12.3 follows Olive (2017b) closely. Consider the model $\boldsymbol{y}_i = E(\boldsymbol{y}_i|\boldsymbol{x}_i) + \boldsymbol{e}_i = m(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$. A practical method for producing a prediction region for $\boldsymbol{y}_f$ is to create pseudodata $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_1, ..., \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_n$ using the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ and the predicted value $\hat{\boldsymbol{y}}_f$. Then apply prediction region (5.12) to the pseudodata but modify $c = k_n = \lceil n(1-\delta) \rceil$ so that the coverage is better for moderate samples. Often the nonparametric prediction region (5.17) will work.

There is little competition for the nonparametric prediction region for multivariate regression if $m > 1$. For $m = 1$ and multiple linear regression, the prediction intervals of Olive (2007) based on the short should be shorter when the error distribution is not symmetric. The parametric MVN region works if the errors $\boldsymbol{\epsilon}_i$ come from a MVN distribution, but this region tends to have volume that is too small if the error distribution is not MVN. For m not much larger than two, the Lei et al. (2013) prediction region or the Hyndman (1996) prediction region (5.8) could be used on the pseudodata $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$, possibly modifying (5.8) with a value of δ_n that increases to δ as $n \to \infty$. For $m = 1$, a similar procedure is done by Lei and Wasserman (2014).

Some robust estimators are described by Wilcox (2009) and Rousseeuw et al. (2004), where the practical FLTS and FMCD methods are not yet backed by theory and should be replaced by the methods in Section 12.6. Plugging in robust dispersion estimators in place of the covariance matrices, as done in Section 12.6.2, is not a new idea. Maronna and Morgenthaler (1986) used M–estimators when $m = 1$. Problems can occur if the error distribution is not elliptically contoured. See Nordhausen and Tyler (2015).

If $m = 1$ and n is not much larger than p, then Hoffman et al. (2015) gave a robust Partial Least Squares–Lasso type estimator that uses a clever weighting scheme.

The *R* software was used to make plots and software. See R Core Team (2016). The function `mpredsim` was used to simulate the prediction regions, `mregsim` was used to simulate the tests of hypotheses, and `mregddsim` simulated the DD plots for various distributions. The function `mltreg` makes the response and residual plots and computes the F_j, MANOVA F, and MANOVA partial F test pvalues, while the function `ddplot4` makes the DD plots.

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. There is 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**. Variable selection for multivariate linear regression is discussed in Fujikoshi et al. (2014). *R* programs are needed to make variable selection easy. Forward selection would be especially useful.

To do crude variable selection, fit the model, leave out the variable with the largest F_j test pvalue > 0.1, and fit the model, and repeat. The statistic $C_p(I) = (p - k)(F_I - 1) + k$ may also be useful. Here p is the number of variables in the full model, k is the number of variables in the candidate model I, and F_I is the MANOVA partial F statistic for testing whether the $p - k$ variables $\boldsymbol{x}_O$ (in the full model but not in the candidate model I) can be deleted. Models that have $C_p(I) \leq k$ are certainly interesting. Check the final 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 $\hat{Y}_{Ij}$ for $j = 1, ..., m$.

If $n < 10p$, do forward selection until there are $J \approx n/10$ predictors. Check that the model with J predictors is reasonable. Then compute $C_p(I)$ for each model considered in the forward selection.

The theory for multivariate linear regression assumes that the model is known before gathering data. If variable selection and response transformations 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 regions will be wrong. When $m = 1$, 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) using variables suggested by the pilot study.

Khattree and Naik (1999, pp. 91–98) discussed 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. Johnstone and Nadler (2017) gave useful approximations for Roy's largest root test when the error vector distribution is multivariate normal.

12.10 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

12.1*. Consider the Hotelling Lawley test statistic. Let

$$T(\boldsymbol{W}) = n \quad [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{LWL}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})].$$

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. Consider 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 that $(n-p)U_j = T_j$.

12.3. Consider the Hotelling Lawley test statistic. Using the Searle (1982, p. 333) identity

$$tr(\boldsymbol{A}\boldsymbol{G}^T\boldsymbol{D}\boldsymbol{G}\boldsymbol{C}) = [vec(\boldsymbol{G})]^T[\boldsymbol{C}\boldsymbol{A} \otimes \boldsymbol{D}^T][vec(\boldsymbol{G})],$$

$$\text{show} \quad (\mathrm{n}-\mathrm{p})\mathrm{U}(\boldsymbol{L}) = \mathrm{tr}[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{B}}^{\mathrm{T}}\boldsymbol{L}^{\mathrm{T}}[\boldsymbol{L}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{L}^{\mathrm{T}}]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}]$$

$= [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$ by identifying $\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{D}$, and $\boldsymbol{C}$.

```
$Ftable    Fj        pvals  #Output for problem 12.4.
[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 output above 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?

12.6. The output below 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.

```
y<-USJudgeRatings[,c(9,10,12)] #See problem 12.6.
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
```

12.7. 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 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, e.g., *ddplot*, will display the code for the function. Use the args command, e.g., *args(ddplot)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/mrsashw.txt) into *R*.

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 makes the response and residual plots for each of the two response 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 part makes a DD plot of the residual vectors and adds the lines corresponding to the three prediction regions of Section 12.3. The robust cutoff is larger than the semiparametric cutoff. Place the plot in *Word.* Do the residual vectors appear to follow a multivariate normal distribution? (Right click *Stop* once.)

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 the SAS Institute (1985, p. 146) Fitness Club 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 makes 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 part makes a DD plot of the residual vectors and adds the lines corresponding to the three prediction regions of Section 12.3. The robust cutoff is larger than the semiparametric cutoff. Place the plot in *Word.* Are there any outliers? (Right click *Stop* once.)

12.10. This problem uses the *mpack* function `mregsim` to simulate the Wilks' Λ 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 $\mathbf{0}^T$. 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 $\mathbf{0}^T$. 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) etc., 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 $\sigma_{ii} = \sigma^2[1 + (m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = \sigma^2[2\psi + (m-2)\psi^2]$ where $\psi = 0.10$. Terms like *Wilkcov* give the percentage of times the Wilks' 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 12.3.

5000 runs will be used so the simulation may take several minutes. Sample sizes $n = (m+p)^2, n = 3(m+p)^2$, and $n = 4(m+p)^2$ were interesting. We 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 where $n = 20, m = 2$, and $p = 4$. Here H_0 is true except for the F_1 test. Wilks' 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 where $n = 20, m = 2$, and $p = 4$. 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 may take several minutes. The *R* command for this problem generates iid lognormal errors then subtracts the mean, producing $\boldsymbol{z}_i$. Then the $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{z}_i$ are generated as in Problem 12.10 with 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

Clustering is used to classify the n cases into k groups. Unlike discriminant analysis, it is not known to which group the cases in the training data belong, and often the number of clusters k is unknown. Discriminant analysis is a type of supervised classification while clustering is a type of unsupervised classification. Factor analysis grouped highly correlated variables X_j together (columns of the data matrix $\boldsymbol{W}$). Clustering groups cases $\boldsymbol{x}_i$ together (rows of the data matrix).

13.1 Hierarchical and *k*-Means Clustering

Two common methods of clustering are k-means clustering and hierarchical clustering. A wide variety of distances or similarities have been suggested. We will focus on Euclidean distances.

For the simplest version of k-means clustering, there are four 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 each 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. It is often useful to run the k-means clustering program with several randomly drawn partitions or seeds, and to use several values of k.

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_13

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 dendrogram. Use the dendrogram to select k reasonable clusters of cases.

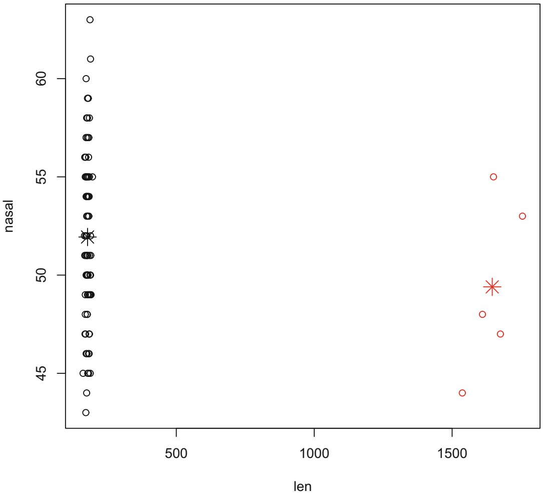

Fig. 13.1 Two clusters from k-means clustering with $k = 2$

Example 13.1. Often the clean data and outliers form two clusters. The R function kmeans was used on the Buxton (1920) data to produce Figure 13.1. See the R commands below. The DD plot of the Buxton data shown in Figure 5.10 also suggests a cluster of outliers and a cluster of clean data.

```
x <- cbind(buxx,buxy)
out<-kmeans(x,2,nstart=25)
plot(x, col = out$cluster)
points(out$centers, col = 1:2, pch = 8, cex=2)
```

Using five clusters does not change the appearance of the plot much. Try the commands below.

```
out5<-kmeans(x,5,nstart=25)
plot(x, col = out5$cluster)
points(out5$centers, col = 1:5, pch = 8, cex=2)
```

Removing the outliers and trying five clusters seem to show one cluster. Try the commands below.

```
xc <-x[-c(61,62,63,64,65),]
out<-kmeans(xc,5,nstart=25)
plot(xc, col = out$cluster)
points(out$centers, col = 1:5, pch = 8, cex=2)
```

The following commands suggest that the clustering was done using values of buxy = height.

```
plot(xc[,c(1,5)],col = out$cluster)
points(out$centers[,c(1,5)],col=1:5,pch=8,cex=2)
```

Example 13.2. *R* functions for hierarchical clustering include `hclust` and `agnes`. See MathSoft (1999b, ch. 4) and Kaufman and Rousseeuw (1990, ch. 5). One problem with hierarchical clustering is that it can be hard to read the labels on the dendrogram unless n is small. The dendrogram for the Buxton (1920) data is shown in Figure 13.2. The very top of the dendrogram is a cluster containing all of the data. Then two clusters are formed, one containing the five outlying cases (the five cases furthest to the left on the bottom of the plot) and one cluster containing all of the remaining cases. Outliers often appear among the last clusters formed in the dendrogram, corresponding to the clusters near the top of the dendrogram.

```
x <- cbind(buxx,buxy)
out <- hclust(dist(x),"complete")
#complete is the default
plot(out)
plot(out,hang=-1)
```

Following James et al. (2014, pp. 391–392), to interpret the dendrogram, each *leaf* on the bottom of Figure 13.2 represents one of the 87 cases of the Buxton data. As we move up the tree, some leaves begin to fuse into branches corresponding to cases that are similar to each other. Moving further up the tree causes branches to fuse with other branches or leaves. The lower in the tree that the fusions occur, the more similar the group of cases are to each other. Cases that fuse near the top of the tree can be quite different. The outliers fused together quickly, and the clean cases fused together quickly. The outliers and clean cases fused together last since the outliers and clean cases are quite different.

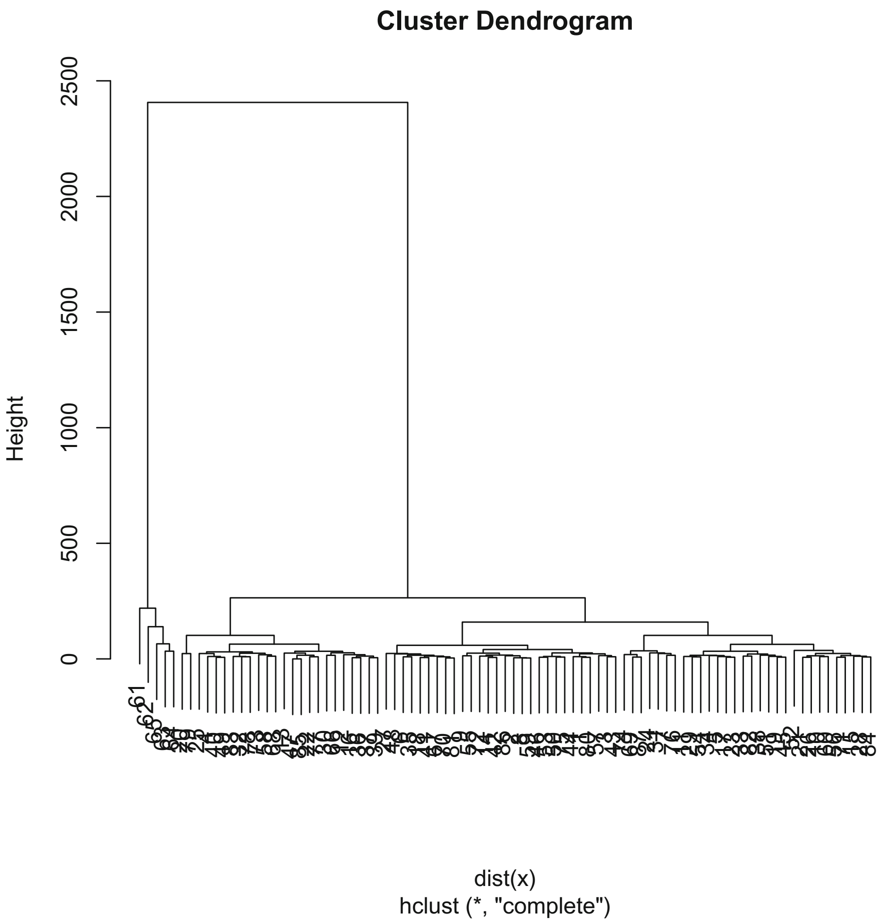

Fig. 13.2 Dendrogram for Buxton (1920) data

Example 13.3. Following James et al. (2014, pp. 392–393), observations that are close together horizontally are not necessarily similar. Case 5 and 7 are similar and cases 1 and 6 are similar since they fuse together at the lowest points in the dendrogram shown in Figure 13.3. Cases 9 and 2 are located close together horizontally, cases 2, 5, 7, and 8 fuse with case 9 at the same height. Hence case 9 is about as similar to cases 5, 7, and 8 as case 9 is to case 2. Plot the raw data to help see this. See Problem 13.3. The height of the fusion determines similarity. A horizontal line at 1.5 gives two clusters, while a horizontal line at 1.0 gives five clusters: i) 1, 6, and 4; ii) 3; iii) 2; iv) 5, 7, and 8; and v) 9. See the *R* code shown below to produce Figure 13.3.

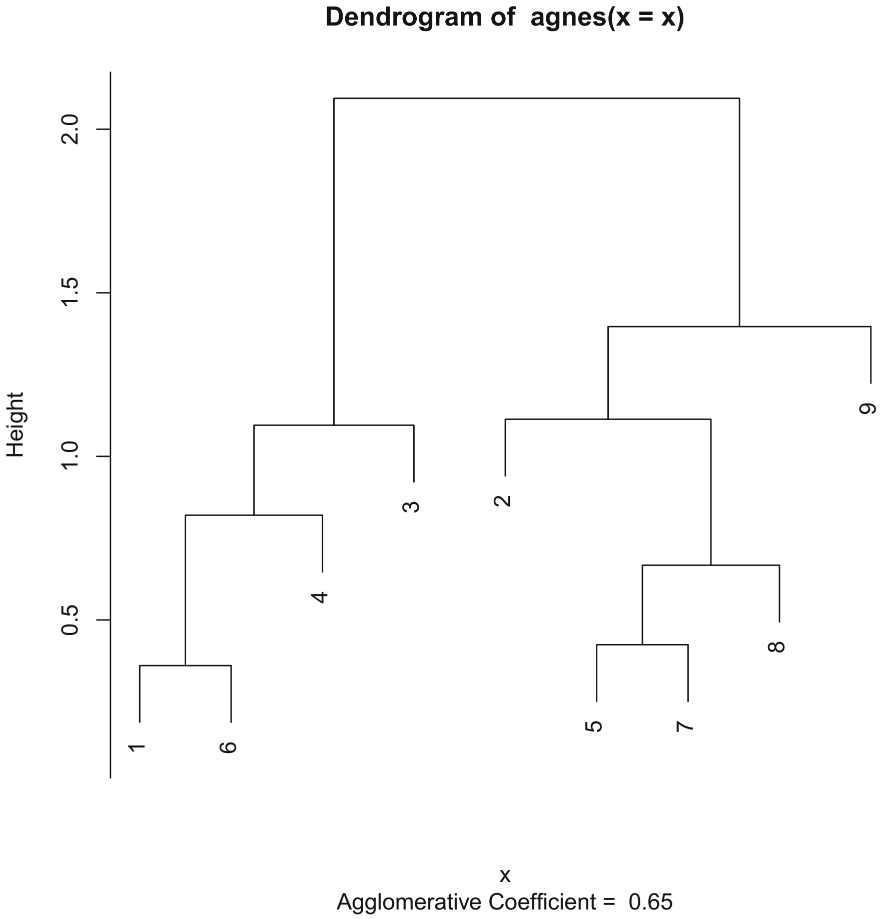

Fig. 13.3 9 and 2 are close in horizontal distance, but 2, 5, 7, and 8 fuse with 9 at the same height

```
x1 <- c(-0.6,0.1,-1.5,-1.4,1.1,-0.9,1.4,0.6,0)
x2 <- c(-1,-0.75,-0.4,-1.6,-0.3,-1.2,0,-0.2,0.7)
x <- cbind(x1,x2)
##out<-hclust(x) #errors
out <- hclust(dist(x))
plot(out)
plot(x[,1],x[,2])
library(cluster)
out<-agnes(x)
plot(out) #right click twice
```

13.2 Complements

Atkinson et al. (2004, ch. 7) gave some interesting ideas. Also see Kaufman and Rousseeuw (1990), and Farcomeni and Greco (2015), and Ritter (2014). A good review for robust methods is García-Escudero et al. (2010). For high dimensional clustering, see Jin and Wang (2016).

13.3 Problems

R Problems

For some of the following problems, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/mrsashw.txt) into *R*.

13.1. Enter the commands for Example 13.1 to reproduce Figure 13.1.

13.2. Enter the commands for Example 13.2 to reproduce Figure 13.2.

13.3. Enter the commands for Example 13.3 to reproduce Figure 13.3. Also plot X_1 versus X_2 to see that case 9 is about as similar to case 2 as case 9 is to cases 5, 7, and 8.

13.4. a) Obtain the file `mbb1415.csv` from (http://lagrange.math.siu.edu/Olive/multbk.htm), and save it on a flash drive (F, say). This file contains comma-separated variables. The commands for this problem show how to read the file into *R*.

The file, obtained and analyzed by Nicole Staples and Philip Kains, contains variables on male basketball players from the Missouri Valley conference 2014–2015 season. The first variable $x_1 = position$ where 0 means position is unknown, 1 for guard, 2 for guard-forward, 3 for forward, 4 for forward-center, and 5 for center. The variable x_2 is games played, x_3 is number of minutes played, x_4 is sst (an efficiency rating), x_5 is sst.ex.pts (an efficiency rating excluding points), x_6 is points, x_7 is assists, x_8 is turnovers, x_9 is assists to turn over ratio, x_{10} is steals, x_{11} is stl.pos (stolen possessions, a ball handling rating), x_{12} is blocks, x_{13} is rebounds, x_{14} is offensive rebounds, x_{15} is defensive rebounds, x_{16} is games played $= x_2$, x_{17} is field goal (FG) attempts, x_{18} is field goals made, x_{19} is FGs missed, x_{20} is field goal percentage, x_{21} is adjusted field goal percentage, x_{22} is two point field goal attempts, x_{23} is two point field goals made, x_{24} is two point FGs missed, x_{25} is two point field goal percentage, x_{26} is three point field goal attempts, x_{27} is three point field goals made, x_{28} is three point FGs missed, x_{29} is three point field goal percentage, x_{30} is free throws attempted, x_{31} is free throws made, x_{32} is free throws missed, x_{33} is free throw percentage, x_{34} is related to the number of "and one plays" (free throw after a made shot), x_{35} is personal fouls taken, and x_{36} is personal fouls committed.

Note that $\boldsymbol{X}$ will not be full rank since, for example $x_{16} = x_2$, and offensive rebounds + defensive rebounds = rebounds.

b) Sometimes the classes are known and you want to see how well clustering works. The commands for this problem use assists and rebounds to form the clusters. The second dendrogram uses positions as labels. We would like each cluster to have one position or neighboring positions (all labels are i's or all labels are i's and $(i+1)$'s). Include the second plot in *Word*.

c) Many basketball players do not play much so all of their statistics are near zero (and could be regarded as near point mass outliers). The commands for this problem delete about 25% of the players who had the fewest minutes and then uses assists and rebounds to form the clusters. Include the plot in *Word*.

13.5. a) Obtain the file `wbb1415.csv` from (http://lagrange.math.siu.edu/Olive/multbk.htm), and save it on a flash drive (F, say). This file contains comma-separated variables. The commands for this problem show how to read the file into *R*.

The file, obtained and analyzed by Nicole Staples and Philip Kains, contains variables on female basketball players from the Missouri Valley conference 2014–2015 season.

The variables are almost the same as those in Problem 13.4. The only difference is that this file does not have two games played variables. Hence variables $x_1, ..., x_{15}$ are the same, but x_i for the `wbb1415` data set are variables x_{i+1} for the `mbb1415` data set for $i = 16, ..., 35$.

b) Sometimes the classes are known and you want to see how well clustering works. The commands for this problem use assists and rebounds to form the clusters. The second dendrogram uses positions as labels. We would like each cluster to have one position or neighboring positions (all labels are i's or all labels are i's and $(i+1)$'s). Include the second plot in *Word*.

c) Many basketball players do not play much so all of their statistics are near zero (and could be regarded as near point mass outliers). The commands for this problem delete about 25% of the players who had the fewest minutes and then uses assists and rebounds to form the clusters. Include the plot in *Word*.

Chapter 14
Other Techniques

This chapter suggests several other techniques using robust estimators. From the literature, often the "robust method" can be improved by replacing the plug in estimator (often FMCD, FS, FMM, or FMVE) with RFCH or RMVN. Using the RMVN set U can also be useful. A short list of some techniques is given in Section 14.1, and then more details are given for robust regression and 1D regression. See Table 1.1 for acronyms.

Three *mpack* functions are useful for cleaning data with the RMVN set U described in Section 4.6. i) The function `getu` gets the RMVN set U. ii) If there are g groups ($g = G$ for discriminant analysis, $g = 2$ for binary regression, and $g = p$ for one-way MANOVA), the function `getubig` gets the RMVN set U_i for each group and combines the g RMVN sets into one large set $U_{big} = U_1 \cup U_2 \cup \cdots \cup U_g$. iii) If there are g groups and it can be assumed that the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ only differs by g location vectors, then `getuc` subtracts the group coordinatewise median from each group, combines the centered data into one data set $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$, and gets the RMVN set U_c applied to the $\boldsymbol{z}_i$. All three functions return *indx*, the indices of the cases in the cleaned data set (U, U_{big}, or U_c). Functions `getbigu` and `getuc` also return *grp* which has the group to which each case in the cleaned data set belongs.

Two functions are useful for cleaning data with the `covmb2` set B, which can be useful even if $p > n$. See Section 4.7. The function `getB` gets B, and the function `getBbig` is like `getubig` except it gets the `covmb2` set B_i for each group, and $B_{big} = B_1 \cup B_2 \cup \cdots \cup B_g$.

14.1 A List of Techniques

To find some techniques in the literature that can be robustified, Google terms like *robust binary regression, robust cluster analysis, robust errors in variables, robust functional data, robust generalized partial linear models, robust independent component analysis, robust invariant coordinates, robust longitu-*

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_14

dinal data analysis, robust orthogonal regression, robust principal components regression, robust quality control, robust regression, and *robust singular value decomposition.* Some more techniques are listed below, and some techniques are covered in more detail in the remaining sections of this chapter.

i) *Resistant regression:* Suppose the regression model has an $m \times 1$ response vector $\boldsymbol{y}$, and a $p \times 1$ vector of predictors $\boldsymbol{x}$. Assume that predictor transformations have been performed to make $\boldsymbol{x}$, and that $\boldsymbol{w}$ consists of $k \leq p$ continuous predictor variables that are linearly related. Find the RMVN set based on the $\boldsymbol{w}$ to obtain n_u cases $(\boldsymbol{y}_{ci}, \boldsymbol{x}_{ci})$ and then run the regression method on the cleaned data. Often the theory of the method applies to the cleaned data set since $\boldsymbol{y}$ was not used to pick the subset of the data. Efficiency can be much lower since n_u cases are used where $n/2 \leq n_u \leq n$, and the trimmed cases tend to be the "farthest" from the center of $\boldsymbol{w}$.

The method will have the most outlier resistance if $k = p$ (or $k = p - 1$ if there is a trivial predictor $X_1 \equiv 1$). If $m = 1$, make the response plot of $\hat{Y}_c$ versus Y_c with the identity line added as a visual aid and make the residual plot of $\hat{Y}_c$ versus $r_c = Y_c - \hat{Y}_c$.

In *R*, assume Y is the vector of response variables, x is the data matrix of the predictors (often not including the trivial predictor), and w is the data matrix of the $\boldsymbol{w}_i$. Then the following *R* commands can be used to get the cleaned data set. We could use the `covmb2` set B instead of the RMVN set U computed from the $\boldsymbol{w}$ by replacing the command *getu(w)* by `getB(w)`.

```
indx <- getu(w)$indx  #often w = x
Yc <- Y[indx]
Xc <- x[indx,]
#example
indx <- getu(buxx)$indx
Yc <- buxy[indx]
Xc <- buxx[indx,]
outr <- lsfit(Xc,Yc)
MLRplot(Xc,Yc) #right click Stop twice
```

Two special cases are listed below.

a) *Resistant additive error regression*: An additive error regression model has the form $Y = g(\boldsymbol{x}) + e$ where there is $m = 1$ response variable Y and the $p \times 1$ vector of predictors $\boldsymbol{x}$ is assumed to be known and independent of the additive error e. An enormous variety of regression models have this form, including multiple linear regression, nonlinear regression, nonparametric regression, partial least squares, lasso, ridge regression. As described above, find the RMVN set (or `covmb2` set) based on the $\boldsymbol{w}$ to obtain n_U cases $(Y_{ci}, \boldsymbol{x}_{ci})$, and then run the additive error regression method on the cleaned data.

b) *Resistant Additive Error Multivariate Regression*

Assume $\boldsymbol{y} = g(\boldsymbol{x}) + \boldsymbol{\epsilon} = E(\boldsymbol{y}|\boldsymbol{x}) + \boldsymbol{\epsilon}$ where $g : \mathbb{R}^p \to \mathbb{R}^m$, $\boldsymbol{y} = (Y_1, ..., Y_m)^T$, and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_m)^T$. Many models have this form, including multivariate linear regression, seemingly unrelated regressions, partial envelopes, partial least squares, and the models in a) with $m = 1$ response variable. Clean the data as in a), but let the cleaned data be stored in $(\boldsymbol{Z}_c, \boldsymbol{X}_c)$. Again, the theory of the method tends to apply to the method applied to the cleaned data since the response variables were not used to select the cases, but the efficiency is often much lower. In the *R* code below, assume the $\boldsymbol{y}$ are stored in z.

```
indx <- getu(w)$indx  #often w = x
Zc <- z[indx]
Xc <- x[indx,]
#example
ht <- buxy
t <- cbind(buxx,ht);
z <- t[,c(2,5)];
x <- t[,c(1,3,4)]
indx <- getu(x)$indx
Zc <- z[indx,]
Xc <- x[indx,]
mltreg(Xc,Zc) #right click Stop four times
```

ii) *Multivariate time series:* see Croux et al. (2010). Need to replace FMCD by RFCH or RMVN.

iii) *Partial least squares:* Cook et al. (2013) is a good reference for partial least squares and competing envelope estimators. Technique i) b) is applicable with the RMVN set if $n > 20(m + p)$. Since partial least squares is a competitor of multivariate linear regression, clean the data as in Section 12.6.2: Combine the nontrivial predictors and response variables into one vector per case and then get the RMVN set. Run the partial least squares method on the set. This method needs theory and $n > 20(m + p)$. Try the covmb2 set if p is large compared to n, e.g., $p > n$. Also see Hubert and Vanden Branden (2003) where FMCD should be replaced by RFCH or RMVN. The Hoffman et al. (2015) robust PLS–Lasso type method is likely superior.

```
t <- cbind(x,z) #could use t <- cbind(w,z) or t <- w
#if w are the continuous linearly related predictors
indx <- getu(t)$indx
Xc <- x[indx,]
Zc <- z[indx,]
#Then plug Xc and Zc in place of x and z into the
#PLS program if n > 20 (m + p).
```

iv) *Transforming Data Toward an elliptically contoured distribution:* See Cook and Nachtsheim (1994). The first step is to obtain a compact ellipsoidal region. Use RMVN or RFCH instead of FMVE to get this region.

14.2 Resistant Multiple Linear Regression

Consider the multiple linear regression model, written in matrix form as $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$. This model is a special case of the multivariate linear regression model with $m = 1$. (Ordinary) least squares (OLS) is the classical regression method. The OLS response and residual plots are very useful for detecting outliers and checking the model. **See Table 1.1 for acronyms.**

Resistant estimators are useful for detecting certain types of outliers. Some good resistant regression estimators are `rmreg2` from Section 12.6.2, the `hbreg` estimator from Section 14.4, and the Olive (2005a) MBA and trimmed views estimators described below.

Resistant estimators are often created by computing several trial fits $\boldsymbol{b}_i$ that are estimators of $\boldsymbol{\beta}$. Then a criterion is used to select the trial fit to be used in the resistant estimator. Suppose $c \approx n/2$. The LMS(c) criterion is $Q_{LMS}(\boldsymbol{b}) = r^2_{(c)}(\boldsymbol{b})$ where $r^2_{(1)} \leq \cdots \leq r^2_{(n)}$ are the ordered squared residuals, and the LTS(c) criterion is $Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^c r^2_{(i)}(\boldsymbol{b})$. The LTA($c$) criterion is $Q_{LTA}(\boldsymbol{b}) = \sum_{i=1}^c |r(\boldsymbol{b})|_{(i)}$ where $|r(\boldsymbol{b})|_{(i)}$ is the ith ordered absolute residual. Three impractical high breakdown robust estimators are the Hampel (1975) least median of squares (LMS) estimator, the Rousseeuw (1984) least trimmed sum of squares (LTS) estimator, and the Hössjer (1991) least trimmed sum of absolute deviations (LTA) estimator. Also see Hawkins and Olive (1999a, b). These estimators correspond to the $\hat{\boldsymbol{\beta}}_L \in \mathbb{R}^p$ that minimizes the corresponding criterion. LMS, LTA, and LTS have $O(n^p)$ or $O(n^{p+1})$ complexity. See Bernholt (2005), Hawkins and Olive (1999b), Klouda (2015), and Mount et al. (2014). Estimators with $O(n^4)$ or higher complexity take too long to compute. LTS and LTA are $\sqrt{n}$ consistent, while LMS has the lower $n^{1/3}$ rate. See Kim and Pollard (1990), Čížek (2006, 2008), and Mašíček (2004). If $c = n$, the LTS and LTA criteria are the OLS and L_1 criteria.

A good resistant estimator is the Olive (2005a) *median ball algorithm* (MBA or `mbareg`). The Euclidean distance of the ith vector of predictors $\boldsymbol{x}_i$ from the jth vector of predictors $\boldsymbol{x}_j$ is

$$D_i(\boldsymbol{x}_j) = D_i(\boldsymbol{x}_j, \boldsymbol{I}_p) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)}.$$

For a fixed $\boldsymbol{x}_j$, consider the ordered distances $D_{(1)}(\boldsymbol{x}_j), ..., D_{(n)}(\boldsymbol{x}_j)$. Next let $\hat{\boldsymbol{\beta}}_j(\alpha)$ denote the OLS fit to the $\min(p + 3 + \lfloor \alpha n/100 \rfloor, n)$ cases with the smallest distances where the approximate percentage of cases used is $\alpha \in \{1, 2.5, 5, 10, 20, 33, 50\}$. (Here $\lfloor x \rfloor$ is the greatest integer function so $\lfloor 7.7 \rfloor = 7$. The extra $p+3$ cases are added so that OLS can be computed for small n and α.) This yields seven OLS fits corresponding to the cases with predictors closest to $\boldsymbol{x}_j$. A fixed number of K cases are selected at random without replacement to use as the $\boldsymbol{x}_j$. Hence $7K$ OLS fits are generated. We use $K = 7$ as the default. A robust criterion Q is used to evaluate the $7K$ fits and the OLS fit to all of the data. Hence $7K + 1$ OLS fits are generated, and the MBA estimator is the fit that minimizes the criterion. The median squared residual is a good choice for Q.

Three ideas motivate this estimator. First, $\boldsymbol{x}$-outliers, which are outliers in the predictor space, tend to be much more destructive than Y-outliers which are outliers in the response variable. Suppose that the proportion of outliers is γ and that $\gamma < 0.5$. We would like the algorithm to have at least one "center" $\boldsymbol{x}_j$ that is not an outlier. The probability of drawing a center that is not an outlier is approximately $1 - \gamma^K > 0.99$ for $K \geq 7$, and this result is free of p. Second, by using the different percentages of coverages, for many data sets there will be a center and a coverage that contains no outliers. Third, the MBA estimator is a $\sqrt{n}$ consistent estimator of the same parameter vector $\boldsymbol{\beta}$ estimated by OLS under mild conditions on the zero mean error distribution.

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

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq D^2_{(j)}\}. \tag{14.1}$$

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. Ellipsoidal trimming differs from technique i) in Section 14.1 that uses the RMVN set on the $\boldsymbol{x}_i$, since the RMVN set uses a random amount of trimming. (The ellipsoidal trimming technique can also be used for other regression models, and the theory of the regression method tends to apply to the method applied to the cleaned data that was not trimmed since the response variables were not used to select the cases.)

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 $Y_i - \hat{Y}_i = r_i$. Note that the response and residual plots are made using all of the data, not just the cleaned data that was not trimmed.

The resistant trimmed views estimator combines ellipsoidal trimming and the response plot. First, compute $(T, \boldsymbol{C})$ on the $\boldsymbol{x}_i$, perhaps using the RMVN estimator. 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, pp. 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 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

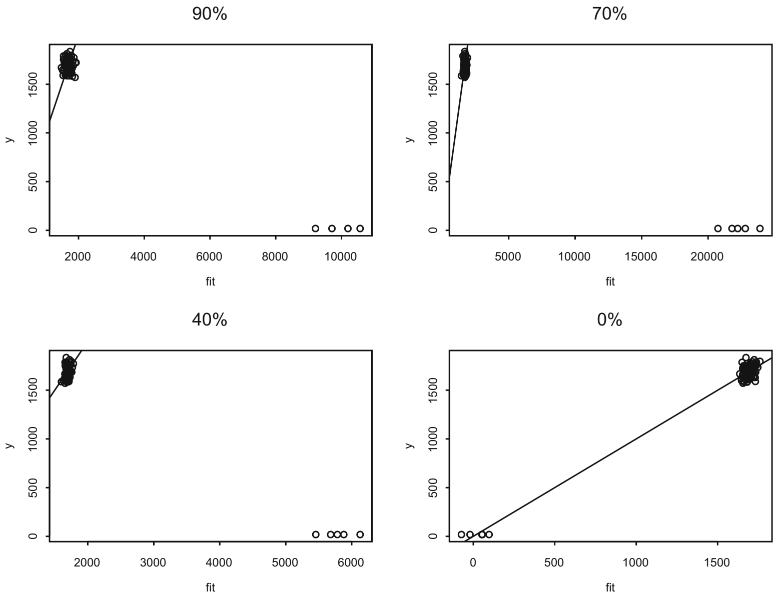

Fig. 14.1 4 Trimmed Views for the Buxton Data

$$\hat{\boldsymbol{\beta}}_{M,n} = (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1} \boldsymbol{X}_{M,n}^T \boldsymbol{Y}_{M,n} \sim N_p(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1})$$

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.

Warning: When $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e$, MLR estimators tend to estimate the same slopes $\beta_2, ..., \beta_p$, but the constant β_1 tends to depend on the estimator unless the errors are symmetric. The MBA and trimmed views estimators do estimate the same $\boldsymbol{\beta}$ as OLS asymptotically, but samples may need to be huge before the MBA and trimmed views estimates of the constant are close to the

OLS estimate of the constant. If the trimmed views estimator is modified so that the LTS, LTA, or LMS criterion is used to select the final estimator, then for the modified trimmed views estimator and the MBA estimator, it is likely that $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} \sum_{i=1}^{k} \pi_i N_p(\mathbf{0}, \sigma^2 \boldsymbol{W}_i)$ where $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^{k} \pi_i = 1$. The index i corresponds to the fits considered by the MBA estimator with $k = 7K + 1$ where often $K = 7$ so $k = 50$ or the modified trimmed views estimator with $k = 10$. For the MBA estimator and the modified trimmed views estimator, the prediction region method, described in Section 5.3, may be useful for testing hypotheses. Large sample sizes may be needed if the error distribution is not symmetric since the constant $\hat{\boldsymbol{\beta}}_1$ needs large samples. See Section 14.7 for an explanation for why large sample sizes may be needed to estimate the constant.

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.

The conditions under which the `rmreg2` estimator of Section 12.6.2 has been shown to be $\sqrt{n}$ consistent are quite strong, but it seems likely that the estimator is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under mild conditions where the parameter vector $\boldsymbol{\beta}$ is not, in general, the parameter vector estimated by OLS. For MLR, the *mpack* function `rmregboot` bootstraps the `rmreg2` estimator, and the function `rmregbootsim` can be used to simulate `rmreg2`. Both functions use the residual bootstrap where the residuals come from OLS. See the *R* code below.

```
out<-rmregboot(belx,bely)
plot(out$betas)
ddplot4(out$betas) #right click Stop

out<-rmregboot(cbrainx,cbrainy)
ddplot4(out$betas) #right click Stop
```

14.3 MLR Outlier Detection

For multiple linear regression, the OLS response and residual plots are very useful for detecting outliers. The DD plot of the continuous predictors is also useful. Use the *mpack* functions `MLRplot` and `ddplot4`.

Huber and Ronchetti (2009, p. 154) noted that efficient methods for identifying leverage groups are needed. Such groups are often difficult to detect with regression diagnostics and residuals, but often have outlying fitted values and responses that can be detected with response and residual plots. The following *rules of thumb* are useful for finding influential cases and outliers. Look for points with large absolute residuals and for points far away from $\overline{Y}$. Also look for gaps separating the data into clusters. The OLS fit often passes through a cluster of outliers, causing a large gap between a cluster corresponding to the bulk of the data and the cluster of outliers. When such a gap appears, it is possible that the smaller cluster corresponds to good leverage points: the cases follow the same model as the bulk of the data. To determine whether small clusters are outliers or good leverage points, give zero weight to the clusters, and fit an MLR estimator such as OLS to the bulk of the data. Denote the weighted estimator by $\hat{\boldsymbol{\beta}}_w$. Then plot $\hat{Y}_w$ versus Y using the entire data set. If the identity line passes through the cluster, then the cases in the cluster may be good leverage points; otherwise, they may be outliers. The trimmed views estimator of the previous section is also useful. Dragging the plots, so that they are roughly square, can be useful.

Definition 14.3. Suppose that some analysis to detect outliers is performed. *Masking* occurs if the analysis suggests that one or more outliers are in fact good cases. *Swamping* occurs if the analysis suggests that one or more good cases are outliers. Suppose that a subset of h cases is selected from the n cases making up the data set. Then the subset is *clean* if none of the h cases are outliers.

Influence diagnostics such as Cook's distances CD_i from Cook (1977) and the weighted Cook's distances WCD_i from Peña (2005) are sometimes useful. Although an index plot of Cook's distance CD_i may be useful for flagging influential cases, the index plot provides no direct way of judging the model against the data. As a remedy, cases in the response and residual plots with $CD_i > \min(0.5, 2p/n)$ are highlighted with open squares, and cases with $|WCD_i - \mathrm{median(WCD_i)}| > 4.5\mathrm{MAD(WCD_i)}$ are highlighted with crosses, where the median absolute deviation $\mathrm{MAD}(w_i) = \mathrm{median}(|w_i - \mathrm{median(w_i)}|)$.

Example 14.2. Figure 14.2 shows the response plot and residual plot for the Buxton (1920) data. Notice that the OLS fit passes through the outliers, but the response plot is resistant to Y-outliers since Y is on the vertical axis. Also notice that although the outlying cluster is far from $\overline{Y}$, only two of the outliers had large Cook's distance and only one case had a large WCD_i. Hence *masking* occurred for the Cook's distances, the WCD_i, and for the OLS residuals, but not for the OLS fitted values. Plots using `lmsreg` and `ltsreg` were similar, but trimmed views and MBA were effective. See Figures 14.1 and 14.6. Figure 14.2 was made with the following R commands.

```
source("G:/mpack.txt"); source("G:/mrobdata.txt")
mlrplot4(buxx,buxy) #right click Stop twice
```

High leverage outliers are a particular challenge to conventional numerical MLR diagnostics such as Cook's distance, but can often be visualized using the response and residual plots. (Using the trimmed views of Section 14.2 is also effective for detecting outliers and other departures from the MLR model.)

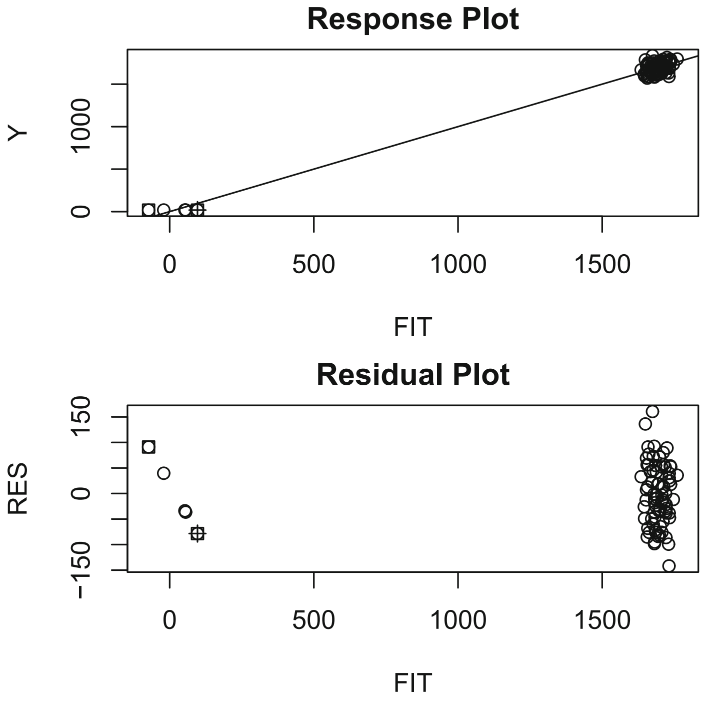

Fig. 14.2 Plots for Buxton Data

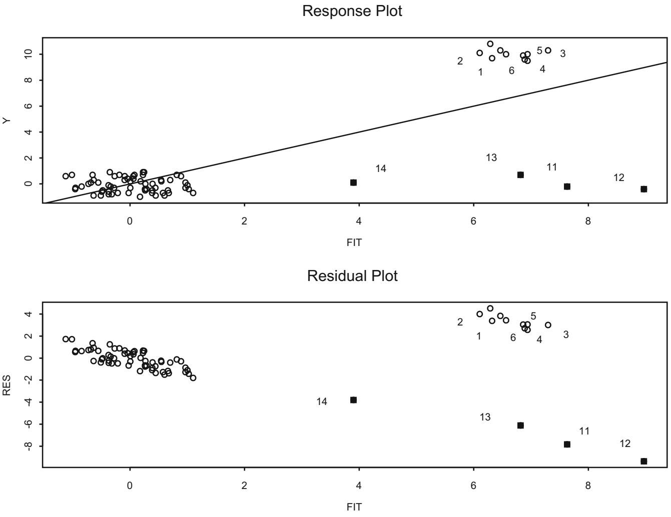

Fig. 14.3 Plots for HBK Data

Example 14.3. Hawkins et al. (1984) gave a well-known artificial data set where the first 10 cases are outliers, while cases 11–14 are good leverage points. Figure 14.3 shows the residual and response plots based on the OLS estimator. The highlighted cases have Cook's distance $> \min(0.5, 2p/n)$, and the identity line is shown in the response plot. Since the good cases 11–14 have the largest Cook's distances and absolute OLS residuals, *swamping* has occurred. (Masking has also occurred since the outliers have small Cook's distances, and some of the outliers have smaller OLS residuals than clean cases.) To determine whether both clusters are outliers or if one cluster consists of good leverage points, cases in both clusters could be given weight zero and the resulting response plot created. (Alternatively, response plots based on the `tvreg` estimator of Section 14.2 could be made where the cases with weight one are highlighted. For high levels of trimming, the identity line often passes through the good leverage points.)

The above example is typical of many "benchmark" outlier data sets for MLR. In these data sets, traditional OLS diagnostics such as Cook's distance and the residuals often fail to detect the outliers, but the combination of the response plot and residual plot is usually able to detect the outliers. The CD_i and WCD_i are the most effective when there is a single cluster about the identity line. If there is a second cluster of outliers or good leverage points or if there is nonconstant variance, then these numerical diagnostics tend to fail.

Often, practical "robust estimators" generate a sequence of K trial fits called *attractors*: $\boldsymbol{b}_1, ..., \boldsymbol{b}_K$. Then some criterion is evaluated, and the attractor $\boldsymbol{b}_A$ that minimizes the criterion is used in the final estimator. One way to obtain attractors is to generate trial fits called *starts* and then use the *concentration* technique. Let $\boldsymbol{b}_{0,j}$ be the jth start, and compute all n residuals $r_i(\boldsymbol{b}_{0,j}) = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}_{0,j}$. At the next iteration, the OLS estimator $\boldsymbol{b}_{1,j}$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest squared residuals $r_i^2(\boldsymbol{b}_{0,j})$. This iteration can be continued for k steps resulting in the sequence of estimators $\boldsymbol{b}_{0,j}, \boldsymbol{b}_{1,j}, ..., \boldsymbol{b}_{k,j}$. Then $\boldsymbol{b}_{k,j}$ is the jth attractor for $j = 1, ..., K$. Using $k = 10$ concentration steps often works well, and the basic resampling algorithm is a special case with $k = 0$; i.e., the attractors are the starts. Elemental starts are the fits from randomly selected "elemental sets" of p cases. Such an algorithm is called a CLTS concentration algorithm or CLTS.

A CLTA concentration algorithm would replace the OLS estimator by the L_1 estimator and the smallest c_n squared residuals by the smallest c_n absolute residuals. Many other variants are possible, but obtaining theoretical results may be difficult.

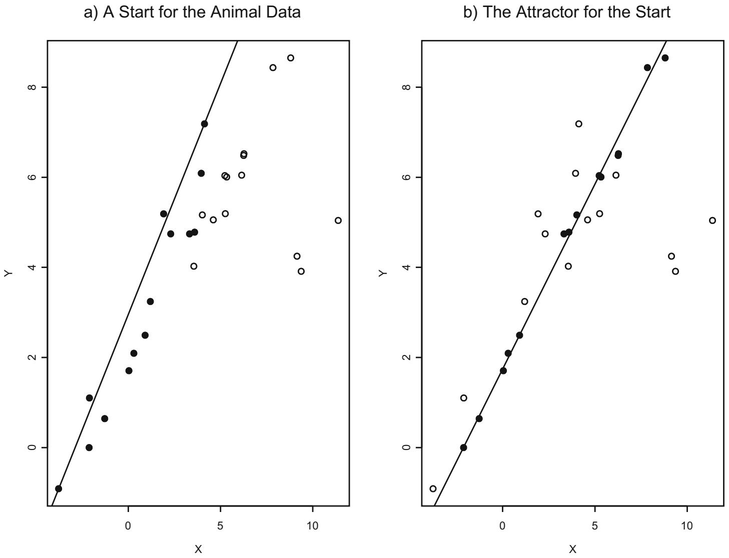

Fig. 14.4 The Highlighted Points are More Concentrated about the Attractor

Example 14.4. As an illustration of the CLTA concentration algorithm, consider the animal data from Rousseeuw and Leroy (1987, p. 57). The response Y is the *log brain weight*, and the predictor x is the *log body weight* for 25 mammals and 3 dinosaurs (outliers with the highest body weight).

Suppose that the first elemental start uses cases 20 and 14, corresponding to mouse and man. Then the start $\boldsymbol{b}_{s,1} = \boldsymbol{b}_{0,1} = (2.952, 1.025)^T$ and the sum of the $c = 14$ smallest absolute residuals $\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{0,1}) = 12.101$. Figure 14.4a shows the scatterplot of x and y. The start is also shown, and the 14 cases corresponding to the smallest absolute residuals are highlighted. The L_1 fit to these c highlighted cases is $\boldsymbol{b}_{1,1} = (2.076, 0.979)^T$ and $\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{1,1}) = 6.990$. The iteration consists of finding the cases corresponding to the c smallest absolute residuals, obtaining the corresponding L_1 fit and repeating. The attractor $\boldsymbol{b}_{a,1} = \boldsymbol{b}_{7,1} = (1.741, 0.821)^T$ and the LTA(c) criterion evaluated at the attractor is $\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{a,1}) = 2.172$. Figure 14.4b shows the attractor and that the c highlighted cases corresponding to the smallest absolute residuals are much more concentrated than those in Figure 14.4a. Figure 14.5a shows 5 randomly selected starts, while Figure 14.5b shows the corresponding attractors.

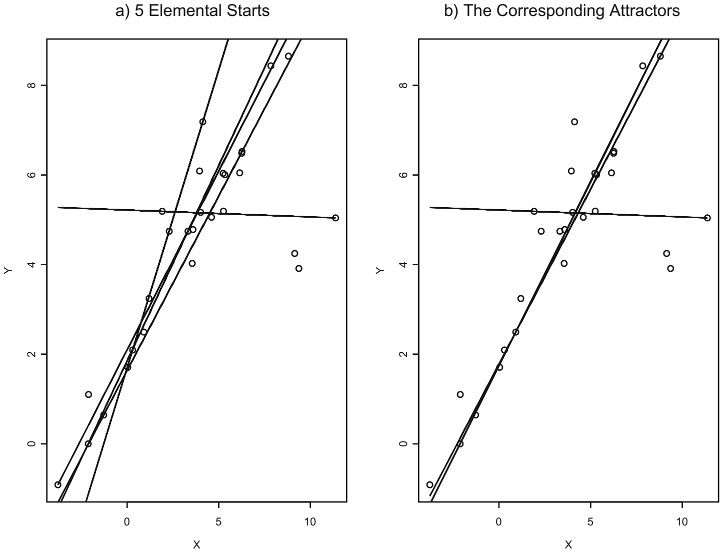

Fig. 14.5 Starts and Attractors for the Animal Data

Notice that the elemental starts have more variability than the attractors, but if the start passes through an outlier, so does the attractor.

Suppose the data set has n cases where d are outliers and $n - d$ are "clean" (not outliers). The outlier proportion $\gamma = d/n$. Suppose that K elemental sets are chosen with replacement and that it is desired to find K such that the probability P(that at least one of the elemental sets is clean) $\equiv P_1 \approx 1 - \alpha$ where $\alpha = 0.05$ is a common choice. Then $P_1 = 1-$ P(none of the K elemental sets is clean) $\approx 1 - [1 - (1 - \gamma)^p]^K$ by independence. Hence $\alpha \approx [1 - (1 - \gamma)^p]^K$ or

$$K \approx \frac{\log(\alpha)}{\log([1 - (1 - \gamma)^p])} \approx \frac{\log(\alpha)}{-(1 - \gamma)^p} \tag{14.2}$$

using the approximation $\log(1 - x) \approx -x$ for small x. Since $\log(0.05) \approx -3$, if $\alpha = 0.05$, then $K \approx \dfrac{3}{(1 - \gamma)^p}$. Frequently a clean subset is wanted even if the contamination proportion $\gamma \approx 0.5$. Then for a 95% chance of obtaining at least one clean elemental set, $K \approx 3\ (2^p)$ elemental sets need to be drawn. If the start passes through an outlier, so does the attractor. For concentration algorithms for multivariate location and dispersion, if the start passes through a cluster of outliers, sometimes the attractor would be clean.

Notice that the number of subsets K needed to obtain a clean elemental set with high probability is an exponential function of the number of predictors p but is free of n. Hawkins and Olive (2002) showed that if K is fixed and free of n, then the resulting elemental or concentration algorithm (that uses k concentration steps) is inconsistent and zero breakdown. See Theorem P.1 in the preface. Nevertheless, many practical estimators tend to use a value of

Table 14.1 Largest p for a 95% Chance of a Clean Subsample

				K				
γ	500	3000	10000	10^5	10^6	10^7	10^8	10^9
0.01	509	687	807	1036	1265	1494	1723	1952
0.05	99	134	158	203	247	292	337	382
0.10	48	65	76	98	120	142	164	186
0.15	31	42	49	64	78	92	106	120
0.20	22	30	36	46	56	67	77	87
0.25	17	24	28	36	44	52	60	68
0.30	14	19	22	29	35	42	48	55
0.35	11	16	18	24	29	34	40	45
0.40	10	13	15	20	24	29	33	38
0.45	8	11	13	17	21	25	28	32
0.50	7	9	11	15	18	21	24	28

K that is free of both n and p (e.g., $K = 500$ or $K = 3000$). Such algorithms include ALMS = FLMS = `lmsreg` and ALTS = FLTS = `ltsreg`. The "A" denotes that an algorithm was used. The "F" means that a fixed number of trial fits (K elemental fits) was used and the criterion (LMS or LTS) was used to select the trial fit used in the final estimator.

To examine the outlier resistance of such inconsistent zero breakdown estimators, fix both K and the contamination proportion γ and then find the largest number of predictors p that can be in the model such that the probability of finding at least one clean elemental set is high. Given K and γ, P(at least one of K subsamples is clean) $= 0.95 \approx 1 - [1 - (1-\gamma)^p]^K$. Thus the largest value of p satisfies $\dfrac{3}{(1-\gamma)^p} \approx K$, or

$$p \approx \left\lfloor \frac{\log(3/K)}{\log(1-\gamma)} \right\rfloor \tag{14.3}$$

if the sample size n is very large. Again, $\lfloor x \rfloor$ is the greatest integer function: $\lfloor 7.7 \rfloor = 7$.

Table 14.1 shows the largest value of p such that there is a 95% chance that at least one of K subsamples is clean using the approximation given by Equation (14.3). Hence if $p = 28$, even with one billion subsamples, there is a 5% chance that none of the subsamples will be clean if the contamination proportion $\gamma = 0.5$. Since clean elemental fits have great variability, an algorithm needs to produce many clean fits in order for the best fit to be good. When contamination is present, all K elemental sets could contain outliers. Hence basic resampling and concentration algorithms that only use K elemental starts are doomed to fail if γ and p are large.

The outlier resistance of elemental algorithms that use K elemental sets decreases rapidly as p increases. However, for $p < 10$, such elemental algorithms are often useful for outlier detection. They can perform better than MBA, trimmed views, and `rmreg2` if p is small and the outliers are close to the bulk of the data or if p is small and there is a mixture distribution: the bulk of the data follows one MLR model, but "outliers" and some of the clean data are fit well by another MLR model. For example, if there is one nontrivial predictor, suppose the plot of x versus Y looks like the letter X. Such a mixture distribution is not really an outlier configuration since outliers lie far from the bulk of the data. All practical estimators have outlier configurations where they perform poorly. If p is small, elemental algorithms tend to have trouble when there is a weak regression relationship for the bulk of the data and a cluster of outliers that are not good leverage points (do not fall near the hyperplane followed by the bulk of the data). The Buxton (1920) data set is an example.

Proposition 14.1. Let $h = p$ be the number of randomly selected cases in an elemental set, and let γ_o be the highest percentage of massive outliers that a resampling algorithm can detect reliably. If n is large, then

$$\gamma_o \approx \min\left(\frac{n-c}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h}\right) 100\%. \tag{14.4}$$

Proof. As in Remark 4.1, if the contamination proportion γ is fixed, then the probability of obtaining at least one clean subset of size h with high probability (say $1 - \alpha = 0.8$) is given by $0.8 = 1 - [1 - (1 - \gamma)^h]^K$. Fix the number of starts K, and solve this equation for γ. □

The value of γ_o depends on $c \geq n/2$ and h. To maximize γ_o, take $c \approx n/2$ and $h = p$. For example, with $K = 500$ starts, $n > 100$, and $h = p \leq 20$, the resampling algorithm should be able to detect up to 24% outliers provided every clean start is able to at least partially separate inliers (clean cases) from outliers. However, if $h = p = 50$, this proportion drops to 11%.

Definition 14.4. Let $\boldsymbol{b}_1, ..., \boldsymbol{b}_J$ be J estimators of $\boldsymbol{\beta}$. Assume that $J \geq 2$ and that OLS is included. A *fit-fit* (FF) plot is a scatterplot matrix of the fitted values $\widehat{Y}(\boldsymbol{b}_1), ..., \widehat{Y}(\boldsymbol{b}_J)$. Often Y is also included in the top or bottom row of the FF plot to see the response plots. A *residual-residual* (RR) plot is a scatterplot matrix of the residuals $r(\boldsymbol{b}_1), ..., r(\boldsymbol{b}_J)$. Often $\hat{Y}$ is also included in the top or bottom row of the RR plot to see the residual plots.

If the multiple linear regression model holds, if the predictors are bounded, and if all J regression estimators are consistent estimators of $\boldsymbol{\beta}$, then the subplots in the FF and RR plots should be linear with a correlation tending to one as the sample size n increases. To prove this claim, let the ith residual from the jth fit $\boldsymbol{b}_j$ be $r_i(\boldsymbol{b}_j) = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}_j$ where $(Y_i, \boldsymbol{x}_i^T)$ is the ith observation. Similarly, let the ith fitted value from the jth fit be $\widehat{Y}_i(\boldsymbol{b}_j) = \boldsymbol{x}_i^T \boldsymbol{b}_j$. Then

$$\|r_i(\boldsymbol{b}_1) - r_i(\boldsymbol{b}_2)\| = \|\widehat{Y}_i(\boldsymbol{b}_1) - \widehat{Y}_i(\boldsymbol{b}_2)\| = \|\boldsymbol{x}_i^T(\boldsymbol{b}_1 - \boldsymbol{b}_2)\|$$

$$\leq \|\boldsymbol{x}_i\| \; (\|\boldsymbol{b}_1 - \boldsymbol{\beta}\| + \|\boldsymbol{b}_2 - \boldsymbol{\beta}\|). \tag{14.5}$$

The FF plot is a powerful way for comparing fits. The commonly suggested alternative is to look at a table of the estimated coefficients, but coefficients can differ greatly while yielding similar fits if some of the predictors are highly correlated or if several of the predictors are independent of the response.

The *mpack* functions ffplot2 and rrplot2 make FF and RR plots using OLS, ALMS from lmsreg, ALTS from ltsreg, mbareg, an outlier detector mbalata, BB, and rmreg2 described in Section 12.6.2. OLS, BB, and mbareg are the three trial fits used by the default version of the $\sqrt{n}$ consistent high breakdown hbreg estimator. See Section 14.4.2. The top row of ffplot2 shows the response plots. The *R* code below is useful and shows how to get some of the text's data sets into *R*.

```
library(MASS)
rrplot2(buxx,buxy)
ffplot2(buxx,buxy)
#The following three data sets can be obtained with
#the source("G:/mrobdata.txt") command
#if the data file is on flash drive G.
rmreg2(buxx,buxy)       #right click Stop twice
rmreg2(cbrainx,cbrainy)
rmreg2(gladox,gladoy)

hbk <- matrix(scan(),nrow=75,ncol=5,byrow=T)
hbk <- hbk[,-1]
rmreg2(hbk[,1:3],hbk[,4]) #Outliers are clear
#but fit avoids good leverage points.

nasty <- matrix(scan(),nrow=32,ncol=6,byrow=T)
nasty <- nasty[,-1]
rmreg2(nasty[,1:4],nasty[,5])

wood <- matrix(scan(),nrow=20,ncol=7,byrow=T)
wood <- wood[,-1]
rmreg2(wood[,1:5],wood[,6]) #failed to find
#the outliers

major <- matrix(scan(),nrow=112,ncol=7,byrow=T)
major <- major[,-1]
rmreg2(major[,1:5],major[,6])
```

Example 14.2**, continued.** The RR and FF plots for the Buxton (1920) data are shown in Figures 14.6 and 14.7. Note that only the last four estimators give large absolute residuals to the outliers. The top row of Figure 14.6 gives the response plots for the estimators. If there are two clusters, one

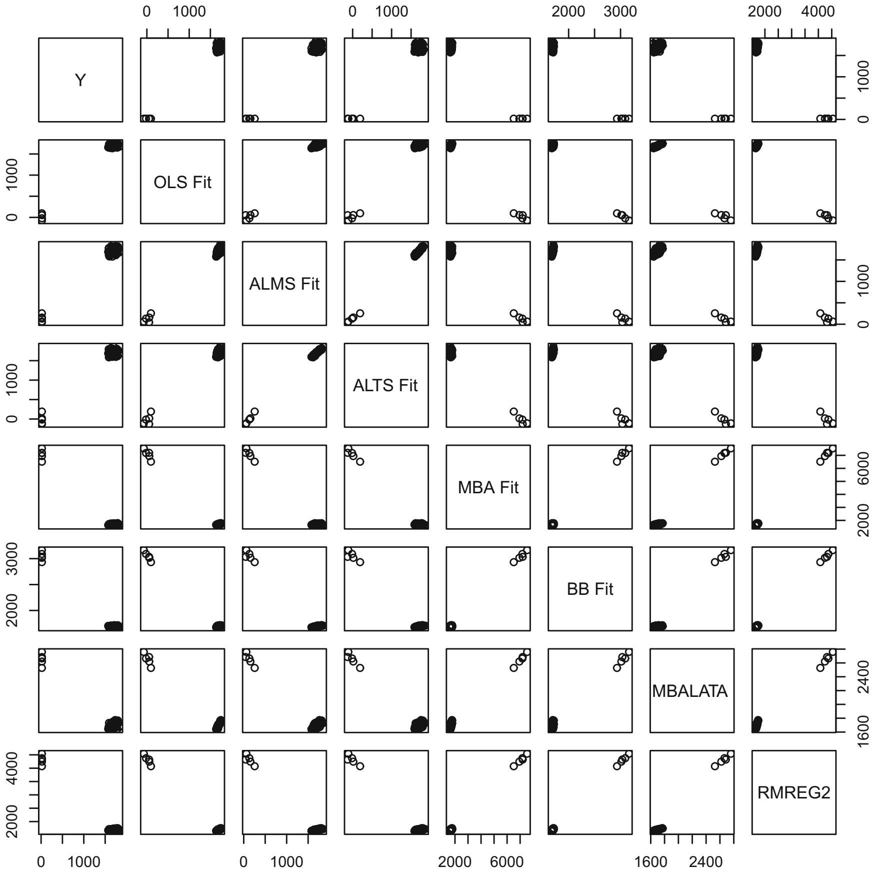

Fig. 14.6 FF Plots for Buxton Data

in the upper right and one in the lower left of the response plot, then the identity line goes through both clusters. Hence the fit passes through the outliers. One feature of the MBA estimator is that it depends on the sample of 7 centers drawn and changes each time the function is called. In ten runs, about seven plots will look like Figures 14.6 and 14.7, but in about three plots the MBA estimator will also pass through the outliers.

Table 14.2 compares the TV, MBA (for MLR), `lmsreg`, `ltsreg`, L_1, and OLS estimators on seven data sets available from the text's website. The column headers give the file name, while the remaining rows of the table give

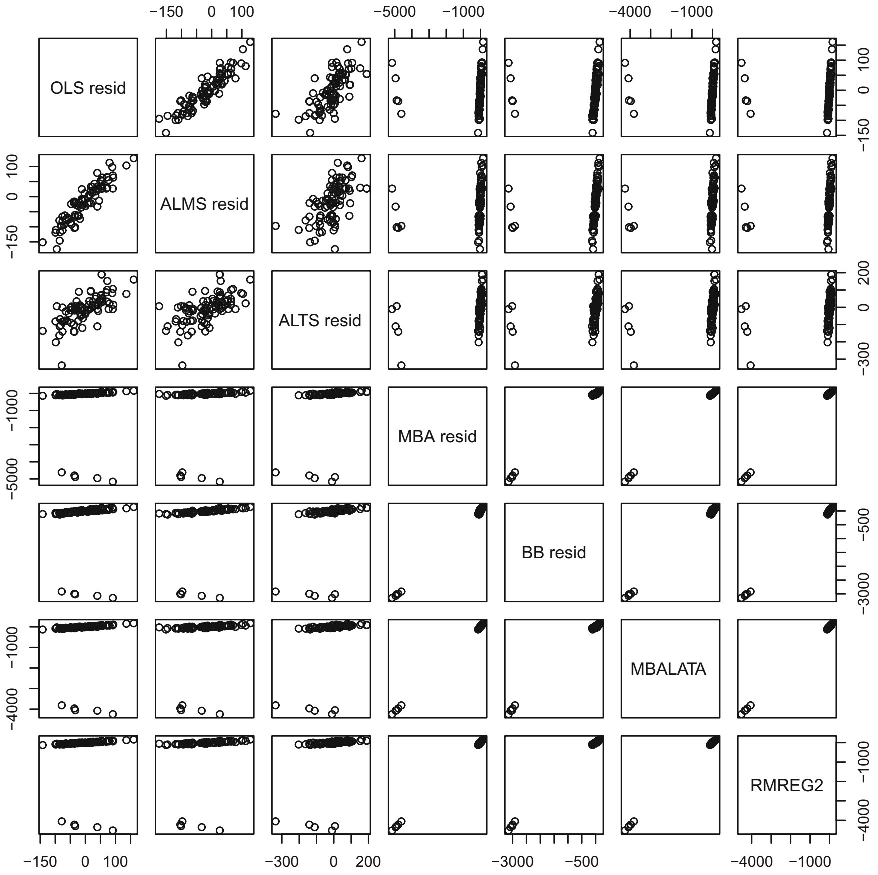

Fig. 14.7 RR Plots for Buxton Data

Table 14.2 Summaries for Seven Data Sets, the Correlations of the Residuals from TV(M), and the Alternative Method are Given in the First Five 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	115	1-10	3,44	2,6,...,30	4,6,8,19
n	87	267	267	75	112	32	20
p	5	7	7	4	6	5	6
M	70	0	30	90	0	90	20

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 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. Each model included a constant.

Notice that the TV, MBA, and OLS estimators were the same for the Gladstone (1905) data and for the Tremearne (1911) *major* data 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 115, 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` detected the outlier, but `ltsreg` did not. These two functions change often, not always for the better.

14.4 Robust Regression

This section will consider the breakdown of a regression estimator and then develop the practical high breakdown `hbreg` estimator.

14.4.1 MLR Breakdown and Equivariance

Breakdown and equivariance properties have received considerable attention in the literature. Several of these properties involve transformations of the data and are discussed below. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the original data, then the vector of the coefficient estimates is

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = T(\boldsymbol{X}, \boldsymbol{Y}), \tag{14.6}$$

the vector of predicted values is

$$\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}), \tag{14.7}$$

and the vector of residuals is

$$\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}. \tag{14.8}$$

If the design matrix $\boldsymbol{X}$ is transformed into $\boldsymbol{W}$ and the vector of dependent variables $\boldsymbol{Y}$ is transformed into $\boldsymbol{Z}$, then $(\boldsymbol{W}, \boldsymbol{Z})$ is the new data set.

Definition 14.5. Regression Equivariance: Let $\boldsymbol{u}$ be any $p \times 1$ vector. Then $\widehat{\boldsymbol{\beta}}$ is regression equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u}. \quad (14.9)$$

Hence if $\boldsymbol{W} = \boldsymbol{X}$ and $\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}$, then $\widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{Y}} + \boldsymbol{X}\boldsymbol{u}$ and $\boldsymbol{r}(\boldsymbol{W}, \boldsymbol{Z}) = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. Note that the residuals are invariant under this type of transformation, and note that if $\boldsymbol{u} = -\widehat{\boldsymbol{\beta}}$, then regression equivariance implies that we should not find any linear structure if we regress the residuals on $\boldsymbol{X}$. Also see Problem 14.3.

Definition 14.6. Scale Equivariance: Let c be any scalar. Then $\widehat{\boldsymbol{\beta}}$ is scale equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, c\boldsymbol{Y}) = T(\boldsymbol{X}, c\boldsymbol{Y}) = cT(\boldsymbol{X}, \boldsymbol{Y}) = c\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \quad (14.10)$$

Hence if $\boldsymbol{W} = \boldsymbol{X}$ and $\boldsymbol{Z} = c\boldsymbol{Y}$, then $\widehat{\boldsymbol{Z}} = c\widehat{\boldsymbol{Y}}$ and $\boldsymbol{r}(\boldsymbol{X}, c\boldsymbol{Y}) = c\,\boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. Scale equivariance implies that if the Y_i's are stretched, then the fits and the residuals should be stretched by the same factor.

Definition 14.7. Affine Equivariance: Let $\boldsymbol{A}$ be any $p \times p$ nonsingular matrix. Then $\widehat{\boldsymbol{\beta}}$ is affine equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = T(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}T(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \quad (14.11)$$

Hence if $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{A}$ and $\boldsymbol{Z} = \boldsymbol{Y}$, then $\widehat{\boldsymbol{Z}} = \boldsymbol{W}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{Y}}$, and $\boldsymbol{r}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. Note that both the predicted values and the residuals are invariant under an affine transformation of the predictor variables.

Definition 14.8. Permutation Invariance: Let $\boldsymbol{P}$ be an $n \times n$ permutation matrix. Then $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is an $n \times n$ identity matrix and the superscript T denotes the transpose of a matrix. Then $\widehat{\boldsymbol{\beta}}$ is permutation invariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) = T(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) = T(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \quad (14.12)$$

Hence if $\boldsymbol{W} = \boldsymbol{PX}$ and $\boldsymbol{Z} = \boldsymbol{PY}$, then $\widehat{\boldsymbol{Z}} = \boldsymbol{P}\widehat{\boldsymbol{Y}}$ and $\boldsymbol{r}(\boldsymbol{PX}, \boldsymbol{PY}) = \boldsymbol{P}\ \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. If an estimator is not permutation invariant, then swapping rows of the $n \times (p+1)$ augmented matrix $(\boldsymbol{X}, \boldsymbol{Y})$ will change the estimator. Hence the case number is important. If the estimator is permutation invariant, then the position of the case in the data cloud is of primary importance. Resampling algorithms are not permutation invariant because permuting the data causes different subsamples to be drawn.

The remainder of this subsection gives a standard definition of breakdown and then shows that if the median absolute residual is bounded in the presence of high contamination, then the regression estimator has a high breakdown value. The following notation will be useful. Let $\boldsymbol{W}$ denote the data matrix where the ith row corresponds to the ith case. For regression, $\boldsymbol{W}$ is the $n \times (p+1)$ matrix with ith row $(\boldsymbol{x}_i^T, Y_i)$. Let $\boldsymbol{W}_d^n$ denote the data matrix where any d_n of the cases have been replaced by arbitrarily bad contaminated cases. Then the contamination fraction is $\gamma \equiv \gamma_n = d_n/n$, and the breakdown value of $\hat{\boldsymbol{\beta}}$ is the smallest value of γ_n needed to make $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large.

Definition 14.9. Let $1 \le d_n \le n$. If $T(\boldsymbol{W})$ is a $p \times 1$ vector of regression coefficients, then the *breakdown value* of T is

$$B(T, \boldsymbol{W}) = \min \left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty \right\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$.

Definition 14.10. *High breakdown* regression estimators have $\gamma_n \to 0.5$ as $n \to \infty$ if the clean (uncontaminated) data are in *general position*: any p clean cases give a unique estimate of $\boldsymbol{\beta}$. Estimators are *zero breakdown* if $\gamma_n \to 0$ and *positive breakdown* if $\gamma_n \to \gamma > 0$ as $n \to \infty$.

The following result greatly simplifies some breakdown proofs and shows that a regression estimator basically breaks down if the median absolute residual MED($|r_i|$) can be made arbitrarily large. The result implies that if the breakdown value ≤ 0.5, breakdown can be computed using the median absolute residual MED($|r_i|(\boldsymbol{W}_d^n)$) instead of $\|T(\boldsymbol{W}_d^n)\|$. Similarly, $\hat{\boldsymbol{\beta}}$ is high breakdown if the median squared residual or the c_nth largest absolute residual $|r_i|_{(c_n)}$ or squared residual $r^2_{(c_n)}$ stay bounded under high contamination where $c_n \approx n/2$. Note that $\|\hat{\boldsymbol{\beta}}\| \equiv \|\hat{\boldsymbol{\beta}}(\boldsymbol{W}_d^n)\| \le M$ for some constant M that depends on T and $\boldsymbol{W}$ but not on the outliers if the number of outliers d_n is less than the smallest number of outliers needed to cause breakdown.

Proposition 14.2. If the breakdown value ≤ 0.5, computing the breakdown value using the median absolute residual MED$(|r_i|(\boldsymbol{W}_d^n))$ instead of $\|T(\boldsymbol{W}_d^n)\|$ is asymptotically equivalent to using Definition 14.9.

Proof. Consider any contaminated data set $\boldsymbol{W}_d^n$ with ith row $(\boldsymbol{w}_i^T, Z_i)^T$. If the regression estimator $T(\boldsymbol{W}_d^n) = \hat{\boldsymbol{\beta}}$ satisfies $\|\hat{\boldsymbol{\beta}}\| \leq M$ for some constant M if $d < d_n$, then the median absolute residual MED$(|Z_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{w}_i|)$ is bounded by $\max_{i=1,\ldots,n} |Y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i| \leq \max_{i=1,\ldots,n}[|Y_i| + \sum_{j=1}^p M|x_{i,j}|]$ if $d_n < n/2$.

If the median absolute residual is bounded by M when $d < d_n$, then $\|\hat{\boldsymbol{\beta}}\|$ is bounded, provided fewer than half of the cases line on the hyperplane (and so have absolute residual of 0), as shown next. Now suppose that $\|\hat{\boldsymbol{\beta}}\| = \infty$. Since the absolute residual is the vertical distance of the observation from the hyperplane, the absolute residual $|r_i| = 0$ if the ith case lies on the regression hyperplane, but $|r_i| = \infty$ otherwise. Hence MED$(|r_i|) = \infty$ if fewer than half of the cases lie on the regression hyperplane. This will occur unless the proportion of outliers $d_n/n > (n/2 - q)/n \rightarrow 0.5$ as $n \rightarrow \infty$ where q is the number of "good" cases that lie on a hyperplane of lower dimension than p. In the literature, it is usually assumed that the original data are in *general position*: $q = p - 1$. □

Suppose that the clean data are in general position and that the number of outliers is less than the number needed to make the median absolute residual and $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large. If the $\boldsymbol{x}_i$ are fixed, and the outliers are moved up and down by adding a large positive or negative constant to the Y values of the outliers, then for high breakdown (HB) estimators, $\hat{\boldsymbol{\beta}}$ and MED$(|r_i|)$ stay bounded where the bounds depend on the clean data $\boldsymbol{W}$ but not on the outliers even if the number of outliers is nearly as large as $n/2$. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large.

If the Y_i's are fixed, arbitrarily large $\boldsymbol{x}$-outliers tend to drive the slope estimates to 0, not ∞. If both $\boldsymbol{x}$ and Y can be varied, then a cluster of outliers can be moved arbitrarily far from the bulk of the data but may still have small residuals. For example, move the outliers along the regression hyperplane formed by the clean cases.

If the $(\boldsymbol{x}_i^T, Y_i)$ are in general position, then the contamination could be such that $\hat{\boldsymbol{\beta}}$ passes exactly through $p-1$ "clean" cases and d_n "contaminated" cases. Hence $d_n + p - 1$ cases could have absolute residuals equal to zero with $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large (but finite). Nevertheless, if T possesses reasonable equivariant properties and $\|T(\boldsymbol{W}_d^n)\|$ is replaced by the median absolute residual in the definition of breakdown, then the two breakdown values are asymptotically equivalent. (If $T(\boldsymbol{W}) \equiv \boldsymbol{0}$, then T is neither regression nor affine equivariant. The breakdown value of T is one, but the median absolute residual can be made arbitrarily large if the contamination proportion is greater than $n/2$.)

If the Y_i's are fixed, arbitrarily large $\boldsymbol{x}$-outliers will rarely drive $\|\hat{\boldsymbol{\beta}}\|$ to ∞. The $\boldsymbol{x}$-outliers can drive $\|\hat{\boldsymbol{\beta}}\|$ to ∞ if they can be constructed so that the estimator is no longer defined, e.g., so that $\boldsymbol{X}^T\boldsymbol{X}$ is nearly singular. The examples following some results on norms may help illustrate these points.

Definition 14.11. Let $\boldsymbol{y}$ be an $n \times 1$ vector. Then $\|\boldsymbol{y}\|$ is a *vector norm* if
vn1) $\|\boldsymbol{y}\| \geq 0$ for every $\boldsymbol{y} \in \mathbb{R}^n$ with equality iff $\boldsymbol{y}$ is the zero vector,
vn2) $\|a\boldsymbol{y}\| = |a| \; \|\boldsymbol{y}\|$ for all $\boldsymbol{y} \in \mathbb{R}^n$ and for all scalars a, and
vn3) $\|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$.

Definition 14.12. Let $\boldsymbol{G}$ be an $n \times p$ matrix. Then $\|\boldsymbol{G}\|$ is a *matrix norm* if
mn1) $\|\boldsymbol{G}\| \geq 0$ for every $n \times p$ matrix $\boldsymbol{G}$ with equality iff $\boldsymbol{G}$ is the zero matrix,
mn2) $\|a\boldsymbol{G}\| = |a| \; \|\boldsymbol{G}\|$ for all scalars a, and
mn3) $\|\boldsymbol{G}+\boldsymbol{H}\| \leq \|\boldsymbol{G}\| + \|\boldsymbol{H}\|$ for all $n \times p$ matrices $\boldsymbol{G}$ and $\boldsymbol{H}$.

Example 14.5. The *q-norm* of a vector $\boldsymbol{y}$ is $\|\boldsymbol{y}\|_q = (|y_1|^q + \cdots + |y_n|^q)^{1/q}$. In particular, $\|\boldsymbol{y}\|_1 = |y_1| + \cdots + |y_n|$, the *Euclidean norm* $\|\boldsymbol{y}\|_2 = \sqrt{y_1^2 + \cdots + y_n^2}$, and $\|\boldsymbol{y}\|_\infty = \max_i \; |y_i|$. Given a matrix $\boldsymbol{G}$ and a vector norm $\|\boldsymbol{y}\|_q$, the *q-norm* or *subordinate matrix norm* of matrix $\boldsymbol{G}$ is $\|\boldsymbol{G}\|_q = \max\limits_{\boldsymbol{y} \neq \boldsymbol{0}} \dfrac{\|\boldsymbol{G}\boldsymbol{y}\|_q}{\|\boldsymbol{y}\|_q}$. It can be shown that the *maximum column sum norm* $\|\boldsymbol{G}\|_1 = \max\limits_{1 \leq j \leq p} \sum\limits_{i=1}^{n} |g_{ij}|$, the *maximum row sum norm* $\|\boldsymbol{G}\|_\infty = \max\limits_{1 \leq i \leq n} \sum\limits_{j=1}^{p} |g_{ij}|$, and the *spectral norm* $\|\boldsymbol{G}\|_2 = \sqrt{\text{maximum eigenvalue of } \boldsymbol{G}^T\boldsymbol{G}}$. The *Frobenius norm*

$$\|\boldsymbol{G}\|_F = \sqrt{\sum_{j=1}^{p}\sum_{i=1}^{n} |g_{ij}|^2} = \sqrt{\text{trace}(\boldsymbol{G}^T\boldsymbol{G})}.$$

Several useful results involving matrix norms will be used. First, for any subordinate matrix norm, $\|\boldsymbol{G}\boldsymbol{y}\|_q \leq \|\boldsymbol{G}\|_q \; \|\boldsymbol{y}\|_q$. Let $J = J_m = \{m_1, ..., m_p\}$ denote the p cases in the mth elemental fit $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$. Then for any elemental fit $\boldsymbol{b}_J$ (suppressing $q = 2$),

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}(\boldsymbol{X}_J\boldsymbol{\beta} + \boldsymbol{e}_J) - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}\boldsymbol{e}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \; \|\boldsymbol{e}_J\|. \quad (14.13)$$

The following results (Golub and Loan 1989, pp. 57, 80) on the Euclidean norm are useful. Let $0 \le \sigma_p \le \sigma_{p-1} \le \cdots \le \sigma_1$ denote the singular values of $\boldsymbol{X}_J = (x_{mi,j})$. Then

$$\|\boldsymbol{X}_J^{-1}\| = \frac{\sigma_1}{\sigma_p \|\boldsymbol{X}_J\|}, \tag{14.14}$$

$$\max_{i,j} |x_{mi,j}| \le \|\boldsymbol{X}_J\| \le p \ \max_{i,j} |x_{mi,j}|, \text{ and} \tag{14.15}$$

$$\frac{1}{p \ \max_{i,j} |x_{mi,j}|} \le \frac{1}{\|\boldsymbol{X}_J\|} \le \|\boldsymbol{X}_J^{-1}\|. \tag{14.16}$$

From now on, unless otherwise stated, we will use the spectral norm as the matrix norm and the Euclidean norm as the vector norm.

Example 14.6. Suppose the response values Y are near 0. Consider the fit from an elemental set: $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1} \boldsymbol{Y}_J$ and examine Equations (14.14), (14.15), and (14.16). Now $\|\boldsymbol{b}_J\| \le \|\boldsymbol{X}_J^{-1}\| \ \|\boldsymbol{Y}_J\|$, and *since x-outliers make $\|\boldsymbol{X}_J\|$ large, x-outliers tend to drive $\|\boldsymbol{X}_J^{-1}\|$ and $\|\boldsymbol{b}_J\|$ toward zero not toward ∞.* The x-outliers may make $\|\boldsymbol{b}_J\|$ large if they can make the trial design $\|\boldsymbol{X}_J\|$ nearly singular. Notice that Euclidean norm $\|\boldsymbol{b}_J\|$ can easily be made large if one or more of the elemental response variables is driven far away from zero.

Example 14.7. Without loss of generality, assume that the clean Y's are contained in an interval $[a, f]$ for some a and f. Assume that the regression model contains an intercept β_1. Then there exists an estimator $\hat{\boldsymbol{\beta}}_M$ of $\boldsymbol{\beta}$ such that $\|\hat{\boldsymbol{\beta}}_M\| \le \max(|a|, |f|)$ if $d_n < n/2$.

Proof. Let $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ and $\text{MAD}(n) = \text{MAD}(Y_1, ..., Y_n)$. Take $\hat{\boldsymbol{\beta}}_M = (\text{MED}(n), 0, ..., 0)^T$. Then $\|\hat{\boldsymbol{\beta}}_M\| = |\text{MED}(n)| \le \max(|a|, |f|)$. Note that the median absolute residual for the fit $\hat{\boldsymbol{\beta}}_M$ is equal to the median absolute deviation $\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, i = 1, ..., n) \le f - a$ if $d_n < \lfloor (n+1)/2 \rfloor$. □

Note that $\hat{\boldsymbol{\beta}}_M$ is a poor high breakdown estimator of $\boldsymbol{\beta}$ and $\hat{Y}_i(\hat{\boldsymbol{\beta}}_M)$ tracks the Y_i very poorly. If the data are in general position, a high breakdown regression estimator is an estimator which has a bounded median absolute residual even when close to half of the observations are arbitrary. Rousseeuw and Leroy (1987, pp. 29, 206) conjectured that high breakdown regression estimators can not be computed cheaply and that if the algorithm is also affine equivariant, then the complexity of the algorithm must be at least $O(n^p)$. The following theorem shows that these two conjectures are false.

Theorem 14.3. If the clean data are in general position and the model has an intercept, then a scale and affine equivariant high breakdown estimator $\hat{\boldsymbol{\beta}}_w$ can be found by computing OLS on the set of cases that have $Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm w \ \text{MAD}(Y_1, ..., Y_n)]$ where $w \geq 1$ (so at least half of the cases are used).

Proof. Note that $\hat{\boldsymbol{\beta}}_w$ is obtained by computing OLS on the set J of the n_j cases which have

$$Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm w\text{MAD}(Y_1, ..., Y_n)] \equiv [\text{MED}(n) \pm w\text{MAD}(n)]$$

where $w \geq 1$ (to guarantee that $n_j \geq n/2$). Consider the estimator $\hat{\boldsymbol{\beta}}_M = (\text{MED}(n), 0, ..., 0)^T$ which yields the predicted values $\hat{Y}_i \equiv \text{MED}(n)$. The squared residual $r_i^2(\hat{\boldsymbol{\beta}}_M) \leq (w \ \text{MAD}(n))^2$ if the ith case is in J. Hence the weighted LS fit $\hat{\boldsymbol{\beta}}_w$ is the OLS fit to the cases in J and has

$$\sum_{i \in J} r_i^2(\hat{\boldsymbol{\beta}}_w) \leq n_j (w \ \text{MAD}(n))^2.$$

Thus

$$\text{MED}(|r_1(\hat{\boldsymbol{\beta}}_w)|, ..., |r_n(\hat{\boldsymbol{\beta}}_w)|) \leq \sqrt{n_j} \ w \ \text{MAD}(n) < \sqrt{n} \ w \ \text{MAD}(n) < \infty.$$

Thus the estimator $\hat{\boldsymbol{\beta}}_w$ has a median absolute residual bounded by $\sqrt{n} \ w \ \text{MAD}(Y_1, ..., Y_n)$. Hence $\hat{\boldsymbol{\beta}}_w$ is high breakdown, and it is affine equivariant since the design is not used to choose the observations. It is scale equivariant since for constant $c = 0$, $\hat{\boldsymbol{\beta}}_w = \mathbf{0}$, and for $c \neq 0$ the set of cases used remains the same under scale transformations and OLS is scale equivariant. □

Note that if w is huge and $\text{MAD}(n) \neq 0$, then the high breakdown estimator $\hat{\boldsymbol{\beta}}_w$ and $\hat{\boldsymbol{\beta}}_{OLS}$ will be the same for most data sets. Thus high breakdown estimators can be very nonrobust. Even if $w = 1$, the HB estimator $\hat{\boldsymbol{\beta}}_w$ only resists large Y outliers.

An ALTA concentration algorithm uses the L_1 estimator instead of OLS in the concentration step and uses the LTA criterion. Similarly, an ALMS concentration algorithm uses the L_∞ estimator and the LMS criterion.

Theorem 14.4. If the clean data are in general position and if a high breakdown start is added to an ALTA, ALTS, or ALMS concentration algorithm, then the resulting estimator is HB.

Proof. Concentration reduces (or does not increase) the corresponding HB criterion that is based on $c_n \geq n/2$ absolute residuals, so the median absolute residual of the resulting estimator is bounded as long as the criterion applied to the HB estimator is bounded. □

For example, consider the LTS(c_n) criterion. Suppose the ordered squared residuals from the high breakdown mth start $\boldsymbol{b}_{0m}$ are obtained. If the data are in general position, then $Q_{LTS}(\boldsymbol{b}_{0m})$ is bounded even if the number of outliers d_n is nearly as large as $n/2$. Then $\boldsymbol{b}_{1m}$ is simply the OLS fit to the cases corresponding to the c_n smallest squared residuals $r^2_{(i)}(\boldsymbol{b}_{0m})$ for $i = 1, ..., c_n$. Denote these cases by $i_1, ..., i_{c_n}$. Then $Q_{LTS}(\boldsymbol{b}_{1m}) =$

$$\sum_{i=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{0m}) = \sum_{j=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{0m}) = Q_{LTS}(\boldsymbol{b}_{0m})$$

where the second inequality follows from the definition of the OLS estimator. Hence concentration steps reduce or at least do not increase the LTS criterion. If $c_n = (n+1)/2$ for n odd and $c_n = 1 + n/2$ for n even, then the LTS criterion is bounded iff the median squared residual is bounded.

Theorem 14.4 can be used to show that the following two estimators are high breakdown. The estimator $\hat{\boldsymbol{\beta}}_B$ is the high breakdown attractor used by the $\sqrt{n}$ consistent high breakdown `hbreg` estimator of Definition 14.14, while the estimator $\boldsymbol{b}_{k,B}$ is the high breakdown attractor used by the $\sqrt{n}$ consistent high breakdown `CLTA` estimator of Theorem 14.11.

Definition 14.13. Make an OLS fit to the $c_n \approx n/2$ cases whose Y values are closest to the MED$(Y_1, ..., Y_n) \equiv$ MED(n), and use this fit as the start for concentration. Define $\hat{\boldsymbol{\beta}}_B$ to be the attractor after k concentration steps. Define $\boldsymbol{b}_{k,B} = 0.9999\hat{\boldsymbol{\beta}}_B$.

Theorem 14.5. If the clean data are in general position, then $\hat{\boldsymbol{\beta}}_B$ and $\boldsymbol{b}_{k,B}$ are high breakdown regression estimators.

Proof. The start can be taken to be $\hat{\boldsymbol{\beta}}_w$ with $w = 1$ from Theorem 14.3. Since the start is high breakdown, so is the attractor $\hat{\boldsymbol{\beta}}_B$ by Theorem 14.4. Multiplying a HB estimator by a positive constant does not change the breakdown value, so $\boldsymbol{b}_{k,B}$ is HB. □

The following result shows that it is easy to make a HB estimator that is asymptotically equivalent to a consistent estimator on a large class of iid zero mean symmetric error distributions, although the outlier resistance of the HB estimator is poor. The following result may not hold if $\hat{\boldsymbol{\beta}}_C$ estimates $\boldsymbol{\beta}_C$ and $\hat{\boldsymbol{\beta}}_{LMS}$ estimates $\boldsymbol{\beta}_{LMS}$ where $\boldsymbol{\beta}_C \neq \boldsymbol{\beta}_{LMS}$. Then $\boldsymbol{b}_{k,B}$ could have a smaller median squared residual than $\hat{\boldsymbol{\beta}}_C$ even if there are no outliers. The two parameter vectors could differ because the constant term is different if the error distribution is not symmetric. For a large class of symmetric error distributions, $\boldsymbol{\beta}_{LMS} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_C \equiv \boldsymbol{\beta}$, then the ratio $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}))/\text{MED}(r_i^2(\boldsymbol{\beta})) \to 1$ as $n \to \infty$ for any consistent estimator of $\boldsymbol{\beta}$. The estimator below has two attractors, $\hat{\boldsymbol{\beta}}_C$ and $\boldsymbol{b}_{k,B}$, and the probability that the final estimator $\hat{\boldsymbol{\beta}}_D$ is equal to $\hat{\boldsymbol{\beta}}_C$ goes to one under the strong assumption that the error distribution is such that both $\hat{\boldsymbol{\beta}}_C$ and $\hat{\boldsymbol{\beta}}_{LMS}$ are consistent estimators of $\boldsymbol{\beta}$.

Theorem 14.6. Assume the clean data are in general position and that the LMS estimator is a consistent estimator of $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}_C$ be any practical consistent estimator of $\boldsymbol{\beta}$, and let $\hat{\boldsymbol{\beta}}_D = \hat{\boldsymbol{\beta}}_C$ if $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}_C)) \leq \text{MED}(r_i^2(\boldsymbol{b}_{k,B}))$. Let $\hat{\boldsymbol{\beta}}_D = \boldsymbol{b}_{k,B}$, otherwise. Then $\hat{\boldsymbol{\beta}}_D$ is a HB estimator that is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$.

Proof. The estimator is HB since the median squared residual of $\hat{\boldsymbol{\beta}}_D$ is no larger than that of the HB estimator $\boldsymbol{b}_{k,B}$. Since $\hat{\boldsymbol{\beta}}_C$ is consistent, $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}_C)) \to \text{MED}(e^2)$ in probability where $\text{MED}(e^2)$ is the population median of the squared error e^2. Since the LMS estimator is consistent, the probability that $\hat{\boldsymbol{\beta}}_C$ has a smaller median squared residual than the biased estimator $\hat{\boldsymbol{\beta}}_{k,B}$ goes to 1 as $n \to \infty$. Hence $\hat{\boldsymbol{\beta}}_D$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$. □

14.4.2 A Practical High Breakdown Consistent Estimator

Olive and Hawkins (2011) showed that the practical hbreg estimator is a high breakdown $\sqrt{n}$ consistent robust estimator that is asymptotically equivalent to the least squares estimator for many error distributions. The hbreg estimator was used in Section 12.6.1 to make the somewhat outlier resistant rmreg estimator of multivariate linear regression that was asymptotically equivalent to the classical estimator.

Suppose the algorithm estimator uses some criterion to choose an attractor as the final estimator where there are K attractors and K is fixed, e.g., $K = 500$, so K does not depend on n. The following theorem is powerful 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 such as elemental concentration algorithms where all K of the attractors are inconsistent are especially untrustworthy.

Suppose there are K consistent estimators $\hat{\boldsymbol{\beta}}_j$ of $\boldsymbol{\beta}$, each with the same rate n^δ. If $\hat{\boldsymbol{\beta}}_A$ is an estimator obtained by choosing one of the K estimators, then $\hat{\boldsymbol{\beta}}_A$ is a consistent estimator of $\boldsymbol{\beta}$ with rate n^δ by Pratt (1959). See Proposition 3.15.

Theorem 14.7. 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, then the algorithm estimator is consistent.

ii) If all of the attractors are consistent with the same rate, e.g., n^δ where $0 < \delta \leq 0.5$, then the algorithm estimator is consistent with the same rate as the attractors.

iii) If all of the attractors are high breakdown, then the algorithm estimator is high breakdown.

Proof. i) Choosing from K consistent estimators results in a consistent estimator and ii) follows from Pratt (1959). iii) Let $\gamma_{n,i}$ be the breakdown value of the ith attractor if the clean data 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 $n \to \infty$. □

The consistency of the algorithm estimator changes dramatically if K is fixed but the start size $h = h_n = g(n)$ where $g(n) \to \infty$. In particular, if K starts with rate $n^{1/2}$ are used, the final estimator also has rate $n^{1/2}$. The drawback to these algorithms is that they may not have enough outlier resistance. Notice that the basic resampling result below is free of the criterion.

Proposition 14.8. Suppose $K_n \equiv K$ starts are used and that all starts have subset size $h_n = g(n) \uparrow \infty$ as $n \to \infty$. Assume that the estimator applied to the subset has rate n^δ.

i) For the h_n-set basic resampling algorithm, the algorithm estimator has rate $[g(n)]^\delta$.

ii) Under regularity conditions (e.g., given by He and Portnoy 1992), the k-step CLTS estimator has rate $[g(n)]^\delta$.

Proof. i) The $h_n = g(n)$ cases are randomly sampled without replacement. Hence the classical estimator applied to these $g(n)$ cases has rate $[g(n)]^\delta$. Thus all K starts have rate $[g(n)]^\delta$, and the result follows by Pratt (1959). ii) By He and Portnoy (1992), all K attractors have $[g(n)]^\delta$ rate, and the result follows by Pratt (1959). □

Remark 14.1. Proposition 14.2 shows that $\hat{\boldsymbol{\beta}}$ is HB if the median absolute or squared residual (or $|r(\hat{\boldsymbol{\beta}})|_{(c_n)}$ or $r^2_{(c_n)}$ where $c_n \approx n/2$) stays bounded under high contamination. Let $Q_L(\hat{\boldsymbol{\beta}}_H)$ denote the LMS, LTS, or LTA criterion for an estimator $\hat{\boldsymbol{\beta}}_H$; therefore, the estimator $\hat{\boldsymbol{\beta}}_H$ is high breakdown if and only if $Q_L(\hat{\boldsymbol{\beta}}_H)$ is bounded for d_n near $n/2$ where $d_n < n/2$ is the number of outliers. The concentration operator refines an initial estimator by successively reducing the LTS criterion. If $\hat{\boldsymbol{\beta}}_F$ refers to the final estimator (attractor) obtained by applying concentration to some starting estimator $\hat{\boldsymbol{\beta}}_H$ that is high breakdown, then since $Q_{LTS}(\hat{\boldsymbol{\beta}}_F) \leq Q_{LTS}(\hat{\boldsymbol{\beta}}_H)$, applying concentration to a high breakdown start results in a high breakdown attractor. See Theorem 14.4.

High breakdown estimators are, however, not necessarily useful for detecting outliers. Suppose $\gamma_n < 0.5$. On the one hand, if the $\boldsymbol{x}_i$ are fixed, and the outliers are moved up and down parallel to the Y axis, then for high breakdown estimators, $\hat{\boldsymbol{\beta}}$ and MED$(|r_i|)$ will be bounded. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large, suggesting that the high breakdown estimator is useful for outlier detection. On the other hand, if the Y_i's are fixed at any values and the $\boldsymbol{x}$ values perturbed, sufficiently large $\boldsymbol{x}$-outliers tend to drive the slope estimates to 0, not ∞. For many estimators, including LTS, LMS, and LTA, a cluster of Y outliers can be moved arbitrarily far from the bulk of the data but still, by perturbing their $\boldsymbol{x}$ values, have arbitrarily small residuals.

Our practical high breakdown procedure is made up of three components.
1) A practical estimator $\hat{\boldsymbol{\beta}}_C$ that is consistent for clean data. Suitable choices would include the full-sample OLS and L_1 estimators.
2) A practical estimator $\hat{\boldsymbol{\beta}}_A$ that is effective for outlier identification. Suitable choices include the `mbareg`, `rmreg2`, `lmsreg`, or FLTS estimators.
3) A practical high breakdown estimator such as $\hat{\boldsymbol{\beta}}_B$ from Definition 14.13 with $k = 10$.

By selecting one of these three estimators according to the features each of them uncovers in the data, we may inherit some of the good properties of each of them.

Definition 14.14. The `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ is defined as follows. Pick a constant $a > 1$ and set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_C$. If $aQ_L(\hat{\boldsymbol{\beta}}_A) < Q_L(\hat{\boldsymbol{\beta}}_C)$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_A$. If $aQ_L(\hat{\boldsymbol{\beta}}_B) < \min[Q_L(\hat{\boldsymbol{\beta}}_C), aQ_L(\hat{\boldsymbol{\beta}}_A)]$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_B$.

That is, find the smallest of the three scaled criterion values $Q_L(\hat{\boldsymbol{\beta}}_C)$, $aQ_L(\hat{\boldsymbol{\beta}}_A)$, $aQ_L(\hat{\boldsymbol{\beta}}_B)$. According to which of the three estimators attain this minimum, set $\hat{\boldsymbol{\beta}}_H$ to $\hat{\boldsymbol{\beta}}_C, \hat{\boldsymbol{\beta}}_A$, or $\hat{\boldsymbol{\beta}}_B$ respectively.

Large sample theory for `hbreg` is simple and given in the following theorem. Let $\hat{\boldsymbol{\beta}}_L$ be the LMS, LTS, or LTA estimator that minimizes the criterion Q_L. Note that the impractical estimator $\hat{\boldsymbol{\beta}}_L$ is never computed. The following theorem shows that $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$ on a large class of zero mean finite variance symmetric error distributions. Thus if $\hat{\boldsymbol{\beta}}_C$ is $\sqrt{n}$ consistent or asymptotically efficient, so is $\hat{\boldsymbol{\beta}}_H$. Notice that $\hat{\boldsymbol{\beta}}_A$ does not need to be consistent. This point is crucial since `lmsreg` is not consistent and it is not known whether FLTS is consistent. The clean data are in *general position* if any p clean cases give a unique estimate of $\hat{\boldsymbol{\beta}}$.

Theorem 14.9. Assume the clean data are in general position, and suppose that both $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_C$ are consistent estimators of $\boldsymbol{\beta}$ where the regression model contains a constant. Then the `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ is high breakdown and asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$.

Proof. Since the clean data are in general position and $Q_L(\hat{\boldsymbol{\beta}}_H) \leq aQ_L(\hat{\boldsymbol{\beta}}_B)$ is bounded for γ_n near 0.5, the `hbreg` estimator is high breakdown. Let $Q_L^* = Q_L$ for LMS and $Q_L^* = Q_L/n$ for LTS and LTA. As $n \to \infty$, consistent estimators $\hat{\boldsymbol{\beta}}$ satisfy $Q_L^*(\hat{\boldsymbol{\beta}}) - Q_L^*(\boldsymbol{\beta}) \to 0$ in probability. Since LMS, LTS, and LTA are consistent and the minimum value is $Q_L^*(\hat{\boldsymbol{\beta}}_L)$, it follows that $Q_L^*(\hat{\boldsymbol{\beta}}_C) - Q_L^*(\hat{\boldsymbol{\beta}}_L) \to 0$ in probability, while $Q_L^*(\hat{\boldsymbol{\beta}}_L) < aQ_L^*(\hat{\boldsymbol{\beta}})$ for any estimator $\hat{\boldsymbol{\beta}}$. Thus with probability tending to one as $n \to \infty$, $Q_L(\hat{\boldsymbol{\beta}}_C) < a \min(Q_L(\hat{\boldsymbol{\beta}}_A), Q_L(\hat{\boldsymbol{\beta}}_B))$. Hence $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$. □

Remark 14.2. i) Let $\hat{\boldsymbol{\beta}}_C = \hat{\boldsymbol{\beta}}_{OLS}$. Then `hbreg` is asymptotically equivalent to OLS when the errors e_i are iid from a large class of zero mean finite variance symmetric distributions, including the $N(0, \sigma^2)$ distribution, since the probability that `hbreg` uses OLS instead of $\hat{\boldsymbol{\beta}}_A$ or $\hat{\boldsymbol{\beta}}_B$ goes to one as $n \to \infty$.

ii) The above theorem proves that practical high breakdown estimators with 100% asymptotic Gaussian efficiency exist; however, such estimators are not necessarily good.

iii) The theorem holds when both $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_C$ are consistent estimators of $\boldsymbol{\beta}$, for example, when the iid errors come from a large class or zero mean finite variance symmetric distributions. For asymmetric distributions, $\hat{\boldsymbol{\beta}}_C$ estimates $\boldsymbol{\beta}_C$ and $\hat{\boldsymbol{\beta}}_L$ estimates $\boldsymbol{\beta}_L$ where the constants usually differ. The theorem holds for some distributions that are not symmetric because of the penalty a. As $a \to \infty$, the class of asymmetric distributions where the theorem holds greatly increases, but the outlier resistance decreases rapidly as a increases for $a > 1.4$.

iv) The default hbreg estimator used OLS, mbareg, and $\hat{\boldsymbol{\beta}}_B$ with $a = 1.4$ and the LTA criterion. For the simulated data with symmetric error distributions, $\hat{\boldsymbol{\beta}}_B$ appeared to give biased estimates of the slopes. However, for the simulated data with right skewed error distributions, $\hat{\boldsymbol{\beta}}_B$ appeared to give good estimates of the slopes but not the constant estimated by OLS, and the probability that the hbreg estimator selected $\hat{\boldsymbol{\beta}}_B$ appeared to go to one.

v) Both MBA and OLS are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$, even for a large class of skewed distributions. Using $\hat{\boldsymbol{\beta}}_A = \hat{\boldsymbol{\beta}}_{MBA}$ and removing $\hat{\boldsymbol{\beta}}_B$ from the hbreg estimator results in a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ when $\hat{\boldsymbol{\beta}}_C = \text{OLS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$, but massive sample sizes were still needed to get good estimates of the constant for skewed error distributions. For skewed distributions, if OLS needed $n = 1000$ to estimate the constant well, mbareg might need $n >$ one million to estimate the constant well.

The situation is worse for multivariate linear regression when hbreg is used instead of OLS, since there are m constants to be estimated. If the distribution of the iid error vectors $\boldsymbol{e}_i$ is not elliptically contoured, getting all m mbareg estimators to estimate all m constants well needs even larger sample sizes. See the rmreg estimator of Section 12.6.1.

vi) The outlier resistance of hbreg is not especially good and could likely be improved by letting the constant a_p depend on p. Let a_p be very near 1 for low p and increase a_p to 1.4 as p increases.

vii) Note that for a large class of symmetric elliptically contoured distributions for $\boldsymbol{\epsilon}$, the hbreg estimator could be used instead of OLS in multivariate linear regression, but the classical tests could still be applied since the rmreg and classical estimators were asymptotically equivalent. See Section 12.6.1.

viii) The estimator $\hat{\boldsymbol{\beta}}_A$ only needs to be practical to compute, and it does not need to be consistent. Hence hbreg can be used to fix the estimators that are zero breakdown and inconsistent, but the maximal bias of hbreg will not be as good as that of some impractical high breakdown estimators.

The family of hbreg estimators is enormous and depends on i) the practical high breakdown estimator $\hat{\boldsymbol{\beta}}_B$, ii) $\hat{\boldsymbol{\beta}}_C$, iii) $\hat{\boldsymbol{\beta}}_A$, iv) a, and v) the criterion Q_L. Note that the theory needs the error distribution to be such that both $\hat{\boldsymbol{\beta}}_C$ and $\hat{\boldsymbol{\beta}}_L$ are consistent. Sufficient conditions for LMS, LTS, and LTA to be consistent are rather strong. To have reasonable sufficient conditions for the hbreg estimator to be consistent, $\hat{\boldsymbol{\beta}}_C$ should be consistent under weak conditions. Hence OLS is a good choice that results in 100% asymptotic Gaussian efficiency.

We suggest using the LTA criterion since in simulations, hbreg behaved like $\hat{\boldsymbol{\beta}}_C$ for smaller sample sizes than those needed by the LTS and LMS criteria. We want a near 1 so that hbreg has outlier resistance similar to $\hat{\boldsymbol{\beta}}_A$, but we want a large enough so that hbreg performs like $\hat{\boldsymbol{\beta}}_C$ for moderate

n on clean data. Simulations suggest that $a = 1.4$ is a reasonable choice. The default `hbreg` program from *mpack* uses the $\sqrt{n}$ consistent outlier resistant estimator `mbareg` as $\hat{\boldsymbol{\beta}}_A$.

There are at least three reasons for using $\hat{\boldsymbol{\beta}}_B$ as the high breakdown estimator. First, $\hat{\boldsymbol{\beta}}_B$ is high breakdown and simple to compute. Second, the fitted values roughly track the bulk of the data. Lastly, although $\hat{\boldsymbol{\beta}}_B$ has rather poor outlier resistance, $\hat{\boldsymbol{\beta}}_B$ does perform well on several outlier configurations where some common alternatives fail. See Figures 14.6 and 14.7.

Next, we will show that the `hbreg` estimator implemented with $a = 1.4$ using Q_{LTA}, $\hat{\boldsymbol{\beta}}_C =$ OLS, and $\hat{\boldsymbol{\beta}}_B$ can greatly improve the estimator $\hat{\boldsymbol{\beta}}_A$. We will use $\hat{\boldsymbol{\beta}}_A =$ `ltsreg` in *R* and *Splus 2000*. Depending on the implementation, the `ltsreg` estimators use the elemental resampling algorithm, the elemental concentration algorithm, or a genetic algorithm. Coverage is 50%, 75%, or 90%. The *Splus 2000* implementation is an unusually poor genetic algorithm with 90% coverage. The *R* implementation appears to be the zero breakdown inconsistent elemental basic resampling algorithm that uses 50% coverage. The *ltsreg* function changes often.

Simulations were run in *R* with the x_{ij} (for $j > 1$) and e_i iid $N(0, \sigma^2)$ and $\boldsymbol{\beta} = \mathbf{1}$, the $p \times 1$ vector of ones. Then $\hat{\boldsymbol{\beta}}$ was recorded for 100 runs. The mean and standard deviation of the $\hat{\beta}_j$ were recorded for $j = 1, ..., p$. For $n \geq 10p$ and OLS, the vector of means should be close to $\mathbf{1}$ and the vector of standard deviations should be close to $\mathbf{1}/\sqrt{n}$. The $\sqrt{n}$ consistent high breakdown `hbreg` estimator performed like OLS if $n \approx 35p$ and $2 \leq p \leq 6$, if $n \approx 20p$ and $7 \leq p \leq 14$, or if $n \approx 15p$ and $15 \leq p \leq 40$. See Table 14.3 for $p = 5$ and 100 runs. ALTS denotes `ltsreg`, HB denotes `hbreg`, and BB denotes $\hat{\boldsymbol{\beta}}_B$. In the simulations, `hbreg` estimated the slopes well for the highly skewed lognormal data, but not the OLS constant. Use the *mpack* function `hbregsim`.

As implemented in *mpack*, the `hbreg` estimator is a practical $\sqrt{n}$ consistent high breakdown estimator that appears to perform like OLS for moderate n if the errors are unimodal and symmetric and to have outlier resistance comparable to competing practical "outlier resistant" estimators.

The `hbreg`, `lmsreg`, `ltsreg`, OLS, and $\hat{\boldsymbol{\beta}}_B$ estimators were compared on the same 25 benchmark data sets. Also see Park et al. (2012). The HB estimator $\hat{\boldsymbol{\beta}}_B$ was surprisingly good in that the response plots showed that it was the best estimator for two data sets and that it usually tracked the data, but it performed poorly in seven of the 25 data sets. The `hbreg` estimator performed well, but for a few data sets `hbreg` did not pick the attractor with the best response plot, as illustrated in the following example.

Example 14.8. The LMS, LTA, and LTS estimators are determined by a "narrowest band" covering half of the cases. Hawkins and Olive (2002) suggested that the fit will pass through outliers if the band through the outliers is narrower than the band through the clean cases. This behavior

Table 14.3 MEAN $\hat{\beta}_i$ and SD($\hat{\beta}_i$)

n	method	mn or sd	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
25	HB	mn	0.9921	0.9825	0.9989	0.9680	1.0231
		sd	0.4821	0.5142	0.5590	0.4537	0.5461
	OLS	mn	1.0113	1.0116	0.9564	0.9867	1.0019
		sd	0.2308	0.2378	0.2126	0.2071	0.2441
	ALTS	mn	1.0028	1.0065	1.0198	1.0092	1.0374
		sd	0.5028	0.5319	0.5467	0.4828	0.5614
	BB	mn	1.0278	0.5314	0.5182	0.5134	0.5752
		sd	0.4960	0.3960	0.3612	0.4250	0.3940
400	HB	mn	1.0023	0.9943	1.0028	1.0103	1.0076
		sd	0.0529	0.0496	0.0514	0.0459	0.0527
	OLS	mn	1.0023	0.9943	1.0028	1.0103	1.0076
		sd	0.0529	0.0496	0.0514	0.0459	0.0527
	ALTS	mn	1.0077	0.9823	1.0068	1.0069	1.0214
		sd	0.1655	0.1542	0.1609	0.1629	0.1679
	BB	mn	1.0184	0.8744	0.8764	0.8679	0.8794
		sd	0.1273	0.1084	0.1215	0.1206	0.1269

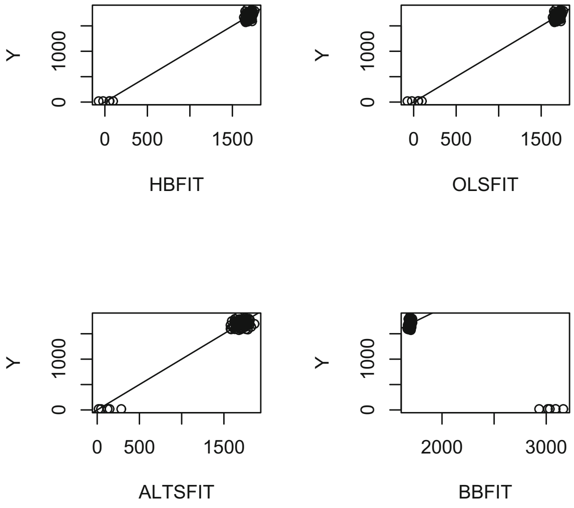

Fig. 14.8 Response Plots Comparing Robust Regression Estimators

tends to occur if the regression relationship is weak, and if there is a tight cluster of outliers where $|Y|$ is not too large. As an illustration, Buxton (1920, pp. 232–5) gave 20 measurements of 88 men. Consider predicting *stature* using an intercept, *head length, nasal height, bigonal breadth*, and *cephalic index*. One case was deleted since it had missing values. Five individuals, numbers 6–65, were reported to be about 0.75 inches tall with head lengths well over five feet! Figure 14.8 shows the response plots for `hbreg`, OLS, `ltsreg`, and $\hat{\boldsymbol{\beta}}_B$. Notice that only the fit from $\hat{\boldsymbol{\beta}}_B$ (BBFIT) did not pass through the outliers, but `hbreg` selected the OLS attractor. There are always outlier configurations where an estimator will fail, and `hbreg` should fail on configurations where LTA, LTS, and LMS would fail.

14.5 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 14.15. In a *1D regression model*, the response Y is conditionally independent of $\boldsymbol{x}$ given the real-valued function $h(\boldsymbol{x})$, written $Y \perp\!\!\!\perp \boldsymbol{x}|h(\boldsymbol{x})$. If $h(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}$ is a single linear combination $\boldsymbol{\beta}^T\boldsymbol{x}$ of the predictors, then

$$Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x} \quad \text{or} \quad \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x}|(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}). \tag{14.17}$$

For the remainder of this section, unless told otherwise, assume $h(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}$. An important 1D regression model, introduced by Li and Duan (1989), has the form

$$Y = g(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}, e) \tag{14.18}$$

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 (14.17) include many important *generalized linear models* (GLMs) and the additive error *single index model*

$$Y = m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) + e. \tag{14.19}$$

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

$$g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e) = t^{-1}(\alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e) \tag{14.20}$$

and t^{-1} is a one-to-one (typically monotone) function. Hence

$$t(Y) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e.$$

Definition 14.16. *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* $\text{VAR}(Y|\boldsymbol{x})$. 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 with $h(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x}$,

$$E(Y|\boldsymbol{x} = \boldsymbol{x}_o) = E(Y|\boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}^T \boldsymbol{x}_o) \equiv M(\boldsymbol{\beta}^T \boldsymbol{x}_o)$$

and

$$\text{VAR}(Y|\boldsymbol{x} = \boldsymbol{x}_o) = \text{VAR}(Y|\boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}^T \boldsymbol{x}_o) \equiv V(\boldsymbol{\beta}^T \boldsymbol{x}_o)$$

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*

$$Y = M(\boldsymbol{\beta}_1^T \boldsymbol{x}) + \sqrt{V(\boldsymbol{\beta}_2^T \boldsymbol{x})} \;\; e \tag{14.21}$$

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 with $h(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x}$ 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) defined 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 14.17. If the 1D regression model with $h(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x}$ 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.

14.6 Visualizing 1D Regression

Consider the OLS estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$. Li and Duan (1989, p. 1031) showed 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 (14.17) holds, then the response plot of

$$\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x} \quad \text{versus} \quad \text{Y}$$

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,** $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ 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 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 m is symmetric about the population median of $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$.

Definition 14.18. If the 1D regression model (14.17) 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 (sliced inverse regression), may have similar labels.

Example 14.9. 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 14.9 shows the sufficient summary plot of $\boldsymbol{\beta}^T\boldsymbol{x}$ versus Y, and Figure 14.10 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* commands.

```
X <- matrix(rnorm(300),nrow=100,ncol=3)
SP <- X%*%1:3
Y <- (SP)^3 + rnorm(100)
plot(SP,Y)
plot(-SP,Y)
```

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

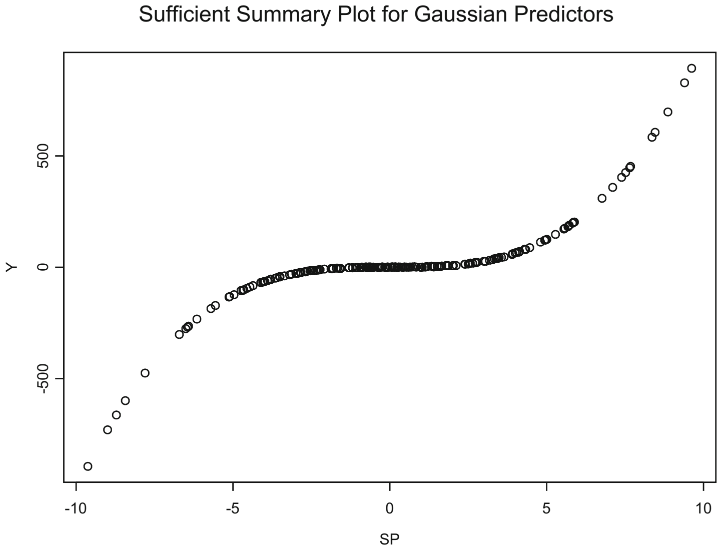

Fig. 14.9 SSP for $m(u) = u^3$

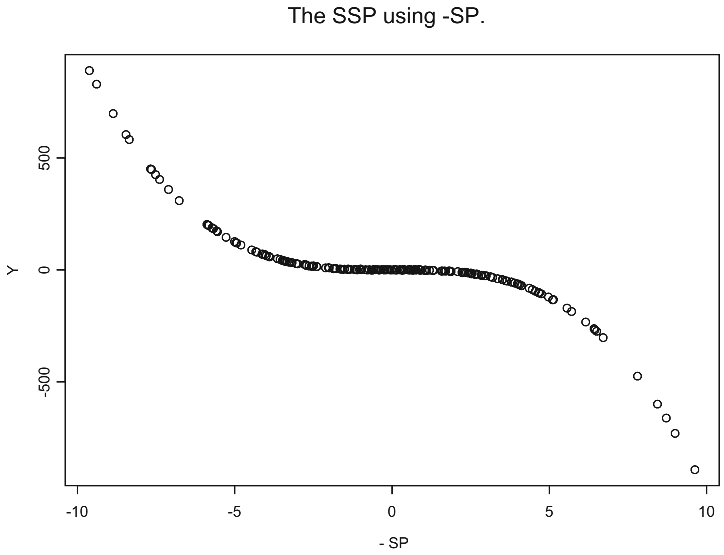

Fig. 14.10 Another SSP for $m(u) = u^3$

$\hat{\boldsymbol{\beta}} \approx c\boldsymbol{\beta}$ for some nonzero constant c. Nice results exist if the additive error single index model is appropriate. Recall that

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

Definition 14.19. Suppose that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid observations and that the positive definite $k \times k$ matrix $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$ and the $k \times 1$ vector $\text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$. 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

$$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x},Y}. \tag{14.22}$$

The following notation will be useful for studying the OLS estimator. Let the sufficient predictor $z = \boldsymbol{\beta}^T \boldsymbol{x}$, and let $\boldsymbol{w} = \boldsymbol{x} - E(\boldsymbol{x})$. Let $\boldsymbol{r} = \boldsymbol{w} - (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T \boldsymbol{w}$.

Theorem 14.10. In addition to the conditions of Definition 14.19, 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

$$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x},Y} = c_{m,\boldsymbol{x}} \boldsymbol{\beta} + \boldsymbol{u}_{m,\boldsymbol{x}} \tag{14.23}$$

where the scalar

$$c_{m,\boldsymbol{x}} = E[\boldsymbol{\beta}^T(\boldsymbol{x} - E(\boldsymbol{x}))\ m(\boldsymbol{\beta}^T\boldsymbol{x})] \tag{14.24}$$

and the bias vector

$$\boldsymbol{u}_{m,\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} E[m(\boldsymbol{\beta}^T\boldsymbol{x})\boldsymbol{r}]. \tag{14.25}$$

Moreover, $\boldsymbol{u}_{m,\boldsymbol{x}} = \boldsymbol{0}$ if $\boldsymbol{x}$ is from an EC distribution with nonsingular $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, and $c_{m,\boldsymbol{x}} \neq 0$ unless $\text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{0}$. If the multiple linear regression model holds, then $c_{m,\boldsymbol{x}} = 1$, and $\boldsymbol{u}_{m,\boldsymbol{x}} = \boldsymbol{0}$.

The proof of the above result is outlined in Problem 14.1 using an argument due to Aldrin et al. (1993). See related results in Cook et al. (1992). If the 1D regression model with $h(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}$ is appropriate, then typically $\text{Cov}(\boldsymbol{x}, Y) \neq \boldsymbol{0}$ unless $\boldsymbol{\beta}^T\boldsymbol{x}$ follows a symmetric distribution and m is symmetric about the median of $\boldsymbol{\beta}^T\boldsymbol{x}$. Often the bias vector is small if there are no strong nonlinearities among the predictors.

Definition 14.20. 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 14.3. 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 with $h(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}$ 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 $c_{m,\boldsymbol{x}} \neq 0$ (e.g., 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 next, and the discussion 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

$$D_i^2 = (\boldsymbol{x}_i - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x}_i - T) \tag{14.26}$$

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

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T C^{-1} (\boldsymbol{x} - T) \leq D^2_{(j)}\}. \tag{14.27}$$

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 RMVN or RFCH. See Section 4.4. When $j \approx n/2$, the RFCH estimator attempts to make the volume of the hyperellipsoid given by Equation (14.27) 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, e.g,. $M = 10$ 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 14.10, any of these three reasons should produce a better estimated sufficient predictor.

For example, examine Figure 5.7. The data are not EC, but the data within the resistant covering ellipsoid are approximately EC.

Example 14.10. Cook and Weisberg (1999a, pp. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. The variables 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 14.11 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.

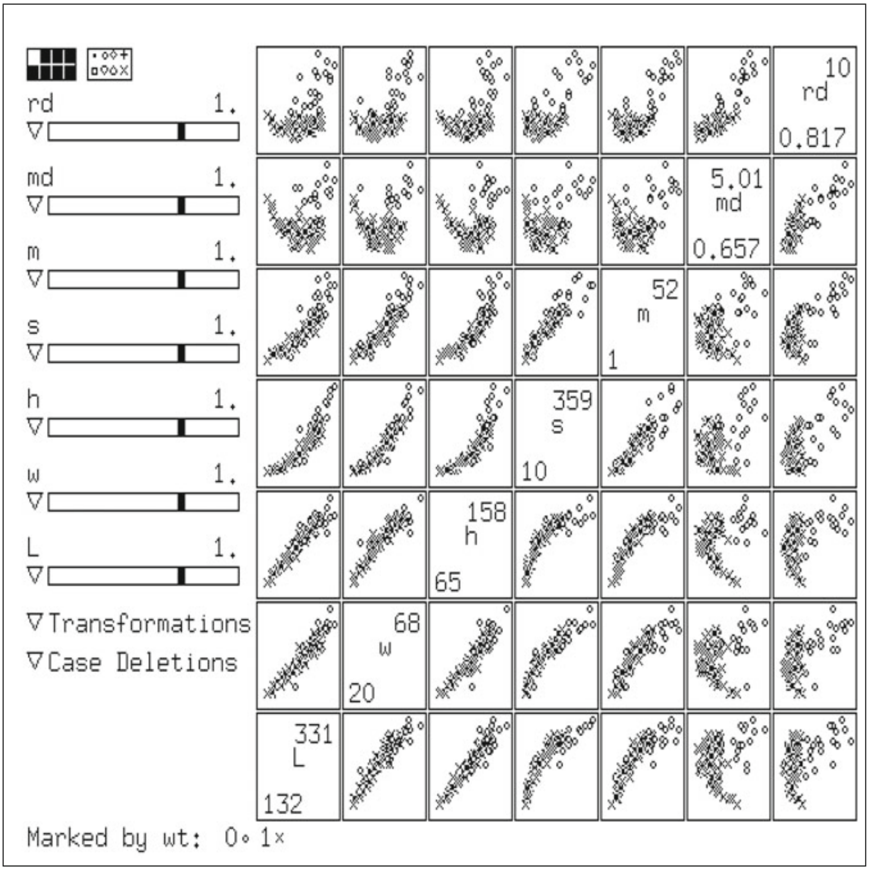

Fig. 14.11 Scatterplot for Mussel Data, o Corresponds to Trimmed Cases

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 (1999a, ch. 8) 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 14.21. 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* function `trviews` that is used to produce the ESSP. The command

```
library(MASS)
```

needs to be entered to access the function `cov.mcd` called by `trviews`. The function `trviews` is used in Problem 14.11. Also notice the `trviews` estimator is basically the same as the `tvreg` estimator described in Section 14.2. 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 $Y = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) + e$ may be reasonable.

Example 14.9 **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 fifth 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 $(\hat{\alpha}_0, \hat{\boldsymbol{\beta}}_0^T)$. The trimmed views with 50% and 90% trimming were very good. 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 14.12 and is nearly identical to the sufficient summary plot shown in Figure 14.9. 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 $|\text{corr(ESP,SP)}| > 0.95$) and follow a line through the origin,

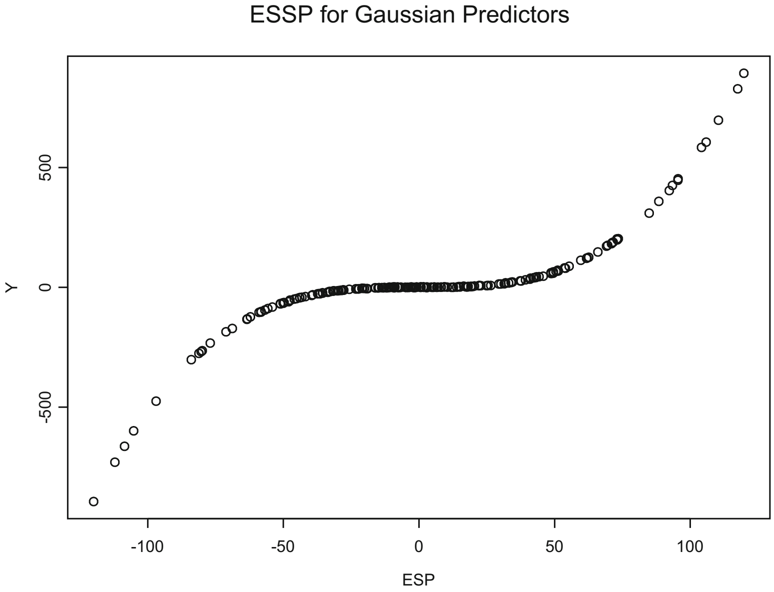

Fig. 14.12 Best View for Estimating $m(u) = u^3$

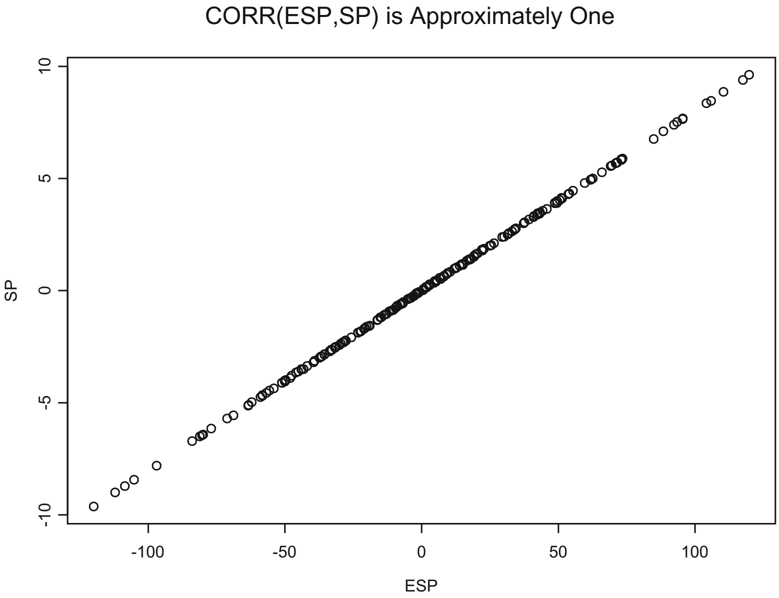

Fig. 14.13 The angle between the SP and the ESP is nearly zero.

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 14.13.

Example 14.11. 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 14.14d shows the OLS view, and Figure 14.15d 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 14.15d, temporarily omit cases that have ESP less than

0.3 and greater than 0.75. From the untrimmed cases, obtain 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 additive error single index model (14.19) 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*.

Application 14.3. The best trimmed view can also be used as a diagnostic for linearity and monotonicity.

For example in Figure 14.12, if ESP $= 0$, then $\hat{Y} = 0$, and if ESP $= 100$, then $\hat{Y} = 500$. Figure 14.12 suggests that the mean function is monotone but not linear, and Figure 14.15 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. Li Cook and Li (2002) summarized 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 14.11 (continued). Figure 14.14 shows that the response plots for SIR, PHD, SAVE, and OLS are not very good, while Figure 14.15 shows that trimming improved the SIR, SAVE, and OLS methods.

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 dispersion estimator and which regression estimator should

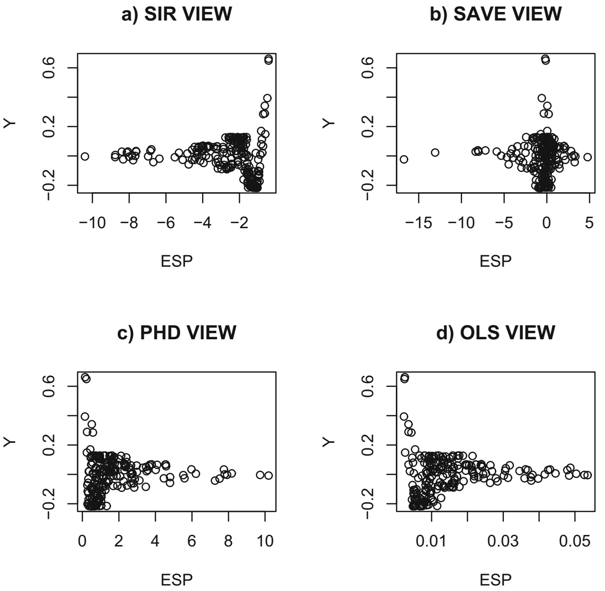

Fig. 14.14 Estimated Sufficient Summary Plots Without Trimming

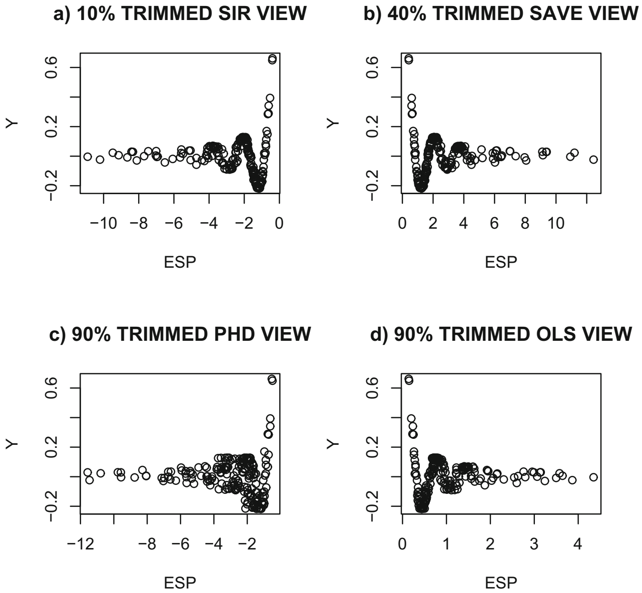

Fig. 14.15 1D Regression with Trimmed Views

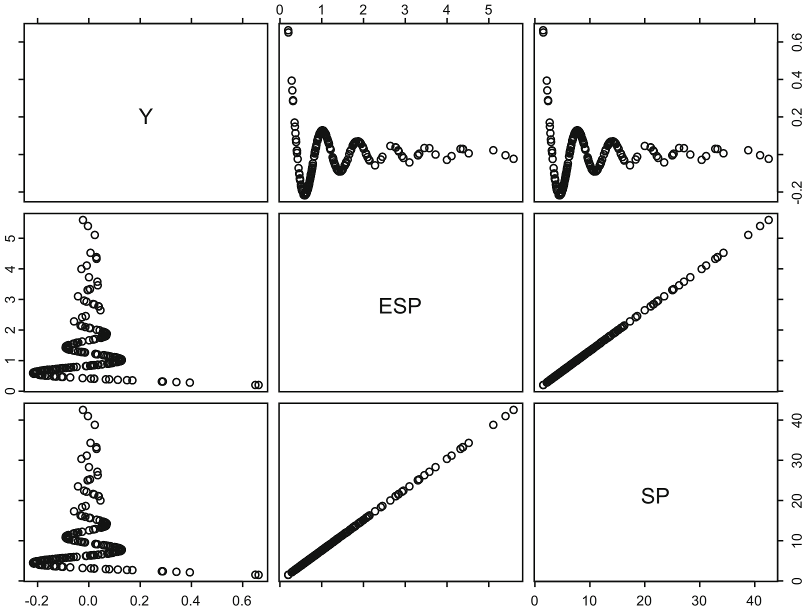

Fig. 14.16 1D Regression with `lmsreg`

be used. We suggest using the RFCH or RMVN 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 14.11 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 14.16 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` instead of OLS. The top row of Figure 14.16 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 14.4 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 14.17 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

Table 14.4 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

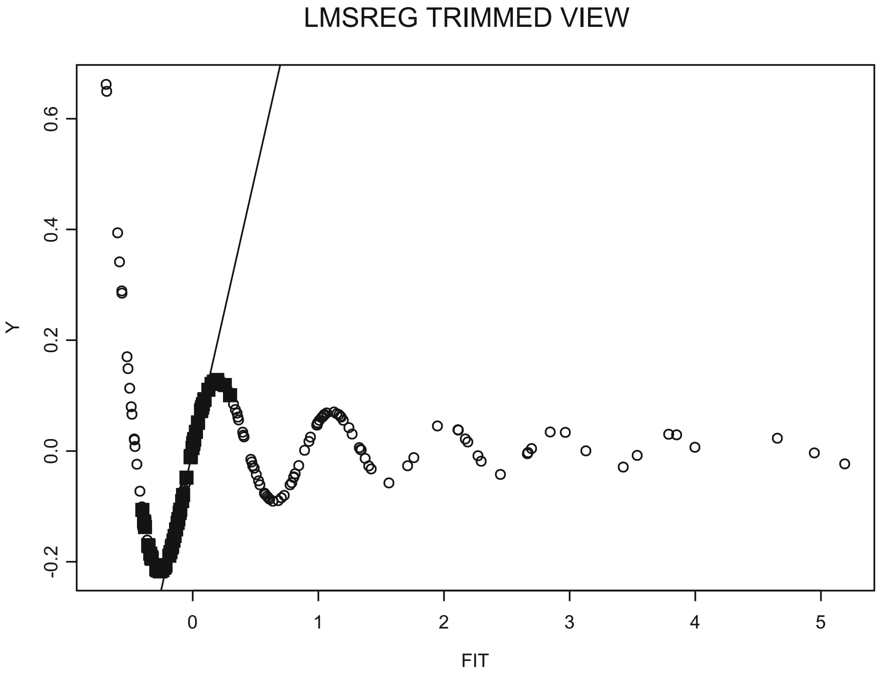

Fig. 14.17 The Weighted `lmsreg` Fitted Values Versus Y

as a whole. Also, `lmsreg` will give half of the highlighted cases zero weight, further linearizing the function. In Figure 14.17, 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 14.12. This insulation data was contributed by Ms. Spector. A box with insulation was heated for 20 minutes and then allowed to cool down. The response variable Y = *temperature* in the middle of the 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 14.18 is shown below. The solid curve corresponds to $(\boldsymbol{x}^T\hat{\boldsymbol{\beta}}, (\boldsymbol{x}^T\hat{\boldsymbol{\beta}})^3) = (FIT, (FIT)^3)$ 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 curves 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 14.17 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)
```

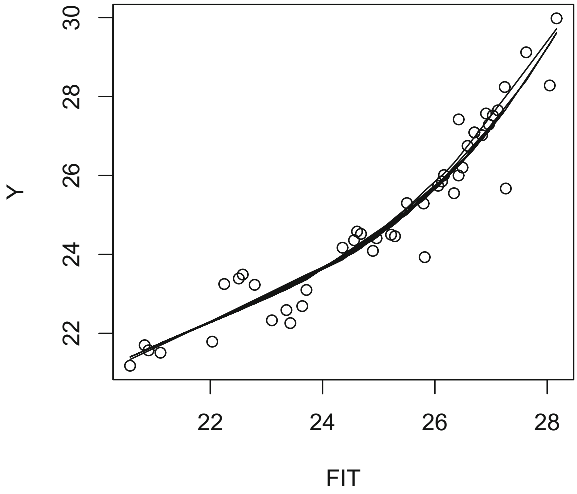

Fig. 14.18 Response Plot for Insulation Data

```
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)
```

14.7 Complements

The fact that response plots are extremely useful for model assessment and for detecting influential cases and outliers for an enormous variety of statistical models does not seem to be well known. Certainly, in any multiple linear regression analysis, the response plot and the residual plot of $\hat{Y}$ versus r should always be made. Cook and Olive (2001) used response plots to select a response transformation graphically. Olive (2005a) suggested using residual, response, RR, and FF plots to detect outliers, while Hawkins and Olive (2002, pp. 141, 158) suggested using the RR and FF plots. The four plots are best for $n \geq 5p$. Olive (2008: ∮ 6.4, 2017a: ch. 5–9) showed that the residual and response plots are useful for experimental design models. Park et al. (2012) showed response plots are competitive with the best robust regression methods for outlier detection on some outlier data sets that have appeared in the literature.

Olive (2004b, 2013b) used response plots for 1D regression, including generalized linear models and generalized additive models. Olive (2017a) covered 1D regression, multiple linear regression, and prediction intervals using the shorth. Olive and Hawkins (2003) showed that regression residuals behave well.

Rousseeuw and Zomeren (1990) suggested that Mahalanobis distances based on "robust estimators" of location and dispersion can be more useful than the distances based on the sample mean and covariance matrix. They show that a plot of robust Mahalanobis distances RD$_i$ versus residuals from "robust regression" can be useful.

Hampel et al. (1986, pp. 96–98) and Donoho and Huber (1983) provided some history for breakdown. Maguluri and Singh (1997) gave interesting examples on breakdown. Morgenthaler (1989) and Stefanski (1991) conjectured that high breakdown estimators with high efficiency are not possible. Theorems 14.6, 14.9, and 14.11 show that these conjectures are false. The cross-checking estimator uses the classical estimator if it is "close" to an impractical high breakdown consistent estimator and uses the high breakdown estimator, otherwise. See He and Portnoy (1992). The estimator in Theorem 14.6 is similar and has problems since the practical robust estimator is bad, and it is hard to define "close" when the robust and classical estimators are not estimating the same $\boldsymbol{\beta}$, e.g., if the error distribution is not symmetric.

This paragraph will explain why `mbareg` needs large samples to give a good estimate of the constant for highly skewed error distributions. Note that the LMS, LTA, and LMS criteria use half sets. For simplicity, consider the LMS criterion that minimizes the median squared residual. Heuristically, for highly right skewed data, let the "left tail half set" shift the constant of the OLS hyperplane down so that the half set of cases closest to the plane are the half set with the smallest OLS residual values. These cases will have negative residuals and residuals close to zero, which are roughly the cases corresponding to the half set of errors in the left tail of the error distribution. Let the "OLS half set" correspond to the half set of cases with the smallest absolute OLS residuals, so the cases closest to the OLS hyperplane. Since the distribution is highly right skewed, the "OLS half set" has much more variability than the "left tail half set." (For the location model, OLS is the sample mean which is greater than the sample median for right skewed data. The "left tail half set" shifts the mean down to the midpoint c of the minimum value and the median value, and often $c \approx$ median/2 if the support of the highly right skewed distribution is $(0, \infty)$.) A trial fit that uses the same OLS slope estimates but which shifts the intercept down to use the "left tail half set" will have a smaller median squared residual than the median squared residual using OLS. The trial fits for `mbareg` are $\sqrt{n}$ consistent, so estimate the OLS intercept eventually. However, the trial fit that uses 1% of the data has less than 1% efficiency since the $\boldsymbol{x}$ values are close in distance rather than spread out. Unless the sample size is large, the `mbareg` estimator tends to produce some trial fits that shift the intercept down, and one of these trial fits is selected to be the final `mbareg` estimator since it has a smaller median squared residual than the other trial fits.

The TV estimator was proposed by Olive (2002, 2005a) and is similar to an estimator proposed by Rousseeuw and 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) suggested a method of adaptive trimming such that the resulting estimator is asymptotically equivalent to the OLS estimator.

Introductions to 1D regression and regression graphics are Cook and Weisberg (1999a, ch. 18, 19, and 20) and Cook and Weisberg (1999b), while Olive (2008: ch. 12, 2010: ch. 15, 2017a) considers 1D regression. Chang (2006) and Chang and Olive (2007, 2010) extended least squares theory to 1D regression models where $h(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x}$.

If n is not much larger than p, then Hoffman et al. (2015) gave a robust Partial Least Squares–Lasso type estimator that uses a clever weighting scheme.

14.7.1 More on Robust Regression

Theorems similar to Theorems 14.6 and 14.9 are easy to derive, but depend on the strong assumption that the robust estimator that minimizes the robust criterion and OLS estimate the same $\boldsymbol{\beta}$. The basic resampling `lmsreg` estimator is an inconsistent zero breakdown estimator by Hawkins and Olive (2002), but the modification in Theorem 14.11 ii) is HB and asymptotically equivalent to OLS for a large class of zero mean finite variance symmetric error distributions. Hence the modified estimator has a $\sqrt{n}$ rate which is higher than the $n^{1/3}$ rate of the LMS estimator. The maximum bias function of the resulting estimator is not the same as that of the LMS estimator.

Theorem 14.11. Suppose that the algorithm uses $K_n \equiv K$ randomly selected elemental starts (e.g., K = 500) with k concentration steps and the two additional attractors $\hat{\boldsymbol{\beta}}_{OLS}$ and $\boldsymbol{b}_{k,B}$. Assume that $\hat{\boldsymbol{\beta}}_{LTS}$ and $\hat{\boldsymbol{\beta}}_{OLS}$ are both consistent estimators of $\boldsymbol{\beta}$.

i) Then the resulting CLTS estimator is a $\sqrt{n}$ consistent HB estimator if $\hat{\boldsymbol{\beta}}_{OLS}$ is $\sqrt{n}$ consistent, and the estimator is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{OLS}$.

ii) Suppose that a HB criterion is used on the $K + 2$ attractors such that the resulting estimator is HB if a HB attractor is used. Also, assume that the global minimizer of the HB criterion is a consistent estimator for $\boldsymbol{\beta}$ (e.g., LMS or LTA). The resulting HB estimator is asymptotically equivalent to the OLS estimator.

Proof. i) Theorems 14.4 and 14.5 show that a LTS concentration algorithm that uses a HB start is HB and that $\boldsymbol{b}_{k,B}$ is a HB biased estimator. The LTS estimator is consistent by Mašiček (2004). As $n \to \infty$, consistent estimators $\hat{\boldsymbol{\beta}}$ satisfy $Q_{LTS}(\hat{\boldsymbol{\beta}})/n - Q_{LTS}(\boldsymbol{\beta})/n \to 0$ in probability. Since $\boldsymbol{b}_{k,B}$ is a biased estimator of $\boldsymbol{\beta}$, OLS will have a smaller criterion value with probability tending to one. With probability tending to one, OLS will also have a smaller criterion value than the criterion value of the attractor from a randomly drawn elemental set (by He and Portnoy (1992)). Since K randomly chosen elemental sets are used, the CLTS estimator is asymptotically equivalent to OLS.

ii) As in the proof of i), the OLS estimator will minimize the criterion value with probability tending to one as $n \to \infty$. □

Researchers are starting to use intelligently chosen trial fits. Maronna and Yohai (2015) used OLS and 500 elemental sets as the 501 trial fits to produce an FS estimator used as the initial estimator for an FMM estimator. Since the 501 trial fits are zero breakdown, so is the FS estimator. Since the FMM estimator has the same breakdown as the initial estimator, the FMM estimator is zero breakdown. For regression, they show that the FS estimator is consistent on a large class of zero mean finite variance symmetric distributions. The theory is very similar to that of Theorems 14.6, 14.9, and 14.11. Consistency follows since the elemental fits and OLS are unbiased estimators of $\boldsymbol{\beta}_{OLS}$, but an elemental fit is an OLS fit to p cases. Hence the elemental fits are very variable, and the probability that the OLS fit has a smaller S estimator criterion than a randomly chosen elemental fit (or K randomly chosen elemental fits) goes to one as $n \to \infty$. (OLS and the S estimator are both $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$, so the ratio of their criterion values goes to one, and the S estimator minimizes the criterion value.) Hence the FMM estimator is asymptotically equivalent to the MM estimator that has the smallest criterion value for a large class of iid zero mean finite variance symmetric error distributions. This FMM estimator is asymptotically equivalent to the FMM estimator that uses OLS as the initial estimator. When the error distribution is skewed, the S estimator and OLS population constant are not the same, and the probability that an elemental fit is selected is close to one for a skewed error distribution as $n \to \infty$. (The OLS estimator $\hat{\boldsymbol{\beta}}$ gets very close to $\boldsymbol{\beta}_{OLS}$, while the elemental fits are highly variable unbiased estimators of $\boldsymbol{\beta}_{OLS}$, so one of the elemental fits is likely to have a constant that is closer to the S estimator constant while still having good slope estimators.) Hence the FS estimator is inconsistent, and the FMM estimator is likely inconsistent for skewed distributions. No practical method is known for computing a $\sqrt{n}$ consistent FS or FMM estimator that has the same breakdown and maximum bias function as the S or MM estimator that has the smallest S or MM criterion value.

Olive (2008, ch. 7, 8, 9, 12, and 13) contains much more information about 1D regression and resistant and robust regression. The least squares response and residual plots are very useful for detecting outliers. For more on the behavior of fits from randomly selected elemental sets, see Hawkins and Olive (2002) and Olive and Hawkins (2007a). For the `hbreg` estimator, see Olive and Hawkins (2011). More on the MBA LATA estimator can be found in Olive (2008, ch. 8).

For the dominant robust statistics paradigm, most of the problems for high breakdown multivariate location and dispersion, discussed in Section 4.9, are also problems for high breakdown regression: the impractical “brand-name” estimators have at least $O(n^p)$ complexity, while the practical estimators used in the software have not been shown to be both high breakdown and consistent. See Hawkins and Olive (2002), Hubert et al. (2002), and Maronna and Yohai (2002). Huber and Ronchetti (2009, pp. xiii, 8–9, 152–154, 196–197) suggested that high breakdown regression estimators do not provide an adequate remedy for the ill effects of outliers, that their statistical and computational properties are not adequately understood, that high breakdown estimators "break down for all except the smallest regression problems by failing to provide a timely answer!", and that "there are no known high breakdown point estimators of regression that are demonstrably stable."

A large number of impractical high breakdown regression estimators have been proposed, including LTS, LMS, LTA, S, LQD, τ, constrained M, repeated median, cross-checking, one step GM, one step GR, t-type, and regression depth estimators. See Rousseeuw and Leroy (1987) and Maronna et al. (2006). The practical algorithms used in the software use a brand-name criterion to evaluate a fixed number of trial fits and should be denoted as an F-brand-name estimator such as FLTS. Two stage estimators, such as the MM estimator, that need an initial consistent high breakdown estimator often have the same breakdown value and consistency rate as the initial estimator. These estimators are typically implemented with a zero breakdown inconsistent initial estimator and hence are zero breakdown with zero efficiency.

The practical estimators can be i) used in RR and FF plots for outlier detection or ii) used as $\hat{\boldsymbol{\beta}}_A$ to create an `hbreg` estimator that is asymptotically equivalent to least squares on a large class of symmetric error distributions.

Some of the theory for the impractical robust regression estimators is useful for determining when the robust estimator and OLS estimate the same $\boldsymbol{\beta}$. Many regression estimators $\hat{\boldsymbol{\beta}}$ satisfy

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(0, V(\hat{\boldsymbol{\beta}}, F)\ \boldsymbol{W}) \tag{14.28}$$

when $\dfrac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{W}^{-1}$, and when the errors e_i are iid with a cdf F and a unimodal pdf f that is symmetric with a unique maximum at 0. When the variance $V(e_i)$ exists,

$$V(OLS, F) = V(e_i) = \sigma^2 \ \text{ while } \ \mathrm{V(L_1, F)} = \frac{1}{4[\mathrm{f(0)}]^2}.$$

See Bassett and Koenker (1978).

Theorem 14.12. Under regularity conditions similar to those in Conjecture 14.1 below, a) the LMS(τ) converges at a cubed root rate to a non-Gaussian limit. b) The estimator $\hat{\boldsymbol{\beta}}_{LTS}$ satisfies Equation (14.28) and

$$V(LTS(\tau), F) = \frac{\int_{F^{-1}(1/2-\tau/2)}^{F^{-1}(1/2+\tau/2)} w^2 dF(w)}{[\tau - 2F^{-1}(1/2+\tau/2) f(F^{-1}(1/2+\tau/2))]^2}. \tag{14.29}$$

The proof of Theorem 14.12a is given in Kim and Pollard (1990). The proof of b) is given in Mašiček (2004), Čížek (2006), and Víšek (2006).

Conjecture 14.1. Let the iid errors e_i have a cdf F that is continuous and strictly increasing on its interval support with a symmetric, unimodal, differentiable density f that strictly decreases as $|x|$ increases on the support.

Then the estimator $\hat{\boldsymbol{\beta}}_{LTA}$ satisfies Equation (14.28) and

$$V(LTA(\tau), F) = \frac{\tau}{4[f(0) - f(F^{-1}(1/2+\tau/2))]^2}. \tag{14.30}$$

See Tableman (1994b, p. 392) and Hössjer (1994).

Čížek (2008) showed that LTA is $\sqrt{n}$ consistent, but did not prove that LTA is asymptotically normal. *Assume Conjecture 14.1 is true for the following LTA remarks in this section.* Then as $\tau \to 1$, the efficiency of LTS approaches that of OLS and the efficiency of LTA approaches that of L_1. Hence for τ close to 1, LTA will be more efficient than LTS when the errors come from a distribution for which the sample median is more efficient than the sample mean (Koenker and Bassett 1978). The results of Oosterhoff (1994) suggest that when $\tau = 0.5$, LTA will be more efficient than LTS only for sharply peaked distributions such as the double exponential. To simplify computations for the asymptotic variance of LTS, we will use truncated random variables (see Definition 1.8).

Lemma 14.13. Under the symmetry conditions given in Conjecture 14.1,

$$V(LTS(\tau), F) = \frac{\tau \sigma^2_{TF}(-k,k)}{[\tau - 2kf(k)]^2} \tag{14.31}$$

and

$$V(LTA(\tau), F) = \frac{\tau}{4[f(0) - f(k)]^2} \tag{14.32}$$

where

$$k = F^{-1}(0.5 + \tau/2). \tag{14.33}$$

Proof. Let W have cdf F and pdf f. Suppose that W is symmetric about zero, and by symmetry, $k = F^{-1}(0.5 + \tau/2) = -F^{-1}(0.5 - \tau/2)$. If W has been truncated at $a = -k$ and $b = k$, then the variance of the truncated random variable W_T is $V(W_T) = \sigma^2_{TF}(-k, k) = \dfrac{\int_{-k}^{k} w^2 dF(w)}{F(k) - F(-k)}$ by Definition 1.8. Hence

$$\int_{F^{-1}(1/2-\tau/2)}^{F^{-1}(1/2+\tau/2)} w^2 dF(w) = \tau\sigma^2_{TF}(-k, k)$$

and the result follows from the definition of k.

This result is useful since formulas for the truncated variance have been given in Section 1.7. The following examples illustrate the result. See Hawkins and Olive (1999b). The *mpack* functions `cltv`, `deltv`, and `nltv` are useful for computing the asymptotic variance of the LTS and LTA estimators for the Cauchy, double exponential, and normal error distributions, as given in the following three examples. See Problems 14.6, 14.7, and 14.8.

Example 14.13: N(0,1) Errors. If Y_T is a $N(0, \sigma^2)$ truncated at $a = -k\sigma$ and $b = k\sigma$, $V(Y_T) = \sigma^2[1 - \dfrac{2k\phi(k)}{2\Phi(k) - 1}]$. At the standard normal

$$V(LTS(\tau), \Phi) = \frac{1}{\tau - 2k\phi(k)} \tag{14.34}$$

$$\text{while} \quad \mathrm{V(LTA(\tau), \Phi)} = \frac{\tau}{4[\phi(0) - \phi(\mathrm{k})]^2} = \frac{2\pi\tau}{4[1 - \exp(-\mathrm{k}^2/2)]^2} \tag{14.35}$$

where ϕ is the standard normal pdf and $k = \Phi^{-1}(0.5 + \tau/2)$. Thus for $\tau \geq 1/2$, LTS(τ) has breakdown value of $1 - \tau$ and Gaussian efficiency

$$\frac{1}{V(LTS(\tau), \Phi)} = \tau - 2k\phi(k). \tag{14.36}$$

The 50% breakdown estimator LTS(0.5) has a Gaussian efficiency of 7.1%. If it is appropriate to reduce the amount of trimming, we can use the 25% breakdown estimator LTS(0.75) which has a much higher Gaussian efficiency of 27.6% as reported in Ruppert (1992, p. 255). Also see the column labeled "Normal" in table 1 of Hössjer (1994).

Example 14.14: Double Exponential Errors. The double exponential (Laplace) distribution is interesting since the L_1 estimator corresponds to maximum likelihood and so L_1 beats OLS, reversing the comparison of the normal case. For a double exponential $DE(0,1)$ random variable,

$$V(LTS(\tau), DE(0,1)) = \frac{2-(2+2k+k^2)\exp(-k)}{[\tau - k\exp(-k)]^2}$$

$$\text{while } \ \mathrm{V(LTA(\tau), DE(0,1))} = \frac{\tau}{4[0.5-0.5\exp(-\mathrm{k})]^2} = \frac{1}{\tau}$$

where $k = -\log(1-\tau)$. Note that LTA(0.5) and OLS have the same asymptotic efficiency at the double exponential distribution. Also see Tableman (1994a, b).

Example 14.15: Cauchy Errors. Although the L_1 estimator and the trimmed estimators have finite variance when the errors are Cauchy, the OLS estimator has infinite variance (because the Cauchy distribution has infinite variance). If X_T is a Cauchy $C(0,1)$ random variable symmetrically truncated at $-k$ and k, then $V(X_T) = \dfrac{k - \tan^{-1}(k)}{\tan^{-1}(k)}$. Hence

$$V(LTS(\tau), C(0,1)) = \frac{2k - \pi\tau}{\pi[\tau - \frac{2k}{\pi(1+k^2)}]^2}$$

$$\text{and } \ V(LTA(\tau), C(0,1)) = \frac{\tau}{4[\frac{1}{\pi} - \frac{1}{\pi(1+k^2)}]^2}$$

where $k = \tan(\pi\tau/2)$. The LTA sampling variance converges to a finite value as $\tau \to 1$ while that of LTS increases without bound. LTS(0.5) is slightly more efficient than LTA(0.5), but LTA pulls ahead of LTS if the amount of trimming is very small.

14.8 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

14.1. (See Aldrin et al. 1993.) Suppose

$$Y = m(\boldsymbol{\beta}^T \boldsymbol{x}) + e \tag{14.37}$$

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 $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{x}) = \text{Cov}(\boldsymbol{w})$. Let $\boldsymbol{r} = \boldsymbol{w} - (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w}$.

a) Recall that $\text{Cov}(\boldsymbol{x}, \boldsymbol{Y}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$ 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 $\text{Cov}(\boldsymbol{r}, z) = 0$ 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 $\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\eta} = 1$.)

14.2. Referring to Definition 14.4, let $\hat{Y}_{i,j} = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_j = \hat{Y}_i(\hat{\boldsymbol{\beta}}_j)$ and let $r_{i,j} = r_i(\hat{\boldsymbol{\beta}}_j)$. Show that $\|r_{i,1} - r_{i,2}\| = \|\hat{Y}_{i,1} - \hat{Y}_{i,2}\|$.

14.3. Assume that the model has a constant β_1 so that the first column of $\boldsymbol{X}$ is $\boldsymbol{1}$. Show that if the regression estimator is regression equivariant, then adding $\boldsymbol{1}$ to $\boldsymbol{Y}$ changes $\hat{\beta}_1$ but does not change the slopes $\hat{\beta}_2, ..., \hat{\beta}_p$.

R 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, e.g., *trviews*, will display the code for the function. Use the `args` command, e.g., *args(trviews)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/mrsashw.txt) into *R*.

14.4. Paste the command for this problem into *R* to produce the second column of Table 14.1. Include the output in *Word.*

14.5. a) To get an idea for the amount of contamination, a basic resampling or concentration algorithm can tolerate, enter or download the `gamper` function (with the *source("G:/mpack.txt")* command) that evaluates Equation (14.4) at different values of $h = p$.

b) Next enter the following commands and include the output in *Word.*

```
zh <- c(10,20,30,40,50,60,70,80,90,100)
for(i in 1:10) gamper(zh[i])
```

The "asymptotic variance" for LTA in Problems 14.6, 14.7, and 14.8 is actually the conjectured asymptotic variance for LTA if the multiple linear regression model is used instead of the location model. See Section 14.7.

14.6. a) Download the *R* function `nltv` that computes the asymptotic variance of the LTS and LTA estimators if the errors are N(0,1).
b) Enter the commands *nltv(0.5), nltv(0.75), nltv(0.9),* and *nltv(0.9999).* Write a table to compare the asymptotic variance of LTS and LTA at these coverages. Does one estimator always have a smaller asymptotic variance?

14.7. a) Download the *R* function `deltv` that computes the asymptotic variance of the LTS and LTA estimators if the errors are double exponential DE(0,1).
b) Enter the commands *deltv(0.5), deltv(0.75), deltv(0.9),*, and *deltv(0.9999).* Write a table to compare the asymptotic variance of LTS and LTA at these coverages. Does one estimator always have a smaller asymptotic variance?

14.8. a) Download the *R* function `cltv` that computes the asymptotic variance of the LTS and LTA estimators if the errors are Cauchy C(0,1).

b) Enter the commands *cltv(0.5), cltv(0.75), cltv(0.9),* and *cltv(0.9999).* Write a table to compare the asymptotic variance of LTS and LTA at these coverages. Does one estimator always have a smaller asymptotic variance?

14.9*. a) If necessary, use the commands *source("G:/mpack.txt")* and *source("G:/mrobdata.txt").*

b) Enter the command *mbamv(belx,bely)* in *R.* Click on the rightmost mouse button (and in *R*, click on *Stop*). You need to do this seven times before the program ends. There is one predictor x and one response Y. The function makes a scatterplot of x and Y and cases that get weight one are shown as highlighted squares. Each MBA sphere covers half of the data. When you find a good fit to the bulk of the data, hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word.*

c) Enter the command *mbamv2(buxx,buxy)* in *R*. Click on the rightmost mouse button (and in *R*, click on *Stop*). You need to do this 14 times before the program ends. There are four predictors $x_1, ..., x_4$ and one response Y. The function makes the response and residual plots based on the OLS fit to the highlighted cases. Each MBA sphere covers half of the data. When you find a good fit to the bulk of the data, hold down the *Ctrl* and *c* keys to make a copy of the two plots. Then paste the plots in *Word*.

14.10. This problem compares the MBA estimator that uses the median squared residual MED(r_i^2) criterion with the MBA estimator that uses the LATA criterion. On clean data, both estimators are $\sqrt{n}$ consistent since both use 50 $\sqrt{n}$ consistent OLS estimators. The MED(r_i^2) criterion has trouble with data sets where the multiple linear regression relationship is weak and there is a cluster of outliers. The LATA criterion tries to give all x-outliers, including good leverage points, zero weight.

a) If necessary, use the commands *source("G:/mpack.txt")* and *source("G:/mrobdata.txt")*. The `mlrplot2` function is used to compute both MBA estimators. Use the rightmost mouse button to advance the plot (and in *R*, highlight stop).

b) Use the command *mlrplot2(belx,bely)* and include the resulting plot in *Word*. Is one estimator better than the other, or are they about the same?

c) Use the command *mlrplot2(cbrainx,cbrainy)* and include the resulting plot in *Word*. Is one estimator better than the other, or are they about the same? (The infants are likely good leverage cases instead of outliers.)

d) Use the command *mlrplot2(museum[,3:11],museum[,2])* and include the resulting plot in *Word*. For this data set, most of the cases are based on humans, but a few are based on apes. The MBA LATA estimator will often give the cases corresponding to apes larger absolute residuals than the MBA estimator based on MED(r_i^2), but the apes appear to be good leverage cases.

e) Use the command *mlrplot2(buxx,buxy)* until the outliers are clustered about the identity line in one of the two response plots. (This will usually happen within 10 or fewer runs. Pressing the "up arrow" will bring the previous command to the screen and save typing.) Then include the resulting plot in *Word*. Which estimator went through the outliers and which one gave zero weight to the outliers?

f) Use the command *mlrplot2(hx,hy)* several times. Usually both MBA estimators fail to find the outliers for this artificial Hawkins data set that is also analyzed by Atkinson and Riani (2000, section 3.1). The *lmsreg* estimator can be used to find the outliers. In *R*, use the commands *library(MASS)* and *ffplot2(hx,hy)*. Include the resulting plot in *Word*.

14.11. Use the following R 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 highlight *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 viewed, the output will show 10 estimated predictors. The last estimate is the OLS (least squares) view and might look like the following output.

```
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 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.

14.12. At the beginning of your *R* session, use *source("G:/mpack.txt")* command and *library(MASS).*

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 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 command "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 command "Paste." Click the rightmost mouse button and 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.

14.13. Warning: This problem may take too much time, but makes a good project. This problem is like Problem 14.12 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 14.9 and 14.10 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 (in *R*, first type `library(MASS)`). 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 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 command "Paste."

d) To make a plot like Figure 14.13, use the following commands. Let *tem* = $\hat{\boldsymbol{\beta}}$ obtained from the *trviews* output. In Example 14.9 (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

14.14. Warning: this problem may take too much time, but makes a good project. Repeat Problem 14.13 but replace `trviews` with a) `lmsviews`, b) `symviews` (that creates views that sometimes work even when symmetry is present), and c) `ctrviews`.

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 command "Paste."

14.15. a) In addition to the *source("G:/mpack.txt")* command, also use the *source("G:/mrobdata.txt")* command, and type the *library(MASS)* command).

b) Type the command *tvreg(buxx,buxy,ii=1).* Click the rightmost mouse button and 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.*

14.16. a) After entering the two *source* commands above Problem 14.4, enter the following command.

```
MLRplot(buxx,buxy)
```

Click the rightmost mouse button (and in *R* click on *Stop*). The response plot should appear. Again, click the rightmost mouse button (and in *R* click on *Stop*). The residual plot should appear. Hold down the *Ctrl* and *c* keys to make a copy of the two plots. Then paste the plots in *Word.*

b) The response variable is *height*, but 5 cases were recorded with heights about 0.75 inches tall. The highlighted squares in the two plots correspond to cases with large Cook's distances. With respect to the Cook's distances, what is happening, swamping, or masking?

c) *RR plots:* One feature of the MBA estimator is that it depends on the sample of seven centers drawn and changes each time the function is called. In ten runs, about seven plots will look like Figure 14.7, but in about three plots the MBA estimator will also pass through the outliers. Make the RR plot by pasting the commands for this problem into *R* and include the plot in *Word.*

d) *FF plots: the plots in the top row will cluster about the identity line if the MLR model is good or if the fit passes through the outliers.* Make the FF plot by pasting the commands for this problem into *R*, and include the plot in *Word.*

Problem using ARC

14.17. a) Activate the *insulation.lsp* data set of Example 14.12 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 and then click on OK. Again, from the insulation menu select *Make interactions*, click on {F}type and time^2 and 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 much of the author's research in multivariate analysis from 1997–2017, 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 et al. (1975, 1981) introduced the concentration technique. Rousseeuw (1984) extended the MCD location estimator to the MCD estimator of multivariate location and dispersion and the LTS estimator and popularized the LMS estimator. 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 brand-name HB robust estimators, such as MCD and MVE estimators, are impractical to compute. The S estimator is 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 practical estimator of multivariate location and dispersion has been shown to be both consistent and high breakdown for a large class of nonspherical distributions. (FCH is consistent and

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_15

T_{FCH} is high breakdown, but $\boldsymbol{C}_{FCH}$ is only conjectured to be HB. The Olive (2004a) MB estimator is HB, but not a consistent estimator of $c\boldsymbol{\Sigma}_{\boldsymbol{x}}$ except for a class of spherical distributions. The MB estimator is conjectured to be a consistent estimator of its population analog. See Conjecture 4.3.) Initial practical consistent HB regression estimators, such as `hbreg`, were first developed by Olive and Hawkins (2007b, 2008, 2011).

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, RFCH DD plot, RMVN 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) gave an estimator that has large sample theory. Better yet, use ideas from this book. See Chapter 9 and Rupasinghe Arachchige Don and Pelawa Watagoda (2017).

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 this text, Hawkins and Olive (2002) and Olive and Hawkins (2007b, 2008, 2010, 2011). 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 (www.sciencedirect.com), to the *Web of Science* (www.webofknowledge.com), or to the *Current Index to Statistics* (www.statindex.org). These resources allow you to search for literature by author, e.g., Olive, or by topic, e.g., robust statistics. Both of these methods search for recent papers.

The search engines (www.google.com), (www.ask.com), (www.msn.com), (www.yahoo.com), and (www.info.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*.

Websites for researchers or research groups can be useful. Rousseeuw, Hubert, Croux, and Van Aelst have lots of papers. STATLIB used to be very useful. 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)
```

Familiarity with a high level programming language such as *R* is essential. A very useful *R* link is (www.r-project.org/). See R Core Team (2016).

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 et al. (1983), Hamada and Sitter (2004), Rubin (2004), and Smith (1997).

15.2 R and Arc

R is the free version of *Splus* available from the **CRAN** website (https://cran.r-project.org/). The website (http://www.stat.umn.edu) has useful links for *Arc* which is the software developed by Cook and Weisberg (1999a). As of June 2017, the author's personal computer has Version 3.3.1 (June 21, 2016) of *R*, and Version 1.06 (July 2004) of *Arc*. Several of the text *R/Splus* functions and figures were made in the 1990s using *Splus* (see MathSoft 1999a, b) on a workstation.

Downloading the book's data.lsp files into Arc

Many of the data sets in the book's website (http://lagrange.math.siu.edu/Olive/multbk.htm) can easily be downloaded into the Cook and Weisberg (1999a) *Arc* software. 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*.

Many of the homework problems use *R* functions contained in the book's website (http://lagrange.math.siu.edu/Olive/multbk.htm) under the file name *mpack.txt*. The following two commands can be copied and pasted into *R* from near the top of the file (http://lagrange.math.siu.edu/Olive/mrsashw.txt).

Downloading the book's R functions *mpack.txt* and data files *mrobdata.txt* into *R*: The commands

```
source("http://lagrange.math.siu.edu/Olive/mpack.txt")
source("http://lagrange.math.siu.edu/Olive/mrobdata.txt")
```

can be used to download the *R* functions and data sets into *R*. Type *ls()*. Over 130 *R* 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 in *R*, but the functions and data are easily obtained with the source commands). For Windows, the files can be saved on a flash drive G, say. Then use the following commands.

```
source("G:/mpack.txt"); source("G:/mrobdata.txt")
```

The remainder of this section gives tips on using *R* but is no replacement for books such as Becker et al. (1988), Braun and Murdoch (2007), Crawley (2005, 2013), or Venables and Ripley (2003). Also see MathSoft (1999a, b) 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*.

Least squares regression is done with the function *lsfit* or *lm*.

The commands *help(fn)* and *args(fn)* give information about the function fn, e.g., 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* 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 flash drive from the webpage for this book, open *cyp.lsp* in *Word.* It has 76 rows and 8 columns. In *R*, 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.* 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
205.40825985   0.94653718   0.17514405   0.23415181
          X4           X5           X6
  0.75927197  -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.*

Making functions in R 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 this book. To remove unwanted items from the worksheet, e.g., 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.

Transferring Data to and from *Arc* and *R*.
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*. 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* and paste them below *new.lsp.* Then modify the file *new.lsp* and save it on a flash drive 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 below.

```
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 the source of the data, e.g., from the STATLIB website. Note that the first variable starts with *Col 0.*

To transfer a data set from Arc to R, select the item “Display data” from the data set’s menu. Select the variables you want to save, and then click 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 *R* using the *read.table* command. The file *bodfat.Rd* saves the documentation about the data set in a standard format for *R.*

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, e.g., 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 in the Weisberg (2002) dimension reduction library *dr*. Following Crawley (2013, p. 8), you may need to "Run as administrator" before you can install packages (right click on the *R* icon to find this). Then use the following command to install the *dr* package.

```
install.packages("dr")
```

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 foolproof. 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 Versions 2.4.1 and later. Also, some functions in *mpack* may no longer work in new versions of *R*.

15.3 Projects

Straightforward Projects

1) Run an *mpack* simulation function for a range of values of n, p, error distributions, estimators, data sets, etc. Functions problem pairs include (`rcovsim`, 4.3), (`concmv`, 4.5), (`ddmv`, 4.6), (`covesim`, 4.8), (`ddsim`, 5.2), (`corrsim`, 5.3), (`predsim`, 5.9), (`pcasim`, 6.9), (`mregsim`, 12.10), and (`mpredsim`, 12.11).

Also see the *mpack* functions `concsim`, `corrbootsim`, `corrsim2`, `covcheck`, `covsim2`, `ddsim3`, `drsim5`, `drsim6`, `drsim7`, `hbregsim`, `mbsim`, `medhotsim`, `mldsim`, `mldsim6`, `MLRsim`, `mregddsim`, `pcabootsim`, `regbootsim`, `rhotsim`, `rhotsim2`, `rmbsim`, `rmpredsim`, `rmregbootsim`, `rmregddsim`, `rmregsim`, `vsbootsim`, and `vsbootsim2`.

2) 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.

3) Read Bentler and Yuan (1998) and Cattell (1966). These papers used 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.

4) 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.

5) 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, RMVN, 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?

6) Examine the DD plot as a diagnostic for multivariate normality and elliptically contoured distributions. Use real and simulated data.

7) Resistant regression: examine `tvreg` by using OLS–FCH instead of OLS–cov.mcd. (L_1–cov.mcd and L_1–covfch are also interesting.) Two other projects would use RFCH or RMVN. Use type=3 for FCH and type=4 for RMVN.

8) *Using ESP to Search for the Missing Link*: Compare `trimmed views` which uses OLS and `cov.mcd` with another regression–MLD combo. There are 5 possible projects: i) OLS–FCH, ii) OLS–Classical (use `ctrviews`), iii) lmsreg–cov.mcd (`lmsviews`), iv) lmsreg–FCH, and v) lmsreg–classical. Do Problem 14.12ac (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 with `cov.mcd` is better, and when the picked procedure i)–v) is better. If you can invent interesting 1D functions, do so. See Problem 14.13.

9) 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.

10) Find some large data sets or data sets with $p > n$ and try to detect outliers using $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p) = \|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|$, the Euclidean distance of $\boldsymbol{x}_i$ from the coordinatewise median MED($\boldsymbol{W}$). Also use `covmb2` and `ddplot5`. See Section 4.7.

11) Study the diagnostics in and below Problem 5.10. See Remark 5.8.

Harder Projects

12) Which estimator is better FCH, RFCH, CMBA, or RCMBA?

13) For large data sets, make the DD plot of the DGK estimator vs. the 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 effective outlier diagnostics for iid multivariate data.

14) *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 for a large class of elliptically contoured distributions that have a 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.

15) 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) showed 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.

16) Compare outliers and missing values, especially missing and outlying at random. See Little and Rubin (2002).

17) Suppose that the data set contains missing values. Code the missing value as ±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).

18) 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 FLTS 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 LTS 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.) A similar conjecture exists for the FMCD concentration algorithm.
- 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 Hawkin 2007b, 2008.)

  It is easy to create consistent algorithm estimators that use $O(n)$ randomly chosen elemental sets. He and Wang (1997) showed that the all elemental subset approximation to S estimators for MLD is consistent for $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ (likely for a large class of elliptically contoured distributions). 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 2007b, 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. For regression, cross-checking is likely to run into problems when the error distribution is not symmetric.
- Does the bootstrap prediction region method of Section 5.3 work under mild conditions for variable selection methods? Are the Machado and Parente (2005) sufficient conditions for estimating an asymptotic covariance matrix of a statistic also sufficient conditions for the prediction region method?

15.4 Hints for Selected Problems

Chapter 1

1.1 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
2.8 Several of the marginal relationships are nonlinear, including $E(M|H)$.

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

3.20 a)

$$N_2\left(\begin{pmatrix} 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} \right).$$

b) $X_2 \perp\!\!\!\perp X_4$ and $X_3 \perp\!\!\!\perp X_4$.
c) $\dfrac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{33}}} = \dfrac{1}{\sqrt{2}\sqrt{3}} = 1/\sqrt{6} = 0.4082$.

Chapter 4

4.4 a) The 4 plots should look nearly identical to 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 c) This is simulated data, but for most plots, the slope is near 2 to 2.5.

Chapter 6

6.1 Note that $o_P(1)O_P(1)$=$[(\hat{\boldsymbol{\Sigma}} - \hat{\lambda}_i) - c(\boldsymbol{\Sigma} - \lambda_i)]\hat{\boldsymbol{e}}_i = c(\boldsymbol{\Sigma} - \lambda_i)\hat{\boldsymbol{e}}_i \xrightarrow{P} \boldsymbol{0}$.

Chapter 8

8.5 See Example 8.6.

Chapter 11

11.2. See Example 11.5.

Chapter 14

14.1

a) Since Y is a (random) scalar and $E(\boldsymbol{w}) = \boldsymbol{0}$, $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(Y - E(Y))^T] = E[\boldsymbol{w}(Y - E(Y))] = E(\boldsymbol{w}Y) - E(\boldsymbol{w})E(Y) = E(\boldsymbol{w}Y)$.

b) Using the definition of z and $\boldsymbol{r}$, note that $Y = m(z) + e$ and $\boldsymbol{w} = \boldsymbol{r} + (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w}$. Hence $E(\boldsymbol{w}Y) = E[(\boldsymbol{r} + (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})(m(z) + e)] = E[(\boldsymbol{r} + (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})m(z)] + E[\boldsymbol{r} + (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w}]E(e)$ since e is independent of $\boldsymbol{x}$. Since $E(e) = 0$, the latter term drops out. Since $m(z)$ and $\boldsymbol{\beta}^T\boldsymbol{w}m(z)$ are (random) scalars, $E(\boldsymbol{w}Y) = E[m(z)\boldsymbol{r}] + E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}$.

c) Using result b), $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}E[m(z)\boldsymbol{r}] + \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}$ $= E[\boldsymbol{\beta}^T\boldsymbol{w}\,m(z)]\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} + \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}E[m(z)\boldsymbol{r}] = E[\boldsymbol{\beta}^T\boldsymbol{w}\,m(z)]\boldsymbol{\beta} + \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}E[m(z)\boldsymbol{r}]$ and the result follows.

d) $E(\boldsymbol{w}z) = E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{x}^T\boldsymbol{\beta}] = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x}^T - E(\boldsymbol{x}^T) + E(\boldsymbol{x}^T))\boldsymbol{\beta}]$ $= E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x}^T - E(\boldsymbol{x}^T))]\boldsymbol{\beta} + E[\boldsymbol{x} - E(\boldsymbol{x})]E(\boldsymbol{x}^T)\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}$.

e) If $m(z)$=z, then $c(\boldsymbol{x})$=$E(\boldsymbol{\beta}^T\boldsymbol{w}z)$=$\boldsymbol{\beta}^T E(\boldsymbol{w}z) = \boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$ by result d).

f) Since z is a (random) scalar, $E(z\boldsymbol{r}) = E(\boldsymbol{r}z) = E[(\boldsymbol{w} - (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})z] =$ $E(\boldsymbol{w}z) - (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T E(\boldsymbol{w}z)$. Using result d), $E(\boldsymbol{r}z) = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} =$ $\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = \boldsymbol{0}$.

g) Since z and $\boldsymbol{r}$ are linear combinations of $\boldsymbol{x}$, the joint distribution of z and $\boldsymbol{r}$ is multivariate normal. Since $E(\boldsymbol{r}) = \boldsymbol{0}$, z and $\boldsymbol{r}$ are uncorrelated and thus independent. Hence $m(z)$ and $\boldsymbol{r}$ are independent and $\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}E[m(z)\boldsymbol{r}] =$ $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}E[m(z)]E(\boldsymbol{r}) = \boldsymbol{0}$.

14.2 $\|r_{i,1} - r_{i,2}\| = \|Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_1 - (Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_2)\| = \|\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_2 - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_1\| =$ $\|\hat{Y}_{2,i} - \hat{Y}_{1,i}\| = \|\hat{Y}_{1,i} - \hat{Y}_{2,i}\|$.

14.3 Adding $\boldsymbol{1}$ to $\boldsymbol{Y}$ is equivalent to using $\boldsymbol{u} = (1, 0, ..., 0)^T$ in Equation (14.9), and the result follows.

14.9 b) The line should go through the left and right cluster but not through the middle cluster of outliers.

c) The identity line should NOT PASS through the cluster of outliers with Y near 0 and the residuals corresponding to these outliers should be large in magnitude.

14.10 e) Usually the MBA estimator based on the median squared residual will pass through the outliers, while the MBA LATA estimator gives zero weight to the outliers (so that the outliers are large in magnitude).

14.11. a) No strong nonlinearities for MVN data but there should be some nonlinearities present for the non–EC data.

b) The plot should look like a cubic function.

c) The plot should use 0% trimming and resemble the plot in b), but may not be as smooth.

d) The plot should be linear, and for many students, some of the trimmed views should be better than the OLS view.

e) The response plot should look like a cubic with trimming greater than 0%.

f) The plot should be linear.

14.12. b) and c) It is possible that none of the trimmed views look much like the sinc(ESP) = sin(ESP)/ESP function.

d) Now at least one of the trimmed views should be good.

e) More lmsreg trimmed views should be good than the views from the other two methods, but since simulated data is used, one of the plots from b) or c) could be as good or even better than the plot in d).

14.15 b) 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.

14.16 b) Masking since three outliers are good cases with respect to Cook's distances.

c) and d) Usually, the MBA residuals will be large in magnitude, but for some students MBA, ALMS, and ALTS will be highly correlated.

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 were 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

References

Aggarwal, C. C. (2017). *Outlier analysis* (2nd ed.). New York: Springer.

Agulló, J. (1996). Exact iterative computation of the multivariate minimum volume ellipsoid estimator with a branch and bound algorithm. In A. Prat (Ed.), *Proceedings in computational statistics* (pp. 175–180). Heidelberg: Physica.

Agulló, J. (1998). Computing the minimum covariance determinant estimator, unpublished manuscript, Universidad de Alicante.

Aldrin, M., Bølviken, E., & Schweder, T. (1993). Projection pursuit regression for moderate non-linearities. *Computational Statistics & Data Analysis*, *16*, 379–403.

Alkenani, A., & Yu, K. (2013). A comparative study for robust canonical correlation methods. *Journal of Statistical Computation and Simulation*, *83*, 690–718.

Alqallaf, F. A., Konis, K. P., Martin, R. D., & 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* (pp. 14–23). Edmonton: ACM.

Alrawashdeh, M. J., Sabri, S. R. M., & Ismail, M. T. (2012). Robust linear discriminant analysis with financial ratios in special interval. *Applied Mathematical Sciences*, *6*, 6021–6034.

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). *An introduction to multivariate statistical analysis* (2nd ed.). New York: Wiley.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). New York: Wiley.

Arcones, M. A. (1995). Asymptotic normality of multivariate trimmed means. *Statistics & Probability Letters*, *25*, 43–53.

Ash, R. B. (1972). *Real analysis and probability.* San Diego: Academic Press.

Atkinson, A., & Riani, R. (2000). *Robust diagnostic regression analysis.* New York: Springer.

Atkinson, A., Riani, R., & Cerioli, A. (2004). *Exploring multivariate data with the forward search.* New York: Springer.

Bali, J. L., Boente, G., Tyler, D. E., & Wang, J. L. (2011). Robust functional principal components: A projection-pursuit approach. *The Annals of Statistics*, *39*, 2852–2882.

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2

Bassett, G. W., & Koenker, R. W. (1978). Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, *73*, 618–622.

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). *The new S language: A programming environment for data analysis and graphics*. Pacific Grove: Wadsworth and Brooks/Cole.

Becker, R. A., & Keller-McNulty, S. (1996). Presentation myths. *The American Statistician*, *50*, 112–115.

Bentler, P. M., & Yuan, K. H. (1998). Tests for linear trend in the smallest eigenvalues of the correlation matrix. *Psychometrika*, *63*, 131–144.

Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, *83*, 686–697.

Beran, R., & Srivastava, M. S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *The Annals of Statistics*, *35*, 95–115.

Berk, K. N. (1978). Comparing subset regression procedures. *Technometrics*, *20*, 1–6.

Berndt, E. R., & Savin, N. E. (1977). Conflict among criteria for testing hypotheses in the multivariate linear regression model. *Econometrika*, *45*, 1263–1277.

Bernholt, T. (2005). Computing the least median of squares estimator in time $O(n^d)$. In *Proceedings of ICCSA 2005* (Vol. 3480, pp. 697–706). LNCS.

Bernholt, T., & Fischer, P. (2004). The complexity of computing the MCD-estimator. *Theoretical Computer Science*, *326*, 383–398.

Bhatia, R., Elsner, L., & 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.

Bianco, A. M., Boente, G., & Rodrigues, I. M. (2017). Conditional tests for elliptical symmetry using robust estimators. *Communications in Statistics: Theory and Methods*, *46*, 1744–1765.

Bickel, P. J., & Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 1196–1217.

Bickel, P. J., & Ren, J.-J. (2001). The bootstrap in hypothesis testing. In M. de Gunst, C. Klaassen, & A. van der Vaart (Eds.), *State of the Art in Probability and Statistics: Festschrift for William R. van Zwet* (pp. 91–112). Hayward: The Institute of Mathematical Statistics.

Billor, N., Hadi, A., & 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*. New York: Springer.

Bloch, D. A., & Gastwirth, J. L. (1968). On a simple estimate of the reciprocal of the density function. *The Annals of Mathematical Statistics*, *39*, 1083–1085.

Boente, G. (1987). Asymptotic theory for robust principal components. *Journal of Multivariate Analysis*, *21*, 67–78.

Boente, G., Fraiman, R., Discussion of 'robust principal component analysis for functional data' by Locantore et al. (1999). *Test*, *8*, 28–35.

Bogdan, M. (1999). Data driven smooth tests for bivariate normality. *Journal of Multivariate Analysis*, *68*, 26–53.

Boudt, K., Rousseeuw, P. J., Vanduffel, S., & Verdonck, T. (2017). The minimum regularized covariance determinant estimator, unpublished document at arXiv:1701.07086.pdf.

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.

Braun, W. J., & Murdoch, D. J. (2007). *A first course in statistical programming with R*. New York: Cambridge University Press.

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 P. J. Bickel, K. A. Doksum, & J. L. Hodges (Eds.), *A Festschrift for Erich L. Lehmann* (pp. 97–114). Pacific Grove: Wadsworth.

Brown, B. L., Hendrix, S. B., Hedges, D. W., & Smith, T. B. (2012). *Multivariate analysis for the biobehavioral and social sciences, a graphical approach.* Hoboken: Wiley.

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, *30*, 927–961.

Budny, K. (2014). A generalization of Chebyshev's inequality for Hilbert-space-valued random variables. *Statistics & Probability Letters*, *88*, 62–65.

Butler, R. W., Davies, P. L., & 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., & Liu, W. (2011). A direct approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, *106*, 1566–1577.

Cambanis, S., Huang, S., & Simons, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, *11*, 368–385.

Carroll, R. J., & Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society, B*, *55*, 693–706.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Belmont: Duxbury.

Cator, E. A., & Lopuhaä, H. P. (2010). Asymptotic expansion of the minimum covariance determinant estimators. *Journal of Multivariate Analysis*, *101*, 2372–2388.

Cator, E. A., & Lopuhaä, H. P. (2012). Central limit theorem and influence function for the MCD estimators at general multivariate distributions. *Bernoulli*, *18*, 520–551.

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., & Tukey, P. (1983). *Graphical methods for data analysis.* Boston: Duxbury.

Chang, J. (2006). Resistant dimension reduction, Ph.D. thesis, Southern Illinois University. http://lagrange.math.siu.edu/Olive/sjingth.pdf.

Chang, J., & Olive, D. J. (2007). Resistant dimension reduction. http://lagrange.math.siu.edu/Olive/preprints.htm.

Chang, J., & Olive, D. J. (2010). OLS for 1D regression models. *Communications in Statistics: Theory and Methods*, *39*, 1869–1882.

Chen, C., He, X., & Wei, Y. (2008). Lower rank approximation of matrices based on fast and robust alternating regression. *Journal of Computational and Graphical Statistics*, *17*, 186–200.

Chen, X. (2011). A new generalization of Chebyshev inequality for random vectors. arXiv:0707.0805v2.

Chew, V. (1966). Confidence, prediction and tolerance regions for the multivariate normal distribution. *Journal of the American Statistical Association*, *61*, 605–617.

Chmielewski, M. A. (1981). Elliptically symmetric distributions: A review and bibliography. *International Statistical Review*, *49*, 67–74.

Čížek, P. (2006). Least trimmed squares under dependence. *Journal of Statistical Planning and Inference*, *136*, 3967–3988.

Čížek, P. (2008). General trimmed estimation: Robust approach to nonlinear and limited dependent variable models. *Econometric Theory*, *24*, 1500–1529.

Collett, D. (1999). *Modelling binary data* (1st ed.). Boca Raton: Chapman & Hall/CRC.

Collett, D. (2003). *Modelling binary data* (2nd ed.). Boca Raton: Chapman & Hall/CRC.

Cook, R. D. (1977). Deletion of influential observations in linear regression. *Technometrics*, *19*, 15–18.

Cook, R. D. (1998). *Regression graphics: Ideas for studying regression through graphics.* New York: Wiley.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science*, *22*, 1–26. (with discussion).

Cook, R. D., & 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., & 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., & Weisberg, S. (1993). Exact iterative computation of the robust multivariate minimum volume ellipsoid estimator. *Statistics & Probability Letters*, *16*, 213–218.

Cook, R. D., Helland, I. S., & Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society, B*, *75*, 851–877.

Cook, R. D., & Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics*, *30*, 455–474.

Cook, R. D., & Nachtsheim, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association*, *89*, 592–599.

Cook, R. D., & Olive, D. J. (2001). A note on visualizing response transformations in regression. *Technometrics*, *43*, 443–449.

Cook, R. D., & Setodji, C. M. (2003). A model-free test for reduced rank in multivariate regression. *Journal of the American Statistical Association*, *98*, 340–351.

Cook, R. D., & Su, Z. (2013). Scaled envelopes: Scale-invariant and efficient estimation in multivariate linear regression. *Biometrika*, *100*, 929–954.

Cook, R. D., & Weisberg, S. (1999a). *Applied regression including computing and graphics.* New York: Wiley.

Cook, R. D., & Weisberg, S. (1999b). Graphs in statistical analysis: is the medium the message? *The American Statistician*, *53*, 29–37.

Cook, R. D., & Zhang, X. (2015). Foundations of envelope models. *Journal of the American Statistical Association*, *110*, 599–611.

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: Princeton University Press.

Crawley, M. J. (2005). *Statistics an introduction using R.* Hoboken: Wiley.

Crawley, M. J. (2013). *The R book* (2nd ed.). Hoboken: Wiley.

Croux, C., Dehon, C., Rousseeuw, P. J., & Van Aelst, S. (2001). Robust estimation of the conditional median function at elliptical models. *Statistics & Probability Letters*, *51*, 361–368.

Croux, C., Dehon, C., & Yadine, A. (2010). The k-step spatial sign covariance matrix. *Advances in Data Analysis and Classification*, *4*, 137–150.

Croux, C., Filzmoser, P., & Fritz, H. (2013). Robust sparse principal component analysis. *Technometrics*, *55*, 202–214.

Croux, C., Filzmoser, P., & Oliveira, M. R. (2007). Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *87*, 218–225.

Croux, C., Filzmoser, P., Pison, P., & Rousseeuw, P. J. (2003). Fitting multiplicative models by robust alternating regressions. *Statistics and Computing*, *13*, 23–36.

Croux, C., Gelper, G., & Mahieu, K. (2010). Robust exponential smoothing of multivariate time series. *Computational Statistics & Data Analysis*, *54*, 2999–3006.

Croux, C., & 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.

Dahiya, R. C., Staneski, P. G., & Chaganty, N. R. (2001). Maximum likelihood estimation of parameters of the truncated cauchy distribution. *Communications in Statistics: Theory and Methods*, *30*, 1737–1750.

DasGupta, A. (2008). *Asymptotic theory of statistics and probability*. New York: Springer.

Datta, B. N. (1995). *Numerical linear algebra and applications*. Pacific Grove: Brooks/Cole Publishing Company.

Davidson, J. (1994). *Stochastic limit theory*. Oxford: Oxford University Press.

Davies, P. L. (1992). Asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *The Annals of Statistics*, *20*, 1828–1843.

Dehon, C., Filzmoser, P., & Croux, C. (2000). Robust methods for canonical correlation analysis. In H. A. L. Kiers, J. P. Rasson, P. J. F. Groenen, & M. Schrader (Eds.), *Data analysis, classification, and related methods* (pp. 321–326). Berlin: Springer.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, *62*, 531–545.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, *76*, 354–362.

Devroye, L., & Wagner, T. J. (1982). Nearest neighbor methods in discrimination. In P. R. Krishnaiah & L. N. Kanal (Eds.), *Handbook of statistics* (Vol. 2, pp. 193–197). Amsterdam: North Holland.

Donoho, D. L., & Huber, P. J. (1983). The notion of breakdown point. In P. J. Bickel, K. A. Doksum, & J. L. Hodges (Eds.), *A Festschrift for Erich L. Lehmann* (pp. 157–184). Pacific Grove: Wadsworth.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd ed.). New York: Wiley.

Easton, G. S., & 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., & 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.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: SIAM.

Efron, B. (2014). Estimation and accuracy after model selection, (with discussion). *Journal of the American Statistical Association*, *109*, 991–1007.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference*. New York: Cambridge University Press.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall/CRC.

Ehrenberg, A. S. C. (1982). Writing technical papers or reports. *The American Statistician*, *36*, 326–329.

Fan, R. (2017). A squared correlation coefficient of the correlation matrix, unpublished manuscript at http://lagrange.math.siu.edu/Olive/sfan.pdf.

Fang, K. T., & Anderson, T. W. (Eds.). (1990). *Statistical inference in elliptically contoured and related distributions*. New York: Allerton Press.

Fang, K. T., Kotz, S., & Ng, K. W. (1990). *Symmetric multivariate and related distributions*. New York: Chapman & Hall.

Farcomeni, A., & Greco, L. (2015). *Robust methods for data reduction.* Boca Rotan: Chapman & Hall/CRC.

Feng, L., & Sun, F. (2015). A note on high-dimensional two-sample test. *Statistics & Probability Letters*, *105*, 29–36.

Feng, X., & He, X. (2014). Statistical inference based on robust low-rank data matrix approximation. *The Annals of Statistics*, *42*, 190–210.

Ferguson, T. S. (1996). *A course in large sample theory.* New York: Chapman & Hall.

Filzmoser, P., Joossens, K., & Croux, C. (2006). Multiple group linear discriminant analysis: Robustness and error rate. In A. Rizzi & M. Vichi (Eds.), *Compstat 2006: Proceedings in computational statistics* (pp. 521–532). Heidelberg: Physica.

Flury, B., & Riedwyl, H. (1988). *Multivariate statistics: A practical approach.* London: Chapman & Hall.

Fogel, P., Hawkins, D. M., Beecher, C., Luta, G., & Young, S. (2013). A tale of two matrix factorizations. *The American Statistician*, *67*, 207–218.

Freeman, D. H., Gonzalez, M. E., Hoaglin, D. C., & Kilss, B. A. (1983). Presenting statistical papers. *The American Statistician*, *37*, 106–110.

Frey, J. (2013). Data-driven nonparametric prediction intervals. *Journal of Statistical Planning and Inference*, *143*, 1039–1048.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, *84*, 165–175.

Friedman, J. H., & Hall, P. (2007). On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, *137*, 669–683.

Fujikoshi, Y. (2002). Asymptotic expansions for the distributions of multivariate basic statistics and one-way MANOVA tests under nonnormality. *Journal of Statistical Planning and Inference*, *108*, 263–282.

Fujikoshi, Y., Sakurai, T., & Yanagihara, H. (2014). Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression. *Journal of Multivariate Analysis*, *123*, 184–200.

García-Escudero, L. A., & Gordaliza, A. (2005). Generalized radius processes for elliptically contoured distributions. *Journal of the American Statistical Association*, *100*, 1036–1045.

García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Clustering*, *4*, 89–109.

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). *Methods for statistical data analysis of multivariate observations* (1st ed.). New York: Wiley.

Gnanadesikan, R. (1997). *Methods for statistical data analysis of multivariate observations* (2nd ed.). New York: Wiley.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, *28*, 81–124.

Golub, G. H., & Van Loan, C. F. (1989). *Matrix computations* (2nd ed.). Baltimore: John Hopkins University Press.

Good, P. I. (2012). *A practitioner's guide to resampling for data analysis, data mining, and modeling.* Boca Raton: Chapman & Hall/CRC.

Gregory, K. B., Carroll, R. J., Baladandayuthapani, V., & Lahari, S. N. (2015). A two-sample test for equality of means in high dimension. *Journal of the American Statistical Association*, *110*, 837–849.

Grimm, L. G., & Yarnold, P. R. (Eds.). (1995). *Reading and understanding multivariate statistics.* Washington: American Psychological Association.

Grimm, L. G., & Yarnold, P. R. (Eds.). (2000). *Reading and understanding more multivariate statistics.* Washington: American Psychological Association.

Gupta, A. K., Varga, T., & Bodnar, T. (2013). *Elliptically contoured models in statistics and portfolio theory* (2nd ed.). New York: Springer.

Hair, J. F., Black, B., Babin, B. J., & Anderson, R. E. (2009). *Multivariate data analysis* (7th ed.). Upper Saddle River: Prentice Hall.

Hall, P. (1988). Theoretical comparisons of bootstrap confidence intervals, (with discussion). *The Annals of Statistics*, *16*, 927–985.

Hamada, M., & Sitter, R. (2004). Statistical research: Some advice for beginners. *The American Statistician*, *58*, 93–101.

Hampel, F. R. (1975). Beyond location parameters: Robust concepts and methods. *Bulletin of the International Statistical Institute*, *46*, 375–382.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics*. New York: Wiley.

Hand, D. J. (2006). Classifier technology and the illusion of progress, (with discussion). *Statistical Science*, *21*, 1–34.

Hand, D. J., & Taylor, C. C. (1987). *Multivariate analysis of variance and repeated measures: A practical approach for behavioral scientists*. London: Chapman & Hall.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton: CRC Press Taylor & Francis.

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 & Data Analysis*, *17*, 197–210.

Hawkins, D. M., Bradu, D., & Kass, G. V. (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics*, *26*, 197–208.

Hawkins, D. M., & McLachlan, G. J. (1997). High breakdown linear discriminant analysis. *Journal of the American Statistical Association*, *92*, 136–143.

Hawkins, D. M., & Olive, D. J. (1999a). Improved feasible solution algorithms for high breakdown estimation. *Computational Statistics & Data Analysis*, *30*, 1–11.

Hawkins, D. M., & Olive, D. (1999b). Applications and algorithms for least trimmed sum of absolute deviations regression. *Computational Statistics & Data Analysis*, *32*, 119–134.

Hawkins, D. M., & 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., & Simonoff, J. S. (1993). High breakdown regression and multivariate estimation. *Applied Statistics*, *42*, 423–432.

He, X., & Portnoy, S. (1992). Reweighted LS estimators converge at the same rate as the initial estimator. *The Annals of Statistics*, *20*, 2161–2167.

He, X., & Wang, G. (1996). Cross-checking using the minimum volume ellipsoid estimator. *Statistica Sinica*, *6*, 367–374.

He, X., & Wang, G. (1997). Qualitative robustness of S*- estimators of multivariate location and dispersion. *Statistica Neerlandica*, *51*, 257–268.

Hebbler, B. (1847). Statistics of prussia. *Journal of the Royal Statistical Society, A*, *10*, 154–186.

Henderson, H. V., & Searle, S. R. (1979). Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics. *The Canadian Journal of Statistics*, *7*, 65–81.

Hesterberg, T. (2014). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. arXiv:1411.5279v1.pdf. (An abbreviated version was published (2015), *The American Statistician*, 69, 371–386.).

Hoffman, I., Serneels, S., Filzmoser, P., & Croux, C. (2015). Sparse partial robust M regression. *Chemometrics and Intelligent Laboratory Systems Part A*, *149*, 50–59.

Horowitz, J. L. (2001). The Bootstrap. In J.J. Heckman, & E. Leamer (Eds.), *Handbook of econometrics,* (Vol. 5), Chap. 52. Amsterdam: Elsevier Science.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Hössjer, O. (1991). Rank-based estimates in the linear model with high breakdown point, Ph.D. thesis, Report 1991:5, Department of Mathematics, Uppsala University, Uppsala, Sweden.

Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics* (2nd ed.). Hoboken: Wiley.

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., & 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., & Van Aelst, S. (2008). High breakdown multivariate methods. *Statistical Science*, *23*, 92–119.

Hubert, M., Rousseeuw, P. J., & Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, *21*, 618–637.

Hubert, M., & Vanden Branden, K. (2003). Robust methods for partial least squares regression. *Journal of Chemometrics*, *17*, 537–549.

Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis* (2nd ed.). Hoboken: Wiley.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, *50*, 120–126.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer.

Janssen, A., & Pauls, T. (2003). How do bootstrap and permutation tests work? *The Annals of Statistics*, *31*, 768–806.

Jiang, J. (2010). *Large sample techniques for statistics*. New York: Springer.

Jin, J., & Wang, W. (2016). Influential features PCA for high dimensional clustering. *The Annals of Statistics*, *44*, 2323–2359.

Johnson, M. E. (1987). *Multivariate statistical simulation*. New York: Wiley.

Johnson, N. L., & Kotz, S. (1970a). *Distributions in statistics: Continuous univariate distributions* (Vol. 1). Boston: Houghton Mifflin Company.

Johnson, N. L., & Kotz, S. (1970b). *Distributions in statistics: Continuous univariate distributions* (Vol. 2). Boston: Houghton Mifflin Company.

Johnson, N. L., & Kotz, S. (1972). *Distributions in statistics: Continuous multivariate distributions*. New York: Wiley.

Johnson, R. A., & Wichern, D. W. (1988). *Applied multivariate statistical analysis* (2nd ed.). Englewood Cliffs: Prentice Hall.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Englewood Cliffs: Prentice Hall.

Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, 4. www.amstat.org/publications/jse/.

Johnstone, I. M., & Nadler, B. (2017). Roy's largest root test under rank-one alternatives. *Biometrika*, *104*, 181–193.

Jolliffe, I. T. (2010). *Principal component analysis* (2nd ed.). New York: Springer.

Kachigan, S. K. (1991). *Multivariate statistical analysis: A conceptual introduction* (2nd ed.). New York: Radius Press.

Kakizawa, Y. (2009). Third-order power comparisons for a class of tests for multivariate linear hypothesis under general distributions. *Journal of Multivariate Analysis*, *100*, 473–496.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.

Kay, R., & Little, S. (1987). Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika*, *74*, 495–501.

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.). New York: Macmillan Publishing.

Khattree, R., & Naik, D. N. (1999). *Applied multivariate statistics with SAS software* (2nd ed.). Cary: SAS Institute.

Kim, J. (2000). Rate of convergence of depth contours: With application to a multivariate metrically trimmed mean. *Statistics & Probability Letters*, *49*, 393–400.

Kim, J., & Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, *18*, 191–219.

Klouda, K. (2015). An exact polynomial time algorithm for computing the least trimmed squares estimate. *Computational Statistics & Data Analysis*, *84*, 27–40.

Koch, I. (2014). *Analysis of multivariate and high-dimensional data*. New York: Cambridge University Press.

Koenker, R. W., & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*, 33–50.

Koltchinskii, V. I., & Li, L. (1998). Testing for spherical symmetry of a multivariate distribution. *Journal of Multivariate Analysis*, *65*, 228–244.

Konietschke, F., Bathke, A. C., Harrar, S. W., & Pauly, M. (2015). Parametric and nonparametric bootstrap methods for general MANOVA. *Journal of Multivariate Analysis*, *140*, 291–301.

Kosfeld, R. (1996). Robust exploratory factor analysis. *Statistical Papers*, *37*, 105–122.

Kowalski, C. J. (1973). Non-normal bivariate distributions with normal marginals. *The American Statistician*, *27*, 103–106.

Krzanowski, W. J. (1988). *Principles of multivariate analysis: A user's perspective*. Oxford: Oxford University Press.

Kshirsagar, A. M. (1972). *Multivariate analysis*. New York: Marcel Dekker.

Lattin, J., Carroll, J. D., & Green, P. E. (2003). *Analyzing multivariate data*. Belmont: Brooks/Cole.

Lehmann, E. L. (1999). *Elements of large-sample theory*. New York: Springer.

Lei, J., Robins, J., & Wasserman, L. (2013). Distribution free prediction sets. *Journal of the American Statistical Association*, *108*, 278–287.

Lei, J., & Wasserman, L. (2014). Distribution free prediction bands. *Journal of the Royal Statistical Society, B*, *76*, 71–96.

Leon, S. J. (1986). *Linear algebra with applications* (2nd ed.). New York: Macmillan Publishing Company.

Li, K. C., & Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics*, *17*, 1009–1052.

Li, R., Fang, K., & 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., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Liu, L., Hawkins, D. M., Ghosh, S., & Young, S. S. (2003). Robust singular value decomposition analysis of microarray data. *Proceedings of the national academy of sciences*, *100*, 13167–13172.

Liu, R. Y., Parelius, J. M., & Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics, and inference. *The Annals of Statistics*, *27*, 783–858.

Liu, X., & Zuo, Y. (2014). Computing projection depth and its associated estimators. *Statistics and Computing*, *24*, 51–63.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., & 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.

Machado, J. A. F., & Parente, P. (2005). Bootstrap estimation of covariance matrices via the percentile method. *Econometrics Journal*, *8*, 70–78.

MacKinnon, J. G. (2009). Bootstrap hypothesis testing. In D. Belsey & E. Kontoghioghes (Eds.), *Handbook of Computational Econometrics*, Chap. 6. Hoboken: Wiley.

Maguluri, G., & Singh, K. (1997). On the fundamentals of data analysis. In G. S. Maddela & C. R. Rao (Eds.), *Robust inference* (pp. 537–549). Amsterdam: Elsevier Science.

Mai, Q., Zou, H., & Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, *99*, 29–42.

Mammen, E. (1992). Bootstrap, wild bootstrap, and asymptotic normality. *Probability Theory and Related Fields*, *93*, 439–455.

Manzotti, A., Pérez, F. J., & 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., & Bibby, J. M. (1979). *Multivariate analysis*. London: Academic Press.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. Hoboken: Wiley.

Maronna, R. A., & Morgenthaler, S. (1986). Robust regression through robust covariances. *Communications in Statistics: Theory and Methods*, *15*, 1347–1365.

Maronna, R. A., & 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., & Yohai, V. J. (2015). High-sample efficiency and robustness based on distance-constrained maximum likelihood. *Computational Statistics & Data Analysis*, *83*, 262–274.

Maronna, R. A., & Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, *44*, 307–317.

Mašíček, L. (2004). Optimality of the least weighted squares estimator. *Kybernetika*, *40*, 715–734.

MathSoft. (1999a). *S-plus 2000 user's guide*. Data Analysis Products Division, MathSoft, Seattle.

MathSoft. (1999b). *S-plus 2000 guide to statistics* (Vol. 2). Data Analysis Products Division, MathSoft, Seattle.

McDonald, G. C., & 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*. Hoboken: Wiley.

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.

Minor, R. (2012). Poverty, productivity, and public health: The effects of "right to work" laws on key standards of living. *Thought & Action: the NEA Higher Education Journal* (pp. 16–28). http://www.nea.org/home/52880.htm.

Møller, S. F., von Frese, J., & Bro, R. (2005). Robust methods for multivariate data analysis. *Journal of Chemometrics*, *19*, 549–563.

Moore, D. S. (2000). *The basic practice of statistics* (2nd ed.). New York: Freeman.

Morgenthaler, S. (1989). Comment on Yohai and Zamar. *Journal of the American Statistical Association*, *84*, 636.

Morrison, D. F. (1967). *Multivariate statistical methods*. New York: McGraw-Hill.

Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2014). On the least trimmed squares estimator. *Algorithmica*, *69*, 148–183.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory* (1st ed.). New York, NY: Wiley.

Muirhead, R. J. (2005). *Aspects of multivariate statistical theory* (2nd ed.). New York: Wiley.

Muirhead, R. J., & Waternaux, C. M. (1980). Asymptotic distribution in canonical correlation analysis and other multivariate procedures for nonnormal populations. *Biometrika*, *67*, 31–43.

Navarro, J. (2014). Can the Bounds in the multivariate Chebyshev inequality be attained? *Statistics & Probability Letters*, *91*, 1–5.

Navarro, J. (2016). A very simple proof of the multivariate Chebyshev's inequality. *Communications in Statistics: Theory and Methods*, *45*, 3458–3463.

Nishi, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, *12*, 758–765.

Nordhausen, K., & Tyler, D. E. (2015). A cautionary note on robust covariance plug-in methods. *Biometrika*, *102*, 573–588.

Norman, G. R., & Streiner, D. L. (1986). *PDQ statistics*. Philadelphia: B.C. Decker.

Obozinski, G., Wainwright, M. J., & Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, *39*, 1–47.

Oja, H. (2010). *Multivariate nonparametric methods with R: An approach based on spatial signs and ranks*. New York: Springer.

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 & Data Analysis*, *46*, 99–102.

Olive, D. J. (2004b). Visualizing 1D regression. In M. Hubert, G. Pison, A. Struyf, & S. Van Aelst (Eds.), *Theory and applications of recent robust methods* (pp. 221–233). Basel: Birkhäuser.

Olive, D. J. (2005a). Two simple resistant regression estimators. *Computational Statistics & Data Analysis*, *49*, 809–819.

Olive, D. J. (2005b). A simple confidence interval for the median, unpublished manuscript available from http://lagrange.math.siu.edu/Olive/ppmedci.pdf.

Olive, D. J. (2007). Prediction intervals for regression models. *Computational Statistics & Data Analysis*, *51*, 3115–3122.

Olive, D. J. (2008). Applied robust statistics, unpublished online text, see http://lagrange.math.siu.edu/Olive/ol-bookp.htm.

Olive, D. J. (2010). Multiple linear and 1D regression models, unpublished online text available from http://lagrange.math.siu.edu/Olive/regbk.htm.

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*, 2610–2628.

Olive, D. J. (2014). *Statistical theory and inference*. New York: Springer.

Olive, D. J. (2017a). *Linear regression*. New York: Springer.

Olive, D.J. (2017b). Applications of hyperellipsoidal prediction regions, *Statistical Papers*, to appear.

Olive, D.J. (2017c), Prediction and statistical learning, unpublished notes available from http://lagrange.math.siu.edu/Olive/slearnbk.htm.

Olive, D. J. (2017d). Bootstrapping hypothesis tests and confidence regions, preprint, see http://lagrange.math.siu.edu/Olive/ppvselboot.pdf.

Olive, D. J., & Hawkins, D. M. (2003). Robust regression with high coverage. *Statistics & Probability Letters*, *63*, 259–266.

Olive, D. J., & Hawkins, D. M. (2005). Variable selection for 1D regression models. *Technometrics*, *47*, 43–50.

Olive, D. J., & Hawkins, D. M. (2007a). Behavior of elemental sets in regression. *Statistics & Probability Letters*, *77*, 621–624.

Olive, D. J., & Hawkins, D. M. (2007b). Robustifying robust estimators, preprint, see http://lagrange.math.siu.edu/Olive/preprints.htm.

Olive, D. J., & Hawkins, D. M. (2008). High breakdown multivariate estimators, preprint, see http://lagrange.math.siu.edu/Olive/preprints.htm.

Olive, D. J., Hawkins, D. M. (2010). Robust multivariate location and dispersion, preprint, see http://lagrange.math.siu.edu/Olive/pphbmld.pdf.

Olive, D. J., & Hawkins, D. M. (2011). Practical high breakdown regression, preprint, see http://lagrange.math.siu.edu/Olive/pphbreg.pdf.

Olive, D. J., Pelawa Watagoda, L. C. R., & Rupasinghe Arachchige Don, H. S. (2015). Visualizing and testing the multivariate linear regression model. *International Journal of Statistics and Probability*, *4*, 126–137.

Oosterhoff, J. (1994). Trimmed mean or sample median? *Statistics & Probability Letters*, *20*, 401–409.

Park, Y., Kim, D., & Kim, S. (2012). Robust regression using data partitioning and M-estimation. *Communications in Statistics: Simulation and Computation*, *8*, 1282–1300.

Pelawa Watagoda, L. C. R. (2013). Plots and testing for multivariate linear regression, Master's research paper, Southern Illinois University. http://lagrange.math.siu.edu/Olive/slasanthi.pdf.

Pelawa Watagoda, L. C. R. (2017). Inference after variable selection, Ph.D. thesis, Southern Illinois University. http://lagrange.math.siu.edu/Olive/slasanthiphd.pdf.

Pelawa Watagoda, L. C. R., & Olive, D. J. (2017). Inference for multiple linear regression after model or variable selection, preprint at http://lagrange.math.siu.edu/Olive/ppvsinf.pdf.

Peña, D. (2005). A new statistic for influence in regression. *Technometrics*, *47*, 1–12.

Peña, D., & 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 W. Gaul & H. Locarek-Junge (Eds.), *Classification in the Information Age, Proceedings of the 22nd Annual GfKl Conference, Dresden 1998* (pp. 225–232). Berlin: Springer.

Pires, A. M., & Branco, J. A. (2010). Projection-pursuit approach to robust linear discriminant analysis. *Journal of Multivariate Analysis*, *101*, 2464–2485.

Pison, G., Rousseeuw, P. J., Filzmoser, P., & Croux, C. (2003). Robust factor analysis. *Journal of Multivariate Analysis*, *84*, 145–172.

Polansky, A. M. (2011). *Introduction to statistical limit theory*. Boca Rotan: CRC Press.

Poor, H. V. (1988). *An introduction to signal detection and estimation*. New York: Springer.

Pourahmadi, M. (2013). *High-dimensional covariance estimation*. Hoboken: Wiley.

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.). Mineola: Dover.

R Core Team. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org.

Rao, C. R. (1965). *Linear statistical inference and its applications* (1st ed.). New York: Wiley.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

Raveh, A. (1989). A nonparametric approach to linear discriminant analysis. *Journal of the American Statistical Association*, *84*, 176–183.

Rencher, A., & Pun, F. (1980). Inflation of R^2 in best subset regression. *Technometrics*, *22*, 49–53.

Rencher, A. C., & Christensen, W. F. (2012). *Methods of multivariate analysis* (3rd ed.). Hoboken, NJ: Wiley.

Reyen, S. S., Miller, J. J., & Wegman, E. J. (2009). Separating a mixture of two normals with proportional covariances. *Metrika*, *70*, 297–314.

Riani, M., Atkinson, A. C., & Cerioli, A. (2009). Finding an unknown number of outliers. *Journal of the Royal Statistical Society, B*, *71*, 447–466.

Ritter, G. (2014). *Robust cluster analysis and variable selection*. Boca Rotan: Chapman & Hall/CRC Press.

Ro, K., Zou, C., Wang, W., & Yin, G. (2015). Outlier detection for high-dimensional data. *Biometrika*, *102*, 589–599.

Rocke, D. M., & Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, *91*, 1047–1061.

Rocke, D. M., & 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*. New York, NY: Wiley.

Rohatgi, V. K. (1984). *Statistical inference*. New York: Wiley.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*, 871–880.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Rousseeuw, P. J., Van Aelst, S., Van Driessen, K., & Agulló, J. (2004). Robust multivariate regression. *Technometrics*, *46*, 293–305.

Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*, 212–223.

Rousseeuw, P. J., & 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., & van Zomeren, B. C. (1992). A comparison of some quick algorithms for robust regression. *Computational Statistics & Data Analysis*, *14*, 107–116.

Rubin, D. B. (2004). On advice for beginners in statistical research. *The American Statistician*, *58*, 196–197.

Rupasinghe Arachchige Don, H.S. (2013), "Robust Multivariate Linear Regression," Master's Research Paper, Southern Illinois University, at (http://lagrange.math.siu.edu/Olive/shasthika.pdf).

Rupasinghe Arachchige Don, H. S. (2017). Bootstrapping analogs of the one way MANOVA test, Ph.D. thesis, Southern Illinois University. http://lagrange.math.siu.edu/Olive/shasthikaphd.pdf.

Rupasinghe Arachchige Don, H. S., & Olive, D. J. (2017). Bootstrapping analogs of the one way MANOVA test, preprint at http://lagrange.math.siu.edu/Olive/ppmanova.pdf.

Rupasinghe Arachchige Don, H. S., & Pelawa Watagoda, L. C. R. (2017). Bootstrapping analogs of the two sample Hotelling's T^2 test. *Communications and Statistics: Theory and Methods,* to appear. http://lagrange.math.siu.edu/Olive/stwosample.pdf.

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*. Cary: SAS Institute.

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: Wiley.

Seber, G. A. F., & Lee, A. J. (2003). *Linear regression analysis* (2nd ed.). New York: Wiley.

Sen, P. K., & Singer, J. M. (1993). *Large sample methods in statistics: An introduction with applications.* New York: Chapman & Hall.

Sen, P. K., Singer, J. M., & Pedrosa De Lima, A. C. (2010). *From finite sample to asymptotic methods in statistics.* New York: Cambridge University Press.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics.* New York: Wiley.

Severini, T. A. (2005). *Elements of distribution theory.* New York: Cambridge University Press.

Shevlyakov, G. L., & Oja, H. (2016). *Robust correlation: Theory and applications.* Hoboken: Wiley.

Silverman, B. A. (1986). *Density estimation for statistics and data analysis.* New York: Chapman and Hall.

Smith, W. B. (1997). Publication is as Easy as C-C-C. *Communications in Statistics: Theory and Methods, 26*, vii–xii.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72–101.

Srivastava, D. K., & 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.

Srivastava, M. S., & Khatri, C. G. (1979). *An introduction to multivariate statistics.* New York: North Holland.

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing.* New York: Wiley.

Stefanski, L. A. (1991). A note on high-breakdown estimators. *Statistics & Probability Letters*, *11*, 353–358.

Stewart, G. M. (1969). On the continuity of the generalized inverse. *SIAM Journal on Applied Mathematics*, *17*, 33–45.

Su, Z., & Cook, R. D. (2012). Inner envelopes: Efficient estimation in multivariate linear regression. *Biometrika*, *99*, 687–702.

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston: Pearson Education.

Tableman, M. (1994a). The influence functions for the least trimmed squares and the least trimmed absolute deviations estimators. *Statistics & Probability Letters*, *19*, 329–337.

Tableman, M. (1994b). The asymptotics of the least trimmed absolute deviations (LTAD) estimator. *Statistics & Probability Letters*, *19*, 387–398.

Tallis, G. M. (1963). Elliptical and radial truncation in normal populations. *The Annals of Mathematical Statistics*, *34*, 940–944.

Tarr, G., Müller, S., & Weber, N. C. (2016). Robust estimation of precision matrices under cellwise contamination. *Computational Statistics & Data Analysis*, *93*, 404–420.

Taskinen, S., Koch, I., & Oja, H. (2012). Robustifying principal component analysis with spatial sign vectors. *Statistics & Probability Letters*, *82*, 765–774.

Thode, H. C. (2002). *Testing for normality.* New York: Marcel Dekker.

Tiao, G. C., & Tsay, R. S. (1983). Consistency properties of least squares estimates of autoregressive parameters in ARMA models. *The Annals of Statistics*, *11*, 856–871.

Todorov, V. (2007). Robust selection of variables in linear discriminant analysis. *Statistical Methods and Applications*, *15*, 395–407.

Todorov, V., & Pires, A. M. (2007). Comparative performance of several robust linear discriminant analysis methods. *REVSTAT, Statistical Journal*, *5*, 63–83.

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. (1977). *Exploratory data analysis.* Reading: Addison-Wesley Publishing Company.

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 Aelst, S., & Willems, G. (2011). Robust and efficient one-way MANOVA tests. *Journal of the American Statistical Association*, *106*, 706–718.

van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge: Cambridge University Press.

Velilla, S. (1993). A note on the multivariate Box-Cox transformation to normality. *Statistics & Probability Letters*, *17*, 259–263.

Venables, W. N., & Ripley, B. D. (2003). *Modern applied statistics with S* (4th ed.). New York: Springer.

Víšek, J. Á. (2006). The least trimmed squares - Part III: Asymptotic normality. *Kybernetika*, *42*, 203–224.

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical statistics with applications* (7th ed.). Belmont: Thomson Brooks/Cole.

Wakaki, H., Yanagihara, H., & Fujikoshi, Y. (2002). Asymptotic expansions of the null distributions of test statistics for multivariate linear hypotheses under nonnormality. *Hiroshima Mathematics Journal*, *32*, 17–50.

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. www.jstatsoft.org.

White, H. (1984). *Asymptotic theory for econometricians.* San Diego: Academic Press.

Wilcox, R. R. (2008). Robust principal components: A generalized variance perspective. *Behavior Research Methods*, *40*, 102–108.

Wilcox, R. R. (2009). Robust multivariate regression when there is heteroscedasticity. *Communications in Statistics: Simulation and Computation*, *38*, 1–13.

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). New York: Academic Press Elsevier.

Wilk, M. B., Gnanadesikan, R., Huyett, M. J., & 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., & Van Aelst, S. (2002). A robust Hotelling test. *Metrika*, *55*, 125–138.

Wisnowski, J. W., Simpson, J. R., & 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., & 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.

Witten, D. M., & Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society, B*, *73*, 753–772.

Wood, S. N. (2006). *Generalized additive models: An introduction with R.* Boca Rotan: Chapman & Hall/CRC.

Woodruff, D. L., & Rocke, D. M. (1993). Heuristic search algorithms for the minimum volume ellipsoid. *Journal of Computational and Graphical Statistics*, *2*, 69–95.

Woodruff, D. L., & 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.

Xu, H., Caramanis, C., & Mannor, S. (2011). Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *99*, 1–9.

Yao, J., Zheng, S., & Bai, Z. (2015). *Large sample covariance matrices and high-dimensional data analysis*. New York: Cambridge University Press.

Yao, Y. (1965). An approximate degrees of freedom solution to the multivariate Behrens Fisher problem. *Biometrika*, *52*, 139–147.

Zhang, J. (2011). Applications of a robust dispersion estimator, Ph.D. thesis, Southern Illinois University. http://lagrange.math.siu.edu/Olive/szhang.pdf.

Zhang, J., & Olive, D. J. (2009). Applications of a Robust Dispersion Estimator, online at http://lagrange.math.siu.edu/Olive/pprcovm.pdf.

Zhang, J., Olive, D. J., & Ye, P. (2012). Robust covariance matrix estimation with canonical correlation analysis. *International Journal of Statistics and Probability*, *1*, 119–136.

Zhang, J.-T., & Liu, X. (2013). A modified Bartlett test for heteroscedastic one-way MANOVA. *Metrika*, *76*, 135–152.

Zhang, J.-T., Zhou, B., Guo, J., & Liu, X. (2016). A modified Bartlett test for heteroscedastic two-way MANOVA. *Journal of Advanced, Statistics*, *1*, 94–108.

Zou, H., Hastie, T., & Tibshirani, R. (1993). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*, 265–286.

Index

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2