

# 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  $i$ th case is  $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})^T$ . 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 *mrobddata.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 *mrobddata.txt* into *R*: The commands

```
source("http://lagrange.math.siu.edu/Olive/mpack.txt")
source("http://lagrange.math.siu.edu/Olive/mrobddata.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- $\tau$ , 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  $(\bar{\mathbf{x}}, \mathbf{S})$  where  $\bar{\mathbf{x}}$  is the sample mean and  $\mathbf{S}$  is the sample covariance matrix. Computing the determinant of  $\mathbf{S}$  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  $(\bar{\mathbf{x}}, \mathbf{S})$  and  $(\mathbf{0}, \mathbf{I}_p)$  where  $\mathbf{x}$  is a  $p \times 1$  vector,  $\mathbf{0}$  is the zero vector, and  $\mathbf{I}_p$  is the  $p \times p$  identity matrix. Since the determinant  $\det(\mathbf{I}_p) = 1$ , the fit with the smallest determinant will not be the classical estimator if  $\det(\mathbf{S}) > 1$ . 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  $(T_i, \mathbf{C}_i)$  (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  $\|\hat{\beta}\|$ ) arbitrarily large or to drive the smallest or largest eigenvalues of the dispersion estimator  $\mathbf{C}$  to 0 or  $\infty$ . High breakdown estimators have  $\gamma_n = d_n/n \rightarrow 0.5$  and zero breakdown estimators have  $\gamma_n \rightarrow 0$  as  $n \rightarrow \infty$ .

Note that an estimator cannot be consistent for  $\theta$  unless the number of randomly selected cases goes to  $\infty$ , except in degenerate situations. The following theorem shows the widely used elemental estimators are zero breakdown estimators. (If  $K = K_n \rightarrow \infty$ , 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  $K_n \rightarrow \infty$ .)

**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  $\infty$  for consistency except in degenerate situations.

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

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  $K = K_n \rightarrow \infty$  or  $h = h_n \rightarrow \infty$  as  $n \rightarrow \infty$ . 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  $Cov(\mathbf{x}) = c_x \Sigma$  for elliptically contoured distributions. But for  $S$  estimators, the global minimizer is estimating  $d_x \Sigma$  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.



Robust Multivariate Analysis

J. Olive, D.

2017, XVI, 501 p. 76 illus., 6 illus. in color., Hardcover

ISBN: 978-3-319-68251-8