Robust Multivariate Analysis

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

February 1, 2013

Contents

Preface

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

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

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

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

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

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

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

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

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

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

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

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

What is in the Book?

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

veloped.

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

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

Some of the applications in this text include the following.

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

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

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

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

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

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

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

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

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

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

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

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

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

Background

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

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

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

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

Need for the book:

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

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

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

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

Some Theory and Conjectures for F-Brand Name Estimators

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

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

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

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

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

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

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

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

Acknowledgments

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