Multivariate Outlier Detection and Robust Clustering with Minimum Covariance Determinant Estimation and S-Estimation

By

Johanna Sarah Hardin

B.A. (Pomona College) 1995
M.S. (University of California, Davis) 1997

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

Davis

Approved:

Committee in Charge

2000

To

Peter, Kris, Dad, Lame

and

everyone else who believed in me.

Acknowledgments

First of all, I would like to thank my advisor Dr. David Rocke whose continued support has been generous and encouraging throughout many years. I am lucky to have found a mentor who is understanding and knows how to push me so that I achieve my best.

Also, I would like to thank Professor Don Bentley at Pomona College for showing me that Statistics is a wonderful use of Mathematics. He also introduced me to the world of teaching where I feel completely at home. I hope someday to be able to inspire other students as he inspired me.

Many thanks go to Dr. David Woodruff and Torsten Reiners for the use of their programs and their insightful comments into my work. My thesis was much improved by the conversations I had with them.

I would like to thank my committee members Drs. Robert Shumway, Wolfgang Polonik, and Jessica Utts for their advice and comments on my work.

A special thanks goes to Dr. Wesley Johnson for accepting me into the graduate program and always believing in me. Also, Larry Tai and Dr. Zhi-Wei Lu deserve many thanks for helping with the computational aspects of the dissertation. And I am indebted to Dr. Janko Gravner for always being available to work through any problems I might have.

On the personal side, I would like to thank Peter Littig whose love and support has helped me through all the hard times. And many thanks to my sister, Jennefer Asperheim, who has been with me every step of the way.

And last but certainly not least, I would like to thank my parents who always told me that I could do anything I put my mind to. Also, it was nice that they gave me good math genes.

This work was supported in part by grants from the National Science Foundation: NSF ACI 96-19020 and NSF DMS 98-70172.

Johanna Sarah Hardin
June 2000
Statistics

Abstract

Mahalanobis-type distances in which the shape matrix is derived from the Minimum Covariance Determinant estimator, a consistent, high-breakdown robust multivariate location and scale estimator, have an asymptotic chi-squared distribution as is the case with those derived from the ordinary covariance matrix. However, even in quite large samples, the chi-squared approximation is poor. We provide an improved F approximation that gives accurate outlier rejection points for various sample sizes.

The robust distances are also used for classifying data into different clusters and identifying outlying points in a mixture population. This work is innovative as it allows an unspecified number of points to be unclassified in the estimation process. Since the outlying points are not used in the estimation, the method is both a robust clustering method and an outlier identification method which uses an F approximation to determine cutoff points. We provide results on both type I and type II errors of the method and a comparable chi-squared method.

The translated biweight S-estimator is also used in the robust Mahalanobis-type distances to identify outliers using an F approximation in the one cluster case and to both cluster and identify outliers using an F approximation in the multiple cluster case. The distances which use the S-estimator metric also have an asymptotic chi-squared distribution, but the F approximation is superior especially in small data sets and datasetwise rejection.

Contents

1 Introduction **1**

1.1. Importance of Outlier Detection 1

1.1.1. What is an outlier? 2

1.1.2. What are the reasons to keep outliers in mind? 2

1.1.3. How do outliers arise? 4

1.1.4. What are the hypotheses for the model? 4

1.1.5. What consequences arise if the outliers are ignored? 5

1.2. Historical Background on Multivariate Outliers 6

1.2.1. Structured Data 8

1.3. Overview of the Thesis 8

2 The Distribution of Robust Distances **10**

2.1. Introduction 10

2.2. Robust Estimators for Outlier Detection 16

2.2.1. Affine Equivariant Estimators 17

2.2.2. Minimum Covariance Determinant 17

2.2.3. Estimating the MCD 18

2.3. Distance Distributions 21

2.3.1. Finding the Degrees of Freedom for the F Distribution . . . 29

2.3.2. Estimating m and c 31

2.4. Results 32

2.5. Conclusion 33

3 Outlier Detection in the Multiple Cluster Setting Using the Minimum Covariance Determinant 35

3.1. Introduction 35

3.2. A Look at Various Clustering Methods 39

3.2.1. Hierarchical Agglomeration 40

3.2.2. Optimization Methods in Clustering 41

3.2.3. Robust Optimization Clustering 42

3.3. Robust Estimators in a Cluster Setting 44

3.3.1. Minimum Covariance Determinant 45

3.3.2. Estimating the MCD 45

3.4. Distance Distributions 47

3.4.1. Distances, a Review 47

3.4.2. Robust Cutoff in a Cluster Setting 49

3.5. Results 50

3.5.1. Clean Data 51

3.5.2. Contaminated Data 52

3.6. Conclusion 54

4 Outlier Detection and Clustering with S-Estimation 56

4.1. Introduction 56

4.2. M- and S- Estimation 58

4.2.1. M- Estimation 58

4.2.2. S- Estimation 60

4.2.3. Comparison with the MCD 67

4.3. Computing the S-Estimate 67

4.3.1. The One Cluster Setting 68

4.3.2. The Multiple Cluster Setting 69

4.4. Identifying Outliers 70

4.4.1. F Cutoff 71

4.4.2. Degrees of Freedom 72

4.5. Results 73

4.5.1. One Cluster 74

4.5.2. Multiple Clusters 76

4.6. Conclusion 77

5 Conclusion **79**

6 Results of Simulations **81**

6.1. MCD – 5% Cutoff (1 Cluster) 82

6.2. MCD – 1% Cutoff (1 Cluster) 83

6.3. MCD – 0.1% Cutoff (1 Cluster) 84

6.4. MCD – Clean Data (2 & 3 Clusters) 85

6.5. MCD – Cluster Outliers (2 Clusters) 86

6.6. MCD – Radial Outliers (2 Clusters) 87

6.7. MCD – Diffuse Outliers (2 Clusters) 88

6.8. S-estimation – Clean Data (1 Cluster – Pointwise) 89

6.9. S-estimation – Clean Data (1 Cluster – Datasetwise) 90

6.10. S-Estimates – Cluster Outliers (1 cluster) 91

6.11. S-Estimates – Radial Outliers (1 cluster) 92

6.12. S-estimation – Clean Data (2 & 3 Clusters) 93

6.13. S-estimation – Cluster Outliers (2 Clusters) 94

6.14. S-estimation – Radial Outliers (2 Clusters) 95

6.15. S-estimation – Diffuse Outliers (2 Clusters) 96

7 Appendix: Theoretical Formulas for Wishart Degrees of Freedom 97

A. MCD Formula 98

B. S-Estimate Formula 99

8 Bibliography 101

Chapter 1

Introduction

1.1. Importance of Outlier Detection

Statisticians have been interested in finding “outlying”, “unusual”, or “unrepresentative” observations for many years. Data that have been incorrectly entered or that do not belong to the population from which the rest of the data came can bias estimates and give misleading results. Methods have been devised to identify and/or accommodate outlying observations in a variety of situations. With recent advances in technology, scientists are collecting larger data sets, and the analyst is getting further and further from the data (in the sense that she no longer writes or even sees every data point.) So, it is important to have good methodology for dealing with rogue observations that might not be noticed in a typical data analysis. Before discussing any specific statistical methodology that would be applied to data, we will ask (and answer) the following questions:

1. What is an outlier?

2. What are the reasons to keep outliers in mind?

3. How do outliers arise?

4. What are the hypotheses for the model?

5. What consequences arise if the outliers are ignored?

1.1.1. What is an outlier?

The basic definition of an outlying observation is a data point or points that do not fit the model of the rest of the data. Specific definitions are given by Barnett and Lewis (1994), Grubbs (1969) and Hawkins (1980):

> An outlier is a point such that “in observing a set of observations in some practical situation one (or more) of the observations 'jars' stands out in contrast to other observations, as extreme value.” (Barnett and Lewis, 1994) (pg 32)

> An outlying observation, or 'outlier', is one that appears to deviate markedly from other members of the sample in which it occurs. (Grubbs, 1969)

> An outlier is “an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism.” (Hawkins, 1980) (pg 1)

These definitions all refer to a point (or points) that is surprisingly different from the rest of the data. However, the words “stands out”, “appears to deviate”, and “arouse suspicions” imply some kind of subjectivity or preconceived ideas about what the data *should* look like. Though formal methods also often rely on distributional assumptions, formal methods will cut down on the amount of subjectivity used in data analyses that employ outlier detection methods.

1.1.2. What are the reasons to keep outliers in mind?

There are two basic reasons to search for outliers: 1. Because there is interest in the outliers for their own sake, and 2. Because the outliers could influence the results from the rest of the data.

In 1949 in England, the case of Hadlum vs. Hadlum gives a good example of interest in outliers for their own sake. Mr. Hadlum appealed the rejection of an earlier petition for divorce on grounds of adultery. Mrs. Hadlum had given birth to a child (who she claimed was fathered by her husband) on August 12, 1945, 349 days after Mr. Hadlum had left the country. The average gestation period for a human female is 280 days, and so the question arose: Is 349 days simply a large observation or does that data point belong to another population, namely one of women who conceived much later than August 28, 1944? (Barnett and Lewis, 1994) (pg 4)

Conversely, imagine a scientist studying a certain type of mosquito. If there were other types of mosquitos in his data collection, he would not be interested in their characteristics, he would simply want to remove the observations or ensure that the observations do not influence the statistical estimates of the original population. In such a situation, the techniques should accommodate the outliers but need not detect and reject them. Accommodation will be used in the sense that estimates will not be seriously affected or distorted by outlying observations and so will "accommodate" them in the estimation. Techniques that accommodate outliers are called robust.

> The word 'robust' is loaded with many - sometimes inconsistent - connotations. We use it in a relatively narrow sense: for our purposes, robustness signifies insensitivity to small deviations from the assumptions (Huber, 1981) (pg 1).

Robust methods guard against unique "bad" points, but they also give "protection against various types of uncertainty of knowledge of the data generating mechanism." (Barnett and Lewis, 1994) (pg 35) Two quantities that are related to robustness are resistance and breakdown. "A resistant method produces results that change only slightly when a small part of the data is replaced by new numbers, possibly very different from the original ones" (Hoaglin et al., 1983) (pg 2). The slight distinction in robustness and resistance is that "robustness generally implies insensitivity to departures from assumptions surrounding an underlying probability model" (Hoaglin et al., 1983) (pg 2). We will use the two terms interchangeably.

The breakdown of an estimator on n points is the fraction m/n, where m is the smallest number of points that can be replaced by arbitrary values to make the estimator unbounded. The breakdown will be discussed in detail in chapter 4.

1.1.3. How do outliers arise?

Along with identifying or accommodating outliers, we should have some idea of why or how the outliers arose. Barnett and Lewis (1994) (pgs 33-34) classify the types of variation into three groups.

- *Inherent Variability* – This is the natural variability in any data set.
- *Measurement Error* – This includes the limitation of the measuring device as well as any recording error done by the scientist.
- *Execution Error* – This includes situations with observations which are not in the population of interest or situations when a biased or misjudged sample is used.

When the analyst is deciding whether or not to reject an extreme observation, she should consider these types of variability. If the variability is due to measurement error or execution error, the point should probably be removed from the sample. However, if the variability is due to inherent variation, the point should remain. The interesting questions occur when the source of variation is unknown.

1.1.4. What are the hypotheses for the model?

In order to understand the source of variation, the analyst should define both the null and alternative hypotheses that will be applied to the model. The null hypothesis is that the data all come from a specified population, and the unspecified alternative hypothesis is that some of the observations do not conform to the structure of that population.

> If outliers arise because our initial model does not reflect the appropriate degree of inherent variation (we really need, say, a fatter-tailed distribution than the ubiquitous normal distribution initially adopted) then omission of extreme values to 'protect against outliers' is hardly a robust policy for estimating some measure of dispersion, say the variance. (Barnett and Lewis, 1994) (pg 36)

Obviously, in this case, rejecting extreme points would cause an underestimation of the variability. (Rocke, 1992)

The alternative hypothesis can come in different forms depending on the structure of the data and the research questions. Some alternative hypothesis possibilities are: (Barnett and Lewis, 1994) (pgs 45-52)

- *Deterministic* – some observations are erroneous due to measurement, recording, or transcribing.
- *Inherent* – some observations are large due to larger inherent variation in the model than was originally thought.
- *Mixture* – some observations come from a different population that has 'contaminated' the original population.

1.1.5. What consequences arise if the outliers are ignored?

With these questions in mind, it bears importance to ask, what if the outliers are ignored and the data analysis is done using conventional non-robust methods?

Masking and swamping refer to effects on the data procedure due to one or more outlying observations. Masking

> arises when a sample contains more than one outlier. These outliers so increase the spread of the sample that the removal of a single outlier makes little improvement in the appearance of the sample. ... The practical consequences of the masking effect is that any attempt to remove these outliers one at a time will prove fruitless, and so there is a need for more sophisticated methods that will detect all outliers (Hawkins, 1980). (pg 12)

Swamping “may affect a block discordancy test of a group of 2 or more outliers.” (Barnett and Lewis, 1994) (pg 109) In effect, an outlying observation can swamp one or more non-outlying observations to appear discordant.

As most analysts have come to realize, when using such sensitive procedures such as the sample mean and covariance, it is important to use outlier identification techniques prior to the data analysis. It may, however, be more efficient and logical to use robust techniques that can accommodate (and may identify) outlying observations.

1.2. Historical Background on Multivariate Outliers

“The study of outliers is as important for multivariate data as it is for univariate samples.” (Barnett and Lewis, 1994) (pg 269) It can be argued that the study of outliers is, in fact, *more important* for multivariate data because unusual observations are difficult to detect when they do not “stick out the end” of a dataset. The difficulty increases as the dimension increases because the outliers can be extreme in any of a growing number of directions.

Because there is no natural ordering to multivariate data, to find discordant points an ordering must be imposed. Distance measures can be used to give an ordering of the data. Let

$$d^2(X;\theta,\Omega) = (X-\theta)'\Omega^{-1}(X-\theta)$$

be a squared distance for some vector θ and positive definite symmetric matrix Ω. Robust estimates of θ and Ω give procedures that accommodate outliers, but little work has been done with robust estimates to identify outliers formally.

Given multivariate normal data, $N(\mu,\Sigma)$, it is well known that $d^2(X;\mu,\Sigma)$ has a χ_p^2 distribution where p is the dimension. Healy proposes that $d^2(X;\overline{X},S)$ can be approximated by a χ_p^2 distribution (Healy, 1968). Another possible distance, $d^2(X_i;\overline{X}^{(i)},S^{(i)})$ (where $\overline{X}^{(i)}$ and $S^{(i)}$ are the mean and covariance of the data without data point i), slightly reduces the risk of masking and has an asymptotic χ_p^2 distribution.

Another technique for identifying outliers is based on the ratio of scatter matrices with and without discordant points. Let $S=\sum_{j=1}^{n}(X_j-\overline{X})(X_j-\overline{X})'$ and $S^{(i)}=\sum_{j\neq i}(X_j-\overline{X})(X_j-\overline{X})'$.

$$\mathcal{R}_i=\frac{|S^{(i)}|}{|S|}$$

is the scatter ratio, and Wilks (Wilks, 1963) shows that the $\mathcal{R}_i$ are identically distributed Beta$((n-p-1)/2,p/2)$ variables if the data come from a multivariate normal sample.

Because $\overline{X}$ and S are not robust to outlying points, the above techniques are successful only when there is at most one outlying point. Wilks modifies his ratio test to test for 2, 3, or 4 outlying points.

$$\mathcal{R}_{i_1,i_2,\ldots,i_s}=\frac{|S^{(i_1,i_2,\ldots,i_s)}|}{|S|}$$

where $s=2,3,$ or 4 and $S^{(i_1,i_2,\ldots,i_s)}$ is the scatter when $X_{i_1},X_{i_2},\ldots,X_{i_s}$ are omitted from the sample. Wilks gives tables for his statistics, but the analyst must know in advance the number of outliers, and the procedures are still limited to 4 outlying points. Also, swamping must be kept in mind when applying any outlier identification procedure that rejects 2, 3, or 4 points simultaneously. As the number of data points and the number of outliers grow, this method runs into severe combinatorial problems.

1.2.1. Structured Data

There has been much work done on outliers in regression, linear models, and designed experiments. (For a survey see Barnett and Lewis (1994: pgs 315-394.)) In the regression context, two types of outliers can arise. First, points can be outlying in the Y space (the dependent variables). Second, points can be outlying in the X space (the independent variables.) Outliers of the first type can be handled using univariate outlier identification techniques because the linear model reduces the parameter space. Outliers of the second type, usually called leverage points because they can be strong influences on the regression problem, can be handled using the multivariate techniques used with unstructured data. Though our work does not directly deal with outliers in a regression context, our methods can easily be applied.

1.3. Overview of the Thesis

> In the analysis of multivariate data we frequently need to employ statistical methods which will be relatively unaffected by the possible presence of contaminants in the sample under investigation. That is, we need methods for the accommodation of outliers. For multivariate data such methods are less well developed than for the univariate case and furthermore they tend to be directed to general robustness considerations rather than to be designed specifically with outliers in mind. (Barnett and Lewis, 1994) (pg 273)

This work will use existing robust techniques but will apply them with outliers in mind. We will modify and enhance accommodating procedures to become outlier identification procedures. As we have seen in the case of Hadlum vs. Hadlum, it is often important to have outlier identification methods.

In Chapter 2 we will employ the use of a robust estimator, the Minimum Covariance Determinant (Rousseeuw, 1984) to find location and scale estimates for the data. In a single cluster situation, we can calculate the Mahalanobis Squared Distances from the robust MCD estimates, and points with large distances will be thought of as outlying. An F distribution is applied to the distances to find outliers based on percentages.

In Chapter 3 we apply the techniques developed in Chapter 2 to a multiple cluster setting. Here we find the MCD for each cluster based on an initial clustering of the data. One important idea from this chapter is that not every point is forced into a cluster, which makes the clustering procedure robust to outlying points. The F distribution techniques are again applied to the distances, but in this chapter we apply the cutoffs on a cluster by cluster basis.

In Chapter 4 we use the translated-biweight S-estimator (Rocke, 1996) in place of the MCD. The translated-biweight is a robust estimator, and we can use it to find robust location and scale estimates for data in a single cluster setting or a multiple cluster setting. Again, the F distribution is applied to identify outliers.

Finally, we conclude in Chapter 5 with some discussions on our future projects.

Chapter 2

The Distribution of Robust Distances

2.1. Introduction

Outlier detection is important to researchers in many fields. For non-statisticians, the ability to detect outlying data is often vital to the interpretation of the final results of a study. For statisticians, detecting outlying points is itself a necessary area of research. Outliers are points that differ from the rest of the data for some reason. Outlying points can be the result of miscalibrated equipment, wrongly entered data, or a normal but rare environmental effect.

In one or two dimensions, outliers are easily identified from simple plots, but detection of outliers is more challenging in higher dimensions. In multivariate applications, with three or more dimensions, outliers can be difficult or impossible to identify from plots of observed data. Various methods for detecting outliers have been studied (Atkinson, 1994; Barnett and Lewis, 1994; Gnanadesikan and Kettenring, 1972; Hadi, 1992; Hawkins, 1980; Maronna and Yohai, 1995; Penny, 1995; Rocke and Woodruff, 1996; Rousseeuw and VanZomeren, 1990).

One way to identify possible multivariate outliers is to calculate a distance from each point to a “center” of the data. An outlier would be a point with a distance larger than some predetermined value. A conventional measurement of quadratic distance from a point X to a location Y given a shape S, in the multivariate setting is:

$$d_S^2(X, Y) = (X - Y)'S^{-1}(X - Y)$$

This quadratic form is often called the Mahalanobis Squared Distance (MSD). In the normal setting, large values of $d_S^2(x_i, \overline{X})$, where $\overline{X}$ and S are the conventional sample mean and covariance matrix, indicate that the point x_i is an outlier (Barnett and Lewis, 1994). The distribution of the MSD with both the true location and shape parameters and the conventional location and shape parameters is well known (Gnanadesikan and Kettenring, 1972). However, the conventional location and shape parameters are not robust to outliers, and the distributional fit to the distance breaks down when robust measures of location and shape are used in the MSD (Rousseeuw and VanZomeren, 1991). Determining exact cutoff values for outlying distances continues to be a difficult problem.

In trying to detect single outliers in a multivariate normal sample, $d_S^2(x_i, \overline{X})$ will identify sufficiently outlying points. In data with clusters of outliers, however, the distance measure $d_S^2(x_i, \overline{X})$ breaks down (Rocke and Woodruff, 1996). Data sets with multiple outliers or clusters of outliers are subject to problems of masking and swamping (Pearson and Chandra Sekar, 1936). Masking occurs when a group of outlying points skews the mean and covariance estimates toward it, and the resulting distance of the outlying point from the mean is small. Swamping occurs when a group of outlying points skews the mean and covariance estimates toward it and away from other inlying points, and the resulting distance from the inlying points to the mean is large. As an example, consider a data set due to Hawkins, Bradu, and Kass (Hawkins et al., 1984). These data consist of 75 points in dimension three. We can only see one outlying point, but 14 of the points were constructed to be outliers. By using the mean and variance of all the data, we have masked the remaining 13 outliers. (See Figure 2.1)

Problems of masking and swamping can be resolved by using robust estimates of

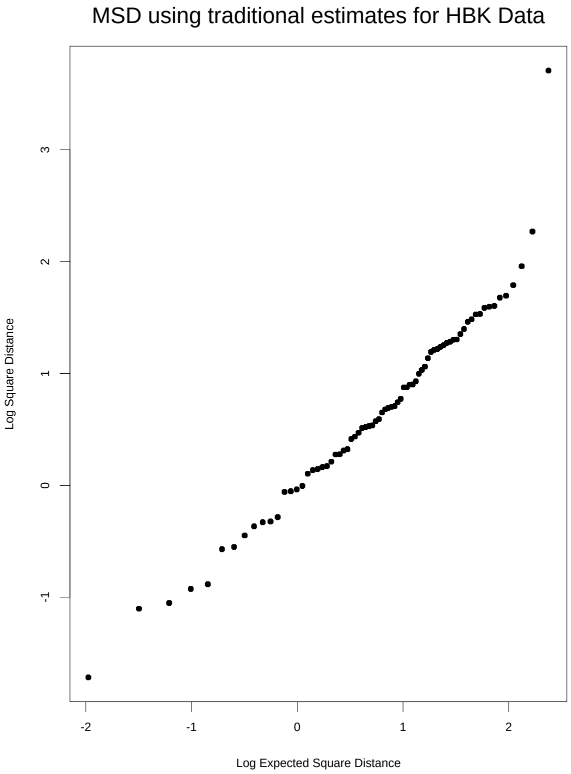

Figure 2.1: *Mahalanobis squared distances for the HBK data plotted against the* χ^2_3 *expected order statistics using the traditional mean and covariance matrix. The data are constructed with 14 outliers which are masked in the traditional computations.*

shape and location; robust location and shape estimates are not affected by outliers. Outlying points will not enter into the calculation of the robust statistics, so they will not be able to influence the parameters used in the MSD. The inlying points, which all come from the underlying distribution, will completely determine the estimate of the location and shape of the data. Some robust estimators of location and shape include the minimum covariance determinant (MCD) and the minimum volume ellipsoid (MVE) of Rousseeuw (Hampel et al., 1986; Rousseeuw, 1984; Rousseeuw and Leroy, 1987) and M-estimators and S-estimators. (Campell, 1980; Campell, 1982; Huber, 1981; Kent and Tyler, 1991; Lopuhaä, 1992; Maronna, 1976; Rocke, 1996; Tyler, 1983; Tyler, 1988; Tyler, 1991). By using a robust location and shape estimate in the MSD, outlying points will not skew the estimates and can be identified as outliers by large values of the MSD.

The MSD can take as its arguments any location and shape estimates. In this chapter we are interested in robust location and shape estimates, which are better suited for detecting outliers. In particular in this chapter, we are interested in the MCD location and shape estimates. Given n data points, the MCD of those data is the mean and covariance matrix based on that sample of size h ($h \leq n$) that minimizes the determinant of the covariance matrix (Rocke and Woodruff, 1996; Rousseeuw, 1984).

$$\begin{aligned} MCD &= (\overline{X}^*_J, S^*_J) \\ \text{where} \qquad \text{J} &= \{\text{set of } h \text{ points} : |S^*_J| \leq |S^*_K| \quad \forall \text{ sets K s.t. } |K| = h\} \\ \overline{X}^*_J &= \frac{1}{h}\sum_{i\in J} X_i \\ S^*_J &= \frac{1}{h}\sum_{i\in J}(X_i - \overline{X}^*_J)(X_i - \overline{X}^*_J)^t \end{aligned}$$

The value h can be thought of as the minimum number of points which must not be outlying. The MCD has its highest possible break down at $h = \lfloor \frac{(n+p+1)}{2} \rfloor$

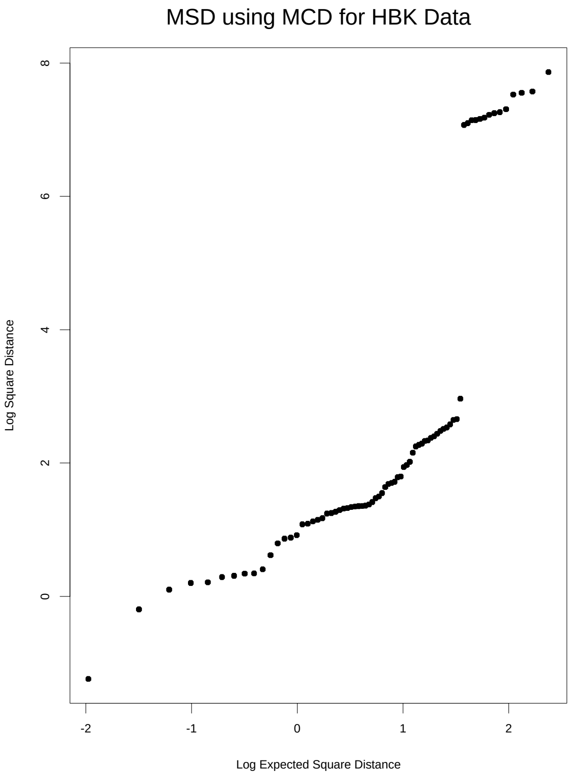

Figure 2.2: *Mahalanobis squared distances for the HBK data plotted against the χ^2_3 expected order statistics using the MCD mean and covariance matrix. All 14 outlying points are clearly visible as outlying.*

where $\lfloor \cdot \rfloor$ is the greatest integer function (Rousseeuw and Leroy, 1987; Lopuhaä and Rousseeuw, 1991). We will use $h = \lfloor \frac{(n+p+1)}{2} \rfloor$ in our calculations and refer to a sample of size h as a half sample. The MCD is computed from the "closest" half sample, and therefore, the outlying points will not skew the MCD location or shape estimate. Large values of MSDs, using the MCD location ($\overline{X}^*$) and shape estimate (S^*), will be robust estimates of distance and will correctly identify points as outlying. Recall the constructed data by Hawkins, Bradu, and Kass. Using the MCD estimates, the distances give a clear identification of the 14 outlying points. (See Figure 2.2)

Not every data set will give rise to an obvious separation between the outlying and non-outlying points. Consider the data given by Daudin, Dauby and Trecourt and analyzed by Atkinson (Daudin et al., 1988; Atkinson, 1994). The data are

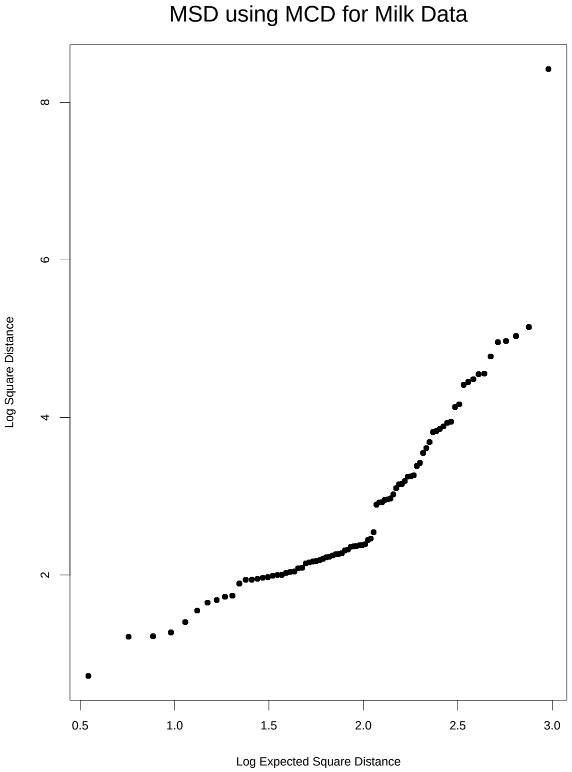

Figure 2.3: *Mahalanobis squared distances for the Milk data plotted against the* χ^2_8 *expected order statistics using the MCD mean and covariance matrix. One outlier is apparent, but how many outlying points are there? One? Five? Six?*

eight measurements on 85 bottles of milk. Using the robust MCD estimates, we are not subject to masking or swamping, but we are not sure which group of points should be considered as outlying. (See Figure 2.3)

In Figure 2.2, points were identified as obvious outliers, but in many situations (including Figure 2.3) it will be important to construct a minimum outlying distance. For some known constant c, $c \cdot d^2_{S^*}(X_i, \overline{X}^*)$ are asymptotically distributed as χ^2_p, but the asymptotic convergence is very slow, and the χ^2_p quantiles will be smaller than the corresponding MSD quantiles for even quite large samples. Use of χ^2_p quantiles as cutoff points will often lead to identifying too many points as outliers (Rousseeuw and VanZomeren, 1991).

Finding a good approximation to the distribution of $d^2_{S^*}(X_i, \overline{X}^*)$ will lead to cutoff values that identify minimum outlying values, even for clusters of outliers. If $\{X_i\}$ come from a multivariate normal sample, and mS comes from an independent Wishart, then $d_S^2(X_i, \mu)$ will have a distribution that is a multiple of an F statistic (Mardia et al., 1979). Since the MCD shape and location estimates are calculated using only the inlying points, $\overline{X}^*$ and S^* can be thought of as asymptotically independent from the extreme values in the sample. We can also approximate the distribution of S^* by matching the first two moments of a Wishart. Accordingly, the $d_{S^*}^2(X_i, \overline{X}^*)$ will be approximately distributed as a multiple of an F statistic. This insight allows us to find cutoff values for outlying points is in the estimation of the degrees of freedom associated with the F statistic. We will examine various cutoff values for MSD with MCD shape and location estimates for multivariate normal data given different values of n and p.

2.2. Robust Estimators for Outlier Detection

The estimation of multivariate location and shape is one of the most difficult problems in robust statistics (Campell, 1980; Campell, 1982; Davies, 1987; Devlin et al., 1981; Donoho, 1982; Hampel et al., 1986; Huber, 1981; Lopuhaä, 1989; Maronna, 1976; Rocke and Woodruff, 1993; Rousseeuw, 1985; Rousseeuw and Leroy, 1987; Stahel, 1981; Tyler, 1983; Tyler, 1991). For some statistical procedures, it is relatively straightforward to obtain estimates that are resistant to a reasonable fraction of outliers– for example, one dimensional location (Andrews et al., 1972) and regression with error-free predictors (Huber, 1981). The multivariate location and shape problem is more difficult, because many known methods will break down if the fraction of outliers is larger than 1/(p+1), where p is the dimension of the data (Donoho, 1982; Maronna, 1976; Stahel, 1981). This means that in high dimensions, a small amount of outliers can result in arbitrarily bad estimates.

2.2.1. Affine Equivariant Estimators

We are particularly interested in affine equivariant estimators of the data. A location estimator $\mathbf{y}_n \in \mathbb{R}^p$ is affine equivariant if and only if for any vector $\mathbf{v} \in \mathbb{R}^p$ and any nonsingular $p \times p$ matrix $\mathbf{M}$,

$$\mathbf{y}_n(\mathbf{M}X + \mathbf{v}) = \mathbf{M}\mathbf{y}_n(X) + \mathbf{v}.$$

A shape estimator $\mathbf{S}_n \in PDS(p)$ (the set of $p \times p$ positive definite symmetric matrices) is affine equivariant if and only if for any vector $\mathbf{v} \in \mathbb{R}^p$ and any nonsingular $p \times p$ matrix $\mathbf{M}$,

$$\mathbf{S}_n(\mathbf{M}X + \mathbf{v}) = \mathbf{M}\mathbf{S}_n(X)\mathbf{M}'.$$

Stretching or rotating the data will appropriately change an affine equivariant estimate of the data. The Mahalanobis Squared Distance is an affine equivariant estimator, which means the shape of the data determines the distances between the points. It is important to have affine equivariant estimators so that the measurement scale, location, and orientation can be ignored in the data analysis. Since MSD's are affine equivariant, the properties and procedures that use the MSD can be calculated for the purpose of simulations without loss of generality for standardized distributions. For the properties under normality, we can use N(0,I).

2.2.2. Minimum Covariance Determinant

The Minimum Covariance Determinant (MCD) location and shape estimates are resistant to outliers because the outliers will not be involved in the location and shape calculations. From the MCD sample, the sample mean and covariance matrix, which are robust estimates of the location and shape of the underlying population, can be computed.

Finding the exact MCD sample can be time consuming and difficult. The only known method for finding the exact MCD is to search every half sample and calculate the determinant of the covariance matrix of that sample. For n=20, the search would require computing about 184,756 determinants; for n=100, the search would require computing about 10^{29} determinants. With any conceivable computer, it is clear that finding the exact MCD is intractable by enumeration.

2.2.3. Estimating the MCD

Since the exact MCD is often impossible to find, the algorithm used to estimate of the MCD is, in some sense, the estimator. Various algorithms have been suggested for estimating the MCD.

Hawkins proposed a method based on swapping points in and out of a sample of size h. The basic algorithm is as follows.

- Start with a subsample of size h, H_1.
- Swap points $x_{i'} \in H_1$ and $x_{j'} \notin H_1$ and call the new subsample H_2 if:

  $$\Delta_i = \det(\text{cov}(H_1)) - \det(\text{cov}(H_2)) > 0$$

  AND the above difference, Δ_i, is maximized over swapping all possible pairs of points $x_i \in H_1$ and $x_j \notin H_1$
- Let H_2 be the new subsample of size h
- Repeat the process until no swap lowers the $\det(\text{cov}(H_i))$ (or equivalently, until no swap gives $\Delta_i > 0$) (Hawkins, 1994; Hawkins, 1999).

A faster algorithm was found independently by Hawkins (1999) and Rousseeuw and Van Driessen (1999). The core of the algorithm is based on what Rousseeuw calls the C-step. Instead of swapping one pair of points in and out, the C-step allows for many points to be interchanged at each step. Again, we start with a subset of the data of size h, H_1. We can compute $\overline{X}_{H_1}, S_{H_1}$, and $d^2_{S_{H_1}}(x_i, \overline{X}_{H_1}) =$

$d^2_{H_1}(i)$ for each point $i = 1, \ldots n$ based on the sample H_1. We can then sort the distances based on a permutation π so that:

$$d^2_{H_1}(\pi(1)) \leq d^2_{H_1}(\pi(2)) \leq \ldots \leq d^2_{H_1}(\pi(n))$$

We will then assign $\{\pi(1), \pi(2), \ldots, \pi(h)\}$ to H_2. Using H_2 we can calculate $\overline{X}_{H_2}, S_{H_2},$ and $d^2_{S_{H_2}}(x_i, \overline{X}_{H_2})$ and repeat the process until the permutation, π, does not change. Rouseeuw and VanDriessen (1999) showed that the process will converge.

The question remains for both algorithms, where does the initial H_1 come from? Previously, Hawkins used a random subset of size h from the data. If the data have large amounts of contamination, a random subset of size h will almost never look like the true underlying (uncontaminated) population, so it will be hard for either swapping algorithm to converge to the true uncontaminated shape of the data. For contaminated data, Rousseeuw proposed starting with a random subset of size p+1 (the minimum number of points needed to define a nonsingular covariance matrix) and adding points until a subset of h points is constructed (Rousseeuw and VanDriessen, 1999). Points are added to the initial random subset based on their distances to the mean of the initial subset. The algorithm is as follows.

- Let H_0 be a random subset of $p + 1$ points
- Find $\overline{X}_{H_0}$ and S_{H_0} (If $\det(S_{H_0}) = 0$ then add random points until $\det(S_{H_0}) > 0$)
- Compute the distances $d^2_{S_{H_0}}(x_i, \overline{X}_{H_0}) = d^2_{H_0}(i)$ and sort them for some permutation π such that,

  $$d^2_{H_0}(\pi(1)) \leq d^2_{H_0}(\pi(2)) \leq \ldots \leq d^2_{H_0}(\pi(n))$$
- $H_1 := \{\pi(1), \pi(2), \ldots, \pi(h)\}$

A random subset of p+1 points is much more likely to be representative of the uncontaminated data, and so H_1 here will be closer to the true data than a random subset of h points. In this paper we are interested in finding quantiles for distances based on MCD estimates. Our simulations are all done with pure multivariate normal data, and therefore, we are able to find quantiles using data that were not at all contaminated. Since our data were not contaminated, it was more effective to start with random subsets of size h, since diversity of starting points is of less value when there is no contamination. This is valid only for normal simulations and would not be used in practice.

The algorithm we used to estimate the MCD begins with a series of random starts, each of which is a randomly chosen half sample (or sample of size h) of the data points. We then use the algorithm referred to above as the C-step. For each random start, the procedure for calculating the MCD sample is as follows.

1. Compute the mean and covariance of the current half sample.
2. Calculate the MSD, based on the mean and covariance from step 1, for each point in the entire data set.
3. Choose a half sample of those points with the smallest MSDs from step 2.
4. Repeat steps 1-3 until the half sample no longer changes.

MCD sample will then be the half sample (in 3) with the minimum covariance determinant of all the random starts. A robust estimator like $d^2_{S^*}(x_i, \overline{X}^*)$, where S^* and $\overline{X}^*$ are the MCD shape and location estimates, is likely to detect outliers because outlying points will not affect the MCD shape and location estimates. For points x_i that are extreme, $d^2_{S^*}(x_i, \overline{X}^*)$ will be large, and for points x_i that are not extreme, $d^2_{S^*}(x_i, \overline{X}^*)$ will not be large. Here we are not subject to problems of masking and swamping.

2.3. Distance Distributions

Mahalanobis squared distances give a one-dimensional measure of how far a point is from a location with respect to a shape. Using MSD we can find points that are unusually far away from a location and call those points outlying. We have discussed the importance of using robust affine equivariant estimates for the location and shape of the data. Unfortunately, using robust estimates gives MSDs with unknown distributional properties. Consider n multivariate data points in dimension p, $X_i \sim N(\mu, \Sigma)$. Let S be an estimate of Σ such that, $mS \sim \text{Wishart}_p(\Sigma, m)$. Below are three distributional results for distances based on the above type of multivariate normal data.

1. The first distance distribution is based on the true parameters μ and Σ. We know that if the data are normal, these distances have an exact χ^2_p distribution (Mardia et al., 1979).

$$d^2_\Sigma(X_i, \mu) \quad \sim \quad \chi^2_p$$

   Which gives:

$$\begin{aligned} \text{E}[d^2_\Sigma(X_i, \mu)] &= p \\ \text{Var}[d^2_\Sigma(X_i, \mu)] &= 2p \end{aligned}$$

2. The second distance distribution is based on the usual mean and covariance estimates. These distances have an exact Beta distribution (Gnanadesikan and Kettenring, 1972; Wilks, 1962). It is interesting to note that the unbiased estimator has a smaller variance than the estimator which takes μ and Σ as parameters.

   Given,

$$\overline{X} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^t$$

then,

$$\frac{(n-1)^2}{n} d_S^2(X_i, \overline{X}) \sim \mathrm{Beta}\left(\frac{p}{2}, \frac{(n-p-1)}{2}\right)$$

Which gives:

$$\begin{aligned} \mathrm{E}\left[\frac{n d_S^2(X_i, \overline{X})}{(n-1)}\right] &= p \\ \mathrm{Var}\left[\frac{n d_S^2(X_i, \overline{X})}{(n-1)}\right] &= 2p\frac{(n-p-1)}{(n+1)} \end{aligned}$$

3. The third distance distribution is based on an estimate of S that is independent of the X_i. These distances have an exact F distribution when μ is the location parameter (Mardia et al., 1979), and an approximate F distribution when $\overline{X}$ is the location parameter (using a Slutsky type proof (Serfling, 1980)). It is interesting (in contrast to 1.) to note here that the unbiased estimator has a larger variance than the estimator which takes μ and Σ as its parameters.

   Given S and X_i independent,

$$\frac{np}{n-p} d_S^2(X_i, \mu) \sim F_{p,n-p}$$

   Using Slutsky's Theorem,

$$\frac{np}{n-p} d_S^2(X_i, \overline{X}) \dot{\sim} F_{p,n-p}$$

   Which gives:

$$\begin{aligned} \mathrm{E}\left[\frac{(n-p-2)}{n} d_S^2(X_i, \overline{X})\right] &\doteq p \\ \mathrm{Var}\left[\frac{(n-p-2)}{n} d_S^2(X_i, \overline{X})\right] &\doteq 2p\frac{(n-2)}{(n-p-4)} \end{aligned}$$

We will refer to the standard location and shape estimates ($\overline{X}$ and S) as within sample estimates and the MCD location and shape estimates ($\overline{X}^*$ and S^*) as out of sample estimates because extreme observations will not be used to calculate the MCD (with high probability). Our interest is in the extreme points which enter into the within sample calculations but not the out of sample calculations.

Since $\overline{X}$ and $\overline{X}^*$ are consistent estimators for μ, and S and $c^{-1}S^*$ (for some constant c) are consistent estimators for Σ, we know that the within sample and out of sample MSD are both asymptotically χ^2_p statistics (Mardia et al., 1979; Serfling, 1980). χ^2_p quantiles are often used (sometimes inappropriately) for identifying MSD extrema (Rousseeuw and VanZomeren, 1991)

Using χ^2_p quantiles to describe $d^2_S(X_i, \overline{X})$ (instead of Beta quantiles) will erroneously lead to identifying too few points as extreme (Rocke and Woodruff, 1996). The misidentification happens because the χ^2_p distribution is more variable than the true Beta distribution, and because the outliers skew the location and shape estimates.

The distribution of distances with out of sample estimates will be similar to the independent covariance structure distribution (the F distribution in 3). The MSD with MCD location and shape estimates will have a distribution which is approximately a multiple of an F distribution. Using χ^2_p quantiles to identify extrema of MSD with MCD estimates will erroneously lead to identifying too many points as extreme. The misidentification happens because the χ^2_p distribution is less variable than the F distribution. The overidentification can be seen in a picture of ordered MSD and ordered χ^2_p quantiles. The "elbow" happens where data points go from being included in the MCD sample to being outside the MCD sample (Rousseeuw and VanZomeren, 1991). (See Figures 2.4 & 2.5) We see that the points in the MCD sample do appear to have a χ^2_p distribution, but the out of sample points do not have a χ^2_p distribution.

Theorem 2.3.1 *Given n points, X_i, independently and identically distributed (iid)*

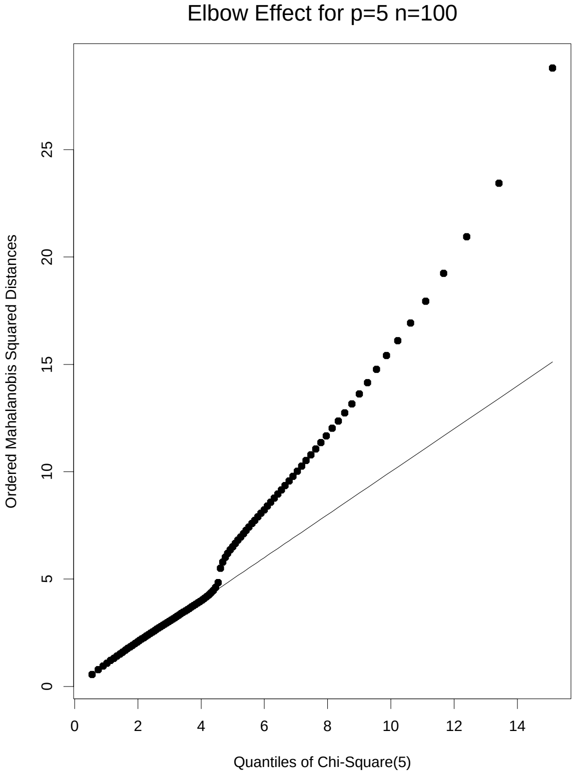

Figure 2.4: *Average Mahalanobis squared distances for simulated* $(n = 100, p = 5)$ *data plotted against the* χ^2_5 *expected order statistics using the MCD mean and covariance matrix. The points that are in the MCD sample appear to have a* χ^2_5 *distribution, but the out of sample points are definitely not distributed* χ^2_5*.*

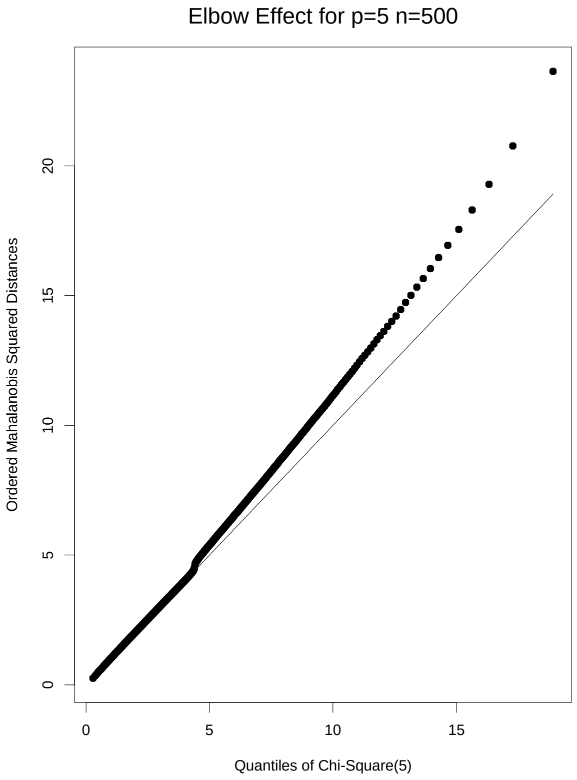

Figure 2.5: *Average Mahalanobis squared distances for simulated* $(n = 500, p = 5)$ *data plotted against the* χ^2_5 *expected order statistics using the MCD mean and covariance matrix. Again, the points that are in the MCD sample appear to have a* χ^2_5 *distribution, but the out of sample points, and especially the furthest outlying points, are not distributed* χ^2_5*. Even in large samples, there is still an elbow effect.*

$N_p(\mu, \Sigma)$, find the MCD sample based on $\epsilon = \frac{h}{n}$ of the sample, and choose δ such that $\epsilon < \delta < 1$. Conditional on X_i such that $(X_i - \mu)'\Sigma^{-1}(X_i - \mu) > \chi^2_{p,\delta}$, X_i will be asymptotically independent of the MCD sample.

Proof. The proof will be given in steps.

1. We can think of the iid sample as coming from 3 truncated Normal distributions.

   - Let n_1, n_2, n_3 come from a Multinomial $(n, \epsilon, \delta - \epsilon, 1 - \delta)$ distribution.
   - Let $x_1, x_2, \ldots, x_{n_1}$ come from a truncated normal distribution. The truncated normal distribution will be $N_p(\mu, \Sigma)$ with a truncation so that each of the points have $(x - \mu)'\Sigma^{-1}(x - \mu) \leq \chi^2_{p,\epsilon}$.
   - Let $x_{n_1+1}, x_{n_1+2}, \ldots, x_{n_1+n_2}$ come from a truncated normal distribution. The truncated normal distribution will be $N_p(\mu, \Sigma)$ with a truncation so that each of the points have $\chi^2_{p,\epsilon} < (x - \mu)'\Sigma^{-1}(x - \mu) \leq \chi^2_{p,\delta}$.
   - Let $x_{n_1+n_2+1}, x_{n_1+n_2+2}, \ldots, x_{n_1+n_2+n_3=n}$ come from a truncated normal distribution. The truncated normal distribution will be $N_p(\mu, \Sigma)$ with a truncation so that each of the points have $\chi^2_{p,\delta} < (x - \mu)'\Sigma^{-1}(x - \mu)$.

   The sample, $x_1, \ldots, x_{n_1}, x_{n_1+1}, \ldots, x_{n_1+n_2}, x_{n_1+n_2+1}, \ldots, x_{n_1+n_2+n_3=n}$, is an iid sample of size n from $N_p(\mu, \Sigma)$.

   We can define the ellipsoid regions R_1, R_2, such that

$$\begin{aligned}
\text{if} \quad (x - \mu)'\Sigma^{-1}(x - \mu) \leq \chi^2_{p,\epsilon} \quad &\Rightarrow \quad x \in R_1 \\
\text{if} \quad \chi^2_{p,\epsilon} < (x - \mu)'\Sigma^{-1}(x - \mu) \leq \chi^2_{p,\delta} \quad &\Rightarrow \quad x \in R_2 \\
\text{if} \quad (x - \mu)'\Sigma^{-1}(x - \mu) > \chi^2_{p,\delta} \quad &\Rightarrow \quad x \in R_3
\end{aligned}$$

2. Letting the MCD location and shape matrix be denoted by $\overline{X}^*$ and S^*, we know,

$$\overline{X}^* \quad \rightarrow \quad \mu$$

$$\frac{1}{c}S^* \quad \rightarrow \quad \Sigma \quad \text{for some } c \quad \text{(Tyler, 1983)}$$

which gives,

$$c(X - \overline{X^*})'{S^*}^{-1}(X - \overline{X}^*) \rightarrow (X-\mu)'\Sigma^{-1}(X-\mu) \qquad \forall X.$$

Let X_i be a $N(\mu, \Sigma)$ random variable. Conditional on $X_i \in R_1$, we have the following.

$$c(X_i - \overline{X}^*)'{S^*}^{-1}(X_i - \overline{X}^*) \leq \chi^2_{p,\epsilon} + O_p(n^{-1/2})$$

Let X_i be a $N(\mu, \Sigma)$ random variable. Conditional on $X_i \in R_3$, we have the following.

$$c(X_i - \overline{X}^*)'{S^*}^{-1}(X_i - \overline{X}^*) \geq \chi^2_{p,\epsilon} - O_p(n^{-1/2})$$

3. Let R_1^* be the ellipsoid containing the MCD points, the radius of R_1^* will be $\chi^2_{p,\epsilon} + O_p(n^{-1/2})$. We want to say that the points in R_3 will almost never, for large n, have MCD distances that will be in R_1^*.

   Let X_i be a $N(\mu, \Sigma)$ random variable, conditional on $X_i \in R_3$. So,

$$(X_i - \mu)'\Sigma^{-1}(X_i - \mu) \geq \chi^2_{p,\delta}.$$

   Then, $(X_i - \mu)'\Sigma^{-1}(X_i - \mu) = \chi^2_{p,\delta} + W$, where W is a positive random variable.

   Also, $c(X_i - \overline{X}^*)'{S^*}^{-1}(X_i - \overline{X}^*) = \chi^2_{p,\delta} + W + O_p(n^{-1/2})$.

$$\begin{aligned}
P(X_i \in \text{ MCD sample}) &= P(c(X_i - \overline{X}^*)'{S^*}^{-1}(X_i - \overline{X}^*) \leq \chi^2_{p,\epsilon} + O_p(n^{-1/2})) \\
&= P(\chi^2_{p,\delta} + W + O_p(n^{-1/2}) \leq \chi^2_{p,\epsilon} + O_p(n^{-1/2})) \\
&= P(\chi^2_{p,\delta} - \chi^2_{p,\epsilon} + W \leq O_p(n^{-1/2})) \\
&\leq P(\chi^2_{p,\delta} - \chi^2_{p,\epsilon} \leq O_p(n^{-1/2})) \\
&\rightarrow 0.
\end{aligned}$$

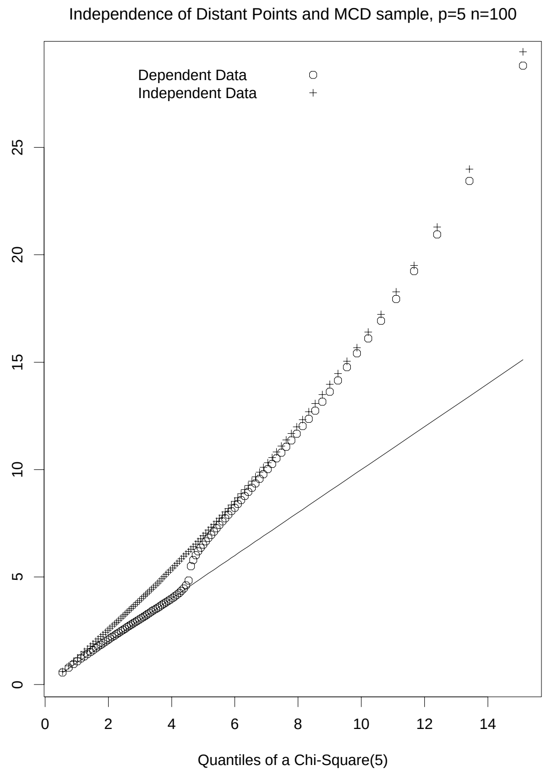

Figure 2.6: *This figure illustrates the lack of dependence of extreme points on the MCD estimates. The distances for the dependent data set, the "o"'s, are calculated using the MCD estimates from the "o" data. Independent data is then simulated, the "+"'s, and the distances are calculated using the MCD estimates from the "o" data. For both sets of data, the points are averages of the ordered distances for 1000 repetitions of dimension 5 size 100 data. It is apparent that the extreme distances are not affected by whether the MCD was calculated using the same sample or a different one.*

For $X_i \in R_3$, if the MCD never involved the point X_i, X_i would be exactly independent of the MCD sample. Any failure of independence involves a point $X_i \in R_3$ being in the MCD sample. We showed that the event of any $X_i \in R_3$ being in the MCD sample becomes exceedingly improbable. □

The independence of the extreme points and the MCD sample can also be seen in Figure 2.6. The picture shows average distances of two sets of independently simulated data sets whose distances were computed using the same MCD estimates. The first set contains the MCD sample, the second set was generated completely independently of the first sample and the MCD estimates. The out of sample estimates are approximately independent of the extreme points, and so the extrema behave like F quantiles. If X_i is multivariate normal data, and $\overline{X}^*$ and S_Y^* are the MCD mean and covariance, then

1. $X_1, \ldots, X_n \sim N_p(\mu, \Sigma)$,
2. the distribution of S_X^* can be approximated by,
$$mc^{-1} S_X^* \,\dot{\sim}\, \text{Wishart}_p(m, \Sigma), \tag{3.1}$$
where m is an unknown degrees of freedom, and c is a constant satisfying $\text{E}[S_X^*] = c\Sigma$ (which holds for some c because S_X^* is an affine equivariant shape estimator of Σ (Tyler, 1983)), and
3. the tail elements of X_i are approximately independent of S_X^*. (See Theorem 3.1 and Figure 2.6)

Then, using $\overline{X}^* \to \mu$,

$$\frac{c(m-p+1)}{pm} d^2_{S_X^*}(X_i, \overline{X}^*) \,\dot{\sim}\, F_{p,m-p+1}. \tag{3.2}$$

Using the above F distribution to calculate cutoff values for distances based on the MCD sample is a robust way of identifying outliers. The problem, then, is to estimate c and m correctly.

2.3.1. Finding the Degrees of Freedom for the F Distribution

Using a method of moments identification by the coefficient of variation (CV), Welch and Satterthwaite (Welch, 1937; Welch, 1947; Satterthwaite, 1946) estimated the degrees of freedom (df) for the well-known hypothesis test $H_o : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$ (when $\sigma_1 \neq \sigma_2$) which has a test statistic,

$$\frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \dot{\sim} t_{df}$$

$$df = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}}$$

Where $\quad V_1 = \dfrac{S_1^2}{n_1}$ and $V_2 = \dfrac{S_2^2}{n_2}$

Using a similar method of moments idea, we can estimate the degrees of freedom associated with the F distribution of $\frac{c(m-p+1)}{pm} d^2_{S_Y^*}(Y_i, \overline{Y}^*)$. Assuming that, for some m, S_Y^* has a multiple of a Wishart distribution (3.1) implies

$$mc^{-1} s_{ii}^* \dot{\sim} \chi^2_m \sigma_{ii}, \tag{3.3}$$

where s_{ii}^* are the diagonal elements of S_Y^*. Since the estimators are affine equivariant, we can assume the data are N(0,I), and therefore $\sigma_{ii} = 1$ and the diagonal elements are identically distributed (Grübel and Rocke, 1990). S_Y^* is an affine equivariant estimator, so the distribution of it will have the same Wishart degrees of freedom for any Σ. Therefore, the estimates of c and m in the N(μ, Σ) case will be the same as the estimates of c and m in the N(0,I) case.

From (3.3),

$$\text{E}[mc^{-1} s_{ii}^*] = m \Rightarrow \text{E}[s_{ii}^*] = c$$

and

$$\text{Var}[mc^{-1} s_{ii}^*] = 2m \Rightarrow \text{Var}[s_{ii}^*] = \frac{2c^2}{m}$$

which gives:

$$\text{CV} = \frac{\sqrt{\text{Var}[s_{ii}^*]}}{\text{E}[s_{ii}^*]} = \frac{c\sqrt{2/m}}{c} = \sqrt{\frac{2}{m}}.$$

So we can estimate m and c by

$$\hat{m} = \frac{2}{\hat{\mathrm{CV}}^2} \qquad\qquad \hat{c} = \frac{1}{h}\sum_{i=1}^{h} s^*_{ii}$$

where CV ($\hat{\mathrm{CV}}$) is the (estimated) coefficient of variation of the diagonal elements of the MCD shape estimator. If we can estimate the distribution of the MCD shape matrix, we could estimate c and m theoretically, and if we do not know the distribution of the MCD shape matrix, we can estimate c and m through simulations of the MCD shape matrix.

2.3.2. Estimating m and c

Simulation

Since multivariate normality is assumed, and the estimates are affine equivariant, hundreds of Monte Carlo N(0,I) data sets for pairs of n and p are simulated, and the MCD shape estimate is computed for each data set. Using the simulated MCD shape estimates, c is estimated from the average of the diagonal elements of S^*_Y, and m is estimated from the $\hat{CV}$ of the diagonal elements of S^*_Y.

Asymptotics

Simulation is often time consuming and tedious, so it would be preferable to have formulas for calculating m and c. An asymptotic expression for c exists that is good even for small samples.

$$c = \frac{P(\chi^2_{p+2} < \chi^2_{(p,h/n)})}{\frac{h}{n}}$$

where χ^2_ν is a Chi-Square random variable with ν degrees of freedom, and $\chi^2_{\nu,\epsilon}$ is the ϵ cutoff point for a χ^2_ν random variable (Croux and Haesbroeck, 2000).

For m there exists an asymptotic expression that is good in large samples and only moderately accurate in small samples (Croux and Haesbroeck, 2000). For small samples, simulation may be necessary to estimate m accurately. Croux and Haesbroeck used influence functions to determine an asymptotic expression for the variance elements of the MCD sample. Details are given in Appendix A.

2.4. Results

A common and reasonable method for identifying clusters of outliers is to find robust distances and then compute distributional quantiles to determine cutoffs. Any point with a distance larger than the cutoff point will be an outlier. Three distributional cutoff choices have been described,

1. χ^2_p (which is known to reject too few points when usual methods are used),
2. F (from (2)) with degrees of freedom calculated from the asymptotic formulas, and
3. F (from (2)) with degrees of freedom calculated from simulations.

We examined the performance of these methods in the null case by a Monte Carlo study with p=5,10,20 and n=50,100,500,1000. First, simulations of the MCD shape estimators with 1000 trials were undertaken to obtain values for m and c for each pair of n and p. Then the cutoff values for 5%, 1%, and 0.1% rejection for each of the three distribution choices were calculated.

Next, 1000 sets of independent data for each pair of dimension and size were simulated, and the number of points the cutoffs identified as outlying was counted. For the 5% nominal test, the percent identified for a subset of the data is shown. (See Table 1) It is seen that the Chi-Square cutoff points are too liberal, the asymptotic cutoff points are too conservative, and the simulated cutoff points are correct to simulation accuracy. Though the asymptotic cutoff is not perfect, it is superior to the Chi-Square because it is conservative and closer to the nominal values. In small samples, the simulated finite sample moments more accurately describe the data than either the asymptotic moments or the χ_p^2 distribution.

Results for 1% and 0.1% nominal tests are in Tables 2 and 3. Again, the Chi-Square cutoffs are too liberal, the asymptotic cutoffs are too conservative, and the simulated cutoffs are correct to simulation accuracy.

From the tables, we can see that the asymptotic accuracy depends primarily on n and not on p. As expected, the asymptotic cutoff becomes more accurate as n increases. These results lead to the following recommendations:

1. For large values of n (at least 500 observations), asymptotic formulas should be used for cutoff values of outlying MCD distances.
2. For smaller values of n (less than 500 observations), the asymptotic formula for c can be used, but simulation will be necessary to find m more accurately. For very small values of n, the simulation cutoffs are still superior to the currently used Chi-Square cutoff values; the simulation cutoffs are somewhat conservative for small samples.

2.5. Conclusion

A new method for determining outlying points in a multivariate normal sample has been derived. The methods presented here are superior to the commonly used Chi-Square cutoff. Asymptotic values for the cutoffs work well in samples of 1000 or larger, while a somewhat more computationally intensive simulation method can be used for smaller samples.

Because this work concerns clusters of outliers, there are implications for clustering as well as outlier identification. It is possible that robust distances may be able to identify outlying points in populations that are made up of two or more different clusters.

Also, the only robust method discussed in depth here is the MCD. The above methods can probably be extended to other robust methods like the Rousseuw's Minimum Volume Ellipsoid (Hampel et al., 1986; Rousseeuw, 1984; Rousseeuw and Leroy, 1987), S-estimation, and M-estimation.

Chapter 3

Outlier Detection in the Multiple Cluster Setting Using the Minimum Covariance Determinant

3.1. Introduction

As discussed previously, it is important to be able to distinguish outliers in a variety of situations. Various methods for finding outliers in a one dimensional setting have been developed. These methods, such as box plots, histograms, and Q-Q plots, tend to be graphical and difficult to modify to work in higher dimensions. We have given a method for finding outliers in a multiple dimensional setting that uses algebraic instead of graphical techniques. The method involves calculating a robust distance and then comparing it to the quantiles of a specified F distribution. All of these methods, however, assume that the data come from one population and not multiple populations.

The multiple cluster, multiple dimension case is another important place where outliers should be able to be identified. A few stray points can easily result in incorrect estimates, or worse, they could mask the number and shape of the clusters present. The previous multiple dimensional one cluster outlier detection method can be generalized to a multiple cluster case.

Outliers can have a particularly detrimental effect on non-robust clustering algorithms. An outlier or a cluster of outliers can influence the shape of a particular cluster, which in turn might mask separation between clusters. It is important to use robust methods in clustering and also to be able to identify the data points that are not in line with the bulk of the data. Robust cluster methods in combination with identification procedures will lead to more reliable data analyses.

Along with the importance of using robust clustering methods, it is also important to continue to use affine equivariant methods in any clustering algorithm which distinguishes between clusters using a distance metric. A non-affine equivariant distance metric, such as Euclidean, might not have the capability to discriminate between clusters if their shape matrices diverge strongly from the underlying spherical assumption of that distance. As an example, consider the data in figure 3.1. These data show two well separated clusters which both have a cigar shape. Even if we can somehow correctly identify the two cluster centers, using Euclidean distances will not give a good separation of these two groups, and the clustering method will probably not be able to identify that there are two clusters within the data structure. Figure 3.2 shows that the ellipses that cover 70% of the data do overlap. Figure 3.3 shows that the 99% ellipses do not overlap when the correct metrics are used. In this case, using Euclidean distances could incorrectly classify some of the points from cluster 1 into cluster 2 and vice versa. As seen previously, affine equivariance leads to a data dependent metric which is better at finding cluster shapes and therefore separating clusters.

In the one cluster case we used a metric that was based on the Minimum Covariance Determinant (MCD), a robust measure of location and shape. In this section, a similar idea is used, but instead of a minimum covariance determinant for the whole dataset, the MCD of each cluster is calculated. The MCD for each cluster determines the metric for that cluster and the metric can then be used to

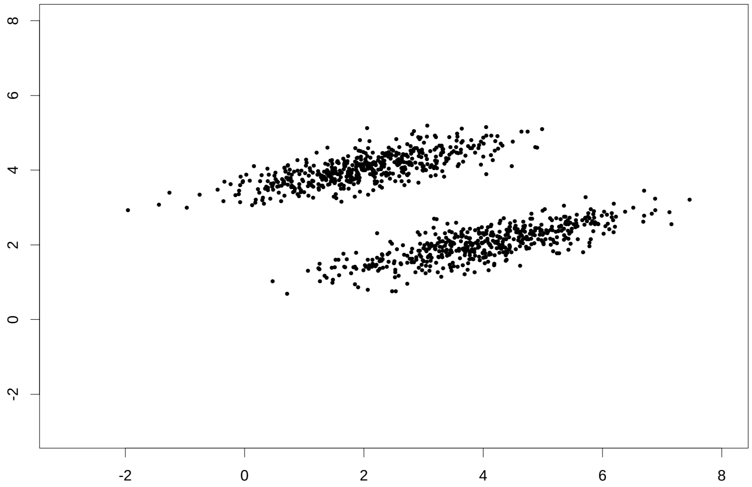

Figure 3.1: *Two independent bivariate normal data sets centered at (2,4) and (4,2) with equal covariances.*

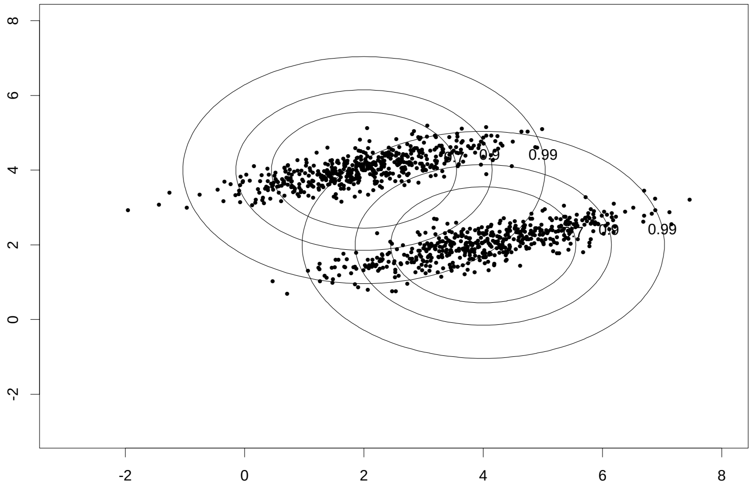

Figure 3.2: *Using Euclidean distances (with the correct centers identified) the probability (χ^2_2) ellipses overlap at the 70% level.*

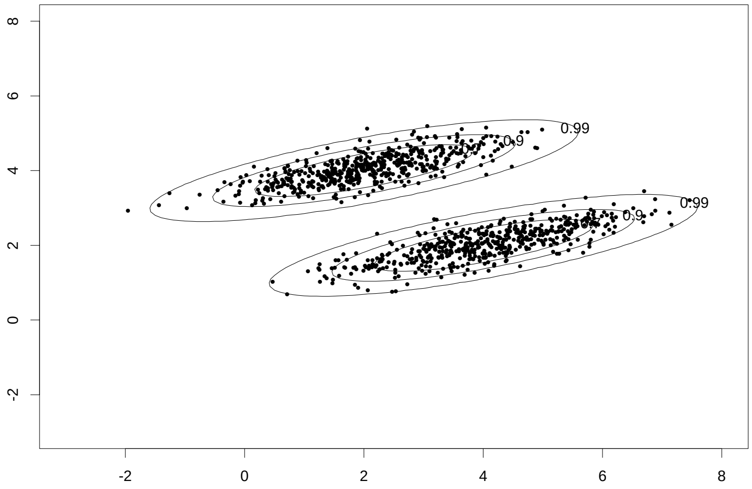

Figure 3.3: *Using the data specific metrics (with the correct centers identified) the probability (χ^2_2) ellipses do not overlap even at the 99% level.*

calculate a robust distance from each point in the dataset to the cluster center. The size of each cluster is determined by the number of points which are closer to that cluster's center than any other cluster center. The robust distance is then compared against the F distribution (modified slightly from chapter 2) with the size of the cluster as the "n" parameter.

The solution for the one cluster MCD starts with a random set of points and iterates to a solution using an algorithm that removes and replaces points according to their distances. As discussed in chapter 2, the random points give initial shape and location estimates. Using those estimates, the distances for all the points are computed, and with the closest "half sample" of those points, a new shape and location are computed. The process is repeated until convergence. A solution is obtained for many random starts, and the solution with the smallest covariance determinant is retained. This algorithm works well in the one sample case because even with 20-40% outliers, the probability of selecting $p+1$ random "good" points is still high. However, when $g \geq 2$, where g is the number of groups, (especially when the dimension is high, $p \geq 3$) it does not makes sense to start with g random subsets of points because with high probability all the subsets will have a similar location and shape to the entire dataset. Once the estimates resemble the entire data metric, it is difficult to converge back to the individual cluster metrics using the distances to remove and replace points. In this work, we start from a robust estimate of the clusters given by a program due to Reiners and Woodruff (2000). Their program gives a robust initialization of the points to groups (which also, consequently, gives both an initial metric and an initial size of each of the groups.) The method of finding a starting point is important; however, it is not what we are focusing on in this chapter.

Note that throughout this chapter, the number of groups is assumed to be fixed and known. This assumption is malleable in that additional small clusters will be ignored in the robust analysis, so they will not have an impact on estimating the g principal clusters. If an analyst is unsure of the number of populations present in the data, it would be wise to try the analysis on a variety of values for the number of clusters.

We will apply cutoff values to multi-cluster, multivariate normal data given different values of g, n, p, and different arrangements and percentages of outlying points.

3.2. A Look at Various Clustering Methods

Many clustering algorithms exist for clustering various types of data. These algorithms use data, multivariate or univariate, as input, and as output the algorithm gives each datum a classification into a particular group. Some algorithms require that the number of populations be pre-specified, and some algorithms allow for an unknown number of populations. Those algorithms that do not require as input a number of groups usually give results for a variety of values for the number of groups. The user then picks the result that most accurately fits the rest of the problem. Those algorithms that do require as input a number of groups can be run with different values for the number of groups. The user can then choose the result that makes the most sense according to the problem or according to some statistical criterion. Finding an appropriate criterion may prove to be a hard problem. These methods can be used to find a best fit to a problem with a given number of groups. Finding the correct number of groups for a particular data set is beyond the scope of this work.

3.2.1. Hierarchical Agglomeration

Hierarchical agglomeration can be done by different approaches, but the basic idea is to start with small clusters (or individual points) and agglomerate to larger clusters. Sometimes the agglomeration is done one pair at a time, and sometimes multiple clusterings happen simultaneously. The agglomerations are done according to some minimum distance. At least initially, the distance metric cannot be cluster-specific data-dependent because there is no way to measure variability of a single point. Only once a cluster has more than $p+1$ points can the metric for that cluster be data-dependent. Since the clusters are allowed to grow at different rates, it can take many agglomerations before all the clusters have at least $p+1$ points. When data-dependent distances are unavailable, either entire-data-set distances, Euclidean distances, Manhattan distances, or absolute distances can be used. Note that the latter three distances are not affine equivariant.

Hierarchical clustering imposes a hierarchical structure on data that may not be hierarchical. Clustering a dataset with a non-hierarchical structure may lead to clustering of “individuals linked by a series of intermediates” that do not seem to belong together (Everitt, 1993). Consider a two dimensional dataset where each point is of roughly the same absolute distance from its neighbors. The hierarchical method might group the points in a “chain” type formation when in fact the extreme points are not close to one another in any metric. A clustering done on the Iris data (which is not hierarchical) shows that the hierarchical procedure mclust (in S-Plus) clusters with a large number of errors. Furthermore, using this procedure, each permutation of the data gives a different clustering (Coleman et al., 1999), which would not happen if the data had a hierarchical structure.

"A hierarchical method suffers from the defect that it can never repair what it has done in previous steps" (Kauffman and Rousseeuw, 1990). Once a set of points is linked, each subsequent step can link more points to that cluster, but the algorithm does not allow for that set of points to become unlinked. For example, two points on the border of two different clusters may get linked accidentally, when on further inspection they would not seem closer to each other than to the cluster centers of their respective clusters. (This can happen when Euclidean distances are initially used, and data-dependent distances are subsequently used, see figure 3.3.)

It is important to note that for data with hierarchical structure and standardized variables, hierarchical agglomeration can provide much insight into the clustering structure of the dataset. However, it is also important to have other methods that are reliable in situations with data that has non-hierarchical or unknown structure. We use partitioning methods that require no internal data structure.

3.2.2. Optimization Methods in Clustering

Optimization methods use model based assumptions to derive different criteria which, when optimized, define a clustering of the dataset. Frequently used methods require that all points be assigned to a particular group. (See the clustering methods available in the software: mclust, kmeans, pam, clara, and fanny in S-Plus version 4.5 and proc cluster, proc fastclus, and proc varclus in SAS version 6.) Some methods, such as k-means, also use Euclidean distances or some other non-affine equivariant distance.

The model based assumptions can lead to inaccurate results if the data do not follow the assumptions. Let E_k be the set of all points in cluster k, then W_k is the sample cross-product matrix for the k^{th} cluster,

$$W_k = \sum_{i \in E_k} (x_i - \overline{x}_k)(x_i - \overline{x}_k)^T$$

and $W = \sum_{k=1}^{g} W_k$. One clustering method optimizes $\mathrm{tr}(W)$ (Ward, 1963), but this method treats the clusters as spherical and the same size. Another method optimizes $|W|$ (Friedman and Rubin, 1967). The $|W|$ method allows for ellipsoidal distributions, but requires that the clusters have the same orientation, size, and shape. A third method optimizes $\sum_{k=1}^{g} n_k \log \mathrm{tr}\ (W_k/n_k)$ (Banfield and Raftery, 1993) which allows for different size spherical shaped clusters. A final method optimizes $\sum_{k=1}^{g} n_k \log |\frac{W_k}{n_k}|$ (Scott and Symons, 1971) which allows for different ellipsoidal configurations for each cluster. Note that each of these optimizations criterion requires that at least $p+1$ points be in each cluster in order to calculate a non-singular W_k. This assumption may not always hold, but if fewer than $p+1$ points make up a particular cluster, maybe that "cluster" is simply a small group of outlying points. Since these above methods require that all points be allocated to a cluster, any small groups of outlying points might skew the estimates.

3.2.3. Robust Optimization Clustering

The clustering method we used, which will be described, assumes only that the clusters are elliptical. (The outlier identification methods, however, treat the data as having some moments follow certain distributions.) Since the cluster shape is estimated from assigned points, it is required that $p+1$ points be assigned to each of the main clusters. However, this method allows for unassigned points, so there could easily be allowed a cluster of points which is smaller than $p+1$ included in the group of outlying points. The algorithm and the program are due to Reiners and Woodruff (Reiners and Woodruff, 2000), and the program is called CLUSTER.

The optimization criterion used in this clustering algorithm is the same as that due to Scott and Symons (1971). However, Reiners and Woodruff use a blackboard architecture that improves the efficiency of the algorithm and allows for implementation on a parallel processor. The blackboard consists of three levels:

Level 1. This level contains the best solution found so far. (In our work a "best solution" consists of g sets of location, shape, and cluster assignment.)

Level 2. This level contains the best solutions for each subsample.

Level 3. This level contains all the solutions for the current subsample, and it is erased when the current subsample is optimized

The clustering algorithm is as follows:

1. Select an evaluation sample. This sample will be used to evaluate the solutions based on the optimization criterion. (With small datasets, the evaluation sample is the entire data set.)

2. Select a subsample from the entire dataset.

3. Randomly select one seed point that has not been used in a seed. Find a location and shape pair from the blackboard, and use them to assign the closest $2p$ points to the seed point to create a seed. Repeat this process $g-1$ times (except that the seed points are selected so as to be mutually distant from the seeds formed so far.) All $2gp$ points used in the seeds are then marked as having been used in a seed, and they will not be used as a seed point in the next iteration.

4. Calculate the means and covariances of the seeds, and again find the $2p$ closest points to improve the original seeds.

5. Add points from the rest of the subsample to the seeds based on the closest distances (with the seed metrics) to a cluster. Again, find cluster means and covariances, and improve the clusterings by adding or removing points based on the new cluster means and covariances.

6. Store the information on the third level of the blackboard. Return to step 3 unless a stopping criteria is reached. There are two stopping criteria: all the points in the subsample have been used in the seeds, or the current clustering is exactly the same as the previous clustering.

7. Project all the clusterings onto the evaluation sample, and store the best one on the second level of the blackboard. Erase the third level of the blackboard. Return to step 2, and repeat as many times as desired.

It is worth pointing out that with small data sets, it may not be necessary to subsample. Also, in the first iteration, there are no metrics available on the blackboard. In this case, the entire data metric is used. Though it may not be representative of the clusterings, the entire data metric does give an affine equivariant estimate of the parameters. It has been discussed that this clustering method allows for a specified number of points to be removed in the optimization step. For example, say T points are allowed to be "unclustered." Every time an estimate is evaluated on the evaluation sample, the closest $n - T$ points are found, and those points are used in the evaluation criteria. This way, outliers are not forced into clusters which could have the effect of skewing the estimates. This method accommodates any number of outlying points up to T. For our simulations we used $T = 100$ for large clusters (of size 200 - 700) and $T = 20$ for small clusters (of size 50-100.)

3.3. Robust Estimators in a Cluster Setting

Estimating cluster location and shape is a difficult problem in robust statistics. Most known methods make various distributional or shape assumptions that fail in the presence of outliers. This means that these clustering methods break down in the presence of a small number of outliers, particularly if the outliers are in their own cluster. The cluster shapes may be ellipsoidal with similar shape and size, but a cluster of outliers can skew the shape of one of the clusters, making it difficult to detect.

As in chapter 2, there is still interest in affine equivariant estimators. It is important to ensure that the estimates are not effected by measurement scale, location, or orientation.

3.3.1. Minimum Covariance Determinant

As discussed in chapter 2, we will use the Minimum Covariance Determinant (MCD) location and shape estimates as robust estimates of the location and shape of the clusters. Points that are outliers with respect to a particular cluster will not be involved in the location and shape calculations of that cluster, and points that are outliers with respect to all clusters will not be involved in the calculations of any clusters. The difference between the single population case and the multiple cluster case is that, in the latter, MCD samples need to be computed for each cluster. This important difference leads to a need for a good robust starting point in the clustering situation. Note that we also needed a good starting point for the one cluster MCD, but we bypassed that problem by using many small random samples. As previously mentioned, when the number of clusters grows, it becomes increasingly more difficult to find random samples that reflect the true layout of the cluster data.

3.3.2. Estimating the MCD

Again, the exact MCD is impossible to find except in small samples or trivial cases. So, the algorithm used to estimate the MCD will be the estimator. The algorithm used in the multiple cluster case will be similar to the single population algorithm with the one exception that the starting point of the algorithm will no longer be a random sub-sample of the data. The reason that it is important to have a non-random starting point for robust clustering is that random starts often give rise to shapes that are more representative of the entire data metric than the individual cluster metrics. Even with random samples of only $g \times (p+1)$ points (where g is the number of clusters and p is the dimension), it is highly unlikely that a random starting point would partition the points into their g clusters respectively. From a starting point which reflects the entire data metric, it is difficult to separate the points into the correct g clusters.

For a robust start, we used the program due to Reiners and Woodruff (2000) discussed previously. One characteristic of their program is that a parameter is set for the number of points that are allowed to be left out of the initial clustering. Because of the optimization criteria they use, the specified number of points will always be left out. The parameter should be set to be larger than the estimated number of outliers. However, if the parameter is set too high, the program will have trouble estimating the cluster shapes from the remaining data. This number should be a function of n.

The outlier detection methods described in this paper are not dependent on this particular robust clustering algorithm. Any robust initialization would give similar results. Even random starts could be used if a condition was added to prevent the clusters from converging to the large dataset shape.

For each dataset, the procedure for calculating the MCDs for each cluster is as follows.

1. Decide from how many populations the data came.

2. Use the program CLUSTER to find an initial robust clustering of the data.

3. From the initial clustering, calculate the mean and covariance of each of the clusters. (Each point belongs to at most one cluster, use the points belonging to a particular cluster to calculate its mean and covariance in the usual way.)

4. Calculate the MSD to each cluster, based on the most recently calculated mean and covariance, for each point in the dataset.

5. Assign each point to the cluster for which it has the smallest MSD. Also, assign a cluster size (n_j) to each cluster based on the number of points that are closest to that cluster.

6. For each cluster, choose a "half sample" ($= \lfloor (n_j + p + 1)/2 \rfloor$) of those points with the smallest MSDs from step 4.

7. For each cluster, compute the mean and covariance of the current half sample.

8. Repeat steps 4-7 until the half sample no longer changes.

9. Report estimates

For each cluster, the MCD sample will then be the final half sample (step 6). For each cluster (j), a robust distance like $d^2_{S^*_j}(x_i, \overline{X}^*_j)$, where S^*_j and $\overline{X}^*_j$ are the MCD shape and location estimates for cluster j, is likely to detect outliers because outlying points will not affect the MCD shape and location estimates. For points x_i that are extreme, $d^2_{S^*_j}(x_i, \overline{X}^*_j)$ will be large for all j, and for points x_i that are not extreme, $d^2_{S^*_j}(x_i, \overline{X}^*_j)$ will not be large for a particular j. Here we are not subject to problems of masking and swamping.

3.4. Distance Distributions

3.4.1. Distances, a Review

Mahalanobis squared distances give a one-dimensional measure of how far a point is from a location with respect to a shape. Using MSD we can find points that are unusually far away from a location and call those points outlying. We have discussed the importance of using robust affine equivariant estimates for the location and shape of the data. Unfortunately, using robust estimates gives MSDs with unknown distributional properties.

In chapter 2, an approximate distributional result for MSD distances based on location and shape derived from an MCD sample is given. Although the robust distances are asymptotically χ^2_p, an F distribution fits the extreme points much more accurately across all sample sizes, but especially in small samples. The distances based on the MCD metric can be expressed as,

$$\frac{c(m-p+1)}{pm} d^2_{S^*_X}(X_i, \overline{X}^*) \dot{\sim} F_{p,m-p+1}. \tag{4.1}$$

where $\overline{X}^*$andS^*_X are the location and shape estimates of the MCD sample, p is the dimension of the sample, and m and c are both parameters based on the shape of the MCD sample. The unknown parameters, m and c, can be estimated in two ways, using simulations or using an asymptotic result. The simulation results are the most accurate but also the most time consuming. In this chapter, we will use the asymptotic F distribution approach which gives results that are more accurate than the χ^2_p distribution and use much less computing time than the F distribution with simulated estimates. The parameter c can be estimated as in chapter 2,

$$c = \frac{P(\chi^2_{p+2} < \chi^2_{(p,h/n)})}{\frac{h}{n}}$$

where χ^2_ν is a Chi-Square random variable with ν degrees of freedom, and $\chi^2_{\nu,\epsilon}$ is the ϵ cutoff point for a χ^2_ν random variable (Croux and Haesbroeck, 2000). Again, we assume that the extreme points are asymptotically independent of the MCD estimates because they do not enter into the calculations. A similar independence argument to that of the one cluster model works for each cluster in the multiple cluster model.

For m there exists an asymptotic expression that is good in large samples and only moderately accurate in small samples (Croux and Haesbroeck, 2000). For small samples, simulation may be necessary to estimate m accurately. Croux and Haesbroeck used influence functions to determine an asymptotic expression for the variance elements of the MCD sample. Details are given in Appendix A. In this chapter we use only the theoretical parameter estimate in the interest of computing time.

3.4.2. Robust Cutoff in a Cluster Setting

Using the same arguments from the single population setting in the cluster setting, an F distribution can be used to approximate distances which are large with respect to a cluster location and shape. However, in this setting there are new factors to consider such as how many points are in each cluster and whether extreme points are simply members of another cluster.

In the single cluster case, the sample size of the dataset is used in the F cutoff calculation. Therefore, in the multiple cluster case, a sample size must be known or estimated for each cluster. The sample size also determines the "h" factor used in the MCD calculation. Recall, $\frac{n-h}{n}$ is the breakdown of the MCD estimator. The sizes from the final MCD iteration (step 5 in section 3.3.2) will be used as the sizes of each of the clusters. The last MCD iteration also provides a robust location and shape for each cluster, these estimates are used to compute distances from each cluster. For a particular point, the distance from each cluster center will be found, and a point will be counted in the cluster for which its distance is the smallest.

$$\#\text{ in cluster } j \;= n_j = \sum_{i=1}^{n} I(d^2_{S^*_j}(X_i, \overline{X}^*_j) \;\le\; d^2_{S^*_k}(X_i, \overline{X}^*_k) \quad \forall k = 1, \ldots, g \text{ groups})$$

Let $n_1, n_2, \ldots, n_g$ be the sizes of the respective clusters, $n = \sum_{j=1}^{g} n_j$. The breakdown for the MCD clustering method is $\frac{n^*-h^*}{n^*}$ where $n^* = \min\{n_1, n_2, \ldots, n_g\}$ and $h^* = \lfloor \frac{n^*+p+1}{2} \rfloor$. With these constructions in mind, the distances of interest are those associated with the cluster to which a point is closest. Let $\tilde{d}_i$ be the distance from point i to the closest cluster. An outlying point, i, will be one with $\tilde{d}_i$ greater than some cutoff value.

Consider g groups of n_j multivariate data points in dimension p, and let $X_{ij} \sim N_p(\mu_j, \Sigma_j)$ where i =observation and j =cluster. Let S_j be an estimate of Σ_j such that, $m_j S_j \sim \text{Wishart}_p(\Sigma_j, m_j)$. For the multiple cluster case,

$$c_j = \frac{P(\chi^2_{p+2} < \chi^2_{p,h_j/n_j})}{h_j/n_j}$$

and

$$\frac{c_j(m_j - p + 1)}{pm_j} d^2_{S_j^*}(X_i, \overline{X_j}^*) \dot{\sim} F_{p,m_j-p+1}.$$

Distributional cutoff results for distances based on the above type of clustered data with four different types of outlier arrangements: none, cluster, radial, and diffuse are given.

3.5. Results

As in the one cluster case, outliers can be identified as points with robust distances that exceed some cutoff value. The cutoffs are computed from distributional quantiles of χ^2 and F distributions. Because finding estimates for the parameters m and c by simulation is so computationally intensive, results are provided only for the F cutoff with theoretical estimates for m and c and for the χ^2_p cutoff. To compare the accuracy of these two estimates, we ran experiments on clean multivariate normal data and also on contaminated multivariate normal data. The contamination was done in three ways:

1. As a distinct cluster of outliers.
2. As radial outliers.
3. As diffuse outliers generated using a covariance structure equal to the entire dataset.

The Monte Carlo experiments were done at $p = 4, 7, 10$ with $g = 2, 3$ for the clean data and $g = 2$ for the contaminated data. Let n_g^p be the size of the groups simulated for a particular dimension and number of groups.

$$\begin{aligned} n_2^4 &= 300, 300 \text{ and } 200, 400 \\ n_3^4 &= 300, 300, 300 \text{ and } 200, 300, 400 \end{aligned}$$

$$\begin{aligned} n_2^7 &= 500, 500 \text{ and } 300, 700 \\ n_3^7 &= 500, 500, 500 \text{ and } 300, 500, 700 \\ n_2^{10} &= 500, 500 \text{ and } 300, 700 \\ n_3^{10} &= 500, 500, 500 \text{ and } 300, 500, 700 \end{aligned}$$

The cluster centers were separated by a distance of $2D$ where $D = \sqrt{\frac{\chi^2_{p,.99}}{p}}$. The value $\sqrt{\chi^2_{p,.99}}$ is the asymptotic radius of a 99% containment ellipsoid around a cluster of standard normal data. By separating the clusters at a distance of $2D$, we have clusters that do not overlap.

In a cluster situation the size of the entire dataset is known, but the size of the individual clusters in not known. Therefore, the clustering algorithm must be applied before outliers can be identified. In the clustering algorithm n_j, the number of points belonging to cluster j, is determined. From n_j we apply the theoretical formulas given for m_j and c_j. The cutoff values for 5%, 1% and 0.1% rejection are calculated for both F_{p,m_j-p+1} and χ^2_p.

3.5.1. Clean Data

For each combination of size, dimension, and number of clusters 100 sets of independent data were simulated, and the number of points the cutoffs identified as outlying was counted. (The number of independent data sets is small due to the computational intensity of the clustering algorithm.)

The tabulated results in table 4 report the percentage of data identified as outlying for each nominal level at $p = 4, 7, 10$ and cluster sizes as mentioned above. As in the one cluster situation, the F cutoff results are much closer to the target significance level than the χ^2_p cutoffs. Also, because the cluster sizes are relatively large, the asymptotic F cutoff is only slightly conservative.

3.5.2. Contaminated Data

Three methods were employed to generate outliers in each dataset. Since we have outlying points, both type I and type II errors can be measured. The type II error estimates what percentage of true outlying points would be identified as non-outlying. The type I error measures the percentage of points which were generated under the null hypothesis that were allocated to a cluster.

The outliers which were generated as a separate cluster were placed along the axis that was used to separate the two clusters at the same distance the two clusters were separated. E.g. cluster 1 is distributed $N(0, I)$, cluster 2 was distributed $N(2D, I)$, and the cluster of outliers was distributed $N(4D, I)$. (Where $D = \sqrt{\frac{\chi^2_{p,.99}}{p}}$).

The radial outliers were generated with the same center as their respective clean clusters but with a covariance of 5 times the cluster's covariance. Again, we are interested in points that are truly outlying so that we can calculate the type II error of our method, but some points that are generated with a large covariance matrix may still be close to the cluster's center. In dimension 4, 38% of points generated with $N(0, 5I)$ will be within the $\chi^2_{4,.99}$ bound of a $N(0, I)$ dataset. Using an acceptance-rejection algorithm, each outlying radial point that was generated was accepted if and only if it was outside the ellipse of clean data. The outliers were constructed to form an annulus around the clean data such that the average squared distance of an outlying point was $2\chi^2_{p,.99}$ away from the center of the clean data. By creating an annulus of outliers, we can correctly measure our type II error as those outlying points that get clustered.

The diffuse outliers were generated using the location and shape of the entire dataset. Again, we are only interested in points that are truly outlying, so we use the same acceptance-rejection algorithm to reject points that fall among the clusters of good data. In dimension 4, $n_2^4 = 300, 300$, 82.7% of the disperse points would have fallen within $\chi^2_{4,.99}$ of one of the two clusters.

The number of outliers simulated for each combination of size and dimension was constructed relative to the smallest cluster. For dimension 4, $n_2^4 = 200, 400$, if the number of outliers is 20% of the larger cluster (80 points), it would be difficult to distinguish between the good cluster of 200 points and the bad cluster of 80 points. So, the percentage of outliers of each pair of size and dimension is relative to the smaller of the two clusters. For each dimension, size, and outlier type we simulated outliers of size 20% of the smallest cluster. (In preliminary work different outlier percentages were tried and results were virtually identical.)

For each combination of size, dimension, and type of outliers, 100 sets of independent data were simulated, and the number of points the asymptotic F and the χ^2 cutoffs identified as outlying was counted.

Cluster Outliers

Table 5 gives the results for both type I and type II errors for data contaminated by a cluster of outliers. These simulations were done with two different cluster configurations, but the total sample size of the clean data stayed constant across dimension (600 total clean points in dimension 4 and 1000 total clean points in dimensions 7 and 10.) As we saw in the one cluster experiments for type I error, the theoretical F cutoff is a little conservative, but it is an improvement over the χ^2 cutoff. For type II error, there is no clustering of the outlying points except in dimension 4 at the extreme percentages.

Radial Outliers

Like the cluster outliers, the total sample size of the clean data stayed constant across dimension, and the outliers were always 20% of the smallest cluster size. In table 6 we see again that the theoretical F is an improvement over the χ^2 which is far too liberal at identifying outliers. There is more type II error with the radial outliers, but the amount is negligible except for dimension 4 with an extreme cutoff.

Diffuse Outliers

The results for diffuse outliers in table 7 are similar to those for radial outliers. Once again, the the theoretical F cutoff gives an improved but conservative bound for the type I error. The type II error is seen as a problem again only in dimension 4 with extreme cutoffs.

As in the one cluster experiments, the results do not appear to depend on p. Further simulations would probably show the dependence to be primarily on the sample size. We are encouraged to notice that the results are similar for balanced and unbalanced clusterings. The large type II error may be disconcerting, but it is important to remember that as the level of significance gets smaller, eventually the type II error will get larger. These results lead to the following recommendations:

1. With a robust initial starting value, the MCD estimator can be modified to find clusters of data.
2. With the MCD estimates, the theoretical F cutoff is superior to the χ^2_p cutoff.
3. The F cutoff method works well under the null hypothesis that all the data come from g multivariate normal populations (where g is the number of clusters). When the data are contaminated the F cutoff method still performs adequately in measures of both the type I and type II error.

3.6. Conclusion

A method for finding outliers in a robust cluster situation is given. The method relies heavily on a robust clustering, and reasons are given for the importance of a robust method for finding clusters. The F cutoff is still clearly superior to the χ^2_p cutoff which is commonly used to determine outliers.

One robust clustering method is given here, other robust methods could be applied in the same manner. Specifically, M-estimates and S-estimates could be used in place of the MCD estimate.

Also, this method could be further explored by using more clusters, different outlier configurations, and the simulated values of m and c as parameters for the F cutoff.

Chapter 4

Outlier Detection and Clustering with S-Estimation

4.1. Introduction

In the first chapters of this work, robust methods for finding outliers in the one and multiple cluster case were given. The robust estimator used was the minimum covariance determinant (MCD) (Hampel et al., 1986; Rousseeuw, 1984; Rousseeuw and Leroy, 1987) which was modified in our work for the multiple cluster case. An extension to this outlier detection work is to use a different robust estimator in place of the MCD.

M-estimation was developed as a generalization of maximum likelihood estimation (thus the "M".) The idea is to solve a system of equations which can be independent of the underlying probability distribution function whereas maximum likelihood maximizes the likelihood function. The optimization can then be developed into an iteration scheme which assigns weights to data points. The weights are smaller for extreme points and can be used to calculate robust estimates for the mean and covariance. With certain optimization functions, a constraint is needed in the M-estimation to keep the estimates from imploding. S-estimation was developed as a constrained optimization problem and has since been shown to be a subset of M-estimation with a particular type of constraint.

M- and S-estimators which give smaller weights to extreme points and zero weight to extreme points which are larger than some finite value are called redescending and suffer from the problem that as more points are given zero weight, the determinant of the covariance decreases which leads to large Mahalanobis distances which leads to more points getting zero weight. If no constraints are put on the algorithm to compute a redescending estimator, eventually all the points will be given zero weight. In order to converge to a solution some constraint needs to be added to the problem.

In chapter 2, we used a robust technique to estimate location and shape for a one cluster dataset. Once the estimates were defined, we used them in the metric to compute Mahalanobis Squared Distances (MSDs) for each point from the robust center estimate. With an F distribution cutoff, points were determined as outlying if they had an MSD larger than some predetermined cutoff. In chapter 2, we used a robust technique to estimate location and shape for clusters ($g \geq 2$, where g is the number of clusters). With the metrics defined, we calculated the distances of points to clusters, and using those distances we determined which points belonged to which cluster. With an F-distribution cutoff, points were determined as outlying if they had a distance larger than some predetermined bound. This method of finding outliers in the one or multiple cluster case can be generalized by substituting the robust MCD estimates with any other robust shape and location estimator.

Properties of S-estimates are generally well known, and the asymptotics associated with them are often good in small samples. S-estimators are relatively more efficient than MCD estimators (Rocke and Woodruff, 1997). With similar sample sizes as used with the MCD, the S-estimators will perform better in outlier detection when using the same type of F-cutoff.

We will investigate the quality of finding outliers for various values of n, p, g, and different arrangements and percentages of outlying points for the translated biweight S-estimator.

4.2. M- and S- Estimation

4.2.1. M- Estimation

M-estimation of a one-dimensional location parameter was introduced by Huber (Huber, 1964). Later, Maronna defined M-estimates for multivariate location and shape (Maronna, 1976). Huber extended Maronna's definition to a solution of simultaneous equations (Huber, 1981). The idea for an M-estimator comes from the equations used to solve for maximum likelihood estimators.

Consider the maximum likelihood estimation of μ and Σ for a multivariate family of densities $\mathrm{g}(X) = |\Sigma|^{-\frac{1}{2}}\ \mathrm{f}(d^2)$ where $d^2 = (X-\mu)^t\Sigma^{-1}(X-\mu)$. The maximum likelihood estimates $\tilde{\mu}$ and $\tilde{\Sigma}$ will maximize

$$\prod_i |\Sigma|^{-\frac{1}{2}}\ \mathrm{f}(d_i^2).$$

where $d_i^2 = (x_i-\mu)^t\Sigma^{-1}(x_i-\mu)$ is the argument based on the realization x_i. Letting $\rho(x) = \log(\mathrm{f})$ and $\psi(x) = -\frac{d}{dx}\rho(x)$, the above maximization can be represented as

$$\begin{aligned}\sum_i \psi(d_i) &= 0\\ \sum_i [d_i\psi(d_i) - 1] &= 0\end{aligned}$$

M-estimation is a generalization of maximum likelihood estimation. In M-estimation ψ and ρ are not restricted by the presumptive probability density function, and estimates for μ and Σ are found by solving

$$\begin{aligned}\sum_i \psi(d_i) &= 0\\ \sum_i \chi(d_i) &= 0\end{aligned}$$

where, usually, ψ is an even function and χ is an odd function (here $\chi(d) = d \cdot \psi(d) - 1$.)

In maximum likelihood, the optimization equations can be written as

$$0 = \sum_i w(d_i)\Sigma^{-1}(x_i - \mu)$$

$$0 = -\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\sum_i w(d_i)\Sigma^{-1}(x_i-\mu)(x_i-\mu)^t\Sigma^{-1}$$

where $w(x) = \frac{\psi(x)}{x}$ and $\psi(x) = -\frac{d}{dx}f(x)$. From the equations, the iterated estimates for μ and Σ can be written as

$$\begin{aligned} \tilde{\mu}^{(j+1)} &= \frac{\sum_i w(d_i^{(j)})x_i}{\sum_i w(d_i^{(j)})} \\ \tilde{\Sigma}^{(j+1)} &= n^{-1}\sum_i w(d_i^{(j)})(x_i-\tilde{\mu}^{(j+1)})(x_i-\tilde{\mu}^{(j+1)})^t \end{aligned}$$

Similarly, in M-estimation the optimization conditions can be represented as

$$\begin{aligned} 0 &= n^{-1}\sum_i v_1(d_i)(x_i-\mu) \\ 0 &= n^{-1}\sum_i [v_2(d_i)(x_i-\mu)(x_i-\mu)^t - v_3(d_i)\Sigma] \end{aligned}$$

where v_1, v_2, and v_3 are the weight functions (which depend on ψ and reduce to $v_1(d) = v_2(d) = \frac{1}{d}\cdot\frac{-\partial\log \mathrm{f}(d)}{\partial d} = \frac{1}{d}\cdot\frac{-\partial\rho(d)}{\partial d}$ and $v_3(d) = 1$ for maximum likelihood.) The iterated parameter M-estimates can be written as

$$\tilde{\mu}^{(j+1)} = \frac{\sum_i v_1(d_i^{(j)})x_i}{\sum_i v_1(d_i^{(j)})} \tag{2.1}$$

$$\tilde{\Sigma}^{(j+1)} = \frac{\sum_i v_2(d_i^{(j)})(x_i-\tilde{\mu}^{(j+1)})(x_i-\tilde{\mu}^{(j+1)})^t}{\sum_i v_3(d_i^{(j)})} \tag{2.2}$$

Our interest in M-estimates is due to their robust qualities. A redescending ψ function will give smaller weights to extreme points thus insuring outlying points from heavily influencing the estimates. A choice of ψ that is not redescending will not be as robust to outlying points. (E.g., the sample mean gives equal weight to all points and is highly sensitive to extrema.) However, a choice of ψ that redescends to zero can cause problems in the estimation. As the estimate for $|\Sigma|$ shrinks, the d_i will grow, and fewer and fewer points will have non-zero weight. Each additional point that has zero weight decreases the determinant of the covariance estimate which increases all of the MSDs of the data. Eventually, if enough points have zero weight, $\tilde{\Sigma}$ will be singular, and the iteration process will fail to converge.

The iteration scheme can easily wander into one of these non-converging regions if there are no constraints placed on the distances. For any M-estimator that has a redescending ψ function, a constraint is needed in the estimation algorithm.

A possibility for constraining the estimates is to constrain the MSD according to some expectation under multivariate normality. Some possibilities are:

$$\begin{aligned} n^{-1}\sum_i \rho(d_i) &= E(\rho(d)) \\ n^{-1}\sum_i \psi(d_i) &= E(\psi(d)) \\ n^{-1}\sum_i \psi(d_i)d_i &= E(\psi(d)d) \\ n^{-1}\sum_i \psi^2(d_i) &= E(\psi^2(d)) \\ n^{-1}\sum_i w(d_i) &= E(w(d)) \\ \text{median}(d_i) &= \text{median}(d) \end{aligned}$$

where ρ is a given optimization function, $\psi(d) = \frac{\partial\rho(d)}{\partial d}$, and $w(d) = \frac{\psi(d)}{d}$. For each data point, the distance will be scaled so that the chosen constraint is satisfied. With any of the above constraints, we can choose a redescending ψ function to obtain robust estimates of shape and location for multivariate data.

4.2.2. S- Estimation

S-estimation originated in the regression context (Rousseeuw and Yohai, 1984) as a constrained optimization problem. Later, it was applied to the multivariate scale and location estimation problem (Davies, 1987). Lopuhaä showed that an S-estimate of location and scale is a type of constrained M-estimate (Lopuhaä, 1989). An S-estimate of multivariate location and shape is defined as follows:

Definition 4.2.1 *(Rousseeuw and Yohai, 1984) Let $\rho : I\!R^+ \rightarrow I\!R^+$ be a twice continuously differentiable, symmetric, nondecreasing function which has $\rho(0) = 0$*

and is constant at $\rho(x) = \rho(c)\ \ \forall x \geq c$. *Given a dataset of* n *points in* $I\!R^p$*, let the S-estimator,* $(\tilde{\mu}, \tilde{\Sigma})$*, be defined by minimizing* $|\tilde{\Sigma}|$ *subject to*

$$n^{-1} \sum_i \rho(d_i) \quad = \quad b_0.$$

Lopuhaä (1989) showed that S-estimates are a case of constrained M-estimates where

$$\begin{aligned}
v_1(d) &= w(d) \\
v_2(d) &= w(d) \\
v_3(d) &= v(d) \\
\psi(d) &= \frac{\partial \rho(d)}{\partial d} \\
w(d) &= v(d) = \frac{\psi(d)}{d}
\end{aligned}$$

with constraint $n^{-1} \sum_i \rho(d_i) = b_0 = E(\rho(d))$. As S-estimates are simply a type of constrained M-estimate, this work will address some common redescending S-estimators.

Examples

Below are five examples of S-estimators that contain redescending ψ functions. Each estimator is determined by the ρ function, but ρ, ψ and the weights, w, are given for completeness.

<u>Andrew's Wave (c)</u> (Andrews et al., 1972):

$$\rho_{aw} = \begin{cases} \frac{c}{\pi^2}(1 - \cos \pi \frac{d}{c}) & d \leq c \\ \frac{2c}{\pi^2} & d > c \end{cases}$$

$$\psi_{aw} = \begin{cases} \frac{1}{\pi} \sin \pi \frac{d}{c} & d \leq c \\ 0 & d > c \end{cases}$$

$$w_{aw} = \begin{cases} \frac{1}{\pi d} \sin \pi \frac{d}{c} & d \leq c \\ 0 & d > c \end{cases}$$

Hampel, three-part redescending (a, b, c) (Andrews et al., 1972) (where $0 < a \leq b \leq c$):

$$\rho_h = \begin{cases} \frac{1}{2}d^2 & d \leq a \\ ad - \frac{1}{2}a^2 & a < d \leq b \\ ab - \frac{1}{2}a^2 + (c-b)\frac{a}{2}[1 - (\frac{c-d}{c-b})^2] & b < d \leq c \\ ab - \frac{1}{2}a^2 + (c-b)\frac{a}{2} & d > c \end{cases}$$

$$\psi_h = \begin{cases} d & d \leq a \\ a & a < d \leq b \\ a\frac{c-d}{c-b} & b < d \leq c \\ 0 & d > c \end{cases}$$

$$w_h = \begin{cases} 1 & d \leq a \\ \frac{a}{d} & a < d \leq b \\ \frac{a(c-d)}{d(c-b)} & b < d \leq c \\ 0 & d > c \end{cases}$$

Tukey's Biweight(c) (Tukey, 1972):

$$\rho_b = \begin{cases} \frac{c^2}{6}[1 - (1 - (\frac{d}{c})^2)^3] & d \leq c \\ \frac{c^2}{6} & d > c \end{cases}$$

$$\psi_b = \begin{cases} d(1 - (\frac{d}{c})^2)^2 & d \leq c \\ 0 & d > c \end{cases}$$

$$w_b = \begin{cases} (1 - (\frac{d}{c})^2)^2 & d \leq c \\ 0 & d > c \end{cases}$$

Least Winsorized Squares (c) (Rousseeuw and Leroy, 1987):

$$\rho_{lws} = \begin{cases} \frac{d^2}{2} & d \leq c \\ \frac{c^2}{2} & d > c \end{cases}$$

$$\psi_{lws} = \begin{cases} d & d \leq c \\ 0 & d > c \end{cases}$$

$$w_{lws} = \begin{cases} 1 & d \leq c \\ 0 & d > c \end{cases}$$

<u>Translated-Biweight (a, b)</u> (Rocke, 1996):

$$\rho_{tb} = \begin{cases} \frac{d^2}{2} & d < a \\ \frac{a^2}{2} - \frac{a^2(a^4-5a^2b^2+15b^4)}{30b^4} + & \\ d^2(\frac{1}{2} + \frac{a^4}{2b^4} - \frac{a^2}{b^2}) + d^3(\frac{4a}{3b^2} - \frac{4a^3}{3b^4}) + & \\ d^4(\frac{3a^2}{2b^4} - \frac{1}{2b^2}) - \frac{d^5 4a}{5b^4} + \frac{d^6}{6b^4} & a \leq d \leq a+b \\ \frac{a^2}{2} + \frac{b(5b+16a)}{30} & d > a+b \end{cases}$$

$$\psi_{tb} = \begin{cases} d & d < a \\ d(1-(\frac{d-a}{b})^2)^2 & a \leq d \leq a+b \\ 0 & d > a+b \end{cases}$$

$$w_{tb} = \begin{cases} 1 & d < a \\ (1-(\frac{d-a}{b})^2)^2 & a \leq d \leq a+b \\ 0 & d > a+b \end{cases}$$

The five above examples are all redescending estimators, but each is slightly different. One common trait across these functions is that the ψ function is linear close to zero. Winsor's principle, quoted by Tukey (p. 457) (Tukey, 1960), states "all distributions are normal in the middle." The ψ function for the maximum likelihood estimate of the mean for normal data is linear. So, a good S-estimate allows for Winsor's principle and has a ψ function which, near zero, resembles that which is best for Gaussian data.

An important difference in the above S-estimators is the number of parameters in each. Hampel's wave takes three parameters, Rocke's translated-biweight takes two parameters, and the others each take one parameter. Two factors in choosing an appropriate estimator for robust estimation are the breakdown and the asymptotic rejection point. The number of allowable parameters in an S-estimate determines the flexibility in the values of these two factors.

The breakdown of an estimator can be of two types, the replacement breakdown and the additive breakdown. The two breakdowns are similar in many contexts but must be defined as different quantities. Replacement breakdown is a number associated with replacing one or more of the data points with any possible value. Additive breakdown is a number associated with adding data to the original set. Both types measure how well the estimator behaves with these restructured datasets. As a robust criterion, we use the replacement breakdown.

Definition 4.2.1 *(Lopuhaä and Rousseeuw, 1991) The replacement breakdown point of a location estimator t_n at a collection X is defined as the smallest fraction $\frac{m}{n}$ of outliers that can take the estimator over all bounds:*

$$\epsilon^*(t_n, X) = \min_{1 \leq m \leq n} \left\{ \frac{m}{n} : \sup_{Y_m} ||t_n(X) - t_n(Y_m)|| = \infty \right\} \tag{2.3}$$

where the supremum is taken over all possible corrupted collections Y_m that can be obtained from X by replacing m points of X by arbitrary values.

The breakdown of a covariance estimator, C_n, at a collection X is defined as the smallest fraction $\frac{m}{n}$ of outliers that can either take the largest eigenvalue $\lambda_1(C_n)$ over all bounds, or take the smallest eigenvalue $\lambda_p(C_n)$ arbitrarily close to 0:

$$\epsilon^*(C_n, X) = \min_{1 \leq m \leq n} \left\{ \frac{m}{n} : \sup_{Y_m} D(C_n(X) - C_n(Y_m)) = \infty \right\} \tag{2.4}$$

where the supremum is taken over the same corrupted collections Y_m as in (2.3) and where $D(A, B) = \max\{|\lambda_1(A) - \lambda_1(B)|, |\lambda_p(A)^{-1} - \lambda_p(B)^{-1}|\}$ with $\lambda_1 \geq \ldots \geq \lambda_p$ being the ordered eigenvalues.

In robust estimation we are interested in estimators that have high breakdown. The mean has a breakdown of $1/n$. In dimension 1 the median has a breakdown of

$\frac{\lfloor (n+1)/2 \rfloor}{n}$. It is important to have estimators that are resistant to a certain amount of contamination.

Theorem 4.2.1 *(Lopuhaä and Rousseeuw, 1991) Let X be a set of $n \geq p+1$ points in $\mathbb{R}^p$. Write $r = b_0/\sup(\rho)$, where b_0 is the value used in the S-estimation constraint. If $r \leq (n-p)/(2n)$ then S-estimates defined by a function ρ that satisfies Definition 2.1 have a breakdown point*

$$\epsilon^*(t_n, X) = \epsilon^*(C_n, X) = \frac{\lfloor nr \rfloor}{n} \approx r.$$

When $r = (n-p)/(2n)$, the breakdown is $\epsilon^* = \frac{\lfloor (n-p)/2 \rfloor}{n}$.

Lopuhaä showed that S-estimators are a subclass of M-estimators, and Lopuhaä and Rousseeuw showed that the breakdown of the S-estimators is r ; yet there is a commonly cited (Davies, 1987; Lopuhaä, 1989; Rocke, 1996; Barnett and Lewis, 1994) result that M-estimators have a breakdown of $1/(p+1)$ (Maronna, 1976). These results appear contradictory, but they are not. Because M-estimates are not constrained, in high dimensions it is possible to have more than one root to the optimization equations. Maronna showed that there exists a root that solves the optimization criteria and breaks down at $1/(p+1)$. He did not, however, show that there exists no root with a higher breakdown. The S-estimation work shows that it is possible that an estimator of higher breakdown can occur. (Rocke, 1998)

Another important concept is that of the asymptotic rejection point.

Definition 4.2.1 *(Rocke, 1996) Consider a redescending M- or S-estimator, in which $c_0 = inf\{d_0 | w(d) = 0, \forall d > d_0\}$ where $w = \psi(d)/d$. The asymptotic rejection probability (ARP) of this estimator is then defined as the probability in large samples under a reference distribution (usually multivariate normal) that a MSD exceeds c_0. If the estimator is normed to the normal distribution, the ARP is $1 - F_{\chi^2(p)}(c_0^2)$.*

The ARP gives the percentage of data points that would be given zero weight if in fact the data were uncontaminated (and in our case distributed multivariate normal.) It is clear that we can choose c_0 to give any value for the ARP, but manipulating the value of the ARP will also change the breakdown of the estimator.

Since the biweight estimator is a one parameter family, it is not possible to set both the breakdown and the ARP by manipulating the parameter. As shown by Rocke (1996), when the breakdown is set to accommodate robust situations, the ARP drops to such a low level that the biweight would fail to identify even many pathological outliers. The maximum breakdown for S-estimators is $\frac{\lfloor (n-p)/2 \rfloor}{n}$, which approaches 50% of the data. When we apply this breakdown in high dimensions to the biweight, the ARP becomes unacceptable for robust estimation. For example, when $p = 20$ and $r = 0.5$ (large sample) then $c = 9.72$ and the ARP $= 1 - F_{\chi^2(20)}(94.5) = 10^{-11}$. This means that in order for a data point to be declared an outlier (and thus given zero weight) the point must have $d^2 \geq 94.5$ which is 12 standard deviations away from the mean (Rocke, 1996). It should be agreed that a point might be considered outlying even if it is closer to the bulk of the data than 12 standard deviations away.

In order to have control over both the ARP and the breakdown, we need to use an estimator that has at least two adjustable parameters. Both the translated-biweight (t-biweight) and Hampel's wave allow for the dual constraints. Though Hampel's functions probably perform equally well in simulations, we chose to use the t-biweight in this work because of its continuous first derivatives.

Though we are not constrained by only one parameter, not every combination of breakdown and ARP is possible with the t-biweight. Our programs are set to have the maximum breakdown, and the user is able to choose the ARP. If the user chooses an ARP that is too large, the parameter b is reduced, and as $b \to 0$ the limit of the t-biweight is the least Winsorized squares estimator. If the user chooses an ARP that is too small, the parameter a is reduced, and as $a \to 0$, the limit of the t-biweight is the biweight estimator. Using the t-biweight, the two parameters a and b can be chosen to give the desired breakdown and ARP subject to the estimator changing in the limit. In other words, given a particular breakdown point, the ARP of the t-biweight cannot be larger than that of the least Winsorized squares estimator or smaller than that of the biweight estimator (Rocke, 1996).

4.2.3. Comparison with the MCD

In the first chapters of this work, we have studied behavior of the MCD estimator. The MCD has some favorable properties, namely, it is computable without a robust starting value and it has a high breakdown. Also, we know some of the asymptotic properties of the estimator, and these properties were used in describing the behavior of the extreme data points using the MCD metric. Though the MCD and the S-estimator are both $n^{-1/2}$ estimators (Butler et al., 1993), the relative efficiency of the S-estimator is much higher than the MCD (Rocke and Woodruff, 1997). This means that in order for the MCD to work as well as the S-estimator, more data are needed.

4.3. Computing the S-Estimate

For reasons discussed above, we use the t-biweight in our simulations. For the one cluster case, the algorithm for computing the S-estimate is fairly straight forward. For the multiple cluster case, the algorithm is modified to allow for any number of clusters. One potential problem with S-estimates is that they are sensitive to a good starting value. As with the MCD we use the CLUSTER program to initialize our estimates when the data come from two or more populations.

4.3.1. The One Cluster Setting

S-estimates have been used as robust estimates of location and shape of multivariate data (Davies, 1987). Davies' optimization equations can be solved as a system of iterative equations, see eqs. 2.1 and 2.2. The solutions give values for $\tilde{\mu}$ and $\tilde{\Sigma}$. The algorithm is as follows.

1. Start with an initial estimate of $\tilde{\mu}$ and $\tilde{\Sigma}$.
2. Compute $d_i^2 = (x_i - \tilde{\mu})^t \tilde{\Sigma}^{-1} (x_i - \tilde{\mu})$ using the most recent estimates of $\tilde{\mu}$ and $\tilde{\Sigma}$.
3. Compute k such that $n^{-1} \sum_i \rho(d_i/k) = b_0$. (Where $b_0 = E[\rho(d)]$ under multivariate normality.)
4. Replace d_i with $\tilde{d}_i = d_i/k$.
5. Find
$$\begin{aligned} \tilde{\mu} &= \frac{\sum_i w(\tilde{d}_i) x_i}{\sum_i w(\tilde{d}_i)} \\ \tilde{\Sigma} &= \frac{p \sum_i w(\tilde{d}_i)(x_i - \tilde{\mu})(x_i - \tilde{\mu})^t}{\sum_i w(\tilde{d}_i)}. \end{aligned}$$
6. If $\tilde{\mu}$ and $\tilde{\Sigma}$ have changed, repeat 2-5. Otherwise, report $\tilde{\mu}$ and $\tilde{\Sigma}$.

Our program uses the t-biweight, and therefore $w = v$. With the t-biweight function a point which is far from the mean of a cluster will have a large MSD for that cluster. The t-biweight takes arguments a and b as well as the MSD for a point using the current metric. Beyond $a + b$ data points will be given zero weight in the estimating equations. The ARP defines the value for $a + b$ and can be set by the user. The value a will then automatically be calculated to give the correct breakdown. In this work we set $r = \frac{n-p}{2n}$, as this allows for the maximum breakdown possible for S-estimators.

Step 1 requires an initial estimate for $\tilde{\mu}$ and $\tilde{\Sigma}$ to be available. In the one cluster setting we used both uncontaminated and contaminated data. To get an initial estimate of parameters, we used the ordinary mean and covariance estimates of the entire clean dataset. (Usually, the clean part of the data would, obviously, not be known.) In practice, some sort of robust estimates should be used to initialize the algorithm. One option is to run multiple starts with randomly selected points and then use some criterion to determine which initialization is best. Another option is to initialize the program with the MCD estimates. Our choice of using the sample mean and covariance as initial estimates was done in the interest of time.

4.3.2. The Multiple Cluster Setting

The algorithm for finding the t-biweight estimates in the multiple cluster setting is similar to that in the one cluster setting with these modifications:

(a) The initialization is done with the program CLUSTER.

(b) The sample size of the clusters is re-estimated at each iteration because the number of points allocated to each cluster can change throughout the program.

(c) The estimates for each group are calculated one cluster at a time.

(d) The number of clusters in the dataset must be determined prior to the analysis.

The algorithm is as follows:

1. Use the CLUSTER program to give an initial clustering of the data based on n, p, and g.

2. Using the CLUSTER assignments, calculate a mean and covariance for each cluster. (E.g., the mean for cluster 1 will be the average of those points which have been assigned to cluster 1 by the CLUSTER program.)

3. Calculate the MSDs, based on the most recently calculated cluster means and covariances, for each point in the entire dataset.

4. Assign each point to the cluster for which it has the smallest MSD. Also, assign a cluster size to each cluster based on the number of points that are closest to that cluster.

5. For each cluster find the t-biweight S-estimate of the mean and covariance based on the above distances and cluster sizes.

6. Repeat steps 3-5 until the parameter estimates no longer change.

7. Report $\tilde{\mu_i}$ and $\tilde{\Sigma}_i$ for each cluster.

Let the cluster sizes be $n_1, n_2, \ldots, n_g$, for g clusters. Then the breakdown parameter is set at $r_j = \frac{\lfloor (n_j - p)/2 \rfloor}{n_j} \quad j = 1, g$. The other parameters will be set to have a specified ARP (ARP = 0.01 in our work) for each cluster.

4.4. Identifying Outliers

The goal of finding S-estimates of shape and location is to use the estimates in identifying outliers in the data. In chapters 2 and 3 we used the F distribution to approximate the tail behavior of distances based on MCD estimates. The motivation was that an MCD covariance can be thought of to have Wishart-like first two moments, and that property is used to establish degrees of freedom appropriate for an F distribution which describes the related distances.

The theorem in chapter 2 states that the MCD estimates are asymptotically independent of the tail distances which motivates the use of an F statistic. The theorem can be directly applied in the S-estimation situation only to points outside the ARP bound. Points that are extreme will be given zero weight and will not be used in the estimates. Those extreme points are the subject of interest in this paper as they will be identified as outlying. It can be argued that any extreme points are independent of the S-estimates if extreme is defined to mean outside the ARP rejection bound.

The ARP has been defined as the percentage of points that would be given zero weight if in fact the data were uncontaminated. Extreme points that are outside the ARP cutoff will be given zero weight and therefore will not be strong influences on the S-estimates. In the MCD case, the extreme points also did not directly influence the estimates. Since the interest of this work is in the extreme points, we can use the logic from chapter 2 to apply an F statistic as a cutoff for an MSD with an S-estimate metric. Points that are outside the ARP bound will be treated like the MCD extrema from chapters 2 and 3. Points that are inside the ARP bound but also extreme will not be highly correlated with the estimates, and so heuristically the methods should approximate the behavior of these semi-extreme points. We will analyze points inside and outside the ARP bound in the interest of examining many different situations.

4.4.1. F Cutoff

As with the MCD estimates, we are interested in finding a statistic that would allow identification of outliers at a particular level. The derivation of approximating the MSDs with an F statistic uses the idea that the first few moments of the covariance resemble those of a Wishart distribution. Though the MCD does not have an exact Wishart distribution, the approximation was OK in small samples and seemed to be quite good in large samples. The sample covariance matrix does have a Wishart distribution (if the data come from a multivariate normal distribution), and the definition of an S-estimate is more like the sample covariance than that of the MCD. This fact, along with the relative efficiency of the S-estimate being larger than the MCD, leads us to think that the Wishart approximation will be better for the S-estimate than it was for the MCD.

4.4.2. Degrees of Freedom

A Wishart approximation is incomplete without an estimate for the degrees of freedom. Using a method of moments approximation in chapter 2, we found we could use the moments of the diagonal elements of the covariance matrix to estimate the degrees of freedom. In the MCD work we used both simulation and a theoretical formula due to Croux and Haesbrock (2000) to estimate the moments of the MCD shape matrix which gave the appropriate degrees of freedom.

S-estimates of covariance are scaled to be consistent for the population covariance (under multivariate normality assumption) (Davies, 1987). Consistency gives us asymptotic unbiasedness, so only the second moment is needed to find the desired degrees of freedom. Applying the logic from the development of the MCD degrees of freedom to the S-estimate degrees of freedom, we get:

$$\tilde{m} = \frac{2}{\mathrm{Var}[\tilde{s_{ii}}]}$$

where Var $[\tilde{s_{ii}}]$ is the variance of the diagonal elements of the S-estimate covariance matrix.

We can estimate the variance in two ways: through simulation and through an asymptotic formula. (Lopuhaä, 1989; Davies, 1987). Since the diagonal elements are identically distributed and uncorrelated, we can simulate N copies of the $p \times p$ S-estimate shape matrix from the n data points in each sample and then estimate m from the variance of the Np diagonal elements. Lopuhaä (1989) derives the value of the variance of the diagonal elements under standard normality as:

$$\begin{aligned}
\mathrm{Var}[\tilde{s_{ii}}] &= \frac{2\sigma_1 + \sigma_2}{n} \\
\text{where} & \\
\sigma_1 &= \frac{p(p+2)E_{0,I}[\psi^2(||X||)||X||^2]}{\{E_{0,I}[\psi'(||X||)||X||^2 + (p+1)\psi(||X||)||X||]\}^2}, \\
\sigma_2 &= \frac{-2}{p}\sigma_1 + \frac{4E_{0,I}[(\rho(||X||) - b_0)^2]}{\{E_{0,I}[\psi(||X||)||X||]\}^2},
\end{aligned}$$

$$b_0 \;=\; E_{0,I}[\rho(||X||)],$$

$E_{0,I}$ is the expectation under standard multivariate normality, and $||X||$ is the norm of a vector X in $I\!R^p$. Details for the computable form of the theoretical formulas for $\tilde{m}$ are given in Appendix B. From the results in the one cluster case, it is seen that the simulated and theoretical estimates give similar values. In the multiple cluster case, only the theoretical estimates are used due to computing constraints. Because the cluster sizes are unknown in the multiple cluster case, it is difficult to simulate the m parameter for every situation. (Even though the simulated cluster sizes are known to the user, these sometimes change slightly in the clustering step.)

Using the estimated variance we get an estimate for the approximate degrees of freedom for a Wishart distribution that describes an S-estimate. The degrees of freedom is then used, with the size and dimension, to calculate the correct cutoff for the MSDs with S-estimate metrics. As in the MCD work, the final approximation is:

$$\frac{(\tilde{m}-p+1)}{p\tilde{m}} d^2_{\tilde{\Sigma}}(x_i, \tilde{\mu}) \;\dot{\sim}\; F_{p,\tilde{m}-p+1}$$

where $\tilde{\mu}$ and $\tilde{\Sigma}$ are respectively the location and shape S-estimates, and $\tilde{m}$ is the estimated degrees of freedom for the approximate Wishart distribution.

4.5. Results

In this chapter we have attempted to do two things, cluster the data points using S-estimation and establish outlier rejection formulas for various cluster settings. As in previous sections we use robust distances, now with the metric defined by an S-estimate, to identify outliers based on distributional quantiles. Experiments were done for various dimensions and sizes as well as number of clusters, cluster configurations, and outlier configurations.

4.5.1. One Cluster

The one cluster (or one population) data were generated as clean data from a multivariate normal distribution for various sizes and dimensions. Since in our analyses, we know that there is only one cluster, the interest is in outlier detection at various cutoff percentages instead of clustering. The S-estimates were calculated as in section 4.3.1, and MSDs were determined using the S-estimate metric. For each combination of dimension and size three cutoff values were calculated,

1. χ_p^2,
2. F with degrees of freedom calculated from the asymptotic formulas, and
3. F with degrees of freedom calculated from simulations.

The quantiles for the distributional cutoffs were set for both pointwise and dataset-wise rejection at levels 5%, 1%, and 0.1%. We discuss the distinction between the two types of rejection.

Pointwise Rejection

Pointwise rejection refers to the usual type I error encountered when the data fall into some natural ordering. It is rejecting a point because the distance to the point is too large (or small) to reasonably belong to the null data. If our focus is on stray points that might require special attention, it is best to use pointwise rejection. The results based on 1000 trials of independent uncontaminated multivariate normal data for 5%, 1%, and 0.1% levels for all cutoffs at $p = 4, 7, 10$, and $n = 50, 100, 500, 1000$ are provided in table 8.

As seen before, the χ_p^2 cutoff is liberal, the theoretical F is conservative, and the simulated F is quite good. However, it seems that the asymptotics work well for the S-estimators, and all three cutoffs give reasonable values at the specified nominal levels.

Datasetwise Rejection

Datasetwise rejection is more conservative and focuses on the dataset as a whole rather than the individual points. If the interest is in anything unusual happening in the *sample* that may require special attention, it is best to use datasetwise rejection. Here, the significance is determined so that the probability that an outlier is identified in the data is set to the desired level. Using a Bonferonni approximation to obtain the desired level of significance, the cutoffs are set at levels of $.05/n, .01/n$, and $.001/n$ to give datasetwise rejection levels of $.05, .01$, and $.001$. The results for 5%, 1%, and 0.1% datasetwise rejection for all cutoffs and $p = 4, 7, 10$ and $n = 50, 100, 500, 1000$ are provided in table 9.

Again, the χ^2_p cutoff is liberal, the theoretical F is conservative, and the simulated F is quite accurate. However, in these simulations there appears to be a lot of variability in the percentages. At the extreme levels it is not surprising that the tail behavior is sporadic. Though the χ^2_p cutoff did well at rejecting at the pointwise level, very large data sets are needed to get accurate χ^2_p cutoff values for datasetwise rejection.

From these simulations it appears that if moments are simulated, the first two moments of the covariance of an S-estimate behave like the first few moments of a Wishart variable.

The results for 1000 trials of contaminated data for $p = 4, 7, 10$ and $n = 50, 100, 500, 1000$ are provided. Table 10 gives the results for data that have been contaminated using a cluster of outliers. Table 11 gives results for data that have been contaminated by radial outliers. The contamination was done identically to that in chapter 3, and the radial outliers were again simulated using the acceptance-rejection algorithm. The outliers are always 20% of the good data (n). Since the data are contaminated, interest is now in both the correct classification of the good data and the misclassification of the contaminated data. Like in chapter 3, we call type I error the percentage of points which have been generated by a multivariate normal distribution but are classified as outlying. Type II error is the percentage of points which have been generated as contamination yet are classified into one of the clusters.

Both tables 10 and 11 show that all the cutoffs are conservative, and the χ^2_p seems to be superior over the theoretical F as all nominative levels and in both the type I and type II error.

From the clean and contaminated one cluster simulations it is seen that the asymptotic F cutoff is better under the null hypothesis (especially in considering datasetwise rejection), but the methods are equal (or maybe the χ^2_p is a little better) when the data are contaminated.

4.5.2. Multiple Clusters

The data in this section were simulated in the same way as was done in chapter 3. The simulations were of clean, cluster outliers, radial outliers, and diffuse outliers at $p = 4, 7, 10$ and

$$\begin{aligned}
n_2^4 &= 50, 50 \text{ and } 300, 300 \text{ and } 200, 400 \\
n_3^4 &= 50, 50, 50 \text{ and } 300, 300, 300 \text{ and } 200, 300, 400 \\
n_2^7 &= 80, 80 \text{ and } 500, 500 \text{ and } 300, 700 \\
n_3^7 &= 80, 80, 80 \text{ and } 500, 500, 500 \text{ and } 300, 500, 700 \\
n_2^{10} &= 100, 100 \text{ and } 500, 500 \text{ and } 300, 700 \\
n_3^{10} &= 100, 100, 100 \text{ and } 500, 500, 500 \text{ and } 300, 500, 700
\end{aligned}$$

where n_g^p is the cluster sizes and configurations in dimension p with g groups. With S-estimation, smaller data sets were simulated because the initial work showed that the estimator did well even with small data sets.

As seen in table 12, both the theoretical F cutoff and the χ^2_p cutoff fairly accurately identified the correct percentage of outliers for nominal levels 5%, 1%, 0.1%. The results for two clusters and three clusters are similar, so we can hypothesize that this method would be successful with more than three clusters. The simulations of small sample sizes show the χ_p^2 to be more liberal than the F, especially at the extreme levels.

Tables 13, 14, and 15 give the results for the contaminated data. Again, we are interested in both type I and type II error for the contaminated data. It is seen that the theoretical F and the χ_p^2 cutoff give very similar results for all contamination types and sample sizes. The only type II error of any significance is seen in small data sets, and this error is seen with both the χ_p^2 and the F cutoffs.

The results from the S-estimate simulations show that the methods do not appear to depend on p. In both the one and multiple cluster setting, the χ_p^2 cutoff is good and the theoretical cutoff is excellent when the data is uncontaminated. When the data is contaminated, both methods work well at classifying the clean data (though they are both slightly conservative), and both methods have some type II error when the sample size is quite small. These results lead to the following recommendations:

1. With a robust initial starting value, the translated biweight S-estimator can be used to cluster the data.
2. If the S-estimate metric is used, in most cases the asymptotic F cutoff works just as well as the χ_p^2 cutoff.
3. With small data sets, the asymptotic F cutoff is superior to the χ_p^2 cutoff, but type II error must be considered.

4.6. Conclusion

A robust method for clustering data and identifying outliers is given. This method is dependent on a robust starting point, but any good robust starting point should perform equally well.

The S-estimate was examined because its asymptotic properties are more well known than those of the MCD. The performance of the method is excellent when the data are uncontaminated, but the MCD appears to have better type II error with contaminated data.

This method could continue to be explored using other S-estimators or other robust estimators. Also, we could use different initial clustering methods to examine how the starting estimates affect the algorithm.

Chapter 5

Conclusion

In this thesis we have explored new robust methods of treating multivariate data. Our philosophy is that though robust methods may not always be necessary, there are often times when they are invaluable. Also, it is often difficult to discern whether or not a robust method should be used, and it is wise to err on the side of being cautious.

We began with a method that identified outliers based on an F cutoff which was derived from a method of moments estimate and was compared to distances which had a Minimum Covariance Determinant metric. The MCD was an important part of the analysis because of its robust properties. We knew that the distances with the MCD metric had an asymptotic χ^2_p distribution, but the χ^2_p quantiles were very liberal in outlier identification for all but very large sample sizes. The simulated F cutoff worked quite well, but finding the correct degrees of freedom was computationally intensive. The F cutoff based on theoretical degrees of freedom gave good but conservative results and were easy to compute.

The robust properties of the MCD estimator led us to think that it could be used in clustering data that had significant contamination. Since the F with the simulated degrees of freedom was difficult to compute, our results were based on F with theoretical degrees of freedom and χ^2_p cutoff values. The MCD distances provided good tools for clustering. Also, the F cutoff value did an excellent job of correctly identifying good data in both the clean and contaminated data sets. The type II error of this method was minimal. The only disadvantage was that it was dependent on a robust initial estimate, which was provide by the programs due to Reiners and Woodruff (2000).

Since the robust MCD estimator worked so well, we repeated our analyses using a different robust estimator, the translated biweight S-estimator. The distances with S-estimate metrics also have an asymptotic χ^2_p distribution, and the χ^2_p cutoff values appear to work much better in the small sample simulations than they did in the MCD small sample simulations. The draw back to the S-estimation is that the type II error is somewhat large in small samples.

Some future projects have been discussed. Different robust estimators can be analyzed. Also, since the MCD is highly robust, we would like to improve the F cutoff to be more accurate in small samples. We have thought that in small samples the degrees of freedom associated with the F distribution might be more heavily dependent on the sample size or the dimension. If we can find this relationship or a model for this relationship, a more accurate formula for the degrees of freedom would give better cutoff values for identifying outliers.

It would also be worth investigating the relationship between the initial estimates and the final clustering and outlier detection. Also, a very interesting project would be to find some goodness-of-fit statistic that would help decide the correct number of clusters for a particular data set.

Chapter 6

Results of Simulations

6.1. MCD – 5% Cutoff (1 Cluster)

Percentiles from Chi-Square(p) cutoff values

		n			
		50	100	500	1000
p	5	19.77	13.50	7.12	6.07
	10	29.93	21.74	8.54	6.79
	20	26.75	32.54	12.25	8.46

Percentiles from Asymptotic cutoff values

		n			
		50	100	500	1000
p	5	0.14	1.41	4.45	4.76
	10	0.06	0.82	4.21	4.71
	20	0.01	0.36	3.61	4.43

Percentiles from Monte Carlo cutoff values

		n			
		50	100	500	1000
p	5	3.33	3.83	4.91	4.92
	10	1.89	3.20	4.83	4.94
	20	1.82	2.59	4.46	4.83

Table 1. Each entry represents the percent of simulated data that was above a specific 5% cutoff value. (Ideally, an entry in a cell would be 5.) The cutoff values were determined by dimension, size, and method of analysis. We can see that the Chi-Square cutoffs consistently reject too many points as outlying. The asymptotic method is quite conservative, but it appears to become quite accurate as n increases. The simulation method is very good for medium to large samples, and it has the best performance of the three for small samples. The analysis was done using the MCD estimates on one cluster of data.

6.2. MCD – 1% Cutoff (1 Cluster)

Percentiles from Chi-Square(p) cutoff values

		n			
		50	100	500	1000
p	5	10.21	5.46	1.83	1.44
	10	20.97	10.33	2.38	1.64
	20	23.63	24.77	3.95	2.26

Percentiles from Asymptotic cutoff values

		n			
		50	100	500	1000
p	5	0	0.10	0.79	0.93
	10	0	0.06	0.77	0.91
	20	0	0.03	0.60	0.84

Percentiles from Monte Carlo cutoff values

		n			
		50	100	500	1000
p	5	0.45	0.65	0.95	0.99
	10	0.26	0.5	0.96	0.98
	20	0.35	0.42	0.83	0.96

Table 2. Each entry represents the percent of simulated data that was above a specific 1% cutoff value. (Ideally, an entry in a cell would be 1.) The cutoff values were determined by dimension, size, and method of analysis. Again, we see the same results, the Chi-Square cutoffs consistently reject too many points as outlying. The asymptotic method is quite conservative, but it appears to become quite accurate as n increases. The simulation method is very good for medium to large samples, and it has the best performance of the three for small samples. The analysis was done using the MCD estimates on one cluster of data.

6.3. MCD – 0.1% Cutoff (1 Cluster)

Percentiles from Chi-Square(p) cutoff values

		n			
		50	100	500	1000
p	5	4.48	1.69	0.29	0.19
	10	12.10	3.86	0.41	0.22
	20	19.82	15.17	0.78	0.35

Percentiles from Asymptotic cutoff values

		n			
		50	100	500	1000
p	5	0	0	0.06	0.08
	10	0	0	0.07	0.09
	20	0	0	0.05	0.08

Percentiles from Monte Carlo cutoff values

		n			
		50	100	500	1000
p	5	0.02	0.04	0.09	0.10
	10	0.02	0.03	0.10	0.10
	20	0.01	0.04	0.08	0.10

Table 3. Each entry represents the percent of simulated data that was above a specific 0.1% cutoff value. (Ideally, an entry in a cell would be 0.1.) The cutoff values were determined by dimension, size, and method of analysis. Again, we see the same results, the Chi-Square cutoffs consistently reject too many points as outlying. The asymptotic method is quite conservative, but it appears to become quite accurate as n increases. The simulation method is very good for medium to large samples, and it has the best performance of the three for small samples. The analysis was done using the MCD estimates on one cluster of data.

6.4. MCD – Clean Data (2 & 3 Clusters)

		2 Clusters				3 Clusters	
		Bal.	Unbal.			Bal.	Unbal.
5% F				5% F			
	4	3.91	4.02		4	3.86	4.06
P	7	4.44	4.44	P	7	4.37	4.27
	10	4.32	4.33		10	4.29	4.31
5% Chisq				5% Chisq			
	4	38.11	37.99		4	37.83	38.01
P	7	34.72	34.74	P	7	34.75	34.71
	10	34.74	34.23		10	34.52	34.71
1% F				1% F			
	4	0.63	0.67		4	0.64	0.70
P	7	0.81	0.83	P	7	0.82	0.75
	10	0.80	0.82		10	0.81	0.77
1% Chisq				1% Chisq			
	4	21.66	21.96		4	21.57	21.73
P	7	17.94	17.98	P	7	17.88	17.76
	10	17.59	17.16		10	17.30	17.43
0.1% F				0.1% F			
	4	0.04	0.03		4	0.04	0.04
P	7	0.07	0.06	P	7	0.08	0.06
	10	0.07	0.09		10	0.07	0.07
0.1% Chisq				0.1% Chisq			
	4	9.82	9.87		4	9.73	9.85
P	7	6.88	6.97	P	7	6.79	6.67
	10	6.40	6.32		10	6.31	6.40

Table 4. Each entry represents the percent of simulated data that was misclassified according to a specified cutoff and percentage. The first column of tables reports for data that come from two populations and the second column reports for data that come from three populations. Balanced refers to data that consist of equal sized clusters; unbalanced refers to data that consist of unequal sized clusters. These data were generated as clusters of multivariate normal data with no contamination. The analysis was done using the MCD estimates.

6.5. MCD – Cluster Outliers (2 Clusters)

		Type I Error				Type II Error	
		Bal.	Unbal.			Bal.	Unbal.
5% F				5% F			
	4	2.87	2.91		4	0.00	0.00
P	7	3.36	3.71	P	7	0.00	0.00
	10	3.31	3.61		10	0.00	0.00
5	4	33.40	34.70		4	0.00	0.00
P	7	30.75	32.34	P	7	0.00	0.00
	10	30.42	32.01		10	0.00	0.00
1% F				1% F			
	4	0.42	0.42		4	0.17	0.00
P	7	0.57	0.61	P	7	0.00	0.00
	10	0.53	0.62		10	0.00	0.00
1% Chisq				1% Chisq			
	4	18.03	18.73		4	0.00	0.00
P	7	15.02	16.01	P	7	0.00	0.00
	10	14.39	15.49		10	0.00	0.00
0.1% F				0.1% F			
	4	0.02	0.03		4	2.29	0.66
P	7	0.04	0.06	P	7	0.06	0.00
	10	0.04	0.05		10	0.00	0.00
0.1% Chisq				0.1% Chisq			
	4	7.38	7.84		4	0.00	0.00
P	7	5.25	5.84	P	7	0.00	0.00
	10	4.80	5.42		10	0.00	0.00

Table 5. Each entry represents the percent of simulated data that was misclassified according to a specified cutoff and percentage. The first column of tables reports the type I error for the procedure, and the second column of tables reports the type II error. Balanced refers to data that consist of equal sized clusters; unbalanced refers to data that consist of unequal sized clusters. These data were generated as clusters of multivariate normal data with a cluster of contamination of size 20% of the smallest clean cluster. The analysis was done using the MCD estimates on two clusters of contaminated data.

6.6. MCD – Radial Outliers (2 Clusters)

		Type I Error				Type II Error	
		Bal.	Unbal.			Bal.	Unbal.
5% F				5% F			
	4	2.93	3.39		4	0.67	0.88
P	7	3.26	3.68	P	7	0.01	0.02
	10	3.20	3.61		10	0.00	0.00
5% Chisq				5% Chisq			
	4	33.72	35.00		4	0.07	0.08
P	7	30.75	32.26	P	7	0.00	0.00
	10	30.34	31.83		10	0.00	0.00
1% F				1% F			
	4	0.37	0.51		4	1.73	2.50
P	7	0.47	0.63	P	7	0.06	0.05
	10	0.48	0.61		10	0.00	0.02
1% Chisq				1% Chisq			
	4	18.02	19.20		4	0.12	0.23
P	7	14.72	15.85	P	7	0.00	0.00
	10	14.31	15.32		10	0.00	0.00
0.1% F				0.1% F			
	4	0.01	0.03		4	11.68	13.71
P	7	0.04	0.05	P	7	0.76	1.73
	10	0.03	0.04		10	0.09	0.61
0.1% Chisq				0.1% Chisq			
	4	7.49	8.33		4	0.40	0.61
P	7	5.02	5.61	P	7	0.00	0.02
	10	4.67	5.28		10	0.00	0.00

Table 6. Each entry represents the percent of simulated data that was misclassified according to a specified cutoff and percentage. The first column of tables reports the type I error for the procedure, and the second column of tables reports the type II error. Balanced refers to data that consist of equal sized clusters; unbalanced refers to data that consist of unequal sized clusters. These data were generated as clusters of multivariate normal data with radial outliers of size 20% of the smallest clean cluster. The analysis was done using the MCD estimates on two clusters of contaminated data.

6.7. MCD – Diffuse Outliers (2 Clusters)

		Type I Error				Type II Error	
		Bal.	Unbal.			Bal.	Unbal.
5% F				5% F			
	4	2.86	3.18		4	0.00	0.00
P	7	3.18	3.78	P	7	0.00	0.00
	10	3.20	3.58		10	0.00	0.00
5% Chisq				5% Chisq			
	4	33.26	34.90		4	0.00	0.00
P	7	30.36	32.49	P	7	0.00	0.00
	10	30.27	31.56		10	0.00	0.00
1% F				1% F			
	4	0.37	0.46		4	0.32	0.66
P	7	0.51	0.67	P	7	0.00	0.02
	10	0.49	0.61		10	0.00	0.00
1% Chisq				1% Chisq			
	4	17.59	18.93		4	0.00	0.00
P	7	14.53	16.16	P	7	0.00	0.00
	10	14.20	15.47		10	0.00	0.00
0.1% F				0.1% F			
	4	0.02	0.03		4	17.98	12.80
P	7	0.04	0.05	P	7	0.63	1.16
	10	0.04	0.05		10	0.11	0.32
0.1% Chisq				0.1% Chisq			
	4	7.27	8.23		4	0.00	0.00
P	7	4.97	5.91	P	7	0.00	0.00
	10	4.63	5.29		10	0.00	0.00

Table 7. Each entry represents the percent of simulated data that was misclassified according to a specified cutoff and percentage. The first column of tables reports the type I error for the procedure, and the second column of tables reports the type II error. Balanced refers to data that consist of equal sized clusters; unbalanced refers to data that consist of unequal sized clusters. These data were generated as clusters of multivariate normal data with diffuse outliers of size 20% of the smallest clean cluster. The analysis was done using the MCD estimates on two clusters of contaminated data.

6.8. S-estimation – Clean Data (1 Cluster – Pointwise)

		n			
		50	100	500	1000
5% F – simulated	4	4.59	4.47	4.89	4.93
P	7	3.04	4.18	4.66	4.83
	10	0.21	1.34	4.12	4.52
5% F – theoretical	4	4.99	4.52	4.88	4.94
P	7	0.79	2.67	4.13	4.54
	10	0.21	1.34	4.11	4.53
5% Chisq	4	7.42	6.11	5.26	5.12
P	7	3.26	5.62	5.25	5.13
	10	2.71	3.87	4.84	4.94
1% F – simulated	4	1.08	0.96	1.00	0.99
P	7	0.97	1.15	1.03	1.00
	10	0.21	1.02	1.04	1.01
1% F – theoretical	4	1.32	0.98	1.00	1.00
P	7	0.04	0.34	0.83	0.90
	10	0.21	1.00	1.04	1.01
1% Chisq	4	3.42	1.90	1.15	1.07
P	7	3.04	2.82	1.28	1.12
	10	0.21	1.34	1.37	1.18
0.1% F – simulated	4	0.10	0.10	0.10	0.10
P	7	0.09	0.14	0.10	0.10
	10	0.10	0.13	0.10	0.10
0.1% F – theoretical	4	0.14	0.11	0.09	0.10
P	7	0.00	0.02	0.07	0.08
	10	0.13	0.13	0.10	0.10
0.1% Chisq	4	0.95	0.29	0.12	0.11
P	7	1.35	0.58	0.14	0.12
	10	0.21	0.72	0.17	0.13

Table 8. Each entry represents the percent of simulated data that was above a cutoff value of a particular distribution at a specified significance level. The three distributional methods give similar results. The analysis was done using the S-estimates on one cluster of data.

6.9. S-estimation – Clean Data (1 Cluster – Datasetwise)

		n			
		50	100	500	1000
5% F – simulated	4	4.50	4.50	5.10	5.20
P	7	4.30	6.10	4.70	4.90
	10	5.20	5.50	6.30	5.60
5% F – theoretical	4	6.50	4.60	5.00	5.30
P	7	0.00	0.10	2.70	4.00
	10	6.30	5.40	6.30	5.60
5% Chisq	4	34.90	15.80	7.00	6.30
P	7	48.90	30.30	7.50	7.10
	10	9.60	35.50	10.50	8.80
1% F – simulated	4	1.00	0.80	1.10	1.00
P	7	1.00	1.20	0.90	1.10
	10	1.10	1.50	1.10	1.40
1% F – theoretical	4	1.40	1.00	1.10	1.00
P	7	0.00	0.00	0.20	1.00
	10	1.50	1.40	1.10	1.50
1% Chisq	4	17.90	5.80	1.90	1.60
P	7	24.50	13.40	2.20	1.50
	10	9.60	15.50	2.70	2.00
0.1% F – simulated	4	0.30	0.00	0.00	0.00
P	7	0.10	0.10	0.10	0.10
	10	0.20	0.20	0.20	0.30
0.1% F – theoretical	4	0.50	0.00	0.00	0.00
P	7	0.00	0.00	0.00	0.10
	10	0.20	0.20	0.20	0.30
0.1% Chisq	4	4.80	1.10	0.10	0.20
P	7	8.90	3.00	0.10	0.20
	10	9.60	4.10	0.30	0.30

Table 9. Each entry represents the percent of data sets that had a point above a cutoff value of a particular distribution at a specified datasetwise significance level. As with the MCD simulations, the χ^2_p cutoff appears to be too liberal, the F with asymptotic degrees of freedom is a bit too conservative, and the F with theoretical degrees of freedom is accurate. The analysis was done using S-estimates on one cluster of data.

6.10. S-Estimates – Cluster Outliers (1 cluster)

		Type I Error				Type II Error			
		n				n			
		50	100	500	1000	50	100	500	1000
5% F – simulated	4	0.98	1.10	1.22	1.26	0.04	0.01	0.00	0.00
P	7	0.41	0.79	1.08	1.10	0.01	0.00	0.00	0.00
	10	0.00	0.12	0.73	0.88	0.00	0.00	0.00	0.00
5% F – asymptotic	4	1.05	1.11	1.23	1.26	0.04	0.01	0.00	0.00
P	7	0.10	0.41	0.90	1.01	0.21	0.01	0.00	0.00
	10	0.00	0.12	0.73	0.88	0.00	0.00	0.00	0.00
5% Chisq	4	1.92	1.68	1.36	1.33	0.00	0.00	0.00	0.00
P	7	0.41	1.19	1.26	1.20	0.00	0.00	0.00	0.00
	10	0.07	0.39	0.92	0.99	0.00	0.00	0.00	0.00
1% F – simulated	4	0.11	0.12	0.14	0.14	0.55	0.16	0.05	0.05
P	7	0.12	0.11	0.14	0.13	0.15	0.04	0.00	0.00
	10	0.00	0.10	0.11	0.12	0.05	0.00	0.00	0.00
1% F – asymptotic	4	0.13	0.12	0.14	0.14	0.44	0.15	0.05	0.05
P	7	0.01	0.02	0.11	0.11	5.00	0.12	0.00	0.00
	10	0.00	0.11	0.11	0.12	0.04	0.00	0.00	0.00
1% Chisq	4	0.62	0.30	0.17	0.15	0.10	0.07	0.05	0.05
P	7	0.41	0.44	0.19	0.15	0.01	0.01	0.00	0.00
	10	0.00	0.12	0.16	0.14	0.00	0.00	0.00	0.00
.1% F – simulated	4	0.01	0.00	0.01	0.00	7.56	2.10	0.69	0.60
P	7	0.02	0.01	0.01	0.00	2.64	0.42	0.05	0.03
	10	0.00	0.00	0.00	0.01	1.04	0.08	0.00	0.00
.1% F – asymptotic	4	0.02	0.00	0.01	0.00	6.43	2.03	0.69	0.60
P	7	0.00	0.00	0.00	0.00	44.29	2.38	0.07	0.04
	10	0.00	0.00	0.00	0.01	0.94	0.08	0.00	0.00
.1% Chisq	4	0.08	0.03	0.01	0.01	0.70	0.70	0.54	0.52
P	7	0.17	0.04	0.01	0.01	0.10	0.06	0.03	0.03
	10	0.00	0.06	0.01	0.01	0.01	0.00	0.00	0.00

Table 10. Each entry represents the percent of simulated data that was misclassified according to some specified cutoff and percentage. The first column of tables reports the type I error, the second reports the type II error. The data were contaminated with a cluster of outliers which were 20% of the sample size. The analysis was done using S-estimates on one cluster of data.

6.11. S-Estimates – Radial Outliers (1 cluster)

		Type I Error				Type II Error			
		n				n			
		50	100	500	1000	50	100	500	1000
5% F – simulated	4	1.07	1.09	1.22	1.24	0.13	0.00	0.00	0.00
P	7	0.44	0.78	1.06	1.14	0.05	0.00	0.00	0.00
	10	0.00	0.14	0.75	0.87	0.13	0.00	0.00	0.00
5% F – asymptotic	4	1.13	1.10	1.22	1.24	0.10	0.00	0.00	0.00
P	7	0.08	0.40	0.89	1.05	5.98	0.02	0.00	0.00
	10	0.00	0.14	0.75	0.87	0.12	0.00	0.00	0.00
5% Chisq	4	2.06	1.67	1.36	1.31	0.00	0.00	0.00	0.00
P	7	0.45	1.20	1.24	1.24	0.04	0.00	0.00	0.00
	10	0.08	0.39	0.94	0.98	0.03	0.00	0.00	0.00
1% F – simulated	4	0.12	0.13	0.14	0.14	7.39	0.96	0.00	0.00
P	7	0.11	0.12	0.14	0.14	4.57	0.36	0.00	0.00
	10	0.00	0.11	0.12	0.11	3.92	0.17	0.00	0.00
1% F – asymptotic	4	0.13	0.13	0.14	0.14	6.38	0.92	0.00	0.00
P	7	0.00	0.03	0.10	0.12	45.23	3.60	0.00	0.00
	10	0.00	0.11	0.12	0.11	3.58	0.17	0.00	0.00
1% Chisq	4	0.68	0.32	0.17	0.15	0.49	0.12	0.00	0.00
P	7	0.44	0.43	0.18	0.16	0.05	0.02	0.00	0.00
	10	0.00	0.14	0.17	0.14	0.03	0.00	0.00	0.00
.1% F – simulated	4	0.01	0.00	0.01	0.01	43.38	24.31	8.17	5.92
P	7	0.00	0.01	0.00	0.01	32.63	13.35	0.55	0.05
	10	0.00	0.01	0.01	0.01	26.19	9.25	0.05	0.00
.1% F – asymptotic	4	0.01	0.00	0.01	0.01	40.03	23.84	8.14	5.88
P	7	0.00	0.00	0.00	0.01	87.86	40.33	1.69	0.17
	10	0.00	0.01	0.01	0.01	25.08	9.02	0.05	0.00
.1% Chisq	4	0.08	0.03	0.01	0.01	10.01	8.32	4.88	4.15
P	7	0.18	0.05	0.01	0.01	2.79	1.51	0.14	0.01
	10	0.00	0.07	0.01	0.01	0.98	0.45	0.01	0.00

Table 11. Each entry represents the percent of simulated data that was misclassified according to some specified cutoff and percentage. The first column of tables reports the type I error, the second reports the type II error. The data were contaminated with radial outliers which were 20% of the sample size. The analysis was done using S-estimates on one cluster of data.

6.12. S-estimation – Clean Data (2 & 3 Clusters)

		2 Clusters			3 Clusters		
		Small	Bal.	Unbal.	Small	Bal.	Unbal.
5% F							
	4	5.16	4.93	4.85	4.71	4.82	4.77
P	7	4.15	4.65	4.78	4.36	4.67	4.76
	10	3.21	4.11	4.23	3.32	4.03	4.08
5% Chisq							
	4	7.38	5.50	5.47	7.13	5.45	5.34
P	7	5.14	5.17	5.32	5.34	5.22	5.31
	10	3.90	4.84	4.94	4.38	4.78	4.83
1% F							
	4	1.26	0.98	1.11	0.97	1.04	0.96
P	7	1.22	1.00	1.08	1.12	0.99	0.98
	10	1.19	1.03	1.02	1.55	1.00	0.98
1% Chisq							
	4	3.61	1.22	1.34	3.07	1.30	1.19
P	7	3.19	1.21	1.30	3.30	1.21	1.21
	10	3.18	1.37	1.39	3.25	1.32	1.33
0.1% F							
	4	0.23	0.11	0.10	0.07	0.10	0.09
P	7	0.18	0.11	0.11	0.12	0.11	0.10
	10	0.21	0.11	0.09	0.29	0.10	0.11
0.1% Chisq							
	4	0.91	0.16	0.17	0.66	0.16	0.14
P	7	0.70	0.15	0.16	0.68	0.15	0.16
	10	0.80	0.16	0.16	1.02	0.15	0.16

Table 12. Each entry represents the percent of simulated data that was misclassified according to a specified cutoff and percentage. The first column of tables reports for data that come from two populations and the second column reports for data that come from three populations. Small refers to the smallest combination of cluster sizes; balanced refers to data that consist of equal sized clusters; unbalanced refers to data that consist of unequal sized clusters. These data were generated as clusters of multivariate normal data with no contamination. The analysis was done using the S-estimates.

6.13. S-estimation – Cluster Outliers (2 Clusters)

		Type I Error			Type II Error		
		Small	Bal.	Unbal.	Small	Bal.	Unbal.
5% F							
	4	3.08	3.03	3.17	0.00	0.02	0.00
P	7	2.64	3.27	3.27	0.00	0.00	0.00
	10	1.87	2.46	2.81	45.89	0.00	0.00
5% Chisq							
	4	4.59	3.40	3.63	0.00	0.02	0.00
P	7	3.15	3.70	3.707	0.00	0.00	0.00
	10	2.99	2.92	3.37	44.46	0.00	0.00
1% F							
	4	0.70	0.57	0.55	0.30	0.12	0.00
P	7	0.72	0.65	0.65	0.06	0.00	0.00
	10	0.99	0.57	0.61	47.86	0.00	0.00
1% Chisq							
	4	2.03	0.71	0.72	0.10	0.08	0.00
P	7	2.10	0.83	0.83	0.00	0.00	0.00
	10	1.74	0.76	0.86	46.49	0.00	0.00
0.1% F							
	4	0.10	0.04	0.06	3.80	0.70	0.25
P	7	0.11	0.05	0.05	0.12	0.02	0.02
	10	0.12	0.05	0.06	49.64	0.00	0.00
0.1% Chisq							
	4	0.57	0.07	0.09	0.50	0.53	0.20
P	7	0.47	0.07	0.07	0.06	0.02	0.02
	10	0.54	0.08	0.11	48.93	0.00	0.00

Table 13. Each entry represents the percent of simulated data that was misclassified according to a specified cutoff and percentage. The first column of tables reports the type I error for the procedure, and the second column of tables reports the type II error. Small refers to the smallest combination of cluster sizes; balanced refers to data that consist of equal sized clusters; unbalanced refers to data that consist of unequal sized clusters. These data were generated as clusters of multivariate normal data with a cluster of outliers of size 20% of the smallest clean cluster. The analysis was done using the S-estimates on two clusters of contaminated data.

6.14. S-estimation – Radial Outliers (2 Clusters)

		Type I Error			Type II Error		
		Small	Bal.	Unbal.	Small	Bal.	Unbal.
5% F							
	4	2.59	2.56	3.19	0.90	0.48	0.62
P	7	2.21	2.47	3.28	0.06	0.02	0.02
	10	0.97	1.94	2.71	0.00	0.00	0.00
5% Chisq							
	4	3.96	2.94	3.60	0.80	0.45	0.60
P	7	2.58	2.78	3.66	0.06	0.02	0.02
	10	1.93	2.31	3.20	0.00	0.00	0.00
1% F							
	4	0.56	0.39	0.54	2.30	1.25	1.65
P	7	0.50	0.41	0.62	0.12	0.06	0.08
	10	0.46	0.38	0.59	0.00	0.01	0.00
1% Chisq							
	4	1.64	0.51	0.70	1.20	1.20	1.53
P	7	1.57	0.51	0.76	0.12	0.06	0.08
	10	0.97	0.51	0.78	0.00	0.01	0.00
0.1% F							
	4	0.07	0.03	0.03	15.30	3.03	3.57
P	7	0.04	0.02	0.06	2.50	0.31	0.27
	10	0.04	0.03	0.05	1.31	0.05	0.03
0.1% Chisq							
	4	0.34	0.05	0.06	3.60	2.50	3.13
P	7	0.24	0.04	0.08	0.38	0.26	0.27
	10	0.30	0.05	0.09	0.00	0.02	0.00

Table 14. Each entry represents the percent of simulated data that was misclassified according to a specified cutoff and percentage. The first column of tables reports the type I error for the procedure, and the second column of tables reports the type II error. Small refers to the smallest combination of cluster sizes; balanced refers to data that consist of equal sized clusters; unbalanced refers to data that consist of unequal sized clusters. These data were generated as clusters of multivariate normal data with radial outliers of size 20% of the smallest clean cluster. The analysis was done using the S-estimates on two clusters of contaminated data.

6.15. S-estimation – Diffuse Outliers (2 Clusters)

		Type I Error			Type II Error		
		Small	Bal.	Unbal.	Small	Bal.	Unbal.
5% F							
	4	2.43	2.58	3.35	0.00	0.00	0.00
P	7	2.11	2.47	3.29	0.94	0.00	0.00
	10	1.49	1.95	2.75	21.74	0.00	0.00
5% Chisq							
	4	3.87	2.95	3.84	0.00	0.00	0.00
P	7	2.67	2.77	3.70	0.94	0.00	0.00
	10	2.43	2.34	3.27	21.11	0.00	0.00
1% F							
	4	0.61	0.41	0.59	2.00	0.00	0.00
P	7	0.47	0.43	0.64	1.06	0.00	0.00
	10	0.77	0.39	0.59	22.58	0.00	0.00
1% Chisq							
	4	1.47	0.53	0.77	0.40	0.00	0.00
P	7	1.53	0.52	0.78	0.94	0.00	0.00
	10	1.36	0.52	0.82	21.84	0.00	0.00
0.1% F							
	4	0.04	0.04	0.04	22.70	2.43	0.33
P	7	0.01	0.03	0.06	8.69	0.01	0.00
	10	0.13	0.02	0.06	26.37	0.00	0.00
0.1% Chisq							
	4	0.37	0.06	0.08	4.70	1.02	0.00
P	7	0.24	0.04	0.08	1.81	0.00	0.00
	10	0.48	0.04	0.10	23.47	0.00	0.00

Table 15. Each entry represents the percent of simulated data that was misclassified according to a specified cutoff and percentage. The first column of tables reports the type I error for the procedure, and the second column of tables reports the type II error. Small refers to the smallest combination of cluster sizes; balanced refers to data that consist of equal sized clusters; unbalanced refers to data that consist of unequal sized clusters. These data were generated as clusters of multivariate normal data with diffuse outliers of size 20% of the smallest clean cluster. The analysis was done using the S-estimates on two clusters of contaminated data.

Chapter 7

Appendix: Theoretical Formulas for Wishart Degrees of Freedom

A. MCD Formula

The parameter m, the degrees of freedom for the estimated Wishart distribution, is calculated using a series of steps. This estimation is due to Croux and Haesbroeck (1999).

$$\begin{aligned}
\alpha &= \frac{n-h}{n} && \text{(A.1)}\\
&\quad \text{where } n \text{ is the sample size and } h = \left\lfloor \frac{(n+p+1)}{2} \right\rfloor . \\
q_\alpha &\text{ is } \text{such that: } 1-\alpha = P(\chi^2_p \le q_\alpha) && \text{(A.2)}\\
c_\alpha &= \frac{1-\alpha}{P(\chi^2_{p+2} \le q_\alpha)} && \text{(A.3)}\\
c_2 &= \frac{-P(\chi^2_{p+2} \le q_\alpha)}{2} && \text{(A.4)}\\
c_3 &= \frac{-P(\chi^2_{p+4} \le q_\alpha)}{2} && \text{(A.5)}\\
c_4 &= 3 \cdot c_3 && \text{(A.6)}\\
b_1 &= \frac{c_\alpha(c_3-c_4)}{1-\alpha} && \text{(A.7)}\\
b_2 &= 0.5 + \frac{c_\alpha}{(1-\alpha)}\left(c_3 - \frac{q_\alpha}{p}\left(c_2 + \frac{(1-\alpha)}{2}\right)\right) && \text{(A.8)}\\
v_1 &= (1-\alpha)b_1^2(\alpha(\frac{c_\alpha q_\alpha}{p}-1)^2-1) - 2c_3c_\alpha^2(3(b_1-pb_2)^2 && \text{(A.9)}\\
&\quad + (p+2)b_2(2b_1-pb_2)) \\
v_2 &= n(b_1(b_1-pb_2)(1-\alpha))^2c_\alpha^2 && \text{(A.10)}\\
v &= \frac{v_1}{v_2} && \text{(A.11)}\\
\hat{m} &= \frac{2}{c_\alpha^2 v} && \text{(A.12)}
\end{aligned}$$

B. S-Estimate Formula

In this appendix we provide for completeness the useable forms of the formulas due to Lopuaä (1989) needed to estimate the degrees of freedom parameter m of the Wishart approximation.

$$r = \frac{(n-p)}{2n} \text{ where } n \text{ is the sample size and } p \text{ is the dimension} \tag{B.1}$$

$$a_0 = \frac{a^2}{2} + \frac{b(5b+16a)}{30} \tag{B.2}$$

$$b_0 = E_{0,I}(\rho(||X||)) \text{ where } \rho \text{ is defined as in the t-biweight}(a,b) \tag{B.3}$$

$$c_{1.1} = \int_0^{a^2} e^{-x/2}x^{(p-1)/2}\sqrt{x}\ \mathrm{d}x \tag{B.4}$$

$$c_{1.2} = \int_{a^2}^{(a+b)^2} e^{-x/2}x^{(p-1)/2}\Big\{\sqrt{x}\Big(1-\Big(\frac{(\sqrt{x}-a)^2}{b}\Big)^2\Big)\Big\}\ \mathrm{d}x \tag{B.5}$$

$$c_1 = \frac{2^{-p/2}}{\Gamma(\frac{p}{2})}(c_{1.1}+c_{1.2}) \tag{B.6}$$

$$c_{2.1} = \int_0^{a^2} e^{-x/2}x^{p/2}\ \mathrm{d}x \tag{B.7}$$

$$\begin{aligned} c_{2.2} = & \int_{a^2}^{(a+b)^2} e^{-x/2}x^{p/2}\Big\{\Big(1-\Big(\frac{(\sqrt{x}-a)^2}{b}\Big)^2\Big) \\ & -4\sqrt{x}(\sqrt{x}-a)\frac{1-\Big(\frac{(\sqrt{x}-a)}{b}\Big)^2}{b^2}\Big\}\ \mathrm{d}x \end{aligned} \tag{B.8}$$

$$c_2 = \frac{2^{-p/2}}{\Gamma(\frac{p}{2})}(c_{2.1}+c_{2.2}) \tag{B.9}$$

$$c_{3.1} = \int_0^{a^2} e^{-x/2}x^{p/2}x\ \mathrm{d}x \tag{B.10}$$

$$c_{3.2} = \int_0^{a^2} e^{-x/2}x^{p/2}\Big\{\sqrt{x}\Big(1-\Big(\frac{(\sqrt{x}-a)^2}{b}\Big)^2\Big)\Big\}^2\ \mathrm{d}x \tag{B.11}$$

$$c_3 = \frac{2^{-p/2}}{\Gamma(\frac{p}{2})}(c_{3.1}+c_{3.2}) \tag{B.12}$$

$$c_{4.0} = \int_0^\infty e^{-x/2} x^{(p-2)/2} \ \mathrm{d}x \tag{B.13}$$

$$c_{4.1} = \int_0^{a^2} e^{-x/2} x^{(p-2)/2} \frac{x}{2} \ \mathrm{d}x \tag{B.14}$$

$$\begin{aligned} c_{4.2} = & \int_{a^2}^{(a+b)^2} e^{-x/2} x^{(p-2)/2} \left\{ \frac{5a^4b^2 - a^6}{30b^4} + x\left(\frac{1}{2} + \frac{a^4}{2b^4} - \frac{a^2}{b^2}\right) + \right. \\ & \left. x^{3/2}\left(\frac{4a}{3b^2} - \frac{4a^3}{3b^4}\right) + x^2\left(\frac{3a^2}{2b^4} - \frac{1}{2b^2}\right) - x^{5/2}\left(\frac{4a}{5b^4}\right) + \frac{x^3}{6b^4} \right\} \ \mathrm{d}x \end{aligned} \tag{B.15}$$

$$c_{4.3} = \int_{(a+b)^2}^\infty e^{-x/2} x^{(p-2)/2} \left\{ \frac{a^2}{2} + \frac{5b^2 + 16ab}{30} \right\} \ \mathrm{d}x \tag{B.16}$$

$$c_{4.4} = \int_0^{a^2} e^{-x/2} x^{(p-2)/2} \frac{x^2}{4} \ \mathrm{d}x \tag{B.17}$$

$$\begin{aligned} c_{4.5} = & \int_{a^2}^{(a+b)^2} e^{-x/2} x^{(p-2)/2} \left\{ \frac{5a^4b^2 - a^6}{30b^4} + x\left(\frac{1}{2} + \frac{a^4}{2b^4} - \frac{a^2}{b^2}\right) + \right. \\ & \left. x^{3/2}\left(\frac{4a}{3b^2} - \frac{4a^3}{3b^4}\right) + x^2\left(\frac{3a^2}{2b^4} - \frac{1}{2b^2}\right) - x^{5/2}\left(\frac{4a}{5b^4}\right) + \frac{x^3}{6b^4} \right\}^2 \ \mathrm{d}x \end{aligned} \tag{B.18}$$

$$c_{4.6} = \int_{(a+b)^2}^\infty e^{-x/2} x^{(p-2)/2} \left\{ \frac{a^2}{2} + \frac{5b^2 + 16ab}{30} \right\}^2 \ \mathrm{d}x \tag{B.19}$$

$$c_4 = \frac{2^{-p/2}}{\Gamma(\frac{p}{2})} \{ b_0^2 c_{4.0} - 2b_0(c_{4.1} + c_{4.2} + c_{4.3}) + (c_{4.4} + c_{4.5} + c_{4.6}) \} \tag{B.20}$$

$$\sigma_1 = \frac{p(p+2)c_3}{(c_2 + (p+1)c_1)^2} \tag{B.21}$$

$$\sigma_2 = \frac{-2\sigma_1}{p} + \frac{4c_4}{c_1^2} \tag{B.22}$$

$$\mathrm{Var}[\tilde{s_{ii}}] = \frac{2\sigma_1 + \sigma_2}{n} \tag{B.23}$$

$$\tilde{m} = \frac{2}{\mathrm{Var}[\tilde{s_{ii}}]} \tag{B.24}$$

Chapter 8

Bibliography

Bibliography

Andrews, D., Bickel, P., Hampel, F., Huber, P., Rogers, W., and Tukey, J. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton University Press, Princeton, NJ.

Atkinson, A. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, 89:1329–1339.

Banfield, J. and Raftery, A. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data.* John Wiley.

Butler, R., Davies, P., and Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, 21:1385–1400.

Campell, N. (1980). Robust procedures in multivariate analysis I: Robust canonical variate analysis. *Applied Statistics*, 29:1–8.

Campell, N. (1982). Robust procedures in multivariate analysis II: Robust canonical variate analysis. *Applied Statistics*, 31:1–8.

Coleman, D., Dong, X., Hardin, J., Rocke, D., and Woodruff, D. (1999). Some computational issues in cluster analysis with no a priori metric. *Computational Statistics and Data Analysis*, 31:1–11.

Croux, C. and Haesbroeck, G. (2000). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis.*

Daudin, J., Duby, C., and Trecourt, P. (1988). Stability of principal component analysis studied by the bootstrap method. *Statistics*, 19:241–258.

Davies, P. (1987). Asymptotic behavior of s-estimators of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15:1269–1292.

Devlin, S., Gnanadesikan, R., and Kettenring, J. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76:354–362.

Donoho, D. (1982). *Breakdown Properties of Multivariate Location Estimators.* PhD thesis, Harvard University, Department of Statistics.

Everitt, B. S. (1993). *Cluster Analysis.* John Wiley.

Friedman, H. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62:1159–1178.

Gnanadesikan, R. and Kettenring, J. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28:81–124.

Grubbs, F. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11:1–21.

Grübel, R. and Rocke, D. (1990). On the cumulants of affine equivariant estimators in elliptical families. *Journal of Multivariate Analysis*, 35:203–222.

Hadi, A. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society B*, 54:761–771.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Functions.* John Wiley.

Hawkins, D. (1980). *Identification of Outliers.* Chapman and Hall.

Hawkins, D. (1994). The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Computational Statistics and Data Analysis*, 17:197–210.

Hawkins, D. (1999). Improved feasible solution algorithms for high breakdown estimation. *Computational Statistics and Data Analysis*, 30:1–11.

Hawkins, D., Bradu, D., and Kass, G. (1984). Location of several outliers in multiple-regression data using elemental sets. *Technometrics*, 26:197–208.

Healy, M. (1968). Multivariate normal plotting. *Applied Statistics*, 17:157–161.

Hoaglin, D., Mosteller, F., and Tukey, J. (1983). *Understanding Robust and Exploratory Data Analysis.* John Wiley.

Huber, P. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, pages 73–101.

Huber, P. (1981). *Robust Statistics.* John Wiley.

Kauffman, L. and Rousseeuw, P. (1990). *Finding Groups in Data.* John Wiley.

Kent, J. and Tyler, D. (1991). Redescending m-estimates of multivariate location and scatter. *The Annals of Statistics*, 19:2102–2119.

Lopuhaä, H. (1989). On the relation between s-estimators and m-estimators of multivariate location and covariance. *The Annals of Statistics*, 17:1662–1683.

Lopuhaä, H. (1992). Highly efficient estimators of multivariate location with high breakdown point. *The Annals of Statistics*, 20:398–413.

Lopuhaä, H. and Rousseeuw, P. (1991). Breakdown of points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19:229–248.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis.* Academic Press.

Maronna, R. (1976). Robust m-estimators of multivariate location and scatter. *The Annals of Statistics*, 4:51–67.

Maronna, R. and Yohai, V. (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90:330–341.

Pearson, E. and Chandra Sekar, C. (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, 28:308–320.

Penny, K. (1995). Appropriate critical values when testing for a single multivariate outlier by using the mahalanobis distance. *Applied Statistics*, 45:73–81.

Reiners, T. and Woodruff, D. (2000). A blackboard architecture applied to maximum likelihood clustering. Working Paper. Graduate School of Management, UC Davis, Davis, CA 95616.

Rocke, D. (1992). Estimation of variation after outlier rejection. *Computational Statistics and Data Analysis*, 13:9–20.

Rocke, D. (1996). Robustness properties of s-estimators of multivariate location and shape in high dimension. *The Annals of Statistics*, 24:1327–1345.

Rocke, D. (1998). Constructive statistics: Estimators, algorithms, and asymptotics. In Weisberg, S., editor, *Computing Science and Statistics*, volume 30, pages 3–14.

Rocke, D. and Woodruff, D. (1993). Computation of robust estimates of multivariate location and shape. *Statistica Neerlandica*, 47:27–42.

Rocke, D. and Woodruff, D. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91:1047–1061.

Rocke, D. and Woodruff, D. (1997). Robust estimation of multivariate location and shape. *Journal of Statistical Planning and Inference*, 57:245–255.

Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In Grossmann, W., Pflug, G., Vincze, I., and Werz, W., editors, *Mathematical Statistics and Applications, Volume B*. Dordrecht:Reidel.

Rousseeuw, P. and Leroy, A. (1987). *Robust Regression and Outlier Detection.* John Wiley.

Rousseeuw, P. and VanDriessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, pages 212–223.

Rousseeuw, P. and VanZomeren, B. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85:633–639.

Rousseeuw, P. and VanZomeren, B. (1991). Robust distances: Simulations and cutoff values. In Stahel, W. and Weisberg, S., editors, *Directions in Robust Statistics and Diagnostics Part 2*, pages 195–203.

Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of s-estimators. In *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics*, volume 26, pages 256– 272. Springer, New York.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2:110–114.

Scott, A. and Symons, M. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, 27:387–397.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley.

Stahel, W. (1981). *Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen.* PhD thesis, ETH Zurich.

Tukey, J. (1960). A survey of sampling from contaminated distributions. In Olkin, I., Ghurye, S., Hoeffding, W., Madow, W., and Mann, H., editors, *Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling.* Stanford University Press, Stanford, CA.

Tukey, J. (1972). Data analysis, computation, and mathematics. *Quarterly of Applied Mathematics*, 30:51–65.

Tyler, D. (1983). Robust and efficiency properties of scatter matrices. *Biometirka*, 70:411–420.

Tyler, D. (1988). Some results on the existence, uniqueness, and computation of the m-estimates of multivariate location and scatter. *SIAM Journal on Scientific and Statistical Computing*, 9:354–362.

Tyler, D. (1991). Some issues in the robust estimation of multivariate location and scatter. In *Directions in Robust Statistics and Diagnositcs Part III*. Springer-Verlag.

Ward, J. (1963). Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244.

Welch, B. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29:350–362.

Welch, B. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34:28–35.

Wilks, S. (1962). *Mathematical Statistics*. John Wiley.

Wilks, S. (1963). Multivariate statistical outliers. *Sankhyã, A*, 25:407–426.