Chapter 3
Multivariate Location and Scale Estimation

Introduction

We have already discussed the classical OLS regression procedure as well as the low breakdown M and BI robust regression procedures. In order to make the transition to current high breakdown regression techniques, this chapter offers a literature search of various multivariate location and scale estimation procedures. The estimator for scale in the multivariate setting is often referred to as either a dispersion matrix or, more commonly, a covariance matrix.

The situation is such that the data has $n$ observations covering

i. the $k$ regressor variables or

ii. the $k$ regressor variables plus the response variable.

Estimation is then in either $k$ dimensions (case i.) or $p$ dimensions (case ii.). Note that in either case, an intercept variable (column of ones) is not incorporated. In the following discussions, it will be assumed that the estimation is in $p$ dimensions. The data is contained in the $n \times p$ $Z_r$ matrix. Removing the response variable from any desired analysis is an easy modification, just replace any references to $p$ with $k$ and use the $Z$ matrix instead.

It is desired to obtain a multivariate location estimator, denoted by $\mathbf{m}$, as well as a multivariate dispersion estimator, given as $\mathbf{C}$. The relevance of this analysis to the field of robust regression lies in its extraction of outlier and/or high leverage information. This translates into another, hopefully more resistant, method of assigning observation weights.
Finally, any mention of “half” the data will correspond to $h = \left\lfloor (n + p + 1)/2 \right\rfloor$ observations, due to optimal breakdown point considerations (Rousseeuw and Leroy (1987)). Where applicable, these methods are described in step-by-step detail so that the reader can more easily distinguish the differences between the competing methods. This will also allow the reader to fully appreciate the computations that are involved in each of the procedures.

§3.1 Classical Estimation

Assuming that the data is drawn from a population whose distribution is multivariate normal, then the optimal estimators for location and dispersion are found, respectively, as the $p\times 1$ sample mean vector,

$$m = \frac{1}{n} \sum_{i=1}^{n} z_{y,i},$$

and $p\times p$ sample covariance matrix,

$$C = \frac{1}{n-1} \sum_{i=1}^{n} (z_{y,i} - m)(z_{y,i} - m)^\prime.$$

These are, obviously, mean-based estimators, so any unusual or extreme observation can arbitrarily inflate either of them. As mentioned earlier, this fact has severe consequences, namely that diagnostics based on these statistics, i.e. the hat diagonals for example, are potentially misleading and generally very unreliable when the data are not well behaved. In terms of high breakdown regression analysis, it becomes necessary to consider alternative estimators. It should be noted, however, that the sample mean vector is the only affine equivariant estimator that can be calculated by coordinatewise location estimators (Donoho (1982), Rousseeuw and Leroy (1987)). Affine equivariance can be interpreted as the situation where any linear translation of the data is paralleled by a similar translation of the estimator. Affine equivariance of regression estimators will be discussed in greater detail in Chapter 6.
§3.2 Outlier Resistant Methods

There are numerous alternative estimators available to replace the classical sample mean vector (see Lopuhaa (1992)) and covariance matrix estimators. Some are more computationally extensive than others, and differences with regard to various theoretical properties may be exhibited as well. The following is a list of the major outlier resistant estimators for multivariate location and dispersion currently mentioned in the literature.

§3.2.1 Coordinatewise Median

Perhaps the simplest way to create a resistant multivariate location estimator is to address each coordinate individually. Following the lead of univariate location estimation, replace the sample mean by the more resistant median for each of the $p$ variables. The dispersion matrix estimator becomes a covariance calculation that is centered by this coordinatewise median. This approach can be extended to other resistant univariate location statistics, such as the M estimator, for example. Convenience seems to have dictated the selection of the median in the past.

An inherent problem with the coordinatewise median is that in the multivariate setting, this location estimator does not necessarily lie within the general data cloud (Rousseeuw and Leroy (1987, pg. 250) state that “it does not have to lie in the convex hull of the sample when $p \geq 3$”). In addition, this estimator is not affine equivariant, meaning that linear translations of the data are not paralleled by a similar translation of the estimator.

§3.2.2 Stahel-Donoho Estimator

A projection-based estimation procedure was developed independently by Stahel (1981) and Donoho (1982), and is mentioned in Rousseeuw and Leroy (1987, pg. 256-257). Simplistically, the idea here is that an outlier or high leverage point will separate out and away from the bulk of the data when viewed from the right perspective. There are two stages to the formation of the robust multivariate location and dispersion estimators. First, robust distances are determined via a projection computation. These distances become the arguments in a weight
function that is used to calculate a weighted mean vector and weighted covariance matrix. This procedure is affine equivariant, and attains a 50% (asymptotic) breakdown point when \( n > 2p + 1 \) and the data are in \textit{general position} (Donoho (1982), Rousseeuw and Leroy (1987)), which means that no more than \( p + 1 \) points lie in any \( p \) dimensional affine subspace. The algorithm itself is described in more detail in Appendix B.1.

While the definition of the Stahel-Donoho estimator requires the supremum over all possible directional vectors, Rousseeuw and van Zomeren (1990) propose a shortcut method which uses just \( n \) directional vectors, one vector in the direction of each centered observation (centered by the coordinatewise median, vector starting from the origin). The projections of the original data on these \( n \) directional vectors produce the robust distances. This algorithm is very computationally inexpensive.

The Stahel-Donoho estimator proceeds backwards from many methods used for outlier detection purposes. Usually, the calculation of robust distances is performed after the robust location and dispersion estimators are determined. Here, this order is reversed.

As a word of caution, it is noted that Cook and Hawkins (1990) suggest that this projection method can typically produce “outliers everywhere.” The meaning here is that the location may become exaggerated for some good observations not in the central data region, so not every observation deemed as extreme is necessarily a detriment to the analysis.

\section*{§3.2.3 Transformed One-step Weighted Dispersion Estimator}

A computationally easy method for computing a robust dispersion estimator was given by Ruiz-Gazen (1996). In this procedure a pre-defined robust location estimator is used to create an intermediate covariance matrix. A kernel function is used to obtain observation weights and a one-step covariance matrix is formed. Finally, a transformation is performed to attain a consistent estimator. It is suggested that the breakdown point for this procedure is roughly 20%, and it
appears as though the choice for the tuning parameter in the kernel function has a rather large impact on whether extreme observations are detected. This procedure is not investigated further, but details of its calculation are found in Appendix B.2.

§3.2.4 Minimum Volume Ellipsoid Estimator

One could consider multivariate location and dispersion estimation in high breakdown ideology as representing the most compact set of half the data. Specifically, determine the ellipsoid with the smallest volume that covers \( h \) observations. From this ellipsoid, the estimates of the \( p \times 1 \) location vector, \( \mathbf{m} \) (or \( \text{MVE}_1 \)), and the \( p \times p \) covariance matrix, \( \mathbf{C} \) (or \( \text{MVE}_2 \)), can be obtained, these estimates being called minimum volume ellipsoid (MVE) estimators. It should be noted that the exact MVE estimators for a given set of data is not necessarily the sample mean vector and associated covariance matrix for one particular subset of observations (Hawkins (1993)). The MVE estimators provide, when used as the basis for a set of robust distances, a robust yardstick for determining what constitutes an outlier and/or a high leverage point. The problem is that there is no closed-form solution for obtaining the MVE estimators.

The first algorithm was offered by Rousseeuw and Leroy (1987), in which a large number of elemental subsets (whose size equals \( p + 1 \), which represents one more observation than the dimension of the estimation problem, the minimum required to compute a full rank covariance matrix) are randomly drawn. Selecting the elemental subset size as \( p + 1 \) does have another theoretical rationalization in that it is guaranteed that at least \( p + 1 \) observations will fall exactly on the boundary of the ellipsoid defined by the MVE. Anyhow, a mechanism inflates (or deflates) each subsequent, randomly drawn, ellipsoid to cover half the data, with the objective function being evaluated and compared among all subsets analyzed. A probabilistic argument guarantees the selection of at least one subset that contains only good observations with “high probability.” However, a particular elemental subset of good observations may not represent the data very well, so obtaining a decent elemental subset may not occur. This tends to underestimate the number of required subsets determined by the probabilistic argument and, for decent results a much larger
number of subsets should be analyzed. Also, since the exact MVE does not have to be defined by the sample mean vector and sample covariance matrix of some subset of the data, the results obtained from this random subsampling algorithm are generally very approximate. Details for this algorithm can be found in Appendix B.4.

With no closed-form solution, MVE estimation is extremely computationally intensive. However, an exact iterative computation algorithm for the MVE does exist, having been developed by Cook, Hawkins, and Weisberg (1993). This algorithm requires the consideration of all $C_h^n$ subsets, or halfsets as each contains $h$ observations, for which the covariance matrix and its determinant are computed. During the search, certain subsets lead to an iteration routine that allows the overall algorithm to converge to the exact solution for the MVE objective function. The combinatoric nature of this approach leads to an enormous amount of computation time, even in moderate sample sizes.

To avoid the exhaustive evaluation of every possible halfset, the feasible solution algorithm (FSA) of Hawkins (1993) was developed. This procedure uses $N$ random starting halfsets, and iterates each to a local minima via a swapping technique. Then the MVE is selected from among the several potential candidates (the local minima), each one being referred to as a feasible solution. Details of the FSA algorithm for MVE estimation is found in Appendix B.3.

It seems as though the resampling algorithm is the most widely used for obtaining MVE estimators, even though the results may generally not be as good as with the other procedures. One reason is in the parallel of this particular algorithm to that of certain high breakdown regression estimators mentioned later in Chapter 4. Thus, one set of random searches can be used for two different purposes.

§3.2.5 Minimum Covariance Determinant Estimator

Another viewpoint on multivariate location and dispersion estimation in high breakdown ideology is based on the determinant of the covariance matrix. Since the covariance matrix is an
\( n \times n \) symmetric positive definite matrix, all \( p \) eigenvalues are positive (Morrison (1990), Johnson and Wichern (1992)). Furthermore, the determinant of a covariance matrix equals the product of the \( p \) eigenvalues. Near linear dependencies among the set of \( p \) variables produce near-zero eigenvalues. Thus, a small value in the determinant reflects some linear patterns in the data. Consider all \( C_h^n \) subsets, and compute the determinant of the covariance matrix for each subset. The subset with the smallest determinant is used to calculate the usual \( p \times 1 \) mean vector, \( \mathbf{m} \) (or \( \text{MCD}_1 \)), and corresponding \( p \times p \) covariance matrix, \( \mathbf{C} \) (or \( \text{MCD}_2 \)). These estimators are called the minimum covariance determinant estimators (MCD’s). Hawkins (1994) illustrates that the MCD’s are actually maximum likelihood estimators when \( h \) observations are from the correct multivariate normal distribution, while the other \( n - h \) observations are from a different, mean-shifted, multivariate normal distribution. As with the MVE’s, these estimators also provide a robust yardstick for determining what constitutes an outlier and/or a high leverage point when used to formulate a set of robust distances.

The MCD estimation process is computationally intensive as well. By definition, the exact MCD solution is found when all \( C_h^n \) subsets are analyzed, which again becomes an enormous undertaking as the sample size increases. *This is in contrast to the MVE algorithm, which still requires iteration of the observation weights.*

As in the MVE case, the feasible solution algorithm (FSA) of Hawkins (1994) allows for a reduction in the computation process by using \( N \) random starts. A probabilistic argument based on a simulation analysis is offered by Hawkins (1994) to contend that only a small number of random subsets is required in order to obtain the exact MCD estimators with “high probability.” However, \( N \) depends on the particular data and the number of local minima that might be reached (again referred to as feasible solutions by Hawkins (1994)), so one might still take \( N \) as large as realistically possible. The FSA algorithm for MCD computation is listed in Appendix B.5.
In terms of asymptotic properties, the MCD is preferred over the MVE. Both have an 
\[ \left( \frac{n - p - 1}{2} \right) / n \] breakdown point, which is 50% asymptotically. Both are also affine equivariant. However, the MCD converges as \( n^{-1/2} \) while MVE converges as \( n^{-1/3} \) (Rousseeuw and Leroy (1987), Davies (1992)). Thus, MCD has higher efficiency than does MVE. Butler, Davies, and Jhun (1993) prove consistency for the MCD location and dispersion, as well as asymptotic normality for the MCD location estimator. Woodruff and Rocke (1994) “strongly support” the use of MCD over MVE. It is also noted that Hawkins and Olive (1999) offer improved FSA algorithms.

§3.2.6 Stalactite Plot Analysis

The stalactite plot for unmasking multiple outliers in the regression setting was previously discussed in Section 2.7.2. This procedure also adapts itself to multivariate location and dispersion estimation by viewing the activity of the subset-based Mahalanobis distances as one increases the size, \( m \), of the subset used to calculate the mean vector and covariance matrix. Again, this is a random subsampling type algorithm, and the end result is a stalactite plot. Actual estimators for location and dispersion can be acquired by scanning across all sequences and determining which subset of size \( h \) produced the largest robust distance. This algorithm is detailed in Appendix B.6.

§3.2.7 Hadi’s Forward Search

Developed by Hadi (1992), this procedure follows the general flow of the stalactite plot, except that only one sequence of computations is required. Appendix B.7 contains specifics on this particular algorithm. According to Rocke and Woodruff (1996), “the algorithm ... breaks down if the contamination is extremely far away from the good data in the correct metric.” They also point out that the procedure is not affine equivariant because of the use of the coordinatewise median.
§3.3 Chapter Summary

The discussion of bounded influence regression in Chapter 2 included mention of leverage weights. Recall that the Welsch weights of Section 2.1 are based on ordinary hat diagonals, and as such are not resistant to data with extreme leverage. By obtaining robust multivariate location and dispersion estimators, \( \mathbf{m} \) and \( \mathbf{C} \), respectively, a robust squared Mahalanobis distance measure, \( RD_i^2 \) (with \( d_i^2 \) also used in a broad sense), can then be defined as

\[
RD_i^2 = (z_{y,d} - \mathbf{m})' \mathbf{C}^{-1} (z_{y,d} - \mathbf{m}), \quad i = 1, 2, ..., n.
\]

Leverage can now be measured in a more outlier resistant metric.

In replacing the \( \pi_i \)'s in bounded influence regression, one possible weighting scheme, commonly referred to as Mallows weights, has the form

\[
w_i = \min \left[ 1, \left( \frac{b}{(z_{y,d} - \mathbf{m})' \mathbf{C}^{-1} (z_{y,d} - \mathbf{m})} \right)^{-\alpha/2} \right],
\]

with \( b \) and \( \alpha \) being predefined scalars. Mallows weights can be understood by considering the perimeter of the ellipsoid defined by setting \( RD_i^2 \) equal to \( b \). If a regressor location is outside of this ellipsoid, then it has higher leverage and the observation should be downweighted as a function of how far outside the ellipsoid this observation really lays. If a regressor location falls inside of this ellipsoid, then leverage is minimal and no downweighting is necessary.

This weighting scheme is incorporated into some of the current high breakdown regression procedures that are mentioned in the next chapter. For example, the MVE estimates are generally employed to determine the Mallows weights, with the constants \( \alpha \) and \( b \) generally taken to be 2 and \( \chi^2_{0.95, p-1} \) respectively. The choice of \( \alpha = 2 \) keeps the standard errors of certain methods from breaking down (Simpson, Ruppert, and Carroll (1992)). Viewing the robust squared Mahalanobis distances as having an asymptotic \( \chi^2_{p-1} \) distribution provides a rough value for the boundary parameter, \( b \). With a viable weighting scheme in place, discussion can now shift to the topic of high breakdown regression procedures that are currently available.