5-2004

A Resistant Estimator of Multivariate Location and Dispersion

David J. Olive
Southern Illinois University Carbondale, dolive@math.siu.edu

Published in *Computational Statistics & Data Analysis*, 46, 99-101.

A Resistant Estimator of Multivariate Location and Dispersion

David J. Olive*

Southern Illinois University

May 29, 2003

Abstract

This paper presents a simple resistant estimator of multivariate location and dispersion. The DD plot is a plot of Mahalanobis distances from the classical estimator versus the distances from a resistant estimator and can be used to detect outliers and as a diagnostic for multivariate normality. The new estimator can be used in the DD plot, is easy to compute and provides insights about several useful robust algorithm techniques.

KEY WORDS: DD plot, minimum covariance determinant estimator.

---

*David J. Olive is Assistant Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. E-mail address: dolive@math.siu.edu. The author is grateful to the editors and referees for a number of helpful suggestions for improvement in the article. This research was supported by NSF grant DMS 0202922.

1 INTRODUCTION

A *multivariate location and dispersion model* is a joint distribution for a $p \times 1$ random vector $\boldsymbol{x}$ that is completely specified by a $p \times 1$ population *location* vector $\boldsymbol{\mu}$ and a $p \times p$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. The observations $\boldsymbol{x}_i$ for $i = 1, ..., n$ are collected in an $n \times p$ matrix $\boldsymbol{W}$ with n rows $\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T$.

Let the $p \times 1$ column vector $T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{W})$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \tag{1.1}$$

for each observation $\boldsymbol{x}_i$. Notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. The classical Mahalanobis distance corresponds to the sample mean and sample covariance matrix

$$T(\boldsymbol{W}) = \overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \text{ and } \boldsymbol{C}(\boldsymbol{W}) = \boldsymbol{S} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\boldsymbol{x}_{\mathrm{i}} - \mathrm{T}(\boldsymbol{W}))(\boldsymbol{x}_{\mathrm{i}} - \mathrm{T}(\boldsymbol{W}))^{\mathrm{T}}$$

and will be denoted by MD_i.

There is an enormous literature on the detection of outliers and influential cases for the multivariate location and dispersion model. Robust estimators are often computed by applying the classical estimator to a subset of the data. Consider the subset J_o of $c_n \approx n/2$ observations whose sample covariance matrix has the smallest determinant among all $C(n, c_n)$ subsets of size c_n. Let T_{MCD} and $\boldsymbol{C}_{MCD}$ denote the sample mean and sample covariance matrix of the c_n cases in J_o. Then the *minimum covariance determinant* MCD(c_n) estimator is $(T_{MCD}(\boldsymbol{W}), \boldsymbol{C}_{MCD}(\boldsymbol{W}))$.

This estimator is impractical to compute, so algorithm estimators are used instead. The "basic resampling" or "elemental set" algorithm for robust estimators uses K_n "elemental starts" – randomly selected subsets of $p+1$ cases where p is the number of variables. The jth elemental fit is the classical estimator $(T_j, \boldsymbol{C}_j)$ computed from the jth elemental set. For each fit a criterion function that depends on all n cases is computed. Then the algorithm returns the elemental fit that optimizes the criterion. The efficiency and resistance properties of the basic resampling algorithm estimator turn out to depend strongly on the number of starts K_n used – see Hawkins and Olive (2002).

Another important algorithm technique is *concentration*. Starts are again used, but they are not necessarily elemental. Let $(T_{0,j}, \boldsymbol{C}_{0,j})$ be the jth start and compute all n Mahalanobis distances $D_i(T_{0,j}, \boldsymbol{C}_{0,j})$. At the next iteration, the classical estimator $(T_{1,j}, \boldsymbol{C}_{1,j})$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest distances. This iteration can be continued for k steps resulting in the sequence of estimators $(T_{0,j}, \boldsymbol{C}_{0,j}), (T_{1,j}, \boldsymbol{C}_{1,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. The result of the iteration $(T_{k,j}, \boldsymbol{C}_{k,j})$ is called the jth attractor. Using $k = 5$ concentration steps works well, and iterating until convergence is usually fast. In a concentration algorithm, the final estimator is the attractor that optimizes the criterion. The basic resampling algorithm is a special case with $k = 0$.

Concentration has been used by several authors. The DGK estimator (Devlin, Gnanadesikan, and Kettenring 1975, 1981) uses the classical estimator computed from all n cases as the only start, and results from Lopuhaä (1999) show that the DGK estimator is $\sqrt{n}$ consistent. Gnanadesikan and Kettenring (1972, pp. 94–95) provide a similar algorithm. Rousseeuw and Van Driessen (1999, p. 214) prove that the concentration steps make

the determinant $|\boldsymbol{C}_{i+1,j}| \leq |\boldsymbol{C}_{i,j}|$ and provide the FMCD concentration algorithm, implemented by the Splus function `cov.mcd`, for the MCD estimator using $K_n \equiv K = 500$ elemental starts. Hawkins and Olive (1999) provide a similar MCD algorithm while Hawkins and Olive (2002) suggest that the percentage γ_o of distant outliers that can be handled by `cov.mcd` is

$$\gamma_o \approx \min(\frac{1}{2}, 1 - [1 - (0.2)^{1/K}]^{1/h})100\% \tag{1.2}$$

if n is large, $K = 500$ and $h = p + 1$.

In addition to concentration and randomly selecting elemental sets, two additional algorithm techniques will be examined in this paper. He and Wang (1996) suggest computing the classical estimator and a robust estimator. The final estimator is the classical estimator if both estimators are "close," otherwise the final estimator is the robust estimator. He (1991) proposed a similar technique for regression. The second technique was proposed by Gnanadesikan and Kettenring (1972, p. 90). They suggest using the dispersion matrix $\boldsymbol{C} = [c_{i,j}]$ where $c_{i,j}$ is a robust estimator of the covariance of X_i and X_j. Computing the classical estimator on a subset of the data results in an estimator of this form. The identity

$$c_{i,j} = \mathrm{Cov}(\mathrm{X_i}, \mathrm{X_j}) = [\mathrm{Var}(\mathrm{X_i} + \mathrm{X_j}) - \mathrm{Var}(\mathrm{X_i} - \mathrm{X_j})]/4$$

where $\mathrm{Var}(\mathrm{X}) = \sigma^2(\mathrm{X})$ suggests that a robust estimator of dispersion can be created by replacing the sample standard deviation $\hat{\sigma}$ by a robust estimator of scale. Maronna and Zamar (2002) modify this idea to create a fairly fast high breakdown consistent estimator of multivariate location and dispersion.

Robust estimators tend to be judged by their Gaussian efficiency and breakdown value (see Zuo 2001 for references). The following notation will be useful. Let $\boldsymbol{W}_d^n$ denote the data matrix where any $d < n/2$ of the n cases have been replaced by arbitrarily bad contaminated cases. Then the contamination fraction is $\gamma = d/n$.

Consider a fixed data set $\boldsymbol{W}_d^n$. A multivariate location estimator T basically "breaks down" if the d outliers can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|)$ arbitrarily large where $\boldsymbol{w}_i^T$ is the ith row of $\boldsymbol{W}_d^n$. Thus a multivariate location estimator T will not break down if T can not be driven out of some ball of (possibly huge) radius R about the origin.

The estimator $\boldsymbol{C}$ breaks down if the smallest eigenvalue λ_p can be driven to zero or if the largest eigenvalue λ_1 can be driven to ∞. From numerical linear algebra, it is known that the largest eigenvalue of a $p \times p$ matrix $\boldsymbol{C}$ is bounded above by $p \max |c_{i,j}|$ where $c_{i,j}$ is the (i,j) entry of $\boldsymbol{C}$. See Datta (1995, p. 403).

Assume that $(T, \boldsymbol{C})$ is the classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ applied to some subset of $c_n \approx n/2$ cases of the data. Denote these cases by $\boldsymbol{z}_1, ..., \boldsymbol{z}_{c_n}$. Then the (i,j) entry of $\boldsymbol{C}$ is

$$c_{i,j} = \frac{1}{c_n - 1} \sum_{k=1}^{c_n} (z_{i,k} - \overline{z}_i)(z_{j,k} - \overline{z}_j).$$

Hence the maximum eigenvalue λ_1 can not get arbitrarily large if the $\boldsymbol{z}_i$ are all contained in some ball of radius R about the origin, e.g., if none of the c_n cases is an outlier. If all of the $\|\boldsymbol{z}_i\|$ are bounded, then all of the λ_i are bounded, and λ_p can only be driven to zero if the determinant of $\boldsymbol{C}$ can be driven to zero. The determinant $|\boldsymbol{S}|$ of $\boldsymbol{S}$ is known as the *generalized sample variance*. Consider the hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq D^2_{(c_n)}\} \tag{1.3}$$

where $D^2_{(c_n)}$ is the c_nth smallest squared Mahalanobis distance based on $(T, \boldsymbol{C})$. This ellipsoid contains the c_n cases with the smallest D^2_i. The volume of this ellipsoid is proportional to the square root of the determinant $|\boldsymbol{C}|^{1/2}$, and this volume will be positive unless extreme degeneracy is present among the c_n cases. See Johnson and Wichern (1988, pp. 103-104).

Section 2 uses ideas presented in this section to create a simple resistant estimator for multivariate location and dispersion.

2 The Median Ball Algorithm

The simplest form of the *median ball algorithm* (MBA) estimator for multivariate location and dispersion uses two carefully chosen starts. Suppose that the data $\boldsymbol{x}_i$ are iid from an elliptically contoured (EC) distribution with finite second moments and parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The first start $(T_{0,1}, \boldsymbol{C}_{0,1})$ is chosen so that the first attractor $(T_{5,1}, \boldsymbol{C}_{5,1})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where the constant $c > 0$ depends on the EC distribution. The second start $(T_{0,2}, \boldsymbol{C}_{0,2})$ is chosen so that the second attractor $(T_{5,2}, \boldsymbol{C}_{5,2})$ is a high (50%) breakdown estimator. Let $(T_A, \boldsymbol{C}_A) = (T_{5,i}, \boldsymbol{C}_{5,i})$ where $i = 1$ if the determinant $|\boldsymbol{C}_{5,1}| \leq |\boldsymbol{C}_{5,2}|$ and $i = 2$, otherwise. Then the MBA estimator $(T_{MBA}, \boldsymbol{C}_{MBA})$ takes $T_{MBA} = T_A$ and

$$\boldsymbol{C}_{MBA} = \frac{\mathrm{MED}(D^2_i(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{2.1}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with p degrees of freedom. This scaling makes $\boldsymbol{C}_{MBA}$ a better estimate of $\boldsymbol{\Sigma}$ if the data is multivariate normal (MVN). See Olive (2002).

A good choice for the first start is the classical estimator $(T_{0,1}, \boldsymbol{C}_{0,1}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. After five concentration steps, the resulting attractor $(T_{5,1}, \boldsymbol{C}_{5,1})$ is the DGK estimator. The DGK estimator is affine equivariant, $\sqrt{n}$ consistent and very simple to compute.

The choice for the second start is motivated by the results on breakdown given in Section 1. Find the set of $c_n \approx n/2$ cases $\boldsymbol{x}_i$ that are closest to the coordinatewise median MED($\boldsymbol{x}$) in Euclidean distance, and let the second start $(T_{0,2}, \boldsymbol{C}_{0,2})$ be the classical sample mean and covariance of these cases. Arcones (1995) and Kim (2000) showed that $T_{0,2}$ is a high breakdown, $\sqrt{n}$ consistent estimator of multivariate location. Since only cases $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{x})\| \leq \text{MED}(\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{x})\|)$ are used, the largest eigenvalue of $\boldsymbol{C}_{0,2}$ is bounded if fewer than half of the cases are outliers.

The geometric behavior of this start is simple. If the data $\boldsymbol{x}_i$ are MVN (or EC) then the highest density regions of the data are hyperellipsoids. The set of $\boldsymbol{x}$ closest to the coordinatewise median in Euclidean distance is a hypersphere. For EC data the highest density ellipsoid and hypersphere will have approximately the same center, and the hypersphere will be drawn towards the longest axis of the hyperellipsoid. Hence too much data will be trimmed in that direction. For example, if the data are MVN with $\boldsymbol{\Sigma} = \text{diag}(1, 2, ..., p)$ then $\boldsymbol{C}_{0,2}$ may underestimate the largest variances and overestimate the smallest variances. Taking five concentration steps can greatly reduce the bias of $\boldsymbol{C}_{5,2}$ if the data is MVN, and the determinant $|\boldsymbol{C}_{5,2}| < |\boldsymbol{C}_{0,2}|$ unless the attractor is equal to the start. The attractor $(T_{5,2}, \boldsymbol{C}_{5,2})$ is not affine equivariant but is resistant to gross outliers in that they will initially be given weight zero if they are further than the median Euclidean distance from the coordinatewise median. Gnanadesikan and Kettenring (1972, p. 94) suggest an estimator similar to the attractor $(T_{5,2}, \boldsymbol{C}_{5,2})$.

The DD plot (Rousseeuw and Van Driessen 1999) is useful for detecting outliers. This plots MD_i vs. RD_i where the RD_i are Mahalanobis distances based on a resistant estimator. The plotted points should cluster about the identity line if the data is MVN or if the resistant estimator fails so that the resistant estimator is nearly the same as the classical estimator. The resistant estimator will be useful if $(T, \boldsymbol{C}) \approx (\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c > 0$ since scaling by c affects the vertical labels of the RD_i but not the shape of the DD plot.

Examining a special outlier configuration may be useful for comparing the FMCD and MBA concentration algorithms. Assume that the "clean" data is ellipsoidal and highly correlated about the major axis $\boldsymbol{a}_1$. Suppose that there is a group of distant outliers in a direction $\boldsymbol{a}_0$ orthogonal to $\boldsymbol{a}_1$, and that the subset of c_n cases with the smallest distances based on the start is not clean. Heuristically, if the sample mean of the c_n cases with the smallest distances is close enough to the clean cases, then after the concentration step the c_n cases will contain fewer outliers and more clean cases. After several steps the attractor may be clean. When the contamination proportion is high (roughly larger than the level given by Eq. (1.2)), every randomly chosen elemental set of $p + 1$ cases will be contaminated with high probability. Hence the probability is high that the initial subset of c_n cases from each FMCD start will contain more outliers than the second MBA start that uses the coordinatewise median. Thus the attractor from the second MBA start is more likely to be clean than the best attractor from the $K = 500$ FMCD starts. Notice that the DGK estimator can have considerable resistance to a group of distant outliers that is placed on the major axis $\boldsymbol{a}_1$.

Many high breakdown estimators of multivariate location and dispersion have been proposed. Estimators such as the projection, S and minimum volume ellipsoid estimators are difficult to compute and are typically approximated by a basic resampling estimator that uses $K \leq 3000$ starts. Such algorithm estimators are inconsistent and have a breakdown value bounded above by K/n. See Hawkins and Olive (2002). Maronna and Zamar (2002) compare their OGK estimator with the FMCD estimator and concluded that they performed about equally well on several real data sets.

To compare $(T_{MBA}, \boldsymbol{C}_{MBA})$ and $(T_{FMCD}, \boldsymbol{C}_{FMCD})$, we made the MBA and FMCD DD plots for 37 small data sets (several are available from the author's website). On most of the data sets the MBA and FMCD distances were highly correlated but for the "modified wood data" (Rousseeuw and Leroy, 1987) and the "nasty data", contributed by Douglas M. Hawkins, the outliers could be detected from the FMCD DD plot but not from the MBA DD plot. The FMCD covariance estimator was more likely to be singular than the MBA estimator when some of the variables were categorical. For such data sets, the robust estimators should be examined on the full data set and with the categorical variables omitted. The DD plot of the MBA distances vs. the FMCD distances was often V-shaped if one or more of the predictors needed to be transformed in order to make the joint distribution of the predictors approximately elliptically contoured.

A small simulation study was also used to illustrate properties of concentration estimators. We computed the FMCD estimator with the Splus function `cov.mcd` which allows up to 50 predictors. Initially the data sets had no outliers, and all 100 cases were MVN with zero mean vector and $\boldsymbol{\Sigma}$ = diag(1,2, ..., p). We generated 500 runs of this data with $p = 4$. The averaged diagonal elements of $\boldsymbol{C}_{MBA}$ were 1.202, 2.260, 3.237 and

4.204. (In the simulations, the scale factor in Eq. (2.1) appeared to be slightly too large for small n but slowly converged to the correct factor as n increased.) The averaged diagonal elements of $\boldsymbol{C}_{FMCD}$ were 0.838, 1.697, 2.531, and 3.373. The approximation $1.2\boldsymbol{C}_{FMCD} \approx \boldsymbol{\Sigma}$ was good. For both matrices, all off diagonal elements had average values less than 0.034 in magnitude.

Next data sets with $\gamma = 40\%$ outliers were generated. The last 60 cases were MVN with zero mean vector and $\boldsymbol{\Sigma}$ = diag(1,2, ..., p). The first 40 cases were MVN with the same $\boldsymbol{\Sigma}$, but the $p \times 1$ mean vector $\boldsymbol{\mu} = (10, 10\sqrt{2}, ..., 10\sqrt{p})^T$. We generated 500 runs of this data using $p = 4$. Shown below are the averages of the estimators $\boldsymbol{C}_{MBA}$ and $\boldsymbol{C}_{FMCD}$. Notice that $\boldsymbol{C}_{FMCD}$ performed extremely well while the $\boldsymbol{C}_{MBA}$ entries were over inflated by a factor of about 2 since the outliers inflate the scale factor $\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))/\chi^2_{p,0.5}$. Although the MBA estimator is biased, the outliers in the MBA DD plot will have large RD_i since $\boldsymbol{C}_{MBA} \approx 2\boldsymbol{C}_{FMCD} \approx 2\boldsymbol{\Sigma}$.

MBA

$$\begin{bmatrix} 2.120 & -0.031 & -0.069 & 0.004 \\ -0.031 & 4.144 & -0.111 & -0.146 \\ -0.069 & -0.111 & 6.211 & -0.419 \\ -0.138 & 0.008 & -0.419 & 7.933 \end{bmatrix}$$

FMCD

$$\begin{bmatrix} 0.980 & 0.002 & -0.004 & 0.011 \\ 0.002 & 1.977 & -0.008 & -0.014 \\ -0.004 & -0.008 & 2.991 & 0.013 \\ 0.011 & -0.014 & 0.013 & 3.862 \end{bmatrix}$$

When p is increased to 8, the `cov.mcd` estimator was usually not useful for detecting the outliers for this type of contamination. Figure 1 shows that now the FMCD RD_i are highly correlated with the MD_i. The DD plot based on the MBA estimator detects the outliers. See Figure 2.

We also compared the two estimators by simulating the outlier data for various values of p, n and γ. For each configuration, twenty data sets were generated. The criterion was the number of the runs where the minimum distance from the outliers was greater than the maximum distance from the non-outliers. When this is the case, the outliers can be separated from non-outliers in the DD plot with a horizontal line. As a benchmark, a count of 17 or higher suggests that the estimator could usually handle the outlier configuration. Table 1 displays the results. Notice that the count provides an approximate lower bound on the number of runs where the best attractor was clean and that whenever the MBA count was less than twenty, the FMCD count was equal to zero. Table 1 also suggests that Eq. (1.2) does give a rough measure of the proportion of distant outliers that the FMCD algorithm can handle. For $n = 500$, Eq. (1.2) overestimates the proportion slightly for small p and underestimates the proportion slightly for larger p.

The comparison of the two estimators on real and simulated data suggests that for some outlier configurations the MBA estimator is inferior to the FMCD estimator while for other configurations the MBA estimator is superior. The discussion papers by Rocke and Woodruff (2001) and by Hubert (2001) stress the fact that no one estimator can dominate all others for every outlier configuration. These papers and Wisnowski, Simpson, and Montgomery (2002) give outlier configurations that can cause problems for the FMCD estimator. The MBA estimator is most vulnerable to outliers that lie inside the hypersphere based on the median Euclidean distance from the coordinatewise median.

Now we give some rules of thumb for using the ideas presented in this paper to analyze multivariate data. First, make a scatterplot matrix of the predictors if p is

small. Transformations may be needed if strong nonlinearities or outliers are present in the marginal plots. Next, make a scatterplot matrix of the Mahalanobis distances from several estimators including the FMCD, MBA and classical estimators. Since robust estimators will often fail if there are three or more groups of data, a cluster analysis may be needed. Suppose that a group of unexplained outliers is detected (e.g. the outliers are not recording errors and are not impossible values). Run the classical analysis on the full data set, the classical analysis with the outliers deleted, and a weighted analysis with the outliers given weight zero. Perform the usual checks on the classical analysis with the outliers deleted to show that the classical analysis is appropriate for the bulk of the data. A DD plot from the weighted analysis may be useful for showing the proportion and severity of the outliers in the data.

The ideas in this paper can also be used to improve existing algorithms. Adding the classical estimator as a start to the FMCD estimator should greatly stabilize the estimator on clean data with a cost of about a 1% increase in computing time. The Maronna and Zamar (2002) OGK estimator can probably be improved with concentration. A simple modification for the MBA estimator would be to add additional starts. For example, let $T_{0,3}$ be the coordinatewise median and let $\boldsymbol{C}_{0,3} = \mathrm{diag}(\mathrm{k}(\mathrm{MAD}(\mathrm{X}_1))^2, ..., \mathrm{k}(\mathrm{MAD}(\mathrm{X}_\mathrm{p}))^2)$ where $\mathrm{MAD}(X_i)$ is the median absolute deviation of the ith variable X_i and $k = 1$ or $k = (1.483)^2$. Instead of hyperspheres, this start generates hyperellipsoids with axes parallel to the coordinate axes. It may be useful to separate starts that result in affine equivariant attractors from starts that do not. For example, the MBA estimator is permutation invariant but not affine equivariant. Wang and Raftery (2002) discuss the merits of affine and non-affine equivariant estimators.

Finally, suppose that the researcher desires to plug in a robust estimator for the classical estimator. A good choice would be to create an adaptive estimator using the He and Wang (1996) cross–checking technique. First cross–check the MBA estimator and the classical estimator. Then cross–check the result with the FMCD estimator. Finally, cross–check the result with another good estimator such as the OGK estimator. The resulting estimator may have good resistance properties and may be asymptotically equivalent to the classical estimator when the data follows a multivariate normal distribution. Again, make a scatterplot matrix of the distances from the component estimators to recover information that might be lost by only using the final estimator.

The author's website (http://www.math.siu.edu/olive) contains several interesting data sets as well as a file *rpack.txt* that contains several Splus functions. The function `covmba` produces the MBA estimator, and the function `ddcomp` produces the MBA and FMCD DD plots. The function `concmv` illustrates that the initial median ball sphere may contain a large proportion of outliers while the attractor is clean when $p = 2$. The c_n cases with the smallest distances are highlighted at each concentration step. The outliers are placed on the borderline of the MBA screen so that the MBA estimator wins the "tug–of–war" in about nine out of ten runs.

3 References

Arcones, M.A., 1995. Asymptotic normality of multivariate trimmed means. Statist. Probab. Lett. 25, 43-53.

Datta, B.N., 1995. Numerical Linear Algebra and Applications. Brooks/Cole, Pacific

Grove, CA.

Devlin, S.J., Gnanadesikan, R., Kettenring, J.R., 1975. Robust estimation and outlier detection with correlation coefficients. Biometrika. 62, 531-545.

Devlin, S.J., Gnanadesikan, R., Kettenring, J.R., 1981. Robust estimation of dispersion matrices and principal components. J. Amer. Statist. Assoc. 76, 354-362.

Gnanadesikan, R., Kettenring, J.R., 1972. Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics. 28, 81-124.

Hawkins, D.M., Olive, D.J., 1999. Improved feasible solution algorithms for high breakdown estimation. Comput. Statist. Data Anal. 30, 1-11.

Hawkins, D.M., Olive, D.J., 2002. Inconsistency of resampling algorithms for high breakdown regression estimators and a new algorithm. J. Amer. Statist. Assoc. 97, 136-159.

He, X., 1991. A local breakdown property of robust tests in linear regression. J. Multiv. Anal. 38, 294-305.

He, X., Wang, G., 1996. Cross-checking using the minimum volume ellipsoid estimator. Statist. Sinica. 6, 367-374.

Hubert, M., 2001, Discussion of 'Multivariate outlier detection and robust covariance matrix estimation' by D. Peña and F.J. Prieto. Technom. 43, 303-306.

Johnson, R.A., Wichern, D.W., 1988. Applied Multivariate Statistical Analysis. 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

Kim, J., 2000. Rate of convergence of depth contours: with application to a multivariate metrically trimmed mean. Statist. Probab. Lett. 49, 393-400.

Lopuhaä, H.P., 1999. Asymptotics of reweighted estimators of multivariate location and scatter. Annals Statist. 27, 1638-1665.

Maronna, R.A., Zamar, R.H., 2002. Robust estimates of location and dispersion for high-dimensional datasets. Technom. 44, 307-317.

Olive, D.J., 2002. Applications of robust distances for regression. Technom. 44, 64-71.

Rocke, D.M., Woodruff, D.L., 2001. Discussion of 'Multivariate outlier detection and robust covariance matrix estimation' by D. Peña and F.J. Prieto. Technom. 43, 300-303.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. John Wiley and Sons, Inc., NY.

Rousseeuw, P.J., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technom. 41, 212-223.

Wang, N., Raftery, A.E., 2002. Nearest-neighbor variance estimation (NNVE): robust covariance estimation via nearest-neighbor cleaning. J. Amer. Statist. Assoc. 97, 994-1019.

Wisnowski J.W., Simpson J.R., Montgomery D.C., 2002. A performance study for multivariate location and shape estimators. Quality Reliability Engineering Internat. 18, 117-129.

Zuo, Y., 2001. Some quantitative relationships between two types of finite sample breakdown point. Statist. Probab. Lett. 51, 369-375.

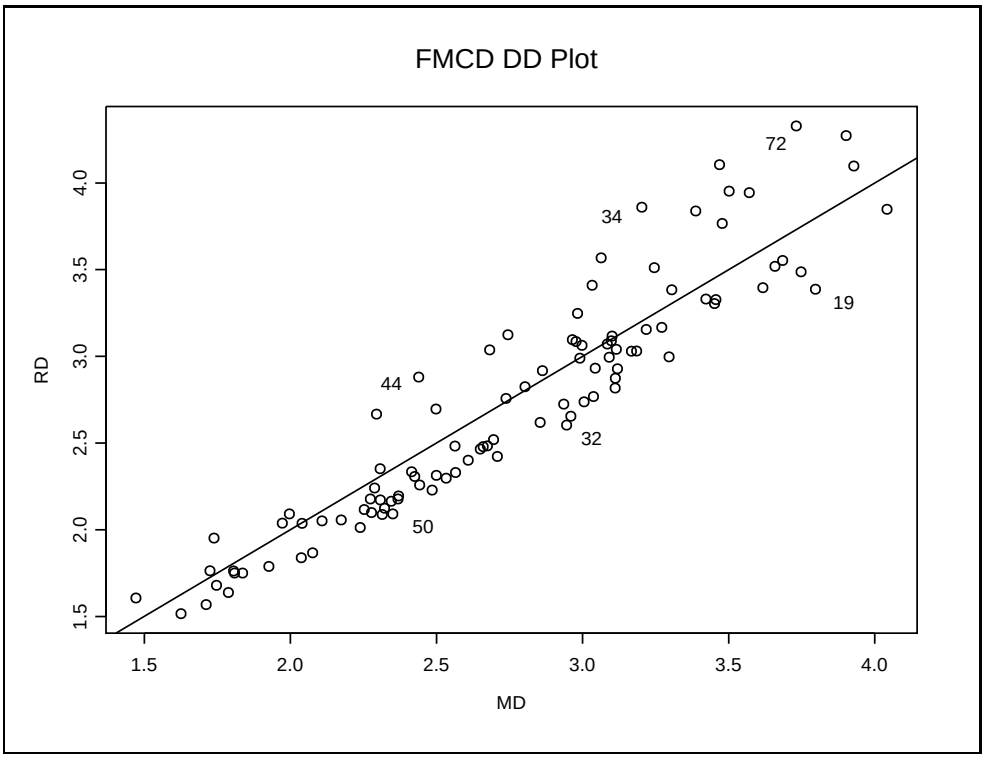

Figure 1: The FMCD Estimator Failed

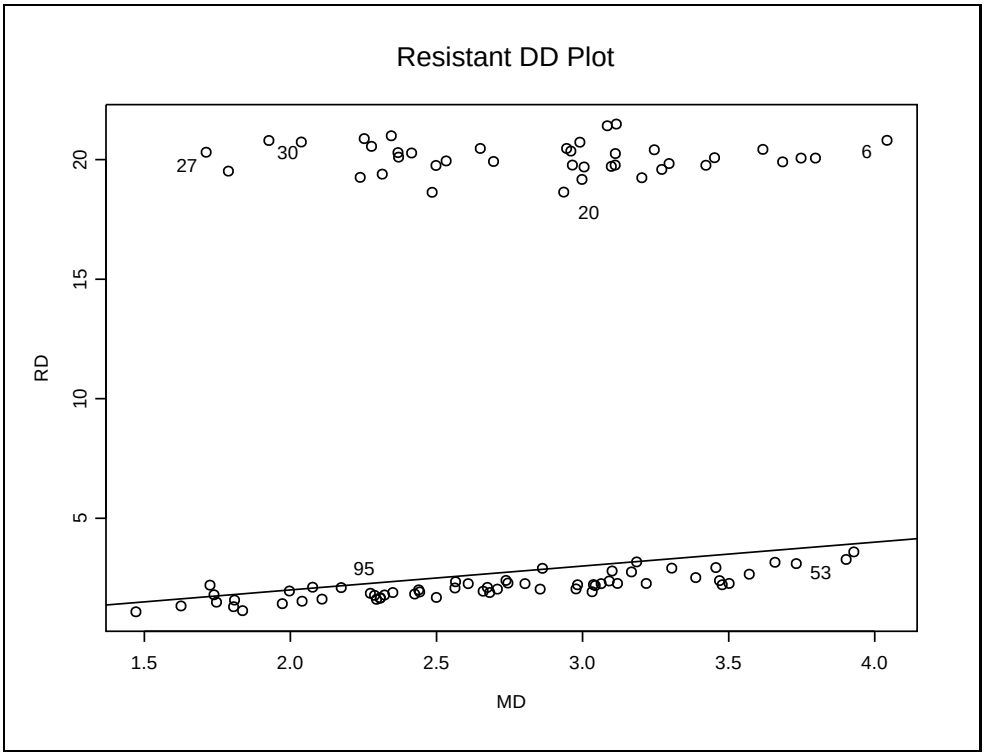

Figure 2: The Outliers are Large in the MBA DD Plot

Table 1: Number of 20 runs where outliers had larger distances than non–outliers.

p	n	γ	MBA Count	FMCD Count
3	100	0.49	20	17
4	20	0.49	18	0
4	200	0.49	20	20
8	500	0.47	20	0
8	500	0.40	20	20
9	500	0.43	20	1
9	500	0.36	20	15
10	100	0.49	19	0
10	100	0.30	20	20
10	500	0.47	20	0
10	500	0.40	20	7
15	500	0.30	20	15
20	100	0.49	12	0
20	100	0.30	20	0
20	500	0.23	20	20
40	500	0.13	20	20
50	400	0.40	19	0
50	500	0.10	20	20
100	700	0.30	17	NA
100	4000	0.40	18	NA