Chapter 7
Robust Regression

This chapter considers outlier detection and then develops robust regression estimators. Robust estimators of multivariate location and dispersion are useful for outlier detection and for developing robust regression estimators. Outliers and dot plots were discussed in Chapter 3.

Definition 7.1 An **outlier** corresponds to a case that is far from the bulk of the data.

Definition 7.2. A *dot plot* of $Z_1, ..., Z_m$ consists of an axis and m points each corresponding to the value of Z_i.

The following plots and techniques will be developed in this chapter. For the location model, use a dot plot to detect outliers. For the multivariate location model with $p = 2$ make a scatterplot. For multiple linear regression with one nontrivial predictor x, plot x versus Y. For the multiple linear regression model, make the residual and response plots. For the multivariate location model, make the DD plot if $n \geq 5p$, and use `ddplot5` if $p > n$. If p is not much larger than 5, elemental sets are useful for outlier detection for multiple linear regression and multivariate location and dispersion.

7.1 The Location Model

The location model is

$$Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{7.1}$$

where $e_1, ..., e_n$ are error random variables, often iid with zero mean. The location model is used when there is one variable Y, such as height, of interest. The location model is a special case of the multiple linear regression model and of the multivariate location and dispersion model, where there are p

variables $x_1, ..., x_p$ of interest, such as height and weight if $p = 2$. The dot plot of Definition 7.2 is useful for detecting outliers in the location model.

The location model is often summarized by obtaining point estimates and confidence intervals for a location parameter and a scale parameter. Assume that there is a sample $Y_1, \ldots, Y_n$ of size n where the Y_i are iid from a distribution with median MED(Y), mean $E(Y)$, and variance $V(Y)$ if they exist. The location parameter μ is often the population mean or median while the scale parameter is often the population standard deviation $\sqrt{V(Y)}$. The ith *case* is Y_i.

Point estimation is one of the oldest problems in statistics and four important statistics for the location model are the sample mean, median, variance, and the median absolute deviation (MAD). Let $Y_1, \ldots, Y_n$ be the random sample; i.e., assume that $Y_1, ..., Y_n$ are iid. The sample mean is a measure of location and estimates the population mean (expected value) $\mu = E(Y)$. The *sample mean* $\overline{Y} = \dfrac{\sum_{i=1}^n Y_i}{n}$. The *sample variance* $S_n^2 = \dfrac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} = \dfrac{\sum_{i=1}^n Y_i^2 - n(\overline{Y})^2}{n-1}$, and the *sample standard deviation* $S_n = \sqrt{S_n^2}$.

If the data set $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the ith order statistic and the $Y_{(i)}$'s are called the *order statistics.* If the data $Y_1 = 1, Y_2 = 4, Y_3 = 2, Y_4 = 5$, and $Y_5 = 3$, then $\overline{Y} = 3$, $Y_{(i)} = i$ for $i = 1, ..., 5$ and MED(n) = 3 where the sample size $n = 5$. The sample median is a measure of location while the sample standard deviation is a measure of spread. The sample mean and standard deviation are vulnerable to outliers, while the sample median and MAD, defined below, are outlier resistant.

Definition 7.3. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \quad \text{if n is odd,} \tag{7.2}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \quad \text{if n is even.}$$

The notation MED(n) = MED($Y_1, ..., Y_n$) will also be used.

Definition 7.4. The *sample median absolute deviation* is

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n). \tag{7.3}$$

Since MAD(n) is the median of n distances, at least half of the observations are within a distance MAD(n) of MED(n) and at least half of the observations are a distance of MAD(n) or more away from MED(n). Like the standard deviation, MAD(n) is a measure of spread.

Example 7.1. Let the data be $1, 2, 3, 4, 5, 6, 7, 8, 9$. Then MED($n$) $= 5$ and MAD(n) $= 2 = \text{MED}\{0, 1, 1, 2, 2, 3, 3, 4, 4\}$.

The trimmed mean is used in Chapter 9. We recommend the 25% trimmed mean. Let $\lfloor x \rfloor$ denote the "greatest integer function" (e.g., $\lfloor 7.7 \rfloor = 7$).

Definition 7.5. The symmetrically trimmed mean or the δ *trimmed mean*

$$T_n = T_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \tag{7.4}$$

where $L_n = \lfloor n\delta \rfloor$ and $U_n = n - L_n$. If $\delta = 0.25$, say, then the δ trimmed mean is called the 25% trimmed mean.

The $(\delta, 1-\gamma)$ *trimmed mean* uses $L_n = \lfloor n\delta \rfloor$ and $U_n = \lfloor n\gamma \rfloor$.

Estimators that use order statistics are common. Theory for the MAD, median, and trimmed mean is given, for example, in Olive (2008), which also gives confidence intervals based on the median and trimmed mean. The shorth estimator of Section 4.3 was used for prediction intervals.

7.2 The Multivariate Location and Dispersion Model

The multivariate location and dispersion (MLD) model is a special case of the multivariate linear model, just like the location model is a special case of the multiple linear regression model. Robust estimators of multivariate location and dispersion are useful for detecting outliers in the predictor variables and for developing an outlier resistant multiple linear regression estimator.

The practical, highly outlier resistant, $\sqrt{n}$ consistent FCH, RFCH, and RMVN estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ are developed along with proofs. The RFCH and RMVN estimators are reweighted versions of the FCH estimator. It is shown why competing "robust estimators" fail to work, are impractical, or are not yet backed by theory. The RMVN and RFCH sets are defined and will be used for outlier detection and to create practical robust methods of multiple linear regression and multivariate linear regression. Many more applications are given in Olive (2017b).

Warning: This section contains many acronyms, abbreviations, and estimator names such as FCH, RFCH, and RMVN. Often the acronyms start with the added letter A, C, F, or R: A stands for *algorithm*, C for *concentration*, F for estimators that use a *fixed* number of trial fits, and R for *reweighted*.

Definition 7.6. The multivariate location and dispersion model is

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{e}_i, \quad i = 1, \ldots, n \tag{7.5}$$

where $\boldsymbol{e}_1, ..., \boldsymbol{e}_n$ are $p \times 1$ error random vectors, often iid with zero mean and covariance matrix $\text{Cov}(\boldsymbol{e}) = \text{Cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma}_{\boldsymbol{Y}} = \boldsymbol{\Sigma}_{\boldsymbol{e}}$.

Note that the location model is a special case of the MLD model with $p = 1$. If $E(\boldsymbol{e}) = \boldsymbol{0}$, then $E(\boldsymbol{Y}) = \boldsymbol{\mu}$. A $p \times p$ dispersion matrix is a symmetric matrix that measures the spread of a random vector. Covariance and correlation matrices are dispersion matrices. One way to get a robust estimator of multivariate location is to stack the marginal estimators of location into a vector. The coordinatewise median MED($\boldsymbol{W}$) is an example. The sample mean $\overline{\boldsymbol{x}}$ also stacks the marginal estimators into a vector, but is not outlier resistant.

Let $\boldsymbol{\mu}$ be a $p \times 1$ location vector and $\boldsymbol{\Sigma}$ a $p \times p$ symmetric dispersion matrix. Because of symmetry, the first row of $\boldsymbol{\Sigma}$ has p distinct unknown parameters, the second row has $p-1$ distinct unknown parameters, the third row has $p - 2$ distinct unknown parameters, ..., and the pth row has one distinct unknown parameter for a total of $1 + 2 + \cdots + p = p(p + 1)/2$ unknown parameters. Since $\boldsymbol{\mu}$ has p unknown parameters, an estimator $(T, \boldsymbol{C})$ of multivariate location and dispersion, needs to estimate $p(p+3)/2$ unknown parameters when there are p random variables. If the p variables can be transformed into an uncorrelated set then there are only $2p$ parameters, the means and variances, while if the dimension can be reduced from p to $p - 1$, the number of parameters is reduced by $p(p+3)/2 - (p-1)(p+2)/2 = p+1$.

The sample covariance or sample correlation matrices estimate these parameters very efficiently since $\boldsymbol{\Sigma} = (\sigma_{ij})$ where σ_{ij} is a population covariance or correlation. These quantities can be estimated with the sample covariance or correlation taking two variables X_i and X_j at a time. Note that there are $p(p + 1)/2$ pairs that can be chosen from p random variables $X_1, ..., X_p$.

Rule of thumb 7.1. For the classical estimators of multivariate location and dispersion, $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ or $(\overline{\boldsymbol{z}} = \boldsymbol{0}, \boldsymbol{R})$, we want $n \geq 10p$. We want $n \geq 20p$ for the robust MLD estimators (FCH, RFCH, or RMVN) described later in this section.

7.2.1 Affine Equivariance

Before defining an important equivariance property, some notation is needed. Assume that the data is collected in an $n \times p$ data matrix $\boldsymbol{W}$. Let $\boldsymbol{B} = \boldsymbol{1}\boldsymbol{b}^T$ where $\boldsymbol{1}$ is an $n \times 1$ vector of ones and $\boldsymbol{b}$ is a $p \times 1$ constant vector. Hence the ith row of $\boldsymbol{B}$ is $\boldsymbol{b}_i^T \equiv \boldsymbol{b}^T$ for $i = 1, ..., n$. For such a matrix $\boldsymbol{B}$, consider the affine transformation $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}$ where $\boldsymbol{A}$ is any nonsingular $p \times p$ matrix. An affine transformation changes $\boldsymbol{x}_i$ to $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b}$ for $i = 1, ..., n$, and affine equivariant multivariate location and dispersion estimators change in natural ways.

Definition 7.7. The multivariate location and dispersion estimator $(T, \boldsymbol{C})$ is *affine equivariant* if

$$T(\boldsymbol{Z}) = T(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}T(\boldsymbol{W}) + \boldsymbol{b}, \tag{7.6}$$

$$\text{and}\ \ \boldsymbol{C}(\boldsymbol{Z}) = \boldsymbol{C}(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}\boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}^T. \tag{7.7}$$

The following theorem shows that the Mahalanobis distances are invariant under affine transformations. See Rousseeuw and Leroy (1987, pp. 252-262) for similar results. Thus if $(T, \boldsymbol{C})$ is affine equivariant, so is $(T, D^2_{(c_n)}(T, \boldsymbol{C})\ \boldsymbol{C})$ where $D^2_{(j)}(T, \boldsymbol{C})$ is the jth order statistic of the D^2_i.

Theorem 7.1. If $(T, \boldsymbol{C})$ is affine equivariant, then

$$D^2_i(\boldsymbol{W}) \equiv D^2_i(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = D^2_i(T(\boldsymbol{Z}), \boldsymbol{C}(\boldsymbol{Z})) \equiv D^2_i(\boldsymbol{Z}). \tag{7.8}$$

Proof. Since $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}$ has ith row $\boldsymbol{z}_i^T = \boldsymbol{x}_i^T\boldsymbol{A}^T + \boldsymbol{b}^T$,

$$D^2_i(\boldsymbol{Z}) = [\boldsymbol{z}_i - T(\boldsymbol{Z})]^T\boldsymbol{C}^{-1}(\boldsymbol{Z})[\boldsymbol{z}_i - T(\boldsymbol{Z})]$$

$$= [\boldsymbol{A}(\boldsymbol{x}_i - T(\boldsymbol{W}))]^T[\boldsymbol{A}\boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}^T]^{-1}[\boldsymbol{A}(\boldsymbol{x}_i - T(\boldsymbol{W}))]$$

$$= [\boldsymbol{x}_i - T(\boldsymbol{W})]^T\boldsymbol{C}^{-1}(\boldsymbol{W})[\boldsymbol{x}_i - T(\boldsymbol{W})] = D^2_i(\boldsymbol{W}). \ \square$$

Definition 7.8. For MLD, an *elemental set* $J = \{m_1, ..., m_{p+1}\}$ is a set of $p+1$ cases drawn without replacement from the data set of n cases. The elemental fit $(T_J, \boldsymbol{C}_J) = (\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ is the sample mean and the sample covariance matrix computed from the cases in the elemental set.

If the data are iid, then the elemental fit gives an unbiased but inconsistent estimator of $(E(\boldsymbol{x}), \text{Cov}(\boldsymbol{x}))$. Note that the elemental fit uses the smallest sample size $p + 1$ such that $\boldsymbol{S}_J$ is nonsingular if the data are in "general position" defined in Definition 7.10. See Definition 4.7 for the sample mean and sample covariance matrix.

7.2.2 Breakdown

This subsection gives a standard definition of breakdown for estimators of multivariate location and dispersion. The following notation will be useful. Let $\boldsymbol{W}$ denote the $n \times p$ data matrix with ith row $\boldsymbol{x}_i^T$ corresponding to the ith case. Let $\boldsymbol{w}_1, ...\boldsymbol{w}_n$ be the contaminated data after d_n of the $\boldsymbol{x}_i$ have been replaced by arbitrarily bad contaminated cases. Let $\boldsymbol{W}^n_d$ denote the $n \times p$ data matrix with ith row $\boldsymbol{w}_i^T$. Then the contamination fraction is $\gamma_n = d_n/n$. Let $(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ denote an estimator of multivariate location and dispersion where the $p \times 1$ vector $T(\boldsymbol{W})$ is an estimator of location and the $p \times p$ symmetric positive semidefinite matrix $\boldsymbol{C}(\boldsymbol{W})$ is an estimator of dispersion.

Theorem 7.2. Let $\boldsymbol{B} > 0$ be a $p \times p$ symmetric matrix with eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p > 0$ and the orthonormal eigenvectors satisfy $\boldsymbol{e}_i^T \boldsymbol{e}_i = 1$ while $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ for $i \neq j$. Let $\boldsymbol{d}$ be a given $p \times 1$ vector and let $\boldsymbol{a}$ be an arbitrary nonzero $p \times 1$ vector.

a) $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T \boldsymbol{d} \boldsymbol{d}^T \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a}} = \boldsymbol{d}^T \boldsymbol{B}^{-1} \boldsymbol{d}$ where the max is attained for $\boldsymbol{a} = c\boldsymbol{B}^{-1}\boldsymbol{d}$ for any constant $c \neq 0$. Note that the numerator $= (\boldsymbol{a}^T \boldsymbol{d})^2$.

b) $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{a}} = \max_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a} = \lambda_1$ where the max is attained for $\boldsymbol{a} = \boldsymbol{e}_1$.

c) $\min_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{a}} = \min_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a} = \lambda_p$ where the min is attained for $\boldsymbol{a} = \boldsymbol{e}_p$.

d) $\max_{\boldsymbol{a} \perp \boldsymbol{e}_1,...,\boldsymbol{e}_k} \dfrac{\boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{a}} = \max_{\|\boldsymbol{a}\|=1, \boldsymbol{a} \perp \boldsymbol{e}_1,...,\boldsymbol{e}_k} \boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a} = \lambda_{k+1}$ where the max is attained for $\boldsymbol{a} = \boldsymbol{e}_{k+1}$ for $k = 1, 2, ..., p-1$.

e) Let $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ be the observed sample mean and sample covariance matrix where $\boldsymbol{S} > 0$. Then $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{n\boldsymbol{a}^T (\overline{\boldsymbol{x}} - \boldsymbol{\mu})(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{S} \boldsymbol{a}} = n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) = T^2$ where the max is attained for $\boldsymbol{a} = c\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu})$ for any constant $c \neq 0$.

f) Let $\boldsymbol{A}$ be a $p \times p$ symmetric matrix. Let $\boldsymbol{C} > 0$ be a $p \times p$ symmetric matrix. Then $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T \boldsymbol{A} \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{C} \boldsymbol{a}} = \lambda_1(\boldsymbol{C}^{-1}\boldsymbol{A})$, the largest eigenvalue of $\boldsymbol{C}^{-1}\boldsymbol{A}$. The value of $\boldsymbol{a}$ that achieves the max is the eigenvector $\boldsymbol{g}_1$ of $\boldsymbol{C}^{-1}\boldsymbol{A}$ corresponding to $\lambda_1(\boldsymbol{C}^{-1}\boldsymbol{A})$. Similarly $\min_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T \boldsymbol{A} \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{C} \boldsymbol{a}} = \lambda_p(\boldsymbol{C}^{-1}\boldsymbol{A})$, the smallest eigenvalue of $\boldsymbol{C}^{-1}\boldsymbol{A}$. The value of $\boldsymbol{a}$ that achieves the min is the eigenvector $\boldsymbol{g}_p$ of $\boldsymbol{C}^{-1}\boldsymbol{A}$ corresponding to $\lambda_p(\boldsymbol{C}^{-1}\boldsymbol{A})$.

Proof Sketch. See Johnson and Wichern (1988, pp. 64-65, 184). For a), note that rank$(\boldsymbol{C}^{-1}\boldsymbol{A}) = 1$, where $\boldsymbol{C} = \boldsymbol{B}$ and $\boldsymbol{A} = \boldsymbol{d}\boldsymbol{d}^T$, since rank$(\boldsymbol{C}^{-1}\boldsymbol{A})$ = rank$(\boldsymbol{A})$ = rank$(\boldsymbol{d})$ = 1. Hence $\boldsymbol{C}^{-1}\boldsymbol{A}$ has one nonzero eigenvalue eigenvector pair $(\lambda_1, \boldsymbol{g}_1)$. Since

$$(\lambda_1 = \boldsymbol{d}^T \boldsymbol{B}^{-1} \boldsymbol{d}, \boldsymbol{g}_1 = \boldsymbol{B}^{-1}\boldsymbol{d})$$

is a nonzero eigenvalue eigenvector pair for $\boldsymbol{C}^{-1}\boldsymbol{A}$, and $\lambda_1 > 0$, the result follows by f).

Note that b) and c) are special cases of f) with $\boldsymbol{A} = \boldsymbol{B}$ and $\boldsymbol{C} = \boldsymbol{I}$.

Note that e) is a special case of a) with $\boldsymbol{d} = (\overline{\boldsymbol{x}} - \boldsymbol{\mu})$ and $\boldsymbol{B} = \boldsymbol{S}$.

(Also note that $(\lambda_1 = (\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}), \boldsymbol{g}_1 = \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}))$ is a nonzero eigenvalue eigenvector pair for the rank 1 matrix $\boldsymbol{C}^{-1}\boldsymbol{A}$ where $\boldsymbol{C} = \boldsymbol{S}$ and $\boldsymbol{A} = (\overline{\boldsymbol{x}} - \boldsymbol{\mu})(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T$.)

For f), see Mardia et al. (1979, p. 480). □

From Theorem 7.2, if $\boldsymbol{C}(\boldsymbol{W}_d^n) > 0$, then $\max_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{C}(\boldsymbol{W}_d^n)\boldsymbol{a} = \lambda_1$ and $\min_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{C}(\boldsymbol{W}_d^n)\boldsymbol{a} = \lambda_p$. A high breakdown dispersion estimator $\boldsymbol{C}$ is positive definite if the amount of contamination is less than the breakdown value. Since $\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a} = \sum_{i=1}^p \sum_{j=1}^p c_{ij} a_i a_j$, the largest eigenvalue λ_1 is bounded as $\boldsymbol{W}_d^n$ varies iff $\boldsymbol{C}(\boldsymbol{W}_d^n)$ is bounded as $\boldsymbol{W}_d^n$ varies.

Definition 7.9. The *breakdown value* of the multivariate location estimator T at $\boldsymbol{W}$ is

$$B(T, \boldsymbol{W}) = \min \left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty \right\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$ and $1 \leq d_n \leq n$. Let $\lambda_1(\boldsymbol{C}(\boldsymbol{W})) \geq \cdots \geq \lambda_p(\boldsymbol{C}(\boldsymbol{W})) \geq 0$ denote the eigenvalues of the dispersion estimator applied to data $\boldsymbol{W}$. The estimator $\boldsymbol{C}$ breaks down if the smallest eigenvalue can be driven to zero or if the largest eigenvalue can be driven to ∞. Hence the *breakdown value* of the dispersion estimator is

$$B(\boldsymbol{C}, \boldsymbol{W}) = \min \left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \max \left[\frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{W}_d^n))}, \lambda_1(\boldsymbol{C}(\boldsymbol{W}_d^n)) \right] = \infty \right\}.$$

Definition 7.10. Let γ_n be the breakdown value of $(T, \boldsymbol{C})$. *High breakdown (HB) statistics* have $\gamma_n \to 0.5$ as $n \to \infty$ if the (uncontaminated) clean data are in *general position*: no more than p points of the clean data lie on any $(p-1)$-dimensional hyperplane. Estimators are *zero breakdown* if $\gamma_n \to 0$ and *positive breakdown* if $\gamma_n \to \gamma > 0$ as $n \to \infty$.

Note that if the number of outliers is less than the number needed to cause breakdown, then $\|T\|$ is bounded and the eigenvalues are bounded away from 0 and ∞. Also, the bounds do not depend on the outliers but do depend on the estimator $(T, \boldsymbol{C})$ and on the clean data $\boldsymbol{W}$.

The following result shows that 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. For an affine equivariant estimator, the largest possible breakdown value is $n/2$ or $(n+1)/2$ for n even or odd, respectively. Hence in the proof of the following result, we could replace $d_n < d_T$ by $d_n < \min(n/2, d_T)$.

Theorem 7.3. Fix n. If nonequivariant estimators (that may have a breakdown value of greater than 1/2) are excluded, then a multivariate location estimator has a breakdown value of d_T/n iff $d_T = d_{T,n}$ is the smallest number of arbitrarily bad cases that can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|)$ arbitrarily large.

Proof. Suppose the multivariate location estimator T satisfies $\|T(\boldsymbol{W}_d^n)\| \leq M$ for some constant M if $d_n < d_T$. Note that for a fixed data set $\boldsymbol{W}_d^n$ with ith row $\boldsymbol{w}_i$, the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) \leq \max_{i=1,\ldots,n} \|\boldsymbol{x}_i - T(\boldsymbol{W}_d^n)\| \leq \max_{i=1,\ldots,n} \|\boldsymbol{x}_i\| + M$ if $d_n < d_T$. Similarly, suppose $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) \leq M$ for some constant M if $d_n < d_T$, then $\|T(\boldsymbol{W}_d^n)\|$ is bounded if $d_n < d_T$. $\square$

Since the coordinatewise median $\text{MED}(\boldsymbol{W})$ is a HB estimator of multivariate location, it is also true that a multivariate location estimator T will not break down if T can not be driven out of some ball of radius r about $\text{MED}(\boldsymbol{W})$. Hence $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ is a HB estimator of MLD.

If a high breakdown estimator $(T, \boldsymbol{C}) \equiv (T(\boldsymbol{W}_d^n), \boldsymbol{C}(\boldsymbol{W}_d^n))$ is evaluated on the contaminated data $\boldsymbol{W}_d^n$, then the location estimator T is contained in some ball about the origin of radius r, and $0 < a < \lambda_p \leq \lambda_1 < b$ where the constants a, r, and b depend on the clean data and $(T, \boldsymbol{C})$, but not on $\boldsymbol{W}_d^n$ if the number of outliers d_n satisfies $0 \leq d_n < n\gamma_n < n/2$ where the breakdown value $\gamma_n \to 0.5$ as $n \to \infty$.

The following theorem will be used to show that if the classical estimator $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is applied to $c_n \approx n/2$ cases contained in a ball about the origin of radius r where r depends on the clean data but not on $\boldsymbol{W}_d^n$, then $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is a high breakdown estimator.

Theorem 7.4. If the classical estimator $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is applied to c_n cases that are contained in some bounded region where $p + 1 \leq c_n \leq n$, then the maximum eigenvalue λ_1 of $\boldsymbol{S}_B$ is bounded.

Proof. The largest eigenvalue of a $p \times p$ matrix $\boldsymbol{A}$ is bounded above by $p \max |a_{i,j}|$ where $a_{i,j}$ is the (i,j) entry of $\boldsymbol{A}$. See Datta (1995, p. 403). Denote the c_n cases by $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{c_n}$. Then the (i,j)th element $a_{i,j}$ of $\boldsymbol{A} = \boldsymbol{S}_B$ is

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

Hence the maximum eigenvalue λ_1 is bounded. $\square$

The determinant $det(\boldsymbol{S}) = |\boldsymbol{S}|$ of $\boldsymbol{S}$ is known as the *generalized sample variance.* Consider the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq D^2_{(c_n)}\} \tag{7.9}$$

where $D^2_{(c_n)}$ is the c_nth smallest squared Mahalanobis distance based on $(T, \boldsymbol{C})$. This hyperellipsoid contains the c_n cases with the smallest D_i^2. Suppose $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}_M, b\, \boldsymbol{S}_M)$ is the sample mean and scaled sample covariance matrix applied to some subset of the data where $b > 0$. The classical, RFCH, and RMVN estimators satisfy this assumption. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)}.$$

If $h^2 = D^2_{(c_n)}$, then the volume is proportional to the square root of the determinant $|\boldsymbol{S}_M|^{1/2}$, and this volume will be positive unless extreme degeneracy is present among the c_n cases. See Johnson and Wichern (1988, pp. 103-104).

7.2.3 The Concentration Algorithm

Concentration algorithms are widely used since impractical brand name estimators, such as the MCD estimator given in Definition 7.11, take too long to compute. The concentration algorithm, defined in Definition 7.12, use K starts and attractors. A *start* is an initial estimator, and an *attractor* is an estimator obtained by refining the start. For example, let the start be the classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$. Then the attractor could be the classical estimator $(T_1, \boldsymbol{C}_1)$ applied to the half set of cases with the smallest Mahalanobis distances. This concentration algorithm uses one concentration step, but the process could be iterated for k concentration steps, producing an estimator $(T_k, \boldsymbol{C}_k)$

If more than one attractor is used, then some criterion is needed to select which of the K attractors is to be used in the final estimator. If each attractor $(T_{k,j}, \boldsymbol{C}_{k,j})$ is the classical estimator applied to $c_n \approx n/2$ cases, then the minimum covariance determinant (MCD) criterion is often used: choose the attractor that has the minimum value of $det(\boldsymbol{C}_{k,j})$ where $j = 1, ..., K$.

The remainder of this section will explain the concentration algorithm, explain why the MCD criterion is useful but can be improved, provide some theory for practical robust multivariate location and dispersion estimators, and show how the set of cases used to compute the recommended RMVN or RFCH estimator can be used to create outlier resistant regression estimators. The RMVN and RFCH estimators are reweighted versions of the practical FCH estimator, given in Definition 7.15.

Definition 7.11. Consider the subset J_o of $c_n \approx n/2$ observations whose sample covariance matrix has the lowest 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 $(T_{MCD}(\boldsymbol{W}), \boldsymbol{C}_{MCD}(\boldsymbol{W}))$.

Here

$$C(n,i) = \binom{n}{i} = \frac{n!}{i! \ \ (n-i)!}$$

is the binomial coefficient.

The MCD estimator is a high breakdown (HB) estimator, and the value $c_n = \lfloor (n+p+1)/2 \rfloor$ is often used as the default. The MCD estimator is the pair

$$(\hat{\beta}_{LTS}, Q_{LTS}(\hat{\beta}_{LTS})/(c_n - 1))$$

in the location model where LTS stands for the least trimmed sum of squares estimator. See Section 7.6. The population analog of the MCD estimator is closely related to the hyperellipsoid of highest concentration that contains $c_n/n \approx$ half of the mass. The MCD estimator is a $\sqrt{n}$ consistent HB asymptotically normal estimator for $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ where a_{MCD} is some positive constant when the data $\boldsymbol{x}_i$ are iid from a large class of distributions. See Cator and Lopuhaä (2010, 2012) who extended some results of Butler et al. (1993).

Computing robust covariance estimators can be very expensive. For example, to compute the exact MCD(c_n) estimator (T_{MCD}, C_{MCD}), we need to consider the $C(n, c_n)$ subsets of size c_n. Woodruff and Rocke (1994, p. 893) noted that if 1 billion subsets of size 101 could be evaluated per second, it would require 10^{33} millenia to search through all $C(200, 101)$ subsets if the sample size $n = 200$. See Section 7.8 for the MCD complexity.

Hence algorithm estimators will be used to approximate the robust estimators. Elemental sets are the key ingredient for both *basic resampling* and *concentration* algorithms.

Definition 7.12. Suppose that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are $p \times 1$ vectors of observed data. For the multivariate location and dispersion model, an *elemental set* J is a set of $p+1$ cases. An elemental start is the sample mean and sample covariance matrix of the data corresponding to J. In a *concentration algorithm,* let $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ be the jth start (not necessarily elemental) and compute all n Mahalanobis distances $D_i(T_{-1,j}, \boldsymbol{C}_{-1,j})$. At the next iteration, the classical estimator $(T_{0,j}, \boldsymbol{C}_{0,j}) = (\overline{\boldsymbol{x}}_{0,j}, \boldsymbol{S}_{0,j})$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest distances. This iteration can be continued for k *concentration steps* resulting in the sequence of estimators $(T_{-1,j}, \boldsymbol{C}_{-1,j}), (T_{0,j}, \boldsymbol{C}_{0,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. The result of the iteration $(T_{k,j}, \boldsymbol{C}_{k,j})$ is called the jth *attractor.* If K_n starts are used, then $j = 1, ..., K_n$. The *concentration attractor*, $(T_A, \boldsymbol{C}_A)$, is the attractor chosen by the algorithm. The attractor is used to obtain the final estimator. A common choice is the attractor that has the smallest determinant $det(\boldsymbol{C}_{k,j})$. The *basic resampling algorithm* estimator is a special case where $k = -1$ so that the attractor is the start: $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j}) = (\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$.

This concentration algorithm is a simplified version of the algorithms given by Rousseeuw and Van Driessen (1999) and Hawkins and Olive (1999a). Using

$k = 10$ concentration steps often works well. The following proposition is useful and shows that $det(\boldsymbol{S}_{0,j})$ tends to be greater than the determinant of the attractor $det(\boldsymbol{S}_{k,j})$.

Theorem 7.5: Rousseeuw and Van Driessen (1999, p. 214). Suppose that the classical estimator $(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ is computed from c_n cases and that the n Mahalanobis distances $D_i \equiv D_i(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ are computed. If $(\overline{\boldsymbol{x}}_{t+1,j}, \boldsymbol{S}_{t+1,j})$ is the classical estimator computed from the c_n cases with the smallest Mahalanobis distances D_i, then $det(\boldsymbol{S}_{t+1,j}) \leq det(\boldsymbol{S}_{t,j})$ with equality iff $(\overline{\boldsymbol{x}}_{t+1,j}, \boldsymbol{S}_{t+1,j}) = (\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$.

Starts that use a consistent initial estimator could be used. K_n is the number of starts and k is the number of concentration steps used in the algorithm. Suppose the algorithm estimator uses some criterion to choose an attractor as the final estimator where there are K attractors and K is fixed, e.g. $K = 500$, so K does not depend on n. A crucial observation is that the theory of the algorithm estimator depends on the theory of the attractors, not on the estimator corresponding to the criterion.

For example, let $(\boldsymbol{0}, \boldsymbol{I}_p)$ and $(\boldsymbol{1}, diag(1, 3, ..., p))$ be the high breakdown attractors where $\boldsymbol{0}$ and $\boldsymbol{1}$ are the $p \times 1$ vectors of zeroes and ones. If the minimum determinant criterion is used, then the final estimator is $(\boldsymbol{0}, \boldsymbol{I}_p)$. Although the MCD criterion is used, the algorithm estimator does not have the same properties as the MCD estimator.

Hawkins and Olive (2002) showed that if K randomly selected elemental starts are used with concentration to produce the attractors, then the resulting estimator is inconsistent and zero breakdown if K and k are fixed and free of n. Note that each elemental start can be made to breakdown by changing one case. Hence the breakdown value of the final estimator is bounded by $K/n \to 0$ as $n \to \infty$. Note that the classical estimator computed from h_n randomly drawn cases is an inconsistent estimator unless $h_n \to \infty$ as $n \to \infty$. Thus the classical estimator applied to a randomly drawn elemental set of $h_n \equiv p + 1$ cases is an inconsistent estimator, so the K starts and the K attractors are inconsistent.

This theory shows that the Maronna et al. (2006, pp. 198-199) estimators that use $K = 500$ and one concentration step ($k = 0$) are inconsistent and zero breakdown. The following theorem is useful because it does not depend on the criterion used to choose the attractor.

Suppose there are K consistent estimators $(T_j, \boldsymbol{C}_j)$ of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ for some constant $a > 0$, each with the same rate n^δ. If $(T_A, \boldsymbol{C}_A)$ is an estimator obtained by choosing one of the K estimators, then $(T_A, \boldsymbol{C}_A)$ is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with rate n^δ by Pratt (1959). See Theorem 1.21.

Theorem 7.6. Suppose the algorithm estimator chooses an attractor as the final estimator where there are K attractors and K is fixed.

i) If all of the attractors are consistent estimators of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$, then the algorithm estimator is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$.

ii) If all of the attractors are consistent estimators of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with the same rate, e.g. n^δ where $0 < \delta \leq 0.5$, then the algorithm estimator is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with the same rate as the attractors.

iii) If all of the attractors are high breakdown, then the algorithm estimator is high breakdown.

iv) Suppose the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid and $P(\boldsymbol{x}_i = \boldsymbol{\mu}) < 1$. The elemental basic resampling algorithm estimator ($k = -1$) is inconsistent.

v) The elemental concentration algorithm is zero breakdown.

Proof. i) Choosing from K consistent estimators for $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ results in a consistent estimator for of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$, and ii) follows from Pratt (1959). iii) Let $\gamma_{n,i}$ be the breakdown value of the ith attractor if the clean data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are in general position. The breakdown value γ_n of the algorithm estimator can be no lower than that of the worst attractor: $\gamma_n \geq \min(\gamma_{n,1}, ..., \gamma_{n,K}) \to 0.5$ as $n \to \infty$.

iv) Let $(\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$ be the classical estimator applied to a randomly drawn elemental set. Then $\overline{\boldsymbol{x}}_{-1,j}$ is the sample mean applied to $p+1$ iid cases. Hence $E(\boldsymbol{S}_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$, $E[\overline{\boldsymbol{x}}_{-1,j}] = E(\boldsymbol{x}) = \boldsymbol{\mu}$, and $\mathrm{Cov}(\overline{\boldsymbol{x}}_{-1,j}) = \mathrm{Cov}(\boldsymbol{x})/(p+1) = \boldsymbol{\Sigma}_{\boldsymbol{x}}/(p+1)$ assuming second moments. So the $(\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$ are identically distributed and inconsistent estimators of $(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Even without second moments, there exists $\epsilon > 0$ such that $P(\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = \delta_\epsilon > 0$ where the probability, ϵ, and δ_ϵ do not depend on n since the distribution of $\overline{\boldsymbol{x}}_{-1,j}$ only depends on the distribution of the iid $\boldsymbol{x}_i$, not on n. Then $P(\min_j \|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = P(\text{all } \ \|\overline{\boldsymbol{x}}_{-1,\mathrm{j}} - \boldsymbol{\mu}\| > \epsilon) \to \delta_\epsilon^{\mathrm{K}} > 0$ as $n \to \infty$ where equality would hold if the $\overline{\boldsymbol{x}}_{-1,j}$ were iid. Hence the "best start" that minimizes $\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\|$ is inconsistent.

v) The classical estimator with breakdown $1/n$ is applied to each elemental start. Hence $\gamma_n \leq K/n \to 0$ as $n \to \infty$. □

Since the FMCD estimator is a zero breakdown elemental concentration algorithm, the Hubert et al. (2008) claim that "MCD can be efficiently computed with the FAST-MCD estimator" is false. Suppose K is fixed, but at least one randomly drawn start is iterated to convergence so that k is not fixed. Then it is not known whether the attractors are inconsistent or consistent estimators, so it is not known whether FMCD is consistent. It is possible to produce consistent estimators if $K \equiv K_n$ is allowed to increase to ∞.

Remark 7.1. Let γ_o be the highest percentage of large outliers that an elemental concentration algorithm can detect reliably. For many data sets,

$$\gamma_o \approx \min\left(\frac{n - c_n}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h}\right) 100\% \tag{7.10}$$

if n is large, $c_n \geq n/2$ and $h = p + 1$.

Proof. Suppose that the data set contains n cases with d outliers and $n - d$ clean cases. Suppose K elemental sets are chosen with replacement.

If W_i is the number of outliers in the ith elemental set, then the W_i are iid hypergeometric$(d, n-d, h)$ random variables. Suppose that it is desired to find K such that the probability P(that at least one of the elemental sets is clean) $\equiv P_1 \approx 1-\alpha$ where $0 < \alpha < 1$. Then $P_1 = 1-$ P(none of the K elemental sets is clean) $\approx 1-[1-(1-\gamma)^h]^K$ by independence. If the contamination proportion γ is fixed, then the probability of obtaining at least one clean subset of size h with high probability (say $1-\alpha = 0.8$) is given by $0.8 = 1-[1-(1-\gamma)^h]^K$. Fix the number of starts K and solve this equation for γ. □

7.2.4 Theory for Practical Estimators

It is convenient to let the $\boldsymbol{x}_i$ be random vectors for large sample theory, but the $\boldsymbol{x}_i$ are fixed clean observed data vectors when discussing breakdown. This subsection presents the FCH estimator to be used along with the classical estimator. Recall from Definition 7.12 that a *concentration algorithm* uses K_n *starts* $(T_{-1,j}, \boldsymbol{C}_{-1,j})$. After finding $(T_{0,j}, \boldsymbol{C}_{0,j})$, each start is refined with k concentration steps, resulting in K_n *attractors* $(T_{k,j}, \boldsymbol{C}_{k,j})$, and the concentration attractor $(T_A, \boldsymbol{C}_A)$ is the attractor that optimizes the criterion.

Concentration algorithms include the *basic resampling algorithm* as a special case with $k = -1$. Using $k = 10$ concentration steps works well, and iterating until convergence is usually fast. The DGK estimator (Devlin et al. 1975, 1981) defined below is one example. The DGK estimator is affine equivariant since the classical estimator is affine equivariant and Mahalanobis distances are invariant under affine transformations by Theorem 7.1. This subsection will show that the Olive (2004a) MB estimator is a high breakdown estimator and that the DGK estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$, the same quantity estimated by the MCD estimator. Both estimators use the classical estimator computed from $c_n \approx n/2$ cases. The breakdown point of the DGK estimator has been conjectured to be "at most $1/p$." See Rousseeuw and Leroy (1987, p. 254).

Definition 7.13. The *DGK estimator* $(T_{k,D}, \boldsymbol{C}_{k,D}) = (T_{DGK}, \boldsymbol{C}_{DGK})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start.

Definition 7.14. The *median ball (MB) estimator* $(T_{k,M}, \boldsymbol{C}_{k,M}) = (T_{MB}, \boldsymbol{C}_{MB})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ as the only start where MED$(\boldsymbol{W})$ is the coordinatewise median. So $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of data closest to MED$(\boldsymbol{W})$ in Euclidean distance.

The proof of the following theorem implies that a high breakdown estimator $(T, \boldsymbol{C})$ has MED$(D_i^2) \le V$ and that the hyperellipsoid $\{\boldsymbol{x} | D^2_{\boldsymbol{x}} \le D^2_{(c_n)}\}$

that contains $c_n \approx n/2$ of the cases is in some ball about the origin of radius r, where V and r do not depend on the outliers even if the number of outliers is close to $n/2$. Also the attractor of a high breakdown estimator is a high breakdown estimator if the number of concentration steps k is fixed, e.g. $k = 10$. The theorem implies that the MB estimator $(T_{MB}, \boldsymbol{C}_{MB})$ is high breakdown.

Theorem 7.7. Suppose $(T, \boldsymbol{C})$ is a high breakdown estimator where $\boldsymbol{C}$ is a symmetric, positive definite $p \times p$ matrix if the contamination proportion d_n/n is less than the breakdown value. Then the concentration attractor $(T_k, \boldsymbol{C}_k)$ is a high breakdown estimator if the coverage $c_n \approx n/2$ and the data are in general position.

Proof. Following Leon (1986, p. 280), if $\boldsymbol{A}$ is a symmetric positive definite matrix with eigenvalues $\tau_1 \geq \cdots \geq \tau_p$, then for any nonzero vector $\boldsymbol{x}$,

$$0 < \|\boldsymbol{x}\|^2 \ \tau_p \leq \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \leq \|\boldsymbol{x}\|^2 \ \tau_1. \tag{7.11}$$

Let $\lambda_1 \geq \cdots \geq \lambda_p$ be the eigenvalues of $\boldsymbol{C}$. By (7.11),

$$\frac{1}{\lambda_1}\|\boldsymbol{x} - T\|^2 \leq (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq \frac{1}{\lambda_p}\|\boldsymbol{x} - T\|^2. \tag{7.12}$$

By (7.12), if the $D^2_{(i)}$ are the order statistics of the $D^2_i(T, \boldsymbol{C})$, then $D^2_{(i)} < V$ for some constant V that depends on the clean data but not on the outliers even if i and d_n are near $n/2$. (Note that $1/\lambda_p$ and $\text{MED}(\|\boldsymbol{x}_i - T\|^2)$ are both bounded for high breakdown estimators even for d_n near $n/2$.)

Following Johnson and Wichern (1988, pp. 50, 103), the boundary of the set $\{\boldsymbol{x}|D^2_{\boldsymbol{x}} \leq h^2\} = \{\boldsymbol{x}|(\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq h^2\}$ is a hyperellipsoid centered at T with axes of length $2h\sqrt{\lambda_i}$. Hence $\{\boldsymbol{x}|D^2_{\boldsymbol{x}} \leq D^2_{(c_n)}\}$ is contained in some ball about the origin of radius r where r does not depend on the number of outliers even for d_n near $n/2$. This is the set containing the cases used to compute $(T_0, \boldsymbol{C}_0)$. Since the set is bounded, T_0 is bounded and the largest eigenvalue $\lambda_{1,0}$ of $\boldsymbol{C}_0$ is bounded by Theorem 7.4. The determinant $det(\boldsymbol{C}_{MCD})$ of the HB minimum covariance determinant estimator satisfies $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_0) = \lambda_{1,0} \cdots \lambda_{p,0}$, and $\lambda_{p,0} > \inf det(\boldsymbol{C}_{MCD})/\lambda^{p-1}_{1,0} > 0$ where the infimum is over all possible data sets with $n - d_n$ clean cases and d_n outliers. Since these bounds do not depend on the outliers even for d_n near $n/2$, $(T_0, \boldsymbol{C}_0)$ is a high breakdown estimator. Now repeat the argument with $(T_0, \boldsymbol{C}_0)$ in place of $(T, \boldsymbol{C})$ and $(T_1, \boldsymbol{C}_1)$ in place of $(T_0, \boldsymbol{C}_0)$. Then $(T_1, \boldsymbol{C}_1)$ is high breakdown. Repeating the argument iteratively shows $(T_k, \boldsymbol{C}_k)$ is high breakdown. □

The following corollary shows that it is easy to find a subset J of $c_n \approx n/2$ cases such that the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to J is a HB estimator of MLD.

Theorem 7.8. Let J consist of the c_n cases $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|)$. Then the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to J is a HB estimator of MLD.

To investigate the consistency and rate of robust estimators of multivariate location and dispersion, review Definitions 1.34 and 1.35.

The following assumption (E1) gives a class of distributions where we can prove that the new robust estimators are $\sqrt{n}$ consistent. Cator and Lopuhaä (2010, 2012) showed that MCD is consistent provided that the MCD functional is unique. Distributions where the functional is unique are called "unimodal," and rule out, for example, a spherically symmetric uniform distribution. Theorem 7.9 is crucial for theory and Theorem 7.10 shows that under (E1), both MCD and DGK are estimating $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$.

Assumption (E1): The $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from a "unimodal" elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with nonsingular covariance matrix $\text{Cov}(\boldsymbol{x}_i)$ where g is continuously differentiable with finite 4th moment: $\int (\boldsymbol{x}^T\boldsymbol{x})^2 g(\boldsymbol{x}^T\boldsymbol{x}) d\boldsymbol{x} < \infty$.

Lopuhaä (1999) showed that if a start $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$, then the classical estimator applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where $a, s > 0$ are some constants. Affine equivariance is not used for $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Also, the attractor and the start have the same rate. If the start is inconsistent, then so is the attractor. The weight function $I(D_i^2(T, \boldsymbol{C}) \leq h^2)$ is an indicator that is 1 if $D_i^2(T, \boldsymbol{C}) \leq h^2$ and 0 otherwise.

Theorem 7.9, Lopuhaä (1999). Assume the number of concentration steps k is fixed. a) If the start $(T, \boldsymbol{C})$ is inconsistent, then so is the attractor.

b) Suppose $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{I}_p)$ with rate n^δ where $s > 0$ and $0 < \delta \leq 0.5$. Assume (E1) holds and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Then the classical estimator $(T_0, \boldsymbol{C}_0)$ applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{I}_p)$ with the same rate n^δ where $a > 0$.

c) Suppose $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where $s > 0$ and $0 < \delta \leq 0.5$. Assume (E1) holds. Then the classical estimator $(T_0, \boldsymbol{C}_0)$ applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with the same rate n^δ where $a > 0$. The constant a depends on the positive constants s, h, p, and the elliptically contoured distribution, but does not otherwise depend on the consistent start $(T, \boldsymbol{C})$.

Let $\delta = 0.5$. Applying Theorem 7.9c) iteratively for a fixed number k of steps produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_j\boldsymbol{\Sigma})$ where the constants $a_j > 0$ depend on s, h, p, and the elliptically contoured distribution, but do not otherwise depend on the consistent start $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$.

The 4th moment assumption was used to simplify theory, but likely holds under 2nd moments. Affine equivariance is needed so that the attractor is affine equivariant, but probably is not needed to prove consistency.

Conjecture 7.1. Change the finite 4th moments assumption to a finite 2nd moments in assumption E1). Suppose $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^{δ} where $s > 0$ and $0 < \delta \leq 0.5$. Then the classical estimator applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with the same rate n^{δ} where $a > 0$.

Remark 7.2. To see that the Lopuhaä (1999) theory extends to concentration where the weight function uses $h^2 = D^2_{(c_n)}(T, \boldsymbol{C})$, note that $(T, \tilde{\boldsymbol{C}}) \equiv (T, D^2_{(c_n)}(T, \boldsymbol{C})\ \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$ where $b > 0$ is derived in (7.14), and weight function $I(D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1)$ is equivalent to the concentration weight function $I(D_i^2(T, \boldsymbol{C}) \leq D^2_{(c_n)}(T, \boldsymbol{C}))$. As noted above Theorem 7.1, $(T, \tilde{\boldsymbol{C}})$ is affine equivariant if $(T, \boldsymbol{C})$ is affine equivariant. Hence Lopuhaä (1999) theory applied to $(T, \tilde{\boldsymbol{C}})$ with $h = 1$ is equivalent to theory applied to affine equivariant $(T, \boldsymbol{C})$ with $h^2 = D^2_{(c_n)}(T, \boldsymbol{C})$.

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^{δ} where $0 < \delta \leq 0.5$, then $D^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) =$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T[\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$$

$$= s^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta}). \tag{7.13}$$

Thus the sample percentiles of $D_i^2(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $s^{-1}D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose $c_n/n \to \xi \in (0, 1)$ as $n \to \infty$, and let $D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the 100ξth percentile of the population squared distances. Then $D^2_{(c_n)}(T, \boldsymbol{C}) \xrightarrow{P} s^{-1}D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $b\boldsymbol{\Sigma} = s^{-1}D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})s\boldsymbol{\Sigma} = D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})\boldsymbol{\Sigma}$. Thus

$$b = D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{7.14}$$

does not depend on $s > 0$ or $\delta \in (0, 0.5]$. □

Concentration applies the classical estimator to cases with $D_i^2(T, \boldsymbol{C}) \leq D^2_{(c_n)}(T, \boldsymbol{C})$. Let $c_n \approx n/2$ and

$$b = D^2_{0.5}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

be the population median of the population squared distances. By Remark 7.2, if $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ then $(T, \tilde{\boldsymbol{C}}) \equiv (T, D^2_{(c_n)}(T, \boldsymbol{C})\ \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$, and $D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1$ is equivalent to $D_i^2(T, \boldsymbol{C}) \leq D^2_{(c_n)}(T, \boldsymbol{C})$). Hence Lopuhaä (1999) theory applied to $(T, \tilde{\boldsymbol{C}})$ with $h = 1$ is equivalent to theory applied to the concentration estimator using the affine equivariant estimator $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$ as the start. Since b does not depend on s, concentration produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where the constant $a > 0$ is the same for $j = 0, 1, ..., k$.

Theorem 7.10 shows that $a = a_{MCD}$ where $\xi = 0.5$. Hence concentration with a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ as a start results in a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with rate n^δ. This result can be applied iteratively for a finite number of concentration steps. Hence DGK is a $\sqrt{n}$ consistent affine equivariant estimator of the same quantity that MCD is estimating. It is not known if the results hold if concentration is iterated to convergence. For multivariate normal data, $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi^2_p$.

Theorem 7.10. Assume that (E1) holds and that $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where the constants $s > 0$ and $0 < \delta \leq 0.5$. Then the classical estimator $(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ computed from the $c_n \approx n/2$ of cases with the smallest distances $D_i(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with the same rate n^δ.

Proof. By Remark 7.2 the estimator is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with rate n^δ. By the remarks above, a will be the same for any consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ and a does not depend on $s > 0$ or $\delta \in (0, 0.5]$. Hence the result follows if $a = a_{MCD}$. The MCD estimator is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ by Cator and Lopuhaä (2010, 2012). If the MCD estimator is the start, then it is also the attractor by Theorem 7.5 which shows that concentration does not increase the MCD criterion. Hence $a = a_{MCD}$. □

Next we define the easily computed robust $\sqrt{n}$ consistent FCH estimator, so named since it is fast, consistent, and uses a high breakdown attractor. The FCH and MBA estimators use the $\sqrt{n}$ consistent DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$ and the high breakdown MB estimator $(T_{MB}, \boldsymbol{C}_{MB})$ as attractors.

Definition 7.15. Let the "median ball" be the hypersphere containing the "half set" of data closest to MED$(\boldsymbol{W})$ in Euclidean distance. The *FCH estimator* uses the MB attractor if the DGK location estimator T_{DGK} is outside of the median ball, and the attractor with the smallest determinant, otherwise. Let $(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_{FCH}, \boldsymbol{C}_{FCH})$ takes $T_{FCH} = T_A$ and

$$\boldsymbol{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{7.15}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with p degrees of freedom.

Remark 7.3. The *MBA estimator* $(T_{MBA}, \boldsymbol{C}_{MBA})$ uses the attractor $(T_A, \boldsymbol{C}_A)$ with the smallest determinant. Hence the DGK estimator is used as the attractor if $det(\boldsymbol{C}_{DGK}) \leq det(\boldsymbol{C}_{MB})$, and the MB estimator is used as the attractor, otherwise. Then $T_{MBA} = T_A$ and $\boldsymbol{C}_{MBA}$ is computed using the right hand side of (7.15). The difference between the FCH and MBA estimators is that the FCH estimator also uses a location criterion to choose the attractor: if the DGK location estimator T_{DGK} has a greater Euclidean distance from MED($\boldsymbol{W}$) than half the data, then FCH uses the MB attractor. The FCH estimator only uses the attractor with the smallest determinant if $\|T_{DGK} - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p))$. Using the location criterion increases the outlier resistance of the FCH estimator for certain types of outliers, as will be seen in Section 7.2.5.

The following theorem shows the FCH estimator has good statistical properties. We conjecture that FCH is high breakdown. Note that the location estimator T_{FCH} is high breakdown and that $det(\boldsymbol{C}_{FCH})$ is bounded away from 0 and ∞ if the data is in general position, even if nearly half of the cases are outliers.

Theorem 7.11. T_{FCH} is high breakdown if the clean data are in general position. Suppose (E1) holds. If $(T_A, \boldsymbol{C}_A)$ is the DGK or MB attractor with the smallest determinant, then $(T_A, \boldsymbol{C}_A)$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. Hence the MBA and FCH estimators are outlier resistant $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c = u_{0.5}/\chi^2_{p,0.5}$, and $c = 1$ for multivariate normal data.

Proof. T_{FCH} is high breakdown since it is a bounded distance from MED($\boldsymbol{W}$) even if the number of outliers is close to $n/2$. Under (E1) the FCH and MBA estimators are asymptotically equivalent since $\|T_{DGK} - \text{MED}(\boldsymbol{W})\| \to 0$ in probability. The estimator satisfies $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_A) \leq det(\boldsymbol{C}_{0,M}) < \infty$ by Theorem 7.7 if up to nearly 50% of the cases are outliers. If the distribution is spherical about $\boldsymbol{\mu}$, then the result follows from Pratt (1959) and Theorem 7.5 since both starts are $\sqrt{n}$ consistent. Otherwise, the MB estimator $\boldsymbol{C}_{MB}$ is a biased estimator of $a_{MCD}\boldsymbol{\Sigma}$. But the DGK estimator $\boldsymbol{C}_{DGK}$ is a $\sqrt{n}$ consistent estimator of $a_{MCD}\boldsymbol{\Sigma}$ by Theorem 7.10 and $\|\boldsymbol{C}_{MCD} - \boldsymbol{C}_{DGK}\| = O_P(n^{-1/2})$. Thus the probability that the DGK attractor minimizes the determinant goes to one as $n \to \infty$, and $(T_A, \boldsymbol{C}_A)$ is asymptotically equivalent to the DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$.

Let $\boldsymbol{C}_F = \boldsymbol{C}_{FCH}$ or $\boldsymbol{C}_F = \boldsymbol{C}_{MBA}$. Let $P(U \leq u_\alpha) = \alpha$ where U is given by (1.35). Then the scaling in (7.15) makes $\boldsymbol{C}_F$ a consistent estimator of $c\boldsymbol{\Sigma}$ where $c = u_{0.5}/\chi^2_{p,0.5}$, and $c = 1$ for multivariate normal data. □

A standard method of reweighting can be used to produce the RMBA and RFCH estimators. RMVN uses a slightly modified method of reweighting so that RMVN gives good estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for multivariate normal data, even when certain types of outliers are present.

Definition 7.16. The *RFCH estimator* uses two standard reweighting steps. Let $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ be the classical estimator applied to the n_1 cases with $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH}) \le \chi^2_{p,0.975}$, and let

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\text{MED}(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1))}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_1.$$

Then let $(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2)$ be the classical estimator applied to the cases with $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1) \le \chi^2_{p,0.975}$, and let

$$\boldsymbol{C}_{RFCH} = \frac{\text{MED}(D_i^2(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2))}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_2.$$

RMBA and RFCH are $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ by Lopuhaä (1999) where the weight function uses $h^2 = \chi^2_{p,0.975}$, but the two estimators use nearly 97.5% of the cases if the data is multivariate normal.

Definition 7.17. The *RMVN estimator* uses $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ and n_1 as above. Let $q_1 = \min\{0.5(0.975)n/n_1, 0.995\}$, and

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\text{MED}(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1))}{\chi^2_{p,q_1}} \tilde{\boldsymbol{\Sigma}}_1.$$

Then let $(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2)$ be the classical estimator applied to the n_2 cases with $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1)) \le \chi^2_{p,0.975}$. Let $q_2 = \min\{0.5(0.975)n/n_2, 0.995\}$, and

$$\boldsymbol{C}_{RMVN} = \frac{\text{MED}(D_i^2(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2))}{\chi^2_{p,q_2}} \tilde{\boldsymbol{\Sigma}}_2.$$

Definition 7.18. Let the n_2 cases in Definition 7.17 be known as the *RMVN set U*. Hence $(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2) = (\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ is the classical estimator applied to the RMVN set U, which can be regarded as the untrimmed data (the data not trimmed by ellipsoidal trimming) or the cleaned data. Also $\boldsymbol{S}_U$ is the unscaled estimated dispersion matrix while $\boldsymbol{C}_{RMVN}$ is the scaled estimated dispersion matrix.

Remark 7.4. Classical methods can be applied to the RMVN subset U to make robust methods. Under (E1), $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c_U\boldsymbol{\Sigma})$ for some constant $c_U > 0$ that depends on the underlying distribution of the iid $\boldsymbol{x}_i$. For a general estimator of multivariate location and dispersion $(T_A, \boldsymbol{C}_A)$, typically a reweight for efficiency step is performed, resulting in a set U such that the classical estimator $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ is the classical estimator applied to a set U. For example, use $U = \{\boldsymbol{x}_i | D_i^2(T_A, \boldsymbol{C}_A) \le \chi^2_{p,0.975}\}$. Then the final estimator is $(T_F, \boldsymbol{C}_F) = (\overline{\boldsymbol{x}}_U, a\boldsymbol{S}_U)$ where scaling is done as in Equation (7.15) in an attempt to make $\boldsymbol{C}_F$ a good estimator of $\boldsymbol{\Sigma}$ if the iid data are from a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Then $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ can be shown to be a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c_U\boldsymbol{\Sigma})$ for a large class of distributions for the RMVN set, for the RFCH set, or if $(T_A, \boldsymbol{C}_A)$ is an affine equivariant $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c_A\boldsymbol{\Sigma})$ on a large class of distributions. The necessary theory is not yet available for other practical robust reweighted estimators such as OGK and Det-MCD.

Table 7.1 Average Dispersion Matrices for Near Point Mass Outliers

RMVN	FMCD	OGK	MB
$\begin{bmatrix} 1.002 & -0.014 \\ -0.014 & 2.024 \end{bmatrix}$	$\begin{bmatrix} 0.055 & 0.685 \\ 0.685 & 122.5 \end{bmatrix}$	$\begin{bmatrix} 0.185 & 0.089 \\ 0.089 & 36.24 \end{bmatrix}$	$\begin{bmatrix} 2.570 & -0.082 \\ -0.082 & 5.241 \end{bmatrix}$

Table 7.2 Average Dispersion Matrices for Mean Shift Outliers

RMVN	FMCD	OGK	MB
$\begin{bmatrix} 0.990 & 0.004 \\ 0.004 & 2.014 \end{bmatrix}$	$\begin{bmatrix} 2.530 & 0.003 \\ 0.003 & 5.146 \end{bmatrix}$	$\begin{bmatrix} 19.67 & 12.88 \\ 12.88 & 39.72 \end{bmatrix}$	$\begin{bmatrix} 2.552 & 0.003 \\ 0.003 & 5.118 \end{bmatrix}$

The RMVN estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ by Lopuhaä (1999) where the weight function uses $h^2 = \chi^2_{p,0.975}$ and $d = u_{0.5}/\chi^2_{p,q}$ where $q_2 \to q$ in probability as $n \to \infty$. Here $0.5 \le q < 1$ depends on the elliptically contoured distribution, but $q = 0.5$ and $d = 1$ for multivariate normal data.

If the bulk of the data is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the RMVN estimator can give useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for certain types of outliers where FCH and RFCH estimate $(\boldsymbol{\mu}, d_E\boldsymbol{\Sigma})$ for $d_E > 1$. To see this claim, let $0 \le \gamma < 0.5$ be the outlier proportion. If $\gamma = 0$, then $n_i/n \xrightarrow{P} 0.975$ and $q_i \xrightarrow{P} 0.5$. If $\gamma > 0$, suppose the outlier configuration is such that the $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH})$ are roughly χ^2_p for the clean cases, and the outliers have larger D_i^2 than the clean cases. Then $\text{MED}(D_i^2) \approx \chi^2_{p,q}$ where $q = 0.5/(1-\gamma)$. For example, if $n = 100$ and $\gamma = 0.4$, then there are 60 clean cases, $q = 5/6$, and the quantile $\chi^2_{p,q}$ is being estimated instead of $\chi^2_{p,0.5}$. Now $n_i \approx n(1-\gamma)0.975$, and q_i estimates q. Thus $\boldsymbol{C}_{RMVN} \approx \boldsymbol{\Sigma}$. Of course consistency cannot generally be claimed when outliers are present.

Simulations suggested $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ gives useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a variety of outlier configurations. Using 20 runs and $n = 1000$, the averages of the dispersion matrices were computed when the bulk of the data are iid $N_2(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = diag(1, 2)$. For clean data, FCH, RFCH, and RMVN give $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}$, while FMCD and the Maronna and Zamar (2002) OGK estimator seem to be approximately unbiased for $\boldsymbol{\Sigma}$. The median ball estimator was scaled using (7.15) and estimated $diag(1.13, 1.85)$.

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_2((0, 15)^T, 0.0001\boldsymbol{I}_2)$, a near point mass at the major axis. FCH, MB, and RFCH estimated $2.6\boldsymbol{\Sigma}$

while RMVN estimated $\boldsymbol{\Sigma}$. FMCD and OGK failed to estimate $d\ \boldsymbol{\Sigma}$. Note that $\chi^2_{2,5/6}/\chi^2_{2,0.5} = 2.585$. See Table 7.1. The following R commands were used where `mldsim` is from *linmodpack*.

```
qchisq(5/6,2)/qchisq(.5,2) = 2.584963
mldsim(n=1000,p=2,outliers=6,pm=15)
```

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_2((20,20)^T, \boldsymbol{\Sigma})$, a mean shift with the same covariance matrix as the clean cases. Rocke and Woodruff (1996) suggest that outliers with mean shift are hard to detect. FCH, FMCD, MB, and RFCH estimated $2.6\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$, and OGK failed. See Table 7.2. The *R command* is shown below.

```
mldsim(n=1000,p=2,outliers=3,pm=20)
```

Remark 7.5. The RFCH and RMVN estimators are recommended. If these estimators are too slow and outlier detection is of interest, try the RMB estimator, the reweighted MB estimator. If RMB is too slow or if $n < 2(p+1)$, the Euclidean distances $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I})$ of $\boldsymbol{x}_i$ from the coordinatewise median $\text{MED}(\boldsymbol{W})$ may be useful. A DD plot of $D_i(\overline{\boldsymbol{x}}, \boldsymbol{I})$ versus $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I})$ is also useful for outlier detection and for whether $\overline{\boldsymbol{x}}$ and $\text{MED}(\boldsymbol{W})$ are giving similar estimates of multivariate location. Also see Section 7.3.

Hubert et al. (2008, 2012) claim that FMCD computes the MCD estimator. This claim is trivially shown to be false in the following theorem.

Theorem 7.12. Neither FMCD nor Det-MCD compute the MCD estimator.

Proof. A necessary condition for an estimator to be the MCD estimator is that the determinant of the covariance matrix for the estimator be the smallest for every run in a simulation. Sometimes FMCD had the smaller determinant and sometimes Det-MCD had the smaller determinant in the simulations done by Hubert et al. (2012). □

Example 7.2. Tremearne (1911) recorded *height* = x[,1] and *height while kneeling* = x[,2] of 112 people. Figure 7.1a shows a scatterplot of the data. Case 3 has the largest Euclidean distance of 214.767 from $\text{MED}(\boldsymbol{W}) = (1680, 1240)^T$, but if the distances correspond to the contours of a covering ellipsoid, then case 44 has the largest distance. For $k = 0$, $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of cases closest to $\text{MED}(\boldsymbol{W})$ in Euclidean distance. The hypersphere (circle) centered at $\text{MED}(\boldsymbol{W})$ that covers half the data is small because the data density is high near $\text{MED}(\boldsymbol{W})$. The median Euclidean distance is 59.661 and case 44 has Euclidean distance 77.987. Hence the intersection of the sphere and the data is a highly correlated clean ellipsoidal region. Figure 7.1b shows the DD plot of the classical distances versus the MB distances. Notice that both the classical and MB estimators give the largest distances to cases 3 and 44. Notice that case 44 could not be detected using marginal methods.

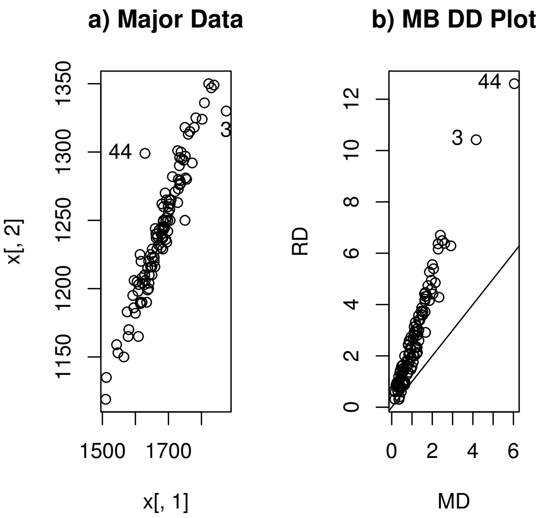

Fig. 7.1 Plots for Major Data

As the dimension p gets larger, outliers that can not be detected by marginal methods (case 44 in Example 7.2) become harder to detect. When $p = 3$ imagine that the clean data is a baseball bat or stick with one end at the SW corner of the bottom of the box (corresponding to the coordinate axes) and one end at the NE corner of the top of the box. If the outliers are a ball, there is much more room to hide them in the box than in a covering rectangle when $p = 2$.

Example 7.3. The estimators can be useful when the data is not elliptically contoured. The Gladstone (1905) data has 11 variables on 267 persons after death. Head measurements were *breadth, circumference, head height, length,* and *size* as well as *cephalic index* and *brain weight. Age*, *height*, and two categorical variables *ageclass* (0: under 20, 1: 20-45, 2: over 45) and *sex* were also given. Figure 7.2 shows the DD plots for the FCH, RMVN, `cov.mcd`, and MB estimators. The DD plots from the DGK, MBA, and RFCH estimators were similar, and the six outliers in Figure 7.2 correspond to the six infants in the data set.

Section 7.3 shows that if a consistent robust estimator is scaled as in (7.15), then the plotted points in the DD plot will cluster about the identity line with unit slope and zero intercept if the data is multivariate normal, and about some other line through the origin if the data is from some other elliptically contoured distribution with a nonsingular covariance matrix. Since multivariate procedures tend to perform well for elliptically contoured data, the DD plot is useful even if outliers are not present.

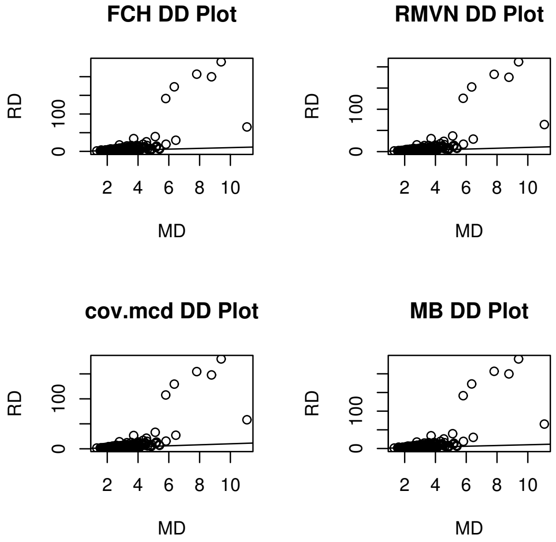

Fig. 7.2 DD Plots for Gladstone Data

7.2.5 Outlier Resistance and Simulations

```
RMVN                             FMCD
 0.996  0.014  0.002 -0.001  0.931  0.017  0.011 0.000
 0.014  2.012 -0.001  0.029  0.017  1.885 -0.003 0.022
 0.002 -0.001  2.984  0.003  0.011 -0.003  2.803 0.010
-0.001  0.029  0.003  3.994  0.000  0.022  0.010 3.752
```

Simulations were used to compare $(T_{FCH}, \boldsymbol{C}_{FCH})$, $(T_{RFCH}, \boldsymbol{C}_{RFCH})$, $(T_{RMVN}, \boldsymbol{C}_{RMVN})$, and $(T_{FMCD}, \boldsymbol{C}_{FMCD})$. Shown above are the averages, using 20 runs and $n = 1000$, of the dispersion matrices when the bulk of the data are iid $N_4(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = diag(1, 2, 3, 4)$. The first pair of matrices used $\gamma = 0$. Here the FCH, RFCH, and RMVN estimators are $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}$, while $\boldsymbol{C}_{FMCD}$ seems to be approximately unbiased for $0.94\boldsymbol{\Sigma}$.

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_4((0, 0, 0, 15)^T, 0.0001\ \boldsymbol{I}_4)$, a near point mass at the major axis. FCH and RFCH estimated $1.93\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$. The FMCD estimator failed to estimate $d\ \boldsymbol{\Sigma}$. Note that $\chi^2_{4,5/6}/\chi^2_{4,0.5} = 1.9276$.

```
RMVN                              FMCD
 0.988 -0.023 -0.007  0.021   0.227 -0.016  0.002 0.049
-0.023  1.964 -0.022 -0.002  -0.016  0.435 -0.014 0.013
-0.007 -0.022  3.053  0.007   0.002 -0.014  0.673 0.179
 0.021 -0.002  0.007  3.870   0.049  0.013  0.179 55.65
```

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_4(15\,\boldsymbol{1}, \boldsymbol{\Sigma})$, a mean shift with the same covariance matrix as the clean cases. Again FCH and RFCH estimated $1.93\boldsymbol{\Sigma}$ while RMVN and FMCD estimated $\boldsymbol{\Sigma}$.

```
RMVN                            FMCD
 1.013  0.008  0.006 -0.026   1.024  0.002  0.003 -0.025
 0.008  1.975 -0.022 -0.016   0.002  2.000 -0.034 -0.017
 0.006 -0.022  2.870  0.004   0.003 -0.034  2.931  0.005
-0.026 -0.016  0.004  3.976  -0.025 -0.017  0.005  4.046
```

Geometrical arguments suggest that the MB estimator has considerable outlier resistance. Suppose the outliers are far from the bulk of the data. Let the "median ball" correspond to the half set of data closest to MED($\boldsymbol{W}$) in Euclidean distance. If the outliers are outside of the median ball, then the initial half set in the iteration leading to the MB estimator will be clean. Thus the MB estimator will tend to give the outliers the largest MB distances unless the initial clean half set has very high correlation in a direction about which the outliers lie. This property holds for very general outlier configurations. The FCH estimator tries to use the DGK attractor if the $det(\boldsymbol{C}_{DGK})$ is small and the DGK location estimator T_{DGK} is in the median ball. Distant outliers that make $det(\boldsymbol{C}_{DGK})$ small also drag T_{DGK} outside of the median ball. Then FCH uses the MB attractor.

Compared to OGK and FMCD, the MB estimator is vulnerable to outliers that lie within the median ball. If the bulk of the data is highly correlated with the major axis of a hyperellipsoidal region, then the distances based on the clean data can be very large for outliers that fall within the median ball. The outlier resistance of the MB estimator decreases as p increases since the volume of the median ball rapidly increases with p.

A simple simulation for outlier resistance is to count the number of times the minimum distance of the outliers is larger than the maximum distance of the clean cases. The simulation used 100 runs. If the count was 97, then in 97 data sets the outliers can be separated from the clean cases with a horizontal line in the DD plot, but in 3 data sets the robust distances did not achieve complete separation. In Spring 2015, Det-MCD simulated much like FMCD, but was more likely to cause an error in R.

The clean cases had $\boldsymbol{x} \sim N_p(\boldsymbol{0}, diag(1, 2, ..., p))$. Outlier types were the mean shift $\boldsymbol{x} \sim N_p(pm\boldsymbol{1}, diag(1, 2, ..., p))$ where $\boldsymbol{1} = (1, ..., 1)^T$ and $\boldsymbol{x} \sim N_p((0, ..., 0, pm)^T, 0.0001\boldsymbol{I}_p)$, a near point mass at the major axis. Notice that the clean data can be transformed to a $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ distribution by multiplying $\boldsymbol{x}_i$ by $\text{diag}(1, 1/\sqrt{2}, ..., 1/\sqrt{p})$, and this transformation changes the location of the near point mass to $(0, ..., 0, pm/\sqrt{p})^T$.

Suppose the attractor is $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j})$ computed from a subset of c_n cases. The MCD(c_n) criterion is the determinant $det(\boldsymbol{S}_{k,j})$. The volume of the hyperellipsoid $\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,j})^T \boldsymbol{S}_{k,j}^{-1} (\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,j}) \le h^2\}$ is equal to

Table 7.3 Number of Times Mean Shift Outliers had the Largest Distances

p	γ	n	pm	MBA	FCH	RFCH	RMVN	OGK	FMCD	MB
10	.1	100	4	49	49	85	84	38	76	57
10	.1	100	5	91	91	99	99	93	98	91
10	.4	100	7	90	90	90	90	0	48	100
40	.1	100	5	3	3	3	3	76	3	17
40	.1	100	8	36	36	37	37	100	49	86
40	.25	100	20	62	62	62	62	100	0	100
40	.4	100	20	20	20	20	20	0	0	100
40	.4	100	35	44	98	98	98	95	0	100
60	.1	200	10	49	49	49	52	100	30	100
60	.1	200	20	97	97	97	97	100	35	100
60	.25	200	25	60	60	60	60	100	0	100
60	.4	200	30	11	21	21	21	17	0	100
60	.4	200	40	21	100	100	100	100	0	100

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{S}_{k,j})}, \tag{7.16}$$

see Johnson and Wichern (1988, pp. 103-104).

For near point mass outliers, a hyperellipsoid with very small volume can cover half of the data if the outliers are at one end of the hyperellipsoid and some of the clean data are at the other end. This half set will produce a classical estimator with very small determinant by (7.16). In the simulations for large γ, as the near point mass is moved very far away from the bulk of the data, only the classical, MB, and OGK estimators did not have numerical difficulties. Since the MCD estimator has smaller determinant than DGK, estimators like FMCD and MBA that use the MCD criterion without using location information will be vulnerable to these outliers. FMCD is also vulnerable to outliers if γ is slightly larger than γ_o given by (7.10).

Table 7.4 Number of Times Near Point Mass Outliers had the Largest Distances

p	γ	n	pm	MBA	FCH	RFCH	RMVN	OGK	FMCD	MB
10	.1	100	40	73	92	92	92	100	95	100
10	.25	100	25	0	99	99	90	0	0	99
10	.4	100	25	0	100	100	100	0	0	100
40	.1	100	80	0	0	0	0	79	0	80
40	.1	100	150	0	65	65	65	100	0	99
40	.25	100	90	0	88	87	87	0	0	88
40	.4	100	90	0	91	91	91	0	0	91
60	.1	200	100	0	0	0	0	13	0	91
60	.25	200	150	0	100	100	100	0	0	100
60	.4	200	150	0	100	100	100	0	0	100
60	.4	200	20000	0	100	100	100	64	0	100

Tables 7.3 and 7.4 help illustrate the results for the simulation. Large counts and small pm for fixed γ suggest greater ability to detect outliers.

Values of p were 5, 10, 15, ..., 60. First consider the mean shift outliers and Table 7.3. For $\gamma = 0.25$ and 0.4, MB usually had the highest counts. For $5 \leq p \leq 20$ and the mean shift, the OGK estimator often had the smallest counts, and FMCD could not handle 40% outliers for $p = 20$. For $25 \leq p \leq 60$, OGK usually had the highest counts for $\gamma = 0.05$ and 0.1. For $p \geq 30$, FMCD could not handle 25% outliers even for enormous values of pm.

In Table 7.4, FCH greatly outperformed MBA although the only difference between the two estimators is that FCH uses a location criterion as well as the MCD criterion. OGK performed well for $\gamma = 0.05$ and $20 \leq p \leq 60$ (not tabled). For large γ, OGK often has large bias for $c\boldsymbol{\Sigma}$. Then the outliers may need to be enormous before OGK can detect them. Also see Table 7.2, where OGK gave the outliers the largest distances for all runs, but $\boldsymbol{C}_{OGK}$ does not give a good estimate of $c\boldsymbol{\Sigma} = c \;\; diag(1,2)$.

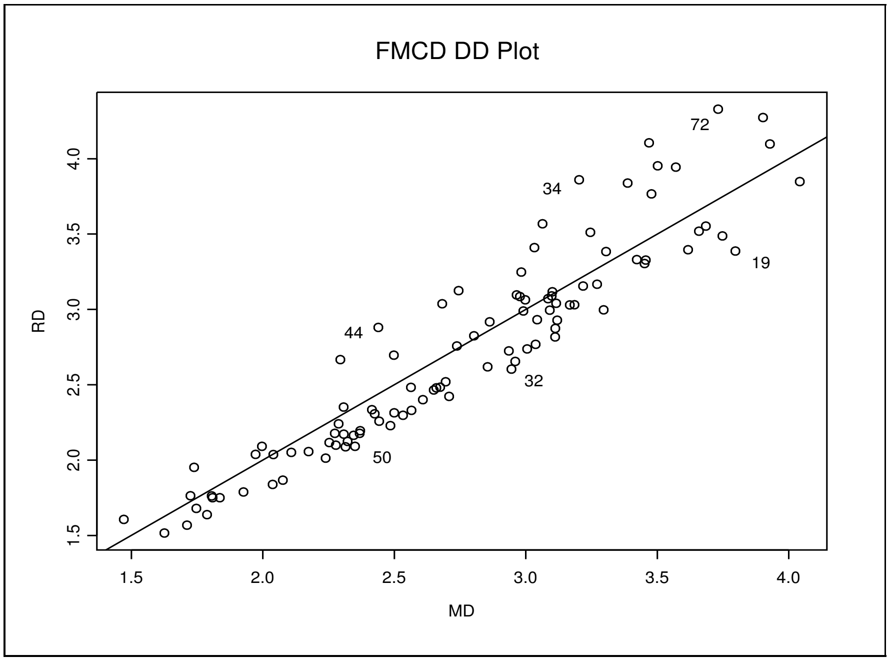

Fig. 7.3 The FMCD Estimator Failed

The DD plot of MD_i versus RD_i is useful for detecting outliers. 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. For the outlier data, the MBA estimator is biased, but the mean shift outliers in the MBA DD plot will have large RD_i since $\boldsymbol{C}_{MBA} \approx 2\boldsymbol{C}_{FMCD} \approx 2\boldsymbol{\Sigma}$.

In an older mean shift simulation, when p was 8 or larger, the `cov.mcd` estimator was usually not useful for detecting the mean shift outliers. Figure

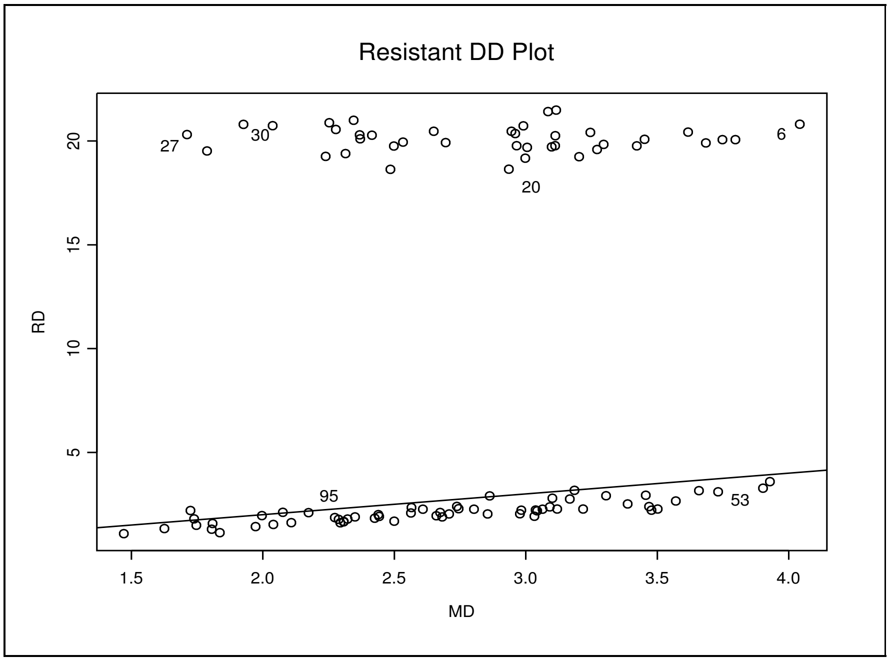

Fig. 7.4 The Outliers are Large in the MBA DD Plot

7.3 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 7.4.

For many data sets, Equation (7.10) gives a rough approximation for the number of large outliers that concentration algorithms using K starts each consisting of h cases can handle. However, if the data set is multivariate and the bulk of the data falls in one compact hyperellipsoid while the outliers fall in another hugely distant compact hyperellipsoid, then a concentration algorithm using a single start can sometimes tolerate nearly 25% outliers. For example, suppose that all $p + 1$ cases in the elemental start are outliers but the covariance matrix is nonsingular so that the Mahalanobis distances can be computed. Then the classical estimator is applied to the $c_n \approx n/2$ cases with the smallest distances. Suppose the percentage of outliers is less than 25% and that all of the outliers are in this “half set.” Then the sample mean applied to the c_n cases should be closer to the bulk of the data than to the cluster of outliers. Hence after a concentration step, the percentage of outliers will be reduced if the outliers are very far away. After the next concentration step the percentage of outliers will be further reduced and after several iterations, all c_n cases will be clean.

In a small simulation study, 20% outliers were planted for various values of p. If the outliers were distant enough, then the minimum DGK distance for the outliers was larger than the maximum DGK distance for the nonoutliers.

Hence the outliers would be separated from the bulk of the data in a DD plot of classical versus robust distances. For example, when the clean data comes from the $N_p(\mathbf{0}, \boldsymbol{I}_p)$ distribution and the outliers come from the $N_p(2000\,\mathbf{1}, \boldsymbol{I}_p)$ distribution, the DGK estimator with 10 concentration steps was able to separate the outliers in 17 out of 20 runs when $n = 9000$ and $p = 30$. With 10% outliers, a shift of 40, $n = 600$, and $p = 50$, 18 out of 20 runs worked. Olive (2004a) showed similar results for the Rousseeuw and Van Driessen (1999) FMCD algorithm and that the MBA estimator could often correctly classify up to 49% distant outliers. The following theorem shows that it is very difficult to drive the determinant of the dispersion estimator from a concentration algorithm to zero.

Theorem 7.13. Consider the concentration and MCD estimators that both cover c_n cases. For multivariate data, if at least one of the starts is nonsingular, then the concentration attractor $\boldsymbol{C}_A$ is less likely to be singular than the high breakdown MCD estimator $\boldsymbol{C}_{MCD}$.

Proof. If all of the starts are singular, then the Mahalanobis distances cannot be computed and the classical estimator can not be applied to c_n cases. Suppose that at least one start was nonsingular. Then $\boldsymbol{C}_A$ and $\boldsymbol{C}_{MCD}$ are both sample covariance matrices applied to c_n cases, but by definition $\boldsymbol{C}_{MCD}$ minimizes the determinant of such matrices. Hence $0 \leq \det(\boldsymbol{C}_{MCD}) \leq \det(\boldsymbol{C}_A)$. □

Software

The `robustbase` library was downloaded from (www.r-project.org/#doc). ∮ 11.1 explains how to use the source command to get the `linmodpack` functions in *R* and how to download a library from *R*. Type the commands `library(MASS)` and `library(robustbase)` to compute the FMCD and OGK estimators with the `cov.mcd` and `covOGK` functions. To use Det-MCD instead of FMCD, change

```
out <- covMcd(x)  to out <- covMcd(x,nsamp="deterministic"),
```

but in Spring 2015 this change was more likely to cause errors.

The `linmodpack` function
mldsim(n=200,p=5,gam=.2,runs=100,outliers=1,pm=15)
can be used to produce Tables 7.1–7.4. Change outliers to 0 to examine the average of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. The function `mldsim6` is similar but does not need the `library` command since it compares the FCH, RFCH, MB estimators, and the `covmb2` estimator of Section 7.3.

The function function *covfch* computes FCH and RFCH, while *covrmvn* computes the RMVN and MB estimators. The function *covrmb* computes MB and RMB where RMB is like RMVN except the MB estimator is reweighted instead of FCH. Functions *covdgk*, *covmba*, and *rmba* compute the scaled DGK, MBA, and RMBA estimators. **Better programs would use MB if DGK causes an error.**

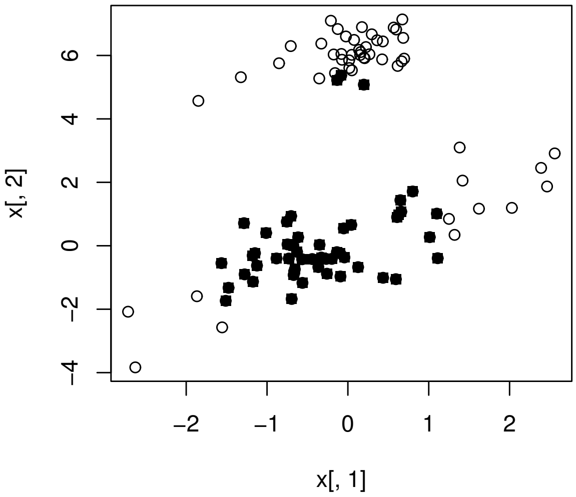

Fig. 7.5 highlighted cases = half set with smallest RD = $(T_0, \boldsymbol{C}_0)$

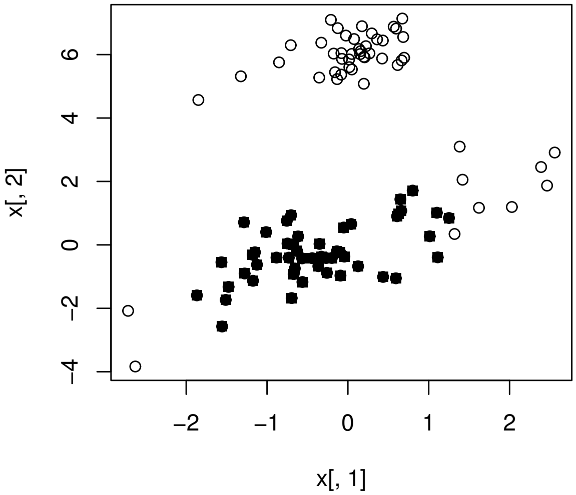

Fig. 7.6 highlighted cases = half set with smallest RD = $(T_1, \boldsymbol{C}_1)$

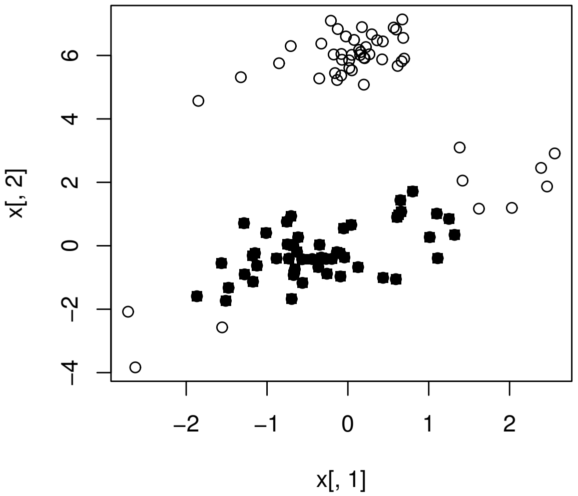

Fig. 7.7 highlighted cases = half set with smallest RD = $(T_2, \boldsymbol{C}_2)$

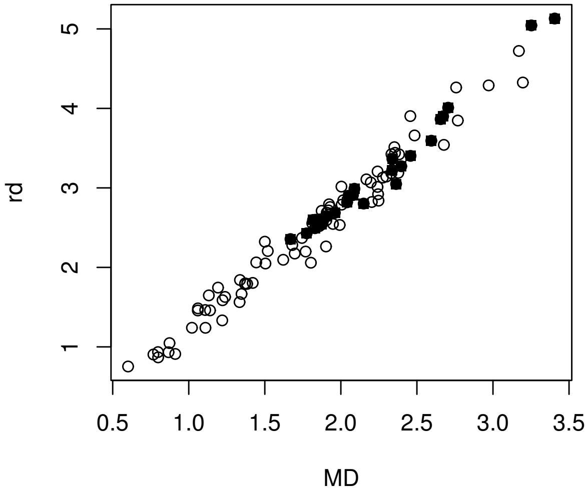

Fig. 7.8 highlighted cases = outliers, RD = $(T_{0,D}, \boldsymbol{C}_{0,D})$

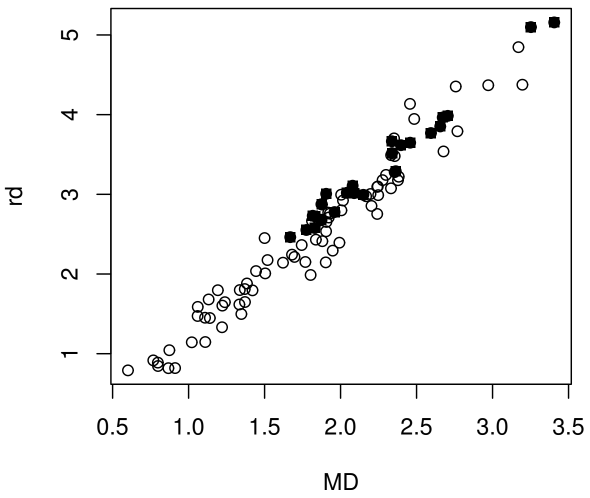

Fig. 7.9 highlighted cases = outliers, RD = $(T_{1,D}, \boldsymbol{C}_{1,D})$

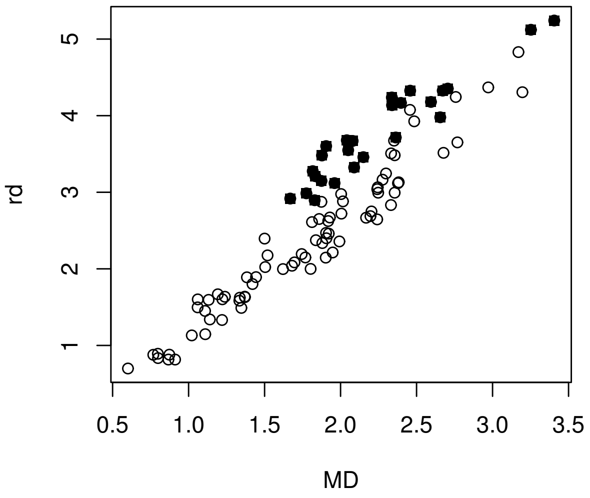

Fig. 7.10 highlighted cases = outliers, RD = $(T_{2,D}, \boldsymbol{C}_{2,D})$

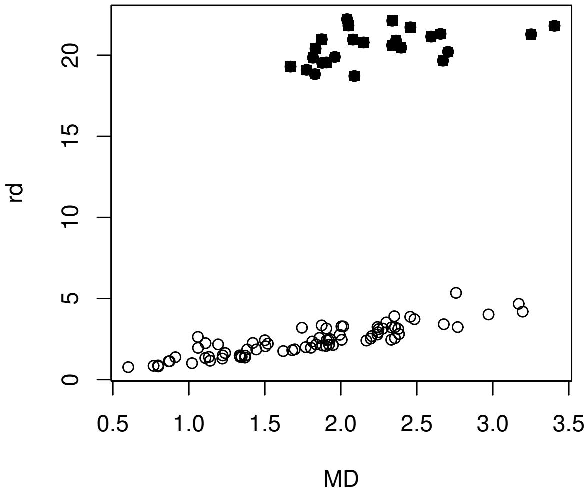

Fig. 7.11 highlighted cases = outliers, RD = $(T_{3,D}, \boldsymbol{C}_{3,D})$

The *concmv* function described in Problem 7.6 illustrates concentration where the start is $(\text{MED}(\boldsymbol{W}), diag([MAD(X_i)]^2))$. In Figures 7.5, 7.6, and 7.7, the highlighted cases are the half set with the smallest distances, and the initial half set shown in Figure 7.5 is not clean, where $n = 100$ and there are 40 outliers. The attractor shown in Figure 7.7 is clean. This type of data set has too many outliers for DGK while the MB starts and attractors are almost always clean.

The *ddmv* function in Problem 7.7 illustrates concentration for the DGK estimator where the start is the classical estimator. Now $n = 100, p = 4$, and there are 25 outliers. A DD plot of classical distances MD versus robust distances RD is shown. See Figures 7.8, 7.9, 7.10, and 7.11. The half set of cases with the smallest RDs is used, and the initial half set shown in Figure 7.8 is not clean. The attractor in Figure 7.11 is the DGK estimator which uses a clean half set. The clean cases $\boldsymbol{x}_i \sim N_4(\boldsymbol{0}, diag(1, 2, 3, 4))$ while the outliers $\boldsymbol{x}_i \sim N_4((10, 10\sqrt{2}, 10\sqrt{3}, 20)^T,\ diag(1, 2, 3, 4))$.

7.2.6 The RMVN and RFCH Sets

Both the RMVN and RFCH estimators compute the classical estimator $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ on some set U containing $n_U \geq n/2$ of the cases. Referring to Definition 7.16, for the RFCH estimator, $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U) = (T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2)$, and then $\boldsymbol{S}_U$ is scaled to form $\boldsymbol{C}_{RFCH}$. Referring to Definition 7.17, for the RMVN estimator, $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U) = (T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2)$, and then $\boldsymbol{S}_U$ is scaled to form $\boldsymbol{C}_{RMVN}$. See Definition 7.18.

The two main ways to handle outliers are i) apply the multivariate method to the cleaned data, and ii) plug in robust estimators for classical estimators. Subjectively cleaned data may work well for a single data set, but we can't get large sample theory since sometimes too many cases are deleted (delete outliers and some nonoutliers) and sometimes too few (do not get all of the outliers). Practical plug in robust estimators have rarely been shown to be $\sqrt{n}$ consistent and highly outlier resistant.

Using the RMVN or RFCH set U is simultaneously a plug in method and an objective way to clean the data such that the resulting robust method is often backed by theory. This result is extremely useful computationally: find the RMVN set or RFCH set U, then apply the classical method to the cases in the set U. This procedure is often equivalent to using $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ as plug in estimators. The method can be applied if $n > 2(p+1)$ but may not work well unless $n > 20p$. The *linmodpack* function `getu` gets the RMVN set U as well as the case numbers corresponding to the cases in U.

The set U is a small volume hyperellipsoid containing at least half of the cases since concentration is used. The set U can also be regarded as the "untrimmed data": the data that was not trimmed by ellipsoidal trimming. Theory has been proved for a large class of elliptically contoured distributions, but it is conjectured that theory holds for a much wider class of distributions. See Olive (2017b, pp. 127-128).

In simulations RFCH and RMVN seem to estimate $c\boldsymbol{\Sigma}_{\boldsymbol{x}}$ if $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu}$ where $\boldsymbol{z} = (z_1, ..., z_p)^T$ and the z_i are iid from a continuous distribution with variance σ^2. Here $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{x}) = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T$. The bias for the MB estimator seemed to be small. It is known that affine equivariant estimators give unbiased estimators of $c\boldsymbol{\Sigma}_{\boldsymbol{x}}$ if the distribution of z_i is also symmetric. DGK is affine equivariant and RFCH and RMVN are asymptotically equivalent to a scaled DGK estimator. But in the simulations the results also held for skewed distributions.

Several illustrative applications of the RMVN set U are given next, where the theory usually assumes that the cases are iid from a large class of elliptically contoured distributions.

i) The classical estimator of multivariate location and dispersion applied to the cases in U gives $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$, a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ for some constant $c > 0$. See Remark 7.4.

ii) The classical estimator of the correlation matrix applied to the cases in U gives $\boldsymbol{R}_U$, a consistent estimator of the population correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}}$.

iii) For multiple linear regression, let Y be the response variable, $x_1 = 1$ and $x_2, ..., x_p$ be the predictor variables. Let $\boldsymbol{z}_i = (Y_i, x_{i2}, ..., x_{ip})^T$. Let U be the RMVN or RFCH set formed using the $\boldsymbol{z}_i$. Then a classical regression estimator applied to the set U results in a robust regression estimator. For least squares, this is implemented with the *linmodpack* function `rmreg3`.

iv) For multivariate linear regression, let $Y_1, ..., Y_m$ be the response variables, $x_1 = 1$ and $x_2, ..., x_p$ be the predictor variables. Let

$$\boldsymbol{z}_i = (Y_{i1}, ...Y_{im}, x_{i2}, ..., x_{ip})^T.$$

Let U be the RMVN or RFCH set formed using the $\boldsymbol{z}_i$. Then a classical least squares multivariate linear regression estimator applied to the set U results in a robust multivariate linear regression estimator. For least squares, this is implemented with the *linmodpack* function `rmreg2`. The method for multiple linear regression in iii) corresponds to $m = 1$. See Section 8.6.

There are also several variants on the method. Suppose there are tentative predictors $Z_1, ..., Z_J$. After transformations assume that predictors $X_1, ..., X_k$ are linearly related. Assume the set U used cases $i_1, i_2, ..., i_{n_U}$. To add variables like $X_{k+1} = X_1^2$, $X_{k+2} = X_3 X_4$, $X_{k+3} = \textit{gender}$, ..., X_p, augment U with the variables $X_{k+1}, ..., X_p$ corresponding to cases $i_1, ..., i_{n_U}$. Adding variables results in cleaned data that is more likely to contain outliers.

If there are g groups ($g = G$ for discriminant analysis, $g = 2$ for binary regression, and $g = p$ for one way MANOVA), the function `getubig` gets the RMVN set U_i for each group and combines the g RMVN sets into one large set $U_{big} = U_1 \cup U_2 \cup \cdots \cup U_g$. Olive (2017b) has many more applications.

7.3 Outlier Detection for the MLD Model

Now suppose the multivariate data has been collected into an $n \times p$ matrix

$$\boldsymbol{W} = \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 \; \boldsymbol{v}_2 \; \dots \; \boldsymbol{v}_p \end{bmatrix}$$

where the ith row of $\boldsymbol{W}$ is the ith case $\boldsymbol{x}_i^T$ and the jth column $\boldsymbol{v}_j$ of $\boldsymbol{W}$ corresponds to n measurements of the jth random variable X_j for $j = 1, ..., p$. Hence the n rows of the data matrix $\boldsymbol{W}$ correspond to the n cases, while the p columns correspond to measurements on the p random variables $X_1, ..., X_p$. For example, the data may consist of n visitors to a hospital where the $p = 2$ variables *height* and *weight* of each individual were measured.

Definition 7.19. The *coordinatewise median* $\text{MED}(\boldsymbol{W}) = (\text{MED}(X_1), ..., \text{MED}(X_p))^T$ where $\text{MED}(X_i)$ is the sample median of the data in column i corresponding to variable X_i and $\boldsymbol{v}_i$.

Example 7.4. Let the data for X_1 be $1, 2, 3, 4, 5, 6, 7, 8, 9$ while the data for X_2 is $7, 17, 3, 8, 6, 13, 4, 2, 1$. Then $\text{MED}(\boldsymbol{W}) = (\text{MED}(X_1), \text{MED}(X_2))^T = (5, 6)^T$.

Definition 7.20: Rousseeuw and Van Driessen (1999). The *DD plot* is a plot of the classical Mahalanobis distances MD_i versus robust Mahalanobis distances RD_i.

The DD plot is used as a diagnostic for multivariate normality, elliptical symmetry, and for outliers. Assume that the data set consists of iid vectors from an $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with second moments. See Section 1.7 for notation. Then the classical sample mean and covariance matrix $(T_M, \boldsymbol{C}_M) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ is a consistent estimator for $(\boldsymbol{\mu}, c_{\boldsymbol{x}}\boldsymbol{\Sigma}) = (E(\boldsymbol{x}), \text{Cov}(\boldsymbol{x}))$. Assume that an alternative algorithm estimator $(T_A, \boldsymbol{C}_A)$ is a consistent estimator for $(\boldsymbol{\mu}, a_A\boldsymbol{\Sigma})$ for some constant $a_A > 0$. By scaling the algorithm estimator, the DD plot can be constructed to follow the identity line with unit slope and zero intercept. Let $(T_R, \boldsymbol{C}_R) = (T_A, \boldsymbol{C}_A/\tau^2)$ denote the scaled algorithm estimator where $\tau > 0$ is a constant to be determined. Notice that $(T_R, \boldsymbol{C}_R)$ is a valid estimator of location and dispersion. Hence the robust distances used in the DD plot are given by

$$\text{RD}_i = \text{RD}_i(T_R, \boldsymbol{C}_R) = \sqrt{(\boldsymbol{x}_i - T_R(\boldsymbol{W}))^T [\boldsymbol{C}_R(\boldsymbol{W})]^{-1} (\boldsymbol{x}_i - T_R(\boldsymbol{W}))}$$

$= \tau \ D_i(T_A, \boldsymbol{C}_A)$ for $i = 1, ..., n$.

The following theorem shows that if consistent estimators are used to construct the distances, then the DD plot will tend to cluster tightly about the line segment through $(0, 0)$ and $(\text{MD}_{n,\alpha}, \text{RD}_{n,\alpha})$ where $0 < \alpha < 1$ and $\text{MD}_{n,\alpha}$ is the 100αth sample percentile of the MD_i. Nevertheless, the variability in the DD plot may increase with the distances. Let $K > 0$ be a constant, e.g. the 99th percentile of the χ^2_p distribution.

Theorem 7.14. Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid observations from a distribution with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix. Let $a_j > 0$ and assume that $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ are consistent estimators of $(\boldsymbol{\mu}, a_j\boldsymbol{\Sigma})$ for $j = 1, 2$.

a) $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = o_P(1)$.

b) Let $0 < \delta \leq 0.5$. If $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - (\boldsymbol{\mu}, a_j\boldsymbol{\Sigma}) = O_p(n^{-\delta})$ and $a_j \hat{\boldsymbol{\Sigma}}_j^{-1} - \boldsymbol{\Sigma}^{-1} = O_P(n^{-\delta})$, then

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = O_P(n^{-\delta}).$$

c) Let $D_{i,j} \equiv D_i(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ be the ith Mahalanobis distance computed from $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$. Consider the cases in the region $R = \{i | 0 \leq D_{i,j} \leq K, \ j = 1, 2\}$. Let r_n denote the correlation between $D_{i,1}$ and $D_{i,2}$ for the cases in R

(thus r_n is the correlation of the distances in the "lower left corner" of the DD plot). Then $r_n \to 1$ in probability as $n \to \infty$.

Proof. Let B_n denote the subset of the sample space on which both $\hat{\boldsymbol{\Sigma}}_{1,n}$ and $\hat{\boldsymbol{\Sigma}}_{2,n}$ have inverses. Then $P(B_n) \to 1$ as $n \to \infty$.

a) and b): $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) =$

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} - \frac{\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

$$= (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{-\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) + (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T (-\boldsymbol{\Sigma}^{-1} + a_j \ \hat{\boldsymbol{\Sigma}}_j^{-1})(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) \ +$$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$

$$+ \frac{2}{a_j} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j) + \frac{1}{a_j} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)$$

$$+ \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T [a_j \hat{\boldsymbol{\Sigma}}_j^{-1} - \boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) \tag{7.17}$$

on B_n, and the last three terms are $o_P(1)$ under a) and $O_P(n^{-\delta})$ under b).

c) Following the proof of a), $D_j^2 \equiv D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) \xrightarrow{P} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) / a_j$ for fixed $\boldsymbol{x}$, and the result follows. □

The above result implies that a plot of the MD_i versus the $D_i(T_A, \boldsymbol{C}_A) \equiv D_i(A)$ will follow a line through the origin with some positive slope since if $\boldsymbol{x} = \boldsymbol{\mu}$, then both the classical and the algorithm distances should be close to zero. We want to find τ such that $\text{RD}_i = \tau \ D_i(T_A, \boldsymbol{C}_A)$ and the DD plot of MD_i versus RD_i follows the identity line. By Theorem 7.14, the plot of MD_i versus $D_i(A)$ will follow the line segment defined by the origin $(0, 0)$ and the point of observed median Mahalanobis distances, $(\text{med}(\text{MD}_i), \text{med}(D_i(A)))$. This line segment has slope

$$\text{med}(D_i(A))/\text{med}(\text{MD}_i)$$

which is generally not one. By taking $\tau = \text{med}(\text{MD}_i)/\text{med}(D_i(A))$, the plot will follow the identity line if $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ is a consistent estimator of $(\boldsymbol{\mu}, c_{\boldsymbol{x}} \boldsymbol{\Sigma})$ and if $(T_A, \boldsymbol{C}_A)$ is a consistent estimator of $(\boldsymbol{\mu}, a_A \boldsymbol{\Sigma})$. (Using the notation from Theorem 7.14, let $(a_1, a_2) = (c_{\boldsymbol{x}}, a_A)$.) The classical estimator is consistent if the population has a nonsingular covariance matrix. The algorithm estimators $(T_A, \boldsymbol{C}_A)$ from Theorem 7.11 are consistent on a large class of EC distributions that have a nonsingular covariance matrix, but tend to be biased for non–EC distributions. We recommend using RFCH or RMVN as the robust estimators in DD plots.

By replacing the observed median med(MD$_i$) of the classical Mahalanobis distances with the target population analog, say MED, τ can be chosen so that the DD plot is *simultaneously* a diagnostic for elliptical symmetry and a diagnostic for the target EC distribution. That is, the plotted points follow the identity line if the data arise from a target EC distribution such as the multivariate normal distribution, but the points follow a line with non-unit slope if the data arise from an alternative EC distribution. In addition the DD plot can often detect departures from elliptical symmetry such as outliers, the presence of two groups, or the presence of a mixture distribution.

Example 7.5. We will use the multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution as the target. If the data are indeed iid MVN vectors, then the $(\text{MD}_i)^2$ are asymptotically χ^2_p random variables, and MED $= \sqrt{\chi^2_{p,0.5}}$ where $\chi^2_{p,0.5}$ is the median of the χ^2_p distribution. Since the target distribution is Gaussian, let

$$\text{RD}_i = \frac{\sqrt{\chi^2_{p,0.5}}}{\text{med}(D_i(A))} D_i(A) \quad \text{so that} \quad \tau = \frac{\sqrt{\chi^2_{\text{p},0.5}}}{\text{med}(\text{D}_\text{i}(\text{A}))}. \tag{7.18}$$

Since every nonsingular estimator of multivariate location and dispersion defines a hyperellipsoid, the DD plot can be used to examine which points are in the robust hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T_R)^T \boldsymbol{C}_R^{-1} (\boldsymbol{x} - T_R) \le RD^2_{(h)}\} \tag{7.19}$$

where $RD^2_{(h)}$ is the hth smallest squared robust Mahalanobis distance, and which points are in a classical hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \le MD^2_{(h)}\}. \tag{7.20}$$

In the DD plot, points below $RD_{(h)}$ correspond to cases that are in the hyperellipsoid given by Equation (7.19) while points to the left of $MD_{(h)}$ are in a hyperellipsoid determined by Equation (7.20). In particular, we can use the DD plot to examine which points are in the nonparametric prediction region (4.24).

Application 7.1. Consider the DD plot with RFCH or RMVN. The DD plot can be used *simultaneously* as a diagnostic for whether the data arise from a multivariate normal distribution or from another EC distribution with nonsingular covariance matrix. EC data will cluster about a straight line through the origin; MVN data in particular will cluster about the identity line. Thus the DD plot can be used to assess the success of numerical transformations towards elliptical symmetry. The DD plot can be used to detect multivariate outliers. Use the DD plot to detect outliers and leverage groups if $n \geq 10p$ for the predictor variables in regression.

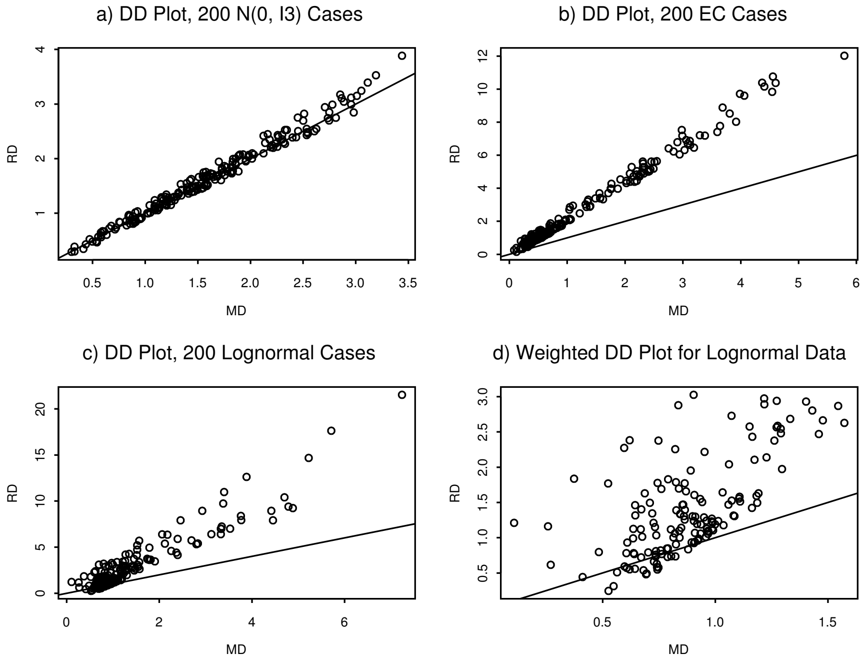

Fig. 7.12 4 DD Plots

For this application, the RFCH and RMVN estimators may be best. For MVN data, the RD_i from the RFCH estimator tend to have a higher correlation with the MD_i from the classical estimator than the RD_i from the FCH estimator, and the `cov.mcd` estimator may be inconsistent.

Figure 7.12 shows the DD plots for 3 artificial data sets using `cov.mcd`. The DD plot for 200 $N_3(\mathbf{0}, \boldsymbol{I}_3)$ points shown in Figure 7.12a resembles the identity line. The DD plot for 200 points from the elliptically contoured distribution $0.6N_3(\mathbf{0}, \boldsymbol{I}_3) + 0.4N_3(\mathbf{0}, 25\ \boldsymbol{I}_3)$ in Figure 7.12b clusters about a line through the origin with a slope close to 2.0.

A *weighted DD plot* magnifies the lower left corner of the DD plot by omitting the cases with $\text{RD}_i \geq \sqrt{\chi^2_{p,.975}}$. This technique can magnify features that are obscured when large RD_i's are present. If the distribution of $\boldsymbol{x}$ is EC with nonsingular $\boldsymbol{\Sigma}$, Theorem 7.14 implies that the correlation of the points in the weighted DD plot will tend to one and that the points will cluster about a line passing through the origin. For example, the plotted points in the weighted DD plot (not shown) for the non-MVN EC data of Figure 7.12b are highly correlated and still follow a line through the origin with a slope close to 2.0.

Figures 7.12c and 7.12d illustrate how to use the weighted DD plot. The ith case in Figure 7.12c is $(\exp(x_{i,1}), \exp(x_{i,2}), \exp(x_{i,3}))^T$ where $\boldsymbol{x}_i$ is the ith case in Figure 7.12a; i.e. the marginals follow a lognormal distribution. The plot does not resemble the identity line, correctly suggesting that the distribution of the data is not MVN; however, the correlation of the plotted points is rather high. Figure 7.12d is the weighted DD plot where cases with $\text{RD}_i \geq \sqrt{\chi^2_{3,.975}} \approx 3.06$ have been removed. Notice that the correlation of the plotted points is not close to one and that the best fitting line in Figure 7.12d may not pass through the origin. These results suggest that the distribution of $\boldsymbol{x}$ is not EC.

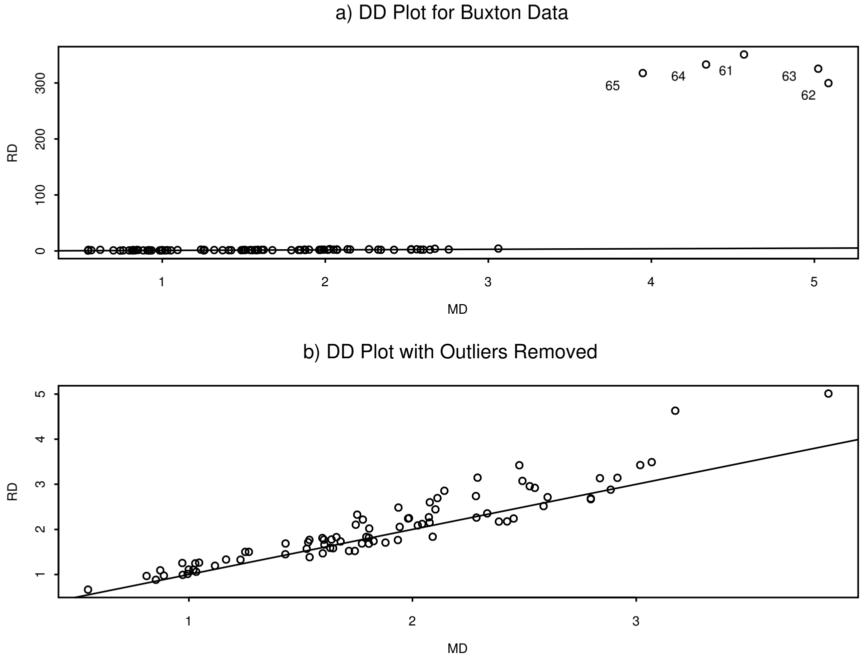

Fig. 7.13 DD Plots for the Buxton Data

Example 7.6. Buxton (1920, pp. 232-5) gave 20 measurements of 88 men. We will examine whether the multivariate normal distribution is a reasonable model for the measurements *head length, nasal height, bigonal breadth,* and *cephalic index* where one case has been deleted due to missing values. Figure 7.13a shows the DD plot. Five head lengths were recorded to be around 5 feet and are massive outliers. Figure 7.13b is the DD plot computed after deleting these points and suggests that the multivariate normal distribution is reasonable. (The recomputation of the DD plot means that the plot is not a weighted DD plot which would simply omit the outliers and then rescale the vertical axis.)

```
library(MASS)
x <- cbind(buxy,buxx)
ddplot(x,type=3) #Figure 7.13a), right click Stop

zx <- x[-c(61:65),]
ddplot(zx,type=3) #Figure 7.13b), right click Stop
```

7.3.1 MLD Outlier Detection if $p > n$

Most outlier detection methods work best if $n \geq 20p$, but often data sets have $p > n$, and outliers are a major problem. One of the simplest outlier detection methods uses the Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median $D_i = D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median MED_j: the coordinatewise median computed from the “half set” of cases $\boldsymbol{x}_i$ with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \boldsymbol{I}_p))$ where $\text{MED}_0 = \text{MED}(\boldsymbol{W})$. We often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, ..., D_n) + k\text{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise. Using $k \geq 0$ insures that at least half of the cases get weight 1. This weighting corresponds to the weighting that would be used in a one sided metrically trimmed mean (Huber type skipped mean) of the distances.

Application 7.2. This outlier resistant regression method uses terms from the following definition. Let the ith case $\boldsymbol{w}_i = (Y_i, \boldsymbol{x}_i^T)^T$ where the continuous predictors from $\boldsymbol{x}_i$ are denoted by $\boldsymbol{u}_i$ for $i = 1, ..., n$. Apply the `covmb2` estimator to the $\boldsymbol{u}_i$, and then run the regression method on the m cases $\boldsymbol{w}_i$ corresponding to the `covmb2` set B indices $i_1, ...i_m$, where $m \geq n/2$.

Definition 7.21. Let the *covmb2 set* B of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. The cases not in set B get weight $W_i = 0$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set B. Hence

$$T = \frac{\sum_{i=1}^n W_i \boldsymbol{x}_i}{\sum_{i=1}^n W_i} \quad \text{and} \quad \boldsymbol{C} = \frac{\sum_{i=1}^n W_i (\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^n W_i \; - \; 1}.$$

Example 7.7. Let the clean data (nonoutliers) be $i\,\mathbf{1}$ for $i = 1, 2, 3, 4$, and 5 while the outliers are $j\,\mathbf{1}$ for $j = 16, 17, 18$, and 19. Here $n = 9$ and $\mathbf{1}$ is $p \times 1$. Making a plot of the data for $p = 2$ may be useful. Then the coordinatewise median $\text{MED}_0 = \text{MED}(\boldsymbol{W}) = 5\,\mathbf{1}$. The median Euclidean distance of the data is the Euclidean distance of $5\,\mathbf{1}$ from $1\,\mathbf{1}$ = the Euclidean distance of $5\,\mathbf{1}$ from $9\,\mathbf{1}$. The *median ball* is the hypersphere centered at the coordinatewise median with radius $r = \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p),\ i = 1, ..., n)$ that tends to contain $(n+1)/2$ of the cases if n is odd. Hence the clean data are in the median ball and the outliers are outside of the median ball. The coordinatewise median of the cases with the 5 smallest distances is the coordinatewise median of the clean data: $\text{MED}_1 = 3\,\mathbf{1}$. Then the median Euclidean distance of the data from MED_1 is the Euclidean distance of $3\,\mathbf{1}$ from $1\,\mathbf{1}$ = the Euclidean distance of $3\,\mathbf{1}$ from $5\,\mathbf{1}$. Again the clean cases are the cases with the 5 smallest Euclidean distances. Hence $\text{MED}_j = 3\,\mathbf{1}$ for $j \geq 1$. For $j \geq 1$, if $\boldsymbol{x}_i = j\,\mathbf{1}$, then $D_i = |j - 3|\sqrt{p}$. Thus $D_{(1)} = 0$, $D_{(2)} = D_{(3)} = \sqrt{p}$, and $D_{(4)} = D_{(5)} = 2\sqrt{p}$. Hence $\text{MED}(D_1, ..., D_n) = D_{(5)} = 2\sqrt{p} = \text{MAD}(D_1, ..., D_n)$ since the median distance of the D_i from $D_{(5)}$ is $2\sqrt{p} - 0 = 2\sqrt{p}$. Note that the 5 smallest absolute distances $|D_i - D_{(5)}|$ are $0, 0, \sqrt{p}, \sqrt{p}$, and $2\sqrt{p}$. Hence $W_i = 1$ if $D_i \leq 2\sqrt{p} + 10\sqrt{p} = 12\sqrt{p}$. The clean data get weight 1 while the outliers get weight 0 since the smallest distance D_i for the outliers is the Euclidean distance of $3\,\mathbf{1}$ from $16\,\mathbf{1}$ with a $D_i = \|16\,\mathbf{1} - 3\,\mathbf{1}\| = 13\sqrt{p}$. Hence the `covmb2` estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix of the clean data. **Note that the distance for the outliers to get zero weight is proportional to the square root of the dimension** $\sqrt{p}$.

The `covmb2` estimator can also be used for $n > p$. The `covmb2` estimator attempts to give a robust dispersion estimator that reduces the bias by using a big ball about MED_j instead of a ball that contains half of the cases. The *linmodpack* function `getB` gives the set B of cases that got weight 1 along with the index `indx` of the case numbers that got weight 1. The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers. An alternative for outlier detection is to replace $\boldsymbol{C}$ by $\boldsymbol{C}_d = diag(\hat{\sigma}_{11}, ..., \hat{\sigma}_{pp})$. For example, use $\hat{\sigma}_{ii} = \boldsymbol{C}_{ii}$. See Ro et al. (2015) and Tarr et al. (2016) for references.

Example 7.8. For the Buxton (1920) data with multiple linear regression, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet! See Problem 7.11 to reproduce the following plots.

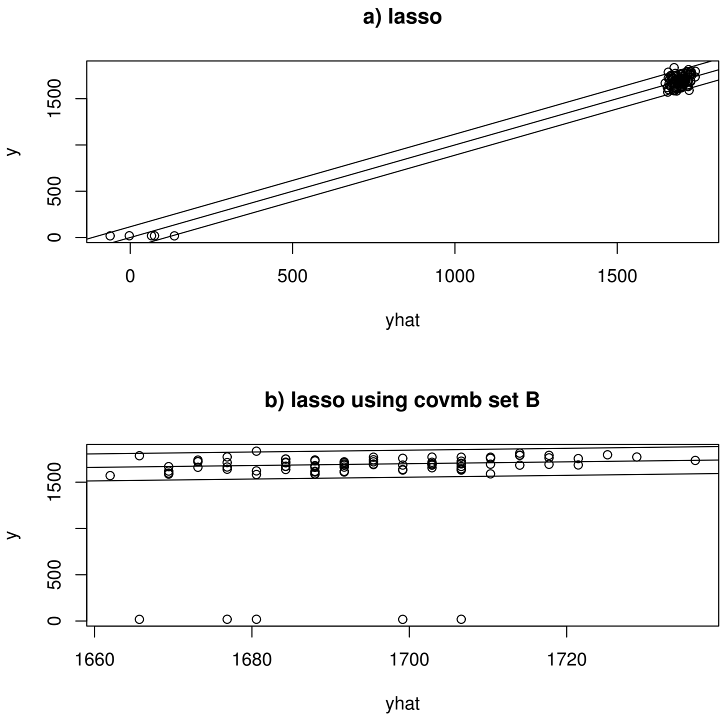

Fig. 7.14 Response plot for lasso and lasso applied to the `covmb2` set B.

Figure 7.14a) shows the response plot for lasso. The identity line passes right through the outliers which are obvious because of the large gap. Figure 7.14b) shows the response plot from lasso for the cases in the `covmb2` set B applied to the predictors, and the set B included all of the clean cases and omitted the 5 outliers. The response plot was made for all of the data, including the outliers. Prediction interval (PI) bands are also included for both plots. Both plots are useful for outlier detection, but the method for plot 7.14b) is better for data analysis: impossible outliers should be deleted or given 0 weight, we do not want to predict that some people are about 0.75 inches tall, and we do want to predict that the people were about 1.6 to 1.8 meters tall. Figure 7.15 shows the DD plot made using `ddplot5`. The five outliers are in the upper right corner.

Also see Problem 7.12 b) for the Gladstone (1905) data where the `covmb2` set B deleted the 8 cases with the largest D_i, including 5 outliers and 3 clean cases.

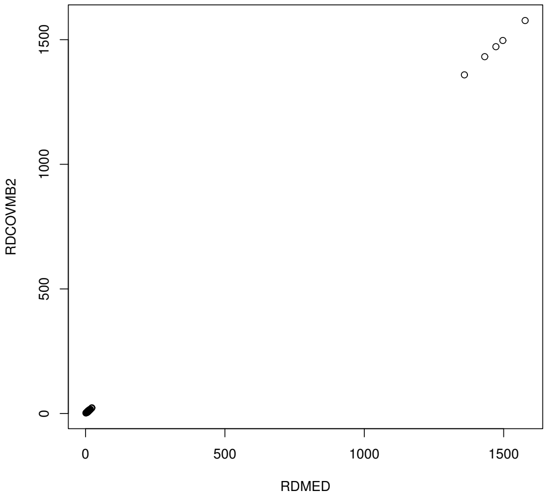

Fig. 7.15 DD plot with ddplot5.

7.4 Outlier Detection for the MLR Model

For multiple linear regression, the OLS response and residual plots are very useful for detecting outliers. The DD plot of the continuous predictors is also useful. Use the *linmodpack* functions `MLRplot` and `ddplot4`. Response and residual plots from outlier resistant methods are also useful. See Figure 7.14.

Huber and Ronchetti (2009, p. 154) noted that efficient methods for identifying leverage groups are needed. Such groups are often difficult to detect with regression diagnostics and residuals, but often have outlying fitted values and responses that can be detected with response and residual plots. The following *rules of thumb* are useful for finding influential cases and outliers. Look for points with large absolute residuals and for points far away from $\overline{Y}$. Also look for gaps separating the data into clusters. The OLS fit often passes through a cluster of outliers, causing a large gap between a cluster corresponding to the bulk of the data and the cluster of outliers. When such a gap appears, it is possible that the smaller cluster corresponds to good leverage points: the cases follow the same model as the bulk of the data. To determine whether small clusters are outliers or good leverage points, give zero weight to the clusters, and fit an MLR estimator such as OLS to the bulk of the data. Denote the weighted estimator by $\hat{\boldsymbol{\beta}}_w$. Then plot $\hat{Y}_w$ versus Y using the entire data set. If the identity line passes through the cluster, then the cases in the cluster may be good leverage points, otherwise they may be outliers. The trimmed views estimator of Section 7.5 is also useful. Dragging the plots, so that they are roughly square, can be useful.

Definition 7.22. Suppose that some analysis to detect outliers is performed. *Masking* occurs if the analysis suggests that one or more outliers are in fact good cases. *Swamping* occurs if the analysis suggests that one or more good cases are outliers. Suppose that a subset of h cases is selected from the n cases making up the data set. Then the subset is *clean* if none of the h cases are outliers.

Influence diagnostics such as Cook's distances CD_i from Cook (1977) and the weighted Cook's distances WCD_i from Peña (2005) are sometimes useful. Although an index plot of Cook's distance CD_i may be useful for flagging influential cases, the index plot provides no direct way of judging the model against the data. As a remedy, cases in the response and residual plots with $CD_i > \min(0.5, 2p/n)$ are highlighted with open squares, and cases with $|WCD_i - \text{median}(\text{WCD}_\text{i})| > 4.5\text{MAD}(\text{WCD}_\text{i})$ are highlighted with crosses, where the median absolute deviation $\text{MAD}(w_i) = \text{median}(|w_i - \text{median}(\text{w}_\text{i})|)$.

Example 7.9. Figure 7.16 shows the response plot and residual plot for the Buxton (1920) data. Notice that the OLS fit passes through the outliers, but the response plot is resistant to Y–outliers since Y is on the vertical axis. Also notice that although the outlying cluster is far from $\overline{Y}$, only two of the outliers had large Cook's distance and only one case had a large WCD_i. Hence *masking* occurred for the Cook's distances, the WCD_i, and for the OLS residuals, but not for the OLS fitted values. Figure 7.16 was made with the following R commands.

```
source("G:/linmodpack.txt"); source("G:/linmoddata.txt")
mlrplot4(buxx,buxy) #right click Stop twice
```

High leverage outliers are a particular challenge to conventional numerical MLR diagnostics such as Cook's distance, but can often be visualized using the response and residual plots. (Using the trimmed views of Section 7.5 is also effective for detecting outliers and other departures from the MLR model.)

Example 7.10. Hawkins et al. (1984) gave a well known artificial data set where the first 10 cases are outliers while cases 11-14 are good leverage points. Figure 7.17 shows the residual and response plots based on the OLS estimator. The highlighted cases have Cook's distance $> \min(0.5, 2p/n)$, and the identity line is shown in the response plot. Since the good cases 11-14 have the largest Cook's distances and absolute OLS residuals, *swamping* has occurred. (Masking has also occurred since the outliers have small Cook's distances, and some of the outliers have smaller OLS residuals than clean cases.) To determine whether both clusters are outliers or if one cluster consists of good leverage points, cases in both clusters could be given weight

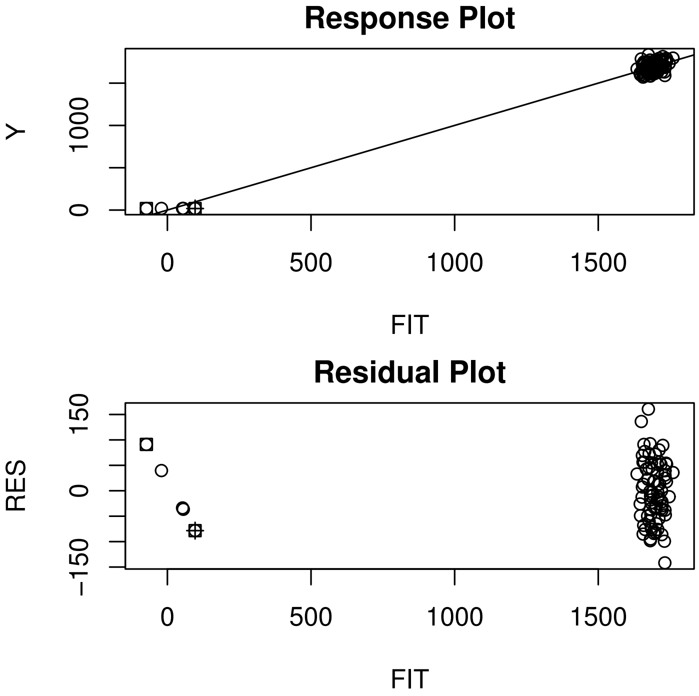

Fig. 7.16 Plots for Buxton Data

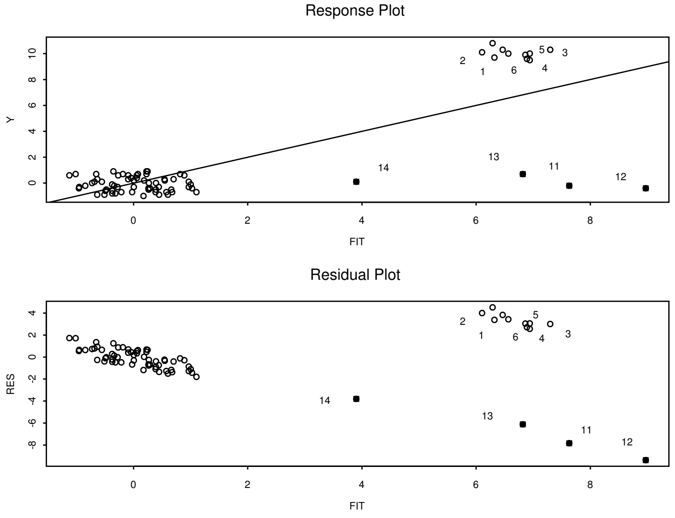

Fig. 7.17 Plots for HBK Data

zero and the resulting response plot created. (Alternatively, response plots based on the `tvreg` estimator of Section 7.5 could be made where the cases with weight one are highlighted. For high levels of trimming, the identity line often passes through the good leverage points.)

The above example is typical of many "benchmark" outlier data sets for MLR. In these data sets traditional OLS diagnostics such as Cook's distance and the residuals often fail to detect the outliers, but the combination of the response plot and residual plot is usually able to detect the outliers. The CD_i and WCD_i are the most effective when there is a single cluster about the identity line. If there is a second cluster of outliers or good leverage points or if there is nonconstant variance, then these numerical diagnostics tend to fail.

7.5 Resistant Multiple Linear Regression

Consider the multiple linear regression model, written in matrix form as $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$. The OLS response and residual plots are very useful for detecting outliers and checking the model. Resistant estimators are useful for detecting certain types of outliers. Some good resistant regression estimators are `rmreg2` from Section 8.6, the `hbreg` estimator from Section 7.7, and the Olive (2005) MBA and trimmed views estimators described below. Also apply a multiple linear regression method such as OLS or lasso to the cases corresponding to the RFCH, RMVN, or `covmb2` set applied to the continuous predictors. See Sections 7.2.6 and 7.3.1.

The L_1 estimator or least absolute deviations estimator is a competitor for OLS. The L_1 estimator $\hat{\boldsymbol{\beta}}_{L_1}$ minimizes the criterion $Q_{L_1}(\boldsymbol{b}) = \sum_{i=1}^n |r_i(\boldsymbol{b})|$ where $r_i(\boldsymbol{b}) = Y_i - \boldsymbol{x}_i^T\boldsymbol{b}$ is the ith residual corresponding to $\boldsymbol{b}$. Response and residual plots from these two estimators are useful for detecting outliers.

Resistant estimators are often created by computing several trial fits $\boldsymbol{b}_i$ that are estimators of $\boldsymbol{\beta}$. Then a criterion is used to select the trial fit to be used in the resistant estimator. Suppose $c \approx n/2$. The LMS(c) criterion is $Q_{LMS}(\boldsymbol{b}) = r^2_{(c)}(\boldsymbol{b})$ where $r^2_{(1)} \leq \cdots \leq r^2_{(n)}$ are the ordered squared residuals, and the LTS(c) criterion is $Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^c r^2_{(i)}(\boldsymbol{b})$. The LTA($c$) criterion is $Q_{LTA}(\boldsymbol{b}) = \sum_{i=1}^c |r(\boldsymbol{b})|_{(i)}$ where $|r(\boldsymbol{b})|_{(i)}$ is the ith ordered absolute residual. Three impractical high breakdown robust estimators are the Hampel (1975) least median of squares (LMS) estimator, the Rousseeuw (1984) least trimmed sum of squares (LTS) estimator, and the Hössjer (1991) least trimmed sum of absolute deviations (LTA) estimator. Also see Hawkins and Olive (1999ab). These estimators correspond to the $\hat{\boldsymbol{\beta}}_L \in \mathbb{R}^p$ that minimizes the corresponding criterion. LMS, LTA, and LTS have $O(n^p)$ or $O(n^{p+1})$ complexity. See Bernholt (2005), Hawkins and Olive (1999b), Klouda (2015), and Mount et al. (2014). Estimators with $O(n^4)$ or higher complexity take too long to compute. LTS and LTA are $\sqrt{n}$ consistent while LMS has the lower $n^{1/3}$ rate. See Kim and Pollard (1990), Čížek (2006, 2008), and Mašíček (2004). If $c = n$, the LTS and LTA criteria are the OLS and L_1 criteria. See Olive (2008, 2017b: ch. 14) for more on these estimators.

A good resistant estimator is the Olive (2005) *median ball algorithm* (MBA or `mbareg`). The Euclidean distance of the ith vector of predictors $\boldsymbol{x}_i$ from the jth vector of predictors $\boldsymbol{x}_j$ is

$$D_i(\boldsymbol{x}_j) = D_i(\boldsymbol{x}_j, \boldsymbol{I}_p) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)}.$$

For a fixed $\boldsymbol{x}_j$ consider the ordered distances $D_{(1)}(\boldsymbol{x}_j), ..., D_{(n)}(\boldsymbol{x}_j)$. Next, let $\hat{\boldsymbol{\beta}}_j(\alpha)$ denote the OLS fit to the $\min(p + 3 + \lfloor \alpha n/100 \rfloor, n)$ cases with the smallest distances where the approximate percentage of cases used is $\alpha \in \{1, 2.5, 5, 10, 20, 33, 50\}$. (Here $\lfloor x \rfloor$ is the greatest integer function so $\lfloor 7.7 \rfloor = 7$. The extra $p+3$ cases are added so that OLS can be computed for small n and α.) This yields seven OLS fits corresponding to the cases with predictors closest to $\boldsymbol{x}_j$. A fixed number of K cases are selected at random without replacement to use as the $\boldsymbol{x}_j$. Hence $7K$ OLS fits are generated. We use $K = 7$ as the default. A robust criterion Q is used to evaluate the $7K$ fits and the OLS fit to all of the data. Hence $7K + 1$ OLS fits are generated and the MBA estimator is the fit that minimizes the criterion. The median squared residual is a good choice for Q.

Three ideas motivate this estimator. First, $\boldsymbol{x}$-outliers, which are outliers in the predictor space, tend to be much more destructive than Y-outliers which are outliers in the response variable. Suppose that the proportion of outliers is γ and that $\gamma < 0.5$. We would like the algorithm to have at least one "center" $\boldsymbol{x}_j$ that is not an outlier. The probability of drawing a center that is not an outlier is approximately $1 - \gamma^K > 0.99$ for $K \geq 7$ and this result is free of p. Secondly, by using the different percentages of coverages, for many data sets there will be a center and a coverage that contains no outliers. Third, by Theorem 1.21, the MBA estimator is a $\sqrt{n}$ consistent estimator of the same parameter vector $\boldsymbol{\beta}$ estimated by OLS under mild conditions.

Ellipsoidal trimming can be used to create resistant multiple linear regression (MLR) estimators. To perform ellipsoidal trimming, an estimator $(T, \boldsymbol{C})$ is computed and used to create the squared Mahalanobis distances D_i^2 for each vector of observed predictors $\boldsymbol{x}_i$. If the ordered distance $D_{(j)}$ is unique, then j of the $\boldsymbol{x}_i$'s are in the ellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq D_{(j)}^2\}. \tag{7.21}$$

The ith case $(Y_i, \boldsymbol{x}_i^T)^T$ is trimmed if $D_i > D_{(j)}$. Then an estimator of $\boldsymbol{\beta}$ is computed from the remaining cases. For example, if $j \approx 0.9n$, then about 10% of the cases are trimmed, and OLS or L_1 could be used on the cases that remain. Ellipsoidal trimming differs from using the RFCH, RMVN, or `covmb2` set since these sets use a random amount of trimming. (The ellipsoidal trimming technique can also be used for other regression models, and the theory of the regression method tends to apply to the method applied to the cleaned data that was not trimmed since the response variables were not used to select the cases. See Chapter 10.)

Use ellipsoidal trimming on the RFCH, RMVN, or `covmb2` set applied to the continuous predictors to get a fit $\hat{\boldsymbol{\beta}}_C$. Then make a response and residual plot using all of the data, not just the cleaned data that was not trimmed.

The resistant trimmed views estimator combines ellipsoidal trimming and the response plot. First compute $(T, \boldsymbol{C})$ on the $\boldsymbol{x}_i$, perhaps using the RMVN estimator. Trim the $M\%$ of the cases with the largest Mahalanobis distances, and then compute the MLR estimator $\hat{\boldsymbol{\beta}}_M$ from the remaining cases. Use $M = 0$, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten response plots of the fitted values $\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_i$ versus Y_i using all n cases. (Fewer plots are used for small data sets if $\hat{\boldsymbol{\beta}}_M$ can not be computed for large M.) These plots are called "trimmed views."

Definition 7.23. The trimmed views (TV) estimator $\hat{\boldsymbol{\beta}}_{T,n}$ corresponds to the trimmed view where the bulk of the plotted points follow the identity line with smallest variance function, ignoring any outliers.

Example 7.11. For the Buxton (1920) data, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet! OLS was used on the cases remaining after trimming, and Figure 7.18 shows four trimmed views corresponding to 90%, 70%, 40%, and 0% trimming. The OLS TV estimator used 70% trimming since this trimmed view was best. Since the vertical distance from a plotted point to the identity line is equal to the case's residual, the outliers had massive residuals for 90%, 70%, and 40% trimming. Notice that the OLS trimmed view with 0% trimming "passed through the outliers" since the cluster of outliers is scattered about the identity line.

The TV estimator $\hat{\boldsymbol{\beta}}_{T,n}$ has good statistical properties if an estimator with good statistical properties is applied to the cases $(\boldsymbol{X}_{M,n}, \boldsymbol{Y}_{M,n})$ that remain after trimming. Candidates include OLS, L_1, Huber's M–estimator, Mallows' GM–estimator, or the Wilcoxon rank estimator. See Rousseeuw and Leroy (1987, pp. 12-13, 150). The basic idea is that if an estimator with $O_P(n^{-1/2})$ convergence rate is applied to a set of $n_M \propto n$ cases, then the resulting estimator $\hat{\boldsymbol{\beta}}_{M,n}$ also has $O_P(n^{-1/2})$ rate provided that the response Y was not used to select the n_M cases in the set. If $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ for $M = 0, ..., 90$ then $\|\hat{\boldsymbol{\beta}}_{T,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ by Theorem 1.21.

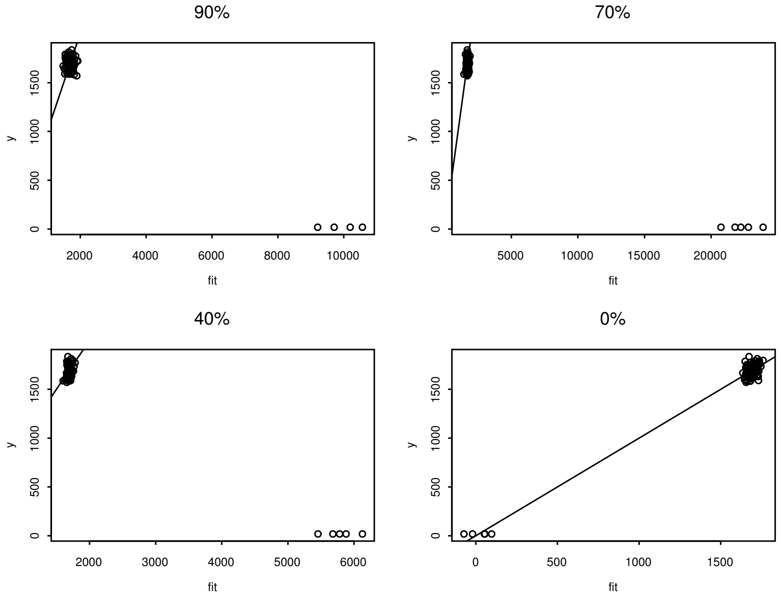

Fig. 7.18 4 Trimmed Views for the Buxton Data

Let $\boldsymbol{X}_n = \boldsymbol{X}_{0,n}$ denote the full design matrix. Often when proving asymptotic normality of an MLR estimator $\hat{\boldsymbol{\beta}}_{0,n}$, it is assumed that

$$\frac{\boldsymbol{X}_n^T \boldsymbol{X}_n}{n} \to \boldsymbol{W}^{-1}.$$

If $\hat{\boldsymbol{\beta}}_{0,n}$ has $O_P(n^{-1/2})$ rate and if for big enough n all of the diagonal elements of

$$\left(\frac{\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n}}{n} \right)^{-1}$$

are all contained in an interval $[0, B)$ for some $B > 0$, then $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$.

The distribution of the estimator $\hat{\boldsymbol{\beta}}_{M,n}$ is especially simple when OLS is used and the errors are iid $N(0, \sigma^2)$. Then

$$\hat{\boldsymbol{\beta}}_{M,n} = (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1} \boldsymbol{X}_{M,n}^T \boldsymbol{Y}_{M,n} \sim N_p(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1})$$

and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}) \sim N_p(\boldsymbol{0}, \sigma^2 (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n}/n)^{-1})$. This result does not imply that $\hat{\boldsymbol{\beta}}_{T,n}$ is asymptotically normal. See the following paragraph for the large sample theory of a modified trimmed views estimator.

Warning: When $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e$, MLR estimators tend to estimate the same slopes $\beta_2, ..., \beta_p$, but the constant β_1 tends to depend on the estimator unless the errors are symmetric. The MBA and trimmed views estimators do estimate the same $\boldsymbol{\beta}$ as OLS asymptotically, but samples may need to be huge before the MBA and trimmed views estimates of the constant are close to the OLS estimate of the constant. If the trimmed views estimator is modified so that the LTS, LTA, or LMS criterion is used to select the final estimator, then a conjecture is that the limiting distribution is similar to that of the variable selection estimator: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MTV} - \boldsymbol{\beta}) \overset{D}{\to} \sum_{i=1}^k \pi_i \boldsymbol{w}_i$ where $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^k \pi_i = 1$. The index i corresponds to the fits considered by the modified trimmed views estimator with $k = 10$. For the MBA estimator and the modified trimmed views estimator, the prediction region method, described in Section 4.5, may be useful for testing hypotheses. Large sample sizes may be needed if the error distribution is not symmetric since the constant $\hat{\boldsymbol{\beta}}_1$ needs large samples. See Olive (2017b, p. 444) for an explanation for why large sample sizes may be needed to estimate the constant.

The conditions under which the `rmreg2` estimator of Section 8.6 has been shown to be $\sqrt{n}$ consistent are quite strong, but it seems likely that the estimator is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under mild conditions where the parameter vector $\boldsymbol{\beta}$ is not, in general, the parameter vector estimated by OLS. For MLR, the *linmodpack* function `rmregboot` bootstraps the `rmreg2` estimator, and the function `rmregbootsim` can be used to simulate `rmreg2`. Both functions use the residual bootstrap where the residuals come from OLS. See the *R* code below.

```
out<-rmregboot(belx,bely)
plot(out$betas)
ddplot4(out$betas) #right click Stop

out<-rmregboot(cbrainx,cbrainy)
ddplot4(out$betas) #right click Stop
```

Often practical "robust estimators" generate a sequence of K trial fits called *attractors*: $\boldsymbol{b}_1, ..., \boldsymbol{b}_K$. Then some criterion is evaluated and the attractor $\boldsymbol{b}_A$ that minimizes the criterion is used in the final estimator.

Definition 7.24. For MLR, an *elemental set* J is a set of p cases drawn with replacement from the data set of n cases. The elemental fit is the OLS estimator $\hat{\boldsymbol{\beta}}_{J_i} = (\boldsymbol{X}_{J_i}^T \boldsymbol{X}_{J_i})^{-1} \boldsymbol{X}_{J_i}^T \boldsymbol{Y}_{J_i} = \boldsymbol{X}_{J_i}^{-1} \boldsymbol{Y}_{J_i}$ applied to the cases corresponding to the elemental set provided that the inverse of $\boldsymbol{X}_{J_i}$ exists. In a *concentration algorithm*, let $\boldsymbol{b}_{0,j}$ be the jth start, not necessarily elemental, and compute all n residuals $r_i(\boldsymbol{b}_{0,j}) = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}_{0,j}$. At the next iteration, the OLS estimator $\boldsymbol{b}_{1,j}$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest squared residuals $r_i^2(\boldsymbol{b}_{0,j})$. This iteration can be continued for

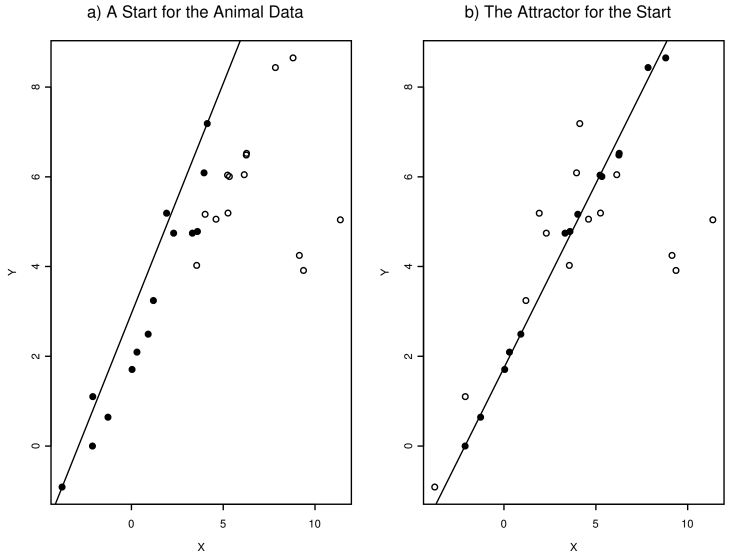

Fig. 7.19 The Highlighted Points are More Concentrated about the Attractor

k steps resulting in the sequence of estimators $\boldsymbol{b}_{0,j}, \boldsymbol{b}_{1,j}, ..., \boldsymbol{b}_{k,j}$. Then $\boldsymbol{b}_{k,j}$ is the jth *attractor* for $j = 1, ..., K$. Then the attractor $\boldsymbol{b}_A$ that minimizes the LTS criterion is used in the final estimator. Using $k = 10$ concentration steps often works well, and the basic resampling algorithm is a special case with $k = 0$, i.e., the attractors are the starts. Such an algorithm is called a CLTS concentration algorithm or CLTS.

A CLTA concentration algorithm would replace the OLS estimator by the L_1 estimator, and the smallest c_n squared residuals by the smallest c_n absolute residuals. Many other variants are possible, but obtaining theoretical results may be difficult.

Example 7.12. As an illustration of the CLTA concentration algorithm, consider the animal data from Rousseeuw and Leroy (1987, p. 57). The response Y is the *log brain weight* and the predictor x is the *log body weight* for 25 mammals and 3 dinosaurs (outliers with the highest body weight). Suppose that the first elemental start uses cases 20 and 14, corresponding to mouse and man. Then the start $\boldsymbol{b}_{s,1} = \boldsymbol{b}_{0,1} = (2.952, 1.025)^T$ and the sum of the $c = 14$ smallest absolute residuals $\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{0,1}) = 12.101$. Figure 7.19a shows the scatterplot of x and y. The start is also shown and the 14 cases corresponding to the smallest absolute residuals are highlighted. The L_1 fit to

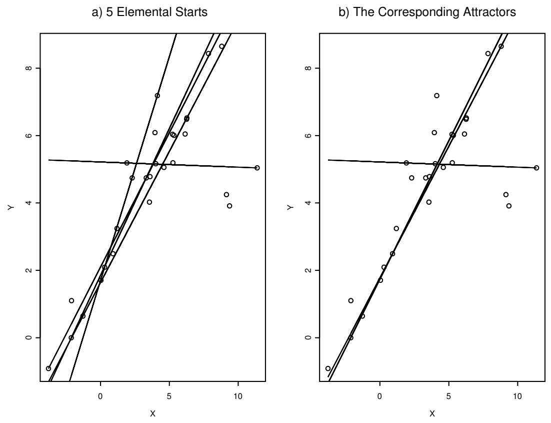

Fig. 7.20 Starts and Attractors for the Animal Data

these c highlighted cases is $\boldsymbol{b}_{1,1} = (2.076, 0.979)^T$ and $\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{1,1}) = 6.990$. The iteration consists of finding the cases corresponding to the c smallest absolute residuals, obtaining the corresponding L_1 fit and repeating. The attractor $\boldsymbol{b}_{a,1} = \boldsymbol{b}_{7,1} = (1.741, 0.821)^T$ and the LTA(c) criterion evaluated at the attractor is $\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{a,1}) = 2.172$. Figure 7.19b shows the attractor and that the c highlighted cases corresponding to the smallest absolute residuals are much more concentrated than those in Figure 7.19a. Figure 7.20a shows 5 randomly selected starts while Figure 7.20b shows the corresponding attractors. Notice that the elemental starts have more variability than the attractors, but if the start passes through an outlier, so does the attractor.

Remark 7.6. Consider drawing K elemental sets $J_1, ..., J_K$ with replacement to use as starts. For multivariate location and dispersion, use the attractor with the smallest MCD criterion to get the final estimator. For multiple linear regression, use the attractor with the smallest LMS, LTA, or LTS criterion to get the final estimator. For $500 \leq K \leq 3000$ and p not much larger than 5, the elemental set algorithm is very good for detecting certain "outlier configurations," including i) a mixture of two regression hyperplanes that cross in the center of the data cloud for MLR (not an outlier configuration since outliers are far from the bulk of the data) and ii) a cluster of outliers that can often be placed close enough to the bulk of the data so that an MB, RFCH, or RMVN DD plot can not detect the outliers. However, the outlier resistance of elemental algorithms decreases rapidly as p increases.

Suppose the data set has n cases where d are outliers and $n-d$ are "clean" (not outliers). The the outlier proportion $\gamma = d/n$. Suppose that K elemental sets are chosen with replacement and that it is desired to find K such that the probability P(that at least one of the elemental sets is clean) $\equiv P_1 \approx 1-\alpha$ where $\alpha = 0.05$ is a common choice. Then $P_1 = 1-$ P(none of the K elemental sets is clean) $\approx 1-[1-(1-\gamma)^p]^K$ by independence. Hence $\alpha \approx [1-(1-\gamma)^p]^K$ or

$$K \approx \frac{\log(\alpha)}{\log([1-(1-\gamma)^p])} \approx \frac{\log(\alpha)}{-(1-\gamma)^p} \tag{7.22}$$

using the approximation $\log(1-x) \approx -x$ for small x. Since $\log(0.05) \approx -3$, if $\alpha = 0.05$, then $K \approx \dfrac{3}{(1-\gamma)^p}$. Frequently a clean subset is wanted even if the contamination proportion $\gamma \approx 0.5$. Then for a 95% chance of obtaining at least one clean elemental set, $K \approx 3\ (2^p)$ elemental sets need to be drawn. If the start passes through an outlier, so does the attractor. For concentration algorithms for multivariate location and dispersion, if the start passes through a cluster of outliers, sometimes the attractor would be clean. See Figure 7.5–7.11.

Table 7.5 Largest p for a 95% Chance of a Clean Subsample.

					K			
γ	500	3000	10000	10^5	10^6	10^7	10^8	10^9
0.01	509	687	807	1036	1265	1494	1723	1952
0.05	99	134	158	203	247	292	337	382
0.10	48	65	76	98	120	142	164	186
0.15	31	42	49	64	78	92	106	120
0.20	22	30	36	46	56	67	77	87
0.25	17	24	28	36	44	52	60	68
0.30	14	19	22	29	35	42	48	55
0.35	11	16	18	24	29	34	40	45
0.40	10	13	15	20	24	29	33	38
0.45	8	11	13	17	21	25	28	32
0.50	7	9	11	15	18	21	24	28

Notice that the number of subsets K needed to obtain a clean elemental set with high probability is an exponential function of the number of predictors p but is free of n. Hawkins and Olive (2002) showed that if K is fixed and free of n, then the resulting elemental or concentration algorithm (that uses k concentration steps), is inconsistent and zero breakdown. See Theorem 7.21. Nevertheless, many practical estimators tend to use a value of K that is free of both n and p (e.g. $K = 500$ or $K = 3000$). Such algorithms include ALMS = FLMS = `lmsreg` and ALTS = FLTS = `ltsreg`. The "A" denotes that an algorithm was used. The "F" means that a fixed number of trial fits (K

elemental fits) was used and the criterion (LMS or LTS) was used to select the trial fit used in the final estimator.

To examine the outlier resistance of such inconsistent zero breakdown estimators, fix both K and the contamination proportion γ and then find the largest number of predictors p that can be in the model such that the probability of finding at least one clean elemental set is high. Given K and γ, P(at least one of K subsamples is clean) $= 0.95 \approx$ $1 - [1 - (1-\gamma)^p]^K$. Thus the largest value of p satisfies $\dfrac{3}{(1-\gamma)^p} \approx K$, or

$$p \approx \left\lfloor \frac{\log(3/K)}{\log(1-\gamma)} \right\rfloor \tag{7.23}$$

if the sample size n is very large. Again $\lfloor x \rfloor$ is the greatest integer function: $\lfloor 7.7 \rfloor = 7$.

Table 7.5 shows the largest value of p such that there is a 95% chance that at least one of K subsamples is clean using the approximation given by Equation (7.23). Hence if $p = 28$, even with one billion subsamples, there is a 5% chance that none of the subsamples will be clean if the contamination proportion $\gamma = 0.5$. Since clean elemental fits have great variability, an algorithm needs to produce many clean fits in order for the best fit to be good. When contamination is present, all K elemental sets could contain outliers. Hence basic resampling and concentration algorithms that only use K elemental starts are doomed to fail if γ and p are large.

The outlier resistance of elemental algorithms that use K elemental sets decreases rapidly as p increases. However, for $p < 10$, such elemental algorithms are often useful for outlier detection. They can perform better than MBA, trimmed views, and `rmreg2` if p is small and the outliers are close to the bulk of the data or if p is small and there is a mixture distribution: the bulk of the data follows one MLR model, but "outliers" and some of the clean data are fit well by another MLR model. For example, if there is one nontrivial predictor, suppose the plot of x versus Y looks like the letter X. Such a mixture distribution is not really an outlier configuration since outliers lie far from the bulk of the data. All practical estimators have outlier configurations where they perform poorly. If p is small, elemental algorithms tend to have trouble when there is a weak regression relationship for the bulk of the data and a cluster of outliers that are not good leverage points (do not fall near the hyperplane followed by the bulk of the data). The Buxton (1920) data set is an example.

Theorem 7.15. Let $h = p$ be the number of randomly selected cases in an elemental set, and let γ_o be the highest percentage of massive outliers that a resampling algorithm can detect reliably. If n is large, then

$$\gamma_o \approx \min\left(\frac{n-c}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h}\right) 100\%. \tag{7.24}$$

Proof. As in Remark 7.1, if the contamination proportion γ is fixed, then the probability of obtaining at least one clean subset of size h with high probability (say $1-\alpha = 0.8$) is given by $0.8 = 1 - [1-(1-\gamma)^h]^K$. Fix the number of starts K and solve this equation for γ. □

The value of γ_o depends on $c \geq n/2$ and h. To maximize γ_o, take $c \approx n/2$ and $h = p$. For example, with $K = 500$ starts, $n > 100$, and $h = p \leq 20$ the resampling algorithm should be able to detect up to 24% outliers provided every clean start is able to at least partially separate inliers (clean cases) from outliers. However, if $h = p = 50$, this proportion drops to 11%.

Definition 7.25. Let $\boldsymbol{b}_1, ..., \boldsymbol{b}_J$ be J estimators of $\boldsymbol{\beta}$. Assume that $J \geq 2$ and that OLS is included. A *fit-fit* (FF) plot is a scatterplot matrix of the fitted values $\widehat{Y}(\boldsymbol{b}_1), ..., \widehat{Y}(\boldsymbol{b}_J)$. Often Y is also included in the top or bottom row of the FF plot to see the response plots. A *residual-residual* (RR) plot is a scatterplot matrix of the residuals $r(\boldsymbol{b}_1), ..., r(\boldsymbol{b}_J)$. Often $\hat{Y}$ is also included in the top or bottom row of the RR plot to see the residual plots.

If the multiple linear regression model holds, if the predictors are bounded, and if all J regression estimators are consistent estimators of $\boldsymbol{\beta}$, then the subplots in the FF and RR plots should be linear with a correlation tending to one as the sample size n increases. To prove this claim, let the ith residual from the jth fit $\boldsymbol{b}_j$ be $r_i(\boldsymbol{b}_j) = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}_j$ where $(Y_i, \boldsymbol{x}_i^T)$ is the ith observation. Similarly, let the ith fitted value from the jth fit be $\widehat{Y}_i(\boldsymbol{b}_j) = \boldsymbol{x}_i^T \boldsymbol{b}_j$. Then

$$\|r_i(\boldsymbol{b}_1) - r_i(\boldsymbol{b}_2)\| = \|\widehat{Y}_i(\boldsymbol{b}_1) - \widehat{Y}_i(\boldsymbol{b}_2)\| = \|\boldsymbol{x}_i^T(\boldsymbol{b}_1 - \boldsymbol{b}_2)\|$$
$$\leq \|\boldsymbol{x}_i\| \ (\|\boldsymbol{b}_1 - \boldsymbol{\beta}\| + \|\boldsymbol{b}_2 - \boldsymbol{\beta}\|). \tag{7.25}$$

The FF plot is a powerful way for comparing fits. The commonly suggested alternative is to look at a table of the estimated coefficients, but coefficients can differ greatly while yielding similar fits if some of the predictors are highly correlated or if several of the predictors are independent of the response. See Olive (2017b, pp. 408-412).

Table 7.6 compares the TV, MBA (for MLR), `lmsreg`, `ltsreg`, L_1, and OLS estimators on 7 data sets available from the text's website. The column headers give the file name while the remaining rows of the table give the sample size n, the number of predictors p, the amount of trimming M used by the TV estimator, the correlation of the residuals from the TV estimator with the corresponding alternative estimator, and the cases that were outliers. If the correlation was greater than 0.9, then the method was effective in detecting the outliers, and the method failed, otherwise. Sometimes the trimming percentage M for the TV estimator was picked after fitting the bulk of the data in order to find the good leverage points and outliers. Each model included a constant.

Table 7.6 Summaries for Seven Data Sets, the Correlations of the Residuals from TV(M) and the Alternative Method are Given in the 1st 5 Rows

Method	Buxton	Gladstone	glado	hbk	major	nasty	wood
MBA	0.997	1.0	0.455	0.960	1.0	-0.004	0.9997
LMSREG	-0.114	0.671	0.938	0.977	0.981	0.9999	0.9995
LTSREG	-0.048	0.973	0.468	0.272	0.941	0.028	0.214
L1	-0.016	0.983	0.459	0.316	0.979	0.007	0.178
OLS	0.011	1.0	0.459	0.780	1.0	0.009	0.227
outliers	61-65	none	115	1-10	3,44	2,6,...,30	4,6,8,19
n	87	267	267	75	112	32	20
p	5	7	7	4	6	5	6
M	70	0	30	90	0	90	20

Notice that the TV, MBA, and OLS estimators were the same for the Gladstone (1905) data and for the Tremearne (1911) *major* data which had two small Y–outliers. For the Gladstone data, there is a cluster of infants that are good leverage points, and we attempt to predict *brain weight* with the head measurements *height, length, breadth, size,* and *cephalic index.* Originally, the variable *length* was incorrectly entered as 109 instead of 199 for case 115, and the *glado* data contains this outlier. In 1997, `lmsreg` was not able to detect the outlier while `ltsreg` did. Due to changes in the *Splus* 2000 code, `lmsreg` detected the outlier but `ltsreg` did not. These two functions change often, not always for the better.

To end this section, we describe resistant regression with the RMVN set U or `covmb2` set B in more detail. Assume that predictor transformations have been performed to make a $p \times 1$ vector of predictors $\boldsymbol{x}$, and that $\boldsymbol{w}$ consists of $k \leq p$ continuous predictor variables that are linearly related. Find the RMVN set based on the $\boldsymbol{w}$ to obtain n_u cases $(\boldsymbol{y}_{ci}, \boldsymbol{x}_{ci})$, and then run the regression method on the cleaned data. Often the theory of the method applies to the cleaned data set since $\boldsymbol{y}$ was not used to pick the subset of the data. Efficiency can be much lower since n_u cases are used where $n/2 \leq n_u \leq n$, and the trimmed cases tend to be the "farthest" from the center of $\boldsymbol{w}$. The method will have the most outlier resistance if $k = p - 1$ if there is a trivial predictor $X_1 \equiv 1$.

In *R*, assume Y is the vector of response variables, x is the data matrix of the predictors (often not including the trivial predictor), and w is the data matrix of the $\boldsymbol{w}_i$. Then the following *R* commands can be used to get the cleaned data set. We could use the `covmb2` set B instead of the RMVN set U computed from the $\boldsymbol{w}$ by replacing the command *getu(w)* by `getB(w)`.

```
indx <- getu(w)$indx  #often w = x
Yc <- Y[indx]
Xc <- x[indx,]
#example
```

```
indx <- getu(buxx)$indx
Yc <- buxy[indx]
Xc <- buxx[indx,]
outr <- lsfit(Xc,Yc)
MLRplot(Xc,Yc) #right click Stop twice
```

7.6 Robust Regression

This section will consider the breakdown of a regression estimator and then develop the practical high breakdown `hbreg` estimator.

7.6.1 MLR Breakdown and Equivariance

Breakdown and equivariance properties have received considerable attention in the literature. Several of these properties involve transformations of the data, and are discussed below. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the original data, then the vector of the coefficient estimates is

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X},\boldsymbol{Y}) = T(\boldsymbol{X},\boldsymbol{Y}), \tag{7.26}$$

the vector of predicted values is

$$\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{Y}}(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{X},\boldsymbol{Y}), \tag{7.27}$$

and the vector of residuals is

$$\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}. \tag{7.28}$$

If the design matrix $\boldsymbol{X}$ is transformed into $\boldsymbol{W}$ and the vector of dependent variables $\boldsymbol{Y}$ is transformed into $\boldsymbol{Z}$, then $(\boldsymbol{W},\boldsymbol{Z})$ is the new data set.

Definition 7.26. Regression Equivariance: Let $\boldsymbol{u}$ be any $p \times 1$ vector. Then $\widehat{\boldsymbol{\beta}}$ is regression equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X},\boldsymbol{Y}+\boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X},\boldsymbol{Y}+\boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X},\boldsymbol{Y}) + \boldsymbol{u} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X},\boldsymbol{Y}) + \boldsymbol{u}. \tag{7.29}$$

Hence if $\boldsymbol{W} = \boldsymbol{X}$ and $\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}$, then $\widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{Y}} + \boldsymbol{X}\boldsymbol{u}$ and $\boldsymbol{r}(\boldsymbol{W},\boldsymbol{Z}) = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{r}(\boldsymbol{X},\boldsymbol{Y})$. Note that the residuals are invariant under this type of transformation, and note that if $\boldsymbol{u} = -\widehat{\boldsymbol{\beta}}$, then regression equivariance implies that we should not find any linear structure if we regress the residuals on $\boldsymbol{X}$. Also see Problem 7.2.

Definition 7.27. Scale Equivariance: Let c be any scalar. Then $\widehat{\boldsymbol{\beta}}$ is scale equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, c\boldsymbol{Y}) = T(\boldsymbol{X}, c\boldsymbol{Y}) = cT(\boldsymbol{X}, \boldsymbol{Y}) = c\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{7.30}$$

Hence if $\boldsymbol{W} = \boldsymbol{X}$ and $\boldsymbol{Z} = c\boldsymbol{Y}$, then $\widehat{\boldsymbol{Z}} = c\widehat{\boldsymbol{Y}}$ and $\boldsymbol{r}(\boldsymbol{X}, c\boldsymbol{Y}) = c\,\boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. Scale equivariance implies that if the Y_i's are stretched, then the fits and the residuals should be stretched by the same factor.

Definition 7.28. Affine Equivariance: Let $\boldsymbol{A}$ be any $p \times p$ nonsingular matrix. Then $\widehat{\boldsymbol{\beta}}$ is affine equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{XA}, \boldsymbol{Y}) = T(\boldsymbol{XA}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}T(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{7.31}$$

Hence if $\boldsymbol{W} = \boldsymbol{XA}$ and $\boldsymbol{Z} = \boldsymbol{Y}$, then $\widehat{\boldsymbol{Z}} = \boldsymbol{W}\widehat{\boldsymbol{\beta}}(\boldsymbol{XA}, \boldsymbol{Y}) =$ $\boldsymbol{XAA}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{Y}}$, and $\boldsymbol{r}(\boldsymbol{XA}, \boldsymbol{Y}) = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. Note that both the predicted values and the residuals are invariant under an affine transformation of the predictor variables.

Definition 7.29. Permutation Invariance: Let $\boldsymbol{P}$ be an $n \times n$ permutation matrix. Then $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{PP}^T = \boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is an $n \times n$ identity matrix and the superscript T denotes the transpose of a matrix. Then $\widehat{\boldsymbol{\beta}}$ is permutation invariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{PX}, \boldsymbol{PY}) = T(\boldsymbol{PX}, \boldsymbol{PY}) = T(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{7.32}$$

Hence if $\boldsymbol{W} = \boldsymbol{PX}$ and $\boldsymbol{Z} = \boldsymbol{PY}$, then $\widehat{\boldsymbol{Z}} = \boldsymbol{P}\widehat{\boldsymbol{Y}}$ and $\boldsymbol{r}(\boldsymbol{PX}, \boldsymbol{PY}) = \boldsymbol{P}\ \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. If an estimator is not permutation invariant, then swapping rows of the $n \times (p+1)$ augmented matrix $(\boldsymbol{X}, \boldsymbol{Y})$ will change the estimator. Hence the case number is important. If the estimator is permutation invariant, then the position of the case in the data cloud is of primary importance. Resampling algorithms are not permutation invariant because permuting the data causes different subsamples to be drawn.

Remark 7.7. OLS has the above invariance properties, but most Statistical Learning alternatives such as lasso and ridge regression do not have all four properties. Hence Remark 5.1 is used to fit the data with $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$. Then obtain $\hat{\boldsymbol{\beta}}$ from $\hat{\boldsymbol{\eta}}$.

The remainder of this subsection gives a standard definition of breakdown and then shows that if the median absolute residual is bounded in the presence of high contamination, then the regression estimator has a high breakdown value. The following notation will be useful. Let $\boldsymbol{W}$ denote the data matrix where the ith row corresponds to the ith case. For regression, $\boldsymbol{W}$ is the $n \times (p+1)$ matrix with ith row $(\boldsymbol{x}_i^T, Y_i)$. Let $\boldsymbol{W}_d^n$ denote the data matrix where any d_n of the cases have been replaced by arbitrarily bad contaminated cases. Then the contamination fraction is $\gamma \equiv \gamma_n = d_n/n$, and the breakdown value of $\hat{\boldsymbol{\beta}}$ is the smallest value of γ_n needed to make $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large.

Definition 7.30. Let $1 \le d_n \le n$. If $T(\boldsymbol{W})$ is a $p \times 1$ vector of regression coefficients, then the *breakdown value* of T is

$$B(T, \boldsymbol{W}) = \min \left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty \right\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$.

Definition 7.31. *High breakdown* regression estimators have $\gamma_n \to 0.5$ as $n \to \infty$ if the clean (uncontaminated) data are in *general position*: any p clean cases give a unique estimate of $\boldsymbol{\beta}$. Estimators are *zero breakdown* if $\gamma_n \to 0$ and *positive breakdown* if $\gamma_n \to \gamma > 0$ as $n \to \infty$.

The following result greatly simplifies some breakdown proofs and shows that a regression estimator basically breaks down if the median absolute residual MED($|r_i|$) can be made arbitrarily large. The result implies that if the breakdown value ≤ 0.5, breakdown can be computed using the median absolute residual MED($|r_i|(\boldsymbol{W}_d^n)$) instead of $\|T(\boldsymbol{W}_d^n)\|$. Similarly $\hat{\boldsymbol{\beta}}$ is high breakdown if the median squared residual or the c_nth largest absolute residual $|r_i|_{(c_n)}$ or squared residual $r^2_{(c_n)}$ stay bounded under high contamination where $c_n \approx n/2$. Note that $\|\hat{\boldsymbol{\beta}}\| \equiv \|\hat{\boldsymbol{\beta}}(\boldsymbol{W}_d^n)\| \le M$ for some constant M that depends on T and $\boldsymbol{W}$ but not on the outliers if the number of outliers d_n is less than the smallest number of outliers needed to cause breakdown.

Theorem 7.16. If the breakdown value ≤ 0.5, computing the breakdown value using the median absolute residual MED($|r_i|(\boldsymbol{W}_d^n)$) instead of $\|T(\boldsymbol{W}_d^n)\|$ is asymptotically equivalent to using Definition 7.30.

Proof. Consider any contaminated data set $\boldsymbol{W}_d^n$ with ith row $(\boldsymbol{w}_i^T, Z_i)^T$. If the regression estimator $T(\boldsymbol{W}_d^n) = \hat{\boldsymbol{\beta}}$ satisfies $\|\hat{\boldsymbol{\beta}}\| \le M$ for some constant M if $d < d_n$, then the median absolute residual MED($|Z_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{w}_i|$) is bounded by $\max_{i=1,\ldots,n} |Y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i| \le \max_{i=1,\ldots,n}[|Y_i| + \sum_{j=1}^p M|x_{i,j}|]$ if $d_n < n/2$.

If the median absolute residual is bounded by M when $d < d_n$, then $\|\hat{\boldsymbol{\beta}}\|$ is bounded provided fewer than half of the cases line on the hyperplane (and so have absolute residual of 0), as shown next. Now suppose that $\|\hat{\boldsymbol{\beta}}\| = \infty$. Since the absolute residual is the vertical distance of the observation from the hyperplane, the absolute residual $|r_i| = 0$ if the ith case lies on the regression hyperplane, but $|r_i| = \infty$ otherwise. Hence MED($|r_i|$) $= \infty$ if fewer than half of the cases lie on the regression hyperplane. This will occur unless the proportion of outliers $d_n/n > (n/2 - q)/n \to 0.5$ as $n \to \infty$ where q is the number of "good" cases that lie on a hyperplane of lower dimension than p.

In the literature it is usually assumed that the original data are in *general position*: $q = p - 1$. □

Suppose that the clean data are in general position and that the number of outliers is less than the number needed to make the median absolute residual and $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large. If the $\boldsymbol{x}_i$ are fixed, and the outliers are moved up and down by adding a large positive or negative constant to the Y values of the outliers, then for high breakdown (HB) estimators, $\hat{\boldsymbol{\beta}}$ and MED($|r_i|$) stay bounded where the bounds depend on the clean data $\boldsymbol{W}$ but not on the outliers even if the number of outliers is nearly as large as $n/2$. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large.

If the Y_i's are fixed, arbitrarily large $\boldsymbol{x}$-outliers tend to drive the slope estimates to 0, not ∞. If both $\boldsymbol{x}$ and Y can be varied, then a cluster of outliers can be moved arbitrarily far from the bulk of the data but may still have small residuals. For example, move the outliers along the regression hyperplane formed by the clean cases.

If the $(\boldsymbol{x}_i^T, Y_i)$ are in general position, then the contamination could be such that $\hat{\boldsymbol{\beta}}$ passes exactly through $p - 1$ "clean" cases and d_n "contaminated" cases. Hence $d_n + p - 1$ cases could have absolute residuals equal to zero with $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large (but finite). Nevertheless, if T possesses reasonable equivariant properties and $\|T(\boldsymbol{W}_d^n)\|$ is replaced by the median absolute residual in the definition of breakdown, then the two breakdown values are asymptotically equivalent. (If $T(\boldsymbol{W}) \equiv \boldsymbol{0}$, then T is neither regression nor affine equivariant. The breakdown value of T is one, but the median absolute residual can be made arbitrarily large if the contamination proportion is greater than $n/2$.)

If the Y_i's are fixed, arbitrarily large $\boldsymbol{x}$-outliers will rarely drive $\|\hat{\boldsymbol{\beta}}\|$ to ∞. The $\boldsymbol{x}$-outliers can drive $\|\hat{\boldsymbol{\beta}}\|$ to ∞ if they can be constructed so that the estimator is no longer defined, e.g. so that $\boldsymbol{X}^T\boldsymbol{X}$ is nearly singular. The examples following some results on norms may help illustrate these points.

Definition 7.32. Let $\boldsymbol{y}$ be an $n \times 1$ vector. Then $\|\boldsymbol{y}\|$ is a *vector norm* if
vn1) $\|\boldsymbol{y}\| \geq 0$ for every $\boldsymbol{y} \in \mathbb{R}^n$ with equality iff $\boldsymbol{y}$ is the zero vector,
vn2) $\|a\boldsymbol{y}\| = |a| \ \|\boldsymbol{y}\|$ for all $\boldsymbol{y} \in \mathbb{R}^n$ and for all scalars a, and
vn3) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$.

Definition 7.33. Let $\boldsymbol{G}$ be an $n \times p$ matrix. Then $\|\boldsymbol{G}\|$ is a *matrix norm* if
mn1) $\|\boldsymbol{G}\| \geq 0$ for every $n \times p$ matrix $\boldsymbol{G}$ with equality iff $\boldsymbol{G}$ is the zero matrix,
mn2) $\|a\boldsymbol{G}\| = |a| \ \|\boldsymbol{G}\|$ for all scalars a, and
mn3) $\|\boldsymbol{G} + \boldsymbol{H}\| \leq \|\boldsymbol{G}\| + \|\boldsymbol{H}\|$ for all $n \times p$ matrices $\boldsymbol{G}$ and $\boldsymbol{H}$.

Example 7.13. The *q-norm* of a vector $\boldsymbol{y}$ is $\|\boldsymbol{y}\|_q = (|y_1|^q + \cdots + |y_n|^q)^{1/q}$. In particular, $\|\boldsymbol{y}\|_1 = |y_1| + \cdots + |y_n|$, the *Euclidean norm* $\|\boldsymbol{y}\|_2 = \sqrt{y_1^2 + \cdots + y_n^2}$, and $\|\boldsymbol{y}\|_\infty = \max_i \ |y_i|$. Given a matrix $\boldsymbol{G}$ and a vector norm $\|\boldsymbol{y}\|_q$ the *q-norm* or *subordinate matrix norm* of matrix $\boldsymbol{G}$ is $\|\boldsymbol{G}\|_q = \max_{\boldsymbol{y} \neq \boldsymbol{0}} \frac{\|\boldsymbol{G}\boldsymbol{y}\|_q}{\|\boldsymbol{y}\|_q}$. It can be shown that the *maximum column sum norm* $\|\boldsymbol{G}\|_1 = \max_{1 \leq j \leq p} \sum_{i=1}^{n} |g_{ij}|$, the *maximum row sum norm* $\|\boldsymbol{G}\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{p} |g_{ij}|$, and the *spectral norm* $\|\boldsymbol{G}\|_2 = \sqrt{\text{maximum eigenvalue of } \boldsymbol{G}^T\boldsymbol{G}}$. The *Frobenius norm*

$$\|\boldsymbol{G}\|_F = \sqrt{\sum_{j=1}^{p} \sum_{i=1}^{n} |g_{ij}|^2} = \sqrt{\text{trace}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})}.$$

Several useful results involving matrix norms will be used. First, for any subordinate matrix norm, $\|\boldsymbol{G}\boldsymbol{y}\|_q \leq \|\boldsymbol{G}\|_q \ \|\boldsymbol{y}\|_q$. Let $J = J_m = \{m_1, ..., m_p\}$ denote the p cases in the mth elemental fit $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$. Then for any elemental fit $\boldsymbol{b}_J$ (suppressing $q = 2$),

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}(\boldsymbol{X}_J\boldsymbol{\beta} + \boldsymbol{e}_J) - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}\boldsymbol{e}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \ \|\boldsymbol{e}_J\|. \quad (7.33)$$

The following results (Golub and Van Loan 1989, pp. 57, 80) on the Euclidean norm are useful. Let $0 \leq \sigma_p \leq \sigma_{p-1} \leq \cdots \leq \sigma_1$ denote the singular values of $\boldsymbol{X}_J = (x_{mi,j})$. Then

$$\|\boldsymbol{X}_J^{-1}\| = \frac{\sigma_1}{\sigma_p \|\boldsymbol{X}_J\|}, \quad (7.34)$$

$$\max_{i,j} |x_{mi,j}| \leq \|\boldsymbol{X}_J\| \leq p \ \max_{i,j} |x_{mi,j}|, \text{ and} \quad (7.35)$$

$$\frac{1}{p \ \max_{i,j} |x_{mi,j}|} \leq \frac{1}{\|\boldsymbol{X}_J\|} \leq \|\boldsymbol{X}_J^{-1}\|. \quad (7.36)$$

From now on, unless otherwise stated, we will use the spectral norm as the matrix norm and the Euclidean norm as the vector norm.

Example 7.14. Suppose the response values Y are near 0. Consider the fit from an elemental set: $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$ and examine Equations (7.34), (7.35), and (7.36). Now $\|\boldsymbol{b}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \ \|\boldsymbol{Y}_J\|$, and *since x-outliers make $\|\boldsymbol{X}_J\|$ large, x-outliers tend to drive $\|\boldsymbol{X}_J^{-1}\|$ and $\|\boldsymbol{b}_J\|$ towards zero not towards ∞.* The x-outliers may make $\|\boldsymbol{b}_J\|$ large if they can make the trial design $\|\boldsymbol{X}_J\|$ nearly singular. Notice that Euclidean norm $\|\boldsymbol{b}_J\|$ can easily be made large if one or more of the elemental response variables is driven far away from zero.

Example 7.15. Without loss of generality, assume that the clean Y's are contained in an interval $[a, f]$ for some a and f. Assume that the regression model contains an intercept β_1. Then there exists an estimator $\hat{\boldsymbol{\beta}}_M$ of $\boldsymbol{\beta}$ such that $\|\hat{\boldsymbol{\beta}}_M\| \leq \max(|a|, |f|)$ if $d_n < n/2$.

Proof. Let $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ and $\text{MAD}(n) = \text{MAD}(Y_1, ..., Y_n)$. Take $\hat{\boldsymbol{\beta}}_M = (\text{MED}(n), 0, ..., 0)^T$. Then $\|\hat{\boldsymbol{\beta}}_M\| = |\text{MED}(n)| \leq \max(|a|, |f|)$. Note that the median absolute residual for the fit $\hat{\boldsymbol{\beta}}_M$ is equal to the median absolute deviation $\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, i = 1, ..., n) \leq f - a$ if $d_n < \lfloor (n+1)/2 \rfloor$. □

Note that $\hat{\boldsymbol{\beta}}_M$ is a poor high breakdown estimator of $\boldsymbol{\beta}$ and $\hat{Y}_i(\hat{\boldsymbol{\beta}}_M)$ tracks the Y_i very poorly. If the data are in general position, a high breakdown regression estimator is an estimator which has a bounded median absolute residual even when close to half of the observations are arbitrary. Rousseeuw and Leroy (1987, pp. 29, 206) conjectured that high breakdown regression estimators can not be computed cheaply, and that if the algorithm is also affine equivariant, then the complexity of the algorithm must be at least $O(n^p)$. The following theorem shows that these two conjectures are false.

Theorem 7.17. If the clean data are in general position and the model has an intercept, then a scale and affine equivariant high breakdown estimator $\hat{\boldsymbol{\beta}}_w$ can be found by computing OLS on the set of cases that have $Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm w \ \text{MAD}(Y_1, ..., Y_n)]$ where $w \geq 1$ (so at least half of the cases are used).

Proof. Note that $\hat{\boldsymbol{\beta}}_w$ is obtained by computing OLS on the set J of the n_j cases which have

$$Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm w\text{MAD}(Y_1, ..., Y_n)] \equiv [\text{MED}(n) \pm w\text{MAD}(n)]$$

where $w \geq 1$ (to guarantee that $n_j \geq n/2$). Consider the estimator $\hat{\boldsymbol{\beta}}_M = (\text{MED}(n), 0, ..., 0)^T$ which yields the predicted values $\hat{Y}_i \equiv \text{MED}(n)$. The squared residual $r_i^2(\hat{\boldsymbol{\beta}}_M) \leq (w \ \text{MAD}(n))^2$ if the ith case is in J. Hence the weighted LS fit $\hat{\boldsymbol{\beta}}_w$ is the OLS fit to the cases in J and has

$$\sum_{i \in J} r_i^2(\hat{\boldsymbol{\beta}}_w) \leq n_j (w \ \text{MAD}(n))^2.$$

Thus

$$\text{MED}(|r_1(\hat{\boldsymbol{\beta}}_w)|, ..., |r_n(\hat{\boldsymbol{\beta}}_w)|) \leq \sqrt{n_j} \ w \ \text{MAD}(n) < \sqrt{n} \ w \ \text{MAD}(n) < \infty.$$

Thus the estimator $\hat{\boldsymbol{\beta}}_w$ has a median absolute residual bounded by $\sqrt{n} \ w \ \text{MAD}(Y_1, ..., Y_n)$. Hence $\hat{\boldsymbol{\beta}}_w$ is high breakdown, and it is affine equivariant since the design is not used to choose the observations. It is scale equivariant since for constant $c = 0$, $\hat{\boldsymbol{\beta}}_w = \mathbf{0}$, and for $c \neq 0$ the set of cases used remains the same under scale transformations and OLS is scale equivariant. □

Note that if w is huge and MAD(n) $\neq 0$, then the high breakdown estimator $\hat{\boldsymbol{\beta}}_w$ and $\hat{\boldsymbol{\beta}}_{OLS}$ will be the same for most data sets. Thus high breakdown estimators can be very nonrobust. Even if $w = 1$, the HB estimator $\hat{\boldsymbol{\beta}}_w$ only resists large Y outliers.

An ALTA concentration algorithm uses the L_1 estimator instead of OLS in the concentration step and uses the LTA criterion. Similarly an ALMS concentration algorithm uses the L_∞ estimator and the LMS criterion.

Theorem 7.18. If the clean data are in general position and if a high breakdown start is added to an ALTA, ALTS, or ALMS concentration algorithm, then the resulting estimator is HB.

Proof. Concentration reduces (or does not increase) the corresponding HB criterion that is based on $c_n \geq n/2$ absolute residuals, so the median absolute residual of the resulting estimator is bounded as long as the criterion applied to the HB estimator is bounded. □

For example, consider the LTS(c_n) criterion. Suppose the ordered squared residuals from the high breakdown mth start $\boldsymbol{b}_{0m}$ are obtained. If the data are in general position, then $Q_{LTS}(\boldsymbol{b}_{0m})$ is bounded even if the number of outliers d_n is nearly as large as $n/2$. Then $\boldsymbol{b}_{1m}$ is simply the OLS fit to the cases corresponding to the c_n smallest squared residuals $r^2_{(i)}(\boldsymbol{b}_{0m})$ for $i = 1, ..., c_n$. Denote these cases by $i_1, ..., i_{c_n}$. Then $Q_{LTS}(\boldsymbol{b}_{1m}) =$

$$\sum_{i=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{0m}) = \sum_{j=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{0m}) = Q_{LTS}(\boldsymbol{b}_{0m})$$

where the second inequality follows from the definition of the OLS estimator. Hence concentration steps reduce or at least do not increase the LTS criterion. If $c_n = (n+1)/2$ for n odd and $c_n = 1+n/2$ for n even, then the LTS criterion is bounded iff the median squared residual is bounded.

Theorem 7.18 can be used to show that the following two estimators are high breakdown. The estimator $\hat{\boldsymbol{\beta}}_B$ is the high breakdown attractor used by the $\sqrt{n}$ consistent high breakdown `hbreg` estimator of Definition 7.35.

Definition 7.34. Make an OLS fit to the $c_n \approx n/2$ cases whose Y values are closest to the MED($Y_1, ..., Y_n$) $\equiv$ MED(n) and use this fit as the start for concentration. Define $\hat{\boldsymbol{\beta}}_B$ to be the attractor after k concentration steps. Define $\boldsymbol{b}_{k,B} = 0.9999\hat{\boldsymbol{\beta}}_B$.

Theorem 7.19. If the clean data are in general position, then $\hat{\boldsymbol{\beta}}_B$ and $\boldsymbol{b}_{k,B}$ are high breakdown regression estimators.

Proof. The start can be taken to be $\hat{\boldsymbol{\beta}}_w$ with $w = 1$ from Theorem 7.17. Since the start is high breakdown, so is the attractor $\hat{\boldsymbol{\beta}}_B$ by Theorem 7.18. Multiplying a HB estimator by a positive constant does not change the breakdown value, so $\boldsymbol{b}_{k,B}$ is HB. □

The following result shows that it is easy to make a HB estimator that is asymptotically equivalent to a consistent estimator on a large class of iid zero mean symmetric error distributions, although the outlier resistance of the HB estimator is poor. The following result may not hold if $\hat{\boldsymbol{\beta}}_C$ estimates $\boldsymbol{\beta}_C$ and $\hat{\boldsymbol{\beta}}_{LMS}$ estimates $\boldsymbol{\beta}_{LMS}$ where $\boldsymbol{\beta}_C \neq \boldsymbol{\beta}_{LMS}$. Then $\boldsymbol{b}_{k,B}$ could have a smaller median squared residual than $\hat{\boldsymbol{\beta}}_C$ even if there are no outliers. The two parameter vectors could differ because the constant term is different if the error distribution is not symmetric. For a large class of symmetric error distributions, $\boldsymbol{\beta}_{LMS} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_C \equiv \boldsymbol{\beta}$, then the ratio $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}))/\text{MED}(r_i^2(\boldsymbol{\beta})) \to 1$ as $n \to \infty$ for any consistent estimator of $\boldsymbol{\beta}$. The estimator below has two attractors, $\hat{\boldsymbol{\beta}}_C$ and $\boldsymbol{b}_{k,B}$, and the probability that the final estimator $\hat{\boldsymbol{\beta}}_D$ is equal to $\hat{\boldsymbol{\beta}}_C$ goes to one under the strong assumption that the error distribution is such that both $\hat{\boldsymbol{\beta}}_C$ and $\hat{\boldsymbol{\beta}}_{LMS}$ are consistent estimators of $\boldsymbol{\beta}$.

Theorem 7.20. Assume the clean data are in general position, and that the LMS estimator is a consistent estimator of $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}_C$ be any practical consistent estimator of $\boldsymbol{\beta}$, and let $\hat{\boldsymbol{\beta}}_D = \hat{\boldsymbol{\beta}}_C$ if $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}_C)) \leq \text{MED}(r_i^2(\boldsymbol{b}_{k,B}))$. Let $\hat{\boldsymbol{\beta}}_D = \boldsymbol{b}_{k,B}$, otherwise. Then $\hat{\boldsymbol{\beta}}_D$ is a HB estimator that is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$.

Proof. The estimator is HB since the median squared residual of $\hat{\boldsymbol{\beta}}_D$ is no larger than that of the HB estimator $\boldsymbol{b}_{k,B}$. Since $\hat{\boldsymbol{\beta}}_C$ is consistent, $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}_C)) \to \text{MED}(e^2)$ in probability where $\text{MED}(e^2)$ is the population median of the squared error e^2. Since the LMS estimator is consistent, the probability that $\hat{\boldsymbol{\beta}}_C$ has a smaller median squared residual than the biased estimator $\hat{\boldsymbol{\beta}}_{k,B}$ goes to 1 as $n \to \infty$. Hence $\hat{\boldsymbol{\beta}}_D$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$. □

The elemental concentration and elemental resampling algorithms use K elemental fits where K is a fixed number that does not depend on the sample size n, e.g. $K = 500$. See Definitions 7.12 and 7.24. Note that an estimator can not be consistent for θ unless the number of randomly selected cases goes to ∞, except in degenerate situations. The following theorem shows the widely used elemental estimators are zero breakdown estimators. (If $K = K_n \to \infty$, then the elemental estimator is zero breakdown if $K_n = o(n)$. A necessary condition for the elemental basic resampling estimator to be consistent is $K_n \to \infty$.)

Theorem 7.21: a) The elemental basic resampling algorithm estimators are inconsistent. b) The elemental concentration and elemental basic resampling algorithm estimators are zero breakdown.

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

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

7.6.2 A Practical High Breakdown Consistent Estimator

Olive and Hawkins (2011) showed that the practical `hbreg` estimator is a high breakdown $\sqrt{n}$ consistent robust estimator that is asymptotically equivalent to the least squares estimator for many error distributions. This subsection follows Olive (2017b, pp. 420-423).

The outlier resistance of the `hbreg` estimator is not very good, but roughly comparable to the best of the practical "robust regression" estimators available in *R* packages as of 2019. The estimator is of some interest since it proved that practical high breakdown consistent estimators are possible. Other practical regression estimators that claim to be high breakdown and consistent appear to be zero breakdown because they use the zero breakdown elemental concentration algorithm. See Theorem 7.21.

The following theorem is powerful because it does not depend on the criterion used to choose the attractor. Suppose there are K consistent estimators $\hat{\boldsymbol{\beta}}_j$ of $\boldsymbol{\beta}$, each with the same rate n^δ. If $\hat{\boldsymbol{\beta}}_A$ is an estimator obtained by choosing one of the K estimators, then $\hat{\boldsymbol{\beta}}_A$ is a consistent estimator of $\boldsymbol{\beta}$ with rate n^δ by Pratt (1959). See Theorem 1.21.

Theorem 7.22. Suppose the algorithm estimator chooses an attractor as the final estimator where there are K attractors and K is fixed.

i) If all of the attractors are consistent, then the algorithm estimator is consistent.

ii) If all of the attractors are consistent with the same rate, e.g., n^δ where $0 < \delta \leq 0.5$, then the algorithm estimator is consistent with the same rate as the attractors.

iii) If all of the attractors are high breakdown, then the algorithm estimator is high breakdown.

Proof. i) Choosing from K consistent estimators results in a consistent estimator, and ii) follows from Pratt (1959). iii) Let $\gamma_{n,i}$ be the breakdown value of the ith attractor if the clean data are in general position. The breakdown value γ_n of the algorithm estimator can be no lower than that of the worst attractor: $\gamma_n \geq \min(\gamma_{n,1}, ..., \gamma_{n,K}) \to 0.5$ as $n \to \infty$. $\square$

The consistency of the algorithm estimator changes dramatically if K is fixed but the start size $h = h_n = g(n)$ where $g(n) \to \infty$. In particular, if K starts with rate $n^{1/2}$ are used, the final estimator also has rate $n^{1/2}$. The drawback to these algorithms is that they may not have enough outlier resistance. Notice that the basic resampling result below is free of the criterion.

Theorem 7.23. Suppose $K_n \equiv K$ starts are used and that all starts have subset size $h_n = g(n) \uparrow \infty$ as $n \to \infty$. Assume that the estimator applied to the subset has rate n^δ.
i) For the h_n-set basic resampling algorithm, the algorithm estimator has rate $[g(n)]^\delta$.
ii) Under regularity conditions (e.g. given by He and Portnoy 1992), the k–step CLTS estimator has rate $[g(n)]^\delta$.

Proof. i) The $h_n = g(n)$ cases are randomly sampled without replacement. Hence the classical estimator applied to these $g(n)$ cases has rate $[g(n)]^\delta$. Thus all K starts have rate $[g(n)]^\delta$, and the result follows by Pratt (1959). ii) By He and Portnoy (1992), all K attractors have $[g(n)]^\delta$ rate, and the result follows by Pratt (1959). □

Remark 7.8. Theorem 7.16 shows that $\hat{\boldsymbol{\beta}}$ is HB if the median absolute or squared residual (or $|r(\hat{\boldsymbol{\beta}})|_{(c_n)}$ or $r^2_{(c_n)}$ where $c_n \approx n/2$) stays bounded under high contamination. Let $Q_L(\hat{\boldsymbol{\beta}}_H)$ denote the LMS, LTS, or LTA criterion for an estimator $\hat{\boldsymbol{\beta}}_H$; therefore, the estimator $\hat{\boldsymbol{\beta}}_H$ is high breakdown if and only if $Q_L(\hat{\boldsymbol{\beta}}_H)$ is bounded for d_n near $n/2$ where $d_n < n/2$ is the number of outliers. The concentration operator refines an initial estimator by successively reducing the LTS criterion. If $\hat{\boldsymbol{\beta}}_F$ refers to the final estimator (attractor) obtained by applying concentration to some starting estimator $\hat{\boldsymbol{\beta}}_H$ that is high breakdown, then since $Q_{LTS}(\hat{\boldsymbol{\beta}}_F) \leq Q_{LTS}(\hat{\boldsymbol{\beta}}_H)$, applying concentration to a high breakdown start results in a high breakdown attractor. See Theorem 7.18.

High breakdown estimators are, however, not necessarily useful for detecting outliers. Suppose $\gamma_n < 0.5$. On the one hand, if the $\boldsymbol{x}_i$ are fixed, and the outliers are moved up and down parallel to the Y axis, then for high breakdown estimators, $\hat{\boldsymbol{\beta}}$ and MED$(|r_i|)$ will be bounded. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large, suggesting that the high breakdown estimator is useful for outlier detection. On the other hand, if the Y_i's are fixed at any values and the $\boldsymbol{x}$ values perturbed, sufficiently large $\boldsymbol{x}$-outliers tend to drive the slope estimates to 0, not ∞. For many estimators, including LTS, LMS, and LTA, a cluster of Y outliers can be moved arbitrarily far from the bulk of the data but still, by perturbing their $\boldsymbol{x}$ values, have arbitrarily small residuals. See Example 7.16.

Our practical high breakdown procedure is made up of three components.
1) A practical estimator $\hat{\boldsymbol{\beta}}_C$ that is consistent for clean data. Suitable choices would include the full-sample OLS and L_1 estimators.
2) A practical estimator $\hat{\boldsymbol{\beta}}_A$ that is effective for outlier identification. Suitable choices include the `mbareg`, `rmreg2`, `lmsreg`, or FLTS estimators.
3) A practical high-breakdown estimator such as $\hat{\boldsymbol{\beta}}_B$ from Definition 7.34 with $k = 10$.

By selecting one of these three estimators according to the features each of them uncovers in the data, we may inherit some of the good properties of each of them.

Definition 7.35. The `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ is defined as follows. Pick a constant $a > 1$ and set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_C$. If $aQ_L(\hat{\boldsymbol{\beta}}_A) < Q_L(\hat{\boldsymbol{\beta}}_C)$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_A$. If $aQ_L(\hat{\boldsymbol{\beta}}_B) < \min[Q_L(\hat{\boldsymbol{\beta}}_C), aQ_L(\hat{\boldsymbol{\beta}}_A)]$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_B$.

That is, find the smallest of the three scaled criterion values $Q_L(\hat{\boldsymbol{\beta}}_C)$, $aQ_L(\hat{\boldsymbol{\beta}}_A)$, $aQ_L(\hat{\boldsymbol{\beta}}_B)$. According to which of the three estimators attains this minimum, set $\hat{\boldsymbol{\beta}}_H$ to $\hat{\boldsymbol{\beta}}_C, \hat{\boldsymbol{\beta}}_A$, or $\hat{\boldsymbol{\beta}}_B$ respectively.

Large sample theory for `hbreg` is simple and given in the following theorem. Let $\hat{\boldsymbol{\beta}}_L$ be the LMS, LTS, or LTA estimator that minimizes the criterion Q_L. Note that the impractical estimator $\hat{\boldsymbol{\beta}}_L$ is never computed. The following theorem shows that $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$ on a large class of zero mean finite variance symmetric error distributions. Thus if $\hat{\boldsymbol{\beta}}_C$ is $\sqrt{n}$ consistent or asymptotically efficient, so is $\hat{\boldsymbol{\beta}}_H$. Notice that $\hat{\boldsymbol{\beta}}_A$ does not need to be consistent. This point is crucial since `lmsreg` is not consistent and it is not known whether FLTS is consistent. The clean data are in *general position* if any p clean cases give a unique estimate of $\hat{\boldsymbol{\beta}}$.

Theorem 7.24. Assume the clean data are in general position, and suppose that both $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_C$ are consistent estimators of $\boldsymbol{\beta}$ where the regression model contains a constant. Then the `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ is high breakdown and asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$.

Proof. Since the clean data are in general position and $Q_L(\hat{\boldsymbol{\beta}}_H) \leq aQ_L(\hat{\boldsymbol{\beta}}_B)$ is bounded for γ_n near 0.5, the `hbreg` estimator is high breakdown. Let $Q_L^* = Q_L$ for LMS and $Q_L^* = Q_L/n$ for LTS and LTA. As $n \to \infty$, consistent estimators $\hat{\boldsymbol{\beta}}$ satisfy $Q_L^*(\hat{\boldsymbol{\beta}}) - Q_L^*(\boldsymbol{\beta}) \to 0$ in probability. Since LMS, LTS, and LTA are consistent and the minimum value is $Q_L^*(\hat{\boldsymbol{\beta}}_L)$, it follows that $Q_L^*(\hat{\boldsymbol{\beta}}_C) - Q_L^*(\hat{\boldsymbol{\beta}}_L) \to 0$ in probability, while $Q_L^*(\hat{\boldsymbol{\beta}}_L) < aQ_L^*(\hat{\boldsymbol{\beta}})$ for any estimator $\hat{\boldsymbol{\beta}}$. Thus with probability tending to one as $n \to \infty$, $Q_L(\hat{\boldsymbol{\beta}}_C) < a \min(Q_L(\hat{\boldsymbol{\beta}}_A), Q_L(\hat{\boldsymbol{\beta}}_B))$. Hence $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$. □

Remark 7.9. i) Let $\hat{\boldsymbol{\beta}}_C = \hat{\boldsymbol{\beta}}_{OLS}$. Then hbreg is asymptotically equivalent to OLS when the errors e_i are iid from a large class of zero mean finite variance symmetric distributions, including the $N(0, \sigma^2)$ distribution, since the probability that hbreg uses OLS instead of $\hat{\boldsymbol{\beta}}_A$ or $\hat{\boldsymbol{\beta}}_B$ goes to one as $n \to \infty$.

ii) The above theorem proves that practical high breakdown estimators with 100% asymptotic Gaussian efficiency exist; however, such estimators are not necessarily good.

iii) The theorem holds when both $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_C$ are consistent estimators of $\boldsymbol{\beta}$, for example, when the iid errors come from a large class or zero mean finite variance symmetric distributions. For asymmetric distributions, $\hat{\boldsymbol{\beta}}_C$ estimates $\boldsymbol{\beta}_C$ and $\hat{\boldsymbol{\beta}}_L$ estimates $\boldsymbol{\beta}_L$ where the constants usually differ. The theorem holds for some distributions that are not symmetric because of the penalty a. As $a \to \infty$, the class of asymmetric distributions where the theorem holds greatly increases, but the outlier resistance decreases rapidly as a increases for $a > 1.4$.

iv) The default hbreg estimator used OLS, mbareg, and $\hat{\boldsymbol{\beta}}_B$ with $a = 1.4$ and the LTA criterion. For the simulated data with symmetric error distributions, $\hat{\boldsymbol{\beta}}_B$ appeared to give biased estimates of the slopes. However, for the simulated data with right skewed error distributions, $\hat{\boldsymbol{\beta}}_B$ appeared to give good estimates of the slopes but not the constant estimated by OLS, and the probability that the hbreg estimator selected $\hat{\boldsymbol{\beta}}_B$ appeared to go to one.

v) Both MBA and OLS are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$, even for a large class of skewed distributions. Using $\hat{\boldsymbol{\beta}}_A = \hat{\boldsymbol{\beta}}_{MBA}$ and removing $\hat{\boldsymbol{\beta}}_B$ from the hbreg estimator results in a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ when $\hat{\boldsymbol{\beta}}_C =$ OLS is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$, but massive sample sizes were still needed to get good estimates of the constant for skewed error distributions. For skewed distributions, if OLS needed $n = 1000$ to estimate the constant well, mbareg might need $n >$ one million to estimate the constant well.

The situation is worse for multivariate linear regression when hbreg is used instead of OLS, since there are m constants to be estimated. If the distribution of the iid error vectors $\boldsymbol{e}_i$ is not elliptically contoured, getting all m mbareg estimators to estimate all m constants well needs even larger sample sizes.

vi) The outlier resistance of hbreg is not especially good.

The family of hbreg estimators is enormous and depends on i) the practical high breakdown estimator $\hat{\boldsymbol{\beta}}_B$, ii) $\hat{\boldsymbol{\beta}}_C$, iii) $\hat{\boldsymbol{\beta}}_A$, iv) a, and v) the criterion Q_L. Note that the theory needs the error distribution to be such that both $\hat{\boldsymbol{\beta}}_C$ and $\hat{\boldsymbol{\beta}}_L$ are consistent. Sufficient conditions for LMS, LTS, and LTA to be consistent are rather strong. To have reasonable sufficient conditions for the hbreg estimator to be consistent, $\hat{\boldsymbol{\beta}}_C$ should be consistent under weak conditions. Hence OLS is a good choice that results in 100% asymptotic Gaussian efficiency.

We suggest using the LTA criterion since in simulations, hbreg behaved like $\hat{\boldsymbol{\beta}}_C$ for smaller sample sizes than those needed by the LTS and LMS criteria. We want a near 1 so that hbreg has outlier resistance similar to $\hat{\boldsymbol{\beta}}_A$, but we want a large enough so that hbreg performs like $\hat{\boldsymbol{\beta}}_C$ for moderate n on clean data. Simulations suggest that $a = 1.4$ is a reasonable choice. The default hbreg program from *linmodpack* uses the $\sqrt{n}$ consistent outlier resistant estimator mbareg as $\hat{\boldsymbol{\beta}}_A$.

There are at least three reasons for using $\hat{\boldsymbol{\beta}}_B$ as the high breakdown estimator. First, $\hat{\boldsymbol{\beta}}_B$ is high breakdown and simple to compute. Second, the fitted values roughly track the bulk of the data. Lastly, although $\hat{\boldsymbol{\beta}}_B$ has rather poor outlier resistance, $\hat{\boldsymbol{\beta}}_B$ does perform well on several outlier configurations where some common alternatives fail.

Next we will show that the hbreg estimator implemented with $a = 1.4$ using Q_{LTA}, $\hat{\boldsymbol{\beta}}_C =$ OLS, and $\hat{\boldsymbol{\beta}}_B$ can greatly improve the estimator $\hat{\boldsymbol{\beta}}_A$. We will use $\hat{\boldsymbol{\beta}}_A =$ ltsreg in *R* and *Splus 2000*. Depending on the implementation, the ltsreg estimators use the elemental resampling algorithm, the elemental concentration algorithm, or a genetic algorithm. Coverage is 50%, 75%, or 90%. The *Splus 2000* implementation is an unusually poor genetic algorithm with 90% coverage. The *R* implementation appears to be the zero breakdown inconsistent elemental basic resampling algorithm that uses 50% coverage. The *ltsreg* function changes often.

Simulations were run in *R* with the x_{ij} (for $j > 1$) and e_i iid $N(0, \sigma^2)$ and $\boldsymbol{\beta} = \mathbf{1}$, the $p \times 1$ vector of ones. Then $\hat{\boldsymbol{\beta}}$ was recorded for 100 runs. The mean and standard deviation of the $\hat{\beta}_j$ were recorded for $j = 1, ..., p$. For $n \geq 10p$ and OLS, the vector of means should be close to $\mathbf{1}$ and the vector of standard deviations should be close to $\mathbf{1}/\sqrt{n}$. The $\sqrt{n}$ consistent high breakdown hbreg estimator performed like OLS if $n \approx 35p$ and $2 \leq p \leq 6$, if $n \approx 20p$ and $7 \leq p \leq 14$, or if $n \approx 15p$ and $15 \leq p \leq 40$. See Table 7.7 for $p = 5$ and 100 runs. ALTS denotes ltsreg, HB denotes hbreg, and BB denotes $\hat{\boldsymbol{\beta}}_B$. In the simulations, hbreg estimated the slopes well for the highly skewed lognormal data, but not the OLS constant. Use the *linmodpack* function hbregsim.

As implemented in *linmodpack*, the hbreg estimator is a practical $\sqrt{n}$ consistent high breakdown estimator that appears to perform like OLS for moderate n if the errors are unimodal and symmetric, and to have outlier resistance comparable to competing practical "outlier resistant" estimators.

The hbreg, lmsreg, ltsreg, OLS, and $\hat{\boldsymbol{\beta}}_B$ estimators were compared on the same 25 benchmark data sets. Also see Park et al. (2012). The HB estimator $\hat{\boldsymbol{\beta}}_B$ was surprisingly good in that the response plots showed that it was the best estimator for 2 data sets and that it usually tracked the data, but it performed poorly in 7 of the 25 data sets. The hbreg estimator performed well, but for a few data sets hbreg did not pick the attractor with the best response plot, as illustrated in the following example.

Table 7.7 MEAN $\hat{\beta}_i$ and SD($\hat{\beta}_i$)

n	method	mn or sd	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
25	HB	mn	0.9921	0.9825	0.9989	0.9680	1.0231
		sd	0.4821	0.5142	0.5590	0.4537	0.5461
	OLS	mn	1.0113	1.0116	0.9564	0.9867	1.0019
		sd	0.2308	0.2378	0.2126	0.2071	0.2441
	ALTS	mn	1.0028	1.0065	1.0198	1.0092	1.0374
		sd	0.5028	0.5319	0.5467	0.4828	0.5614
	BB	mn	1.0278	0.5314	0.5182	0.5134	0.5752
		sd	0.4960	0.3960	0.3612	0.4250	0.3940
400	HB	mn	1.0023	0.9943	1.0028	1.0103	1.0076
		sd	0.0529	0.0496	0.0514	0.0459	0.0527
	OLS	mn	1.0023	0.9943	1.0028	1.0103	1.0076
		sd	0.0529	0.0496	0.0514	0.0459	0.0527
	ALTS	mn	1.0077	0.9823	1.0068	1.0069	1.0214
		sd	0.1655	0.1542	0.1609	0.1629	0.1679
	BB	mn	1.0184	0.8744	0.8764	0.8679	0.8794
		sd	0.1273	0.1084	0.1215	0.1206	0.1269

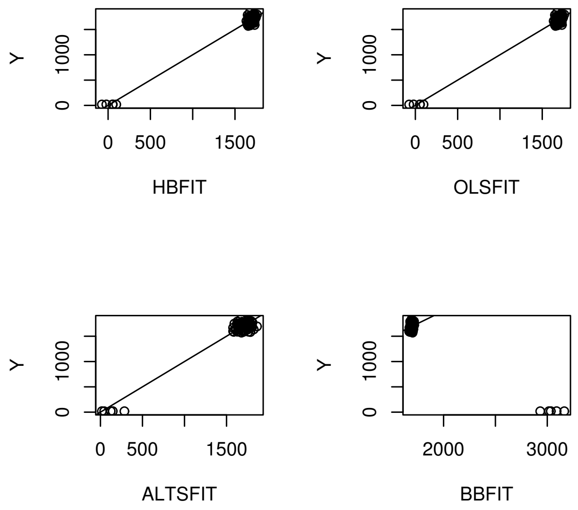

Fig. 7.21 Response Plots Comparing Robust Regression Estimators

Example 7.16. The LMS, LTA, and LTS estimators are determined by a "narrowest band" covering half of the cases. Hawkins and Olive (2002) suggested that the fit will pass through outliers if the band through the outliers is narrower than the band through the clean cases. This behavior tends to occur if the regression relationship is weak, and if there is a tight cluster of outliers where $|Y|$ is not too large. As an illustration, Buxton (1920, pp. 232-5) gave 20 measurements of 88 men. Consider predicting *stature* using an intercept, *head length, nasal height, bigonal breadth*, and *cephalic index*. One case was deleted since it had missing values. Five individuals, numbers 61-65, were reported to be about 0.75 inches tall with head lengths well over five feet! Figure 7.21 shows the response plots for `hbreg`, OLS, `ltsreg`, and $\hat{\boldsymbol{\beta}}_B$. Notice that only the fit from $\hat{\boldsymbol{\beta}}_B$ (BBFIT) did not pass through the outliers, but `hbreg` selected the OLS attractor. There are always outlier configurations where an estimator will fail, and `hbreg` should fail on configurations where LTA, LTS, and LMS would fail.

7.7 Summary

1) For the location model, the sample mean $\overline{Y} = \dfrac{\sum_{i=1}^n Y_i}{n}$, the sample variance $S_n^2 = \dfrac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1}$, and the sample standard deviation $S_n = \sqrt{S_n^2}$. If the data $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the ith order statistic and the $Y_{(i)}$'s are called the *order statistics*. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \ \text{ if n is odd,}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \ \text{ if n is even.}$$

The notation $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ will also be used. The *sample median absolute deviation* is $\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n)$.

2) Suppose the multivariate data has been collected into an $n \times p$ matrix

$$\boldsymbol{W} = \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}.$$

The *coordinatewise median* $\text{MED}(\boldsymbol{W}) = (\text{MED}(X_1), ..., \text{MED}(X_p))^T$ where $\text{MED}(X_i)$ is the sample median of the data in column i corresponding to variable X_i. The **sample mean** $\overline{\boldsymbol{x}} = \dfrac{1}{n}\displaystyle\sum_{i=1}^n \boldsymbol{x}_i = (\overline{X}_1, ..., \overline{X}_p)^T$ where $\overline{X}_i$ is the sample mean of the data in column i corresponding to variable X_i. The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

That is, the ij entry of $\boldsymbol{S}$ is the sample covariance S_{ij}. The *classical estimator of multivariate location and dispersion* is $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$.

3) Let $(T, \boldsymbol{C}) = (T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ be an estimator of multivariate location and dispersion. The ith *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the ith *squared Mahalanobis distance* is $D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) =$ $(\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W}))$.

4) The squared Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median is $D_i^2 = D_i^2(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median MED_j: the coordinatewise median computed from the cases $\boldsymbol{x}_i$ with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \boldsymbol{I}_p))$ where $\text{MED}_0 = \text{MED}(\boldsymbol{W})$. Often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, ..., D_n) + k\text{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise.

5) Let the *covmb2 set* B of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set B. Hence

$$T = \frac{\sum_{i=1}^n W_i \boldsymbol{x}_i}{\sum_{i=1}^n W_i} \quad \text{and} \quad \boldsymbol{C} = \frac{\sum_{i=1}^n W_i (\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^n W_i \; - \; 1}.$$

The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers.

7.8 Complements

Most of this chapter was taken from Olive (2017b). See that text for references to concepts such as breakdown. The fact that response plots are extremely useful for model assessment and for detecting influential cases and outliers for an enormous variety of statistical models does not seem to be well known. Certainly in any multiple linear regression analysis, the response plot and the residual plot of $\hat{Y}$ versus r should always be made. Cook and Olive (2001) used response plots to select a response transformation graphically. Olive (2005) suggested using residual, response, RR, and FF plots to detect outliers while Hawkins and Olive (2002, pp. 141, 158) suggested using the RR and FF plots. The four plots are best for $n \geq 5p$. Olive (2008: $\oint$ 6.4, 2017a: ch. 5-9) showed that the residual and response plots are useful for experimental design models. Park et al. (2012) showed response plots are competitive with the best robust regression methods for outlier detection on some outlier data sets that have appeared in the literature.

Olive (2002) found applications for the DD plot. The TV estimator was proposed by Olive (2002, 2005a). Although both the TV and MBA estimators have the good $O_P(n^{-1/2})$ convergence rate, their efficiency under normality may be very low. Chang and Olive (2010) suggested a method of adaptive trimming such that the resulting estimator is asymptotically equivalent to the OLS estimator.

If n is not much larger than p, then Hoffman et al. (2015) gave a robust Partial Least Squares–Lasso type estimator that uses a clever weighting scheme. See Uraibi et al. (2017, 2019) for robust methods of forward selection and least angle regression.

Robust MLD

For the FCH, RFCH, and RMVN estimators, see Olive and Hawkins (2010), Olive (2017b, ch. 4), and Zhang et al. (2012). See Olive (2017b, p. 120) for the `covmb2` estimator.

The fastest estimators of multivariate location and dispersion that have been shown to be both consistent and high breakdown are the minimum covariance determinant (MCD) estimator with $O(n^v)$ complexity where $v = 1 + p(p+3)/2$ and possibly an all elemental subset estimator of He and Wang (1997). See Bernholt and Fischer (2004). The minimum volume ellipsoid (MVE) complexity is far higher, and **for** $p > 2$ **there may be no known method for computing** S, τ, projection based, and constrained M estimators. For some depth estimators, like the Stahel-Donoho estimator, the exact algorithm of Liu and Zuo (2014) appears to take too long if $p \geq 6$ and $n \geq 100$, and simulations may need $p \leq 3$. It is possible to compute the MCD and MVE estimators for $p = 4$ and $n = 100$ in a few hours using branch and bound algorithms (like estimators with $O(100^4)$ complexity). See Agulló (1996, 1998) and Pesch (1999). These algorithms take too long if both $p \geq 5$ and $n \geq 100$. Simulations may need $p \leq 2$. Two stage estimators such as the MM estimator, that need an initial high breakdown consistent estimator, take longer to compute than the initial estimator. Rousseeuw (1984) introduced the MCD and MVE estimators. See Maronna et al. (2006, ch. 6) for descriptions and references.

Estimators with complexity higher than $O[(n^3+n^2p+np^2+p^3)\log(n)]$ take too long to compute and will rarely be used. Reyen et al. (2009) simulated the OGK and the Olive (2004a) median ball algorithm (MBA) estimators for $p = 100$ and n up to 50000, and noted that the OGK complexity is $O[p^3 + np^2\log(n)]$ while that of MBA is $O[p^3 + np^2 + np\log(n)]$. FCH, RMBA, and RMVN have the same complexity as MBA. FMCD has the same complexity as FCH, but FCH is roughly 100 to 200 times faster.

Robust Regression

For the `hbreg` estimator, see Olive and Hawkins (2011) and Olive (2017b, ch. 14). Robust regression estimators have unsatisfactory outlier resistance and large sample theory. The `hbreg` estimator is fast and high breakdown, but does not provide an adequate remedy for outliers, and the symmetry condition for consistency is too strong. OLS response and residual plots, and

RMVN or RFCH DD plots are useful for detecting multiple linear regression outliers.

Many of the robust statistics for the location model are practical to compute, outlier resistant, and backed by theory. See Huber and Ronchetti (2009). A few estimators of multivariate location and dispersion, such as the coordinatewise median, are practical to compute, outlier resistant, and backed by theory.

For practical estimators for MLR and MCD, `hbreg` and FCH appear to be the only estimators proven to be consistent (for a large class of symmetric error distributions and for a large class of EC distributions, respectively) with some breakdown theory (T_{FCH} is HB). Perhaps all other "robust statistics" for MLR and MLD that have been shown to be both consistent and high breakdown are impractical to compute for $p > 4$: the impractical "brand name" estimators have at least $O(n^p)$ complexity, while the practical estimators used in the software for the "brand name estimators" have not been shown to be both high breakdown and consistent. See Theorems 7.12 and 7.21, Hawkins and Olive (2002), Olive (2008, 2017b), Hubert et al. (2002), and Maronna and Yohai (2002). Huber and Ronchetti (2009, pp. xiii, 8-9, 152-154, 196-197) suggested that high breakdown regression estimators do not provide an adequate remedy for the ill effects of outliers, that their statistical and computational properties are not adequately understood, that high breakdown estimators "break down for all except the smallest regression problems by failing to provide a timely answer!" and that "there are no known high breakdown point estimators of regression that are demonstrably stable."

A large number of impractical high breakdown regression estimators have been proposed, including LTS, LMS, LTA, S, LQD, τ, constrained M, repeated median, cross checking, one step GM, one step GR, t-type, and regression depth estimators. See Rousseeuw and Leroy (1987) and Maronna et al. (2006). The practical algorithms used in the software use a brand name criterion to evaluate a fixed number of trial fits and should be denoted as an F-brand name estimator such as FLTS. Two stage estimators, such as the MM estimator, that need an initial consistent high breakdown estimator often have the same breakdown value and consistency rate as the initial estimator. These estimators are typically implemented with a zero breakdown inconsistent initial estimator and hence are zero breakdown with zero efficiency.

Maronna and Yohai (2015) used OLS and 500 elemental sets as the 501 trial fits to produce an FS estimator used as the initial estimator for an FMM estimator. Since the 501 trial fits are zero breakdown, so is the FS estimator. Since the FMM estimator has the same breakdown as the initial estimator, the FMM estimator is zero breakdown. For regression, they show that the FS estimator is consistent on a large class of zero mean finite variance symmetric distributions. Consistency follows since the elemental fits and OLS are unbiased estimators of $\boldsymbol{\beta}_{OLS}$ but an elemental fit is an OLS fit to p cases.

Hence the elemental fits are very variable, and the probability that the OLS fit has a smaller S-estimator criterion than a randomly chosen elemental fit (or K randomly chosen elemental fits) goes to one as $n \to \infty$. (OLS and the S-estimator are both $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$, so the ratio of their criterion values goes to one, and the S-estimator minimizes the criterion value.) Hence the FMM estimator is asymptotically equivalent to the MM estimator that has the smallest criterion value for a large class of iid zero mean finite variance symmetric error distributions. This FMM estimator is asymptotically equivalent to the FMM estimator that uses OLS as the initial estimator. When the error distribution is skewed the S-estimator and OLS population constant are not the same, and the probability that an elemental fit is selected is close to one for a skewed error distribution as $n \to \infty$. (The OLS estimator $\hat{\boldsymbol{\beta}}$ gets very close to $\boldsymbol{\beta}_{OLS}$ while the elemental fits are highly variable unbiased estimators of $\boldsymbol{\beta}_{OLS}$, so one of the elemental fits is likely to have a constant that is closer to the S-estimator constant while still having good slope estimators.) Hence the FS estimator is inconsistent, and the FMM estimator is likely inconsistent for skewed distributions. No practical method is known for computing a $\sqrt{n}$ consistent FS or FMM estimator that has the same breakdown and maximum bias function as the S or MM estimator that has the smallest S or MM criterion value.

The L_1 CLT is

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{L_1} - \boldsymbol{\beta}) \xrightarrow{D} N_p\left(0, \frac{1}{4[f(0)]^2}\, \boldsymbol{W}\right) \tag{7.37}$$

when $\boldsymbol{X}^T\boldsymbol{X}/n \to \boldsymbol{W}^{-1}$, and when the errors e_i are iid with a cdf F and a pdf f such that the unique population median is 0 with $f(0) > 0$. If a constant β_1 is in the model or if the column space of $\boldsymbol{X}$ contains $\mathbf{1}$, then this assumption is mild, but if the pdf is not symmetric about 0, then the L_1 β_1 tends to differ from the OLS β_1. See Bassett and Koenker (1978). Estimating $f(0)$ can be difficult, so the residual bootstrap using OLS residuals or using $\hat{e}_i = r_i - \overline{r}$ where the r_i are the L_1 residuals with the prediction region method may be useful.

7.9 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

7.1. Referring to Definition 7.25, let $\hat{Y}_{i,j} = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_j = \hat{Y}_i(\hat{\boldsymbol{\beta}}_j)$ and let $r_{i,j} = r_i(\hat{\boldsymbol{\beta}}_j)$. Show that $\|r_{i,1} - r_{i,2}\| = \|\hat{Y}_{i,1} - \hat{Y}_{i,2}\|$.

7.2. Assume that the model has a constant β_1 so that the first column of $\boldsymbol{X}$ is $\boldsymbol{1}$. Show that if the regression estimator is regression equivariant, then adding $\boldsymbol{1}$ to $\boldsymbol{Y}$ changes $\hat{\beta}_1$ but does not change the slopes $\hat{\beta}_2, ..., \hat{\beta}_p$.

R Problems

Use the command *source("G:/linmodpack.txt")* **to download the functions** and the command *source("G:/linmoddata.txt")* **to download the data. See Preface or Section 11.1.** Typing the name of the `linmodpack` function, e.g. *trviews*, will display the code for the function. Use the `args` command, e.g. *args(trviews)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://parker.ad.siu.edu/Olive/mrsashw.txt) into *R*.

7.3. Paste the command for this problem into *R* to produce the second column of Table 7.5. Include the output in *Word*.

7.4. a) To get an idea for the amount of contamination a basic resampling or concentration algorithm for MLR can tolerate, enter or download the `gamper` function (with the *source("G:/linmodpack.txt")* command) that evaluates Equation (7.24) at different values of $h = p$.

b) Next enter the following commands and include the output in *Word*.

```
zh <- c(10,20,30,40,50,60,70,80,90,100)
for(i in 1:10) gamper(zh[i])
```

7.5*. a) Assuming that you have done the two source commands above Problem 7.3 (and the *R* command *library(MASS)*), type the command *ddcomp(buxx)*. This will make 4 DD plots based on the DGK, FCH, FMCD, and median ball estimators. The DGK and median ball estimators are the two attractors used by the FCH estimator. With the leftmost mouse button, move the cursor to an outlier and click. This data is the Buxton (1920) data and cases with numbers 61, 62, 63, 64, and 65 were the outliers with head lengths near 5 feet. After identifying at least three outliers in each plot, hold the rightmost mouse button down (and in *R* click on *Stop*) to advance to the next plot. When done, hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

b) Repeat a) but use the command *ddcomp(cbrainx)*. This data is the Gladstone (1905) data and some infants are multivariate outliers.

c) Repeat a) but use the command *ddcomp(museum[,-1])*. This data is the Schaaffhausen (1878) skull measurements and cases 48–60 were apes while the first 47 cases were humans.

7.6*. (Perform the *source("G:/linmodpack.txt")* command if you have not already done so.) The *concmv* function illustrates concentration with $p = 2$ and a scatterplot of X_1 versus X_2. The outliers are such that the robust estimators can not always detect them. Type the command *concmv()*. Hold the rightmost mouse button down (and in *R* click on *Stop*) to see the DD plot after one concentration step. The start uses the coordinatewise median and $diag([MAD(X_i)]^2)$. Repeat 4 more times to see the DD plot based on the attractor. The outliers have large values of X_2 and the highlighted cases have the smallest distances. Repeat the command *concmv()* several times. Sometimes the start will contain outliers but the attractor will be clean (none of the highlighted cases will be outliers), but sometimes concentration causes more and more of the highlighted cases to be outliers, so that the attractor is worse than the start. Copy one of the DD plots where none of the outliers are highlighted into *Word*.

7.7***.** (Perform the *source("G:/linmodpack.txt")* command if you have not already done so.) The *ddmv* function illustrates concentration with the DD plot. The outliers are highlighted. The first graph is the DD plot after one concentration step. Hold the rightmost mouse button down (and in *R* click on *Stop*) to see the DD plot after two concentration steps. Repeat 4 more times to see the DD plot based on the attractor. In this problem, try to determine the proportion of outliers *gam* that the DGK estimator can detect for $p = 2, 4, 10,$ and 20. Make a table of *p* and *gam*. For example the command *ddmv(p=2,gam=.4)* suggests that the DGK estimator can tolerate nearly 40% outliers with $p = 2$, but the command *ddmv(p=4,gam=.4)* suggest that *gam* needs to be lowered (perhaps by 0.1 or 0.05). Try to make $0 < gam < 0.5$ as large as possible.

7.8***.** a) If necessary, use the commands *source("G:/linmodpack.txt")* and *source("G:/linmoddata.txt")*.

b) Enter the command *mbamv(belx,bely)* in *R*. Click on the rightmost mouse button (and in *R*, click on *Stop*). You need to do this 7 times before the program ends. There is one predictor x and one response Y. The function makes a scatterplot of x and Y and cases that get weight one are shown as highlighted squares. Each MBA sphere covers half of the data. When you find a good fit to the bulk of the data, hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

c) Enter the command *mbamv2(buxx,buxy)* in *R*. Click on the rightmost mouse button (and in *R*, click on *Stop*). You need to do this 14 times before the program ends. There are four predictors $x_1, ..., x_4$ and one response Y. The function makes the response and residual plots based on the OLS fit to the highlighted cases. Each MBA sphere covers half of the data. When you find a good fit to the bulk of the data, hold down the *Ctrl* and *c* keys to make a copy of the two plots. Then paste the plots in *Word*.

7.9. This problem compares the MBA estimator that uses the median squared residual MED(r_i^2) criterion with the MBA estimator that uses the LATA criterion. On clean data, both estimators are $\sqrt{n}$ consistent since both use 50 $\sqrt{n}$ consistent OLS estimators. The MED(r_i^2) criterion has trouble with data sets where the multiple linear regression relationship is weak and there is a cluster of outliers. The LATA criterion tries to give all x–outliers, including good leverage points, zero weight.

a) If necessary, use the commands *source("G:/linmodpack.txt")* and *source("G:/linmoddata.txt")*. The `mlrplot2` function is used to compute both MBA estimators. Use the rightmost mouse button to advance the plot (and in *R*, highlight stop).

b) Use the command *mlrplot2(belx,bely)* and include the resulting plot in *Word*. Is one estimator better than the other, or are they about the same?

c) Use the command *mlrplot2(cbrainx,cbrainy)* and include the resulting plot in *Word*. Is one estimator better than the other, or are they about the same? (The infants are likely good leverage cases instead of outliers.)

d) Use the command *mlrplot2(museum[,3:11],museum[,2])* and include the resulting plot in *Word*. For this data set, most of the cases are based on humans but a few are based on apes. The MBA LATA estimator will often give the cases corresponding to apes larger absolute residuals than the MBA estimator based on MED(r_i^2), but the apes appear to be good leverage cases.

e) Use the command *mlrplot2(buxx,buxy)* until the outliers are clustered about the identity line in one of the two response plots. (This will usually happen within 10 or fewer runs. Pressing the "up arrow" will bring the previous command to the screen and save typing.) Then include the resulting plot in *Word*. Which estimator went through the outliers and which one gave zero weight to the outliers?

f) Use the command *mlrplot2(hx,hy)* several times. Usually both MBA estimators fail to find the outliers for this artificial Hawkins data set that is also analyzed by Atkinson and Riani (2000, section 3.1). The *lmsreg* estimator can be used to find the outliers. In *R* use the commands *library(MASS)* and *ffplot2(hx,hy)*. Include the resulting plot in *Word*.

7.10. a) After entering the two *source* commands above Problem 7.3, enter the following command.

```
MLRplot(buxx,buxy)
```

Click the rightmost mouse button (and in *R* click on *Stop*). The response plot should appear. Again, click the rightmost mouse button (and in *R* click on *Stop*). The residual plot should appear. Hold down the *Ctrl* and *c* keys to make a copy of the two plots. Then paste the plots in *Word*.

b) The response variable is *height*, but 5 cases were recorded with heights about 0.75 inches tall. The highlighted squares in the two plots correspond to cases with large Cook's distances. With respect to the Cook's distances, what is happening, swamping or masking?

7.11. For the Buxton (1920) data with multiple linear regression, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet!

a) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. The identity line passes right through the outliers which are obvious because of the large gap. Prediction interval (PI) bands are also included in the plot.

b) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. This did lasso for the cases in the `covmb2` set B applied to the predictors which included all of the clean cases and omitted the 5 outliers. The response plot was made for all of the data, including the outliers.

c) Copy and paste the commands for this problem into *R*. Include the DD plot in *Word*. The outliers are in the upper right corner of the plot.

7.12. Consider the Gladstone (1905) data set that has 12 variables on 267 persons after death. There are 5 infants in the data set. The response variable was *brain weight.* Head measurements were *breadth, circumference, head height, length,* and *size* as well as *cephalic index* and *brain weight. Age*, *height*, and three categorical variables *cause*, *ageclass* (0: under 20, 1: 20-45, 2: over 45) and *sex* were also given. The constant x_1 was the first variable. The variables *cause* and *ageclass* were not coded as factors. Coding as factors might improve the fit.

a) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. The identity line passes right through the infants which are obvious because of the large gap. Prediction interval (PI) bands are also included in the plot.

b) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. This did lasso for the cases in the `covmb2` set B applied to the nontrivial predictors which are not categorical (omit the *constant, cause, ageclass* and *sex*) which omitted 8 cases, including the 5 infants. The response plot was made for all of the data.

c) Copy and paste the commands for this problem into *R*. Include the DD plot in *Word*. The infants are in the upper right corner of the plot.

7.13. The *linmodpack* function `mldsim6` compares 7 estimators: FCH, RFCH, CMVE, RCMVE, RMVN, `covmb2`, and MB described in Olive (2017b, ch. 4). Most of these estimators need $n > 2p$, need a nonsingular dispersion matrix, and work best with $n > 10p$. The function generates data sets and counts how many times the minimum Mahalanobis distance $D_i(T, \boldsymbol{C})$ of the outliers is larger than the maximum distance of the clean data. The value *pm* controls how far the outliers need to be from the bulk of the data, and *pm* roughly needs to increase with $\sqrt{p}$.

For data sets with $p > n$ possible, the function `mldsim7` used the Euclidean distances $D_i(T, \boldsymbol{I}_p)$ and the Mahalanobis distances $D_i(T, \boldsymbol{C}_d)$ where $\boldsymbol{C}_d$ is the diagonal matrix with the same diagonal entries as $\boldsymbol{C}$ where $(T, \boldsymbol{C})$ is the `covmb2` estimator using j concentration type steps. Dispersion matrices are effected more by outliers than good robust location estimators, so when the outlier proportion is high, it is expected that the Euclidean distances $D_i(T, \boldsymbol{I}_p)$ will outperform the Mahalanobis distance $D_i(T, \boldsymbol{C}_d)$ for many outlier configurations. Again the function counts the number of times the minimum outlier distance is larger than the maximum distance of the clean data.

Both functions used several outlier types. The simulations generated 100 data sets. The clean data had $\boldsymbol{x}_i \sim N_p(\boldsymbol{0}, diag(1, ..., p))$. Type 1 had outliers in a tight cluster (near point mass) at the major axis $(0, ..., 0, pm)^T$. Type 2 had outliers in a tight cluster at the minor axis $(pm, 0, ..., 0)^T$. Type 3 had mean shift outliers $\boldsymbol{x}_i \sim N_p((pm, ..., pm)^T, diag(1, ..., p))$. Type 4 changed the pth coordinate of the outliers to pm. Type 5 changed the 1st coordinate of the outliers to pm. (If the outlier $\boldsymbol{x}_i = (x_{1i}, ..., x_{pi})^T$, then $x_{i1} = pm$.)

Table 7.8 Number of Times All Outlier Distances > Clean Distances, otype=1

n	p	γ	osteps	pm	FCH	RFCH	CMVE	RCMVE	RMVN	covmb2	MB
100	10	0.25	0	20	85	85	85	85	86	67	89

a) Table 7.8 suggests with osteps = 0, `covmb2` had the worst count. When pm is increased to 25, all counts become 100. Copy and paste the commands for this part into R and make a table similar to Table 7.8, but now osteps=9 and $p = 45$ is close to $n/2$ for the second line where $pm = 60$. Your table should have 2 lines from output.

Table 7.9 Number of Times All Outlier Distances > Clean Distances, otype=1

n	p	γ	osteps	pm	covmb2	diag
100	1000	0.4	0	1000	100	41
100	1000	0.4	9	600	100	42

b) Copy and paste the commands for this part into R and make a table similar to Table 7.9, but type 2 outliers are used.

c) When you have two reasonable outlier detectors, there are outlier configurations where one will beat the other. Simulations suggest that "covmb2" using $D_i(T, \boldsymbol{I}_p)$ outperforms "diag" using $D_i(T, \boldsymbol{C}_d)$ for many outlier configurations, but there are some exceptions. Copy and paste the commands for this part into R and make a table similar to Table 7.9, but type 3 outliers are used.

7.14. a) In addition to the *source("G:/linmodpack.txt")* command, also use the *source("G:/linmoddata.txt")* command, and type the *library(MASS)* command).

b) Type the command *tvreg(buxx,buxy,ii=1)*. Click the rightmost mouse button and highlight *Stop*. The response plot should appear. Repeat 10 times and remember which plot percentage M (say M = 0) had the best response plot. Then type the command *tvreg2(buxx,buxy, M = 0)* (except use your value of M, not 0). Again, click the rightmost mouse button (and in *R*, highlight *Stop*). The response plot should appear. Hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

c) The estimated coefficients $\hat{\boldsymbol{\beta}}_{TV}$ from the best plot should have appeared on the screen. Copy and paste these coefficients into *Word*.

7.15. This problem is like Problem 7.11, except elastic net is used instead of lasso.

a) Copy and paste the commands for this problem into *R*. Include the elastic net response plot in *Word*. The identity line passes right through the outliers which are obvious because of the large gap. Prediction interval (PI) bands are also included in the plot.

b) Copy and paste the commands for this problem into *R*. Include the elastic net response plot in *Word*. This did elastic net for the cases in the `covmb2` set B applied to the predictors which included all of the clean cases and omitted the 5 outliers. The response plot was made for all of the data, including the outliers. (Problem 7.11 c) shows the DD plot for the data.)