# Chapter 8
# Robust Statistics

This chapter considers large sample theory for robust statistics. Robust estimators of multivariate location and dispersion are useful for outlier detection and for developing robust regression estimators. This chapter follows Olive (2008, 2017b, 2022c) closely.

**Definition 8.1** An **outlier** corresponds to a case that is far from the bulk of the data.

## 8.1 The Location Model

The location model is

$$Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{8.1}$$

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 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 $\mu$ is often the population mean or median while the scale parameter is often the population standard deviation $\sqrt{V(Y)}$. The $i$th *case* is $Y_i$.

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* $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$. The *sample variance* $S_n^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^{n} Y_i^2 - n(\overline{Y})^2}{n-1}$, 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 $i$th 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 $\text{MED}(n) = 3$ 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 8.2.** The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \text{ if n is odd,} \tag{8.2}$$

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

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

**Definition 8.3.** The *sample median absolute deviation* is

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n). \tag{8.3}$$

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

**Example 8.1.** Let the data be $1, 2, 3, 4, 5, 6, 7, 8, 9$. Then $\text{MED}(n) = 5$ and $\text{MAD}(n) = 2 = \text{MED}\{0, 1, 1, 2, 2, 3, 3, 4, 4\}$.

The population median $\text{MED}(Y)$ and the population median absolute deviation $\text{MAD}(Y)$ are important quantities of a distribution.

**Definition 8.4.** The *population median* is any value $\text{MED}(Y)$ such that

$$P(Y \leq \text{MED}(Y)) \geq 0.5 \text{ and } P(Y \geq \text{MED}(Y)) \geq 0.5. \tag{8.4}$$

**Definition 8.5.** The *population median absolute deviation* is

$$\text{MAD}(Y) = \text{MED}(|Y - \text{MED}(Y)|). \tag{8.5}$$

$\text{MED}(Y)$ is a measure of location while $\text{MAD}(Y)$ is a measure of scale. The median is the middle value of the distribution. Since $\text{MAD}(Y)$ is the me-

**Table 8.1** Some commonly used notation.

| population | sample |
|---|---|
| $E(Y)$, $\mu$, $\theta$ | $\overline{Y}_n$, $E(n)$ $\hat{\mu}$, $\hat{\theta}$ |
| MED$(Y)$, $M$ | MED$(n)$, $\hat{M}$ |
| VAR$(Y)$, $\sigma^2$ | VAR$(n)$, $S^2$, $\hat{\sigma}^2$ |
| SD$(Y)$, $\sigma$ | SD$(n)$, $S$, $\hat{\sigma}$ |
| MAD$(Y)$ | MAD$(n)$ |
| IQR$(Y)$ | IQR$(n)$ |

dian distance from MED$(Y)$, at least half of the mass is inside [MED$(Y)$ − MAD$(Y)$, MED$(Y)$ + MAD$(Y)$] and at least half of the mass of the distribution is outside of the interval (MED$(Y)$ − MAD$(Y)$, MED$(Y)$ + MAD$(Y)$). In other words, MAD$(Y)$ is any value such that

$$P(Y \in [\text{MED}(Y) - \text{MAD}(Y), \text{MED}(Y) + \text{MAD}(Y)]) \geq 0.5,$$

$$\text{and } P(Y \in (\text{MED}(Y) - \text{MAD}(Y), \text{MED}(Y) + \text{MAD}(Y)) \ ) \leq 0.5.$$

**Definition 8.6.** The *sample interquantile range* IQR$(n) = Y_{(\lceil 0.75n \rceil)} - Y_{(\lceil 0.25n \rceil)}$. The *population interquantile range* IQR$(Y) = y_{0.75} - y_{0.25}$ where $P(Y \leq y_\alpha) = \alpha$ if $y_\alpha$ is a continuity point of the cdf $F_Y(y)$.

Notation is needed in order to distinguish between population quantities, random quantities, and observed quantities. For population quantities, capital letters like $E(Y)$ and MAD$(Y)$ will often be used while the estimators will often be denoted by MED$(n)$, MAD$(n)$, MED$(Y_i, i = 1, ..., n)$, or MED$(Y_1, \ldots, Y_n)$. The random sample will be denoted by $Y_1, \ldots, Y_n$. Sometimes the observed sample will be fixed and lower case letters will be used. For example, the observed sample may be denoted by $y_1, ..., y_n$ while the estimates may be denoted by med$(n)$, mad$(n)$, or $\overline{y}_n$. Table 8.1 summarizes some of this notation.

**Definition 8.7.** Let $f_Y(y)$ be the pdf of Y. Then the family of pdfs $f_W(w) = f_Y(w - \mu)$ indexed by the *location parameter* $\mu$, $-\infty < \mu < \infty$, is the *location family* for the random variable $W = \mu + Y$ with *standard pdf* $f_Y(y)$.

**Definition 8.8.** Let $f_Y(y)$ be the pdf of Y. Then the family of pdfs $f_W(w) = (1/\sigma) f_Y(w/\sigma)$ indexed by the *scale parameter* $\sigma > 0$, is the *scale family* for the random variable $W = \sigma Y$ with *standard pdf* $f_Y(y)$.

**Definition 8.9.** Let $f_Y(y)$ be the pdf of Y. Then the family of pdfs $f_W(w) = (1/\sigma) f_Y((w - \mu)/\sigma)$ indexed by the *location and scale parameters* $\mu$, $-\infty < \mu < \infty$, and $\sigma > 0$, is the *location–scale family* for the random variable $W = \mu + \sigma Y$ with *standard pdf* $f_Y(y)$.

Finding MED($Y$) and MAD($Y$) for symmetric distributions and location–scale families is made easier by the following theorem. Let $F(y_\alpha) = P(Y \leq y_\alpha) = \alpha$ for $0 < \alpha < 1$ where the cdf $F(y) = P(Y \leq y)$. Let $D = \text{MAD}(Y)$, $M = \text{MED}(Y) = y_{0.5}$ and $U = y_{0.75}$.

**Theorem 8.1.** a) If $W = a + bY$, then $\text{MED}(W) = a + b\text{MED}(Y)$ and $\text{MAD}(W) = |b|\text{MAD}(Y)$.

b) If $Y$ has a pdf that is continuous and positive on its support and symmetric about $\mu$, then $\text{MED}(Y) = \mu$ and $\text{MAD}(Y) = y_{0.75} - \text{MED}(Y)$. Find $M = \text{MED}(Y)$ by solving the equation $F(M) = 0.5$ for $M$, and find $U$ by solving $F(U) = 0.75$ for $U$. Then $D = \text{MAD}(Y) = U - M$.

c) Suppose that $W$ is from a location–scale family with standard pdf $f_Y(y)$ that is continuous and positive on its support. Then $W = \mu + \sigma Y$ where $\sigma > 0$. First find $M$ by solving $F_Y(M) = 0.5$. After finding $M$, find $D$ by solving $F_Y(M + D) - F_Y(M - D) = 0.5$. Then $\text{MED}(W) = \mu + \sigma M$ and $\text{MAD}(W) = \sigma D$.

**Proof sketch.** a) Assume the probability density function of $Y$ is continuous and positive on its support. Assume $b > 0$. Then

$$1/2 = P[Y \leq \text{MED}(Y)] = P[a + bY \leq a + b\text{MED}(Y)] = P[W \leq \text{MED}(W)].$$

$$\begin{aligned} 1/2 &= P[\text{MED}(Y) - \text{MAD}(Y) \leq Y \leq \text{MED}(Y) + \text{MAD}(Y)] \\ &= P[a + b\text{MED}(Y) - b\text{MAD}(Y) \leq a + bY \leq a + b\text{MED}(Y) + b\text{MAD}(Y)] \\ &= P[\text{MED}(W) - b\text{MAD}(Y) \leq W \leq \text{MED}(W) + b\text{MAD}(Y)] \\ &= P[\text{MED}(W) - \text{MAD}(W) \leq W \leq \text{MED}(W) + \text{MAD}(W)]. \end{aligned}$$

The proofs of b) and c) are similar. □

**Application 8.1.** *The MAD Method:* In analogy with the method of moments, *robust point estimators* can be obtained by solving $\text{MED}(n) = \text{MED}(Y)$ and $\text{MAD}(n) = \text{MAD}(Y)$. In particular, the location and scale parameters of a location–scale family can often be estimated robustly using $c_1\text{MED}(n)$ and $c_2\text{MAD}(n)$ where $c_1$ and $c_2$ are appropriate constants.

Estimators that use order statistics are common. The shorth estimator of Section 4.1 was used for prediction and confidence intervals.

**Definition 8.10.** Consider intervals that contain $c_n$ cases: $[Y_{(1)}, Y_{(c_n)}]$, $[Y_{(2)}, Y_{(c_n+1)}], ..., [Y_{(n-c_n+1)}, Y_{(n)}]$. Denote the set of $c_n$ cases in the $i$th interval by $J_i$, for $i = 1, 2, ..., n - c_n + 1$. Often $c_n = \lfloor n/2 \rfloor + 1$.

i) Let the shorth($c_n$) estimator $= [Y_{(s)}, Y_{(s+c_n-1)}]$ be the shortest such interval. Then the *least median of squares estimator* $\text{LMS}(c_n)$ is $(Y_{(s)} + Y_{(s+c_n-1)})/2$, the midpoint of the shorth($c_n$) interval. The LMS estimator is also called the *least quantile of squares estimator* $\text{LQS}(c_n)$.

ii) Compute the sample mean and sample variance $(\overline{Y}_{J_i}, S^2_{J_i})$ of the $c_n$ cases in the $i$th interval. The *minimum covariance determinant* estimator MCD($c_n$) estimator $(\overline{Y}_{MCD}, S^2_{MCD})$ is equal to the $(\overline{Y}_{J_j}, S^2_{J_j})$ with the smallest $S^2_{J_i}$. The *least trimmed sum of squares estimator* is LTS($c_n$) $= \overline{Y}_{MCD}$.

iii) Compute the sample median $M_{J_i}$ of the $c_n$ cases in the $i$th interval. Let $Q_{LTA}(M_{J_i}) = \sum_{j \in J_i} |y_j - M_{J_i}|$. The *least trimmed sum of absolute deviations estimator* LTA($c_n$) is equal to the $M_{J_j}$ with the smallest $Q_{LTA}(M_{J_i})$.

## *8.1.1* Robust Confidence Intervals

In this subsection, large sample confidence intervals (CIs) for the sample median and 25% trimmed mean are given. Theory is given later in Section 8.1. The following confidence interval provides some resistance to gross outliers while being very simple to compute. The standard error SE(MED($n$)) is due to Bloch and Gastwirth (1968), but the degrees of freedom $p \approx \lceil \sqrt{n} \rceil$) is motivated by the confidence interval for the trimmed mean. Let $\lfloor x \rfloor$ denote the "greatest integer function" (e.g., $\lfloor 7.7 \rfloor = 7$). Let $\lceil x \rceil$ denote the smallest integer greater than or equal to $x$ (e.g., $\lceil 7.7 \rceil = 8$).

**Warning:** Closed intervals should be used instead of open intervals: $a \pm b = [a - b, a + b]$.

**Application 8.2: inference with the sample median.** Let $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4} \rceil$ and use

$$SE(\text{MED}(n)) = 0.5(Y_{(U_n)} - Y_{(L_n+1)}).$$

Let $p = U_n - L_n - 1$. Then a $100(1-\alpha)\%$ confidence interval for the population median is

$$\text{MED}(n) \pm t_{p,1-\alpha/2} SE(\text{MED}(n)). \tag{8.6}$$

**Warning.** This CI is easy to compute by hand, but tends to be long with undercoverage if $n < 100$. See Baszczyńska and Pekasiewicz (2010) for two competitors that work better. We recommend bootstrap confidence intervals for the population median.

The trimmed mean is also useful, and we recommend the 25% trimmed mean. Let $\lfloor x \rfloor$ denote the "greatest integer function" (e.g., $\lfloor 7.7 \rfloor = 7$).

**Definition 8.11.** The symmetrically trimmed mean or the $\delta$ *trimmed mean*

$$T_n = T_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \tag{8.7}$$

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

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

The trimmed mean is estimating a truncated mean $\mu_T$. Assume that $Y$ has a probability density function $f_Y(y)$ that is continuous and positive on its support. Let $y_\delta$ be the number satisfying $P(Y \leq y_\delta) = \delta$. Then

$$\mu_T = \frac{1}{1-2\delta} \int_{y_\delta}^{y_{1-\delta}} y f_Y(y) dy. \tag{8.8}$$

Notice that the 25% trimmed mean is estimating

$$\mu_T = \int_{y_{0.25}}^{y_{0.75}} 2y f_Y(y) dy.$$

To perform inference, find $d_1, ..., d_n$ where

$$d_i = \begin{cases} Y_{(L_n+1)}, & i \leq L_n \\ Y_{(i)}, & L_n + 1 \leq i \leq U_n \\ Y_{(U_n)}, & i \geq U_n + 1. \end{cases}$$

Then the Winsorized variance is the sample variance $S_n^2(d_1, ..., d_n)$ of $d_1, ..., d_n$, and the scaled Winsorized variance

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2}. \tag{8.9}$$

The standard error (SE) of $T_n$ is $SE(T_n) = \sqrt{V_{SW}(L_n, U_n)/n}$.

**Application 8.3: inference with the $\delta$ trimmed mean.** A large sample 100 $(1-\alpha)$% confidence interval (CI) for $\mu_T$ is

$$T_n \pm t_{p,1-\frac{\alpha}{2}} SE(T_n) \tag{8.10}$$

where $P(t_p \leq t_{p,1-\frac{\alpha}{2}}) = 1 - \alpha/2$ if $t_p$ is from a $t$ distribution with $p = U_n - L_n - 1$ degrees of freedom. This interval is the classical t–interval when $\delta = 0$, but $\delta = 0.25$ gives a robust CI.

**Example 8.2.** Let the data be 6, 9, 9, 7, 8, 9, 9, 7. Assume the data came from a symmetric distribution with mean $\mu$, and find a 95% CI for $\mu$.

**Solution.** When computing small examples by hand, the steps are to sort the data from smallest to largest value, find $n$, $L_n$, $U_n$, $Y_{(L_n+1)}$, $Y_{(U_n)}$, $p$, MED$(n)$ and $SE($MED$(n))$. After finding $t_{p,1-\alpha/2}$, plug the relevant quantities into the formula for the CI. The sorted data are 6, 7, 7, 8, 9, 9, 9, 9. Thus MED$(n) = (8+9)/2 = 8.5$. Since $n = 8$, $L_n = \lfloor 4 \rfloor - \lceil \sqrt{2} \rceil = 4 - \lceil 1.414 \rceil = 4 - 2 = 2$ and $U_n = n - L_n = 8 - 2 = 6$. Hence $SE($MED$(n)) = 0.5(Y_{(6)} - Y_{(3)}) = 0.5 * (9 - 7) = 1$. The degrees of free-

dom $p = U_n - L_n - 1 = 6 - 2 - 1 = 3$. The cutoff $t_{3,0.975} = 3.182$. Thus the 95% CI for MED$(Y)$ is

$$\text{MED}(n) \pm t_{3,0.975} SE(\text{MED}(n))$$

$= 8.5 \pm 3.182(1) = [5.318, 11.682]$. The classical t–interval uses $\overline{Y} = (6 + 7 + 7 + 8 + 9 + 9 + 9 + 9)/8$ and $S_n^2 = (1/7)[(\sum_{i=1}^{n} Y_i^2) - 8(8^2)] = (1/7)[(522 - 8(64)] = 10/7 \approx 1.4286$, and $t_{7,0.975} \approx 2.365$. Hence the 95% CI for $\mu$ is $8 \pm 2.365(\sqrt{1.4286/8}) = [7.001, 8.999]$. Notice that the $t$-cutoff $= 2.365$ for the classical interval is less than the $t$-cutoff $= 3.182$ for the median interval and that $SE(\overline{Y}) < SE(\text{MED}(n))$. The parameter $\mu$ is between 1 and 9 since the test scores are integers between 1 and 9. Hence for this example, the t–interval is considerably superior to the overly long median interval.

**Example 8.3.** In the last example, what happens if the 6 becomes 66 and a 9 becomes 99?

**Solution.** Then the ordered data are 7, 7, 8, 9, 9, 9, 66, 99. Hence MED$(n) = 9$. Since $L_n$ and $U_n$ only depend on the sample size, they take the same values as in the previous example and $SE(\text{MED}(n)) = 0.5(Y_{(6)} - Y_{(3)}) = 0.5 * (9 - 8) = 0.5$. Hence the 95% CI for MED$(Y)$ is MED$(n) \pm t_{3,0.975} SE(\text{MED}(n)) = 9 \pm 3.182(0.5) = [7.409, 10.591]$. Notice that with discrete data, it is possible to drive $SE(\text{MED}(n))$ to 0 with a few outliers if $n$ is small. The classical confidence interval $\overline{Y} \pm t_{7,0.975} S/\sqrt{n}$ blows up and is equal to $[-2.955, 56.455]$.

### *8.1.2* Some Two Stage Trimmed Means

Robust estimators are often obtained by applying the sample mean to a sequence of consecutive order statistics. The sample median, trimmed mean, metrically trimmed mean, and two stage trimmed means are examples. For the trimmed mean given in Definition 8.11 and for the Winsorized mean, defined below, the proportion of cases trimmed and the proportion of cases covered are fixed.

**Definition 8.12.** Using the same notation as in Definition 8.11, the *Winsorized mean*

$$W_n = W_n(L_n, U_n) = \frac{1}{n}[L_n Y_{(L_n+1)} + \sum_{i=L_n+1}^{U_n} Y_{(i)} + (n - U_n) Y_{(U_n)}]. \quad (8.11)$$

**Definition 8.13.** A *randomly trimmed mean*

$$R_n = R_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \tag{8.12}$$

where $L_n < U_n$ are integer valued random variables. $U_n - L_n$ of the cases are *covered* by the randomly trimmed mean while $n - U_n + L_n$ of the cases are trimmed.

**Definition 8.14.** The *metrically trimmed mean* (also called the Huber type skipped mean) $M_n$ is the sample mean of the cases inside the interval

$$[\hat{\theta}_n - k_1 D_n,\ \hat{\theta}_n + k_2 D_n]$$

where $\hat{\theta}_n$ is a location estimator, $D_n$ is a scale estimator, $k_1 \geq 1$, and $k_2 \geq 1$.

The proportions of cases covered and trimmed by randomly trimmed means such as the metrically trimmed mean are now random. Typically MED$(n)$ and MAD$(n)$ are used for $\hat{\theta}_n$ and $D_n$, respectively. The amount of trimming will depend on the distribution of the data. For example, if $M_n$ uses $k_1 = k_2 = 5.2$ and the data is normal (Gaussian), about 1% of the data will be trimmed while if the data is Cauchy, about 12% of the data will be trimmed. Hence the upper and lower trimming points estimate lower and upper population percentiles $L(F)$ and $U(F)$ and change with the distribution $F$.

Two stage estimators are frequently used in robust statistics. Often the initial estimator used in the first stage has good resistance properties but has a low asymptotic relative efficiency or no convenient formula for the SE. Ideally, the estimator in the second stage will have resistance similar to the initial estimator but will be efficient and easy to use. The metrically trimmed mean $M_n$ with tuning parameter $k_1 = k_2 \equiv k = 6$ will often be the initial estimator for the two stage trimmed means. That is, retain the cases that fall in the interval

$$[\text{MED}(n) - 6\text{MAD}(n), \text{MED}(n) + 6\text{MAD}(n)].$$

Let $L(M_n)$ be the number of observations that fall to the left of MED$(n) - k_1$ MAD$(n)$ and let $n - U(M_n)$ be the number of observations that fall to the right of MED$(n) + k_2$ MAD$(n)$. When $k_1 = k_2 \equiv k \geq 1$, at least half of the cases will be covered. Consider the set of 51 trimming proportions in the set $C = \{0, 0.01, 0.02, ..., 0.49, 0.50\}$. Alternatively, the coarser set of 6 trimming proportions $C = \{0, 0.01, 0.1, 0.25, 0.40, 0.49\}$ may be of interest. The greatest integer function (e.g. $\lfloor 7.7 \rfloor = 7$) is used in the following definitions.

**Definition 8.15.** Consider the smallest proportion $\alpha_{o,n} \in C$ such that $\alpha_{o,n} \geq L(M_n)/n$ and the smallest proportion $1 - \beta_{o,n} \in C$ such that $1 - \beta_{o,n} \geq 1 - (U(M_n)/n)$. Let $\alpha_{M,n} = \max(\alpha_{o,n}, 1 - \beta_{o,n})$. Then the *two stage*

*symmetrically trimmed mean* $T_{S,n}$ is the $\alpha_{M,n}$ trimmed mean. Hence $T_{S,n}$ is a randomly trimmed mean with $L_n = \lfloor n \ \ \alpha_{M,n} \rfloor$ and $U_n = n - L_n$. If $\alpha_{M,n} = 0.50$, then use $T_{S,n} = \text{MED}(n)$.

**Definition 8.16.** As in the previous definition, consider the smallest proportion $\alpha_{o,n} \in C$ such that $\alpha_{o,n} \geq L(M_n)/n$ and the smallest proportion $1 - \beta_{o,n} \in C$ such that $1 - \beta_{o,n} \geq 1 - (U(M_n)/n)$. Then the *two stage asymmetrically trimmed mean* $T_{A,n}$ is the $(\alpha_{o,n}, 1 - \beta_{o,n})$ trimmed mean. Hence $T_{A,n}$ is a randomly trimmed mean with $L_n = \lfloor n \ \ \alpha_{o,n} \rfloor$ and $U_n = \lfloor n \ \ \beta_{o,n} \rfloor$. If $\alpha_{o,n} = 1 - \beta_{o,n} = 0.5$, then use $T_{A,n} = \text{MED}(n)$.

**Example 8.4.** These two stage trimmed means are almost as easy to compute as the classical trimmed mean, and no knowledge of the unknown parameters is needed to do inference. First, order the data and find the number of cases $L(M_n)$ less than $\text{MED}(n) - k_1\text{MAD}(n)$ and the number of cases $n - U(M_n)$ greater than $\text{MED}(n) + k_2\text{MAD}(n)$. (These are the cases trimmed by the metrically trimmed mean $M_n$, but $M_n$ need not be computed.) Next, convert these two numbers into percentages and round both percentages up to the nearest integer. For $T_{S,n}$ find the maximum of the two percentages. For example, suppose that there are $n = 205$ cases and $M_n$ trims the smallest 15 cases and the largest 20 cases. Then $L(M_n)/n = 0.073$ and $1 - (U(M_n)/n) = 0.0976$. Hence $M_n$ trimmed the 7.3% smallest cases and the 9.76% largest cases, and $T_{S,n}$ is the 10% trimmed mean while $T_{A,n}$ is the (0.08, 0.10) trimmed mean.

**Definition 8.17.** The standard error $\text{SE}_{RM}$ for the two stage trimmed means given in Definitions 8.11, 8.15, or 8.16 is

$$SE_{RM}(L_n, U_n) = \sqrt{V_{SW}(L_n, U_n)/n}$$

where the *scaled Winsorized variance* $V_{SW}(L_n, U_n) =$

$$\frac{[L_n Y^2_{(L_n+1)} + \sum_{i=L_n+1}^{U_n} Y^2_{(i)} + (n - U_n)Y^2_{(U_n)}] - \ n \ [W_n(L_n, U_n)]^2}{(n-1)[(U_n - L_n)/n]^2}. \quad (8.13)$$

**Remark 8.1.** A simple method for computing $V_{SW}(L_n, U_n)$ has the following steps. First, find $d_1, ..., d_n$ where

$$d_i = \begin{cases} Y_{(L_n+1)}, & i \leq L_n \\ Y_{(i)}, & L_n + 1 \leq i \leq U_n \\ Y_{(U_n)}, & i \geq U_n + 1. \end{cases}$$

Then the Winsorized variance is the sample variance $S^2_n(d_1, ..., d_n)$ of $d_1, ..., d_n$, and the scaled Winsorized variance

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2}. \tag{8.14}$$

Notice that the SE given in Definition 8.17 is the SE for the $\delta$ trimmed mean where $L_n$ and $U_n$ are fixed constants rather than random.

**Application 8.4.** Let $T_n$ be the two stage (symmetrically or) asymmetrically trimmed mean that trims the $L_n$ smallest cases and the $n - U_n$ largest cases. Then for the one and two sample procedures described in Section 5.1, use the one sample standard error $\mathrm{SE}_{RM}(L_n, U_n)$ given in Definition 8.17 and the $t_p$ distribution where the degrees of freedom $p = U_n - L_n - 1$.

The CIs and tests for the $\delta$ trimmed mean and two stage trimmed means given by Applications 8.3 and 8.4 are very similar once $L_n$ has been computed. For example, a large sample 100 $(1 - \alpha)\%$ confidence interval (CI) for $\mu_T$ is

$$[T_n - t_{U_n - L_n - 1, 1-\frac{\alpha}{2}} SE_{RM}(L_n, U_n), T_n + t_{U_n - L_n - 1, 1-\frac{\alpha}{2}} SE_{RM}(L_n, U_n)] \tag{8.15}$$

where $P(t_p \leq t_{p,1-\frac{\alpha}{2}}) = 1 - \alpha/2$ if $t_p$ is from a $t$ distribution with $p$ degrees of freedom. Section 8.1.6 provides the asymptotic theory for the $\delta$ and two stage trimmed means and shows that $\mu_T$ is the mean of a truncated distribution. Next Examples 8.2 and 8.3 are repeated using the intervals based on the two stage trimmed means instead of the median.

**Example 8.5.** Let the data be 6, 9, 9, 7, 8, 9, 9, 7. Assume the data came from a symmetric distribution with mean $\mu$, and find a 95% CI for $\mu$.

**Solution.** If $T_{A,n}$ or $T_{S,n}$ is used with the metrically trimmed mean that uses $k = k_1 = k_2$, e.g. $k = 6$, then $\mu_T(a, b) = \mu$. When computing small examples by hand, it is convenient to sort the data:
6, 7, 7, 8, 9, 9, 9, 9.
Thus $\mathrm{MED}(n) = (8 + 9)/2 = 8.5$. The ordered residuals $Y_{(i)} - \mathrm{MED}(n)$ are
-2.5, -1.5, -1.5, 0.5, 0.5, 0.5, 0.5, 0.5.
Find the absolute values and sort them to get
0.5, 0.5, 0.5, 0.5, 0.5, 1.5, 1.5, 2.5.
Then $\mathrm{MAD}(n) = 0.5$, $\mathrm{MED}(n) - 6MAD(n) = 5.5$, and $\mathrm{MED}(n) + 6MAD(n) = 11.5$. Hence no cases are trimmed by the metrically trimmed mean, i.e. $L(M_n) = 0$ and $U(M_n) = n = 8$. Thus $L_n = \lfloor 8(0) \rfloor = 0$, and $U_n = n - L_n = 8$. Since no cases are trimmed by the two stage trimmed means, the robust interval will have the same endpoints as the classical t–interval. To see this, note that $M_n = T_{S,n} = T_{A,n} = \overline{Y} = (6 + 7 + 7 + 8 + 9 + 9 + 9 + 9)/8 = 8 = W_n(L_n, U_n)$. Now $V_{SW}(L_n, U_n) = (1/7)[\sum_{i=1}^{n} Y_{(i)}^2 - 8(8^2)]/[8/8]^2 = (1/7)[(522 - 8(64)] = 10/7 \approx 1.4286$, and $t_{7,0.975} \approx 2.365$. Hence the 95% CI for $\mu$ is $8 \pm 2.365(\sqrt{1.4286/8}) = [7.001, 8.999]$.

**Example 8.6.** In the last example, what happens if a 6 becomes 66 and a 9 becomes 99? Use $k = 6$ and $T_{A,n}$. Then the ordered data are
7, 7, 8, 9, 9, 9, 66, 99.

Thus MED$(n) = 9$ and MAD$(n) = 1.5$. With $k = 6$, the metrically trimmed mean $M_n$ trims the two values 66 and 99. Hence the left and right trimming proportions of the metrically trimmed mean are 0.0 and $0.25 = 2/8$, respectively. These numbers are also the left and right trimming proportions of $T_{A,n}$ since after converting these proportions into percentages, both percentages are integers. Thus $L_n = \lfloor 0 \rfloor = 0$, $U_n = \lfloor 0.75(8) \rfloor = 6$ and the two stage asymmetrically trimmed mean trims 66 and 99. So $T_{A,n} = 49/6 \approx 8.1667$. To compute the scaled Winsorized variance, use Remark 8.3 to find that the $d_i$'s are
7, 7, 8, 9, 9, 9, 9, 9
and

$$V_{SW} = \frac{S_n^2(d_1, ..., d_8)}{[(6-0)/8]^2} \approx \frac{0.8393}{.5625} \approx 1.4921.$$

Hence the robust confidence interval is $8.1667 \pm t_{5,0.975}\sqrt{1.4921/8} \approx 8.1667 \pm 1.1102 \approx [7.057, 9.277]$. The classical confidence interval $\overline{Y} \pm t_{n-1,0.975} S/\sqrt{n}$ blows up and is equal to $[-2.955, 56.455]$.

**Example 8.7.** Use $k = 6$ and $T_{A,n}$ to compute a robust CI using the 87 heights from the Buxton (1920) data that includes 5 outliers. The mean height is $\overline{Y} = 1598.862$ while $T_{A,n} = 1695.22$. The classical 95% CI is [1514.206,1683.518] and is more than five times as long as the robust 95% CI which is [1679.907,1710.532]. In this example the five outliers can be corrected. For the corrected data, no cases are trimmed and the robust and classical estimators have the same values. The results are $\overline{Y} = 1692.356 = T_{A,n}$ and the robust and classical 95% CIs are both [1678.595,1706.118]. Note that the outliers did not have much affect on the robust confidence interval.

### *8.1.3* Asymptotics for Two Stage Trimmed Means

Large sample theory is very important for understanding robust statistics. Truncated and Winsorized random variables are important because they simplify the asymptotic theory of robust estimators. Let $Y$ be a random variable with continuous cdf $F$ and let $\alpha = F(a) < F(b) = \beta$. Thus $\alpha$ is the *left trimming proportion* and $1 - \beta$ is the *right trimming proportion.* Let $F(a-) = P(Y < a)$. (Refer to Section 1.8 for the notation used below.)

**Definition 8.18.** The *truncated random variable* $Y_T \equiv Y_T(a, b)$ with *truncation points* $a$ and $b$ has cdf

$$F_{Y_T}(y|a, b) = G(y) = \frac{F(y) - F(a-)}{F(b) - F(a-)} \tag{8.16}$$

for $a \leq y \leq b$. Also $G$ is 0 for $y < a$ and $G$ is 1 for $y > b$. The mean and variance of $Y_T$ are

$$\mu_T = \mu_T(a,b) = \int_{-\infty}^{\infty} y dG(y) = \frac{\int_a^b y dF(y)}{\beta - \alpha} \tag{8.17}$$

and

$$\sigma_T^2 = \sigma_T^2(a,b) = \int_{-\infty}^{\infty} (y - \mu_T)^2 dG(y) = \frac{\int_a^b y^2 dF(y)}{\beta - \alpha} - \mu_T^2.$$

See Cramér (1946, p. 247).

**Definition 8.19.** The *Winsorized random variable*

$$Y_W = Y_W(a,b) = \begin{cases} a, & Y \leq a \\ Y, & a \leq Y \leq b \\ b, & Y \geq b. \end{cases}$$

If the cdf of $Y_W(a,b) = Y_W$ is $F_W$, then

$$F_W(y) = \begin{cases} 0, & y < a \\ F(a), & y = a \\ F(y), & a < y < b \\ 1, & y \geq b. \end{cases}$$

Since $Y_W$ is a mixture distribution with a point mass at $a$ and at $b$, the mean and variance of $Y_W$ are

$$\mu_W = \mu_W(a,b) = \alpha a + (1-\beta)b + \int_a^b y dF(y)$$

and

$$\sigma_W^2 = \sigma_W^2(a,b) = \alpha a^2 + (1-\beta)b^2 + \int_a^b y^2 dF(y) - \mu_W^2.$$

**Definition 8.20.** The *quantile function*

$$F_Q^{-1}(t) = Q(t) = \inf\{y : F(y) \geq t\}. \tag{8.18}$$

The *sample* $\rho$ *quantile* $\hat{\xi}_{n,\rho} = Y_{(\lceil n\rho \rceil)} = \hat{y}_\rho$. The *population quantile* $y_\rho = \pi_\rho = \xi_\rho = Q(\rho)$ where $0 < \rho < 1$.

**Warning:** Software often uses a slightly different definition of the sample quantile then the one given in Definition 8.20.

Note that $Q(t)$ is the left continuous inverse of $F$ and if $F$ is strictly increasing and continuous, then $F$ has an inverse $F^{-1}$ and $F^{-1}(t) = Q(t)$. The following conditions on the cdf are used.

**Regularity Conditions.** (R1) Let $Y_1, \ldots, Y_n$ be iid with cdf $F$.
(R2) Let F be continuous and strictly increasing at $a = Q(\alpha)$ and $b = Q(\beta)$.

The following theorem is proved in Bickel (1965), Stigler (1973), and Shorack and Wellner (1986, p. 678-679). The $\alpha$ trimmed mean is asymptotically equivalent to the $(\alpha, 1-\alpha)$ trimmed mean. Let $T_n$ be the $(\alpha, 1-\beta)$ trimmed mean. Theorem 8.3 shows that the standard error $\mathrm{SE}_{RM}$ given in the previous section is estimating the appropriate asymptotic standard deviation of $T_n$.

**Theorem 8.2.** If conditions (R1) and (R2) hold and if $0 < \alpha < \beta < 1$, then

$$\sqrt{n}(T_n - \mu_T(a,b)) \xrightarrow{D} N\left[0, \frac{\sigma^2_W(a,b)}{(\beta-\alpha)^2}\right]. \tag{8.19}$$

**Theorem 8.3: Shorack and Wellner (1986, p. 680).** Assume that regularity conditions (R1) and (R2) hold and that

$$\frac{L_n}{n} \xrightarrow{P} \alpha \text{ and } \frac{U_n}{n} \xrightarrow{P} \beta. \tag{8.20}$$

Then

$$V_{SW}(L_n, U_n) \xrightarrow{P} \frac{\sigma^2_W(a,b)}{(\beta-\alpha)^2}.$$

Since $L_n = \lfloor n\alpha \rfloor$ and $U_n = n - L_n$ (or $L_n = \lfloor n\alpha \rfloor$ and $U_n = \lfloor n\beta \rfloor$) satisfy the above lemma, the standard error $\mathrm{SE}_{RM}$ can be used for both trimmed means and two stage trimmed means: $\mathrm{SE}_{RM}(L_n, U_n) = \sqrt{V_{SW}(L_n, U_n)/n}$ where the *scaled Winsorized variance* $V_{SW}(L_n, U_n) =$

$$\frac{[L_n Y^2_{(L_n+1)} + \sum_{i=L_n+1}^{U_n} Y^2_{(i)} + (n-U_n)Y^2_{(U_n)}] - \ n\ [W_n(L_n, U_n)]^2}{(n-1)[(U_n - L_n)/n]^2}.$$

Again $L_n$ is the number of cases trimmed to the left and $n - U_n$ is the number of cases trimmed to the right by the trimmed mean.

The following notation will be useful for finding the asymptotic distribution of the two stage trimmed means. Let $a = \mathrm{MED}(Y) - k\mathrm{MAD}(Y)$ and $b = \mathrm{MED}(Y) + k\mathrm{MAD}(Y)$ where $\mathrm{MED}(Y)$ and $\mathrm{MAD}(Y)$ are the population median and median absolute deviation respectively. Let $\alpha = F(a-) = P(Y < a)$ and let $\alpha_o \in C = \{0, 0.01, 0.02, ..., 0.49, 0.50\}$ be the smallest value in $C$ such that $\alpha_o \geq \alpha$. Similarly, let $\beta = F(b)$ and let $1 - \beta_o \in C$ be the smallest value in the index set $C$ such that $1 - \beta_o \geq 1 - \beta$. Let $\alpha_o = F(a_o-)$, and let $\beta_o = F(b_o)$. Recall that $L(M_n)$ is the number of cases trimmed to the left and that $n - U(M_n)$ is the number of cases trimmed to the right by the metrically trimmed mean $M_n$. Let $\alpha_{o,n} \equiv \hat{\alpha}_o$ be the smallest value in $C$ such that $\alpha_{o,n} \geq L(M_n)/n$, and let $1 - \beta_{o,n} \equiv 1 - \hat{\beta}_o$ be the smallest value in $C$ such that $1 - \beta_{o,n} \geq 1 - (U(M_n)/n)$. Then the robust estimator $T_{A,n}$ is the $(\alpha_{o,n}, 1 - \beta_{o,n})$ trimmed mean while $T_{S,n}$ is the $\max(\alpha_{o,n}, 1 - \beta_{o,n})100\%$

trimmed mean. The following theorem is useful for showing that $T_{A,n}$ is asymptotically equivalent to the $(\alpha_o, 1-\beta_o)$ trimmed mean and that $T_{S,n}$ is asymptotically equivalent to the $\max(\alpha_o, 1-\beta_o)$ trimmed mean. One proof of Theorem 8.5 is to show that $T_{A,n}$ and $T_{S,n}$ are model selection estimators where the probability $T_{A,n}$ selects the $(\alpha_o, 1-\beta_o)$ trimmed mean and the probability that $T_{S,n}$ selects the $\max(\alpha_o, 1-\beta_o)$ trimmed mean goes to one.

**Theorem 8.4: Shorack and Wellner (1986, p. 682-683).** Let F have a strictly positive and continuous derivative in some neighborhood of $\text{MED}(Y) \pm k\text{MAD}(Y)$. Assume that

$$\sqrt{n}(MED(n) - MED(Y)) = O_P(1) \tag{8.21}$$

and

$$\sqrt{n}(MAD(n) - MAD(X)) = O_P(1). \tag{8.22}$$

Then

$$\sqrt{n}(\frac{L(M_n)}{n} - \alpha) = O_P(1) \tag{8.23}$$

and

$$\sqrt{n}(\frac{U(M_n)}{n} - \beta) = O_P(1). \tag{8.24}$$

**Theorem 8.5.** Let $Y_1, ..., Y_n$ be iid from a distribution with cdf $F$ that has a strictly positive and continuous pdf $f$ on its support. Let $\alpha_M = \max(\alpha_o, 1-\beta_o) \leq 0.49$, $\beta_M = 1 - \alpha_M$, $a_M = F^{-1}(\alpha_M)$, and $b_M = F^{-1}(\beta_M)$. Assume that $\alpha$ and $1-\beta$ are not elements of $C = \{0, 0.01, 0.02, ..., 0.50\}$. Then

$$\sqrt{n}[T_{A,n} - \mu_T(a_o, b_o)] \xrightarrow{D} N\left[0, \frac{\sigma^2_W(a_o, b_o)}{(\beta_o - \alpha_o)^2}\right],$$

and

$$\sqrt{n}[T_{S,n} - \mu_T(a_M, b_M)] \xrightarrow{D} N\left[0, \frac{\sigma^2_W(a_M, b_M)}{(\beta_M - \alpha_M)^2}\right].$$

**Proof.** The first result follows from Theorem 8.2 if the probability that $T_{A,n}$ is the $(\alpha_o, 1-\beta_o)$ trimmed mean goes to one as $n$ tends to infinity. This condition holds if $L(M_n)/n \xrightarrow{D} \alpha$ and $U(M_n)/n \xrightarrow{D} \beta$. But these conditions follow from Theorem 8.4. The proof for $T_{S,n}$ is similar. □

### *8.1.4* Asymptotic Theory for the MAD

Let $\text{MD}(n) = \text{MED}(|Y_i - \text{MED}(Y)|,\ i = 1, \ldots, n)$. Since $\text{MD}(n)$ is a median and convergence results for the median are well known, see for example Serfling (1980, p. 74-77) or Theorem 2.6, it is simple to prove conver-

gence results for MAD($n$). Typically MED($n$) = MED($Y$) + $O_P(n^{-1/2})$ and MAD($n$) = MAD($Y$) + $O_P(n^{-1/2})$.

**Theorem 8.6.** If MED($n$) = MED($Y$) + $O_P(n^{-\delta})$ and MD($n$) = MAD($Y$) + $O_P(n^{-\delta})$, then MAD($n$) = MAD($Y$) + $O_P(n^{-\delta})$.

**Proof.** Let $W_i = |Y_i - \text{MED}(n)|$ and let $V_i = |Y_i - \text{MED}(Y)|$. Then

$$W_i = |Y_i - \text{MED}(Y) + \text{MED}(Y) - \text{MED}(n)| \le V_i + |\text{MED}(Y) - \text{MED}(n)|,$$

and

$$\text{MAD}(n) = \text{MED}(W_1, \ldots, W_n) \le \text{MED}(V_1, \ldots, V_n) + |\text{MED}(Y) - \text{MED}(n)|.$$

Similarly

$$V_i = |Y_i - \text{MED}(n) + \text{MED}(n) - \text{MED}(Y)| \le W_i + |\text{MED}(n) - \text{MED}(Y)|$$

and thus

$$\text{MD}(n) = \text{MED}(V_1, \ldots, V_n) \le \text{MED}(W_1, \ldots, W_n) + |\text{MED}(Y) - \text{MED}(n)|.$$

Combining the two inequalities shows that

$$\text{MD}(n) - |\text{MED}(Y) - \text{MED}(n)| \le \text{MAD}(n) \le \text{MD}(n) + |\text{MED}(Y) - \text{MED}(n)|,$$

or

$$|\text{MAD}(n) - \text{MD}(n)| \le |\text{MED}(n) - \text{MED}(Y)|. \tag{8.25}$$

Adding and subtracting MAD($Y$) to the left hand side shows that

$$|\text{MAD}(n) - \text{MAD}(Y) - O_P(n^{-\delta})| = O_P(n^{-\delta}) \tag{8.26}$$

and the result follows. □

The main point of the following theorem is that the joint distribution of MED($n$) and MAD($n$) is asymptotically normal. Hence the limiting distribution of MED($n$) + $k$MAD($n$) is also asymptotically normal for any constant $k$. The parameters of the covariance matrix are quite complex and hard to estimate. The assumptions of $f$ used in Theorem 8.7 guarantee that MED($Y$) and MAD($Y$) are unique.

**Theorem 8.7: Falk (1997).** Let the cdf $F$ of $Y$ be continuous near and differentiable at MED($Y$) = $F^{-1}(1/2)$ and MED($Y$)$\pm$MAD($Y$). Assume that $f = F'$, $f(F^{-1}(1/2)) > 0$, and $A \equiv f(F^{-1}(1/2) - \text{MAD}(Y)) + f(F^{-1}(1/2) + \text{MAD}(Y)) > 0$. Let $C \equiv f(F^{-1}(1/2) - \text{MAD}(Y)) - f(F^{-1}(1/2) + \text{MAD}(Y))$, and let $B \equiv C^2 + 4Cf(F^{-1}(1/2))[1 - F(F^{-1}(1/2) - \text{MAD}(Y)) - F(F^{-1}(1/2) + \text{MAD}(Y))]$. Then

$$\sqrt{n}\left(\begin{pmatrix}\text{MED}(n)\\ \text{MAD}(n)\end{pmatrix} - \begin{pmatrix}\text{MED}(Y)\\ \text{MAD}(Y)\end{pmatrix}\right) \xrightarrow{D}$$

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_M^2 & \sigma_{M,D}\\ \sigma_{M,D} & \sigma_D^2\end{pmatrix}\right) \tag{8.27}$$

where

$$\sigma_M^2 = \frac{1}{4f^2(F^{-1}(\frac{1}{2}))},\ \sigma_D^2 = \frac{1}{4A^2}(1 + \frac{B}{f^2(F^{-1}(\frac{1}{2}))}),$$

and

$$\sigma_{M,D} = \frac{1}{4Af(F^{-1}(\frac{1}{2}))}(1 - 4F(F^{-1}(\frac{1}{2}) + \text{MAD}(Y)) + \frac{C}{f(F^{-1}(\frac{1}{2}))}).$$

Determining whether the population median and mad are unique can be useful. Recall that $F(y) = P(Y \le y)$ and $F(y-) = P(Y < y)$. The median is unique unless there is a flat spot at $F^{-1}(0.5)$, that is, unless there exist $a$ and $b$ with $a < b$ such that $F(a) = F(b) = 0.5$. If $\text{MED}(Y)$ is unique, then $\text{MAD}(Y)$ is unique unless $F$ has flat spots at both $F^{-1}(\text{MED}(Y) - \text{MAD}(Y))$ and $F^{-1}(\text{MED}(Y) + \text{MAD}(Y))$. Moreover, $\text{MAD}(Y)$ is unique unless there exist $a_1 < a_2$ and $b_1 < b_2$ such that $F(a_1) = F(a_2)$, $F(b_1) = F(b_2)$,

$$P(a_i \le Y \le b_i) = F(b_i) - F(a_i-) \ge 0.5,$$

and

$$P(Y \le a_i) + P(Y \ge b_i) = F(a_i) + 1 - F(b_i-) \ge 0.5$$

for $i = 1, 2$. The following theorem gives some simple bounds for $\text{MAD}(Y)$.

**Theorem 8.8.** Assume $\text{MED}(Y)$ and $\text{MAD}(Y)$ are unique. a) Then

$$\min\{\text{MED}(Y) - F^{-1}(0.25), F^{-1}(0.75) - \text{MED}(Y)\} \le \text{MAD}(Y) \le$$

$$\max\{\text{MED}(Y) - F^{-1}(0.25), F^{-1}(0.75) - \text{MED}(Y)\}. \tag{8.28}$$

b) If $Y$ is symmetric about $\mu = F^{-1}(0.5)$, then the three terms in a) are equal.
c) If the distribution is symmetric about zero, then $\text{MAD}(Y) = F^{-1}(0.75)$.
d) If $Y$ is symmetric and continuous with a finite second moment, then

$$\text{MAD}(Y) \le \sqrt{2\text{VAR}(Y)}.$$

e) Suppose $Y \in [a, b]$. Then

$$0 \le \text{MAD}(Y) \le m = \min\{\text{MED}(Y) - a, b - \text{MED}(Y)\} \le (b - a)/2,$$

and the inequalities are sharp.

**Proof.** a) This result follows since half the mass is between the upper and lower quartiles and the median is between the two quartiles.

b) and c) are corollaries of a).

d) This inequality holds by Chebyshev's inequality, since

$$P(\ |Y - E(Y)| \geq \text{MAD}(Y)\ ) = 0.5 \geq P(\ |Y - E(Y)| \geq \sqrt{2\text{VAR}(Y)}\ ),$$

and $E(Y) = \text{MED}(Y)$ for symmetric distributions with finite second moments.

e) Note that if $\text{MAD}(Y) > m$, then either $\text{MED}(Y) - \text{MAD}(Y) < a$ or $\text{MED}(Y) + \text{MAD}(Y) > b$. Since at least half of the mass is between $a$ and $\text{MED}(Y)$ and between $\text{MED}(Y)$ and $b$, this contradicts the definition of $\text{MAD}(Y)$. To see that the inequalities are sharp, note that if at least half of the mass is at some point $c \in [a, b]$, than $\text{MED}(Y) = c$ and $\text{MAD}(Y) = 0$. If each of the points $a, b$, and $c$ has $1/3$ of the mass where $a < c < b$, then $\text{MED}(Y) = c$ and $\text{MAD}(Y) = m$. □

Many other results for $\text{MAD}(Y)$ and $\text{MAD}(n)$ are possible. For example, note that Theorem 8.8 b) implies that when $Y$ is symmetric, $\text{MAD}(Y) = F^{-1}(3/4) - \mu$ and $F(\mu + \text{MAD}(Y)) = 3/4$. Also note that $\text{MAD}(Y)$ and the interquartile range $IQR(Y)$ are related by

$$2\text{MAD}(Y) = IQR(Y) \equiv y_{0.75} - y_{0.25}$$

when $Y$ is symmetric.

### *8.1.5* Truncated Distributions

Truncated distributions can be used to simplify the asymptotic theory of robust estimators of location and regression. This subsection is useful when the underlying distribution is exponential, double exponential, normal, or Cauchy.

Definitions 8.18 and 8.19 defined the truncated random variable $Y_T(a, b)$ and the Winsorized random variable $Y_W(a, b)$. Let $Y$ have cdf $F$ and let the truncated random variable $Y_T(a, b)$ have the cdf $F_{T(a,b)}$. The following lemma illustrates the relationship between the means and variances of $Y_T(a, b)$ and $Y_W(a, b)$. Note that $Y_W(a, b)$ is a mixture of $Y_T(a, b)$ and two point masses at $a$ and $b$. Let $c = \mu_T(a, b) - a$ and $d = b - \mu_T(a, b)$.

**Theorem 8.9.** Let $a = \mu_T(a, b) - c$ and $b = \mu_T(a, b) + d$. Then

a) $\mu_W(a, b) = \mu_T(a, b) - \alpha c + (1 - \beta)d$, and

b) $\sigma^2_W(a, b) = (\beta - \alpha)\sigma^2_T(a, b) + (\alpha - \alpha^2)c^2 + [(1 - \beta) - (1 - \beta)^2]d^2 + 2\alpha(1 - \beta)cd$.

c) If $\alpha = 1 - \beta$ then

$$\sigma_W^2(a,b) = (1-2\alpha)\sigma_T^2(a,b) + (\alpha-\alpha^2)(c^2+d^2) + 2\alpha^2 cd.$$

d) If $c = d$ then

$$\sigma_W^2(a,b) = (\beta-\alpha)\sigma_T^2(a,b) + [\alpha - \alpha^2 + 1 - \beta - (1-\beta)^2 + 2\alpha(1-\beta)]d^2.$$

e) If $\alpha = 1-\beta$ and $c = d$, then $\mu_W(a,b) = \mu_T(a,b)$ and

$$\sigma_W^2(a,b) = (1-2\alpha)\sigma_T^2(a,b) + 2\alpha d^2.$$

**Proof.** We will prove b) since its proof contains the most algebra. Now

$$\sigma_W^2 = \alpha(\mu_T - c)^2 + (\beta-\alpha)(\sigma_T^2 + \mu_T^2) + (1-\beta)(\mu_T + d)^2 - \mu_W^2.$$

Collecting terms shows that

$$\sigma_W^2 = (\beta-\alpha)\sigma_T^2 + (\beta - \alpha + \alpha + 1 - \beta)\mu_T^2 + 2[(1-\beta)d - \alpha c]\mu_T$$
$$+\alpha c^2 + (1-\beta)d^2 - \mu_W^2.$$

From a),

$$\mu_W^2 = \mu_T^2 + 2[(1-\beta)d - \alpha c]\mu_T + \alpha^2 c^2 + (1-\beta)^2 d^2 - 2\alpha(1-\beta)cd,$$

and we find that

$$\sigma_W^2 = (\beta-\alpha)\sigma_T^2 + (\alpha-\alpha^2)c^2 + [(1-\beta) - (1-\beta)^2]d^2 + 2\alpha(1-\beta)cd. \quad \square$$

**The Truncated Exponential Distribution**

Let $Y$ be a (one sided) truncated exponential $TEXP(\lambda, b)$ random variable. Then the pdf of $Y$ is

$$f_Y(y|\lambda, b) = \frac{\frac{1}{\lambda}e^{-y/\lambda}}{1 - \exp(-\frac{b}{\lambda})}$$

for $0 < y \leq b$ where $\lambda > 0$. Let $b = k\lambda$, and let

$$c_k = \int_0^{k\lambda} \frac{1}{\lambda} e^{-y/\lambda} dy = 1 - e^{-k}.$$

Next we will find the first two moments of $Y \sim TEXP(\lambda, b = k\lambda)$ for $k > 0$.

**Theorem 8.10.** If $Y$ is $TEXP(\lambda, b = k\lambda)$ for $k > 0$, then

$$a)\ E(Y) = \lambda \left[ \frac{1 - (k+1)e^{-k}}{1 - e^{-k}} \right],$$

and

$$b)\ E(Y^2) = 2\lambda^2 \left[ \frac{1 - \frac{1}{2}(k^2 + 2k + 2)e^{-k}}{1 - e^{-k}} \right].$$

See Problem 8.6 for a related result.

**Proof.** a) Note that

$$c_k E(Y) = \int_0^{k\lambda} \frac{y}{\lambda} e^{-y/\lambda} dy = -ye^{-y/\lambda}|_0^{k\lambda} + \int_0^{k\lambda} e^{-y/\lambda} dy$$

(use integration by parts). So

$$c_k E(Y) = -k\lambda e^{-k} + (-\lambda e^{-y/\lambda})|_0^{k\lambda} = -k\lambda e^{-k} + \lambda(1 - e^{-k}).$$

Hence

$$E(Y) = \lambda \left[ \frac{1 - (k+1)e^{-k}}{1 - e^{-k}} \right].$$

b) Note that

$$c_k E(Y^2) = \int_0^{k\lambda} \frac{y^2}{\lambda} e^{-y/\lambda} dy.$$

Since

$$\frac{d}{dy}[-(y^2 + 2\lambda y + 2\lambda^2)e^{-y/\lambda}] = \frac{1}{\lambda} e^{-y/\lambda}(y^2 + 2\lambda y + 2\lambda^2) - e^{-y/\lambda}(2y + 2\lambda)$$

$$= y^2 \frac{1}{\lambda} e^{-y/\lambda},$$

we have $c_k E(Y^2) = [-(y^2 + 2\lambda y + 2\lambda^2)e^{-y/\lambda}]_0^{k\lambda} =$
$-(k^2\lambda^2 + 2\lambda^2 k + 2\lambda^2)e^{-k} + 2\lambda^2$. So the result follows. □

Since as $k \to \infty$, $E(Y) \to \lambda$, and $E(Y^2) \to 2\lambda^2$, we have $\text{VAR}(Y) \to \lambda^2$. If $k = 9\log(2) \approx 6.24$, then $E(Y) \approx .998\lambda$, and $E(Y^2) \approx 0.95(2\lambda^2)$.

**The Truncated Double Exponential Distribution**

Suppose that $X$ is a double exponential $DE(\mu, \lambda)$ random variable. Then $\text{MED}(X) = \mu$ and $\text{MAD}(X) = \log(2)\lambda$. Let $c = k\log(2)$, and let the truncation points $a = \mu - k\text{MAD}(X) = \mu - c\lambda$ and $b = \mu + kMAD(X) = \mu + c\lambda$. Let $X_T(a, b) \equiv Y$ be the truncated double exponential $TDE(\mu, \lambda, a, b)$ random variable. Then for $a \le y \le b$, the pdf of $Y$ is

$$f_Y(y|\mu, \lambda, a, b) = \frac{1}{2\lambda(1 - \exp(-c))} \exp(-|y - \mu|/\lambda).$$

**Theorem 8.11.** a) $E(Y) = \mu$.

$$b)\ \mathrm{VAR}(Y) = 2\lambda^2 \left[ \frac{1 - \frac{1}{2}(c^2 + 2c + 2)e^{-c}}{1 - e^{-c}} \right].$$

**Proof.** a) follows by symmetry and b) follows from Theorem 8.10 b) since $\mathrm{VAR}(Y) = E[(Y-\mu)^2] = E(W_T^2)$ where $W_T$ is $TEXP(\lambda, b = c\lambda)$. $\square$

As $c \to \infty$, $\mathrm{VAR}(Y) \to 2\lambda^2$. If $k = 9$, then $c = 9\log(2) \approx 6.24$ and $\mathrm{VAR}(Y) \approx 0.95(2\lambda^2)$.

**The Truncated Normal Distribution**

Now if $X$ is $N(\mu, \sigma^2)$ then let $Y$ be a truncated normal $TN(\mu, \sigma^2, a, b)$ random variable. Then $f_Y(y) = \dfrac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} I_{[a,b]}(y)$ where $\Phi$ is the standard normal cdf. The indicator function

$$I_{[a,b]}(y) = 1 \ \text{ if } \ \mathrm{a} \leq \mathrm{y} \leq \mathrm{b}$$

and is zero otherwise. Let $\phi$ be the standard normal pdf.

**Theorem 8.12.** $E(Y) = \mu + \left[ \dfrac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \right] \sigma$, and

$$V(Y) = \sigma^2 \left[ 1 + \frac{(\frac{a-\mu}{\sigma})\phi(\frac{a-\mu}{\sigma}) - (\frac{b-\mu}{\sigma})\phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \right] - \sigma^2 \left[ \frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \right]^2.$$

(See Johnson and Kotz 1970a, p. 83.)

**Proof.** Let $c = \dfrac{1}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$.

Then $E(Y) = \int_a^b y f_Y(y) dy$. Hence

$$\frac{1}{c}E(Y) = \int_a^b \frac{y}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$= \int_a^b \left(\frac{y-\mu}{\sigma}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy \ + \ \frac{\mu}{\sigma} \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$= \int_a^b \left(\frac{y-\mu}{\sigma}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy \ + \ \mu \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy.$$

Note that the integrand of the last integral is the pdf of a $N(\mu, \sigma^2)$ distribution. Let $z = (y-\mu)/\sigma$. Thus $dz = dy/\sigma$, and $E(Y)/c =$

$$\int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \sigma \frac{z}{\sqrt{2\pi}} e^{-z^2/2} dz + \frac{\mu}{c} = \frac{\sigma}{\sqrt{2\pi}} (-e^{-z^2/2}) \Big|_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} + \frac{\mu}{c}.$$

Multiplying both sides by $c$ gives the expectation result.

$$E(Y^2) = \int_a^b y^2 f_Y(y)dy.$$

Hence

$$\frac{1}{c}E(Y^2) = \int_a^b \frac{y^2}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy$$

$$= \sigma \int_a^b \left(\frac{y^2}{\sigma^2} - \frac{2\mu y}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right)\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy$$

$$+\sigma \int_a^b \frac{2y\mu - \mu^2}{\sigma^2}\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy$$

$$= \sigma \int_a^b \left(\frac{y-\mu}{\sigma}\right)^2 \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy + 2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c}.$$

Let $z = (y-\mu)/\sigma$. Then $dz = dy/\sigma$, $dy = \sigma dz$, and $y = \sigma z + \mu$. Hence

$$\frac{E(Y^2)}{c} = 2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} + \sigma \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \sigma \frac{z^2}{\sqrt{2\pi}} e^{-z^2/2}dz.$$

Next integrate by parts with $w = z$ and $dv = ze^{-z^2/2}dz$. Then $E(Y^2)/c =$

$$2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} + \frac{\sigma^2}{\sqrt{2\pi}}[(-ze^{-z^2/2})|_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} + \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-z^2/2}dz]$$

$$= 2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} + \sigma^2\left[\left(\frac{a-\mu}{\sigma}\right)\phi\left(\frac{a-\mu}{\sigma}\right) - \left(\frac{b-\mu}{\sigma}\right)\phi\left(\frac{b-\mu}{\sigma}\right) + \frac{1}{c}\right].$$

Using

$$\text{VAR}(Y) = c\frac{1}{c}E(Y^2) - (E(Y))^2$$

gives the result. □

**Theorem 8.13.** Let $Y$ be $TN(\mu, \sigma^2, a = \mu - k\sigma, b = \mu + k\sigma)$. Then $E(Y) = \mu$ and $V(Y) = \sigma^2\left[1 - \frac{2k\phi(k)}{2\Phi(k) - 1}\right]$.

**Proof.** Use the symmetry of $\phi$, the fact that $\Phi(-x) = 1 - \Phi(x)$, and the above lemma to get the result. □

Examining $V(Y)$ for several values of $k$ shows that the $TN(\mu, \sigma^2, a = \mu - k\sigma, b = \mu + k\sigma)$ distribution does not change much for $k > 3.0$. See Table 8.2.

**The Truncated Cauchy Distribution**

**Table 8.2** Variances for Several Truncated Normal Distributions

| $k$ | $V(Y)$ |
|---|---|
| 2.0 | $0.774\sigma^2$ |
| 2.5 | $0.911\sigma^2$ |
| 3.0 | $0.973\sigma^2$ |
| 3.5 | $0.994\sigma^2$ |
| 4.0 | $0.999\sigma^2$ |

If $X$ is a Cauchy $C(\mu, \sigma)$ random variable, then MED$(X) = \mu$ and MAD$(X) = \sigma$. If $Y$ is a truncated Cauchy $TC(\mu, \sigma, \mu - a\sigma, \mu + b\sigma)$ random variable, then

$$f_Y(y) = \frac{1}{\tan^{-1}(b) + \tan^{-1}(a)} \; \frac{1}{\sigma[1 + (\frac{y-\mu}{\sigma})^2]}$$

for $\mu - a\sigma < y < \mu + b\sigma$. For the following theorem, see Johnson and Kotz (1970a, p. 162) and Dahiya, Staneski and Chaganty (2001).

**Theorem 8.14.** a)

$$E(Y) = \mu + \sigma \left( \frac{\log(1 + b^2) - \log(1 + a^2)}{2[\tan^{-1}(b) + \tan^{-1}(a)]} \right), \quad \text{and}$$

$$V(Y) = \sigma^2 \left[ \frac{b + a - \tan^{-1}(b) - \tan^{-1}(a)}{\tan^{-1}(b) + \tan^{-1}(a)} - \left( \frac{\log(1 + b^2) - \log(1 + a^2)}{\tan^{-1}(b) + \tan^{-1}(a)} \right)^2 \right].$$

b) If $a = b$, then $E(Y) = \mu$, and $V(Y) = \sigma^2 \left[ \frac{b - \tan^{-1}(b)}{\tan^{-1}(b)} \right]$.

### *8.1.6* Asymptotic Variances for Trimmed Means

The truncated distributions will be useful for finding the asymptotic variances of trimmed and two stage trimmed means. Assume that $Y$ is from a symmetric location–scale family with parameters $\mu$ and $\sigma$ and that the truncation points are $a = \mu - z\sigma$ and $b = \mu + z\sigma$. Recall that for the trimmed mean $T_n$,

$$\sqrt{n}(T_n - \mu_T(a, b)) \xrightarrow{D} N \left[ 0, \frac{\sigma_W^2(a, b)}{(\beta - \alpha)^2} \right].$$

Since the family is symmetric and the truncation is symmetric, $\alpha = F(a) = 1 - \beta$ and $\mu_T(a, b) = \mu$.

**Definition 8.21.** Let $Y_1, ..., Y_n$ be iid random variables and let $D_n \equiv D_n(Y_1, ..., Y_n)$ be an estimator of a parameter $\mu_D$ such that

$$\sqrt{n}(D_n - \mu_D) \xrightarrow{D} N(0, \sigma_D^2).$$

Then the *asymptotic variance* of $\sqrt{n}(D_n - \mu_D)$ is $\sigma_D^2$ and the *asymptotic variance* (AV) of $D_n$ is $\sigma_D^2/n$. If $S_D^2$ is a consistent estimator of $\sigma_D^2$, then the (asymptotic) *standard error* (SE) of $D_n$ is $S_D/\sqrt{n}$.

**Remark 8.2.** In the literature, usually either $\sigma_D^2$ or $\sigma_D^2/n$ is called the asymptotic variance of $D_n$. The parameter $\sigma_D^2$ is a function of both the estimator $D_n$ and the underlying distribution $F$ of $Y_1$. Frequently $n\text{VAR}(D_n)$ converges in distribution to $\sigma_D^2$, but not always. See Staudte and Sheather (1990, p. 51) and Lehmann (1999, p. 232).

**Example 8.8.** If $Y_1, ..., Y_n$ are iid from a distribution with mean $\mu$ and variance $\sigma^2$, then by the central limit theorem,

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Recall that $\text{VAR}(\overline{Y}_n) = \sigma^2/n = AV(\overline{Y}_n)$ and that the standard error $SE(\overline{Y}_n) = S_n/\sqrt{n}$ where $S_n^2$ is the sample variance.

**Remark 8.3.** Returning to the trimmed mean $T_n$ where $Y$ is from a symmetric location–scale family, take $\mu = 0$ since the asymptotic variance does not depend on $\mu$. Then

$$n\ AV(T_n) = \frac{\sigma_W^2(a,b)}{(\beta - \alpha)^2} = \frac{\sigma_T^2(a,b)}{1 - 2\alpha} + \frac{2\alpha(F^{-1}(\alpha))^2}{(1-2\alpha)^2}.$$

See, for example, Bickel (1965). This formula is useful since the variance of the truncated distribution $\sigma_T^2(a,b)$ has been computed for several distributions in the previous subsection.

**Definition 8.22.** An estimator $D_n$ is a *location and scale equivariant estimator* if $D_n(\alpha + \beta Y_1, ..., \alpha + \beta Y_n) = \alpha + \beta D_n(Y_1, ..., Y_n)$ where $\alpha$ and $\beta$ are arbitrary real constants.

**Remark 8.4.** Many location estimators such as the sample mean, sample median, trimmed mean, metrically trimmed mean, and two stage trimmed means are equivariant. Let $Y_1, ..., Y_n$ be iid from a distribution with cdf $F_Y(y)$ and suppose that $D_n$ is an equivariant estimator of $\mu_D \equiv \mu_D(F_Y) \equiv \mu_D(F_Y(y))$. If $X_i = \alpha + \beta Y_i$ where $\beta \neq 0$, then the cdf of $X$ is $F_X(y) = F_Y((y - \alpha)/\beta)$. Suppose that

$$\mu_D(F_X) \equiv \mu_D[F_Y(\frac{y - \alpha}{\beta})] = \alpha + \beta \mu_D[F_Y(y)]. \tag{8.29}$$

Let $D_n(\boldsymbol{Y}) \equiv D_n(Y_1, ..., Y_n)$. If $\sqrt{n}[D_n(\boldsymbol{Y}) - \mu_D(F_Y(y))] \xrightarrow{D} N(0, \sigma_D^2)$, then

$$\sqrt{n}[D_n(\boldsymbol{X}) - \mu_D(F_X)] = \sqrt{n}[\alpha + \beta D_n(\boldsymbol{Y}) - (\alpha + \beta\mu_D(F_Y))] \xrightarrow{D} N(0, \beta^2\sigma_D^2).$$

This result is especially useful when $F$ is a cdf from a location–scale family with parameters $\mu$ and $\sigma$. In this case, Equation (8.29) holds when $\mu_D$ is the population mean, population median, and the population truncated mean with truncation points $a = \mu - z_1\sigma$ and $b = \mu + z_2\sigma$ (the parameter estimated by trimmed and two stage trimmed means).

Refer to the notation for two stage trimmed means below Theorem 8.3. Then from Theorem 8.5,

$$\sqrt{n}[T_{A,n} - \mu_T(a_o, b_o)] \xrightarrow{D} N\left[0, \frac{\sigma_W^2(a_o, b_o)}{(\beta_o - \alpha_o)^2}\right],$$

and

$$\sqrt{n}[T_{S,n} - \mu_T(a_M, b_M)] \xrightarrow{D} N\left[0, \frac{\sigma_W^2(a_M, b_M)}{(\beta_M - \alpha_M)^2}\right].$$

If the distribution of $Y$ is symmetric then $T_{A,n}$ and $T_{S,n}$ are asymptotically equivalent. It is important to note that no knowledge of the unknown distribution and parameters is needed to compute the two stage trimmed means and their standard errors.

The next three theorems find the asymptotic variance for trimmed and two stage trimmed means when the underlying distribution is normal, double exponential and Cauchy, respectively. Assume $a = \text{MED}(Y) - k\text{MAD}(Y)$ and $b = \text{MED}(Y) + k\text{MAD}(Y)$.

**Theorem 8.15.** Suppose that $Y$ comes from a normal $N(\mu, \sigma^2)$ distribution. Let $\Phi(x)$ be the cdf and let $\phi(x)$ be the density of the standard normal. Then for the $\alpha$ trimmed mean,

$$n\ AV = \left(\frac{1 - \frac{2z\phi(z)}{2\Phi(z)-1}}{1-2\alpha} + \frac{2\alpha z^2}{(1-2\alpha)^2}\right)\sigma^2 \tag{8.30}$$

where $\alpha = \Phi(-z)$, and $z = k\Phi^{-1}(0.75)$. For the two stage estimators, round $100\alpha$ up to the nearest integer $J$. Then use $\alpha_J = J/100$ and $z_J = -\Phi^{-1}(\alpha_J)$ in Equation (8.30).

**Proof.** If $Y$ follows the normal $N(\mu, \sigma^2)$ distribution, then $a = \mu - k\text{MAD}(Y)$ and $b = \mu + k\text{MAD}(Y)$ where $\text{MAD}(Y) = \Phi^{-1}(0.75)\sigma$. It is enough to consider the standard N(0,1) distribution since $n\ AV(T_n, N(\mu, \sigma^2)) = \sigma^2\ n\ AV(T_n, N(0,1))$. If $a = -z$ and $b = z$, then by Theorem 8.13,

$$\sigma_T^2(a, b) = 1 - \frac{2z\phi(z)}{2\Phi(z) - 1}.$$

Use Remark 8.3 with $z = k\Phi^{-1}(0.75)$, and $\alpha = \Phi(-z)$ to get Equation (8.30).

**Theorem 8.16.** Suppose that $Y$ comes from a double exponential DE(0,1) distribution. Then for the $\alpha$ trimmed mean,

$$n\ AV = \frac{\frac{2-(z^2+2z+2)e^{-z}}{1-e^{-z}}}{1-2\alpha} + \frac{2\alpha z^2}{(1-2\alpha)^2} \tag{8.31}$$

where $z = k\log(2)$ and $\alpha = 0.5\exp(-z)$. For the two stage estimators, round $100\alpha$ up to the nearest integer $J$. Then use $\alpha_J = J/100$ and let $z_J = -\log(2\alpha_J)$.

**Proof Sketch.** For the $DE(0,1)$ distribution, $\mathrm{MAD}(Y) = \log(2)$. If the DE(0,1) distribution is truncated at $-z$ and $z$, then use Remark 8.3 with

$$\sigma_T^2(-z,z) = \frac{2-(z^2+2z+2)e^{-z}}{1-e^{-z}}.$$

**Theorem 8.17.** Suppose that $Y$ comes from a Cauchy (0,1) distribution. Then for the $\alpha$ trimmed mean,

$$n\ AV = \frac{z-\tan^{-1}(z)}{(1-2\alpha)\tan^{-1}(z)} + \frac{2\alpha(\tan[\pi(\alpha-\frac{1}{2})])^2}{(1-2\alpha)^2} \tag{8.32}$$

where $z = k$ and

$$\alpha = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(z).$$

For the two stage estimators, round $100\alpha$ up to the nearest integer $J$. Then use $\alpha_J = J/100$ and let $z_J = \tan[\pi(\alpha_J - 0.5)]$.

**Proof Sketch.** For the $C(0,1)$ distribution, $\mathrm{MAD}(Y) = 1$. If the C(0,1) distribution is truncated at $-z$ and $z$, then use Remark 8.3 with

$$\sigma_T^2(-z,z) = \frac{z-\tan^{-1}(z)}{\tan^{-1}(z)}.$$

Next we give a theorem for the metrically trimmed mean $M_n$. Lopuhaä (1999) shows the following result. Suppose $(\hat{\boldsymbol{\mu}}_n, \boldsymbol{C}_n)$ is an estimator of multivariate location and dispersion. Suppose that the iid data follow an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution. Let $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ be the classical estimator applied to the set $J$ of cases with squared Mahalanobis distances $D_i^2(\hat{\boldsymbol{\mu}}_n, \boldsymbol{C}_n) \le k^2$. Under regularity conditions, if $(\hat{\boldsymbol{\mu}}_n, \boldsymbol{C}_n) \xrightarrow{P} (\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate $n^\delta$ where $0 < \delta \le 0.5$, then $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J) \xrightarrow{P} (\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ with the same rate $n^\delta$ where $s > 0$ and $d > 0$ are some constants. See Section 8.2 for discussion of the above quantities.

In the univariate setting with $p = 1$, let $\hat{\theta}_n = \hat{\mu}_n$ and let $D_n^2 = \boldsymbol{C}_n$ where $D_n$ is an estimator of scale. Suppose the classical estimator $(\overline{Y}_J, S_J^2) \equiv (\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ is applied to the set $J$ of cases with $\hat{\theta}_n - kD_n \leq Y_i \leq \hat{\theta}_n + kD_n$. Hence $\overline{Y}_J$ is the metrically trimmed mean $M_n$ with $k_1 = k_2 \equiv k$. See Definition 8.14.

The population quantity estimated by $(\overline{Y}_J, S_J^2)$ is the truncated mean and variance $(\mu_T(a,b), \sigma_T^2(a,b))$ of Definition 8.18 where $\hat{\theta}_n - kD_n \xrightarrow{P} a$ and $\hat{\theta}_n + kD_n \xrightarrow{P} b$. In the theorem below, the pdf corresponds to an elliptically contoured distribution with $p = 1$ and $\boldsymbol{\Sigma} = \tau^2$. Each pdf corresponds to a location scale family with location parameter $\mu$ and scale parameter $\tau$. Note that $(\hat{\theta}_n, D_n) = (\text{MED}(n), \text{MAD}(n))$ results in a $\sqrt{n}$ consistent estimator $(M_n, S_J^2)$.

**Assumption E1:** Suppose $Y_1, ..., Y_n$ are iid from an $EC_1(\mu, \tau^2, g)$ distribution with pdf

$$f(y) = \frac{c}{\tau} \; g\left[\left(\frac{y-\mu}{\tau}\right)^2\right]$$

where $g$ is continuously differentiable with finite 4th moment $\int y^4 g(y^2)dy < \infty$, $c > 0$ is some constant, $\tau > 0$ where $y$ and $\mu$ are real.

**Theorem 8.18.** Let $M_n$ be the metrically trimmed mean with $k_1 = k_2 \equiv k$. Assume (E1) holds. If $(\hat{\theta}_n, D_n^2) \xrightarrow{P} (\mu, s\tau^2)$ with rate $n^\delta$ for some constant $s > 0$ where $0 < \delta \leq 0.5$, then $(M_n, S_J^2) \xrightarrow{P} (\mu, \sigma_T^2(a,b))$ with the same rate $n^\delta$.

**Proof.** The result is a special case of Lopuhaä (1999) which shows that $(M_n, S_J^2) \xrightarrow{P} (\mu, d\tau^2)$ with rate $n^\delta$. Since $k_1 = k_2 = k$, $d\tau^2 = \sigma_T^2(a,b)$. □

Note that the classical estimator applied to the set $\tilde{J}$ of cases $Y_i$ between $a$ and $b$ is a $\sqrt{n}$ consistent estimator of $(\mu_T(a,b), \sigma_T^2(a,b))$. Consider the set $J$ of cases with $\text{MED}(n) - k\text{MAD}(n) \leq Y_i \leq \text{MED}(N) + k\text{MAD}(n)$. By Theorem 8.4 sets $\tilde{J}$ and $J$ differ primarily in neighborhoods of $a$ and $b$. This result leads to the following conjecture.

**Conjecture 8.1.** If $Y_1, ..., Y_n$ are iid from a distribution with a pdf that is positive in neighborhoods of $a$ and $b$, and if $\hat{\theta}_n - k_1 D_n \xrightarrow{P} a$ and $\hat{\theta}_n + k_2 D_n \xrightarrow{P} b$ at rate $n^{0.5}$, then $(M_n, S_J^2) \xrightarrow{P} (\mu_T(a,b), \sigma_T^2(a,b))$ with rate $n^{0.5}$.

## 8.2 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. Olive (2017b) shows 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 8.23.** The multivariate location and dispersion model is

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{e}_i, \quad i = 1, \ldots, n \tag{8.33}$$

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

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 $\mathrm{MED}(\boldsymbol{W})$ 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 $p$th 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.

The sample covariance or sample correlation matrices estimate these parameters very efficiently since $\boldsymbol{\Sigma} = (\sigma_{ij})$ where $\sigma_{ij}$ 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 $X_1, ..., X_p$. See

Definition 4.5 for the sample mean $\overline{\boldsymbol{x}}$, the sample covariance matrix $\boldsymbol{S}$, and the sample correlation matrix $\boldsymbol{R}$.

**Rule of thumb 8.1.** 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.

## *8.2.1* 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 $i$th 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 8.24.** The multivariate location and dispersion estimator $(T, \boldsymbol{C})$ is *affine equivariant* if

$$T(\boldsymbol{Z}) = T(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}T(\boldsymbol{W}) + \boldsymbol{b}, \tag{8.34}$$

$$\text{and} \quad \boldsymbol{C}(\boldsymbol{Z}) = \boldsymbol{C}(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}\boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}^T. \tag{8.35}$$

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 $j$th order statistic of the $D_i^2$.

**Theorem 8.19.** If $(T, \boldsymbol{C})$ is affine equivariant, then

$$D_i^2(\boldsymbol{W}) \equiv D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = D_i^2(T(\boldsymbol{Z}), \boldsymbol{C}(\boldsymbol{Z})) \equiv D_i^2(\boldsymbol{Z}). \tag{8.36}$$

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

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

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

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

**Definition 8.25.** 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 $(E(\boldsymbol{x}), \mathrm{Cov}(\boldsymbol{x}))$. 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 8.27.

## *8.2.2* 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 $i$th row $\boldsymbol{x}_i^T$ corresponding to the $i$th 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 $\boldsymbol{W}_d^n$ denote the $n \times p$ data matrix with $i$th 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 8.20.** 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 $\min_{\boldsymbol{a}\neq\boldsymbol{0}} \frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}} = \lambda_p(\boldsymbol{C}^{-1}\boldsymbol{A})$, 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 $\text{rank}(\boldsymbol{C}^{-1}\boldsymbol{A}) = 1$, where $\boldsymbol{C} = \boldsymbol{B}$ and $\boldsymbol{A} = \boldsymbol{d}\boldsymbol{d}^T$, since $\text{rank}(\boldsymbol{C}^{-1}\boldsymbol{A})$ $= \text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{d}) = 1$. 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 8.20, 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 $\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a} = \sum_{i=1}^p \sum_{j=1}^p c_{ij}a_ia_j$, the largest eigenvalue $\lambda_1$ is bounded as $\boldsymbol{W}_d^n$ varies iff $\boldsymbol{C}(\boldsymbol{W}_d^n)$ is bounded as $\boldsymbol{W}_d^n$ varies.

**Definition 8.26.** 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 $\infty$. 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 8.27.** Let $\gamma_n$ 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 $\infty$. 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 $i$th 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 8.21.** 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 $i$th 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 $d_n < d_T$. □

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 8.22.** 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 $\lambda_1$ 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 $\boldsymbol{z}_1, ..., \boldsymbol{z}_{c_n}$. 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 $\lambda_1$ is bounded. □

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

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq D^2_{(c_n)}\} \tag{8.37}$$

where $D^2_{(c_n)}$ is the $c_n$th smallest squared Mahalanobis distance based on $(T, \boldsymbol{C})$. This hyperellipsoid contains the $c_n$ cases with the smallest $D^2_i$. 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).

### *8.2.3* The Concentration Algorithm

Concentration algorithms are widely used since impractical brand name estimators, such as the MCD estimator given in Definition 8.28, take too long to compute. The concentration algorithm, defined in Definition 8.29, 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 8.32.

**Definition 8.28.** 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 Definition 8.10. 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 8.8 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 8.29.** 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 $j$th 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 $j$th *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 8.23: 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 $n^\delta$. 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 $n^\delta$ by Pratt (1959). See Theorem 2.18.

**Theorem 8.24.** 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. $n^\delta$ 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 $i$th attractor if the clean data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are in general position. The breakdown value $\gamma_n$ 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, $\epsilon$, and $\delta_\epsilon$ 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

$P(\min_j \|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = P(\text{all } \|\overline{\boldsymbol{x}}_{-1,\text{j}} - \boldsymbol{\mu}\| > \epsilon) \to \delta_\epsilon^{\text{K}} > 0$ 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 $\infty$.

**Remark 8.5.** Let $\gamma_o$ be the highest percentage of large outliers that an elemental concentration algorithm can detect reliably. For many data sets,

$$\gamma_o \approx \min\left(\frac{n - c_n}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h}\right) 100\% \tag{8.38}$$

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 $i$th 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 $\gamma$ 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 $\gamma$. □

## *8.2.4* 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 8.29 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 8.19. 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 8.30.** 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 8.31.** The *median ball (MB) estimator* $(T_{k,M}, \boldsymbol{C}_{k,M}) = (T_{MB}, \boldsymbol{C}_{MB})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) = (\mathrm{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ as the only start where $\mathrm{MED}(\boldsymbol{W})$ 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 $\mathrm{MED}(\boldsymbol{W})$ in Euclidean distance.

The proof of the following theorem implies that a high breakdown estimator $(T, \boldsymbol{C})$ has $\mathrm{MED}(D_i^2) \leq V$ and that the hyperellipsoid $\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \leq D_{(c_n)}^2\}$ 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 8.25.** 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 $\boldsymbol{x}$,

$$0 < \|\boldsymbol{x}\|^2 \ \tau_p \leq \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \leq \|\boldsymbol{x}\|^2 \ \tau_1. \tag{8.39}$$

Let $\lambda_1 \geq \cdots \geq \lambda_p$ be the eigenvalues of $\boldsymbol{C}$. By (8.39),

$$\frac{1}{\lambda_1}\|\boldsymbol{x} - T\|^2 \leq (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq \frac{1}{\lambda_p}\|\boldsymbol{x} - T\|^2. \tag{8.40}$$

By (8.40), 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 8.22. 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 8.26.** 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 3.5 and 3.6.

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 8.27 is crucial for theory and Theorem 8.28 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 $D^2_i(T, \boldsymbol{C}) \leq h^2$ 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 8.27, 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 $n^\delta$ 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 $n^\delta$ where $a > 0$.

c) Suppose $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate $n^\delta$ 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 $n^\delta$ 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 8.27c) 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 8.2.** 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 $n^\delta$ 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 $n^\delta$ where $a > 0$.

**Remark 8.6.** 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 (8.42), 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 8.19, $(T, \tilde{\boldsymbol{C}})$ 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 $n^\delta$ where $0 < \delta \leq 0.5$, then $D^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$

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

$$= s^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta}). \tag{8.41}$$

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 $100\xi$th percentile of the population squared distances. Then $D_{(c_n)}^2(T, \boldsymbol{C}) \xrightarrow{P} s^{-1} D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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

$$b = D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{8.42}$$

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

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

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

be the population median of the population squared distances. By Remark 8.6, if $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ then $(T, \tilde{\boldsymbol{C}}) \equiv (T, D_{(c_n)}^2(T, \boldsymbol{C})\ \boldsymbol{C})$ 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 $D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C})$). 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 8.28 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 $n^\delta$ as a start results in a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with rate $n^\delta$. 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, $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi_p^2$.

**Theorem 8.28.** Assume that (E1) holds and that $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate $n^\delta$ 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 $n^\delta$.

**Proof.** By Remark 8.6, the estimator is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with rate $n^\delta$. 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 8.23 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 8.32.** 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

$$\boldsymbol{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{8.43}$$

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

**Remark 8.7.** 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 (8.43). 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. See Olive (2017b).

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 $\infty$ if the data is in general position, even if nearly half of the cases are outliers.

**Theorem 8.29.** $T_{FCH}$ 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 8.25 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 8.23 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 8.28 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 (1.62). Then the scaling in (8.43) 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 8.33.** 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 $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH}) \leq \chi^2_{p,0.975}$, 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 $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1) \leq \chi^2_{p,0.975}$, 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 8.34.** 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 $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1)) \leq \chi^2_{p,0.975}$. Let $q_2 = \min\{0.5(0.975)n/n_2, 0.995\}$, and

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

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 $0.5 \leq q < 1$ depends on the elliptically contoured distribution, but $q = 0.5$ and $d = 1$ for multivariate normal data.

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 8.30.** 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). □

The following theorem shows that it is very difficult to drive the determinant of the dispersion estimator from a concentration algorithm to zero.

**Theorem 8.31.** 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). The preface explains how to use the source command to get the `lspack` 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 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.**

### *8.2.5* 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 8.33, 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 8.34, 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 8.35. The RFCH set can be defined similarly.

**Definition 8.35.** Let the $n_2$ cases in Definition 8.34 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 8.8.** 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 $U = \{\boldsymbol{x}_i | D_i^2(T_A, \boldsymbol{C}_A) \leq \chi^2_{p,0.975}\}$. 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 (8.43) 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 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. 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 *lspack* 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).

**Application 8.6.** Outlier resistant regression: Let the $i$th 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 $i = 1, ..., n$. Find the RFCH or RMVN set from the $\boldsymbol{u}_i$, and then run the regression method on the $n_U$ cases $\boldsymbol{w}_i$ corresponding to the set $U$ indices $i_1, ...i_{n_U}$, where $n_U \geq n/2$. Since the response variable was not used to pick the cases, this regression method, conditional on $n_U$ and on the $n_U$ selected cases, has similar large sample theory to the classical regression method that uses all $n$ cases. A similar technique can be used if there is more than one response variable.

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 $\boldsymbol{w}$.

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
```

### *8.2.6* 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 $\text{MED}_j$: 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.

**Definition 8.36.** 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

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

**Example 8.9.** Let the clean data (nonoutliers) be $i$ **1** for $i = 1, 2, 3, 4$, and 5 while the outliers are $j$ **1** for $j = 16, 17, 18$, and 19. Here $n = 9$ and **1** is $p \times 1$. Making a plot of the data for $p = 2$ may be useful. Then the coordinatewise median $\text{MED}_0 = \text{MED}(\boldsymbol{W}) = 5$ **1**. The median Euclidean distance of the data is the Euclidean distance of 5 **1** from 1 **1** = the Euclidean distance of 5 **1** from 9 **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: $\text{MED}_1 = 3$ **1**. Then the median Euclidean distance of the data from $\text{MED}_1$ is the Euclidean distance of 3 **1** from 1 **1** = the Euclidean distance of 3 **1** from 5 **1**. Again the clean cases are the cases with the 5 smallest Euclidean distances. Hence $\text{MED}_j = 3$ **1** for $j \geq 1$. For $j \geq 1$, if $\boldsymbol{x}_i = j$ **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 **1** from 16 **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 $\sqrt{p}$.**

**Application 8.7.** Outlier resistant regression: Let the $i$th 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 $m \geq n/2$.

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 $\text{MED}_j$ instead of a ball that contains half of the cases. The *lspack* 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.

## 8.3 Resistant Multiple Linear Regression

Consider the multiple linear regression model, written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Some good outlier resistant regression estimators are `rmreg2` from Section 8.5, the `hbreg` estimator from Section 8.4, 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 Applications 8.6 and 8.7.

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 $i$th 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 $i$th vector of predictors $\boldsymbol{x}_i$ from the $j$th 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 $\alpha$.) 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 $\gamma$ 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 2.28, 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

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq D_{(j)}^2\}. \tag{8.44}$$

The $i$th 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.)

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 8.37.** 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 8.10.** 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 2.28.

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

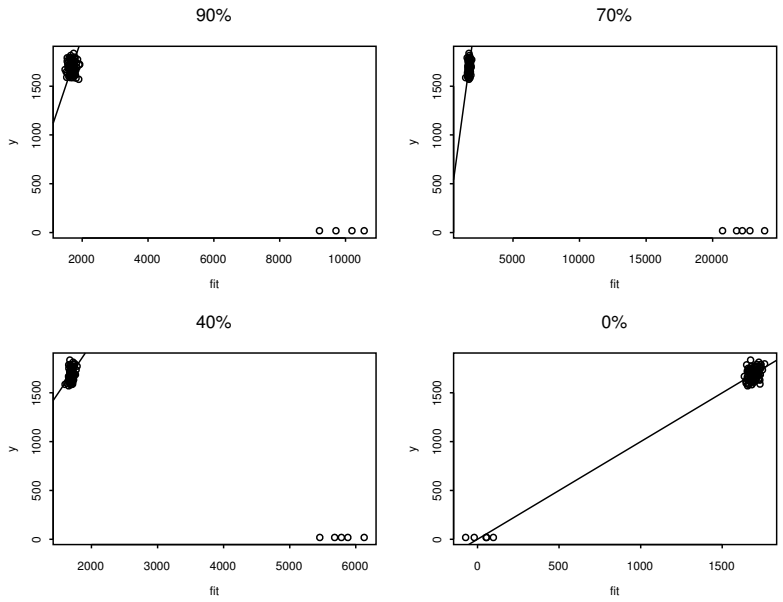


**Fig. 8.1** 4 Trimmed Views for the Buxton Data

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

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.

**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 $\beta_1$ 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. See Olive (2017b, p. 444) for an explanation for why large sample sizes may be needed to estimate the constant.

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 8.38.** 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 $j$th 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 $k$ 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 $j$th *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.

**Remark 8.9.** 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

$$K \approx \frac{\log(\alpha)}{\log([1-(1-\gamma)^p])} \approx \frac{\log(\alpha)}{-(1-\gamma)^p} \tag{8.45}$$

using the approximation $\log(1-x) \approx -x$ for small $x$. Since $\log(0.05) \approx -3$, if $\alpha = 0.05$, then $K \approx \frac{3}{(1-\gamma)^p}$. 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 Olive (2017b: pp. 114-117).

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 8.39. 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 $\gamma$ 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 $\gamma$, $P$(at least one of $K$ subsamples is clean) $= 0.95 \approx$
$1 - [1 - (1-\gamma)^p]^K$. Thus the largest value of $p$ satisfies $\frac{3}{(1-\gamma)^p} \approx K$, or

$$p \approx \left\lfloor \frac{\log(3/K)}{\log(1-\gamma)} \right\rfloor \tag{8.46}$$

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

**Theorem 8.32.** Let $h = p$ be the number of randomly selected cases in an elemental set, and let $\gamma_o$ be the highest percentage of massive outliers that a resampling algorithm can detect reliably. If $n$ is large, then

$$\gamma_o \approx \min\left(\frac{n-c}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h}\right) 100\%. \tag{8.47}$$

**Proof.** As in Remark 8.5, if the contamination proportion $\gamma$ 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 $\gamma$. □

The value of $\gamma_o$ depends on $c \geq n/2$ and $h$. To maximize $\gamma_o$, 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%.

## 8.4 Robust Regression

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

### *8.4.1* 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

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = T(\boldsymbol{X}, \boldsymbol{Y}), \tag{8.48}$$

the vector of predicted values is

$$\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}), \tag{8.49}$$

and the vector of residuals is

$$\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}. \tag{8.50}$$

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

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u}. \tag{8.51}$$

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 $\boldsymbol{X}$.

**Definition 8.40. Scale Equivariance:** Let $c$ be any scalar. Then $\widehat{\boldsymbol{\beta}}$ is scale equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, c\boldsymbol{Y}) = T(\boldsymbol{X}, c\boldsymbol{Y}) = cT(\boldsymbol{X}, \boldsymbol{Y}) = c\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{8.52}$$

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

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = T(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}T(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{8.53}$$

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

**Definition 8.42. Permutation Invariance:** Let $\boldsymbol{P}$ be an $n \times n$ permutation matrix. Then $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}_n$ 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

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) = T(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) = T(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{8.54}$$

Hence if $\boldsymbol{W} = \boldsymbol{P}\boldsymbol{X}$ and $\boldsymbol{Z} = \boldsymbol{P}\boldsymbol{Y}$, then $\widehat{\boldsymbol{Z}} = \boldsymbol{P}\widehat{\boldsymbol{Y}}$ and $\boldsymbol{r}(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) =$ $\boldsymbol{P}\ \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. 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 8.10.** 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 6.2 is used to fit the data with $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. 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 $i$th row corresponds to the $i$th case. For regression, $\boldsymbol{W}$ is the $n \times (p+1)$ matrix with $i$th 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 $\gamma_n$ needed to make $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large.

**Definition 8.43.** Let $1 \leq d_n \leq n$. 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 8.44.** *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 $\leq 0.5$, 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 $c_n$th 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 $\|\hat{\boldsymbol{\beta}}\| \equiv \|\hat{\boldsymbol{\beta}}(\boldsymbol{W}_d^n)\| \leq M$ 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 8.33.** If the breakdown value $\leq 0.5$, 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 8.43.

**Proof.** Consider any contaminated data set $\boldsymbol{W}_d^n$ with $i$th row $(\boldsymbol{w}_i^T, Z_i)^T$. If the regression estimator $T(\boldsymbol{W}_d^n) = \hat{\boldsymbol{\beta}}$ satisfies $\|\hat{\boldsymbol{\beta}}\| \leq M$ 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 $\max_{i=1,\ldots,n} |Y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i| \leq \max_{i=1,\ldots,n}[|Y_i| + \sum_{j=1}^p M|x_{i,j}|]$ 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 $i$th 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 $\infty$. 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 $\infty$. The $\boldsymbol{x}$-outliers can drive $\|\hat{\boldsymbol{\beta}}\|$ to $\infty$ 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 8.45.** 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 8.46.** 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 8.11.** 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 $m$th elemental fit $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$. Then for any elemental fit $\boldsymbol{b}_J$ (suppressing $q = 2$),

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}(\boldsymbol{X}_J\boldsymbol{\beta} + \boldsymbol{e}_J) - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}\boldsymbol{e}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \, \|\boldsymbol{e}_J\|. \quad (8.55)$$

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

$$\|\boldsymbol{X}_J^{-1}\| = \frac{\sigma_1}{\sigma_p \|\boldsymbol{X}_J\|}, \quad (8.56)$$

$$\max_{i,j} |x_{mi,j}| \leq \|\boldsymbol{X}_J\| \leq p \, \max_{i,j} |x_{mi,j}|, \text{ and} \quad (8.57)$$

$$\frac{1}{p \, \max_{i,j} |x_{mi,j}|} \leq \frac{1}{\|\boldsymbol{X}_J\|} \leq \|\boldsymbol{X}_J^{-1}\|. \quad (8.58)$$

*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 8.12.** 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 (8.56), (8.57), and (8.58). 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 $\infty$.* 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 8.13.** 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 $\beta_1$. 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 8.34.** 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 $i$th 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 $L_\infty$ estimator and the LMS criterion.

**Theorem 8.35.** 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 $m$th 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 8.35 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 8.48.

**Definition 8.47.** 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 8.36.** 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 8.34. Since the start is high breakdown, so is the attractor $\hat{\boldsymbol{\beta}}_B$ by Theorem 8.35. 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 8.37.** 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 8.29 and 8.38. Note that an estimator can not be consistent for $\theta$ unless the number of randomly selected cases goes to $\infty$, 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 8.38:** 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 $\infty$ 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. □

## *8.4.2* 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 2022. 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 8.38.

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 $n^\delta$. 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 $n^\delta$ by Pratt (1959). See Theorem 2.18.

**Theorem 8.39.** 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., $n^\delta$ 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 $i$th attractor if the clean data are in general position. The breakdown value $\gamma_n$ 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 $n \to \infty$. □

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 8.40.** 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 $n^\delta$.
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 8.11.** Theorem 8.33 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 8.35.

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 $\infty$. 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.

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 8.47 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 8.48.** 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 8.41.** 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 $\gamma_n$ 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 8.12.** 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.

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.

As implemented in *lspack*, 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.

## 8.5 The Robust rmreg2 Estimator

The robust multivariate linear regression estimator `rmreg2` is the classical multivariate linear regression estimator applied to the RMVN set when RMVN is computed from the vectors $\boldsymbol{u}_i = (x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})^T$ for $i = 1, ..., n$. Hence $\boldsymbol{u}_i$ is the $i$th case with $x_{i1} = 1$ deleted. This regression estimator has considerable outlier resistance, and is one of the most outlier resistant practical robust regression estimator for the $m = 1$ multiple linear regression case. The `rmreg2` estimator has been shown to be consistent if the $\boldsymbol{u}_i$ are iid from a large class of elliptically contoured distributions, which is a much stronger assumption than having iid error vectors $\boldsymbol{\epsilon}_i$.

Let $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$and let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2^T)^T = (\alpha, \boldsymbol{\eta}^T)^T$. Now for multivariate linear regression, $\hat{\boldsymbol{\beta}}_j = (\hat{\alpha}_j, \hat{\boldsymbol{\eta}}_j^T)^T$ where $\hat{\alpha}_j = \overline{Y}_j - \hat{\boldsymbol{\eta}}_j^T \overline{\boldsymbol{u}}$ and $\hat{\boldsymbol{\eta}}_j = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y_j}$. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{uy}} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^T$ which has $j$th column $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y_j}$ for $j = 1, ..., m$. Let

$$\boldsymbol{v} = \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{pmatrix}, \quad E(\boldsymbol{v}) = \boldsymbol{\mu}_{\boldsymbol{v}} = \begin{pmatrix} E(\boldsymbol{u}) \\ E(\boldsymbol{y}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{u}} \\ \boldsymbol{\mu}_{\boldsymbol{y}} \end{pmatrix}, \quad \text{and} \quad \text{Cov}(\boldsymbol{v}) = \boldsymbol{\Sigma}_{\boldsymbol{v}} =$$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{uu}} & \boldsymbol{\Sigma}_{\boldsymbol{uy}} \\ \boldsymbol{\Sigma}_{\boldsymbol{yu}} & \boldsymbol{\Sigma}_{\boldsymbol{yy}} \end{pmatrix}.$$

Let the vector of constants be $\boldsymbol{\alpha}^T = (\alpha_1, ..., \alpha_m)$ and the matrix of slope vectors $\boldsymbol{B}_S = \begin{bmatrix} \boldsymbol{\eta}_1 & \boldsymbol{\eta}_2 & \dots & \boldsymbol{\eta}_m \end{bmatrix}$. Then the population least squares coefficient matrix is

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{\alpha}^T \\ \boldsymbol{B}_S \end{pmatrix}$$

where $\boldsymbol{\alpha} = \boldsymbol{\mu}_{\boldsymbol{y}} - \boldsymbol{B}_S^T \boldsymbol{\mu}_{\boldsymbol{u}}$ and $\boldsymbol{B}_S = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{uy}}$ where $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{\Sigma}_{\boldsymbol{uu}}$.

If the $\boldsymbol{u}_i$ are iid with nonsingular covariance matrix Cov($\boldsymbol{u}$), the least squares estimator

$$\hat{\boldsymbol{B}} = \begin{pmatrix} \hat{\boldsymbol{\alpha}}^T \\ \hat{\boldsymbol{B}}_S \end{pmatrix}$$

where $\hat{\boldsymbol{\alpha}} = \overline{\boldsymbol{y}} - \hat{\boldsymbol{B}}_S^T \overline{\boldsymbol{u}}$ and $\hat{\boldsymbol{B}}_S = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{uy}}$. The least squares multivariate linear regression estimator can be calculated by computing the classical estimator $(\overline{\boldsymbol{v}}, \boldsymbol{S}_{\boldsymbol{v}}) = (\overline{\boldsymbol{v}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}})$ of multivariate location and dispersion on the $\boldsymbol{v}_i$, and then plug in the results into the formulas for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{B}}_S$.

Let $(T, \boldsymbol{C}) = (\tilde{\boldsymbol{\mu}}_{\boldsymbol{v}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}})$ be a robust estimator of multivariate location and dispersion. If $\tilde{\boldsymbol{\mu}}_{\boldsymbol{v}}$ is a consistent estimator of $\boldsymbol{\mu}_{\boldsymbol{v}}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}}$ is a consistent estimator of $c\, \boldsymbol{\Sigma}_{\boldsymbol{v}}$ for some constant $c > 0$, then a robust estimator of multivariate linear regression is the plug in estimator $\tilde{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}} - \tilde{\boldsymbol{B}}_S^T \tilde{\boldsymbol{\mu}}_{\boldsymbol{u}}$ and $\tilde{\boldsymbol{B}}_S = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{uy}}$.

For the `rmreg2` estimator, $(T, \boldsymbol{C})$ is the classical estimator applied to the RMVN set when RMVN is applied to vectors $\boldsymbol{v}_i$ for $i = 1, ..., n$ (could use $(T, \boldsymbol{C})$ = RMVN estimator since the scaling does not matter for this application). Then $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}_{\boldsymbol{v}}, c\, \boldsymbol{\Sigma}_{\boldsymbol{v}})$ if the $\boldsymbol{v}_i$ are iid from a large class of $EC_d(\boldsymbol{\mu}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}, g)$ distributions where $d = m+p-1$. Thus the classical and robust estimators of multivariate linear regression are both $\sqrt{n}$ consistent estimators of $\boldsymbol{B}$ if the $\boldsymbol{v}_i$ are iid from a large class of elliptically contoured distributions. This assumption is very strong, but the robust estimator is useful for detecting outliers. 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. When there are categorical predictors or the joint distribution of $\boldsymbol{v}$ is not elliptically contoured, it is possible that the robust estimator is bad and very different from the good classical least squares estimator. The *lspack* function `rmreg2` computes the `rmreg2` estimator and produces the response and residual plots.

## 8.6 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 $i$th order statistic and the $Y_{(i)}$'s are called the *order statistics*. The *sample median*

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

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

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** $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = (\overline{X}_1, ..., \overline{X}_p)^T$ 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 $i$th *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the $i$th *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 $\text{MED}_j$: 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

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

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.

## 8.7 Complements

**Nearly all of the literature for high breakdown regression and high breakdown multivariate location and dispersion has massive errors**: i) the estimators that have large sample theory tend to be impractical to compute, while ii) estimators that are practical to compute tend to be inconsistent and zero breakdown, or have no proven large sample theory. See Hawkins and Olive (2002). Read Olive (2008, 2017b, 2022c) for practical robust statistics backed by some large sample theory. Sections 8.2 and 8.4 showed that getting large sample theory for practical estimators is very difficult.

**Location Model:** The two stage trimmed means are due to Olive (2001). The confidence interval for the population median appears in Olive (2017b). Huber and Ronchetti (2009) is useful for other estimators.

**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, $\tau$, 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 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 8.30 and 8.38, 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, $\tau$, 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. (2019). 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.

## 8.8 Problems

**PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.**

**8.1.** Use Theorem 2.6 to find the limiting distribution of $\sqrt{n}(\text{MED}(n) - \text{MED}(Y))$.

**8.2.** The interquartile range $\text{IQR}(n) = \hat{\xi}_{n,0.75} - \hat{\xi}_{n,0.25}$ and is a popular estimator of scale. Use Theorem 3.11 to show that

$$\sqrt{n}\frac{1}{2}(IQR(n) - IQR(Y)) \xrightarrow{D} N(0, \sigma_A^2)$$

where

$$\sigma_A^2 = \frac{1}{64}\left[\frac{3}{[f(\xi_{3/4})]^2} - \frac{2}{f(\xi_{3/4})f(\xi_{1/4})} + \frac{3}{[f(\xi_{1/4})]^2}\right].$$

**8.3*.** Let $F$ be the $N(0,1)$ cdf. Show that the ARE of the sample median $\text{MED}(n)$ with respect to the sample mean $\overline{Y}_n$ is $ARE \approx 0.64$.

**8.4*.** Let $F$ be the $DE(0,1)$ cdf. Show that the ARE of the sample median MED$(n)$ with respect to the sample mean $\overline{Y}_n$ is $ARE \approx 2.0$.

**8.5.** If $Y$ is $TEXP(\lambda, b = k\lambda)$ for $k > 0$, show that

$$a) \quad E(Y) = \lambda \left[ 1 - \frac{k}{e^k - 1} \right].$$

$$b) \quad E(Y^2) = 2\lambda^2 \left[ 1 - \frac{(0.5k^2 + k)}{e^k - 1} \right].$$