Chapter 8

Robust Regression Algorithms

Recall from Chapter 7 that high breakdown regression estimators such as LTA, LTS, and LMS are impractical to compute. Hence algorithm estimators are used as approximations. Consider the multiple linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients. Assume that the regression estimator $\hat{\boldsymbol{\beta}}_Q$ is the global minimizer of some criterion $Q(\boldsymbol{b}) \equiv Q(\boldsymbol{b}|\boldsymbol{Y}, \boldsymbol{X})$. In other words, $Q(\hat{\boldsymbol{\beta}}_Q) \leq Q(\boldsymbol{b})$ for all $\boldsymbol{b} \in B \subseteq \Re^p$. Typically $B = \Re^p$, but occasionally B is a smaller set such as the set of OLS fits to $c_n \approx n/2$ of the cases. In this case, B has a huge but finite number $C(n, c_n)$ of vectors $\boldsymbol{b}$. Often Q depends on $\boldsymbol{Y}$ and $\boldsymbol{X}$ only through the residuals $r_i(\boldsymbol{b}) = Y_i - \boldsymbol{x}_i^T\boldsymbol{b}$, but there are exceptions such as the regression depth estimator.

Definition 8.1. In the multiple linear regression setting, an *elemental set* is a set of p cases.

Some notation is needed for algorithms that use many elemental sets. Let

$$J = J_h = \{h_1, ..., h_p\}$$

denote the set of indices for the ith elemental set. Since there are n cases, $h_1, ..., h_p$ are p distinct integers between 1 and n. For example, if $n = 7$ and $p = 3$, the first elemental set may use cases $J_1 = \{1, 7, 4\}$, and the second elemental set may use cases $J_2 = \{5, 3, 6\}$. The data for the ith elemental set is $(\boldsymbol{Y}_{J_h}, \boldsymbol{X}_{J_h})$ where $\boldsymbol{Y}_{J_h} = (Y_{h1}, ..., Y_{hp})^T$ is a $p \times 1$ vector, and the $p \times p$

matrix

$$\boldsymbol{X}_{J_h} = \begin{bmatrix} \boldsymbol{x}_{h1}^T \\ \boldsymbol{x}_{h2}^T \\ \vdots \\ \boldsymbol{x}_{hp}^T \end{bmatrix} = \begin{bmatrix} x_{h1,1} & x_{h1,2} & \cdots & x_{h1,p} \\ x_{h2,1} & x_{h2,2} & \cdots & x_{h2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{hp,1} & x_{hp,2} & \cdots & x_{hp,p} \end{bmatrix}.$$

Definition 8.2. The *elemental fit* from the hth elemental set J_h is

$$\boldsymbol{b}_{J_h} = \boldsymbol{X}_{J_h}^{-1} \boldsymbol{Y}_{J_h}$$

provided that the inverse of $\boldsymbol{X}_{J_h}$ exists.

Definition 8.3. Assume that the p cases in each elemental set are distinct (eg drawn without replacement from the n cases that form the data set). Then the *elemental basic resampling algorithm* for approximating the estimator $\hat{\boldsymbol{\beta}}_Q$ that globally minimizes the criterion $Q(\boldsymbol{b})$ uses K_n elemental sets $J_1, ..., J_{K_n}$ randomly drawn (eg with replacement) from the set of all $C(n,p)$ elemental sets. The *algorithm estimator* $\boldsymbol{b}_A$ is the elemental fit that minimizes Q. That is,

$$\boldsymbol{b}_A = \text{argmin}_{h=1,...,K_n} \;\; Q(\boldsymbol{b}_{J_h}).$$

Several estimators can be found by evaluating all elemental sets. For example, the LTA, L_1, RLTA, LATA, and regression depth estimators can be found this way. Given the criterion Q, the *key parameter* of the basic resampling algorithm is the number K_n of elemental sets used in the algorithm. It is crucial to note that the criterion $Q(\boldsymbol{b})$ is a function of all n cases even though the elemental fit only uses p cases. For example, assume that $K_n = 2$, $J_1 = \{1, 7, 4\}$, $Q(\boldsymbol{b}_{J_1}) = 1.479$, $J_2 = \{5, 3, 6\}$, and $Q(\boldsymbol{b}_{J_2}) = 5.993$. Then $\boldsymbol{b}_A = \boldsymbol{b}_{J_1}$.

To understand elemental fits, the notions of a *matrix norm* and *vector norm* will be useful. We will follow Datta (1995, p. 26-31) and Golub and Van Loan (1989, p. 55-60).

Definition 8.4. The $\boldsymbol{y}$ be an $n \times 1$ vector. Then $\|\boldsymbol{y}\|$ is a *vector norm* if
vn1) $\|\boldsymbol{y}\| \geq 0$ for every $\boldsymbol{y} \in \Re^n$ with equality iff $\boldsymbol{y}$ is the zero vector,
vn2) $\|a\boldsymbol{y}\| = |a| \, \|\boldsymbol{y}\|$ for all $\boldsymbol{y} \in \Re^n$ and for all scalars a, and
vn3) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\Re^n$.

Definition 8.5. 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 8.1. 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*

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

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

8.1 Inconsistency of Resampling Algorithms

We will call algorithms that approximate high breakdown (HB) regression estimators "HB algorithms" although the high breakdown algorithm estimators $\boldsymbol{b}_A$ that have appeared in the literature (that are practical to compute) are typically inconsistent low breakdown estimators. To examine the statistical properties of the basic resampling algorithm, more properties of matrix norms are needed. For the matrix $\boldsymbol{X}_{J_h}$, the subscript h will often be suppressed.

Several useful results involving matrix norms will be used. First, for any subordinate matrix norm,

$$\|\boldsymbol{G}\boldsymbol{y}\|_q \leq \|\boldsymbol{G}\|_q \ \|\boldsymbol{y}\|_q.$$

Hence for any elemental fit $\boldsymbol{b}_J$ (suppressing $q = 2$),

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}(\boldsymbol{X}_J\boldsymbol{\beta} + \boldsymbol{e}_J) - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}\boldsymbol{e}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \ \|\boldsymbol{e}_J\|. \quad (8.1)$$

The following results (Golub and Van Loan 1989, p. 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 $\boldsymbol{X}_J$. Then

$$\|\boldsymbol{X}_J^{-1}\| = \frac{\sigma_1}{\sigma_p \|\boldsymbol{X}_J\|}, \quad (8.2)$$

$$\max_{i,j} |x_{hi,j}| \leq \|\boldsymbol{X}_J\| \leq p \ \max_{i,j} |x_{hi,j}|, \text{ and} \quad (8.3)$$

$$\frac{1}{p \ \max_{i,j} |x_{hi,j}|} \leq \frac{1}{\|\boldsymbol{X}_J\|} \leq \|\boldsymbol{X}_J^{-1}\|. \quad (8.4)$$

The key idea for examining elemental set algorithms is eliminating $\|\boldsymbol{X}_J^{-1}\|$. If there are reasonable conditions under which $\inf \|\boldsymbol{X}_J^{-1}\| > d$ for some constant d that is free of n where the infinum is taken over all $C(n, p)$ elemental sets, then the elemental design matrix $\boldsymbol{X}_J$ will play no role in producing a sequence of consistent elemental fits. We will use the convention that if the inverse $\boldsymbol{X}_J^{-1}$ does not exist, then $\|\boldsymbol{X}_J^{-1}\| = \infty$. The following lemma is crucial.

Lemma 8.1. Assume that the $n \times p$ design matrix $\boldsymbol{X} = [x_{ij}]$ and that the np entries x_{ij} are bounded:

$$\max_{i,j} |x_{ij}| \leq M$$

for some real number $M > 0$ that does not depend on n. Then for any elemental set $\boldsymbol{X}_J$,

$$\|\boldsymbol{X}_J^{-1}\| \geq \frac{1}{pM}. \tag{8.5}$$

Proof. If $\boldsymbol{X}_J$ does not have an inverse, then by the convention $\|\boldsymbol{X}_J^{-1}\| = \infty$, and the result holds. Assume that $\boldsymbol{X}_J$ does have an inverse. Then by Equation (8.4),

$$\frac{1}{pM} \leq \frac{1}{p \ \max_{i,j} |x_{hi,j}|} \leq \frac{1}{\|\boldsymbol{X}_J\|} \leq \|\boldsymbol{X}_J^{-1}\|.$$

QED

In proving consistency results, there is an infinite sequence of estimators that depend on the sample size n. Hence the subscript n will be added to the estimators. Refer to Remark 2.4 for the definition of convergence in probability.

Definition 8.6. Lehmann (1999, p. 53-54): a) A sequence of random variables W_n is *tight* or *bounded in probability,* written $W_n = O_P(1)$, if for every $\epsilon > 0$ there exist positive constants D_ϵ and N_ϵ such that

$$P(|W_n| \leq D_\epsilon) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$. Also $W_n = O_P(X_n)$ if $|W_n/X_n| = O_P(1)$.

b) The sequence $W_n = o_P(n^{-\delta})$ if $n^\delta W_n = o_P(1)$ which means that

$$n^\delta W_n \xrightarrow{P} 0.$$

c) W_n has the same order as X_n in probability, written $W_n \asymp_P X_n$, if for every $\epsilon > 0$ there exist positive constants N_ϵ and $0 < d_\epsilon < D_\epsilon$ such that

$$P(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon) = P(\frac{1}{D_\epsilon} \leq \left|\frac{X_n}{W_n}\right| \leq \frac{1}{d_\epsilon}) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$.

d) Similar notation is used for a $k \times r$ matrix $\boldsymbol{A} = [a_{i,j}]$ if each element $a_{i,j}$ has the desired property. For example, $\boldsymbol{A} = O_P(n^{-1/2})$ if each $a_{i,j} = O_P(n^{-1/2})$.

Definition 8.7. Let $W_n = \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|$.

a) If $W_n \asymp_P n^{-\delta}$ for some $\delta > 0$, then both W_n and $\hat{\boldsymbol{\beta}}_n$ have (tightness) **rate** n^{δ}.

b) If there exists a constant κ such that

$$n^{\delta}(W_n - \kappa) \xrightarrow{D} X$$

for some nondegenerate random variable X, then both W_n and $\hat{\boldsymbol{\beta}}_n$ have *convergence rate* n^{δ}.

If W_n has convergence rate n^{δ}, then W_n has tightness rate n^{δ}, and the term "tightness" will often be omitted. Notice that if $W_n \asymp_P X_n$, then $X_n \asymp_P W_n$, $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$. Notice that if $W_n = O_P(n^{-\delta})$, then n^{δ} is a lower bound on the rate of W_n. As an example, if LMS, OLS or L_1 are used for $\hat{\boldsymbol{\beta}}$, then $W_n = O_P(n^{-1/3})$, but $W_n \asymp_P n^{-1/3}$ for LMS while $W_n \asymp_P n^{-1/2}$ for OLS and L_1. Hence the rate for OLS and L_1 is $n^{1/2}$.

To examine the lack of consistency of the basic resampling algorithm estimator $\boldsymbol{b}_{A,n}$ meant to approximate the theoretical estimator $\hat{\boldsymbol{\beta}}_{Q,n}$, recall that the key parameter of the basic resampling algorithm is the number of elemental sets $K_n \equiv K(n,p)$. Typically K_n is a fixed number, eg $K_n \equiv K = 3000$, that does not depend on n.

Example 8.2. This example illustrates the basic resampling algorithm with $K_n = 2$. Let the data consist of the five (x_i, y_i) pairs (0,1), (1,2), (2,3), (3,4), and (1,11). Then $p = 2$ and $n = 5$. Suppose the criterion Q is the median of the n squared residuals and that $J_1 = \{1, 5\}$. Then observations $(0, 1)$ and $(1, 11)$ were selected. Since $\boldsymbol{b}_{J_1} = (1, 10)^T$, the estimated line is $y = 1 + 10x$, and the corresponding residuals are $0, -9, -18, -27$, and 0. The criterion $Q(\boldsymbol{b}_{J_1}) = 9^2 = 81$ since the ordered squared residuals are $0, 0, 81, 18^2$, and 27^2. If observations $(0, 1)$ and $(3, 4)$ are selected next, then $J_2 = \{1, 4\}$, $\boldsymbol{b}_{J_2} = (1, 1)^T$, and 4 of the residuals are zero. Thus $Q(\boldsymbol{b}_{J_2}) = 0$ and $\boldsymbol{b}_A = \boldsymbol{b}_{J_2} = (1, 1)^T$. Hence the algorithm produces the fit $y = 1 + x$.

Example 8.3. In the previous example the algorithm fit was reasonable, but in general using a fixed $K_n \equiv K$ in the algorithm produces inconsistent estimators. To illustrate this claim, consider the location model $Y_i = \beta + e_i$ where the e_i are iid and β is a scalar (since $p = 1$ in the location model). If β was known, the natural criterion for an estimator b_n of β would be $Q(b_n) = |b_n - \beta|$. For each sample size n, K elemental sets $J_{h,n} = \{h_n\}, h = 1, ..., K$

of size $p = 1$ are drawn with replacement from the integers $1, ..., n$. Denote the resulting elemental fits by

$$b_{J_{h,n}} = Y_{hn}$$

for $h = 1, ..., K$. Then the "best fit" $Y_{o,n}$ minimizes $|Y_{hn} - \beta|$. If $\alpha > 0$, then

$$P(|Y_{o,n} - \beta| > \alpha) \geq [P(|Y_1 - \beta| > \alpha)]^K > 0$$

provided that the errors have mass outside of $[-\alpha, \alpha]$, and thus $Y_{o,n}$ is not a consistent estimator. The inequality is needed since the Y_{hn} may not be distinct: the inequality could be replaced with equality if the $Y_{1n}, ..., Y_{Kn}$ were an iid sample of size K. Since $\alpha > 0$ was arbitrary in the above example, the inconsistency result holds unless the iid errors are degenerate at zero.

The basic idea is from sampling theory. A fixed finite sample can be used to produce an estimator that contains useful information about a population parameter, eg the population mean, but unless the sample size n increases to ∞, the confidence interval for the population parameter will have a length bounded away from zero. In particular, if $\overline{Y}_n(K)$ is a sequence of sample means based on samples of size $K = 100$, then $\overline{Y}_n(K)$ is not a consistent estimator for the population mean.

The following notation is useful for the general regression setting and will also be used for some algorithms that modify the basic resampling algorithm. Let $\boldsymbol{b}_{si,n}$ be the ith elemental fit where $i = 1, ..., K_n$ and let $\boldsymbol{b}_{A,n}$ be the algorithm estimator; that is, $\boldsymbol{b}_{A,n}$ is equal to the $\boldsymbol{b}_{si,n}$ that minimized the criterion Q. Let $\hat{\boldsymbol{\beta}}_{Q,n}$ denote the estimator that the algorithm is approximating, eg $\hat{\boldsymbol{\beta}}_{LTA,n}$. Let $\boldsymbol{b}_{os,n}$ be the "best" of the K elemental fits in that

$$\boldsymbol{b}_{os,n} = \text{argmin}_{i=1,...,K_n} \|\boldsymbol{b}_{si,n} - \boldsymbol{\beta}\| \tag{8.6}$$

where the Euclidean norm is used. Since the algorithm estimator is an elemental fit $\boldsymbol{b}_{si,n}$,

$$\|\boldsymbol{b}_{A,n} - \boldsymbol{\beta}\| \geq \|\boldsymbol{b}_{os,n} - \boldsymbol{\beta}\|.$$

Thus an upper bound on the rate of $\boldsymbol{b}_{os,n}$ is an upper bound on the rate of $\boldsymbol{b}_{A,n}$.

Theorem 8.2. Let the number of *randomly selected elemental sets* $K_n \to \infty$ as $n \to \infty$. Assume that the error distribution possesses a density

f that is positive and continuous in a neighborhood of zero and that $K_n \leq C(n,p)$. Also assume that the predictors are bounded in probability and that the iid errors are independent of the predictors. Then an upper bound on the rate of $\boldsymbol{b}_{os,n}$ is $K_n^{1/p}$.

Proof. Let $J = \{i_1, ..., i_p\}$ be a randomly selected elemental set. Then $\boldsymbol{Y}_J = \boldsymbol{X}_J\boldsymbol{\beta} + \boldsymbol{e}_J$ where the p errors are independent, and the data $(\boldsymbol{Y}_J, \boldsymbol{X}_J)$ produce an estimator

$$\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$$

of $\boldsymbol{\beta}$. Let $0 < \delta \leq 1$. If each observation in J has an absolute error bounded by M/n^δ, then

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}\boldsymbol{e}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \frac{M\sqrt{p}}{n^\delta}.$$

Lemma 8.1 shows that the norm $\|\boldsymbol{X}_J^{-1}\|$ is bounded away from 0 provided that the predictors are bounded. Thus if the predictors are bounded in probability, then $\|\boldsymbol{b}_J - \boldsymbol{\beta}\|$ is small only if all p errors in $\boldsymbol{e}_J$ are small. Now

$$P_n \equiv P(|e_i| < \frac{M}{n^\delta}) \approx \frac{2\ M\ f(0)}{n^\delta} \tag{8.7}$$

for large n. Note that if W counts the number of errors satisfying (8.7) then $W \sim$ binomial(n, P_n), and the probability that all p errors in $\boldsymbol{e}_J$ satisfy Equation (8.7) is proportional to $1/n^{\delta p}$. If $K_n = o(n^{\delta p})$ elemental sets are used, then the probability that the best elemental fit $\boldsymbol{b}_{os,n}$ satisfies

$$\|\boldsymbol{b}_{os,n} - \boldsymbol{\beta}\| \leq \frac{M_\epsilon}{n^\delta}$$

tends to zero regardless of the value of the constant $M_\epsilon > 0$. Replace n^δ by $K_n^{1/p}$ for the more general result. QED

Remark 8.1. It is crucial that the elemental sets were chosen *randomly*. For example the cases within any elemental set could be chosen without replacement, and then the K_n elemental sets could be chosen with replacement. Alternatively, random permutations of the integers 1, ..., n could be selected with replacement. Each permutation generates approximately n/p elemental sets: the jth set consists of the cases $(j-1)p + 1, ..., jp$. Alternatively $g(n)$ cases could be selected without replacement and then all

$$K_n = C(g(n), p) = \binom{g(n)}{p}$$

elemental sets generated. As an example where the elemental sets are not chosen randomly, consider the L_1 criterion. Since there is always an elemental L_1 fit, this fit has $n^{1/2}$ convergence rate and is a consistent estimator of $\boldsymbol{\beta}$. Here we can take $K_n \equiv 1$, but the elemental set was not drawn randomly. Using brain power to pick elemental sets is frequently a good idea.

It is also crucial to note that the $K_n^{1/p}$ rate is only an upper bound on the rate of the algorithm estimator $\boldsymbol{b}_{A,n}$. It is possible that the best elemental set has a good convergence rate while the basic resampling algorithm estimator is inconsistent. Notice that the following result holds regardless of the criterion used.

Theorem 8.3. If the number $K_n \equiv K$ of randomly selected elemental sets is fixed and free of the sample size n, eg $K = 3000$, then the algorithm estimator $\boldsymbol{b}_{A,n}$ is an inconsistent estimator of $\boldsymbol{\beta}$.

Proof. Each of the K elemental fits is an inconsistent estimator. So regardless of how the algorithm chooses the final elemental fit, the algorithm estimator is inconsistent.

Conjecture 8.1. Suppose that the errors possess a density that is positive and continuous on the real line, that $\|\hat{\boldsymbol{\beta}}_{Q,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ and that $K_n \leq C(n,p)$ randomly selected elemental sets are used in the algorithm. Then the algorithm estimator satisfies $\|\boldsymbol{b}_{A,n} - \boldsymbol{\beta}\| = O_P(K_n^{-1/2p})$.

Remark 8.2. This rate can be achieved if the algorithm minimizing Q over all elemental subsets is $\sqrt{n}$ consistent (eg regression depth, see Bai and He 1999). Randomly select $g(n)$ cases and let $K_n = C(g(n),p)$. Then apply the all elemental subset algorithm to the $g(n)$ cases. Notice that an upper bound on the rate of $\boldsymbol{b}_{os,n}$ is $g(n)$ while

$$\|\boldsymbol{b}_{A,n} - \boldsymbol{\beta}\| = O_P((g(n))^{-1/2}).$$

8.2 Theory for Concentration Algorithms

Newer HB algorithms use random elemental sets to generate starting trial fits, but then refine them. One of the most successful subset refinement algorithms is the *concentration algorithm*. Consider the LTA, LTS and LMS criterion that cover $c \equiv c_n \geq n/2$ cases.

Definition 8.8. A *start* is an initial trial fit and an *attractor* is the final fit generated by the algorithm from the start. In a *concentration algorithm*, let $\boldsymbol{b}_{0,j}$ be the jth start 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, a classical estimator $\boldsymbol{b}_{1,j}$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest squared residuals. This iteration can be continued for k steps resulting in the sequence of estimators $\boldsymbol{b}_{0,j}, \boldsymbol{b}_{1,j}, ..., \boldsymbol{b}_{k,j}$. The result of the iteration $\boldsymbol{b}_{k,j}$ is called the jth attractor. The final concentration algorithm estimator is the attractor that optimizes the criterion.

Sometimes the notation $\boldsymbol{b}_{si,n} = \boldsymbol{b}_{0i,n}$ for the ith start and $\boldsymbol{b}_{ai,n} = \boldsymbol{b}_{ki,n}$ for the ith attractor will be used. Using $k = 10$ concentration steps often works well, and iterating until convergence is usually fast (in this case $k = k_i$ depends on i). The "h–set" basic resampling algorithm uses starts that are fits to randomly selected sets of $h \geq p$ cases, and is a special case of the concentration algorithm with $k = 0$.

The notation CLTS, CLMS and CLTA will be used to denote concentration algorithms for LTA, LTS and LMS, respectively. Consider the LTS(c_n) criterion. Suppose the ordered squared residuals from the mth start $\boldsymbol{b}_{0m}$ are obtained. Then $\boldsymbol{b}_{1m}$ is simply the OLS fit to the cases corresponding to the c_n smallest squared residuals. Denote these cases by $i_1, ..., i_{c_n}$. Then

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

where the second inequality follows from the definition of the OLS estimator. Convergence to the attractor tends to occur in a few steps.

A simplified version of the CLTS(c) algorithms of Ruppert (1992), Víšek (1996), Hawkins and Olive (1999a) and Rousseeuw and Van Driessen (2000, 2002, 2006) uses K_n elemental starts. The LTS(c) criterion is

$$Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c} r^2_{(i)}(\boldsymbol{b}) \tag{8.8}$$

where $r^2_{(i)}(\boldsymbol{b})$ is the ith smallest squared residual. For each elemental start find the exact-fit $\boldsymbol{b}_{sj}$ to the p cases in the elemental start and then get the c smallest squared residuals. Find the OLS fit to these c cases and find the resulting c smallest squared residuals, and iterate for k steps. Doing this

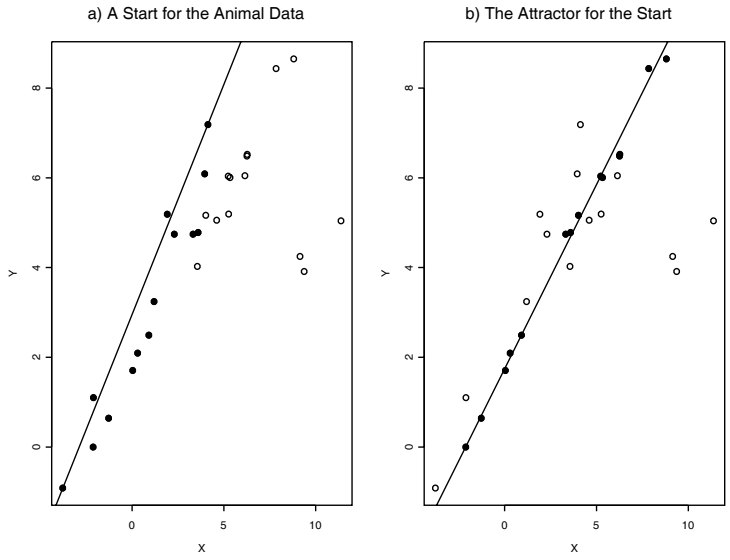

Figure 8.1: The Highlighted Points are More Concentrated about the Attractor

for K_n elemental starts leads to K_n (not necessarily distinct) attractors $\boldsymbol{b}_{aj}$. The algorithm estimator $\hat{\boldsymbol{\beta}}_{ALTS}$ is the attractor that minimizes Q. Substituting the L_1 or Chebyshev fits and LTA or LMS criteria for OLS in the concentration step leads to the CLTA or CLMS algorithm.

Example 8.4. As an illustration of the CLTA concentration algorithm, consider the animal data from Rousseeuw and Leroy (1987, p. 57). The response y is the *log brain weight* and the predictor x is the *log body weight* for 25 mammals and 3 dinosaurs (outliers with the highest body weight). Suppose that the first elemental start uses cases 20 and 14, corresponding to mouse and man. Then the start $\boldsymbol{b}_{s,1} = \boldsymbol{b}_{0,1} = (2.952, 1.025)^T$ and the sum of the $c = 14$ smallest absolute residuals

$$\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{0,1}) = 12.101.$$

Figure 8.1a shows the scatterplot of x and y. The start is also shown and the 14 cases corresponding to the smallest absolute residuals are highlighted.

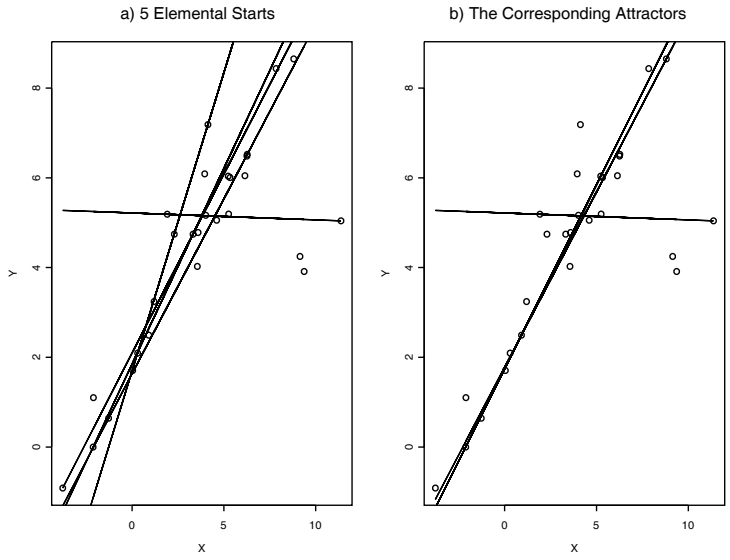

Figure 8.2: Starts and Attractors for the Animal Data

The L_1 fit to these c highlighted cases is $\boldsymbol{b}_{1,1} = (2.076, 0.979)^T$ and

$$\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{1,1}) = 6.990.$$

The iteration consists of finding the cases corresponding to the c smallest residuals, obtaining the corresponding L_1 fit and repeating. The attractor $\boldsymbol{b}_{a,1} = \boldsymbol{b}_{7,1} = (1.741, 0.821)^T$ and the LTA(c) criterion evaluated at the attractor is

$$\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{a,1}) = 2.172.$$

Figure 8.1b shows the attractor and that the c highlighted cases corresponding to the smallest absolute residuals are much more concentrated than those in Figure 8.1a. Figure 8.2a shows 5 randomly selected starts while Figure 8.2b shows the corresponding attractors. Notice that the elemental starts have more variablity than the attractors, but if the start passes through an outlier, so does the attractor.

Notation for the attractor needs to be added to the notation used for the basic resampling algorithm. Let $\boldsymbol{b}_{si,n}$ be the ith start, and let $\boldsymbol{b}_{ai,n}$ be the

ith attractor. Let $\boldsymbol{b}_{A,n}$ be the algorithm estimator, that is, the attractor that minimized the criterion Q. Let $\hat{\boldsymbol{\beta}}_{Q,n}$ denote the estimator that the algorithm is approximating, eg $\hat{\boldsymbol{\beta}}_{LTS,n}$. Let $\boldsymbol{b}_{os,n}$ be the "best" start in that

$$\boldsymbol{b}_{os,n} = \text{argmin}_{i=1,...,K_n} \|\boldsymbol{b}_{si,n} - \boldsymbol{\beta}\|.$$

Similarly, let $\boldsymbol{b}_{oa,n}$ be the best attractor. Since the algorithm estimator is an attractor, $\|\boldsymbol{b}_{A,n} - \boldsymbol{\beta}\| \geq \|\boldsymbol{b}_{oa,n} - \boldsymbol{\beta}\|$, and an upper bound on the rate of $\boldsymbol{b}_{oa,n}$ is an upper bound on the rate of $\boldsymbol{b}_{A,n}$.

Typically the algorithm will use randomly selected elemental starts, but more generally the start could use (eg OLS or L_1) fits computed from h_i cases. Many algorithms will use the same number $h_i \equiv h$ of cases for all starts. If $\boldsymbol{b}_{si,n}, \boldsymbol{b}_{1i,n}, ..., \boldsymbol{b}_{ai,n}$ is the sequence of fits in the iteration from the ith start to the ith attractor, typically c_n cases will be used after the residuals from the start are obtained. However, for LATx algorithms, the jth fit $\boldsymbol{b}_{ji,n}$ in the iteration uses $C_n(\boldsymbol{b}_{j-1,i,n})$ cases where $C_n(\boldsymbol{b})$ is given by Equation (7.5) on p. 230. Since the criterion is evaluated on the attractors, using OLS as an attractor also makes sense.

Remark 8.3. *Failure of zero-one weighting.* Assume that the iteration from start to attractor is bounded by the use of a stopping rule. In other words, $ai, n \leq M$ for some constant M and for all $i = 1, ..., K_n$ and for all n. Then the consistency rate of the best attractor is equal to the rate for the best start for the LTS concentration algorithm if all of the start sizes h_i are bounded (eg if all starts are elemental). For example, suppose the concentration algorithm for LTS uses elemental starts, and OLS is used in each concentration step. If the best start satisfies $\|\boldsymbol{b}_{os,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ then the best attractor satisfies $\|\boldsymbol{b}_{oa,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$. *In particular, if the number of starts $K_n \equiv K$ is a fixed constant (free of the sample size n) and all K of the start sizes are bounded by a fixed constant (eg p), then the algorithm estimator $\boldsymbol{b}_{A,n}$ is inconsistent.*

This result holds because zero-one weighting fails to improve the consistency rate. That is, suppose an initial fit $\hat{\boldsymbol{\beta}}_n$ satisfies $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ where $0 < \delta \leq 0.5$. If $\hat{\boldsymbol{\beta}}_{cn}$ denotes the OLS fit to the $c \approx n/2$ cases with the smallest absolute residuals, then

$$\|\hat{\boldsymbol{\beta}}_{cn} - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{8.9}$$

See Ruppert and Carroll (1980, p. 834 for $\delta = 0.5$), Dollinger and Staudte (1991, p. 714), He and Portnoy (1992) and Welsh and Ronchetti (1993).

These results hold for a wide variety of zero-one weighting techniques. Concentration uses the cases with the smallest c absolute residuals, and the popular "reweighting for efficiency" technique applies OLS to cases that have absolute residuals smaller than some constant. He and Portnoy (1992, p. 2161) note that such an attempt to get a rate $n^{1/2}$ estimator from the rate $n^{1/3}$ initial LMS fit does not in fact improve LMS's rate.

Remark 8.4. While the formal proofs in the literature cover OLS fitting, it is a reasonable conjecture that the result also holds if alternative fits such as L_1 are used in the concentration steps. Heuristically, zero-one weighting from the initial estimator results in a data set with the same "tilt" as the initial estimator, and applying a $\sqrt{n}$ consistent estimator to the cases with the c smallest case distances can not get rid of this tilt.

Remarks 8.3 and 8.4 suggest that the consistency rate of the algorithm estimator is bounded above by the rate of the best elemental start. Theorem 8.2 and the following remark show that the number of random starts is the determinant of the actual performance of the estimator, as opposed to the theoretical convergence rate of $\hat{\boldsymbol{\beta}}_{Q,n}$. Suppose $K_n = O(n)$ starts are used. Then the rate of the algorithm estimator is no better than $n^{1/p}$ which drops dramatically as the dimensionality increases.

Remark 8.5: The wide spread of subsample slopes. Some additional insights into the size h of the start come from a closer analysis of an idealized case – that of normally distributed predictors. Assume that the errors are iid $N(0,1)$ and that the $\boldsymbol{x}_i$'s are iid $N_p(\boldsymbol{0}, \boldsymbol{I})$. Use h observations $(\boldsymbol{X}_h, \boldsymbol{Y}_h)$ to obtain the OLS fit

$$\boldsymbol{b} = (\boldsymbol{X}_h^T \boldsymbol{X}_h)^{-1} \boldsymbol{X}_h^T \boldsymbol{Y}_h \sim N_p(\boldsymbol{\beta}, (\boldsymbol{X}_h^T \boldsymbol{X}_h)^{-1}).$$

Then $\|\boldsymbol{b} - \boldsymbol{\beta}\|^2 = (\boldsymbol{b} - \boldsymbol{\beta})^T (\boldsymbol{b} - \boldsymbol{\beta})$ is distributed as $(p\ F_{p,h-p+1})/(h-p+1)$.

Proof (due to Morris L. Eaton). Let $V = \boldsymbol{X}_h^T \boldsymbol{X}_h$. Then V has the Wishart distribution $W(\boldsymbol{I}_p, p, h)$ while V^{-1} has the inverse Wishart distribution $W^{-1}(\boldsymbol{I}_p, p, h+p-1)$. Without loss of generality, assume $\boldsymbol{\beta} = \boldsymbol{0}$. Let $W \sim W(\boldsymbol{I}_p, p, h)$ and $\hat{\boldsymbol{\beta}}|W \sim N_p(\boldsymbol{0}, W^{-1})$. Then the characteristic function of $\hat{\boldsymbol{\beta}}$ is

$$\phi(\boldsymbol{t}) = E(E[\exp(i\boldsymbol{t}^T \hat{\boldsymbol{\beta}})|W]) = E_W[\exp(-\frac{1}{2}\boldsymbol{t}^T W^{-1} \boldsymbol{t})].$$

Let $\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ and $S \sim W(\boldsymbol{I}_p, p, h)$ be independent. Let $\boldsymbol{Y} = S^{-1/2}\boldsymbol{X}$.

Then the characteristic function of $\boldsymbol{Y}$ is

$$\psi(\boldsymbol{t}) = E(E[\exp(i(S^{-1/2}\boldsymbol{t})^T\boldsymbol{X})|S]) = E_S[\exp(-\frac{1}{2}\boldsymbol{t}^T S^{-1}\boldsymbol{t})].$$

Since $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{Y}$ have the same characteristic functions, they have the same distribution. Thus $\|\hat{\boldsymbol{\beta}}\|^2$ has the same distribution as

$$\boldsymbol{X}^T S^{-1}\boldsymbol{X} \sim (p/(h-p+1))\ F_{p,h-p+1}.$$

QED

This result shows the inadequacy of elemental sets in high dimensions. For a trial fit to provide a useful preliminary classification of cases into inliers and outliers requires that it give a reasonably precise slope. However if p is large, this is most unlikely; the density of $(\boldsymbol{b}-\boldsymbol{\beta})^T(\boldsymbol{b}-\boldsymbol{\beta})$ varies near zero like $[(\boldsymbol{b}-\boldsymbol{\beta})^T(\boldsymbol{b}-\boldsymbol{\beta})]^{(\frac{p}{2}-1)}$. For moderate to large p, this implies that good trial slopes will be extremely uncommon and so enormous numbers of random elemental sets will have to be generated to have some chance of finding one that gives a usefully precise slope estimate. The only way to mitigate this effect of basic resampling is to use larger values of h, but this negates the main virtue elemental sets have, which is that when outliers are present, the smaller the h the greater the chance that the random subset will be clean.

The following two propositions examine increasing the start size. The first result (compare Remark 8.3) proves that increasing the start size from elemental to $h \geq p$ results in a zero breakdown inconsistent estimator. Let the k–step CLTS estimator be the concentration algorithm estimator for LTS that uses k concentration steps. Assume that the number of concentration steps k and the number of starts $K_n \equiv K$ do not depend on n (eg $k = 10$ and $K = 3000$, breakdown is defined in Section 9.4).

Proposition 8.4. Suppose that each start uses h randomly selected cases and that $K_n \equiv K$ starts are used. Then
i) the ("h-set") basic resampling estimator is inconsistent.
ii) The k–step CLTS estimator is inconsistent.
iii) The breakdown value is bounded above by K/n.

Proof. To prove i) and ii), notice that each start is inconsistent. Hence each attractor is inconsistent by He and Portnoy (1992). Choosing from K inconsistent estimators still results in an inconsistent estimator. To prove iii)

replace one observation in each start by a high leverage case (with y tending to ∞). QED

Suppose that $\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_K$ are consistent estimators of $\boldsymbol{\beta}$ each with the same rate $g(n)$. The lemma below shows that 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 $g(n)$.

Lemma 8.5: Pratt (1959). a) Let $X_{1,n}, ..., X_{K,n}$ each be $O_P(1)$ where K is fixed. Suppose $W_n = X_{i_n,n}$ for some $i_n \in \{1, ..., K\}$. Then

$$W_n = O_P(1). \tag{8.10}$$

b) Suppose $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ for $j = 1, ..., K$ where $0 < \delta \leq 1$. Let $T_n^* = T_{i_n,n}$ for some $i_n \in \{1, ..., K\}$ where, for example, $T_{i_n,n}$ is the $T_{j,n}$ that minimized some criterion function. Then

$$\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{8.11}$$

Proof. a) $P(\max\{X_{1,n}, ..., X_{K,n}\} \leq x) = P(X_{1,n} \leq x, ..., X_{K,n} \leq x) \leq$

$$F_{W_n}(x) \leq P(\min\{X_{1,n}, ..., X_{K,n}\} \leq x) = 1 - P(X_{1,n} > x, ..., X_{K,n} > x).$$

Since K is finite, there exists $B > 0$ and N such that $P(X_{i,n} \leq B) > 1 - \epsilon/2K$ and $P(X_{i,n} > -B) > 1 - \epsilon/2K$ for all $n > N$ and $i = 1, ..., K$. Bonferroni's inequality states that $P(\cap_{i=1}^K A_i) \geq \sum_{i=1}^K P(A_i) \ - \ (K-1)$. Thus

$$F_{W_n}(B) \geq P(X_{1,n} \leq B, ..., X_{K,n} \leq B) \geq$$

$$K(1 - \epsilon/2K) - (K-1) = K - \epsilon/2 - K + 1 = 1 - \epsilon/2$$

and

$$-F_{W_n}(-B) \geq -1 + P(X_{1,n} > -B, ..., X_{K,n} > -B) \geq$$

$$-1 + K(1 - \epsilon/2K) - (K-1) = -1 + K - \epsilon/2 - K + 1 = -\epsilon/2.$$

Hence

$$F_{W_n}(B) - F_{W_n}(-B) \geq 1 - \epsilon \ \text{ for } \ \text{n} > \text{N}.$$

b) Use with $X_{j,n} = n^{\delta}\|T_{j,n} - \boldsymbol{\beta}\|$. Then $X_{j,n} = O_P(1)$ so by a), $n^{\delta}\|T_n^* - \boldsymbol{\beta}\| = O_P(1)$. Hence $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$. QED

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 several 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.

Proposition 8.6. Suppose $K_n \equiv K$ starts are used and that all starts have subset size $h_n = g(n) \uparrow \infty$ as $n \to \infty$. Assume that the estimator applied to the subset has rate n^{δ}.
i) For the h_n-set basic resampling algorithm, the algorithm estimator has rate $[g(n)]^{\delta}$.
ii) Under mild regularity conditions (eg 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). QED

These results show that fixed $K_n \equiv K$ elemental methods are inconsistent. Several simulation studies have shown that the versions of the resampling algorithm that use a fixed number of elemental starts provide fits with behavior that conforms with the asymptotic behavior of the $\sqrt{n}$ consistent target estimator. These paradoxical studies can be explained by the following proposition (a recasting of a coupon collection problem).

Proposition 8.7. Suppose that $K_n \equiv K$ random starts of size h are selected and let $Q_{(1)} \leq Q_{(2)} \leq \cdots \leq Q_{(B)}$ correspond to the order statistics of the criterion values of the $B = C(n,h)$ possible starts of size h. Let R be the rank of the smallest criterion value from the K starts. If $P(R \leq R_\alpha) = \alpha$, then

$$R_\alpha \approx B[1 - (1-\alpha)^{1/K}].$$

Proof. If W_i is the rank of the ith start, then $W_1, ..., W_K$ are iid discrete uniform on $\{1, ..., B\}$ and $R = \min(W_1, ..., W_K)$. If r is an integer in $[1, B]$, then

$$P(R \leq r) = 1 - (\frac{B-r}{B})^K.$$

Solve the above equation $\alpha = P(R \leq R_\alpha)$ for R_α. QED

Remark 8.6. If $K = 500$, then with $\alpha = 50\%$ probability about 14 in 10000 elemental sets will be better than the best elemental start found from the elemental concentration algorithm. From Feller (1957, p. 211-212),

$$E(R) \approx 1 + \frac{B}{K+1}, \text{ and VAR(R)} \approx \frac{\mathrm{KB}^2}{(\mathrm{K}+1)^2(\mathrm{K}+2)} \approx \frac{\mathrm{B}^2}{\mathrm{K}^2}.$$

Notice that the median of R is $\mathrm{MED}(R) \approx B[1 - (0.5)^{1/K}]$.

Thus simulation studies that use very small generated data sets, so the probability of finding a good approximation is high, are quite misleading about the performance of the algorithm on more realistically sized data sets. For example, if $n = 100$, $h = p = 3$, and $K = 3000$, then $B = 161700$ and the median rank is about 37. Hence the probability is about 0.5 that only 36 elemental subsets will give a smaller value of Q than the fit chosen by the algorithm, and so using just 3000 starts may well suffice. This is not the case with larger values of p.

If the algorithm evaluates the criterion on trial fits, then these fits will be called the attractors. The following theorem shows that it is simple to improve the CLTS estimator by adding two carefully chosen attractors. Notice that `lmsreg` is an inconsistent zero breakdown estimator but the modification to `lmsreg` is HB and asymptotically equivalent to OLS. Hence the modified estimator has a $\sqrt{n}$ rate which is higher than the $n^{1/3}$ rate of the LMS estimator. Let $\boldsymbol{b}_k$ be the attractor from the start consisting of OLS applied to the c_n cases with Y's closest to the median of the Y_i and let $\hat{\boldsymbol{\beta}}_{k,B} = 0.99\boldsymbol{b}_k$. Then $\hat{\boldsymbol{\beta}}_{k,B}$ is a HB biased estimator of $\boldsymbol{\beta}$. (See Example 9.3. An estimator is HB if its median absolute residual stays bounded even if nearly half of the cases are outliers.)

Theorem 8.8. Suppose that the algorithm uses $K_n \equiv K$ randomly selected elemental starts (eg K = 500) with k LTS concentration steps and the attractors $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{k,B}$.

i) Then the resulting CLTS estimator is a $\sqrt{n}$ consistent HB estimator if $\hat{\boldsymbol{\beta}}_{OLS}$ is $\sqrt{n}$ consistent, and the estimator is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{OLS}$.

ii) Suppose that a HB criterion is used on the $K + 2$ attractors such that the resulting estimator is HB if a HB attractor is used. Also assume that the global minimizer of the HB criterion is a consistent estimator for $\boldsymbol{\beta}$ (eg

LMS). The resulting HB estimator is asymptotically equivalent to the OLS estimator if the OLS estimator is a consistent estimator of $\boldsymbol{\beta}$.

Proof. i) Chapter 9 shows that LTS concentration algorithm that uses a HB start is HB, and that $\hat{\boldsymbol{\beta}}_{k,B}$ is a HB biased estimator. The LTS estimator is consistent by Mašïček (2004). As $n \to \infty$, consistent estimators $\hat{\boldsymbol{\beta}}$ satisfy $Q_{LTS}(\hat{\boldsymbol{\beta}})/n - Q_{LTS}(\boldsymbol{\beta})/n \to 0$ in probability. Since $\hat{\boldsymbol{\beta}}_{k,B}$ is a biased estimator of $\boldsymbol{\beta}$, OLS will have a smaller criterion value with probability tending to one. With probability tending to one, OLS will also have a smaller criterion value than the criterion value of the attractor from a randomly drawn elemental set (by Remark 8.5, Proposition 8.7 and He and Portnoy 1992). Since K randomly chosen elemental sets are used, the CLTS estimator is asymptotically equivalent to OLS.

ii) As in the proof of i), the OLS estimator will minimize the criterion value with probability tending to one as $n \to \infty$. QED

Remark 8.7. The basic resampling algorithm evaluates a HB criterion on K randomly chosen elemental sets. Theorem 8.8 uses k LTS concentration steps on K randomly drawn elemental sets and then evaluates the HB criterion on $\boldsymbol{b}_{k1}, ..., \boldsymbol{b}_{k500}$, the biased HB attractor $\hat{\boldsymbol{\beta}}_{k,B}$ and $\hat{\boldsymbol{\beta}}_{OLS}$. Hence $k = 0$ can be used to improve the basic resampling algorithm. If $\hat{\boldsymbol{\beta}}_{OLS}$ is replaced by another consistent attractor, say $\hat{\boldsymbol{\beta}}_{D,n}$, then the estimator will be HB and asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{D,n}$. In other words, suppose there is a consistent attractor $\hat{\boldsymbol{\beta}}_{D,n}$, one biased HB attractor, and all of the other K attractors $\boldsymbol{b}_{a,n}$ are such that $P(\|\boldsymbol{b}_{a,n} - \boldsymbol{\beta}\| < \epsilon) \to 0$ as $\epsilon \to 0$. Attractors satisfying this requirement include randomly drawn elementals sets, randomly drawn elemental sets after k LTS concentration steps and biased attractors. Then with probability tending to one, the ratios $Q(\hat{\boldsymbol{\beta}}_{D,n})/Q(\boldsymbol{\beta})$ and $Q(\hat{\boldsymbol{\beta}}_{Q,n})/Q(\boldsymbol{\beta})$ converge to 1 as $n \to \infty$. Hence the probability that $\hat{\boldsymbol{\beta}}_{D,n}$ is the attractor that minimizes Q goes to 1, and the resulting algorithm estimator is HB and asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{D,n}$. Using $\hat{\boldsymbol{\beta}}_{D,n} = \hat{\boldsymbol{\beta}}_{OLS}$ makes sense because then the resulting estimator has 100% Gaussian efficiency. Other good choices for $\hat{\boldsymbol{\beta}}_D$ are L_1, the Wilcoxon rank estimator, $\hat{\boldsymbol{\beta}}_{k,OLS}$, the Mallows GM estimator and estimators that perform well when heteroscedasticity is present.

Remark 8.8. To use this theory for the fast LTS algorithm, which uses 500 starts, partitioning, iterates 5 starts to convergence, and then a reweight

for efficiency step, consider the following argument. Add the consistent and high breakdown biased attractors to the algorithm. Suppose the data set has n_D cases. Then the maximum number of concentration steps until convergence is bounded by k_D, say. Assume that for $n > n_D$, no more than k_D concentration steps are used. (This assumption is not unreasonable. Asymptotic theory is meant to simplify matters, not to make things more complex. Also the algorithm is supposed to be fast. Letting the maximum number of concentration steps increase to ∞ would result in an impractical algorithm.) Then the elemental attractors are inconsistent so the probability that the LTS criterion picks the consistent estimator goes to one. The "weight for efficiency step" does not change the $\sqrt{n}$ rate by He and Portnoy (1992).

8.3 Elemental Sets Fit All Planes

The previous sections showed that using a fixed number of randomly selected elemental sets results in an inconsistent estimator while letting the subset size $h_n = g(n)$ where $g(n) \to \infty$ resulted in a consistent estimator that had little outlier resistance. Since elemental sets seem to provide the most resistance, another option would be to use elemental sets, but let $K_n \to \infty$. This section provides an upper bound on the rate of such algorithms.

In the elemental basic resampling algorithm, K_n elemental sets are randomly selected, producing the estimators $\boldsymbol{b}_{1,n}, ..., \boldsymbol{b}_{K_n,n}$. Let $\boldsymbol{b}_{o,n}$ be the "best" elemental fit examined by the algorithm in that

$$\boldsymbol{b}_{o,n} = \text{argmin}_{i=1,...,K_n} \|\boldsymbol{b}_{i,n} - \boldsymbol{\beta}\|. \tag{8.12}$$

Notice that $\boldsymbol{b}_{o,n}$ is not an estimator since $\boldsymbol{\beta}$ is unknown, but since the algorithm estimator is an elemental fit, $\|\boldsymbol{b}_{A,n} - \boldsymbol{\beta}\| \geq \|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\|$, and an upper bound on the rate of $\boldsymbol{b}_{o,n}$ is an upper bound on the rate of $\boldsymbol{b}_{A,n}$. Theorem 8.2 showed that the rate of the $\boldsymbol{b}_{o,n} \leq K_n^{1/p}$, regardless of the criterion Q. *This result is one of the most powerful tools for examining the behavior of robust estimators actually used in practice.* For example, many basic resampling algorithms use $K_n = O(n)$ elemental sets drawn with replacement from all $C(n,p)$ elemental sets. Hence the algorithm estimator $\boldsymbol{b}_{A,n}$ has a rate $\leq n^{1/p}$.

This section will show that the rate of $\boldsymbol{b}_{o,n}$ is $K_n^{1/p}$ and suggests that the number of elemental sets $\boldsymbol{b}_{i,n}$ that satisfy $\|\boldsymbol{b}_{i,n} - \boldsymbol{\beta}\| \leq Mn^{\delta}$ (where $M > 0$ is some constant and $0 < \delta \leq 1$) is proportional to $n^{p(1-\delta)}$.

Two assumptions are used.
(A1) The errors are iid, independent of the predictors, and have a density f that is positive and continuous in a neighborhood of zero.
(A2) Let τ be proportion of elemental sets J that satisfy $\|\boldsymbol{X}_J^{-1}\| \leq B$ for some constant $B > 0$. Assume $\tau > 0$.

These assumptions are reasonable, but results that do not use (A2) are given later. If the errors can be arbitrarily placed, then they could cause the estimator to oscillate about $\boldsymbol{\beta}$. Hence no estimator would be consistent for $\boldsymbol{\beta}$. Note that if $\epsilon > 0$ is small enough, then $P(|e_i| \leq \epsilon) \approx 2\epsilon f(0)$. Equations (8.2) and (8.3) suggest that (A2) will hold unless the data is such that nearly all of the elemental trial designs $\boldsymbol{X}_J$ have badly behaved singular values.

Theorem 8.9. Assume that all $C(n,p)$ elemental subsets are searched and that (A1) and (A2) hold. Then $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(n^{-1})$.

Proof. Let the random variable $W_{n,\epsilon}$ count the number of errors e_i that satisfy $|e_i| \leq M_\epsilon/n$ for $i = 1, ..., n$. For fixed n, $W_{n,\epsilon}$ is a binomial random variable with parameters n and P_n where $nP_n \to 2f(0)M_\epsilon$ as $n \to \infty$. Hence $W_{n,\epsilon}$ converges in distribution to a Poisson$(2f(0)M_\epsilon)$ random variable, and for any fixed integer $k > p$, $P(W_{n,\epsilon} > k) \to 1$ as $M_\epsilon \to \infty$ and $n \to \infty$. Hence if n is large enough, then with arbitrarily high probability there exists an M_ϵ such that at least $C(k,p)$ elemental sets J_{h_n} have all $|e_{h_n i}| \leq M_\epsilon/n$ where the subscript h_n indicates that the sets depend on n. By condition (A2), the proportion of these $C(k,p)$ fits that satisfy $\|\boldsymbol{b}_{J_{h_n}} - \boldsymbol{\beta}\| \leq B\sqrt{p}M_\epsilon/n$ is greater than τ. If k is chosen sufficiently large, and if n is sufficiently large, then with arbitrarily high probability, $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| \leq B\sqrt{p}M_\epsilon/n$ and the result follows. QED

Corollary 8.10. Assume that $H_n \leq n$ but $H_n \uparrow \infty$ as $n \to \infty$. If (A1) and (A2) hold, and if $K_n = H_n^p$ randomly chosen elemental sets are used, then $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(H_n^{-1}) = O_P(K_n^{-1/p})$.

Proof. Suppose H_n cases are drawn without replacement and all $C(H_n,p) \propto H_n^p$ elemental sets are examined. Then by Theorem 8.9, the best elemental set selected by this procedure has rate H_n. Hence if $K_n = H_n^p$ randomly chosen elemental sets are used and if n is sufficiently large, then the probability of drawing an elemental set J_{h_n} such that $\|\boldsymbol{b}_{J_{h_n}} - \boldsymbol{\beta}\| \leq M_\epsilon H_n^{-1}$ goes to one as $M_\epsilon \to \infty$ and the result follows. QED

Suppose that an elemental set J is "good" if $\|\boldsymbol{b}_J - \boldsymbol{\beta}\| \leq M_\epsilon H_n^{-1}$ for some constant $M_\epsilon > 0$. The preceding proof used the fact that with high probability, good elemental sets can be found by a specific algorithm that searches $K_n \propto H_n^p$ distinct elemental sets. Since the total number of elemental sets is proportional to n^p, an algorithm that randomly chooses H_n^p elemental sets will find good elemental sets with arbitrarily high probability. For example, the elemental sets could be drawn with or without replacement from all of the elemental sets. As another example, draw a random permutation of the n cases. Let the first p cases be the 1st elemental set, the next p cases the 2nd elemental set, etc. Then about n/p elemental sets are generated, and the rate of the best elemental set is $n^{1/p}$.

Also note that the number of good sets is proportional to $n^p H_n^{-p}$. In particular, if $H_n = n^\delta$ where $0 < \delta \leq 1$, then the number of "good" sets is proportional to $n^{p(1-\delta)}$. If the number of randomly drawn elemental sets $K_n = o((H_n)^p)$, then $\|\boldsymbol{b}_{A,n} - \boldsymbol{\beta}\| \neq O_P(H_n^{-1})$ since $P(\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| \leq H_n^{-1} M_\epsilon) \rightarrow 0$ for any $M_\epsilon > 0$.

A key assumption to Corollary 8.10 is that the elemental sets are randomly drawn. If this assumption is violated, then the rate of the best elemental set could be much higher. For example, the single elemental fit corresponding to the L_1 estimator could be used, and this fit has a $n^{1/2}$ rate.

The following argument shows that similar results hold if the predictors are iid with a multivariate density that is everywhere positive. For now, assume that the regression model contains a constant: $\boldsymbol{x} = (1, x_2, ..., x_p)^T$. Construct a (hyper) pyramid and place the "corners" of the pyramid into a $p \times p$ matrix $\boldsymbol{W}$. The pyramid defines p "corner regions" $R_1, ..., R_p$. The p points that form $\boldsymbol{W}$ are not actual observations, but the fit $\boldsymbol{b}_J$ can be evaluated on $\boldsymbol{W}$. Define the $p \times 1$ vector $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{\beta}$. Then $\boldsymbol{\beta} = \boldsymbol{W}^{-1}\boldsymbol{z}$, and $\hat{\boldsymbol{z}} = \boldsymbol{W}\boldsymbol{b}_J$ is the fitted hyperplane evaluated at the corners of the pyramid. If an elemental set has one observation in each corner region and if all p absolute errors are less than ϵ, then the absolute deviation $|\delta_i| = |z_i - \hat{z}_i| < \epsilon$, $i = 1, ..., p$.

To fix ideas and notation, we will present three examples. The first two examples consider the simple linear regression model with one predictor and an intercept while the third example considers the multiple regression model with two predictors and an intercept.

Example 8.5. Suppose the design has exactly two distinct predictor values, $(1, x_{1,2})$ and $(1, x_{2,2})$, where $x_{1,2} < x_{2,2}$ and

$$P(Y_i = \beta_1 + \beta_2 x_{1,2} + e_i) = P(Y_i = \beta_1 + \beta_2 x_{2,2} + e_i) = 0.5.$$

Notice that

$$\boldsymbol{\beta} = \boldsymbol{X}^{-1}\boldsymbol{z}$$

where

$$\boldsymbol{z} = (z_1, z_2)^T = (\beta_1 + \beta_2 x_{1,2}, \beta_1 + \beta_2 x_{2,2})^T$$

and

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,2} \\ 1 & x_{2,2} \end{bmatrix}.$$

If we assume that the errors are iid $N(0,1)$, then $P(Y_i = z_j) = 0$ for $j = 1, 2$ and $n \geq 1$. However,

$$\min_{i=1,\ldots,n} |Y_i - z_j| = O_P(n^{-1}).$$

Suppose that the elemental set $J = \{i_1, i_2\}$ is such that $x_{i_j} = x_j$ and $|y_{i_j} - z_j| < \epsilon$ for $j = 1, 2$. Then $\boldsymbol{b}_J = \boldsymbol{X}^{-1}\boldsymbol{Y}_J$ and

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| \leq \|\boldsymbol{X}^{-1}\| \|\boldsymbol{Y}_J - \boldsymbol{z}\| \leq \|\boldsymbol{X}^{-1}\| \sqrt{2}\, \epsilon.$$

Hence $\|\boldsymbol{b}_J - \boldsymbol{\beta}\|$ is bounded by ϵ multiplied by a constant (free of n).

Example 8.6. Now assume that $Y_i = \beta_1 + \beta_2 x_{i,2} + e_i$ where the design points $x_{i,2}$ are iid $N(0,1)$. Although there are no replicates, we can still evaluate the elemental fit at two points, say w_1 and w_2 where $w_2 > 0$ is some number (eg $w_2 = 1$) and $w_1 = -w_2$. Since $p = 2$, the 1-dimensional pyramid is simply a line segment $[w_1, w_2]$ and

$$\boldsymbol{W} = \begin{bmatrix} 1 & w_1 \\ 1 & w_2 \end{bmatrix}.$$

Let region $R_1 = \{x_2 : x_2 \leq w_1\}$ and let region $R_2 = \{x_2 : x_2 \geq w_2\}$. Now a fit $\boldsymbol{b}_J$ will be a "good" approximation for $\boldsymbol{\beta}$ if J corresponds to one observation $x_{i_1,2}$ from R_1 and one observation $x_{i_2,2}$ from R_2 *and* if both absolute errors are small compared to w_2. Notice that the observations with absolute errors $|e_i| < \epsilon$ fall between the two lines $y = \beta_1 + \beta_2 x_2 \pm \epsilon$. If the errors e_i are iid

$N(0,1)$, then the number of observations in regions R_1 and R_2 with errors $|e_i| < \epsilon$ will increase to ∞ as n increases to ∞ provided that

$$\epsilon = \frac{1}{n^\delta}$$

where $0 < \delta < 1$.

Now we use a trick to get bounds. Let $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{\beta}$ be the true line evaluated at w_1 and w_2. Thus $\boldsymbol{z} = (z_1, z_2)^T$ where $z_i = \beta_1 + \beta_2 w_i$ for $i = 1, 2$. Consider any subset $J = \{i_1, i_2\}$ with $x_{i_j,2}$ in R_j and $|e_{i_j}| < \epsilon$ for $j = 1, 2$. The line from this subset is determined by $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$ so

$$\hat{\boldsymbol{z}} = \boldsymbol{W}\boldsymbol{b}_J$$

is the fitted line evaluated at w_1 and w_2. Let the deviation vector

$$\boldsymbol{\delta}_J = (\delta_{J,1}, \delta_{J,2})^T$$

where

$$\boldsymbol{\delta}_{J,i} = z_i - \hat{z}_i.$$

Hence

$$\boldsymbol{b}_J = \boldsymbol{W}^{-1}(\boldsymbol{z} - \boldsymbol{\delta}_J)$$

and

$$|\delta_{J,i}| \leq \epsilon$$

by construction. Thus

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{W}^{-1}\boldsymbol{z} - \boldsymbol{W}^{-1}\boldsymbol{\delta}_J - \boldsymbol{W}^{-1}\boldsymbol{z}\|$$

$$\leq \|\boldsymbol{W}^{-1}\|\|\boldsymbol{\delta}_J\| \leq \|\boldsymbol{W}^{-1}\|\sqrt{2}\ \epsilon.$$

The basic idea is that if a fit is determined by one point from each region and if the fit is good, then the fit has small deviation at points w_1 and w_2 because *lines can't bend.* See Figure 8.3. Note that the bound is true for *every* fit such that one point is in each region and both absolute errors are less than ϵ. The number of such fits can be enormous. For example, if ϵ is a constant, then the number of observations in region R_i with errors less than ϵ is proportional to n for $i = 1, 2$. Hence the number of "good" fits from the two regions is proportional to n^2.

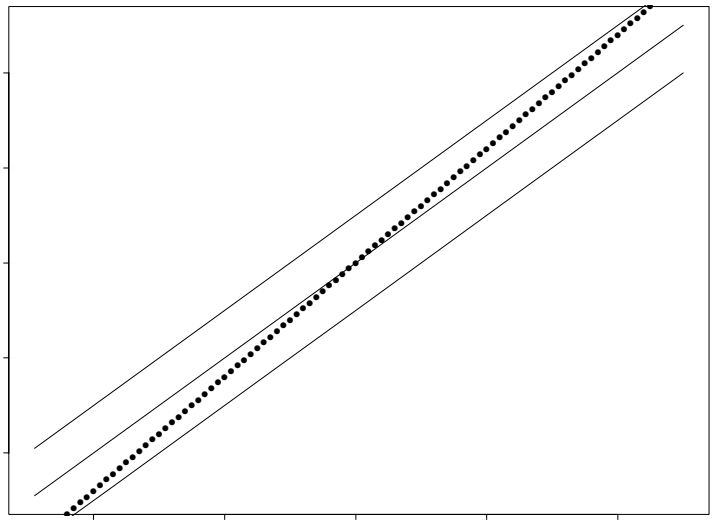

Figure 8.3: The true line is y = x + 0.

Example 8.7. Now assume that $p = 3$ and $Y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + e_i$ where the predictors $(x_{i,2}, x_{i,3})$ are scattered about the origin, eg iid $N(\mathbf{0}, \boldsymbol{I}_2)$. Now we need a matrix $\boldsymbol{W}$ and three regions with many observations that have small errors. Let

$$\boldsymbol{W} = \begin{bmatrix} 1 & a & -a/2 \\ 1 & -a & -a/2 \\ 1 & 0 & a/2 \end{bmatrix}$$

for some $a > 0$ (eg $a = 1$). Note that the three points $(a, -a/2)^T$, $(-a, -a/2)^T$, and $(0, a/2)^T$ determine a triangle. Use this triangle as the pyramid. Then the corner regions are formed by extending the three lines that form the triangle and using points that fall opposite of a corner of the triangle. Hence

$$R_1 = \{(x_2, x_3)^T : x_3 < -a/2 \text{ and } x_2 > a/2 - x_3\},$$

$$R_2 = \{(x_2, x_3)^T : x_3 < -a/2 \text{ and } x_2 < x_3 - a/2\}, \text{ and}$$

$$R_3 = \{(x_2, x_3)^T : x_3 > x_2 + a/2 \text{ and } x_3 > a/2 - x_2\}.$$

See Figure 8.4.

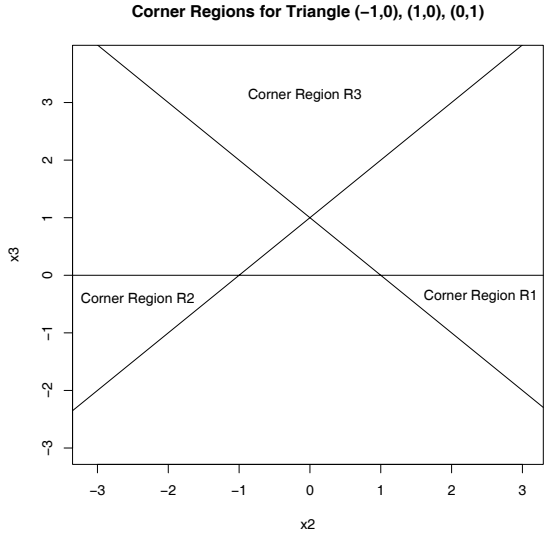

Figure 8.4: The Corner Regions for Two Predictors and a Constant.

Now we can bound certain fits in a manner similar to that of Example 8.6. Again let $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{\beta}$. The notation $\boldsymbol{x} \in R_i$ will be used to indicate that $(x_2, x_3)^T \in R_i$. Consider any subset $J = \{i_1, i_2, i_3\}$ with $\boldsymbol{x}_{i_j}$ in R_j and $|e_{i_j}| < \epsilon$ for $j = 1, 2$, and 3. The plane from this subset is determined by $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$ so

$$\hat{\boldsymbol{z}} = \boldsymbol{W}\boldsymbol{b}_J$$

is the fitted plane evaluated at the corners of the triangle. Let the deviation vector

$$\boldsymbol{\delta}_J = (\delta_{J,1}, \delta_{J,2}, \delta_{J,3})^T$$

where

$$\delta_{J_i} = z_i - \hat{z}_i.$$

Hence

$$\boldsymbol{b}_J = \boldsymbol{W}^{-1}(\boldsymbol{z} - \boldsymbol{\delta}_J)$$

and

$$|\delta_{J,i}| \leq \epsilon$$

by construction. Thus

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{W}^{-1}\boldsymbol{z} - \boldsymbol{W}^{-1}\boldsymbol{\delta}_J - \boldsymbol{W}^{-1}\boldsymbol{z}\|$$

$$\leq \|\boldsymbol{W}^{-1}\|\|\boldsymbol{\delta}_J\| \leq \|\boldsymbol{W}^{-1}\|\sqrt{3}\ \epsilon.$$

For Example 8.7, there is a prism shaped region centered at the triangle determined by $\boldsymbol{W}$. Any elemental subset J with one point in each corner region and with each absolute error less than ϵ produces a plane that cuts the prism. Hence each absolute deviation at the corners of the triangle is less than ϵ.

The geometry in higher dimensions uses hyperpyramids and hyperprisms. When $p = 4$, the $p = 4$ rows that form $\boldsymbol{W}$ determine a 3–dimensional pyramid. Again we have 4 corner regions and only consider elemental subsets consisting of one point from each region with absolute errors less than ϵ. The resulting hyperplane will cut the hyperprism formed by extending the pyramid into 4 dimensions by a distance of ϵ. Hence the absolute deviations will be less than ϵ.

We use the pyramids to insure that the fit from the elemental set is good. Even if all p cases from the elemental set have small absolute errors, the resulting fit can be very poor. Consider a typical scatterplot for simple linear regression. Many pairs of points yield fits almost orthogonal to the "true" line. If the 2 points are separated by a distance d, and the errors are very small compared to d, then *the fit is close* to $\boldsymbol{\beta}$. The separation of the p cases in $p-$space by a $(p-1)$–dimensional pyramid is sufficient to insure that the elemental fit will be good if all p of the absolute errors are small.

Now we describe the pyramids in a bit more detail. Since our model contains a constant, if $p = 2$, then the 1–dimensional pyramid is simply a line segment. If $p = 3$, then the 2–dimensional pyramid is a triangle, and in general the $(p-1)$–dimensional pyramid is determined by p points. We also need to define the p *corner regions* R_i. When $p = 2$, the two regions are to the left and right of the line segment. When $p = 3$, the corner regions are formed by extending the lines of the triangle. In general, there are p corner regions, each formed by extending the $p-1$ surfaces of the pyramid that form the corner. Hence each region looks like a pyramid without a base. (Drawing pictures may help visualizing the geometry.)

The pyramid determines a $p \times p$ matrix $\boldsymbol{W}$. Define the $p \times 1$ vector $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{\beta}$. Hence

$$\boldsymbol{\beta} = \boldsymbol{W}^{-1}z.$$

Note that the p points that determine $\boldsymbol{W}$ are not actual observations, but $\boldsymbol{W}$ will be useful as a tool to obtain a bound as in Examples 8.6 and 8.7.

The notation $\boldsymbol{x} \in R_i$ will be used to indicate that $(x_2, ..., x_p)^T \in R_i$.

Lemma 8.11. Fix the pyramid that determines $(\boldsymbol{z}, \boldsymbol{W})$ and consider any elemental set $(\boldsymbol{X}_J, \boldsymbol{Y}_J)$ with each point $(\boldsymbol{x}_{hi}^T, y_{hi})$ such that $\boldsymbol{x}_{hi} \in$ a corner region R_i and each absolute error $|y_{hi} - \boldsymbol{x}_{hi}^T\boldsymbol{\beta}| \leq \epsilon$. Then the elemental set produces a fit $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$ such that

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| \leq \|\boldsymbol{W}^{-1}\| \sqrt{p}\, \epsilon. \tag{8.13}$$

Proof. Let the $p \times 1$ vector $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{\beta}$, and consider any subset $J = \{h_1, h_2, ..., h_p\}$ with $\boldsymbol{x}_{hi}$ in R_i and $|e_{hi}| < \epsilon$ for $i = 1, 2, ..., p$. The fit from this subset is determined by $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$ so $\hat{\boldsymbol{z}} = \boldsymbol{W}\boldsymbol{b}_J$. Let the $p \times 1$ deviation vector $\boldsymbol{\delta} = (\delta_1, ..., \delta_p)^T$ where $\delta_i = z_i - \hat{z}_i$. Then $\boldsymbol{b}_J = \boldsymbol{W}^{-1}(\boldsymbol{z} - \boldsymbol{\delta})$ and $|\delta_i| \leq \epsilon$ by construction. Thus $\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{W}^{-1}\boldsymbol{z} - \boldsymbol{W}^{-1}\boldsymbol{\delta} - \boldsymbol{W}^{-1}\boldsymbol{z}\|$ $\leq \|\boldsymbol{W}^{-1}\|\|\boldsymbol{\delta}\| \leq \|\boldsymbol{W}^{-1}\|\sqrt{p}\, \epsilon$. QED

Remark 8.9. When all elemental sets are searched, Theorem 8.2 showed that the rate of $\boldsymbol{b}_{o,n} \leq n$. Also, the rate of $\boldsymbol{b}_{o,n} \in [n^{1/2}, n]$ since the L_1 estimator is elemental and provides the lower bound.

Next we will consider all $C(n, p)$ elemental sets and again show that best elemental fit $\boldsymbol{b}_{o,n}$ satisfies $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(n^{-1})$. To get a bound, we need to assume that the number of observations in each of the p corner regions is proportional to n. This assumption is satisfied if the nontrivial predictors are iid from a distribution with a joint density that is positive on the entire $(p-1)-$dimensional Euclidean space. We replace (A2) by the following assumption.

(A3) Assume that the probability that a randomly selected $\boldsymbol{x} \in R_i$ is bounded below by $\alpha_i > 0$ for large enough n and $i = 1, ..., p$.

If U_i counts the number of cases $(\boldsymbol{x}_j^T, y_j)$ that have $\boldsymbol{x}_j \in R_i$ and $|e_i| < M_\epsilon/H_n$, then U_i is a binomial random variable with success probability proportional to M_ϵ/H_n, and the number G_n of elemental fits $\boldsymbol{b}_J$ satisfying Equation (8.13) with ϵ replaced by M_ϵ/H_n satisfies

$$G_n \geq \prod_{i=1}^{p} U_i \propto n^p (\frac{M_\epsilon}{H_n})^p.$$

Hence the probability that a randomly selected elemental set $\boldsymbol{b}_J$ that satisfies $\|\boldsymbol{b}_J - \boldsymbol{\beta}\| \leq \|\boldsymbol{W}^{-1}\| \sqrt{p}\, M_\epsilon/H_n$ is bounded below by a probability that is

proportional to $(M_\epsilon/H_n)^p$. If the number of randomly selected elemental sets $K_n = H_n^p$, then

$$P(\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| \leq \|\boldsymbol{W}^{-1}\| \sqrt{p}\, \frac{M_\epsilon}{H_n}) \to 1$$

as $M_\epsilon \to \infty$. Notice that one way to choose K_n is to draw $H_n \leq n$ cases without replacement and then examine all $K_n = C(H_n, p)$ elemental sets. These remarks prove the following corollary.

Corollary 8.12. Assume that (A1) and (A3) hold. Let $H_n \leq n$ and assume that $H_n \uparrow \infty$ as $n \to \infty$. If $K_n = H_n^p$ elemental sets are randomly chosen then

$$\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(H_n^{-1}) = O_P(K_n^{-1/p}).$$

In particular, if all $C(n, p)$ elemental sets are examined, then $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(n^{-1})$. Note that Corollary 8.12 holds as long as the bulk of the data satisfies (A1) and (A3). Hence if a fixed percentage of outliers are added to clean cases, rather than replacing clean cases, then Corollary 8.12 still holds. The following result shows that elemental fits can be used to approximate any $p \times 1$ vector $\boldsymbol{c}$. Of course this result is asymptotic, and some vectors will not be well approximated for reasonable sample sizes.

Theorem 8.13. Assume that (A1) and (A3) hold and that the error density f is positive and continuous everywhere. Then the closest elemental fit $\boldsymbol{b}_{c,n}$ to any $p \times 1$ vector $\boldsymbol{c}$ satisfies $\|\boldsymbol{b}_{c,n} - \boldsymbol{c}\| = O_P(n^{-1})$.

Proof sketch. The proof is essentially the same. Sandwich the plane determined by $\boldsymbol{c}$ by only considering points such that $|g_i| \equiv |y_i - \boldsymbol{x}_i^T \boldsymbol{c}| < \alpha$. Since the e_i's have positive density, $P(|g_i| < \alpha) \propto 1/\alpha$) (at least for $\boldsymbol{x}_i$ in some ball of possibly huge radius R about the origin). Also the pyramid needs to lie on the $\boldsymbol{c}$-plane and the corner regions will have smaller probabilities. By placing the pyramid so that $\boldsymbol{W}$ is in the "center" of the $\boldsymbol{X}$ space, we may assume that these probabilities are bounded away from zero, and make M_ϵ so large that the probability of a "good" elemental set is larger than $1 - \epsilon$. QED

This result proves that elemental sets can be useful for projection pursuit as conjectured by Rousseeuw and Leroy (1987, p. 145). Normally we will only be interested in insuring that many elemental fits are close to $\boldsymbol{\beta}$. If the

errors have a pdf which is positive only in a neighborhood of 0, eg uniform(-1, 1), then Corollary 8.12 holds, but some slope intercept combinations cannot be realized. If the errors are not symmetric about 0, then many fits may be close to $\boldsymbol{\beta}$, but estimating the constant term without bias may not be possible. If the model does not contain a constant, then results similar to Corollary 8.12 and Theorem 8.13 hold, but a p dimensional pyramid is used in the proofs instead of a $(p-1)$ dimensional pyramid.

8.4 Complements

Olive first proved that the elemental basic resampling algorithm is inconsistent in 1996. My current proof is simple: for a randomly selected set of size h_n to produce a consistent estimator of $\boldsymbol{\beta}$, the size h_n must go to ∞ as $n \to \infty$. An elemental set uses $h_n = p$ for MLR. Thus each elemental fit is an inconsistent estimator of $\boldsymbol{\beta}$, and an algorithm that chooses from K elemental fits is also inconsistent.

For MLR where $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ and $\boldsymbol{\beta}$ is a $p \times 1$ coefficient vector, an elemental fit $\boldsymbol{b}_J$ is the exact fit to p randomly drawn cases. The p cases scatter about the regression plane $\boldsymbol{x}^T\boldsymbol{\beta}$, so a randomly drawn elemental fit $\boldsymbol{b}_{J_n}$ will be a consistent estimator of $\boldsymbol{\beta}$ only if all p absolute errors $|e_i|$ go to zero as $n \to \infty$. For iid errors with a pdf, the probability that a randomly drawn case has $|e_i| < \epsilon$ goes to 0 as $\epsilon \to 0$. Hence if $\boldsymbol{b}_{J_n}$ is the fit from a randomly drawn elemental set, then $P(\|\boldsymbol{b}_{J_n} - \boldsymbol{\beta}\| > \epsilon)$ becomes arbitrarily close to 1 as $\epsilon \to 0$.

The widely used basic resampling and concentration algorithms that use a fixed number K of randomly drawn elemental sets are inconsistent, because each attractor is inconsistent. Theorem 8.8 shows that it is easy to modify some of these algorithms so that the easily computed modified estimator is a $\sqrt{n}$ consistent high breakdown (HB) estimator. The basic idea is to evaluate the criterion on K elemental attractors as well as on a $\sqrt{n}$ consistent estimator such as OLS and on an easily computed HB but biased estimator such as $\hat{\boldsymbol{\beta}}_{k,B}$. Similar ideas will be used to create easily computed $\sqrt{n}$ consistent HB estimators of multivariate location and dispersion. See Section 10.7.

This chapter followed Hawkins and Olive (2002) and Olive and Hawkins (2007ab, 2008) closely. The "basic resampling", or "elemental set" method was used for finding outliers in the regression setting by Rousseeuw (1984), Siegel (1982), and Hawkins, Bradu and Kass (1984). Farebrother (1997)

sketches the history of elemental set methods. Also see Mayo and Gray (1997). Hinich and Talwar (1975) used nonoverlapping elemental sets as an alternative to least squares. Rubin (1980) used elemental sets for diagnostic purposes. The "concentration" technique may have been introduced by Devlin, Gnanadesikan and Kettenring (1975) although a similar idea appears Gnanadesikan and Kettenring (1972, p. 94). The concentration technique for regression was used by Ruppert (1992) and Hawkins and Olive (1999a).

A different generalization of the elemental set method uses for its starts subsets of size greater than p (Atkinson and Weisberg 1991). Another possible refinement is a preliminary partitioning of the cases (Woodruff and Rocke, 1994, Rocke, 1998, Rousseeuw and Van Driessen, 1999, 2002).

If an exact algorithm exists but an approximate algorithm is also used, the two estimators should be distinguished in some manner. For example $\hat{\boldsymbol{\beta}}_{LMS}$ could denote the estimator from the exact algorithm while $\hat{\boldsymbol{\beta}}_{ALMS}$ could denote the estimator from the approximate algorithm. In the literature this distinction is too seldomly made, but there are a few outliers. Portnoy (1987) makes a distinction between LMS and PROGRESS LMS while Cook and Hawkins (1990, p. 640) point out that the AMVE is not the minimum volume ellipsoid (MVE) estimator (which is a high breakdown estimator of multivariate location and dispersion that is sometimes used to define weights in regression algorithms). Rousseeuw and Bassett (1991) find the breakdown point and equivariance properties of the LMS algorithm that searches all $C(n, p)$ elemental sets. Woodruff and Rocke (1994, p. 889) point out that in practice the algorithm *is* the estimator. Hawkins (1993a) has some results when the fits are computed from disjoint elemental sets, and Rousseeuw (1993, p. 126) states that the all subsets version of PROGRESS is a high breakdown algorithm, but the random sampling versions of PROGRESS are *not* high breakdown algorithms.

Algorithms which use one interchange on elemental sets may be competitive. Heuristically, only $p - 1$ of the observations in the elemental set need small absolute errors since the best interchange would be with the observation in the set with a large error and an observation outside of the set with a very small absolute error. Hence $K \propto n^{\delta(p-1)}$ starts are needed. Since finding the best interchange requires $p(n - p)$ comparisons, the run time should be competitive with the concentration algorithm. Another idea is to repeat the interchange step until convergence. We do not know how many starts are needed for this algorithm to produce good results.

Theorems 8.2 and 8.9 are a correction and extension of Hawkins (1993a,

p. 582) which states that if the algorithm uses $O(n)$ elemental sets, then at least one elemental set $\boldsymbol{b}$ is likely to have its jth component b_j close to the jth component β_j of $\boldsymbol{\beta}$.

Note that one-step estimators can improve the rate of the initial estimator. For example, see Simpson, Ruppert, and Carroll (1992). Although the theory for the estimators in this paper requires an initial high breakdown estimator with at least an $n^{1/4}$ rate of convergence, implementations often use an initial inconsistent, low breakdown algorithm estimator. Instead of using `lmsreg` or `ltsreg` as the initial estimator, use the CLTS estimator of Theorem 8.8 (or the MBA or trimmed views estimators of Sections 7.6 and 11.3). The CLTS estimator can also be used to create an asymptotically efficient high breakdown cross checking estimator, but replacing OLS by an efficient estimator as in Remark 8.7 is a better option.

The Rousseeuw and Leroy (1987) data sets are available from the following website

`(www.uni-koeln.de/themen/Statistik/data/rousseeuw/).`

Good websites for Fortran programs of algorithm estimators include

`(www.agoras.ua.ac.be/) and`
`(www.stat.umn.edu/ARCHIVES/archives.html).`

8.5 Problems

8.1. Since an elemental fit $\boldsymbol{b}$ passes through the p cases, a necessary condition for $\boldsymbol{b}$ to approximate $\boldsymbol{\beta}$ well is that all p errors be small. Hence no "good" approximations will be lost when we consider only the cases with $|e_i| < \epsilon$. If the errors are iid, then for small $\epsilon > 0$, case i has

$$P(|e_i| < \epsilon) \approx 2\ \epsilon\ f(0).$$

Hence if $\epsilon = 1/n^{(1-\delta)}$, where $0 \leq \delta < 1$, find how many cases have small errors.

8.2. Suppose that $e_1, ..., e_{100}$ are iid and that $\alpha > 0$. Show that

$$P(\min_{i=1,...,100} |e_i| > \alpha) = [P(|e_1| > \alpha)]^{100}.$$

Splus Problems

For problems 8.3 and 8.4, if the animal or Belgian telephone data sets (Rousseeuw and Leroy 1987) are not available, use the following commands.

```
> zx <- 50:73
> zy <- -5.62 +0.115*zx + 0.25*rnorm(24)
> zy[15:20] <- sort(rnorm(6,mean=16,sd=2))
```

Warning: Use the command *source("A:/rpack.txt")* **to download the programs. See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *conc2*, will display the code for the function. Use the `args` command, eg *args(conc2)*, to display the needed arguments for the function.

8.3. a) Download the *Splus* function `conc2`. This function does not work in *R*.

b) Include the output from the following command in *Word*.

```
conc2(zx,zy)
```

8.4. a) Download the *Splus* function `attract` that was used to produce Figure 8.2. This function does not work in *R*.

b) Repeat the following command five times.

```
> attract(zx,zy)
```

c) Include one of the plots from the command in b) in *Word*.

8.5. This problem will not work in *R*. a) Repeat the following commands five times.

```
> zx <- rnorm(1000)
> zy <- 1 + 4*zx + rnorm(1000,sd=1)
> attract(zx,zy)
```

b) Include one of the plots from the command in a) in *Word*.

The elemental starts are inconsistent, but the attractors are iterated until convergence, and the attractors look good when there are no outliers. It is not known whether a randomly selected elemental set produces a consistent attractor when the iteration is until convergence. Changing sd=1 to sd=5 and sd=10 is informative.