3-2002

Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm

Douglas M. Hawkins
University of Minnesota - Twin Cities

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

Published in *Journal of the American Statistical Association*, 97, 136-148.

Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm

Douglas M. Hawkins and David J. Olive *

University of Minnesota and Southern Illinois University

July 31, 2003

Abstract

Since high breakdown estimators are impractical to compute exactly in large samples, approximate algorithms are used. The algorithm generally produces an estimator with a lower consistency rate and breakdown value than the exact theoretical estimator. This discrepancy grows with the sample size, with the implication that huge computations are needed for good approximations in large high-dimensioned samples.

*Douglas M. Hawkins is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455-0493, USA. David J. Olive is Assistant Professor, Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408, USA. The authors are grateful to the editors and referees for a number of helpful suggestions for improvement in the article. Their work was supported by the National Science Foundation under grants DMS 9803622 and ACI 9619020.

The workhorse for HBE has been the 'elemental set', or 'basic resampling' algorithm. This turns out to be completely ineffective in high dimensions with high levels of contamination. However, enriching it with a "concentration" step turns it into a method that is able to handle even high levels of contamination, provided the regression outliers are located on random cases. It remains ineffective if the regression outliers are concentrated on high leverage cases. We focus on the multiple regression problem, but several of the broad conclusions – notably those of the inadequacy of fixed numbers of elemental starts – are relevant to multivariate location and dispersion estimation as well.

We introduce a new algorithm – the "X-cluster" method – for large high-dimensional multiple regression data sets that are beyond the reach of standard resampling methods. This algorithm departs sharply from current HBE algorithms in that, even at a constant percentage of contamination, it is more effective the larger the sample, making a compelling case for using it in the large-sample situations that current methods serve poorly. A multi-pronged analysis, using both traditional OLS and L_1 methods along with newer resistant techniques, will often detect departures from the multiple regression model that can not be detected by any single estimator.

KEY WORDS: Elemental Sets; LMS; LTA; LTS; MCD; Outliers.

1 Introduction.

Consider the regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{1.1}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, and $\boldsymbol{e}$ is an $n \times 1$ vector of errors. The ith case $(\boldsymbol{x}_i^T, y_i)$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element of $\boldsymbol{Y}$. If $\boldsymbol{e}$ follows a normal $N(\boldsymbol{0}, \boldsymbol{I})$ distribution, then ordinary least squares (OLS) provides the maximum likelihood estimator of $\boldsymbol{\beta}$, but OLS may be arbitrarily bad if $\boldsymbol{e}$ includes outliers.

High breakdown (HB) estimators are used to produce "fits" that resist outliers. The least median of squares (LMS) estimator (Hampel 1975, p. 380), the least trimmed squares (LTS) estimator (Rousseeuw 1984), and the least trimmed absolute deviations (LTA) estimator (Bassett 1991 and Hössjer 1991) all have exact algorithms, and branch and bound algorithms (eg Agulló 1997) can be used to compute these estimators. For a trial regression fit $\boldsymbol{b}$, compute the n residuals $r_1(\boldsymbol{b}), \ldots, r_n(\boldsymbol{b})$ where

$$r_k = r_k(\boldsymbol{b}) = y_k - \boldsymbol{x}_k^T\boldsymbol{b}, \tag{1.2}$$

and let $|r|_{(1)}(\boldsymbol{b}) < |r|_{(2)}(\boldsymbol{b}) < \ldots < |r|_{(n)}(\boldsymbol{b})$ be the absolute residuals ranked from smallest to largest. Then the LTA, LTS and LMS criteria are respectively the L_1, L_2 and Chebyshev (L_∞) norms of the c smallest $|r|_{(i)}$. We will use the symbol Q to refer to any of these three criteria. By implication, the c best-fitting cases are accommodated by the fit, while the remaining $n - c$ are trimmed. The coverage c (at least $n/2$) is conventionally defined to be the value that gives maximum breakdown, but larger values may be appropriate if the data set is expected to be relatively outlier-free.

The asymptotic theory for LTS and LTA has not yet been extended beyond the location model. See Davies (1993), García-Escudero, Gordaliza, and Matrán (1999), Hössjer (1994), Stromberg, Hawkins, and Hössjer (2000), and Rousseeuw (1984) for further discussion and conjectures. Davies (1990) and Kim and Pollard (1990) derived the asymptotic theory for LMS while Yohai and Maronna (1976) and Butler (1982) derived asymptotic theory for LTS in the location model. Tableman (1994a,b) derived asymptotic theory for LTA in the location model.

While we focus on the multiple regression problem, many of our observations have parallels in the multivariate problem of estimating a location vector and dispersion matrix of multivariate data, where the residuals r_k are replaced by Mahalanobis distances of the cases from a trial location vector using a trial dispersion matrix. To stress the broader applicability of our conclusions, we will use the term "case distances" to refer to the residuals in the regression setting, and the Mahalanobis distances in the multivariate setting. The minimum covariance determinant estimator (MCD) is the pair (LTS, Q_{LTS}/n) in the location model. Rousseeuw (1984, p. 877) defines the MCD estimator to be the classical mean and dispersion estimator $(\bar{\boldsymbol{x}}, \boldsymbol{S})$ applied to the set of c cases for which the determinant of $\boldsymbol{S}$ is minimal.

Computing any of these criteria exactly is impractical in all but small data sets, since it involves the combinatorial problem of determining which c cases to cover, followed by the relatively easy problem of performing a L_1, L_2 or Chebyshev fit on these cases. Since exact computation is generally impractical, approximate algorithms are used.

The oldest of these is the "basic resampling", or "elemental set" method (Siegel

1982, Rousseeuw 1984, Hawkins, Bradu and Kass 1984). In this, trial vectors are found by randomly sampling elemental sets (subsets of size p cases for regression, $p+1$ for multivariate location/dispersion). Performing an exact fit of the regression to this subset gives a trial fit $\boldsymbol{b}$. The method consists of sampling many such subsets and using as the approximation that which gives the smallest value of the HB criterion. Evaluating all elemental sets will give the exact LTA fit. It is also a route to the "maximum depth" fit (Rousseeuw and Hubert 1999). This approach is attractive when n and p are sufficiently small that evaluating all $C(n,p)$ elemental fits is tolerable.

The newer HBE algorithms for LTA, LTS and LMS still use random elemental sets to generate starting trial fits, but then refine them using such devices as "concentration," "line search" (Ruppert 1992), and "interchange" (Hawkins and Olive 1999). All of these methods may be characterized as having a "start" – the initial trial fit, and an "attractor" – the final fit to which a start converges. In the "concentration" approach, the cases with the c smallest distances from a trial fit are found. An improved fit is then given by fitting the model to these c cases. The "interchange" approach seeks to swap one covered and one uncovered case to get a smaller criterion value. In both methods, the improvement step is iterated until no further changes reducing the criterion can be found. The resulting fit is an "attractor", which may be reached from more than one starting trial fit.

A simplified version of the LTS(c) algorithms of Ruppert (1992), Hawkins and Olive (1999) and Rousseeuw and Van Driessen (1999b) uses K elemental starts. The LTS(c) criterion is

$$Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c} |r|^2_{(i)}(\boldsymbol{b}) \tag{1.3}$$

where $|r|^2_{(i)}(\boldsymbol{b})$ is the ith smallest squared residual. For each elemental start find the exact-fit $\boldsymbol{b}$ and get the c smallest squared residuals. Find the OLS fit to these c cases and find the resulting c smallest squared residuals, and iterate until convergence. Doing this for K elemental starts leads to K (not necessarily distinct) attractors – the OLS $\boldsymbol{b}$ vectors at each convergence. The algorithm estimator $\hat{\boldsymbol{\beta}}_{ALTS}$ is the attractor that minimizes Q. Substituting the L_1 and Chebyshev criteria for OLS in the concentration step leads to equivalent LTA and LTQ algorithms.

As an illustration of an LTA concentration algorithm, consider the animal data from Rousseeuw and Leroy (1987, p. 57) available from

`http://www.uni-koeln.de/themen/Statistik/data/rousseeuw/`

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 $\boldsymbol{b}_{s,1} = (2.952, 1.025)^T$ and the sum of the $c = 14$ smallest residuals

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

Figure 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. The L_1 fit to these c highlighted cases is $\boldsymbol{b}_{2,1} = (2.076, 0.979)^T$ and

$$\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{2,1}) = 6.990.$$

The iteration consists of finding the cases corresponding to the c smallest residuals, obtaining the corresponding L_1 fit and repeating. The attractor $\boldsymbol{b}_{a,1} = \boldsymbol{b}_{8,1} = (1.741, 0.821)^T$

and the LTA(c) criterion evaluated at the attractor is

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

Figure 1b shows the attractor and that the c highlighted cases corresponding to the smallest absolute residuals are much more concentrated than those in Figure 1a. Figure 2a shows 5 randomly selected starts while Figure 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.

Algorithms for the MCD are similar. The ith start $(\bar{\boldsymbol{x}}_{si}, \boldsymbol{S}_{si})$ consists of the sample mean and covariance computed from $p+1$ cases selected without replacement. Then $(\bar{\boldsymbol{x}}_{2i}, \boldsymbol{S}_{2i})$ is the sample mean and covariance computed from the cases corresponding to the c smallest Mahalanobis distances $\mathrm{MD}_{(1)}(\bar{\boldsymbol{x}}_{si}, \boldsymbol{S}_{si}), ..., \mathrm{MD}_{(c)}(\bar{\boldsymbol{x}}_{si}, \boldsymbol{S}_{si})$. A new set of Mahalanobis distances is generated and the iteration continues. Rousseeuw and Van Driessen (1999, p. 214) prove that the MCD criterion $\det(\boldsymbol{S}_{j+1,i}) \leq \det(\boldsymbol{S}_{j,i})$ with equality iff $(\bar{\boldsymbol{x}}_{j+1,i}, \boldsymbol{S}_{j+1,i}) = (\bar{\boldsymbol{x}}_{ji}, \boldsymbol{S}_{ji})$. Hence the start tends to rapidly converge to the attractor $(\bar{\boldsymbol{x}}_{ai}, \boldsymbol{S}_{ai})$.

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, 1999ab).

For regression we will fit either ordinary least squares (OLS) or least absolute deviations (L_1) to the subset. These two choices allow an enormous range of regression criteria to be approximated. This class includes elemental algorithms and the SURREAL algo-

rithms (Ruppert 1992) for the LTS, LMS, and S estimators. The class also includes the FLTS algorithm (Rousseeuw and Van Driessen 1999b), and algorithms for the least adaptively trimmed sum of squares (LATS) and the least adaptively trimmed sum of absolute deviations (LATA) estimators (Olive and Hawkins 1999).

Section 2 shows that resampling algorithms that use a fixed number K of starts of bounded size (eg elemental) produce inconsistent estimators. Section 3 gives suggestions for the practitioner, and section 4 provides examples and simulations. Section 5 introduces a new algorithm – the "X-cluster" algorithm.

2 Inconsistency of Resampling Algorithms

The following notation is useful. For regression, 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

$$\boldsymbol{b}_{os,n} = \text{argmin}_{i=1,\ldots,K} \|\boldsymbol{b}_{si,n} - \boldsymbol{\beta}\| \tag{2.1}$$

where K is the number of random starts and the Euclidean norm is used. 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 $\boldsymbol{b}_{A,n}$.

Remark 1: Failure of zero-one weighting. 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 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 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 cases with the smallest absolute residuals, then

$$\|\hat{\boldsymbol{\beta}}_{cn} - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{2.2}$$

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 an $O_P(n^{-1/2})$ estimator from the $O_P(n^{-1/3})$ initial LMS fit does not in fact improve LMS's convergence rate.

Similar results for the MCD concentration algorithm hold since Lopuhaä (1999) shows that applying the classical estimator $(\bar{\boldsymbol{x}}, \boldsymbol{S})$ to the cases with the smallest Mahalanobis distances also results in an estimator with the same rate as the affine equivariant start.

Remark 2: While the formal proofs in the literature cover OLS fitting, it is a reasonable conjecture that the result also holds if the L_1 fit is 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 1 and 2 suggest that the consistency rate of the algorithm estimator is bounded above by the rate of the best elemental start. The following lemma shows 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 $K = O(n)$ starts are used. Then the rate of the algorithm estimator is no better than $n^{-1/p}$ which drops dramatically as the dimensionality increases. The lemma is an extension of Hawkins (1993, 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 $\boldsymbol{\beta}$.

Lemma 1. (See appendix for proof.) Let the number of randomly selected elemental starts $K = K(n,p) \to \infty$ 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 $K \leq C(n,p)$. Also assume that the errors are independent of the predictors. Then $\|\boldsymbol{b}_{os,n} - \boldsymbol{\beta}\| \leq O_P(K^{-1/p})$.

Conjecture. 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 $K \leq C(n,p)$ bounded starts are used in the algorithm. Then the algorithm estimator satisfies $\|\boldsymbol{b}_{A,n} - \boldsymbol{\beta}\| = O_P(K^{-1/2p})$.

Remark 3: This rate can be achieved if the algorithm minimizing Q over all elemental subsets is $\sqrt{n}$ consistent (eg maximal depth, see Bai and He 1999). Randomly select $g(n)$

cases and let $K = C(g(n), p)$. Then apply the all elemental subset algorithm to the $g(n)$ cases.

Note that one-step estimators can improve the rate of the initial estimator. See for example Chang, McKean, Naranjo, and Sheather (1999) and Simpson, Ruppert, and Carroll (1992). The theory for the estimators in these two papers requires an initial high breakdown estimator with at least an $n^{-1/4}$ rate of convergence. Implementations though often use an initial inconsistent, low breakdown algorithm estimator. The performance of a one-step estimator when applied to an inconsistent start appears to be an open question.

Remark 4: The wide spread of subsample slopes. Some additional insights into the initial estimator 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 $N_p(\mathbf{0}, \boldsymbol{I})$. 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 $(\boldsymbol{b}-\boldsymbol{\beta})^T(\boldsymbol{b}-\boldsymbol{\beta})$ (see appendix for a proof provided by Morris L. Eaton) is distributed as $(p\ F_{p,h-p+1})/(h-p+1)$.

This 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 h the greater the chance that the random subset will be clean.

Our results show that fixed 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 lemma (a recasting of a coupon collection problem).

Lemma 2. (See appendix for proof.) Suppose that K random starts of size h are selected and let $Q_{(1)} \leq Q_{(2)} \leq ... \leq Q_{(M)}$ correspond to the order statistics of the criterion values of the $M = C(n, h)$ possible starts of size h. Let R be the rank of the smallest criterion value from the K starts. Then with probability ≈ 0.5,

$$R \leq \max(1, M[1 - (0.5)^{1/K}]).$$

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 $M = 161700$ 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.

3 Practical implications

Remark 5: Breakdown. The breakdown value of concentration algorithms that use K elemental starts is bounded above by K/n. For example if 500 starts are used and $n = 50000$, then the breakdown value is at most 1%. To cause a regression algorithm to break down, simply contaminate one observation in each starting elemental set so as to displace the fitted coefficient vector by a large amount. Since K elemental starts are used, at most K points need to be contaminated. Similarly, for MCD algorithms, if the start is computed from a contaminated elemental set, then the attractor can be made arbitrarily bad.

This is a worst-case model, but sobering results on the outlier resistance of such algorithms for a fixed data set with d gross outliers can also be derived. Assume that the LTS algorithm is applied to a fixed data set of size n where $n - d$ of the cases follow a well behaved model and $d < n/2$ of the cases are gross outliers. If $d > n - c$, then every criterion evaluation will use outliers, and every attractor will produce a bad fit even if some of the starts are good. If $d < n - c$ and if the outliers are far enough from the remaining cases, then all "clean" starts (subsets of size h that contain no outliers) will result in clean attractors that could in principle detect the outliers (though, as seen from remark 4, this may require the outliers to be hugely discrepant). If the h cases that form the start are chosen without replacement from the n cases, then the probability that the start is clean is hypergeometric. Let γ_o be the highest percentage of massive outliers that a resampling algorithm can detect reliably. Then

$$\gamma_o \approx \min(\frac{n-c}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h})100\% \tag{3.1}$$

if n is large. (Rousseeuw and Leroy 1987, p. 198 show that if the contamination proportion γ is fixed, then the probability of obtaining at least one clean subset of size h with high probability (say 0.8) is given by $0.8 = 1 - [1 - (1 - \gamma)^h]^K$. Fix the number of starts K and solve this equation for γ.)

For example, with $K = 500$ starts, $n > 100$, and $p \leq 20$ the resampling algorithm should be able to detect up to 24% outliers provided every clean start is able to at least partially separate inliers from outliers. However if $p = 50$, this proportion drops to 11%.

Remark 6: Hybrid Algorithms. More sophisticated algorithms use both concentration and partitioning. Partitioning evaluates the start on a subset of the data, and poor starts are discarded. This technique speeds up the algorithm, but the consistency and outlier resistance still depends on the number of starts. For example, equation (3.1) agrees very well with the Rousseeuw and Van Driessen (1999a) simulation performed on a hybrid MCD algorithm.

Occasionally, motivated by the distribution of $(\boldsymbol{b} - \boldsymbol{\beta})^T(\boldsymbol{b} - \boldsymbol{\beta})$ sketched above, start sizes $h > p$ are suggested. The tradeoff is that elemental sets have the highest chance of being clean, but clean starts of size $h > p$ are more likely to produce fits close to $\boldsymbol{\beta}$. This however is a very poor trade if there is appreciable contamination. Writing $m^{[r]}$ for $m(m-1)...(n-r+1)$, the ratio of the probability of a clean subset of size h to that for size p is $(n-d-p)^{[h-p]}/(n-p)^{[h-p]}$, which rapidly turns finding clean subsets into a search for needles in haystacks.

The above discussion and the results in section 2 suggest several (not necessarily original) guidelines for the practitioner.

1) Do not overlook classical (OLS and L_1) procedures and diagnostics. They often suffice where the errors e_i and their propensity to be outlying are independent of the predictors $\boldsymbol{x}_i$. To see this, suppose the data set satisfies the "no excessive maldistribution of leverage" condition that the off-diagonal elements of the "hat" matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ are $o_P(n^{-1/2})$. The fitted residual of case i is $r_i = (1 - h_{ii})e_i - \sum_{j \neq i} h_{ij}e_j$. If the e_i have a common distribution with finite variance σ^2, then the term $\sum_{j \neq i} h_{ij}e_j$ has variance $\sum_{j \neq i} h_{ij}^2\sigma^2$ and so is $o_P(1)$ – even if the error distribution is outlier-prone. Thus the cases with the largest true errors will tend to have large OLS residuals in large samples and several passes of sequential trimming using OLS should find all large outliers. This is even more true of L_1 fits, which are less susceptible to masking and swamping than OLS. The assumption of a statistical distribution for the true residuals excludes the "games against nature" framework underlying breakdown calculation but covers situations where the residuals are random, even with an outlier prone distribution. This latter framework probably covers the majority of real-world regression outlier data sets.

2) For 3 or fewer variables, use graphical methods such as scatterplots and 3D plots to detect outliers and other model violations.

3) Use several estimators – both classical and robust. Then make a scatterplot matrix of the residuals or Mahalanobis distances from the different fits. The subplots will be strongly linear if consistent estimators are used and can be used to detect a wide variety of violations of model assumptions.

4) Use $\sqrt{n}$ consistent starts (eg $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{L_1}$) for the HBE's, as well as randomly selected subset starts.

5) Ensure that sufficient random starts are used, recognizing that the 1980's recom-

mendations were far too low.

6) Use subset refinement – concentration and/or interchange. It does not improve the theoretical convergence rates, but gives dramatic practical improvement in many data sets.

7) For regression, compute the median absolute deviation of the response variable $\text{mad}(y_i)$ and the median absolute residual $\text{med}(|r|_i(\hat{\boldsymbol{\beta}}))$ from the estimator $\hat{\boldsymbol{\beta}}$. If $\text{mad}(y_i)$ is smaller, then the constant $\text{med}(y_i)$ fits the data better than $\hat{\boldsymbol{\beta}}$ according to the median squared residual criterion. In fact, Rousseeuw and Leroy (1987, p. 44) suggests

$$1 - (\frac{\text{med}(|r|_i)}{\text{mad}(y_i)})^2$$

as a robust R^2.

4 Two Examples

To illustrate these points with existing standard implementations, we examined two moderately-sized data sets with six Splus estimators: OLS, L_1, ALMS = the default version of `lmsreg`, ALTS = the default version of `ltsreg`, KLMS = `lmsreg` with the option "all" which makes $K = \min(C(n,p), 30000)$, and KLTS = `ltsreg` with $K = 100000$.

Gladstone (1905-6) records the brain weight and various head measurements for 276 individuals. This data set, along with the Buxton data set introduced below, can be downloaded from the Web site

`http:\\www.stat.umn.edu\hawkins`

We'll predict brain weight using six head measurements (head height, length, breadth,

size, cephalic index and circumference) as predictors, deleting cases 188 and 239 because of missing values. There are five infants (cases 238, 263-266) of age less than 7 months that are $\boldsymbol{x}$-outliers. Nine toddlers were between 7 months and 3.5 years of age, four of whom appear to be $\boldsymbol{x}$-outliers (cases 241, 243, 267, and 269). (The points are not labeled on the plot, but the five infants and these four toddlers are easy to recognize when discrepant.) There are 6×10^{11} elemental sets, so exhaustive enumeration is impossible.

The "RR plot", a scatterplot matrix of the residuals from several regression fits, is a powerful way of comparing different fits of the same data. We will use this data set, primarily to illustrate the use of the plot as a way of comparing fits – specifically of the non-robust and high breakdown fits. In line with our recommendation of including traditional methods in the mix, we advise always including OLS and L_1 in the RR plot. Tukey (1991) notes that the plot will be linear with slope one if the model assumptions hold. In fact, if $r_{i,j}$ is the ith residual from the jth fit, then by Cauchy-Schwartz

$$|r_{i,1} - r_{i,2}| \leq \|\boldsymbol{x}_i^T\| \ (\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\| + \|\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}\|).$$

Figure 3 shows the RR plot. We dispose of the OLS and L_1 fits by noting that the very close agreement in their residuals implies an operational equivalence in the two fits. ALMS fits the nine $\boldsymbol{x}$-outliers quite differently than OLS, L_1, and ALTS. All fits are highly correlated for the remaining 265 points, showing that all fits agree on these cases, thus focusing attention on the infants and toddlers.

All of the Splus fits except ALMS accommodated the infants. The fundamental reason that ALMS is the "outlier" among the fits is that the infants and toddlers, while well separated from the rest of data, turn out to fit the overall linear model quite well. A

strength of the LMS criterion – that it does not pay much attention to the leverage of cases – is perhaps a weakness here since it leads to the impression that these cases are bad, whereas they are no more than atypical.

Turning to optimization issues, ALMS had an objective function of 52.7 while KLMS had a much higher objective function of 114.7 even though KLMS used ten times as many subsamples. The large difference resulting from changing a run option illustrates that even on a data set that is not very large by current standards, finding "the" LMS solution is not at all reliable. In view of the questions about the adequacy of modest numbers of elemental starts, we ran extensive calculations to have a better idea of what the true LMS solution might be. We began with a total of 15 million starts of the elemental set algorithm, using the Hawkins-Simonoff (1993) code. This oscillated between solutions accommodating the infants and solutions excluding them. From run 62,000 to 966,000, the best solution was one with the infants outlying, but from there on this was dominated by a solution with criterion 43.763 accommodating them.

Finally we ran 16,000 random starts using Hawkins' feasible solution algorithm. This yielded an estimate with criterion 41.793 which accommodated the infants, but gave rather large residuals to some toddlers. Another feasible solution with criterion 42.535 made the infants out as severe outliers. Both feasible solutions beat the best elemental set approximation (criterion 43.763), though not by much. This confirms the ALMS results while showing that the LMS criterion gives quite unstable residuals.

As a second example, Buxton (1920, p. 232-5) gives 20 measurements of 88 men. We chose to predict stature using an intercept, head length, nasal height, bigonal breadth,

and cephalic index. Observation 9 was deleted since it had missing values. Five individuals, numbers 62-66, were reported to be about 0.75 inches tall with head lengths well over five feet! This appears to be a clerical error; these individuals' stature was recorded as head length and the integer 18 or 19 given for stature, making the cases massive outliers with enormous leverage. These absurdly bad observations turned out to confound the standard HBE's. The residual plots in Figure 4 show that five of the six Splus estimators accommodated them. This is a warning that even using the objective of high breakdown will not necessarily protect one from extremely aberrant data. Nor should we take much comfort in the fact that KLMS clearly identified them; the criterion of this fit was worse than that of the ALMS fit, and so should be regarded as inferior.

5 The "X-cluster" algorithm

The results so far show that large data sets in high dimension create a problem, even if they include only modest levels of contamination. Unlike the case with most statistical methods, large sample sizes (assuming a constant fraction of contamination) make things worse and not better since they increase computational loads without improving the performance of the individual starts. See also Woodruff and Rocke (1994) for a parallel assessment of the multivariate location/scatter problem.

The L_1 criterion is an $O_P(n^{-1/2})$ regression estimator that is resistant to regression outliers provided they are located on low-leverage cases; it has been found empirically that L_1 can accommodate as much as 25% contamination with regression outliers on low-leverage cases. L_1 fails though in the face of regression outliers on high leverage

cases. Thus L_1 could be used to provide at the very least good starting values, provided a way could be found to insulate it from regression outliers on high leverage cases. This idea underlies Rousseeuw and van Zomeren's (1992) idea of using a fit confined to the cases whose predictor vectors are covered by the minimum volume ellipsoid.

We now describe an algorithm – the "X-cluster" algorithm – that capitalizes on this property of the L_1 norm. It handles high levels of contamination in high dimensions, even if the regression outliers are on high-leverage cases. Unlike most alternative HBE's, it generates root-n-consistent starting values, and yields a root-n-consistent estimator. This means that it is increasingly successful with increasing sample size.

The X-cluster algorithm

- Apply clustering by reallocation using the heteroscedastic multivariate normal clustering criterion to the $\boldsymbol{X}$ matrix (see for example Hawkins 1982, Rocke and Woodruff 2000). Use a starting allocation that has some random element in it. Break the cases down into a fixed number H of clusters, restricting the reallocation so that these clusters are of approximately equal size. (In our implementation we do this by refusing to remove cases from any cluster whose size is less than half the average size.)

- Carry out an L_1 fit to the cases within each cluster.

- Using this L_1 fit as a starting point, apply the iterated concentration LTS algorithm to all n cases.

The heteroscedastic multivariate normal clustering method repeatedly reallocates the cases to H clusters in such as way as to minimize the doubled negative log likelihood

$$\sum_{k=1}^{H} n_k log|W_k/n_k|$$

where W_k is the matrix of sum of squares and cross products of deviations from the mean vector of the n_k cases allocated to cluster k.

It is an empiric truism that reallocation cluster analysis is programmed to find ellipsoids in data, and will do so whether they are there or not. This property is precisely what is needed for our purposes. We can gain a qualitative understanding of its operation by looking at the change in criterion value if we reallocate a case from cluster k to cluster j. Let D_k^2 and D_j^2 be the squared Mahalanobis distances of the case from the two clusters. Assuming both n_k and n_j are large and using this to make some simplifying approximations, the change in criterion is approximately the heteroscedastic discriminant analysis criterion

$$[log(|W_k/n_k|) - D_k^2] - [log(|W_j/n_j|) - D_j^2]$$

and the swap will improve the criterion if this change is negative. If the two clusters have the same generalized variance $|W/n|$, then the case is allocated to whichever cluster is closer in Mahalanobis distance, as would be the case with homoscedastic cluster analysis. If two clusters are equidistant though, the allocation will be made to the cluster whose generalized variance is larger. The boundaries between clusters are ellipsoids along which the cases from each cluster have the same Mahalanobis distances from their cluster. Since the Mahalanobis distances of the cases from their cluster are proportional to the cases' leverages when used in the subsequent regression, this equality of Mahalanobis distance

along the inter-cluster boundaries will translate into cases with near-equal leverage along with inter-cluster boundaries, and smaller leverages inside the cluster. Only isolated X-outlying points can have large leverage relative to their clusters.

The theoretical properties of the heteroscedastic clustering procedure are not well established. It is well known that if the method is applied to data from a mixture of normal distributions its results do not provide consistent estimators of the parameters of the component distributions (McLachlan and Basford 1988). It is also notorious for having multiple local optima (Symons 1981). Neither of these properties is damaging for our purposes. Rather the second is a positive benefit, since it means that multiple starts of the algorithm usually produce different clusterings of the cases, and therefore different potentially interesting starting values for the search for regression outliers. Another deficiency for clustering purposes is the fact that the likelihood is degenerate and can be made infinite by setting any of the cluster sizes to p. This is avoided in our use by the restriction that the clusters are kept of similar size.

We are not able to give a thorough theoretical analysis of the X-clustering as applied in the regression setting, where there are no distributional requirements on the predictors, but can give a qualitative narrative of how and why, and by implication when the method can be expected to succeed.

Rocke and Woodruff (2000) report successful use of the heteroscedastic cluster analysis for high breakdown estimation of the mean vector and covariance matrix of highly contaminated multivariate normal data. Since high leverage cases are cases outlying in the ellipsoidal metric of the predictors, their success provides some further empiric reason to anticipate that clustering may be successful in breaking the sample up into groups of

cases of comparable leverage.

Turning to the regression part of the problem, if a cluster includes cases that are regression outliers, then they cannot be concentrated on high leverage cases since, to the extent that the X-clustering was able to form roughly ellipsoidal clusters, there are no high leverage cases. The L_1 regression then will provide a slope estimator that is resistant and of relatively high statistical efficiency. Using this good starting value in the iterated concentration algorithm applied to the full data set can therefore be expected to give a good approximation to the true optimum of the criterion, despite the presence of the outliers.

Since each cluster is of size approximately n/H, the regressions fitted in the individual clusters have $O_P(n^{-1/2})$ convergence. They thus provide the square-root convergent starts that we earlier showed to be a key in achieving good performance of refinement algorithms. If many outliers are concentrated in one cluster, then there will be fewer in other clusters. Thus while some of the clusters may produce poor estimates because of the impact of more outliers than L_1 can handle, the collection of L_1 regressions from the different clusters can be expected to include some at least that are good estimates of the underlying $\boldsymbol{\beta}$.

5.1 Simulation of some larger data sets.

To investigate the performance of different methods in a high-dimension seriously contaminated setting, we simulated a number of data sets. All had $n = 1000$ and $p = 51$ (50 non-trivial predictors and an intercept), with 400 mean-shift outliers. The slope vector

$\boldsymbol{\beta}$ was set to $\mathbf{0}$ and σ to 1. The other features of the data set were varied so as to particularly challenge the different estimators being studied.

a. Design form.

Six different choices of the design matrix $\boldsymbol{X}$ were used:-

Sphere (abbreviated 'S'). In this configuration, the $\boldsymbol{x}$ vectors were randomly sampled from a $N(\mathbf{0}, \boldsymbol{I})$ distribution. This configuration should be an easy one for all the methods.

Vslash (abbreviated 'V'). Here, each $\boldsymbol{x}$ vector was a $N(\mathbf{0}, \boldsymbol{I})$ divided by a uniform $U(0, 1)$ variate. This distribution tends to produce a sprinkling of isolated very remote $\boldsymbol{x}$ vectors. Provided the outliers are not concentrated on these remote vectors though, simple full-sample methods like OLS and L_1 should handle this configuration quite effectively.

The 'true' X clusters are concentric spheroidal shells so despite the marked differences in leverage in the full-sample metric, the cases within each cluster except the innermost will have quite similar leverages. X-clustering should therefore work particularly well.

Binary (abbreviated 'B'). This configuration has each component of $\boldsymbol{x}$ either 1 or -1, each with probability 0.5. It is impossible for non-coincident cases to be very close in this configuration, and this should be favorable for elemental set methods since it reduces the frequency of near-singular elemental design matrices, though it may have numbers of singular elemental designs.

The last three configurations also contain 40% x-outlying cases. They are:

Disk and axle (abbreviated 'D&A'). This configuration is based on the example of Huber (1981) demonstrating the breakdown of M estimates of multivariate location and scatter. The non-trivial portion of the predictor vector $\boldsymbol{x}$ comprises a first component

x_1 and a $(p-1)$ component vector $\boldsymbol{x}_{(2)}$. The sample consists of 600 cases whose $\boldsymbol{x}_{(2)}$ is $N(\mathbf{0}, \boldsymbol{I})$ and whose x_1 is $N(0, \epsilon^2)$. For the remaining 400 cases, $\boldsymbol{x}_{(2)}$ is $N(\mathbf{0}, \epsilon^2 \boldsymbol{I})$ while x_1 has a scaled randomly-signed χ_{p-1} distribution. The variance ϵ^2 is chosen just large enough to avoid numeric singularity problems. The overall vector $\boldsymbol{x}$ then has a mean vector of zero and a correlation matrix $\boldsymbol{I}$. Even though the 400 cases in the second group have infinite leverage in relation to the 600-case majority, the conventional "hat matrix" diagnostics do not show them up as remarkable. This makes the configuration particularly intractable for full-sample methods, and subsampling methods that do not happen to stumble upon the 600-case majority group.

Dash and dot (abbreviated 'D&D'). The $\boldsymbol{x}_{(2)}$ vector is $N(\mathbf{0}, \boldsymbol{I})$, while x_1 is uniform U(-3,3) for 600 of the cases, and U(19,20) for the remaining 400 cases. This is a milder version of the situation in the Buxton data set.

Sphere and vslash (abbreviated 'S&V'). This final configuration has 600 cases distributed under the "sphere" model, and 400 under the "vslash" model.

b. Outlier placement.

Random or badly-placed outliers (abbreviated 'R' and 'B' respectively). For the "random" case, the 400 regression outliers were placed on randomly-selected $\boldsymbol{x}$. For the "badly-placed" case, the regression outliers were put on the x-outlying cases. The badly-placed option is possible only for the last three X configurations which have identified x-outlying cases.

c. Outlier size.

Plus, Plus/minus or degenerate (abbreviated '+', '+/-' and 'D' respectively). In all cases, a null N(0,1) $\boldsymbol{y}$ vector was generated, then the mean-shift outliers made. In the

'+' case this was done by adding 6 to the y of each outlying case. In the '+/-' case, either +6 or -6 was added to the y, the sign being determined at random. The 'D' case, which is relevant only for the three configurations with identified x-outlying cases, has a near-exact-fit for the 600 x-inlying cases, and an incompatible near-exact-fit for the 400 x-outlying cases. Any algorithm that can recognize the 600 inlying cases will then identify the remaining cases as effectively infinitely outlying, but algorithms that do not find the inlying cases tend to fail completely.

These design factors give rise to a total of 21 sample configurations. One sample was generated according to each configuration and analyzed. The algorithms investigated were:-

1. OLS.
2. L_1.
3. Random elemental sets.
4. Random elemental sets followed by concentration.
5. X-clustering.

We also made an idealized calculation of performing an OLS fit to the 600 clean cases and seeing how many outliers this fit based on perfect advance knowledge could yield.

In each of the methods, the final phase of outlier identification was made by finding the residuals from the fit and getting the root mean square of the c smallest residuals. This was then multiplied by 2.65, a factor that makes it an unbiased estimate of σ at normal data. Cases whose residuals were more than $3\hat{\sigma}$ were considered to be flagged as outliers.

To evaluate each method, we computed the average number of outliers found per

random start, and the number found in the best start.

The runs used 10,000 simple elemental sets, 100 elemental starts with concentration, and 100 X-clusterings with separate random starting allocations.

5.2 Results

Table 1 shows the percentage of outliers identified by the full-sample OLS fit, the full-sample L_1 fit, the OLS fit to the clean cases, and the results from three iterative algorithms. For each of the three iterative algorithms, we list the average percentage of outliers found per random start, and the number found with the best solution obtained in the run. The 4th column, labeled "clean", is the percentage flagged by the OLS fit to the clean cases. Taking this column first, we see that finding six-standard-deviation outliers in a 50-dimensional regression is not trivial. This has a simple explanation. If there are 400 severe regression outliers, then the median of the absolute residuals is at the 83rd percentile of a half-normal distribution, and not the 50th. Thus when we rescale the trimmed standard deviation by the correction factor of 2.65, rather than unbiased for σ, $\hat{\sigma}$ has expectation $2.65\sigma/1.43 = 1.85\sigma$, so the $3\hat{\sigma}$ cutoff for outliers is actually at 5.6σ, which indeed will flag only about half the 6σ outliers.

Based on this reasoning, we might describe an outlier search as 'successful' if it manages to locate at least half the outliers in the non-degenerate setups.

OLS failed totally in the situation where all outliers are +6 and also in the three degenerate data sets. Only where the outliers were of mixed sign did OLS have any success in detecting outliers, and this success was to say the least modest. L_1 fared

substantially better, though still not well. It too generally failed with the +6 (except, rather surprisingly, the badly-placed disk and axle) and degenerate data sets, but was more successful with the mixed-sign outliers.

Before looking at the detailed results of the three iterative algorithms, we should note that one line is initially unintuitive. In the dash & dot configuration when all outliers are at +6 displacement and placed on the x-outliers, the best fit is not the fit to the 600 clean cases – rather it is a fit accommodating the outliers. The 'clean' fit with its identification of 62% of the outliers actually yields a higher HBE criterion value than do the three HBE estimators. The failure of the HBE's to find regression outliers in this configuration is because there are arguably no regression outliers. This is the same phenomenon seen in more dramatic form with the Buxton data.

The raw elemental set approach did not fare at all well, as the results of the paper would have led one to expect. In none of the 21 setups did the best of the 10,000 random elemental sets reach the 50% threshold suggested for a 'successful' analysis. Adding the concentration step to the elemental start improved results dramatically. In most of the non-degenerate setups, the best of 100 elemental starts with concentration located a majority of the outliers.

Elemental sets with concentration failed on two of the three degenerate setups. This again is predicted by the results of the paper. With the disk & axle, and the dash & dot X configurations, so long as an elemental set contains one or more of the contaminated cases, the elemental set will not find other outliers, and nor will concentration improve matters.

Turning to the final pair of columns, the X-clustering method is, overall, the most

successful of all. In 14 out of the 21 configurations, even the average X-cluster solution detected more than half the outliers and the best of 100 random starts routinely flagged a higher percentage of outliers than even the idealized 'clean' result. (This result suggests that $\hat{\sigma}$ from the best X-cluster fit was typically less than $\hat{\sigma}$ from the clean fit.) The method was spectacularly successful with the three degenerate configurations, where it located the outliers consistently even in individual random starts.

The only configuration where X-clustering was less effective than elemental sets with concentration was the disk & axle X configuration with positive outliers placed randomly. Since even in this case its 37% outlier discovery rate was close to the 43% of the "clean' solution though, it is hard to fault it for even this modest failure.

Next, a smaller simulation of a larger and easier target was run. Here the outlier shift was 10 σ rather than 6 σ. The "clean" solution identified them with close to 100% accuracy, and columns 2 and 3 from Table 2 show the results given by the full-sample OLS and L_1 fits. OLS still did not perform very well. It did not find any outliers when they were of the same sign, and even in the mixed-sign case gave good results in only 4 of the 9 settings. L_1 was considerably more successful, finding most or all of the outliers in all the mixed-sign settings. In the same-sign settings, it was much better than OLS, but still not particularly effective. This simulation tempers the overall favorable comments made earlier about full-sample OLS and L_1 fits with the warning that they are more successful if the outliers have different displacements than if they have the same displacement.

Columns 4 through 9 from Table 2 show the results of 2,000 random elemental sets and 20 concentration starts and 20 X-cluster starts. As with the smaller shift, X-clustering was almost uniformly more effective per start than concentration. It was ineffective

however when mixed sign outliers were placed badly in the "dash and dot" configuration – the reason for this is that since each cluster tends to be confined to either the 'dot' portion or the 'dash' portion of the data, none of the starting L_1 fits spanned the two groups of points. By including points from both groups, the elemental plus concentration method succeeded in finding the outliers. See the third to last row of Table 2.

One important feature comes from comparing Tables 1 and 2. This is that raw elemental sets fared no better in finding 10σ outliers than they did with 6σ.

A referee wondered how the X-clustering algorithm performed on the Gladstone and Buxton data sets. Since X-clustering is a more reliable algorithm for reaching a conventional HBE – LTS – it does no better and no worse than the feasible solution algorithm for LTS applied to these data sets. The X-clustering does of course separate the 4 hugely anomalous cases from the rest of the data, but this does not in and of itself change the final estimate.

Computational complexity.

There is one drawback to the X-cluster method – its computational load is appreciable. The fastest method consists of forming the within-cluster mean vector and covariance matrices of some random starting allocation and getting the inverses of the covariance matrices. Thereafter we compute the impact of moving each case from its current cluster to each other cluster, and if an improvement is possible, update the two inverse covariance matrices. The initial setup involves $O(np^2 + Hp^3)$ computations to get the starting mean vectors, covariance matrices and their inverses. Investigating any one case for a possible move to another cluster involves $O(Hp^2)$ computations to get Mahalanobis distances, and the two required inverse matrix updates another $O(p^2)$. For large n and p, this

computation can be appreciable. In our simulation, for example, each X-clustering with follow-up required some 4 minutes on a 450 MHz Pentium III, making 100 random starts an overnight run.

Since statistical analysis is generally just a small part of the effort and cost of any data gathering and analysis, one should not make too much of this computational load. We consider it clearly far better to use an analysis that takes 10 hours but finds all outliers than one that takes 10 seconds but misses most of them.

One may wonder whether the X-clustering method is foolproof. It is not. Consider a data set in which each $\boldsymbol{x}$ vector, to within a very small random variation, equals one of just H distinct vectors. Then when we cluster the cases, we can expect to recover these H near-point-masses. Fitting an L_1 regression within any single cluster will likely not produce a good starting value because of the near-singularity of the design matrix and the X-clustering method will probably fail. While it is not hard to recognize this eventuality and take steps to evade the resulting problems, it is perhaps better to recognize this as another piece of evidence for the proposition that no one method of analysis solves all problems, and that a variety of approaches will provide a clearer picture than any one of them alone.

6 Conclusion

High breakdown estimation and outlier identification can be defined in terms of optimization problems, but these formulations obscure the fact that the optimization is combinatorially hard. Methods that work well on text-book-size problems may, on closer

examination, turn out to be useless for large problems. The "basic resampling", or "elemental set" method has nice theoretical properties but, as we show in this paper, is unable to handle large, dirty data sets in a tolerable amount of computation. The newer methods that combine elemental starts with refinement have the same theoretical convergence rates as does the start. Their practical performance is frequently much better, but this is not guaranteed. This argues for a multi-prong analysis of large data sets, combining high breakdown methods with traditional approaches such as OLS and L_1 which may fail but often (perhaps even usually) succeed.

We introduce a new approach, based on clustering the data by their $\boldsymbol{x}$ vectors on the heteroscedastic normal reallocation approach and using L_1 fits within clusters. Since this yields clusters of cases of comparable leverage, it is able to accommodate general outlier data sets – even those in which the regression outliers are concentrated on high leverage cases. This resulting method appears to hold considerable promise in data sets where no current algorithm is able to locate the outliers.

Appendix

Mathematical proofs:

Proof of Lemma 1. Let $J = \{c_1, ..., c_p\}$ be a randomly selected elemental set. Then $Y_J = \boldsymbol{X}_J\boldsymbol{\beta} + \boldsymbol{e}_J$ where the p errors are independent, and the data $(Y_J, \boldsymbol{X}_J)$ produce an estimator

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

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

then

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

Note 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

$$P_n \equiv P(|e_i| < \frac{M}{n^\delta}) \approx \frac{2\ M\ f(0)}{n^\delta} \tag{6.1}$$

for large n. Note that if W counts the number of errors satisfying (6.1) then $W \sim$ binomial(n, P_n), and the probability that all p errors in $\boldsymbol{e}_J$ satisfy equation (6.1) is proportional to $1/n^{\delta p}$. If $K = o(n^{\delta p})$ 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 $K^{1/p}$ for the more general result. QED

Proof of remark 4. 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 $\hat{\boldsymbol{\beta}}|W \sim N(\boldsymbol{0}, W^{-1})$. 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

Proof of Lemma 2. If W_i is the rank of the ith start, then $W_1, ..., W_K$ are iid discrete uniform on $\{1, ..., M\}$ and $R = \min(W_1, ..., W_K)$. Thus

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

and the median of R is $\text{MED}(R) \approx M[1 - (0.5)^{1/K}]$. QED

7 References

Agulló, J. (1997), "Exact Algorithms to Compute the Least Median of Squares Estimate in Multiple Linear Regression," in L_1-*Statistical Procedures and Related Topics,* ed. Dodge, Y., Institute of Mathematical Statistics, Hayward, CA, 133-146.

Atkinson, A.C., and Weisberg, S. (1991), "Simulated Annealing for the Detection of Multiple Outliers Using Least Squares and Least Median of Squares Fitting," in *Directions in Robust Statistics and Diagnostics,* Part 1, eds. Stahel, W., and Weisberg, S., Springer-Verlag, NY, 7-20.

Bai, Z.D., and He, X. (1999), "Asymptotic Distributions oof the Maximal Depth Estimators for Regression and Multivariate Location," *The Annals of Statistics,* 27, 1616-1637.

Bassett, G.W. (1991), "Equivariant, Monotonic, 50% Breakdown Estimators," *The American Statistician,* 45, 135-137.

Butler, R.W. (1982), "Nonparametric Interval and Point Prediction Using Data Trim-

ming by a Grubbs-Type Outlier Rule," *The Annals of Statistics,* 10, 197-204.

Buxton, L. H. D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

Chang, W.H., McKean, J.W., Naranjo, J.D., and Sheather, S.J. (1999), "High-Breakdown Rank Regression," *Journal of the American Statistical Association,* 94, 205-219.

Davies, P.L. (1990), "The Asymptotics of S-Estimators in the Linear Regression Model," *The Annals of Statistics,* 18, 1651-1675.

Davies, P.L. (1993), "Aspects of Robust Linear Regression," *The Annals of Statistics,* 21, 1843-1899.

Dollinger, M.B., and Staudte, R.G. (1991), "Influence Functions of Iteratively Reweighted Least Squares Estimators," *Journal of the American Statistical Association,* 86, 709-716.

García-Escudero, L.A., Gordaliza, A., and Matrán, C. (1999), "A Central Limit Theorem for Multivariate Generalized k-Means," *The Annals of Statistics,* 27, 1061-1079.

Gladstone, R. J. (1905-1906), "A Study of the Relations of the Brain to the Size of the Head," *Biometrika,* 4, 105-123.

Hampel, F.R. (1975), "Beyond Location Parameters: Robust Concepts and Methods," *Bulletin of the International Statistical Institute,* 46, 375-382.

Hawkins, D. M., (1982), (ed.), *Topics in Applied Multivariate Analysis*, Cambridge University Press.

Hawkins, D.M. (1993), "The Accuracy of Elemental Set Approximations for Regression," *Journal of the American Statistical Association,* 88, 580-589.

Hawkins, D.M., Bradu, D., and Kass, G.V. (1984), "Location of Several Outliers in Multiple Regression Data Using Elemental Sets," *Technometrics,* 26, 197-208.

Hawkins, D. M., and Olive, D. J. (1999), "Improved Feasible Solution Algorithms for High Breakdown Estimation," *Computational Statistics and Data Analysis,* 30, 1-11.

Hawkins, D. M., and Simonoff, J. S. (1993), "High Breakdown Regression and Multivariate Estimation," *Applied Statistics,* 42, 423-432.

He, X., and Portnoy, S. (1992), " Reweighted LS Estimators Converge at the Same Rate as the Initial Estimator," *The Annals of Statistics,* 20, 2161-2167.

Hössjer, O. (1991), Rank-Based Estimates in the Linear Model with High Breakdown Point, Ph.D. Thesis, Report 1991:5, Department of Mathematics, Uppsala University, Uppsala, Sweden.

Hössjer, O. (1994), "Rank-Based Estimates in the Linear Model with High Breakdown Point," *Journal of the American Statistical Association,* 89, 149-158.

Huber, P. J. (1981), *Robust Statistics*, Wiley, New York.

Kim, J., and Pollard, D. (1990), "Cube Root Asymptotics," *The Annals of Statistics,* 18, 191-219.

Lopuhaä, H. P. (1999), "Asymptotics of Reweighted Estimators of Multivariate Location and Scatter," *The Annals of Statistics,* 27, 1638-1665.

McLachlan, G. J., and Basford, K. E., (1988), *Mixture Models*, Marcel Dekker, New York.

Olive, D.J., and Hawkins, D.M. (1999), Comment on "Regression Depth," by P.J. Rousseeuw and M. Hubert, *Journal of the American Statistical Association,* 94, 416-

417.

Rocke, D. M. (1998), "Constructive Statistics: Estimators, Algorithms, and Asymptotics," in *Computing Science and Statistics,* 30, ed. Weisberg, S., Interface Foundation of North America, Inc., Fairfax Station, Va, 1-14.

Rocke, D. M. and Woodruff, D. L. (2000), "Robust Cluster Analysis and Outlier Identification", Unpublished Manuscript.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871-880.

Rousseeuw, P. J., and Hubert, M. (1999), "Regression Depth," *Journal of the American Statistical Association,* 94, 388-433.

Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection,* New York: John Wiley.

Rousseeuw, P. J., and Van Driessen, K. (1999a), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

Rousseeuw, P. J., and Van Driessen, K. (1999b), "Computing LTS Regression for Large Data Sets," University of Antwerp, Technical Report.

Rousseeuw, P.J., and van Zomeren, B.C. (1992), "A Comparison of Some Quick Algorithms for Robust Regression," *Computational Statistics and Data Analysis,* 14, 107-116.

Ruppert, D. (1992), "Computing S-Estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics,* 1, 253-270.

Ruppert, D., and Carroll, R. J. (1980), "Trimmed Least Squares Estimation in the

Linear Model," *Journal of the American Statistical Association,* 75, 828-838.

Siegel, A.F. (1982), "Robust Regression Using Repeated Medians," *Biometrika,* 69, 242-244.

Simpson, D.G., Ruppert, D., and Carroll, R.J. (1992), "On One-Step GM Estimates and Stability of Inferences in Linear Regression," *Journal of the American Statistical Association,* 87, 439-450.

Stromberg, A.J., Hawkins, D.M., and Hössjer, O. (2000), "The Least Trimmed Differences Regression Estimator and Alternatives," *Journal of the American Statistical Association,* 95, 853-864.

Symons, M., (1981), "Clustering Criteria and Multivariate Normal Mixtures", *Biometrics*, 37, 35-43.

Tableman, M. (1994a), "The Influence Functions for the Least Trimmed Squares and the Least Trimmed Absolute Deviations Estimators," *Statistics and Probability Letters,* 19, 329-337.

Tableman, M. (1994b), "The Asymptotics of the Least Trimmed Absolute Deviations (LTAD) Estimator," *Statistics and Probability Letters,* 19, 387-398.

Tukey, J. W. (1991), "Graphical Displays for Alternative Regression Fits," in *Directions in Robust Statistics and Diagnostics,* Part 2, eds. Stahel, W., and Weisberg, S., Springer-Verlag, NY, 309-326.

Welsh, A. H., and Ronchetti, E. (1993), "A Failure of Intuition: Naive Outlier Deletion in Linear Regression," Preprint.

Woodruff, D. L., and Rocke, D. M. (1994), "Computable Robust Estimation of Multi-

variate Location and Shape in High Dimension Using Compound Estimators," *Journal of the American Statistical Association,* 89, 888-896.

Yohai, V.J. and Maronna, R. (1976), "Location Estimators Based on Linear Combinations of Modified Order Statistics," *Communications in Statistics Theory and Methods,* 5, 481-486.

Table 1: Percentage of 6σ Outliers Detected

Size, Design	OLS	L1	Clean	Elemental		Concentration		X-cluster	
and Placement				Mean	Best	Mean	Best	Mean	Best
+, S, R	0	0	60	1	3	18	67	29	68
+, V, R	0	0	63	13	18	38	58	56	66
+, B, R	0	0	64	0	2	12	71	28	71
+, D&A, R	0	0	43	34	42	22	51	13	37
+, D&A, B	0	33	61	0	0	0	0	53	63
+, D&D, R	0	0	56	1	3	17	63	51	63
+, D&D, B	0	0	62	1	2	1	2	1	3
+, S&V, R	0	0	58	9	14	37	58	47	58
+, S&V, B	0	4	65	19	26	43	60	65	72
+/-, S, R	7	42	51	1	3	62	63	62	64
+/-, V, R	28	53	57	12	16	41	53	56	63
+/-, B, R	3	36	53	0	2	59	61	59	61
+/-, D&A, R	2	37	54	33	42	32	63	59	62
+/-, D&A, B	56	48	53	0	0	53	70	50	64
+/-, D&D, R	0	47	62	1	3	68	70	68	71
+/-, D&D, B	8	49	62	1	4	54	70	49	52
+/-, S&V, R	25	52	64	7	11	56	66	62	67
+/-, S&V, B	43	56	62	21	26	54	66	67	72
D, D&A	0	0	100	0	0	0	0	100	100
D, D&D	0	0	100	1	4	0	0	99	100
D, S&V	0	0	99	40 21	29	36	99	99	99

Table 2: Percentage of 10σ Outliers Detected

Size, Design	OLS	L1	Elemental		Concentration		X-cluster	
and Placement			Mean	Best	Mean	Best	Mean	Best
+, S, R	0	0	1	2	37	100	66	100
+, V, R	0	51	11	15	98	100	99	100
+, B, R	0	0	0	2	16	100	95	100
+, D&A, R	0	8	32	39	92	100	46	100
+, D&A, B	0	47	0	0	0	0	95	100
+, D&D, R	0	0	1	2	27	100	86	100
+, D&D, B	0	0	1	3	1	2	1	1
+, S&V, R	0	40	8	12	89	99	98	99
+, S&V, B	0	40	21	27	85	90	96	98
+/-, S, R	51	100	1	2	100	100	100	100
+/-, V, R	92	97	12	15	98	99	99	100
+/-, B, R	16	100	0	2	100	100	100	100
+/-, D&A, R	48	100	33	40	75	100	97	100
+/-, D&A, B	100	100	0	0	100	100	56	100
+/-, D&D, R	12	100	1	3	100	100	100	100
+/-, D&D, B	47	75	1	5	70	100	51	52
+/-, S&V, R	94	98	8	12	98	100	99	100
+/-, S&V, B	80	81	21	27	90	93	96	98

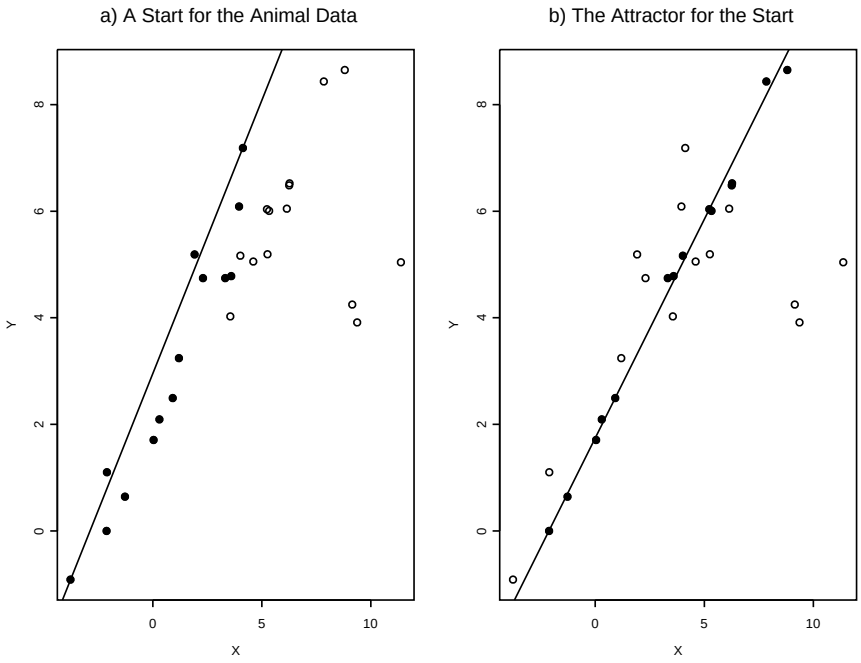

Figure 1: The Highlighted Points are More Concentrated about the Attractor

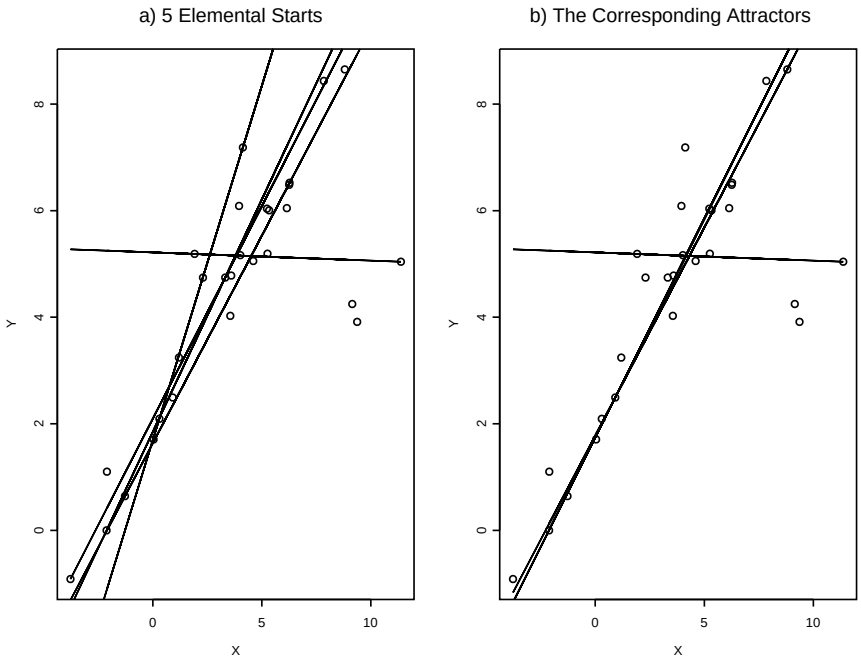

Figure 2: Starts and Attractors for the Animal Data

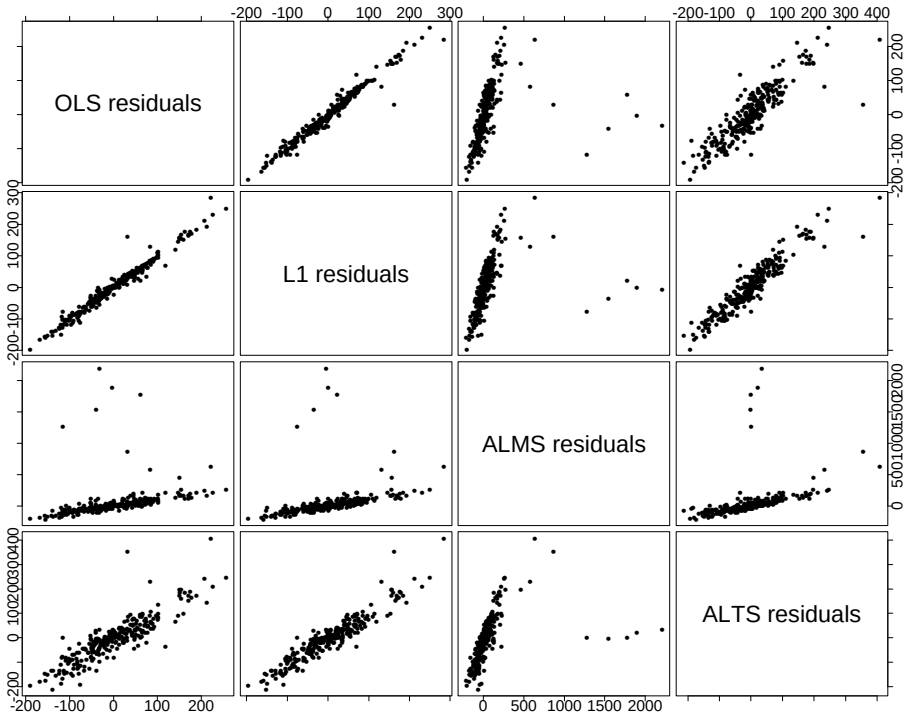

Figure 3: RR Plot for Gladstone Data

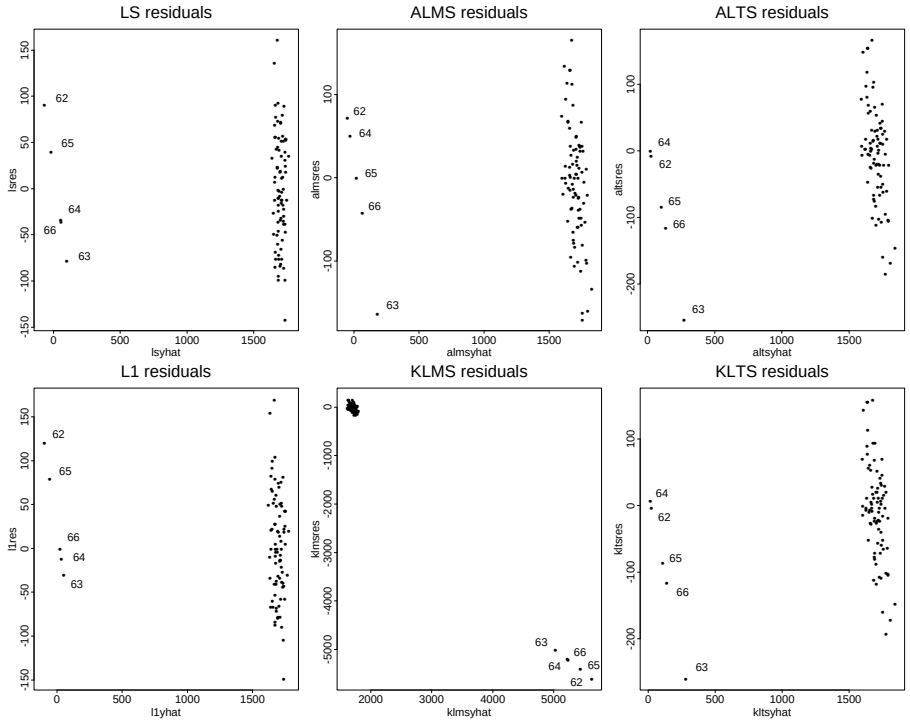

Figure 4: Residuals vs Predicted Values, Buxton Data