Improved Feasible Solution Algorithms for High Breakdown Estimation

Douglas M. Hawkins

David J. Olive

Department of Applied Statistics

University of Minnesota

St Paul, MN 55108

Abstract

High breakdown estimation allows one to get reasonable estimates of the parameters from a sample of data even if that sample is contaminated by large numbers of awkwardly placed outliers. Two particular application areas in which this is of interest are multiple linear regression, and estimation of the location vector and scatter matrix of multivariate data. Standard high breakdown criteria for the regression problem are the least median of squares (LMS) and least trimmed squares (LTS); those for the multivariate location/scatter problem are the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD). All of these present daunting computational problems. The 'feasible solution algorithms' for these criteria have been shown to have excellent performance for text-book sized problems, but their performance on much larger data sets is less impressive. This paper points out a computationally cheaper feasibility condition for LTS, MVE and MCD, and shows how the combination of the criteria leads to improved performance on large data sets. Algorithms incorporating these improvements are available from the author's Web site.

KEY WORDS: Linear model, outliers, high breakdown estimation, least trimmed squares, minimum volume ellipsoid, minimum covariance determinant

1 INTRODUCTION

Two related problems in multi-parameter estimation are that of estimating the coefficient vector in a multiple linear regression, and of estimating the location vector and scatter matrix of multivariate data. Conventional second-moment methods such as least squares for regression, and calculation of the sample mean vector and covariance matrix for the location/scatter problem can fail completely in the presence of even a quite modest numbers of outliers, and this is true even if they are supplemented with conventional diagnostics. The methodology of high breakdown estimation addresses this estimation problem by providing estimates that will have respectable performance despite the possible presence of outliers.

We will deal with both the regression and the multivariate location/scatter problem within the framework of a single notation. Suppose we have a sample of n p-component vectors X_1, X_2, ..., X_n. For the location/scatter problem, a common baseline model might be that these vectors follow a common multivariate normal distribution with mean vector μ and covariance matrix Σ

$$X_i \sim N(\mu, \Sigma), \quad i = 1, ..., n.$$

The high breakdown location/scatter problem comprises estimating μ and $\sum$ if the data vectors include a minority of vectors in which the X_i has been replaced by some (possibly maliciously chosen) contaminating vectors.

For the regression problem, we also have a dependent variable Y_i, which is related to the X_i vector by the linear model

$$Y_i = X_i^t \beta + \epsilon_i$$

where ϵ_i is a disturbance. In the common baseline model, the disturbances ϵ_i are independent $N(0, \sigma^2)$ random variables. The need for high breakdown estimation arises if a minority of the Y_i have been replaced by some other (possibly maliciously chosen) values.

There are close connections between these two problems. For example, in the regression case, outlying X_i correspond to high leverage cases, and so the ability to detect outlying X_i is important for the identification of cases that are highly influential in the regression. Also, if a case is outlying in the regression situation, while we commonly think of it as a case whose Y_i has been corrupted, it is logically equally possible that its Y_i is correct, but that its X_i was corrupted. Rousseeuw and van Zomeren (1990) emphasize the importance of calculating both robust residuals and robust leverages in high breakdown regression case diagnostics.

It is an essential feature of the high breakdown formulation that a majority of the cases do conform to the baseline model. In symbols, there are at least C cases for which

$$X_i \sim N(\mu, \Sigma) \text{ in the location/scatter problem}$$

$$Y_i \sim N(X_i^t \beta, \sigma^2) \text{ in the multiple regression problem.}$$

The required minimum coverage C will not concern us here; however, it is discussed in Rousseeuw and Leroy (1987) and is strictly greater than half the sample size. The basic methodology of high breakdown estimation consists of a two-part process – of finding

which C of the n cases are most plausibly the cases that conform to the baseline model, and then using this identification to estimate the parameters $--\beta$ for the regression problem, and μ and $\sum$ for the location/scatter problem.

For both LTS and MCD, the second step of the process is easy and immediate – for LTS, once one has decided on the C candidate cases to 'cover', the LTS estimate of β is given by the ordinary least squares (OLS) regression of Y on X using just these C cases. The full LTS problem then consists of:

- Explicitly or implicitly consider all possible partitions of the cases into C covered and $n - C$ uncovered cases, and fit the OLS regression of Y on X in each such partition.

- The exact LTS estimator is given by that partition and that OLS fit for which the residual sum of squares is a minimum.

The second step is also easy for LMS and least trimmed absolute deviations (LTA) which replaces the squared residual in the LTS criterion by the absolute residual. LMS is computed by fitting the Chebyshev (L_∞) fit instead of OLS while LTA is computed by fitting least absolute deviations (L_1) instead of OLS.

In the MCD case, the location vector μ is given by the conventional sample mean vector of the C covered cases, and that of the scatter matrix Σ is the conventional sample covariance matrix of the covered cases. The algorithm for the MCD is then:

- Explicitly or implicitly consider all partitions of the cases into C covered and $n-C$ uncovered, and find the mean vector and covariance matrix of the C covered cases, and also the determinant of the covariance matrix.

- The exact MCD estimator is given by the subset for which this determinant is a minimum.

The MVE criterion is fitted by:-

- Consider all partitions, exactly as in the other two methods. For each partition, find the ellipsoid of minimum volume that covers the C trial cases.

- The exact MVE estimator is based on the subset of size C for which the volume of the smallest covering ellipsoid is a minimum.

Finding the minimum volume ellipsoid covering the C cases is however a much more laborious computation than is involved in either the MCD or the LTS problems — see Cook, Hawkins and Weisberg (1993) for a discussion of the problem and its connection to D-optimal design and Titterington (1975) for the algorithm used. Woodruff and Rocke (1994) strongly suggest that the MCD is better than the MVE and give some empirical evidence. Rousseeuw and Van Driessen (1997, p. 2) also state that the MVE should be superseded by the MCD.

2 Necessary conditions and the feasible solution algorithms

Since each of these high breakdown methods involves an explicit or implicit generation of all possible subsets of C cases, all are combinatorially hard. Agulló (1996, 1997) gives branch and bound algorithms for exact computation of LMS and the MVE, and these

algorithms can be modified to compute the MCD, LTS, and LTA. Rousseeuw and Van Driessen (1997) state that the branch and bound algorithm for the MVE can be applied for $n \leq 100$ and $p \leq 5$, but all current resampling methods for problems of greater than trivial size involve some method of sampling to get an approximate solution. The family of feasible solution algorithms proposed by Hawkins (1993a, 1993b, 1993c, 1994) all note that the subset of covered cases giving rise to the exact optimum must satisfy the necessary condition:

The criterion cannot be improved by exchanging any of the currently uncovered cases for any of the currently covered cases.

The feasible solution algorithms then consist of taking candidate subsets of size C at random from the n cases, evaluating the criterion (residual sum of squares for LTS, determinant of covariance matrix for MCD and volume of covering ellipsoid for MVE) and seeing if the criterion can be improved by a case swap. If it can, then the swap that leads to the greatest improvement is made and the test is repeated on the new candidate subset. When the current subset can not be improved by a case swap, it then satisfies the necessary condition, and becomes a 'feasible solution'. The best feasible solution found in a suitably large number of random starts is then taken as the estimate of the global optimum. Rocke and Woodruff (1996, p. 1048) incorporate the FSA in their algorithm.

2.1 Computational complexity of the algorithms.

In the processing of a candidate subset, LTS and MCD both involve the computation of a mean vector and a covariance matrix. This is an $O(Cp^2)$ operation. The subsequent

fitting of the regression in LTS and the evaluation of the determinant in MCD are both $O(p^3)$ operations. There are $C(n-C)$ swaps to consider. By judicious precomputing requiring $O(np)$ time, the amount of work involved in evaluating a MCD swap can be done in $O(1)$ time for many swaps, needing a worst-case $O(p)$ time, so the complexity of the swap phase for MCD is $O(C\{n-C\}p)$. In LTS, evaluating a case swap requires $O(p)$ computation. Thus in evaluating a trial subset, both LTS and MCD require computations of order

$$O(Cp^2) + O(p^3) + O(C\{n-C\}p).$$

The MVE involves fitting a covering ellipsoid to the C covered cases. This is done iteratively, and each step of the initial fitting involves $O(Cp^2)+O(p^3)$ calculations. Evaluation of a swap involves a similar amount of computation, so the full evaluation of a subset involves a computational complexity that is some multiple of

$$O(C^2\{n-C\}p^2) + O(C\{n-C\}p^3)$$

where the multiple is some increasing function of n and p. A referee pointed out that the Titterington (1975) algorithm for finding the smallest covering ellipsoid can be replaced by the Welzl (1991) algorithm. This algorithm has expected $O(n)$ time, but the dependence on p is exponential.

As even this brief sketch suggests, the feasible solution algorithm for MVE is very slow, compared to those for LTS and MCD.

All three of these algorithms involve a computation to be carried out on all pairs of cases with one in and one out of the covered set. If both C and $n-C$ are $O(n)$ while $p = o(n)$, then the overall computational complexity is dominated by the evaluation of

the swaps, and is $O(n^2p)$ for LTS and MCD, and $O(n^3p)$ for MVE. In practical terms, there is a decision to be made in considering the swaps – should one

1. Accept the first swap that leads to an improvement in the criterion,

2. Search for the swap that leads to the greatest improvement, or

3. Search for a while, stopping at the first subset has been found that gives at least some minimum threshold of improvement.

In complexity terms, there is no difference between these three approaches (since in all of them to establish that a trial solution is feasible you need to evaluate all possible swaps), but in practical terms 3 is a clear winner since it leads to many fewer inner iterations than 1, and mostly much faster inner iterations than 2. However the overall complexity of $O(n^2p)$ or $O(n^3p)$ remains, and means that the feasible solution algorithms based on the case-swap necessary condition cannot be used for very large data sets.

2.2 An easier and faster necessary condition.

All three criteria also have another necessary condition for the optimum.

- In the case of LTS, each of the C covered cases has a smaller squared residual than any of the $n - C$ uncovered cases.

- In the case of MCD and MVE, if we calculated the Mahalanobis distance of each case from the location vector using the scatter matrix, each covered case must have smaller distance than any uncovered case.

This condition has been considered by Ruppert (1992, p. 258) for LMS and LTS, and the condition also holds for LTA. Rousseeuw and Van Driessen (1997) prove that the condition holds for the MCD estimator using results from Grübel (1988).

To avoid lengthy repetition, we will use the term 'case distance' in the material that follows to refer to a case's squared residual in the LTS problem, and the case's Mahalanobis distance in the MCD or MVE problems.

These necessary conditions can also be made the basis for a feasible solution approach:-

- Fit the criterion to the current trial subset of C cases, and evaluate the distance of each case from the solution.
- If the distances of all covered cases are smaller than the distances of all uncovered cases, then the current subset and its solution satisfy the weaker necessary condition.
- If however there are uncovered cases that fit the current solution better than do covered cases, replace the current trial subset with the C cases that best fit the current solution – that is, those with the C smallest distances.

The computational complexities of the first parts of this scheme are the same as the case-swapping feasible solution algorithm. But instead of evaluating all $C(n-C)$ possible swaps for the subset-refinement phase, we merely have to find the C smallest distances. This involves finding the Cth order statistic of the distances, which is an $O(n)$ operation, and another $O(n)$ calculation to find which cases have the C smallest distances and

check whether they are the trial subset, so the computational complexity of the inner case replacement phase is $O(np)$. If a sort is used to find the C smallest distances, as in the current implementation of the FSA, then the complexity is $O(np\ \log(n))$.

Lemma. Suppose that both the weak and strong necessary conditions hold for the criterion (eg LMS, LTA, LTS, MCD, or MVE). Then the stronger necessary condition is not satisfied unless the weaker necessary condition is satisfied.

Proof. Let J_k be the index set of C cases covered in the kth step of the case swapping iteration. Let

$$d_{(1)}(J_k) \leq d_{(2)}(J_k) \leq ... \leq d_{(n)}(J_k)$$

denote the ordered case distances when the parameters are computed from the subset J_k. We need to show that a swap can be made which reduces the criterion if J_k does not correspond to the cases with the C smallest case distances. Hence we assume that $J_k = \{i_1, ..., i_C\}$ where

$$d_{i_1} \leq d_{i_2} \leq ... \leq d_{i_C}$$

and $d_{i_C} > d_{(C)}$. Let i_s denote the case not in J_k that has the smallest case distance. Thus $d_{i_s} \leq d_{(C)}$ and if we apply the weak test on cases $i_1, ..., i_C, i_s$ then cases i_C and i_s are swapped. (Let the new sample size $\tilde{n} = C + 1$, and the new coverage size $\tilde{C} = \tilde{n} - 1 = C$. If, for example, the $C + 1$ case distances satisfy

$$d_{i_1} < ... < d_{i_s} < ... < d_{i_{C-1}} < d_{i_C},$$

then the weak test will use every case except i_C.) In other words, since the weak necessary condition holds, a swap of case i_C and case i_s will reduce the criterion. QED

It is a drawback of this necessary condition that it is weaker than the case-swapping necessary condition, so that there may be solutions that pass this screen, but fail the case-swapping necessary condition. We can however synthesize the pair of conditions to try to get the best of both worlds — use the weak condition as a preliminary screen to move quickly into the general area of the optimal subset, and then use the stronger condition to refine the subsets that pass the weaker condition.

Another hybrid possibility is to use the weaker condition to move from an arbitrary starting subset to one in which the weak condition is satisfied, and then to do some checking for possible improvements by case swapping. If instead of evaluating all possible case swaps, however, this second phase evaluates only $O(n)$ of the $C(n-C)$ possible swaps, the overall $O(np)$ complexity of LTS and MCD is retained.

3 Some computational experience.

We have implemented combined algorithms for LTS and MCD in which each candidate subset is first checked using the weak condition. Once it satisfies the weak condition, then it is tested with the strong case-swap condition. If an improvement is made in the strong test, then the weak test is applied again, and the process continued until the algorithm reaches a subset satisfying both conditions. The hope for improvement using the weak condition rests on the possibility that it might reduce the number of swaps in the strong test by cheap refinement of initial poor estimates.

The MVE estimator is substantially harder than either MCD or LTS. This is because the fitting operation involves an iteration with the calculation and inversion of a covari-

ance matrix in its core. This calculation also has to be carried out at least partially to evaluate case swaps. Thus using any case swapping is prohibitively expensive for moderate to large samples. We have implemented an MVE code that uses just the first weak condition. The quality of the estimates returned by it will, of course, be inferior to that given by the stronger necessary condition, but the substantial savings in computer time can be used to good effect to explore more initial subsets, so the end result may still be an improvement.

As an illustrative example, we used the Boston housing data of Harrison and Rubinfeld (1978), omitting the fourth predictor (a binary indicating adjacency to the Charles River) since it leads to degeneracy. This data set was chosen, not because of its previous use in the high breakdown literature, but because it is a real, widely-accessible, moderately large data set.

We computed the MCD for the 12 remaining predictors. This gave us $n = 506$ cases, with $p = 12$. We ran the data using the older FSA codes from our Web site, and the new code. We know of no reason to think that the new procedure would produce good solutions either more often or less often than the old solution, so the primary basis for comparison is the number of swaps made in going from an initial random sample to the final feasible solution, and in the total execution time.

In our runs, we selected C to be the standard default – the value providing the maximum breakdown for the estimators; this is 260 for LTS and 259 for MCD. This is not necessarily a good choice – see for example Cook and Hawkins (1990). A more careful choice based on the perceived maximum number of outliers that could plausibly be present in the data set will provide greater statistical efficiency, as well as possibly

faster execution. Each run used 1,000 random starting subsets.

MCD: The criterion values of the three leading feasible solutions were

New algorithm 22.627 22.631 27.335

Old algorithm 22.627 22.631 22.759

In going from each random starting subset to the local feasible solution, the new algorithm used an average of 32 cycles of refinement using the weak condition, and 2.6 cycles using the stronger condition. The old algorithm, using only the strong condition, used an average of 139 cycles of pairwise swapping.

Clearly the preliminary screen using the weak condition has been highly successful in reducing the number of cycles needed using the much more expensive case swap phase. This greater success is reflected in the execution times. The old code required 65 minutes on a HP 712/60 workstation, compared with 17 minutes for the new code — a roughly four-fold speedup.

LTS: The criterion values returned by the three leading feasible solutions using each algorithm were:-

New algorithm 236.7 249.7 251.6

Old algorithm 222.0 222.5 232.2

The average number of cycles of subset refinement for the new code was 59 for the weak condition, and 30 for the strong condition. The average number of strong condition cycles for the old code was 135. Here too, use of the computationally fast weak condition has led to a very effective screening, dramatically reducing the number of cycles of the case-swapping phase.

The computation times reflect this saving even more clearly than they did in the MCD case. The execution time for the new code was 42 minutes, while the old code was 10 times slower, requiring 420 minutes. The reason for the larger saving than was seen with MCD is that evaluating a possible case swap for the LTS criterion is always an $O(p)$ computation, while bounding allows many of the potential MCD swaps to be rejected without evaluation. We note that the three feasible solutions returned by the old code had lower criterion values than those found by the two-condition algorithm, but hesitate to find a general truth in this.

MVE: The new code for the MVE involves only the weak condition. We ran it with 50 random starts, getting feasible solution criterion values (the common log of the volume of the covering ellipsoid) of 13.886, 13.889 and 13.961. Execution time was 247 minutes, or about 5 minutes per random start.

The old code with the strong necessary condition required nearly *18 hours* per random start – graphic evidence of the earlier comment that the case-swap necessary condition is computationally prohibitive for MVE problems of any but very modest size. However the quality of the solutions was much higher – the very first random start yielded a criterion value of 13.192 and the second a value of 11.209. This covering ellipsoid's volume was smaller by a factor of $10^{2.7}$ than the best solution found using just the weak condition. The hope for reasonable results with the MVE then rests on the possibility that the weak condition, by evaluating many more initial subsets, may be able to match the results of the case-swap condition by sheer repetition. Since 200 weak-condition samples can be run in the same time as one strong-condition sample, this seems a reasonable prospect.

The feasible solution algorithms for MCD, MVE, LMS, LTS, and LTA are at the

following website (go to the software icon).

http://www.stat.umn.edu

Rousseeuw and Van Driessen's algorithm for the MCD is at the website below.

http://win-www.uia.ac.be/u/statis/

Since these algorithms may not yield consistent estimators if the number of starts is fixed, we suggest that at least $\max(500, n/10)$ starts be used.

4 Conclusions

The feasible solution algorithms have shown themselves very effective for high breakdown estimation in modestly-sized data sets. For very large data sets however the $O(n^2p)$ computational complexity of the case-swap necessary condition makes them too slow to be competitive. The weaker necessary condition that the covered cases have the smallest distances from the putative solution has only $O(np)$ complexity, and so runs much faster in large data sets, though its approximations have solutions that are not as good. Combining the two conditions gives many of the benefits of both – the faster weak necessary condition dramatically reduces the number of swaps made in the case-swap necessary condition, and leads to much faster execution.

Croux and Haesbroeck (1997) suggest keeping track of the best elemental subset (a subset of size $p+1$ for location/scatter and of size p for regression) containing case i and then averaging the corresponding parameter estimates over the half set with the smallest criterion values. Their simulations indicated that the averaged estimator may outperform the exact MVE estimator in the univariate case and $p = 2$ case for small n.

This idea may be worth investigating.

Acknowledgements

The author is grateful to David Rocke, Arny Stromberg and Carlos Lopez for highlighting some of the problems with the feasible solution algorithms in data sets whose size is in the thousands. The referees made a number of helpful suggestions for improving the article. The work reported here was supported by the National Science Foundation under grant DMS 9505440.

5 REFERENCES

Agulló, J., Exact iterative computation of the multivariate minimum volume ellipsoid estimator with a branch and bound algorithm, in: A. Prat (Ed.), *Proceedings in Computational Statistics,* (Physica-Verlag, Heidelberg, 1996) 175-180.

Agulló, J., Exact algorithms to compute the least median of squares estimate in multiple linear regression, in: Y. Dodge (Ed.), L_1-*Statistical Procedures and Related Topics,* (Institute of Mathematical Statistics, Hayward, CA, 1997) 133-146.

Cook, R. D. and D.M. Hawkins, discussion of 'Unmasking multivariate outliers and leverage points', *Journal of the American Statistical Association,* **85**, (1990) 640-644.

Cook, R. D., Hawkins, D. M. and S. Weisberg, Exact computation of the robust multivariate minimum volume ellipsoid estimator, *Statistics and Probability Letters,* **16**, (1993) 213-218.

Croux, C., and G. Haesbroeck, An easy way to increase the finite-sample efficiency

of the resampled minimum volume ellipsoid estimator, *Computational Statistics and Data Analysis,* **25**, (1997) 125-141.

Grübel, R., A minimal characterization of the covariance matrix, *Metrika,* **35**, (1988) 49-52.

Harrison, D. and D.L. Rubinfeld, Hedonic prices and the demand for clean air, *Journal of Environmental Economics and Management,* **5**, (1978) 81-102.

Hawkins, D.M., A feasible solution algorithm for the minimum volume ellipsoid estimator in multivariate data, *Computational Statistics* **9**, (1993a) 95-107.

Hawkins, D.M., The feasible solution algorithm for least trimmed squares regression, *Computational Statistics and Data Analysis,* **17**, (1993b) 185-196.

Hawkins, D.M., The feasible set algorithm for least median of squares regression, *Computational Statistics and Data Analysis*, (1993c) **16**, 81-101.

Hawkins, D.M., The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data, *Computational Statistics and Data Analysis*, **17**, (1994) 197-210.

Rocke, D.M. and D.L. Woodruff, D.L., Identification of outliers in multivariate data, *Journal of the American Statistical Association,* **91**, (1996) 1047-1061.

Rousseeuw, P. J. and B.C. van Zomeren, Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association,* **85**, (1990) 633-639.

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

Rousseeuw, P.J., and K. Van Driessen, A fast algorithm for the minimum covariance

determinant estimator, Technical report, (1997).

Ruppert, D., Computing S-estimators for regression and multivariate location/dispersion, *Journal of Computational and Graphical Statistics,* **1**, (1992) 253-270.

Titterington, D. M., Optimal design: some geometrical aspects of *D*-optimality, *Biometrika,* **62**, (1975) 313-320.

Welzl, E., Smallest enclosing disks (balls and ellipsoids), in: H. Maurer (Ed.), *New Results and New Trends in Computer Science, Lecture Notes in Computer Science, 555*, (Springer-Verlag, Berlin, 1991) 359-370.

Woodruff, D.L. and D.M. Rocke, Computable robust estimation of multivariate location and shape in high dimension using compound estimators, *Journal of the American Statistical Association,* **89**, (1994) 888-896.