Computational Connections Between Robust Multivariate Analysis and Clustering

David M. Rocke[1] and David L. Woodruff[2]

[1] Department of Applied Science, University of California at Davis, Davis, CA 95616, USA
[2] Graduate School of Management, University of California at Davis, Davis, CA 95616, USA

1 Introduction

In this paper we examine some of the relationships between two important optimization problems that arise in statistics: robust estimation of multivariate location and shape parameters and maximum likelihood assignment of multivariate data to clusters. We offer a synthesis and generalization of computational methods reported in the literature. These connections are important because they can be exploited to support effective robust analysis of large data sets.

Recognition of the connections between estimators for clusters and outliers immediately yields one important result that is demonstrated by Rocke and Woodruff (2002); namely, the ability to detect outliers can be improved a great deal using a combined perspective from outlier detection and cluster identification. One can achieve practical breakdown values that approach the theoretical limits by using algorithms for both problems. It turns out that many configurations of outliers that are hard to detect using robust estimators are easily detected using clustering algorithms. Conversely, many configurations of small clusters that could be considered outliers are easily distinguished from the main population using robust estimators even though clustering algorithms fail.

There are assumed to be n data points in R^p and we may refer to them sometimes as a set of column vectors, $\{\boldsymbol{x}_i\} = \{\boldsymbol{x}_i | i = 1, 2, \ldots, n\}$. We are concerned here primarily with combinatorial estimators and restrict ourselves to those that are *affine equivariant*.

2 Robust Estimation and Clustering

2.1 Robust Estimation and Outlier Detection

The MCD was defined by Rousseeuw as that sample of size h that results in the lowest covariance determinant. Usually, h is chosen as the "half-sample size" $\lfloor (n + p + 1)/2 \rfloor$, which is the choice that maximizes the breakdown (Rousseeuw and Leroy 1987; Lopuhaä and Rousseeuw 1991). We define the MCD formally as the solution to the problem of selecting a set $H \subset N$ of size h so as to minimize $|W|$, where $N = \{1, 2, \ldots, n\}$ and where

$$W = \sum_{j \in H} (x_j - \bar{x}_H)(x_j - \bar{x}_H)^T,$$

and

$$\bar{x}_H = h^{-1} \sum_{j \in H} x_j .$$

The location and shape estimates are then x_H and $n^{-1}W$.

Rigorous analysis of the theoretical computational complexity of the MCD by Bernholt and Fischer (2001) implies that the problem must be addressed using heuristic algorithms that search for a good solution. The difficulty in constructing such algorithms is that if points that are "outliers" are included in H, they will distort the estimates of shape and location so as to make it difficult to detect that they are outlying.

An analysis of difficult forms is provided by Rocke and Woodruff (1996). An extremely plausible, yet still difficult, form of contamination is referred to as *shift outliers* (see Hawkins 1980 page 104). Shift outliers have the same shape and size as the main population, but a different location.

2.2 Maximum Likelihood Clusters

The problem of finding the maximum likelihood assignment of data points to clusters is similar, but the literature has developed separately for the most part. There is a very large literature devoted to clustering when there is a metric known in advance. However, in order to retain affine equivariance, we rely on the smaller but growing literature related to using metrics gleaned from the data itself.

A thorough examination of criteria based on the likelihood is given by Banfield and Raftery (1993). Their paper proposes a number of criteria that maximize the likelihood conditional on a clustering, under a number of assumptions about the relative sizes and shapes of the clusters. A popular method is to solve problem (MINW), (Friedman and Rubin 1967), which finds the clustering that minimizes the determinant of the pooled covariance $|W|$ where

$$W = \sum_{i=1}^{g} W_i ,$$

$$W_i = \sum_{j \in H_i} (x_j - \bar{x}_H)(x_j - \bar{x}_H)^T ,$$

and where $H_1, H_2, \ldots, H_g$ is a partition of N. This corresponds to maximum classification likelihood under the assumption that the data vectors are multivariate normal with cluster covariances such that $\Sigma_1 = \cdots = \Sigma_g$.

An objective that is similar from a computational standpoint is

$$\sum_{i=1}^{g} h_i \log \left| \frac{W_i}{h_i} \right| ,$$

where $h_i = |H_i|$. The minimum corresponds to a maximum classification likelihood under the assumption of heterogeneous covariance matrices. It was first given by Scott and Symons (1971) and adjusted by Banfield and Raftery (1993). Call the problem with this objective function (MIND). In order to avoid singularities, as a practical matter a parameter $h_{\mathrm{min}} > p$ must be given for the minimum number of points assigned to each cluster. Difficult forms are discussed by Coleman and Woodruff (2000).

3 Neighborhoods

Although most were not written using the terminology of local search, the proposals in the literature for algorithms for robust estimation and cluster finding can be cast in that framework. This facilitates synthesis and some generalization. Local search is defined relative to an evaluation function for an optimization problem and a neighborhood structure.

3.1 Local Search

We define the generic hard problem to which local search algorithms are applied as

$$\begin{array}{c} \min_{\tau} f(\tau) \quad \text{(P)} \\ \text{Subject to: } \tau \in \Xi \end{array}$$

where the set Ξ is intended to summarize the constraints placed on the decision vector τ. Solution vectors that (do not) satisfy the constraints are said to be *(in)feasible*. The constrained optimization literature refers to all data for the problem—the data that specifies the *objective function* $f(\cdot)$ and Ξ—as (P). It is easy to see that the MCD, MINW and MIND estimators can all be stated in this form.

Neighborhoods are based on *moves* from one solution to another. All of the solutions that can be reached from a given solution in one move are said to be in the neighborhood of the solution. We use the notation $\mathcal{N}(\tau)$ to indicate the set of solutions that are neighbors of a solution τ.

Simplifying things somewhat to ease exposition, we can define *Steepest descent* as a general purpose procedure that begins with an initial solution, $\boldsymbol{\tau}^0$, and selects solutions at iteration $k > 0$ using the relation

$$\boldsymbol{\tau}^k = \underset{\boldsymbol{\tau} \in \mathcal{N}(\boldsymbol{\tau}^{k-1})}{\operatorname{argmin}} f(\boldsymbol{\tau})$$

(a tie breaking rule may be needed). The algorithm terminates when there are no lower objective function value solutions in the neighborhood of the current solution. Such a solution is referred to as a *local minimum.* A *first-improving* descent is similar but requires more notation and proceeds through an ordered neighborhood until an improving move is found which is then immediate made. After a move, the traversal of the neighborhood continues using the ordering (or some approximation to it). One possibility (for either steepest descent of first improving) is to repeat the descent many times, restarted at a random starting point each time the algorithm hits a local minimum.

Many general purpose optimization algorithms are based on combinatorial steepest descent (e.g., simulated annealing). An application of some of these methods to computation of the MVE is given by Woodruff and Rocke (1993).

3.2 Exchange Neighborhoods

For the MCD a sensible neighborhood is one where a point in H is exchanged with one not currently in H. We refer to this as an *exchange* or *swap* neighborhood. For (MINW) and (MIND) the corresponding neighborhood is one where a point is moved from one group to another. For solutions where the size constraints are not binding, the neighborhood has $(g-1)n$ solutions. There are fewer neighbors of solutions for which one or more of the size constraints is binding.

3.3 Constructive Neighborhoods

The swap neighborhoods can be classified as *transition* neighborhoods that move from one full solution to another. In contrast, *constructive* neighborhoods move from a partially specified solution to a more complete solution and *destructive* neighborhoods are the opposite. For the problems under consideration here, a constructive neighborhood would correspond to moves that transform a solution with some but not all data points assigned to one with one (or more) additional data point(s) assigned; a destructive neighborhood would correspond to moves that unassign one or more data points. So-called *greedy* algorithms can then be cast as steepest descent with a constructive neighborhood.

A constructive neighborhood for the MCD "surrounds" a set of points, H, that has between $p+1$ and h members. A subset of H (typically either empty or all of H) is required to be included in all neighbors; call this subset $\tilde{H}$. Finally, a subset of N is eligible for inclusion in any of the neighbors (typically all of N); call it $\tilde{N}$. Given a set H, moves to a new set H' must be such that all of the points in $\tilde{H}$ are in H' plus one or more points from $\tilde{N}$. This is summarized as follows:

p	Dimension of the data (given)
n	Number of data points (given)
h_{min}	Minimum cluster size (given)
N	Index set $1, \ldots, n$
H	Subset of N (currently) estimated to be in the majority population
$\tilde{N}$	Subset of N eligible for inclusion in the H during the next iteration
$\tilde{H}$	Subset of N required to be included in H during the next iteration

Algorithms based on steepest descent must specify the method of constructing an initial solution, an evaluation function, $\hat{f}(\cdot)$, and perhaps also a *refresh* period, ψ, that controls how many moves are allowed before corrections are made for the fact that $\hat{f}(\cdot)$ is based on an approximation to the current state of the neighborhood. Some of the algorithms in the literature have started with an initial set as large as a half-sample (e.g., Hawkins 1994), but many use a starting set of size $p+1$, and we have conducted simulation studies confirming this choice of size for computational reasons. There are three affine equivariant possibilities reported in the literature for picking a initial sets H to begin the descent process.

- Select $p+1$ points at random (RAND).
- Select $p+1$ points that are "good" based on a heuristic (HEUR).
- Select the $p+1$ points that have lowest Mahalanobis distance from the result of the last full solution constructed (WALK).

Clearly, use of the last choice results in an iterative algorithm that can be terminated either after some number of constructions, or when it reaches a fixed point. Refer to such an iterative algorithm as a *walking* algorithm. Note that K-means algorithms are generally (non-affine-equivariant) walking algorithms using this convention. Such algorithms are common in clustering, but apparently were first used in calculating the MCD by Hawkins (1999) and Rousseeuw and Van Driessen (1999) independently. A walking algorithm must be started using either RAND or HEUR.

For the estimators of interest to us, there are two move evaluation function commonly in use. One is based on Mahalanobis distances from the mean of points in H using the covariance matrix of points in H as the metric;

Algorithm	$\hat{f}(\cdot)$	$\tilde{H}$	ψ	Start/Restart	Walking?
Fast-MCD Rousseeuw 1999	MAHAL	$\emptyset$	∞	HEUR	Yes
Forward Atkinson 1994	UPDATE	H	1	HEUR	No
Improved FSA Hawkins 1999	UPDATE	$\emptyset$	∞	RAND	Yes
Hadi Hadi 1992	UPDATE	$\emptyset$	1	HEUR	No
Multout Rocke and Woodruff 1996	UPDATE	H	1	HEUR	No

Table 1. Summary of Affine Equivariant Constructive Neighborhoods for MVE/MCD Algorithms Reported in the Literature as cast in a Local Search Framework

i.e., select the point(s) $i \in \tilde{N}$ that minimize(s) $d^2_{S_H}(x_i, \bar{x}_H)$ where $d^2_{S_H}$ is the Mahalanobis distance under covariance matrix S_H, and $\bar{x}_H$ is the mean of the points in H. Call this evaluation method MAHAL. We indicate that multiple points might be selected, because if the refresh period is infinite, then one selects the h or $h - p - 1$ (depending on the makeup of $\tilde{H}$) points that have lowest distance and all of the moves can be made at once. If the refresh period is one, then after each point was added, the values of $\bar{x}_H$ and S_H are updated. An alternative to MAHAL is to use update formulas to predict the effect of a move. The refresh period specifies how often the mean and covariance are recomputed from the current set H. Call this method UPDATE.

Table 1 gives a summary of constructive neighborhoods that have been reported in the literature for the MCD (and/or the MVE). In all cases, $\tilde{N} = N \setminus \tilde{H}$. Of course, this table provides only a summary of the constructive neighborhood used and not a complete description of the algorithms. The start-restart heuristic used by Fast-MCD is as follows: do a large number of random starts each followed by walking that is terminated after two constructive descents; the best ten results are then pursued with a convergent walking algorithm. The heuristic reported for use with Forward is to select the $p + 1$ points closest to the mean of all of the data under the metric for the data. Hadi suggest the use of a non-affine equivariant starting heuristic, but his algorithm is otherwise affine equivariant. Multout uses the result of a lengthy search based on swap neighborhoods to find a starting point for a constructive descent that is very similar to Forward. Many of the methods are being updated so that this table represents only the state of affairs at the time of this writing. Our main goal is to demonstrate that the local search framework is very useful as a means of synthesizing the evolving methods. This notation generalizes to the clustering problems as shown in Rocke and Woodruff (2002).

4 Conclusions

In this paper we have drawn on concepts from local search to demonstrate strong connections between algorithms for two important problems in statistics: robust estimation of multivariate location and shape parameters and maximum likelihood assignment of multivariate data to clusters. We provided a synthesis and generalization of computational methods reported in the literature. These connections are important because they can be exploited to support effective robust analysis of large data sets.

References

Atkinson, A. C. (1994) "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association*, **89**, 1329–1339.

Banfield, J.D. and A.E. Raftery (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, **49**, 803-821

Bernholt, T. and P. Fisher (2001), "The Complexity of Computing the MCD-Estimator," Technical Report, Lehrstuhl Informatik 2, Universität Dortmund, Germany.

Coleman, D.A. and D.L. Woodruff (2000), "Cluster Analysis for Large Data Sets: Efficient Algorithms for Maximizing the Mixture Likelihood," *Journal of Computational and Graphical Statistics*, to appear.

Friedman, H.P. and J. Rubin (1967), "On some Invariant Criteria for Grouping Data", *Journal of American Statistical Association,* **62***, 1159-1178.*

Hadi, A.S. (1992) "Identifying Multiple Outliers in Multivariate Data," *Journal of the Royal Statistical Society, Series B*, 54, 761-771.

Hawkins, D.M., (1980) *The Identification of Outliers*, Chapman and Hall, London.

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

Hennig, C. (1998) "Clustering and Outlier Identification: Fixed Point Cluster Analysis," Rizzi, A., Vichi, M. and Bock, H.H. (eds.): Advances in Data Science and Classification, Springer, Berlin, 37–42

Lopuhaä, H. P. and Rousseeuw, P. J. (1991) "Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices," *Annals of Statistics*, **19**, 229–248.

Rocke, D.M., and D.L. Woodruff (1996), "Identification of Outliers in Multivariate Data" *Journal of the American Statistical Association*, 91, 1047-1061.

Rocke, D.M., and D.L. Woodruff (2002), "Multivariate Outlier Detection and Cluster Identification," submitted for publication.

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

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

Scott A.J. and Symons M.J. (1971), "Clustering based on Likelihood Ratio Criteria," *Biometrics*, **27**, 387-97.

Woodruff, D. L., and Rocke D. M. (1993) "Heuristic Search Algorithms for the Minimum Volume Ellipsoid," *Journal of Computational and Graphical Statistics*, **2**, 69–95.

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