A Synthesis of Outlier Detection and Cluster Identification

David M. Rocke
Center for Image Processing and Integrated Computing
and
David L. Woodruff
Graduate School of Management

University of California, Davis
Davis CA 95616

September 2, 1999

Abstract

We examine relationships between the problem of robust estimation of multivariate location and shape and the problem of maximum likelihood assignment of multivariate data to clusters and 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 we demonstrate in this paper; 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. Computational results are reported that demonstrate the effectiveness of this approach.

1 Introduction

Topics related to multivariate analysis gain importance as scientists, engineers and business managers grapple with the exploding availability of data. Under the rubric *data mining* many ad hoc methods provide quick results but lack the solid foundation provided by statistical models. However, there are numerous places where statistical models can be employed in data mining (see e.g. Glymour et al 1997).

For example, statistical research has a lot to offer in the areas of robust estimation of multivariate location and shape parameters and maximum likelihood assignment of multivariate data to clusters. In this paper we examine some of the relationships between these two problems and 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. For example, we introduce a partial assignment estimator that is needed in order to make use of subsampling and to find clusters in data sets that have extraneous data.

Recognition of the connections between estimators for clusters and outliers immediately yields one important result that we demonstrate in this paper; 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.

It is useful to divide estimators for such problems into two classes that we will call *combinatorial* and *smooth*. Smooth estimators are computed in the parameter space, which is continuous. Combinatorial estimators, on the other hand, work in the combinatorial space of subsets of the set of data points. These estimators are often used to find starting points for iterative algorithms applied to the smooth estimators, or may be used on their own. There are assumed to be n data points in R^p and we may refer to them sometimes as a set of column vectors, $\{\boldsymbol{x}_i\}$. We are concerned here primarily with combinatorial estimators and restrict ourselves to those that are *affine equivariant*.

A location estimator $\boldsymbol{t}_n \in \mathcal{R}^p$ is affine equivariant if and only if for any vector $\boldsymbol{b} \in \mathcal{R}^p$ and any non-singular $p \times p$ matrix $\boldsymbol{A}$

$$\boldsymbol{t}_n(\{\boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b}\}) = \boldsymbol{A}\boldsymbol{t}_n(\{\boldsymbol{x}_i\}) + \boldsymbol{b}. \tag{1}$$

A shape estimator $\boldsymbol{C}_n \in \mathrm{PDS}(p)$ is affine equivariant if and only if for any vector $\boldsymbol{b} \in \mathcal{R}^p$ and any non-singular $p \times p$ matrix $\boldsymbol{A}$

$$\boldsymbol{C}_n(\{\boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b}\}) = \boldsymbol{A}\boldsymbol{C}_n(\{\boldsymbol{x}_i\})\boldsymbol{A}^T \tag{2}$$

Estimators satisfying (1) and (2) transform properly under changes of scale and rotation so that clusterings or identification of outliers do not change under these operations. Perhaps more important, affine equivariant estimators are exactly those for which distances are completely data determined and do not depend on any arbitrary prior metric (such as the Euclidean metric). Although there may be some few instances in which theory or prior knowledge determines the correct metric for distances between data points, we would argue that in almost all cases these should be data determined. Some care must be taken in designing methods, algorithms, and their implementations to maintain this property.

Effective affine equivariant estimators for outlier detection and for finding clusters are hard to compute. Many theoretical estimators are defined in terms of the minima of functions over subsets or partitions, but there are no known algorithms to find or verify such minima in less than geologic amounts of time for even modest sized data sets. Hence, the algorithms that search for approximations to the theoretic statistics are very important. Because the performance of the estimator is intimately tied to the performance of the search algorithm, and since the second best answer if often far worse than the best (correct) answer, we argue that, in essence, the algorithms are the estimators (Rocke 1998; Seidel, Mosler, and Alker 1999). To study the properties of

the estimates, we must separate the statistic that is the *objective function* from the algorithms. Furthermore, we must decompose the algorithms into their constituent parts to identify effective methods.

The next section defines the theoretical estimators. In §3 we synthesize and generalize some of the algorithms and literature by viewing these methods as variations of local search, and studying the neighborhood structures. In §4 we provide a new estimator and some analysis that compares the statistics (or objective functions) and the neighborhood structures. In §5 we explore computational boundary between data classes where one is better off using an outlier model and where one is better using a two cluster model. Furthermore, we show that by using both types of algorithms, the effective breakdown point for outlier detection can be extended to be near theoretical limits. The final section offers concluding remarks and directions for further research.

2 Estimation and Clustering

2.1 Robust Estimation and Outlier Detection

When there is thought to be one main population contributing data points, but when one must guard against the possibility of one or more "contaminating" or "outlying" populations, the problem becomes one of *robust estimation.* In order to remove the outliers, one must estimate the parameters of the main population, but the presence of the outliers can make this difficult. We will follow the bulk of the literature by assuming the main population obeys an elliptical distribution, with estimators calibrated to the multivariate normal.

Smooth estimators for the determination of robust location and shape, such as maximum likelihood and M-estimators (Campbell 1980, 1982; Huber 1981; Kent and Tyler 1991; Lopuhaä 1992; Maronna 1976; Rocke 1996; Tyler 1983, 1988, 1991), and S-estimators (Davies 1987; Hampel et al. 1986; Lopuhaä 1989; Rousseeuw and Leroy 1987) can be computed with a straightforward iteration from a good starting point provided by a combinatorial estimator (Rocke and Woodruff 1993). Combinatorial estimators, such as the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) estimators of Rousseeuw (1985; Hampel et al. 1986; Rousseeuw and Leroy 1987), have been addressed with random search (Rousseeuw and Leroy 1987), steepest descent with random restarts (Hawkins 1993, 1994), and heuristic search optimization efforts (Woodruff and Rocke 1993, 1994). Sequential point addition estimators have been defined algorithmically by Atkinson (1994; Atkinson and Mulira 1993), Hadi (1992), Rousseeuw and Van Driessen (1999), and Hawkins (1999).

It was originally thought that the MVE would be preferable for computational reasons (see Rousseeuw and Van Zomeren 1990), even though the MCD has greater asymptotic efficiency. This was based on the notion that MVE algorithms would make use of random elemental subsets, of which there are much fewer than random half samples. Woodruff and Rocke (1993) demonstrated that heuristic search algorithms that use larger subsample sizes perform better. Given this fact, there is no longer any reason to prefer the MVE to the MCD. Simulations done by Woodruff and Rocke (1994) strongly support the contention that the MCD is in fact the better estimator to

use (also see Rousseeuw and Van Driessen 1999).

Given a sample $\{x_i\}$ of n points in dimension p, the MCD is defined as that sample of size h that results in the lowest covariance determinant. Usually, h is chosen as the "half-sample size" $\lfloor (n+p+1)/2 \rfloor$ (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 $J \subset N$ of size h so as to minimize $|W|$, where $N = \{1, 2, \ldots, n\}$ and where

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

and where

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

The location and shape estimates are then x_J and $n^{-1}W$ (or $(n-1)^{-1}W$ if one prefers).

The only known method of exact solution is complete enumeration. In fact, this is the only known method of verifying an exact solution. As a consequence, one must use heuristic algorithms that search for a good solution. The difficulty in constructing such algorithms is that if points that are "outliers" are included in J, they will distort the estimates of shape and location so as to make it difficult to detect that they should be removed.

An analysis of difficult forms is provided by Rocke and Woodruff (1996) making use of the following definitions. Here, as elsewhere, the word metric (a distance function satisfying several properties) is used interchangeably to refer to the distance itself and to the positive definite symmetric matrix that is used to define the quadratic metric

$$d^2(X, Y) = (X - Y)^T \quad ^{-1}(X - Y),$$

which is sometimes referred to as the squared *Mahalanobis* distance.

Definition 1 *Let be a matrix defining a metric. The* size *of the metric is the determinant* | |. *The* shape *of the metric is the equivalence class of metrics* Ξ *such that* /| | = $\Xi/|\Xi|$. *Equivalently, we may identify the shape as the member of the equivalence class with determinant 1; that is,* /| |.

This leads to similar definition of shape and size for samples.

Definition 2 *Let* $\boldsymbol{X}$ *be an* $n \times p$ *matrix representing a sample of* n *points in* R^p. *Let* $\boldsymbol{S} = n^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})'(\boldsymbol{X} - \overline{\boldsymbol{X}})$ *be the sample covariance matrix. The* size *or* scale *of* $\boldsymbol{X}$ *is the determinant* $|\boldsymbol{S}|$ *of its covariance matrix, and the* shape *of* $\boldsymbol{X}$ *is* $\boldsymbol{S}/|\boldsymbol{S}|$. *By extension, we refer to the size and shape of other covariance-like estimators.*

Rocke and Woodruff (1996) go on to show that the class of outliers composed of a single group with the same shape as the main data are well masked in the sense that under the metric of the majority population, the distribution of Mahalanobis distances from the mean of the majority population to outliers overlaps with the distribution of distances to majority population points. This overlap is made more severe as the scale of the outlying cluster is reduced with the worst overlap occurring for a point

mass. Of course, an actual point mass is easy to detect. 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.

So we summarize by noting that based both on theoretical arguments and on simulations, difficult outlier cases are when the data points that are not optimally included in J are themselves from a population that generates either a point mass or a cluster with a covariance structure that is similar to that for the points that are optimally in J. In some cases one is interested in just the main cluster and wants to simply remove the outlying cluster. In other situations, the outlying cluster can be more interesting. What emerges is that a large class of plausible and difficult outlier problems can be modeled as problems of identifying clusters. In general, of course, one cannot know that clusters are present until they have been identified.

A kind of corollary of the observation that placing the outliers in a cluster is a difficult configuration is that *radial* outliers are a particularly easy case. The outliers have the same radial distribution as the main data but lie further away. A robust estimator such as the MCD or M- or S-estimator has no difficulty with this case. This configuration can be seen as each outlying point forming its own cluster. One can think of parameterizing outlier configurations by the number of clusters. With αn radial outliers, there are $\alpha n + 1$ clusters and with shift outliers there are 2 clusters. In general, the more difficult outlier problems have 2 or 3 clusters. The exact level of difficulty depends, of course, on the configuration of the clusters and the algorithm used for estimation.

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. These methods include, for example, the widely used K-means algorithm. However, in order to retain affine equivariance, we rely on the smaller but growing literature related to using metrics gleaned from the data itself.

To do this with maximized likelihood, a statistical model is assumed. In this model the data are generated in two stages: 1) a cluster distribution is randomly selected from among g possibilities and 2) an observation is drawn from the selected cluster. The probability of selecting cluster i is denoted by π_i. Cluster i is assumed to be modeled by a multivariate normal distribution with density ϕ and mean and covariance $\theta_i = (\mu_i, \Sigma_i)$. This results in the mixture likelihood

$$L(\boldsymbol{X};\ \theta) \quad = \quad \prod_{j=1}^{n} \left[\sum_{i=1}^{g} \pi_i \phi(x_j, \theta_i) \right] \tag{3}$$

where $\theta = (\theta_1, \ldots, \theta_g, \pi_1, \ldots, \pi_g)$.

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

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

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

and where $J_1, J_2, \ldots, J_g$ is a partition of N. This corresponds to maximum classification likelihood under the assumption that the data vectors are multivariate normal with homogeneous but otherwise unrestricted cluster covariances. Here the classification likelihood is

$$L(\boldsymbol{X};\ \theta) \quad = \quad \prod_{i=1}^{g} \prod_{j \in J_i} \phi(x_j, \theta_i). \tag{4}$$

which (unlike the mixture likelihood) uniquely assigns each point to a cluster. Algorithms proposed for this minimization include hierarchical agglomeration (Ward 1963, Murtagh and Raftery 1984) and local search (Späth 1985, Coleman et al. 1999).

The striking similarity with the MCD leads immediately to the idea that for data with shift outliers that are far enough from the main population, one of the clusters for the optimal solution of (MINW) for $g = 2$ is the same as the optimal solution to the MCD. We explore such comparisons in §4.2.

Generally, we have no particular reason to expect homogeneous cluster covariance structures, so we can make use of an objective that is similar from a computational standpoint, which is

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

where $n_i = |J_i|$. 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 $H > p$ must be given for the minimum number of points assigned to each cluster.

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

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

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*. 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.

As an example, consider (MIND). We have as data the number of groups, g, a minimum cluster size, H, and n points in R^p with each point given as x_j for $j = 1, \ldots, n$. Our objective is to select indicators of membership $\tau_{ij},\ i = 1, \ldots, g,\ j = 1, \ldots, n$ so as to

$$\text{minimize} \sum_{i=1}^{g} n_i \log \left| \frac{W_i}{n_i} \right| \quad \text{(MIND)}$$

subject to

$$\begin{array}{lcll} W_i & = & \sum_{j=1}^{n} \tau_{ij}(x_j - \bar{x}_{J_i})(x_j - \bar{x}_{J_i})^T & i = 1, \ldots, g \\ \bar{x}_{J_i} & = & \left(\sum_{j=1}^{n} \tau_{ij} x_j\right) / n_i & i = 1, \ldots, g \\ n_i & = & \sum_{j=1}^{n} \tau_{ij} & i = 1, \ldots, g \\ \sum_{i=1}^{g} \tau_{ij} & = & 1 & j = 1, \ldots, n \quad \text{(placement)} \\ \sum_{j=1}^{n} \tau_{ij} & \geq & H & i = 1, \ldots, g \quad \text{(size)} \\ \tau_{ij} & \in & \{0, 1\} & i = 1, \ldots, g, \\ & & & j = 1, \ldots, n. \end{array}$$

The formulations for (MINW) and the (MCD) are very similar.

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.

An evaluation function, $\hat{f}(\boldsymbol{\tau})$, from a set A to R must be specified for move evaluation. The function usually resembles $f(\cdot)$, but may differ in order to take advantage of special knowledge of the problem. Often, $A = \Xi$ and care must be taken to avoid encountering infeasible solutions during the search. In cases where infeasible solutions are allowed during the search, $A \supsetneq \Xi$ and $\hat{f}(\cdot)$ is constructed to encourage discovery of feasible solutions. In many cases, the function would be more precisely written as $\hat{f}(\mathcal{N}, \tau; \tau')$ defined only for $\tau' \in \mathcal{N}(\tau)$, because the functions are typically constructed to compute changes from one solution to the next for a particular neighborhood structure.

Once an evaluation function and neighborhood have been given, special purpose algorithms can be designed or general purpose algorithms can be adapted. *Steepest descent* is a general purpose procedure begins with an initial solution, $\boldsymbol{\tau}^0$, and selects

solutions at iteration $k > 0$ using the relation

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

(a tie breaking rule may be needed). The algorithm terminates when there are only higher 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 is to repeat the descent many times, restarted at a random starting point each time the algorithm hits a local minimum. Parallelization of such an algorithm, referred to here as *steepest descent with random restarts*, can be classified as embarrassingly easy.

Well-known algorithms such as simulated annealing (see, e.g., Aarts and Korst 1989) and Tabu Search (see, e.g., Glover and Laguna 1997) are clearly in the local search family. Classic, simple genetic algorithms (GA) are not unless the definition of a neighborhood is extended. However, almost all modern GAs used for function minimization make extensive use of steepest descent (see, e.g., Mülenbein 1992) so that their performance often resembles steepest descent with random restarts in that the genetic recombination and mutation select starting points for descent.

3.2 Exchange Neighborhoods

For the MCD a sensible neighborhood is one where a point in J is exchanged with one not currently in J. We refer to this as a *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.

Neighborhoods of this type are used, for example, by (Späth 1985) as well as Coleman and Woodruff (1997) in first-improving local search algorithms for (MINW); Hawkins (1994) as well as Rocke and Woodruff (1996) for the MCD.

Update formulas for the covariance determinant form the basis of the move evaluation functions. The test for the effect of a swap move for the MCD can be done quickly using a method described by Hawkins (1994). We construct a $p+1$ by n matrix $\boldsymbol{Z}$ that has a one in the first element of each vector and matches data vectors in all other elements. If $J \subset N$ let $\boldsymbol{Z}_J$ consist of the rows of $\boldsymbol{Z}$ indexed by J. A matrix $\boldsymbol{B}_J$ is then formed as $\boldsymbol{Z}_J \boldsymbol{Z}_J^T$. Let $\boldsymbol{u}$ be the column (i.e., data point) of $\boldsymbol{Z}_J$ to be swapped out by the move being evaluated, $\boldsymbol{v}$ be the one swapped in. The determinant will change by the factor

$$(1 - \boldsymbol{u}^T \boldsymbol{B}_J^{-1} \boldsymbol{u})(1 + \boldsymbol{v}^T \boldsymbol{B}_J^{-1} \boldsymbol{v}) + (\boldsymbol{u}^T \boldsymbol{B}_J^{-1} \boldsymbol{v})^2. \tag{5}$$

Hence, we can compute the effect of a swap move with an $\mathrm{O}(p^2)$ effort. Furthermore, a shortcut given in Hawkins (1994) allows the best swap to be computed in an $O(np^2)$ effort instead of the obvious $O(n^2p^2)$. For problems such as (MINW) and (MIND), very similar functions can be used.

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. So-called *greedy* algorithms can then be cast as steepest descent with a constructive neighborhood. These sorts of algorithms were no doubt first applied to combinatorial problems before our species developed language. They have a tradition in the literature of combinatorial optimization with too many citations to begin to sample them; one entry into the literature is Korte et al. (1991). A recent example of a general purpose descent algorithm with a constructive neighborhood, a randomized evaluation function and random restarts is commonly called GRASP (see, e.g. Feo and Resende 1995).

In the realm of clustering, constructive neighborhoods have been used as the basis for K-means algorithms (see, e.g., Selim and Ismail 1984). There are many variations of K-means but, roughly, the algorithm proceeds as follows when given as data the number of groups (we call it g and they call it K), a metric and the data points:

1. Find an initial collection of g points that will serve as seed points, one point for each group. These might be randomly selected data points. Assign each seed point to its respective group.

2. Use a constructive neighborhood to assign additional points to each group. Use as the evaluation function the distance from the point to the groups and pick the group that is closest to each point.

3. Use the mean of each cluster as its seed point and repeat step 2 or terminate if none of the groups changed members between the last two executions of step 2.

K-means algorithms are not affine equivariant, but typically use columnwise normalization to avoid sensitivity to simple changes in one of the measurement scales.

To apply these types of neighborhoods to the problems of interest to us, the Mahalanobis distance is a useful metric. Armed with this metric as the basis for an evaluation function, we can create steepest descent with a destructive neighborhood. The algorithm proceeds as follows:

1. Begin with $J = N$

2. Remove from J the point $i \in J$ that maximizes $d^2_{\Sigma_J}(x_i, \bar{x}_J)$ where Σ_J is the covariance matrix and $\bar{x}_J$ is the mean of the points in J.

3. If J contains H points, stop; otherwise go to step 2.

Unfortunately, this algorithm cannot be counted on to reject the outliers (see, e.g., Atkinson and Mulira 1993), since the beginning shape matrix Σ_N can easily be so distorted by outliers, that those outliers do not appear especially distant ("masking").

Constructive neighborhoods have been used with greater success so we proceed to formalize them. A number of parameters are needed to synthesize the various proposals in the literature. Consider first the MCD. A constructive neighborhood "surrounds"

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

Table 1: The parameters given as data and those computed dynamically using the rules of the algorithm

a set of points, J, that has between $p+1$ and H members. A subset of J (typically either empty or all of J) is required to be included in all neighbors; call this subset $\tilde{J}$. 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 J, moves to a new set J' must be such at all of the points in $\tilde{J}$ are in J' plus one more or points from $\tilde{N}$. This is summarized in Table 1.

It sounds a little odd to refer to "construction by steepest descent," but that is the way many algorithms are cast in our framework. 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 J 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 J using the covariance matrix of points in J as the metric; i.e., select the point(s) $i \in \tilde{N}$ that minimize(s)

$$d^2_{\Sigma_J}(x_i, \bar{x}_J)$$

Algorithm	$\hat{f}(\cdot)$	$\tilde{J}$	ψ	Start/Restart	Walking?
Fast-MCD Rousseeuw 1999	MAHAL	$\emptyset$	∞	HEUR	Yes
Forward Atkinson 1994	UPDATE	J	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	J	1	HEUR	No

Table 2: Summary of Affine Equivariant Constructive Neighborhoods for MVE/MCD Algorithms Reported in the Literature as cast in a Local Search Framework

where Σ_J is the covariance matrix and $\bar{x}_J$ is the mean of the points in J. 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 $\tilde{J}$) 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 $\bar{x}_J$ and Σ_J are updated. An alternative to MAHAL is to use Equation 5 to predict the effect of a move. Update formulas can be used for the inverse of the covariance matrix and the refresh period specifies how often the mean and covariance are recomputed from the current set J. Call this method UPDATE.

Table 2 gives a summary of constructive neighborhoods that have been reported in the literature for the MCD (and/or the MVE). In all cases, $\tilde{N} = N \setminus \tilde{J}$. 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. The constructive neighborhood has a long history of use for K-means and for other non-affine-equivariant methods such as the moving center algorithms of Mirkin (1996). The main difficulty is in expressing concisely how to deal with the problem of *contention*, which does not crop up for the MCD because there is only one set J.

When viewed from the perspective of a clustering problem, the MCD construction can be considered *cluster-wise* in that the main population is assigned the points closest to it. Cluster-wise assignment can result in contention problems for clustering. For one thing, unlike the set J for the MCD, the sets J_i for clustering problems typically do not have a fixed number of points. In addition, a point could be one of the closest points to two or more different clusters under the respective metrics of the clusters.

A partial remedy is *point-wise* construction. Each point x is assigned to the cluster given by

$$\operatorname*{argmin}_i d^2_{\Sigma_{J_i}}(x, \bar{x}_{J_i}).$$

We use the words "partial remedy" because a repair problem remains in the presence the constraint that each cluster must have at least $H \geq p+1$ points. Point-wise assignment can result in a solution where some of the clusters will have too few members. Something must be done to *repair* the solution so that all clusters have at least H points. A simple repair heuristic iteratively assigns to clusters with too few members the point closest to the cluster under its metric from among those points that are in clusters with more than H members. Even this simple heuristic requires a fair amount of notation, so we forgo rigorous description of repair heuristics because they are not central to analysis.

4 A (somewhat) New Cluster Estimator

There are two related ways to arrive at the need for an estimator that maximizes a pseudo likelihood for clustering by allowing some points to be ignored. The first is to use the same sort of reasoning that results in fixed point cluster methods (Hennig 1998) or partial discrimination rules (see, e.g., Anderson 1969 and Gessaman and Gessaman 1972). One can simply get better classifications if some points can be ignored. It can be the case that are "stray points" that perhaps come from populations whose probability of inclusion is low so that there are only one or two points from the population. Actually, any points from a population that supplies less than $p + 1$ points cannot be correctly identified as a cluster. We could think of such points as "outliers" to highlight the connection with robust analysis.

The other way to arrive at a need for partial classification is to construct a subsampling algorithm for maximum likelihood cluster assignment. As is the case for the MCD, subsampling is clearly needed for large datasets because of the computational effort required to find a good cluster assignment (see, e.g., Coleman and Woodruff 1997). However, when one constructs a subsample, populations that contributed small but non-negligible numbers of points to the dataset may contribute too few to the subsample. A clustering algorithm that can omit discrepant points can be used in a hierarchical manner to cluster data even with relatively small clusters. This would be done by re-applying the cluster algorithm to the points collectively omitted from the first clustering attempt. This could obviously be done recursively as many times as necessary.

4.1 The Minimization Problem

We refer to the partial classification estimator as (MINO) and define it as the result of a minimization. We have as data the number of groups, g, a minimum cluster size, H, maximum number of points to ignore T, and n points in R^p with each point given as x_j for $j = 1, \ldots, n$. Our objective is to select indicators of membership $\tau_{ij},\ i = 1, \ldots, g+1,\ j = 1, \ldots, n$ so as to

$$\text{minimize} \sum_{i=1}^{g} n_i \log \left| \frac{W_i}{n_i} \right| \quad \text{(MINO)}$$

subject to

$$\begin{array}{lcll}
W_i & = & \sum_{j=1}^{n} \tau_{ij}(x_j - \bar{x}_{J_i})(x_j - \bar{x}_{J_i})^T & i = 1, \ldots, g \\
\bar{x}_{J_i} & = & \left(\sum_{j=1}^{n} \tau_{ij} x_j \right) / n_i & i = 1, \ldots, g \\
n_i & = & \sum_{j=1}^{n} \tau_{ij} & i = 1, \ldots, g \\
\sum_{i=1}^{g+1} \tau_{ij} & = & 1 & j = 1, \ldots, n \quad \text{(placement)} \\
\sum_{j=1}^{n} \tau_{ij} & \geq & H & i = 1, \ldots, g \quad \text{(size)} \\
\sum_{j=1}^{n} \tau_{ij} & \leq & T & i = g+1 \quad \text{(ignore)} \\
\tau_{ij} & \in & \{0, 1\} & i = 1, \ldots, g+1, \\
 & & & j = 1, \ldots, n.
\end{array}$$

Constructive and swap neighborhoods generalize quickly to this problem. For the purpose of analysis and discussion, refer to the $g+1$ group as the *unassigned* points. This group is limited in size by the constraint labeled (ignore). It is easy to see that with data in general position, this constraint will be binding for the optimal solution. A search algorithm can consider solutions that exclude up to T points from assignment. For constructive neighborhoods there are a number of possibilities. The simplest is to use all points during a constructive descent then perform a destructive descent until T points have been removed. A computationally more expedient method is exclude from $\tilde{N}$ the T points with greatest

$$\min_i d^2 \Sigma_{J_i}(x, \bar{x}_{J_i}).$$

4.2 Remarks

If we adopt the MCD and (MINO) as our primary concerns, then in the language of local search we have two problems and two neighborhood structures (transition and constructive). The problems are clearly related, and in this section we provide some insight by making a few remarks concerning comparisons of the theoretical solutions to these problems. We then continue with some remarks concerning the neighborhoods and local minima. This analysis provides guidance for creating an algorithm based on MCD algorithms to search for good solutions to (MINO) as well as insights into connections between the problems.

4.2.1 Comparisons of Global Minima

We begin our discussion by comparing the global minima in order to describe the close relationships between the MCD and the objective functions that maximize cluster like-

lihoods. To put it another way, we are comparing the theoretical estimators consisting of the true global minimum, which is never computationally feasible to locate exactly except in very small cases. Our first remark can be seen immediately and provides motivation for extending (MIND) to become (MINO).

Remark 1 *The objective function for (MINO) with $g = 1$ and $T = n - \lfloor (n+p+1)/2 \rfloor$ imposes the same order on feasible solutions as the objective function for the MCD.*

The next remark provides a demonstration of the strong connection between the two estimator types. A very similar remark can be made concerning the problem (MINW). We make the remark for (MIND) rather than (MINO) because the presence of ignored points does not add to the insights gleaned.

Remark 2 *Consider a dataset of size n with a main population in general position consisting of at least $\lfloor (n+p+1)/2 \rfloor$ points and a group of outliers that are shifted by adding $\lambda\eta$ to each vector (η is a unit vector and λ is a large number). Then the larger of the clusters for the optimal solution of (MIND) for $g = 2$ contains the optimal solution to the MCD with a probability that goes to one as $\lambda \to \infty$. If the main cluster is exactly of size $\lfloor (n+p+1)/2 \rfloor$, then the two coincide.*

This is obvious because inclusion of a point in the far cluster in the MCD group will inflate the covariance determinant by an amount that increases without bound with λ. Remark 2 does not imply that we can do away with the MCD and use only a cluster model. As an extreme example, consider radial outliers. They are easily detected by MCD algorithms, but will generally not be revealed by search for two clusters using a MINW- or MIND-type criteria.

Remark 3 *Suppose that the main population is multivariate normal and that any outliers are radial. Then the optimal solutions of (MIND), (MINW), and (MINO) for $g = 2$ will split the main population between the two clusters with a probability that goes to one as n increases.*

This is not hard to see by looking at the functional definition of the cluster estimators. Consider first the case of an uncontaminated multivariate standard normal. Clusters consisting of a spherical main cluster and the rest of the distribution in the other produce a (MINW) criterion of exactly 1, and a (MIND) criterion strictly greater than 1. On the other hand, clusters that split along a hyperplane through 0 give a (MINW) and (MIND) criterion of $1 - \pi^{-1} = .68$. This generalizes to any radially symmetric distribution: splits into a sphere and the remainder generate criteria at least equal to the expected covariance determinant of the whole distribution, while half-space splits generate criteria strictly less than this. Thus, for large n, the same will be true of the finite sample version.

4.2.2 Neighborhoods and Local Minima

Local minima are defined with respect to a neighborhood structure and evaluation function. For theoretical purposes such as the remarks we make here, it is often useful

to consider the evaluation function to be the objective function. This is not unreasonable, since the evaluation functions for these problems are intended to be good approximations to the objective function and because if all else fails it is computationally reasonable to make use of the objective function as the evaluation function in the neighborhood of local minima. The next remark is an immediate consequence of the problem formulations.

Remark 4 *The global minima are local minima with respect to swap neighborhoods for both the MCD and (MINO).*

A proof for the following remark is given by Rousseeuw and Van Driessen (1999).

Remark 5 *Walking algorithms for the MCD are monotone in the objective function value.*

The proof should carry over to the case of (MINW), where each point can be assigned to the cluster to which it is closest. Unfortunately, it does not carry over to (MIND) and (MINO) because of the constraints on cluster size needed to avoid singularity. The consequence is that MCD and (MINW) walking algorithms can use the objective function value as a termination criteria and be assured of reaching local minima, but (MIND) and (MINO) algorithms generally cannot.

5 The Envelope of Outlier Detection

Our interest in this section is in exploring the efficacy of using a combination of an MCD algorithm and a (MINO) algorithm to find a main population in the presence of outliers. This highlights the connections between the two problems and demonstrates that there can be important benefits in combining them.

When searching for outliers, combinatorial estimators are of value principally to find good starting solutions for iterative algorithms applied to smooth estimators. Hence, we make use of two stages where the first stage is either the result of a search for the MCD or for a solution to (MINO) and the first stage result is fed to an M-Estimator of robust location and shape in the second stage. The M-estimator is held fixed as described in Rocke and Woodruff (1996) so that the two first stage estimators can be compared.

5.1 Search Algorithm

Based on remarks in the previous section, we can see that it is possible to construct an algorithm that can search for both the MCD and solutions to (MINO), with the difference determined entirely by input parameters. This will not result in an algorithm of maximum efficiency (with respect to computer resources) especially for the MCD. However, our interest here is in efficacy and in highlighting connections between the two problems, so we proceed with a computer program developed by Torsten Reiners, which exploits the remarks of the preceding section.

The program uses constructive neighborhoods as well as walking. Since, as we noted, Remark 5 does not hold for (MINO) we hash the solution vectors during walking

and terminate on a hash collision which indicates (with high probability) either a cycle or a local minimum. Details of the algorithm are given in the Appendix, and complete information in Reiners (1998).

For studies such as this one, the trick is to create datasets that are difficult to analyze, but where there is not too much overlap in the data. Algorithms can quickly detect outliers that are very far from the main data as noted by Kosinski (1998), who considered outliers that were much further than the ones we considered here. We tested (MINO) on the class of data sets that he used, and we were able to correctly identify all outliers. On the other hand, if the outliers are so close that they overlap the data in the main population, then the experiments are difficult to interpret and test the efficacy of the M-estimator more than they test the MCD and (MINO) algorithms.

Like other affine-equivariant methods, ours is dependent on discovering from the data the true shapes of the clusters to use in measuring distances. Thus, difficult data sets for our algorithm are those where the shape of the main data and the shape of the entire dataset differ substantially. One such class of data sets has a standard normal main population with outlying clusters arranged on a line (Coleman and Woodruff 1997). The spherical shape of the main data differs then from the shape of the entire data set, which is best described as resembling a cigar along the line.

We generated simulated datasets of $n = 1000$ points in dimension $p = 10$. These datasets all have a main cluster with 550 points, zero, one or two outlying clusters and either zero or 150 radial outliers. When present, the outlying clusters come from a population with a mean on the main diagonal that has a distance from the main population that is a multiple, $D = 2, 4$, of $\sqrt{\chi^2_{p;0.001}}$, which is more or less the radius of the sphere around the mean of the standard-normal main data that contains almost all the good points. The outlying clusters have the same shape as the main cluster, but the size is reduced so that the expected distance from an outlier to mean of the main cluster under the "correct" main cluster metric is equal to the expected distance to a point in the main cluster itself because this makes identification even harder than for shift outliers (see Rocke and Woodruff 1996).

Since the MCD and (MIND) are both special cases of (MINO), we use a (MINO) algorithm for all three. For the MCD, we use $g = 1$ and $T = (1-\alpha)n$. For both (MIND) and (MINO) we use $H = (p+1)$. Results for $D = 2$ are summarized in Table 3, where N_1 gives the number of points in the main cluster, N_2 and N_3 in the outlying clusters, and the column labeled "Radial" gives the number of radial outliers. The parameters for (MINO) are given in the columns labeled g and T. When $g = 1$, the value of T is set for the MCD. For the other two values of g we used both $T = 1/(g+1)$ and $T = 0$, corresponding to (MINO) and (MIND), respectively.

We verified these conclusions with simulated data sets of $n = 8p^2, 12p^2$ points in dimension $p = 5, 10$ each with a standard normal main population and αn outliers, with $\alpha = .2, .3, .4, .48$. The outliers are in one or more clusters. When there is more than one outlier cluster they have the same number of points. For each data set generated, we search for an MCD using (MINO) with $T = (1-\alpha)n$ and for (MIND) with $H = (p+1)$ and the correct value of g.

One can always do better with more time or better algorithms, but that is beside the point. The point is that by using the results of (MIND), we can extend the envelope

N_1	N_2	N_3	Radial	g	T	Estimator	Success Rate
550	450	0	0	1	495	MCD	0.5
				2	333	MINO	0.9
					0	MIND	1.0
				3	250	MINO	0.8
					0	MIND	0.0
	225	225		1	495	MCD	0.0
				2	333	MINO	1.0
					0	MIND	1.0
				3	250	MINO	1.0
					0	MIND	1.0
	0	0	450	1	495	MCD	1.0
				2	333	MINO	1.0
					0	MIND	1.0
				3	250	MINO	1.0
					0	MIND	1.0
	300	0	150	1	495	MCD	0.6
				2	333	MINO	1.0
					0	MIND	0.0
				3	250	MINO	1.0
					0	MIND	1.0
	150	150	150	1	495	MCD	1.0
				2	333	MINO	0.7
					0	MIND	0.0
				3	250	MINO	1.0
					0	MIND	0.7

Table 3: Fraction of Ten Replicates Resulting in an Outlier Misclassification Rate below 0.025 for Spherical Main Cluster with Outlying Clusters and Radial Outliers with $p = 10$

of contamination beyond what is reported in the literature summarized in Table 2 and to the theoretical limits for these cases.

The algorithm for (MIND) was able to find a starting point for the M-estimator that resulted in detection of the outliers for all datasets as was the MCD with $\alpha \leq 0.4$. When $\alpha = 0.48$, the MCD broke down completely with $D = 2$ and $p = 2$. When three groups are generated, the MCD performs well with only a few failures for small n and $\alpha = 0.48$.

Several points are notable from these results:

- The MCD does not work as well when the outliers form a single cluster, while (MINO) and (MIND) work well. Thus it would appear that running a clustering algorithm can protect against this difficult kind of outliers.

- Everything works when the only outliers are radial. This is not necessarily because the cluster estimators work well themselves but rather because the M-estimator has a unique solution in this case so that any starting point will do.

- The algorithm configured for problem (MIND) works well if one guesses g "correctly" or guesses low, but (MINO) is more robust. In some cases, (MIND) is able to treat radial outliers as a cluster, but this is likely because 150 points in dimension 10 are too few to be evenly distributed.

- For spherical clusters, two shrunken outlying clusters on either side of the main cluster are harder for the MCD than one at the same distance. We confirmed this result using Fast-MCD (Rousseeuw and Van Driessen 1999). Although masking is maximized in some sense by a single cluster when the correct metric is known, the correct metric can only be estimated. As noted earlier, stochastic algorithms used as estimators are adversely affected when the all-data metric is significantly different from the correct estimate. The shape of the all-data metric is elongated more by the presence of two outlying clusters than one, which in turn maximizes the chances that any sub sample will result in a misleading shape estimate given that the clusters are spherical.

6 Conclusion

In this paper, we have shown that robust multivariate estimation and outlier detection on the one hand and cluster analysis on the other can be placed in a common conceptual and computational framework. This perspective allows a dramatic increase in the ability to handle outlier problems by subjecting the data both to robust estimation/outlier detection methods and to companion cluster analysis methods. Furthermore, we have defined a robust clustering method that can cope with observations that lie in no cluster.

Acknowledgments

The research reported in this paper was supported by grants from the National Science Foundation (DMS 95-10511, DMS 96-26843, ACI 96-19020, and DMS 98-70172) and the National Institute of Environmental Health Sciences, National Institutes of Health (P42 ES04699). The authors are grateful to Torsten Reiners for the use of his computer code and his help with computational experiments.

Appendix: Algorithmic Details

In this appendix, we describe some details of the algorithm we use for the computations in this paper; a complete description can be found in Reiners (1998). A unique aspect of this algorithm is the way in which the iterations are started for different descent trials.

The algorithm keeps a list $\mathcal{L}$ of length g of cluster means and covariances corresponding to the best solution seen so far. At the very beginning, each list element is the same: the location and shape of all of the data. This list is used to *process a seed point* to form the nucleus of group i (which is always done after groups $i' < i$ have been processed) as follows: Find the member of $\mathcal{L}$ with lowest distance from the list element mean to the seed point using the list element covariance for a metric. Then form a small cluster of the $p+1$ points that are closest to the seed point under that metric (this will always consist of the seed point and p additional points. Use this as a starting point for walking as shown for Fast-MCD in Table 2 with $h = p+1$ and $\tilde{N} = N \setminus \bigcup_{i'<i} J_{i'}$. Call the resulting points J_i.

The algorithm proceeds roughly as follows for each point in the data set, x_j.

1. Process x_j as seed point for group 1.
2. A seed point for each of the subsequent groups, $i = 2, \ldots, g$ is chosen in order and processed. The point for group i is the

$$\underset{x \in \tilde{N}}{\operatorname{argmax}} \Pi_{k<i} d\Sigma_{J_k}; \bar{x}_{J_k}, x)$$

   where $\bar{x}_{J_k}$ and Σ_{J_k} are the mean and covariance matrix respectively for the set of points J_k and $\tilde{N} = N \setminus \bigcup_{i'<i} J_{i'}$.
3. The points that have been assigned to each cluster are used to start a walking algorithm with simple repair as described in §3.3 that is terminated when a duplicate solution is encountered (based on a hash table).

References

Aarts, E.H.L., and J. Korst (1989), Simulated Annealing and Boltzmann Machines, John Wiley and Sons, New York.

Anderson, J.A. (1969) "Constrained Discrimination Between k Populations," *Journal of the Royal Statistical Society, Series B*, **31**, 123-139.

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

Atkinson, A.C. and H.M. Mulira (1993), "The Stalactite Plot for the Detection of Multivariate Outliers" *Statistics and Computing*, **3** 27-35.

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

Campbell, N. A. (1980) "Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation," *Applied Statistics*, **29**, 231–237.

Campbell, N. A. (1982) "Robust Procedures in Multivariate Analysis II: Robust Canonical Variate Analysis," *Applied Statistics*, **31**, 1–8.

Coleman, D.A. and D.L. Woodruff (1997), "Cluster Analysis for Large Data Sets: Efficient Algorithms for Maximizing the Mixture Likelihood," Technical Report, GSM, UC Davis, Davis CA, 95616.

Coleman, D.A., Dong, X., Hardin, J., Rocke, D.M. and Woodruff, D.L. (1999) "Some Computational Issues in Cluster Analysis with no a Priori Metric," *Computational Statistics and Data Analysis*, **31**, 1–11.

Davies, P. L. (1987) "Asymptotic Behavior of S-Estimators of Multivariate Location Parameters and Dispersion Matrices," *Annals of Statistics*, **15**, 1269–1292.

Feo, T.A. and Resende, M.G.C. (1995), "Greedy Randomized Adaptive Procedures," *Journal of Global Optimization*, 6, 109-133.

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

Gessaman, M.P., and P.H. Gessamann (1972), "A Comparison of Some Multivariate Discrimination Procedures," *Journal of the American Statistical Association*, **67**, 468-472.

Glover, F., and M. Laguna (1997) *Tabu Search*, Kluwer, Boston.

Glymour, C., D. Madigan, D. Pregibon, P. Smyth (1997), "Statistical Themes and Lessons for Data Mining," *Data Mining and Knowledge Discovery* **1**, 11-28.

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

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley.

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

Hawkins, D. M. (1993) "A Feasible Solution Algorithm for the Minimum Volume Ellipsoid Estimator," *Computational Statistics*, **9**, 95–107.

Hawkins, D. M. (1994), 'The Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator in Multivariate Data," *Computational Statistics and Data Analysis*, **17**, 197–210.

Hawkins, D.M. (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

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

Kent, J. T. and Tyler, D. E. (1991) "Redescending M-estimates of multivariate location and scatter," *Annals of Statistics*, **19**, 2102–2119.

Korte, B., L. Lovasz, and R. Schrader (1991), *Greedoids*, Springer, Berlin.

Kosinski, A.S. (1998), "A Procedure for the Detection of Multivariate Outliers," *Computational Statistics and Data Analysis*, **29**, 145–161

Lopuhaä, H. P. (1989) "On the Relation between S-Estimators and M-Estimators of Multivariate Location and Covariance," *Annals of Statistics*, **17**, 1662-1683.

Lopuhaä, H. P. (1992) "Highly efficient estimators of multivariate location with high breakdown point," *Annals of Statistics*, **20**, 398–413.

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.

Maronna, R. A. (1976) "Robust M-Estimators of Multivariate Location and Scatter," *Annals of Statistics*, **4**, 51–67.

Mirkin, B. (1997) *Mathematical Classification and Clustering*, Kluwer Academic Publishers, Boston.

Mülenbein, H. (1992), "Parallel Genetic Algorithms in Combinatorial Optimization," in Balci et al, eds., *Computer Science and Operations Research*, Pergamon Press, New York.

Murtagh, F. and A.E. Raftery (1984), "Fitting Straight Lines to Point Patterns," *Pattern Recognition*, **17**, 479-483.

Reiners, T., (1998) *Maximum Likelihood Clustering of Data Sets Using a Multilevel, Parallel Heuristic*, Masters Thesis, Department of Economics, Business Computer Science and Information Management, Technische Universität Braunschweig, D-38106 Germany.

Rocke, D. M. (1996) "Robustness Properties of S-Estimators of Multivariate Location and Shape in High Dimension," *Annals of Statistics*, **24**, 1327–1345.

Rocke, D.M. (1998) "Constructive Statistics: Estimators, Algorithms, and Asymptotics," *Computing Science and Statistics*, **30**, 3–14.

Rocke, D. M. and Woodruff, D. L. (1993) "Computation of Robust Estimates of Multivariate Location and Shape," *Statistica Neerlandica*, **47**, 27-42.

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

Rousseeuw, P. J. (1985), "Multivariate estimation with high breakdown point," in Grossmann, W., Pflug, G., Vincze, I. and Werz, W. *Mathematical Statistics and Applications, Volume B*, Dordrecht: Reidel.

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.

Seidel, W., K. Mosler, and Manfred Alker (1999), "A Cautionary Note on Likelihood Ratio Tests in Mixture Models," *Annals of the Institute of Statistical Mathematics*, to appear.

Selim, S.Z. and Ismail, M.A. (1984), "K-Means-type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality", *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 6, 81-87.

Späth, H. (1985), Cluster Dissection and Analysis, Ellis Horwood (Wiley), Chichester.

Tyler, D. E. (1983) "Robustness and Efficiency Properties of Scatter Matrices," *Biometrika*, **70**, 411–420.

Tyler, D. E. (1988) "Some results on the existence, uniqueness, and computation of the M-estimates of multivariate location and scatter," *SIAM Journal on Scientific and Statistical Computing*, **9**, 354–362.

Tyler, D. E. (1991) "Some Issues in the Robust Estimation of Multivariate Location and Scatter," in Stahel, W. and Weisberg, S. (eds) *Directions in Robust Statistics and Diagnostics Part II,* New York: Springer-Verlag.

Ward, J. (1963) "Hierarchical Grouping to Optimize and Objective Function," *Journal of the American Statistical Association*, 37, 236-244

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.