


A Deterministic Method for Robust Estimation of Multivariate Location and Shape

Wendy L. POSTON, Edward J. WEGMAN,
Carey E. PRIEBE, and Jeffrey L. SOLKA

The existence of outliers in a data set and how to deal with them is an important problem in statistics. The minimum volume ellipsoid (MVE) estimator is a robust estimator of location and covariate structure; however its use has been limited because there are few computationally attractive methods. Determining the MVE consists of two parts—finding the subset of points to be used in the estimate and finding the ellipsoid that covers this set. This article addresses the first problem. Our method will also allow us to compute the minimum covariance determinant (MCD) estimator. The proposed method of subset selection is called the effective independence distribution (EID) method, which chooses the subset by minimizing determinants of matrices containing the data. This method is deterministic, yielding reproducible estimates of location and scatter for a given data set. The EID method of finding the MVE is applied to several regression data sets where the true estimate is known. Results show that the EID method, when applied to these data sets, produces the subset of data more quickly than conventional procedures and that there is less than 6% relative error in the estimates. We also give timing results illustrating the feasibility of our method for larger data sets. For the case of 10,000 points in 10 dimensions, the compute time is under 25 minutes.

Key Words: Minimum covariance determinant; Minimum volume ellipsoid; Outliers; Robust estimators; Subset selection.

1. INTRODUCTION

An important area of research in statistics is the robust estimation of location and covariance structure for a set of data. In this article, robust estimation will refer to those estimators that have high breakdown points (Rousseeuw and Leroy 1987) or estimators

Wendy L. Poston is Aerospace Engineer, Naval Surface Warfare Center, Dahlgren Division, Advanced Processors Group, Dahlgren, VA 22448; e-mail: wposton@nswc.navy.mil. Edward J. Wegman is Professor, Center for Computational Statistics, George Mason University, Fairfax, VA 22030; e-mail: ewegman@gmu.edu. Carey E. Priebe is Assistant Professor, Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD 21218; e-mail: priebe@kronecker.mts.jhu.edu. Jeffrey L. Solka is Mathematician, Naval Surface Warfare Center, Dahlgren Division, Advanced Computation Technology Group, Dahlgren, VA 22448; e-mail: jsolka@nswc.navy.mil.

Journal of Computational and Graphical Statistics, Volume 6, Number 3, Pages 300–313

that will tolerate a large number of outliers before the estimate is affected. The estimator of interest here is called the minimum volume ellipsoid (MVE), an estimator that has desirable robustness properties due to its optimal breakdown point of 50% (Woodruff and Rocke 1993). No computationally reasonable deterministic methods of calculating the MVE exist, especially in high dimensions and for large sample sizes, making the MVE impractical for frequent use by statisticians.

The MVE of a given data set is determined by a subset of m points subject to the constraint that the ellipsoid that covers the points has minimum volume among all ellipsoids constructed using m points (Hawkins 1993; Rousseeuw 1985; Woodruff and Rocke 1993). The size of the subset is a function of the number of data points n and the dimensionality p and is chosen to give an estimate with a breakdown point of 50%. From this description of the MVE, it is apparent that finding a value of the estimator for a given data set has two parts. The first is to find the subset of data that is to be included in the estimate, and the second is to calculate the covering ellipsoid. A computationally efficient algorithm (relative to the expense of finding the set of m points) has been published (Cook, Hawkins, and Weisberg 1993) that will find the exact covering ellipsoid for a set of points. However, finding the MVE still requires exhaustive specification of all possible subsets of size m, making it computationally intractable for large data sets. Thus, the subset selection problem is the more computationally intensive of the two problems, and the one that remains to be solved. It is this issue that will be addressed in this article.

Current methods of subset selection include the basic resampling method described by Rousseeuw and Leroy (1987), which randomly chooses subsets and then retains the one yielding the minimum volume. Improvements on this resampling method include heuristic search algorithms investigated by Woodruff and Rocke (1993). These include simulated annealing, tabu search, and genetic algorithms. Another approach to finding the MVE is that of Hawkins (1993) called the feasible solution algorithm (FSA). These methods are random in that they rely on random starting points and random searches, and they are not guaranteed to find the exact MVE for any finite amount of sampling. Clearly, none of these methods provide reproducible estimates of the MVE for a given data set, unless the methods are implemented with the same random number generator and seed. Additionally, these methods are computationally intensive because one would repeat these for several random starting points and taking the smallest ellipsoid as the MVE. Finally, another problem with the heuristic algorithms is that each one involves several parameters that affect their performance and must be determined for each application.

The effective independence distribution (EID) method (Poston 1994) is proposed as a new solution to the subset selection problem in estimating the MVE. As with the other methods, it may not provide the exact MVE. However, we present results indicating that it does pick subsets that yield ellipsoids approaching the true MVE. Other aspects that make it particularly appealing are the repeatability of an estimate for a given data set due to its deterministic nature, and the fact that it is computationally tractable even for large data sets and high-dimensional problems. Also, one does not have to determine optimal algorithm parameters to implement it.

Section 2 provides some background information on the MVE estimator and describes the algorithm for finding the minimum covering ellipsoid. Section 3 introduces

the EID method for selecting the subset to be covered, and Section 4 describes the procedure for finding the MVE. Section 5 presents results that show the relative error in the volume of the ellipsoid obtained using the EID approach for several regression data sets where the true MVE is known. We also present results indicating the computational feasibility of the algorithm as a function of the sample size and the dimensionality of the data.

2. MINIMUM VOLUME ELLIPSOID ESTIMATOR

The problem of robust estimation of multivariate location and shape is: given a set of n p-dimensional observations, find an estimate of location and shape that is resistant to outliers or contaminated data. The MVE is one such estimator, and it is known that it has a breakdown point that approaches 50% as the number of points in the data set increases (Rousseeuw and van Zomeren 1990). This is the maximum possible breakdown point, and it means that approximately half of the data can be arbitrarily contaminated without affecting the estimate.

The MVE is given by the ellipsoid (Hawkins 1993)

$$(\mathbf{x} - \mathbf{c})^T \Gamma^{-1} (\mathbf{x} - \mathbf{c}) = p, \tag{2.1}$$

where $\mathbf{c}$ and Γ are the location vector and scatter matrix respectively and p is the dimension of the data. The location vector is a weighted mean calculated as

$$\mathbf{c} = \sum_{i=1}^{h} w_i \mathbf{x}_i^*, \tag{2.2}$$

and the covariance or scatter matrix is

$$\Gamma = \sum_{i=1}^{h} w_i (\mathbf{x}_i^* - \mathbf{c})(\mathbf{x}_i^* - \mathbf{c})^T, \tag{2.3}$$

where $\mathbf{x}_i^*$ is a column vector denoting the ith observation in the subset of m points, w_i is the weight for the ith observation, and $h = [(n + p + 1)/2]$ (the brackets denote the greatest integer function). The volume of the covering ellipsoid will be proportional to the determinant of Γ. It is evident from these equations that to find the MVE one must determine which m points should be covered and the corresponding weights to ensure coverage of the points.

The algorithm that will be used to find the weights is credited to Titterington (1975) and described by Hawkins (1993). It will be referred to within as Titterington's algorithm. All of the weights are initially set to $w_i^{(0)} = 1/h,\ i = 1, \ldots, h$, which is the usual weight given to points when calculating the sample mean of a data set of size m. Then, at each iteration k, calculate the weighted mean and covariance from Equations (2.2)–(2.3) and the Mahalanobis distances for each observation given by

$$D_i^{(k)} = \left(\mathbf{x}_i^* - \mathbf{c}^{(k)}\right)^T \Gamma_{(k)}^{-1} \left(\mathbf{x}_i^* - \mathbf{c}^{(k)}\right). \tag{2.4}$$

If $D_i^{(k)} \leq p$ for every i, then the current ellipsoid using $\mathbf{c}^{(k)}$ and $\Gamma_{(k)}^{-1}$ is the MVE covering the m observations. If the Mahalanobis distance for any of the observations exceeds p, then the weights must be adjusted using the following

$$w_i^{(k+1)} = w_i^{(k)} \frac{D_i^{(k)}}{p}, \tag{2.5}$$

and the calculations of Equations (2.2)–(2.4) are repeated until all of the distances are less than p. This procedure enlarges the ellipsoid until all of the m points are covered.

The algorithm for finding the weights can be somewhat computationally intensive for some data sets. Another method for estimating the weights can be found in Rousseeuw and van Zomeren (1990); it is quicker than Titterington's algorithm, but it does not yield the exact minimum covering ellipsoid for a given data set. However, it should be apparent that the real computational burden arises from the determination of which points must be covered by the ellipsoid. The brute force method of exhaustive enumeration is of computational complexity $O\left(\left(\begin{array}{c} n \\ h \end{array}\right)\right)$. The EID algorithm is presented as a means of addressing this problem.

3. EFFECTIVE INDEPENDENCE DISTRIBUTION

The derivation of the EID provided here was first given by Kammer (1991). The EID provides a ranking of each point according to its contribution to the eigenvalues, and hence to the determinant of the information matrix which is defined in the following. It will be shown that the EID offers a direct relationship between the determinants of the information matrix as points are removed from the data set, and can be used to optimize the determinant. It was orignially stated by Kammer (1991) that choosing points based on the EID values will yield the most linearly independent subset of observations. Poston (1995) showed that the EID values rank points in a data set according to how much each one contributes to the linear independence of the parameter space. Thus, the motivation for the name effective independence distribution.

The EID is developed from the set of equations familiar from regression theory (Rousseeuw and Leroy 1987). These are

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\mathbf{y}$ is an n-dimensional vector of responses, $\mathbf{X}$ is an $n \times p$ matrix of predictor variables with each column linearly independent, $\boldsymbol{\beta}$ is a p-dimensional column vector of unobservable parameters that must be estimated from the data, and $\boldsymbol{\epsilon}$ denotes the noise in the measurements. It is further assumed that $\boldsymbol{\mu} = E[\boldsymbol{\epsilon}] = 0$ and $\Sigma = E[(\boldsymbol{\epsilon} - \boldsymbol{\mu})^T(\boldsymbol{\epsilon} - \boldsymbol{\mu})] = E[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}]$. The information matrix is then given by FIM $= \mathbf{X}^T\mathbf{X}$.

The EID is an n-dimensional vector where each element corresponds to one measurement location. The development of the EID method given here will show that the ith term of the EID vector is the contribution of the ith data point to all of the eigenvalues

of the information matrix. Because

$$|\text{FIM}| = \prod_{j=1}^{p} \lambda_j, \tag{3.2}$$

where $(|\bullet|)$ denotes the determinant, then the eigenvalues are also a measure of the information and indicate the contribution of a data point to the determinant of the information matrix.

The EID can be derived from the following eigenvalue problem

$$(\text{FIM} - \lambda_j \mathbf{I})\Psi_j = 0, \tag{3.3}$$

where I is a $p \times p$ identity matrix, λ_j is the jth eigenvalue, and Ψ_j is the jth eigenvector. It follows from the definition that the information matrix is symmetric. Because the columns of $\mathbf{X}$ are linearly independent, this implies that it is also positive definite. Therefore, the eigenvector Ψ_j can be chosen to be orthonormal, and we will denote the matrix of eigenvectors as Ψ.

It can be shown that the jth eigenvalue has the form

$$\lambda_j = \sum_{i=1}^{n} \left(\sum_{j=1}^{p} x_{ik} \psi_{kj} \right)^2, \quad j = 1, \ldots, p. \tag{3.4}$$

The eigenvectors of the information matrix span the p-dimensional parameter space, so they can be used to transform the data matrix $\mathbf{X}$. The following matrix product is now formed

$$\mathbf{G} = (\mathbf{X}\Psi) \circ (\mathbf{X}\Psi), \tag{3.5}$$

where $\circ$ denotes an element-by-element matrix multiplication and $\mathbf{X}\Psi$ represents the transformed data matrix. The ijth element of $\mathbf{G}$ is given by

$$g_{ij} = \left(\sum_{k=1}^{p} x_{ik} \psi_{kj} \right)^2. \tag{3.6}$$

An examination of each element of $\mathbf{G}$ reveals that the sum of the jth column of $\mathbf{G}$ equals the jth eigenvalue given in Equation (3.4), hence

$$\sum_{i=1}^{n} g_{ij} = \lambda_j. \tag{3.7}$$

The next step is to post-multiply $\mathbf{G}$ by Λ^{-1} forming the following matrix

$$\mathbf{E} = \mathbf{G}\Lambda^{-1}, \tag{3.8}$$

which normalizes each column of $\mathbf{G}$ by dividing by the corresponding eigenvalue (i.e., the jth column is divided by the jth eigenvalue). Each column in the matrix $\mathbf{E}$ now sums to one, and the element e_{ij} represents the fractional contribution of the ith data point to the jth eigenvalue.

Finally, the EID is calculated by summing the terms in the ith row of the matrix $\mathbf{E}$

$$\text{EID}_i = \sum_{j=1}^{p} e_{ij}. \tag{3.9}$$

Thus, EID_i represents the contribution of the ith observation to the eigenvalues of the information matrix. Again, note that there are n elements in the EID, one corresponding to each point in the data set.

The diagonal elements of the "hat" matrix from regression theory (Rousseeuw and Leroy 1987) will also yield the EID values for each observation. Thus, an alternative formulation of the EID is given by

$$\text{EID} = \text{diag}(\mathbf{H}) = \text{diag}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T). \tag{3.10}$$

To derive this equation, start with the definition of the ith element of the EID

$$\text{EID}_i = \sum_{j=1}^{p} e_{ij} = \sum_{j=1}^{p} \frac{g_{ij}}{\lambda_j}, \tag{3.11}$$

and substituting for the ijth element of $\mathbf{G}$ from Equation (3.6) yields

$$\text{EID}_i = \sum_{j=1}^{p} \left(\sum_{k=1}^{p} \frac{x_{ik}\psi_{kj}}{\sqrt{\lambda_j}} \right)^2. \tag{3.12}$$

These are the diagonal elements of the following matrix product

$$\mathbf{H} = (\mathbf{X}\Psi\Lambda^{-1/2})(\mathbf{X}\Psi\Lambda^{-1/2})^T, \tag{3.13}$$

where $\Lambda^{1/2}$ is a diagonal matrix containing the square roots of the eigenvalues. The matrix $\mathbf{H}$ can be re-written as

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \tag{3.14}$$

and $\mathbf{H}$ is the usual "hat" matrix from regression.

The matrix given in Equation (3.14) has interesting properties that offer some insight into the nature of the EID. One is that it is an idempotent matrix. These matrices have the property that the trace equals the rank, so

$$\sum_{i=1}^{n} \text{EID}_i = \text{rank}(\mathbf{H}) = \text{rank}(\mathbf{X}) = p. \tag{3.15}$$

The EID can be said to show the contribution of the ith measurement location to the rank of the data matrix and thus also to the linear independence of the parameter space.

It has been shown previously (Poston and Tolson 1992) that the following relationship holds between the determinants of the information matrices as points are removed from a data set

$$\left|\mathbf{X}_{-i}^T\mathbf{X}_{-i}\right| = (1 - \text{EID}_i)\left|\mathbf{X}^T\mathbf{X}\right|, \tag{3.16}$$

where $\mathbf{X}_{-i}$ is the data matrix with the ith point removed and EID_i is the value for the ith point. From this one can see that there is a direct relationship between the determinants as the points are removed from the data set. If the objective is to minimize the determinant, then the observation with the largest EID value should be deleted. This is the case for finding the set of points used to determine the MVE.

The following proposition shows (Kammer 1991; Rousseeuw and Leroy 1987) the range of values that an element of the EID can have.

Proposition 1. *EID_i is in the range $0 \leq EID_i \leq 1$.*

Proof: Because $\mathbf{H}$ is an idempotent matrix, this implies that

$$h_{ii} = (\mathbf{HH})_{ii} = \sum_{j=1}^{n} h_{ij} h_{ji}.$$

Because $\mathbf{H}$ is also symmetric, the diagonal elements can be written

$$h_{ii} = \sum_{j=1}^{n} h_{ij} h_{ji} = \sum_{j=1}^{n} h_{ij}^2.$$

Expanding the sum on the right side yields

$$h_{ii} = h_{ii}^2 + \sum_{i \neq j} h_{ij}^2.$$

This equality can only be true if $h_{ii} \geq h_{ii}^2$ which implies that

$$0 \leq h_{ii} \leq 1$$

or that

$$0 \leq \text{EID}_i \leq 1$$

and the proposition is proved. □

It is instructive to examine what happens if a data point has a corresponding EID value of zero or one. A data point with an EID value of one must be retained to preserve the linear independence of the data matrix $\mathbf{X}$. This is obvious from Equation (3.16). If such a point is deleted, then the determinant of the resulting information matrix is zero and the problem becomes singular. In the regression setting, this means that all of the parameters cannot be estimated. On the other hand, if an observation has an EID value of zero, then the determinant is unchanged and no loss of information occurs.

4. PROCEDURE

Recall that the volume of the MVE is proportional to the determinant of the scatter matrix. This is the rationale for using the EID to select the subset of data points that is used in the MVE. If we use the matrix $\mathbf{X}^T\mathbf{X}$ to approximate Γ, then we can use the relationship in Equation (3.16) to successively remove points until m points remain. These m points will then be used in the algorithm described previously for finding the weights

and the resulting ellipsoid. However, to better approximate the scatter matrix, the data will be centered by subtracting the p-dimensional sample mean from each observation. This is repeated as each point is deleted. The complete procedure consists of the following steps:

1. Calculate the matrix

$$\mathbf{X}'^{(j)} = \left(\mathbf{X}^{(j)} - \bar{\mathbf{X}}^{(j)}\right),$$

   where $\mathbf{X}^{(j)}$ is the set of raw data points at the jth iteration of the method and $\bar{\mathbf{X}}^{(j)}$ is an $(n-j) \times p$ matrix with each row containing the p-dimensional sample mean for the current set of data. Note that at iteration $j = 0$ there are n points in the data set, at iteration $j = 1$ there are $n - 1$ points, and so on.
2. Use the matrix $\mathbf{X}'^{(j)}$ in Equation (3.10) to calculate the EID value for each point in the current data set.
3. Delete the point that corresponds to the maximum EID value.
4. Repeat steps 1–3 until m points remain.
5. Adjust the weights using Titterington's algorithm until the m points are covered by the ellipsoid.

Some care should be taken with Step 3 when implementing this method. It is quite possible that in the very first calculation of the n EID values that a data point has an EID value of one. Such a point must be retained to keep the problem nonsingular (see Eq. 3.16). Instead of deleting this point, one should remove the observation corresponding to the next highest EID value. The chances of an observation having an EID value of one becomes greater as the data set is reduced, and it is obvious from Equation (3.15) that when there are only p points left in the set, then each observation must have a value of one. Thus, this discussion becomes more critical as more points are deleted from the set, and hence it appears that the larger n relative to m, the better.

5. APPLICATIONS AND RESULTS

For illustrative purposes, the EID method is used to estimate the MVE for several data sets where the true MVE is known. The article by Hawkins (1993) gives the correct subset and the resulting volume of the true MVE for these data sets. The relative error in the volume of the ellipsoid based on the subset obtained using the EID method is determined for comparison purposes. The six data sets can be found in Rousseeuw and Leroy (1987), and the parameters of interest are shown in Table 1. From this one can

Table 1. Regression Data Set Parameters Timing Results

Data set	*p*	*n*	*h*	*Time (sec.) to select subset of points using the EID*	*Time (sec.) to find weights in Titterington's algorithm*
Aircraft	4	23	14	.053	.87
Coleman	5	20	13	.053	.33
Delivery	2	25	14	.053	2.64
Education	3	50	27	.170	2.04
Gravity	5	20	13	.053	1.38
Salinity	3	28	16	.053	.27

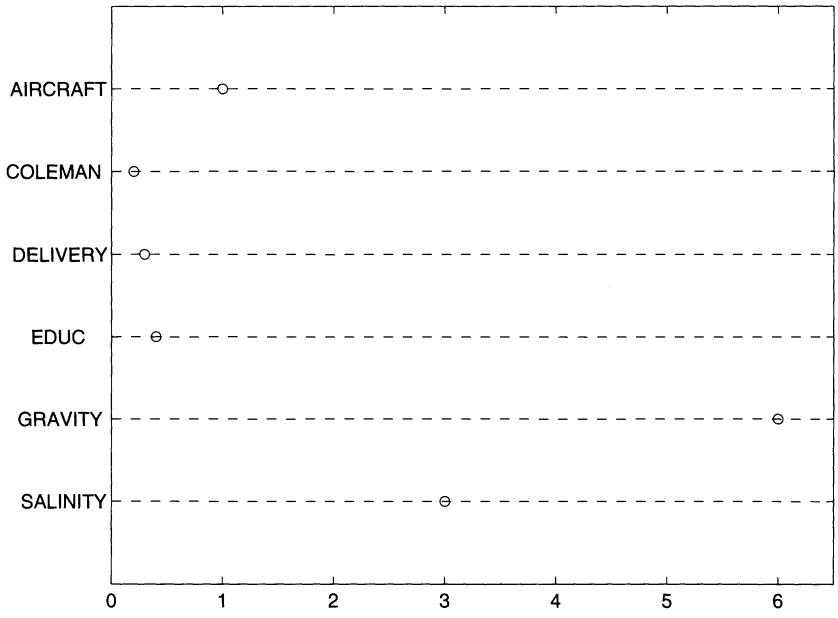

Figure 1. Percent Relative Error in the Volume of the MVE as Determined by the EID Approach.

see that the data sizes are relatively small ranging in size from $n = 20$ to $n = 50$. The dimensionality of the data is also low, from two to five dimensions.

The EID algorithm is implemented in MATLAB on a Pentium, 166MHz computer. The relative errors in the volume of the minimum covering ellipsoid using the EID approach are shown in Figure 1. It is evident from the small error that ours is a feasible approach to finding the MVE. The relatively large error in estimating the MVE for the gravity data set is due to the small size of the data set. It is recommended by Rousseeuw and van Zomeren (1990) that there be at least five observations per dimension.

The times needed to determine the subset of points using the EID method are given in Table 1, along with the time it took to determine the set of weights using Titterington's algorithm. These results show that the MVE can be estimated in under three seconds for the data sets considered here. It should be noted that for those sets with reported times of .053 seconds, the elapsed time to find the subset of points was too fast for the time resolution of the computer. Hence, this time is an upper bound for the execution of the algorithm in these instances.

The two-dimensional delivery data set is shown in Figure 2 to provide a qualitative assessment of the method. From this, it is clear that the bulk of the data is clustered toward the origin. When the EID method is applied to this data set, the first observations that are deleted are the outlying ones in the upper right corner of the plot. It is not until the last points are deleted that the EID algorithm makes an incorrect choice. The optimal set (Hawkins 1993) is shown in Figure 3, and the set chosen by the EID approach is shown in Figure 4. Note the point that is incorrectly retained in the set. One reason for this error is that the point the EID deletes has a larger magnitude than the one that should

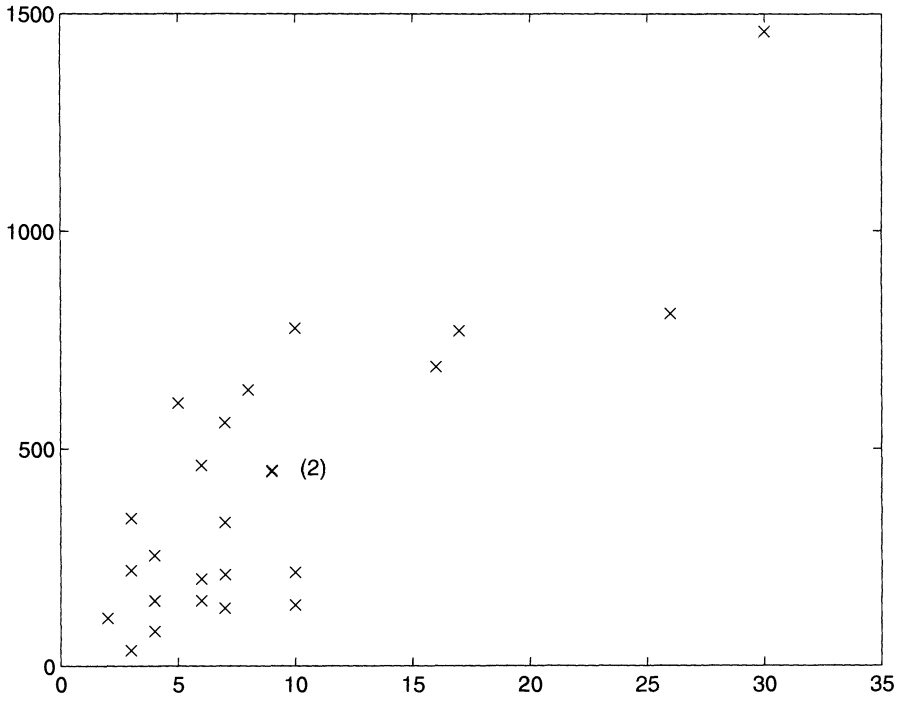

Figure 2. Delivery Data Set, n = 25.

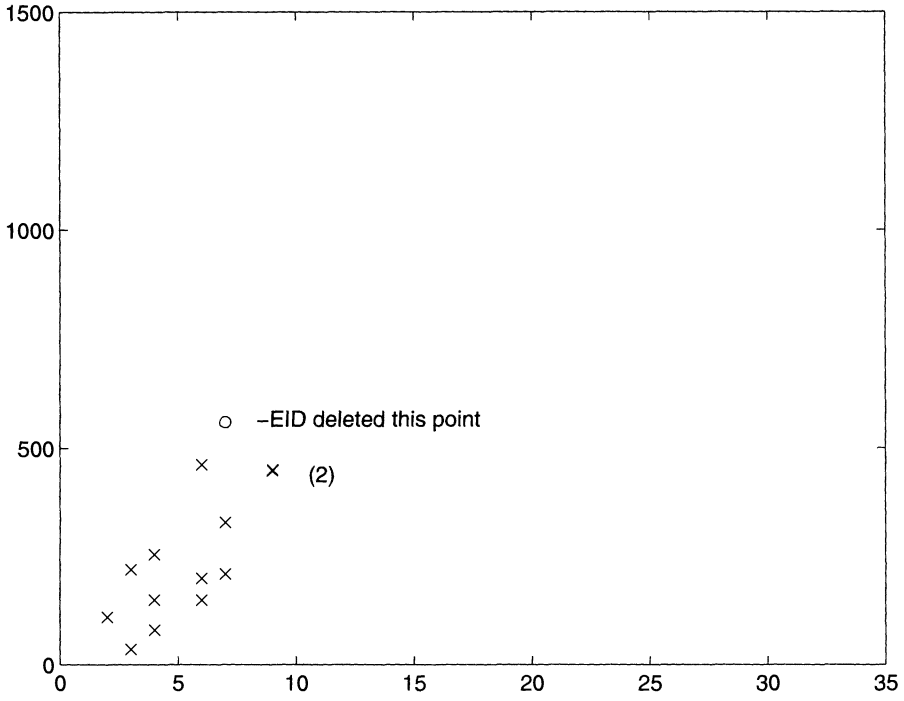

Figure 3. Subset Used in True MVE.

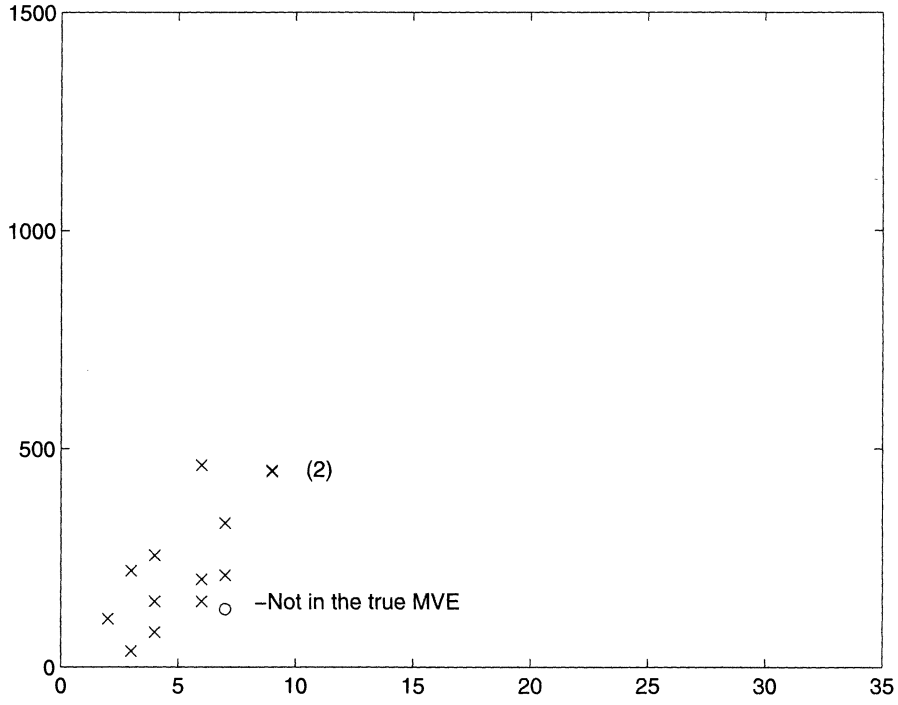

Figure 4. Subset Chosen by the EID Method.

be kept in the set. Previous studies indicate that these will be the points that tend to have a large EID value.

Finally, one last comparison is in order regarding the salinity data set. It is stated in Hawkins (1993) that this set would require approximately 5,000 random starts with the FSA to reliably determine the MVE, which is a computationally intensive task. Note that for this data set, the EID method of subset selection finds a set of points in less than .053 seconds with only 3% error in the volume of the ellipse. Thus, the EID method is a computationally efficient technique that produces a good estimate even for those data sets that trouble other methods.

To further study the efficiency of the EID algorithm for subset selection, we performed a study to examine the computational feasibility of the algorithm particularly as a function of the size of the data set and the dimensionality of the data. As before, MATLAB was used to implement the algorithm. It should be noted that MATLAB is an interpreted language, so this will cause the algorithm to be slower than if it were implemented in a compiled language. To provide a rough comparison to the heuristic algorithms described in Woodruff and Rocke (1993), they indicated that the average time for these algorithms on a DEC Alpha using a compiled language is approximately 30 minutes for $n = 50$ and $p = 10$ and could be as high as six hours in some cases. The results presented in Figure 5 show that our algorithm is very attractive with respect to computational efficiency and is suitable for large data sets. In particular, note that even with $n = 10,000$ and $p = 10$, an interpreted language, and a slower microprocessor, our algorithm has a timing under 25 minutes. Detailed times are given in Table 2.

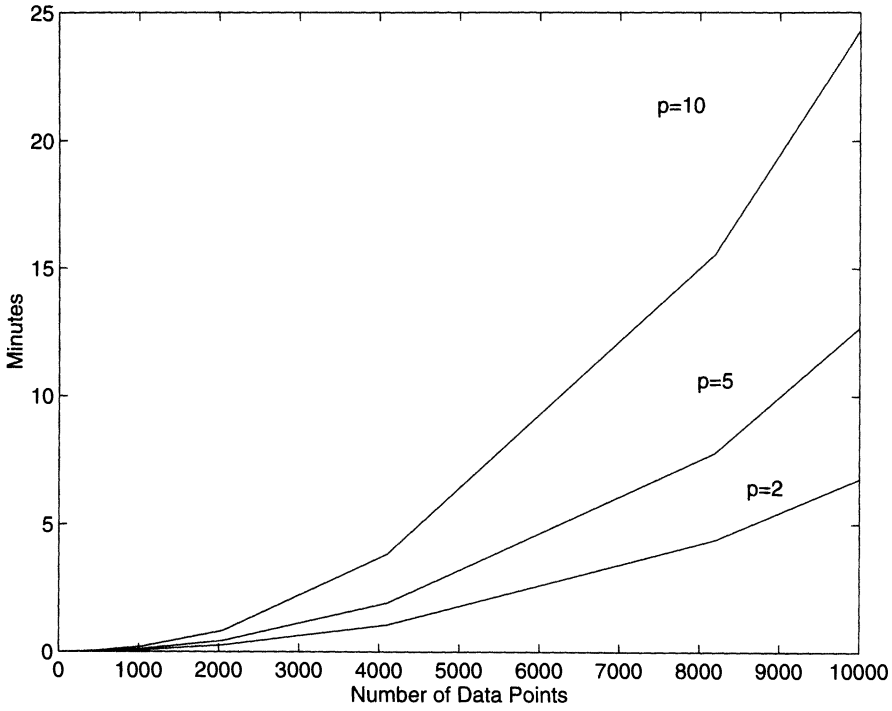

Figure 5. Time to Select Subset of Points.

6. SUMMARY

In this article, the EID method of determining the subset of points used in the MVE has been described. This method is a closed-form, deterministic algorithm as opposed to stochastic methods, which are currently in use. Subset selection is what makes the MVE a computationally expensive method to implement in daily practice. Results indicate that the EID method for selecting the set of points to be included in the MVE estimator is a useful one.

It should be noted that the EID method for subset selection could also be used to estimate the minimum covariance determinant (MCD) estimator (Atkinson 1994; Hawkins

Table 2. Time (minutes) to Select Subset of Points Using the EID

Size of data set	*p = 2*	*p = 5*	*p = 10*
16	0	0	0
32	0	0	0
64	0	0	0
128	0	0	0
256	0	.01	.01
512	.02	.03	.05
1024	.06	.10	.20
2048	.27	.44	.83
4096	1.06	1.91	3.83
8192	4.39	7.78	15.56
10000	6.77	12.67	24.31

1994; Rousseeuw 1985; Woodruff and Rocke 1994). The MCD is defined as the mean and covariance arising from the subset of m points that minimizes the determinant of the estimated covariance matrix. An estimate of the MCD could be obtained directly by the EID method and would preclude the need for finding the minimum covering ellipsoid. In addition to the obvious savings in computational effort, there is some evidence that the MCD is more efficient (Woodruff and Rocke 1994).

Although the EID method is not guaranteed to find the true MVE, it has certain advantages that make it more attractive than the algorithms currently in use. As discussed previously, it involves little computational effort, and thus it is suitable for sets with large n and p. Poston (1995) developed a version of the EID method that has been implemented in parallel. Hence, it is suitable for massive data set applications. Also, due to the iterative nature of the method, it would be easy to get a family of estimators for different values of m which is a useful feature (Hawkins 1993). Finally, this method would be useful as a starting point for other robust estimators such as S or M estimators (Woodruff and Rocke 1994).

ACKNOWLEDGMENTS

The authors thank the reviewers for their helpful comments that resulted in a much improved article. The work of Poston, Solka, and Priebe was supported by the NSWCDD In-house Laboratory Independent Research (ILIR) Program. Wegman's research was supported by the Office of Naval Research under Grant N00014-92-J-1303, the Army Research Office under Contract DAAL03-91-G-0039, and the National Science Foundation under Grant DMS9002237.

[Received August 1994. Revised November 1996.]

REFERENCES

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

Cook, R.D., Hawkins, D.M., and Weisberg, S. (1993), "Exact Iterative Computation of the Robust Multivariate Minimum Volume Ellipsoid Estimator," *Statistics and Probability Letters,* 16, 213–218.

Hawkins, D.M. (1993), "A Feasible Solution for the Minimum Volume Ellipsoid Estimator in Multivariate Data," *Computational Statistics,* 8, 95–107.

——— (1994), "The Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator," *Computational Statistics and Data Analysis,* 17, 197–210.

Kammer, D.C. (1991), "Sensor Placement for On-Orbit Modal Identification and Correlation of Large Space Structures," *AIAA Journal of Guidance, Control and Dynamics,* 14, 251–259.

Poston, W.L. (1995), "Optimal Subset Selection Methods," unpublished Ph.D. dissertation, George Mason University, Dept. of Applied and Engineering Statistics.

Poston, W.L., and Priebe, C.E. (1994), "Finding the Minimum Volume Ellipsoid," *Computing Science and Statistics,* 26, 351–355.

Poston, W.L., and Tolson, R.H. (1992), "Maximizing the Determinant of the Information Matrix With the Effective Independence Distribution Method," *AIAA Journal of Guidance, Control and Dynamics,* 15, 1513–1514.

Rousseeuw, P.J. (1985), "Multivariate Estimation With High Breakdown Point," in *Mathematical Statistics and Applications,* vol. B, eds. W. Grossmann, G. Pflug, I. Vincze, and W. Werz, Dordrecht: Reidel, p. 283–297.

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

Rousseeuw, P.J., and van Zomeren, B. (1990), "Unmasking Multivariate Outliers and Leverage Points" (with discussion), *Journal of the American Statistical Association,* 85, 633–651.

Titterington, D.M. (1975), "Optimal Design: Some Geometrical Aspects of d-Optimality," *Biometrika,* 62, 313–320.

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.

——— (1994), "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators," *Journal of the American Statistical Association,* 89, 888–896.