Grouping all the terms that correspond to the same powers of ω and using $\nu_1(\alpha+\lambda(1-\alpha))-(1-\alpha+\alpha\lambda^2)=(1-\lambda)(\alpha^2\lambda-(1-\alpha)^2)$, the result in (4) is obtained.

[Received March 1999. Revised June 2000.]

REFERENCES

Agulló, J. (1996), "Exact Iterative Computation of the Multivariate Minimum Volume Ellipsoid Estimator With a Branch and Bound Algorithm," in *Proceedings in Computational Statistics*, ed. A. Prat, Heidelberg: Physica-Verlag, pp. 175–180.

Arnold, H. J. (1964), "Permutation Support for Multivariate Techniques," *Biometrika*, 51, 65–70.

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

Balanda, K. P., and MacGillivray, H. L. (1988), "Kurtosis: A Critical Review," *The American Statistician*, 42, 111–119.

Becker, C., and Gather, U. (1999), "The Masking Breakdown Point of Multivariate Outlier Identification Rules," *Journal of the American Statistical Association*, 94, 947–955.

Box, G. E. P., and Tiao, G. C. (1968), "A Bayesian Approach to Some Outlier Problems," *Biometrika*, 55, 119–129.

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

——— (1989), "Bushfire Mapping Using NOAA AVHRR Data," technical report, CSIRO, North Ryde, Australia.

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.

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

Donoho, D. L. (1982), "Breakdown Properties of Multivariate Location Estimators," unpublished Ph.D. qualifying paper, Harvard University, Dept. of Statistics.

Gnanadesikan, R., and Kettenring, J. R. (1972), "Robust Estimates, Residuals, and Outliers Detection with Multiresponse Data," *Biometrics*, 28, 81–124.

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

——— (1994), "A Modification of a Method for the Detection of Outliers in Multivariate Samples," *Journal of the Royal Statistical Society*, Ser. B, 56, 393–396.

Hampel, F. R. (1985), "The Breakdown Point of the Mean Combined With Some Rejection Rules," *Technometrics*, 27, 95–107.

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

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

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

Jones, M. C., and Sibson, R. (1987), "What Is Projection Pursuit?" *Journal of the Royal Statistical Society*, Ser. A, 150, 29–30.

Juan, J., and Prieto, F. J. (1997), "Identification of Point-Mass Contaminations in Multivariate Samples," Working Paper 97–13, Statistics and Econometrics Series, Universidad Carlos III de Madrid.

Malkovich, J. F., and Afifi, A. A. (1973), "On Tests for Multivariate Normality," *Journal of the American Statistical Association*, 68, 176–179.

Mardia, K. V. (1970), "Measures of Multivariate Skewness and Kurtosis with Applications," *Biometrika*, 57, 519–530.

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

Maronna, R. A., and Yohai, V. J. (1995), "The Behavior of the Stahel–Donoho Robust Multivariate Estimator," *Journal of the American Statistical Association*, 90, 330–341.

Posse, C. (1995), "Tools for Two-Dimensional Exploratory Projection Pursuit," *Journal of Computational and Graphical Statistics*, 4, 83–100.

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

——— (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association*, 91, 1047–1061.

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

——— (1993), "A Resampling Design for Computing High-Breakdown Point Regression," *Statistics and Probability Letters*, 18, 125–128.

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

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

Rousseeuw, P. J., and van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633–639.

Ruppert, D. (1987), "What Is Kurtosis," *The American Statistician*, 41, 1–5.

Stahel, W. A. (1981), "Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen," unpublished Ph.D. thesis, Eidgenössische Technische Hochschule, Zurich.

Tyler, D. E. (1991), "Some Issues in the Robust Estimation of Multivariate Location and Scatter," in *Directions in Robust Statistics and Diagnostics, Part II*, eds. W. Stahel and S. Weisberg, New York: Springer-Verlag, pp. 327–336.

Wilks, S. S. (1963), "Multivariate Statistical Outliers," *Sankhya*, Ser. A, 25, 407–426.

Discussion

David M. Rocke

Department of Applied Science
University of California
Davis, CA 95616
(*dmrocke@ucdavis.edu*)

David L. Woodruff

Graduate School of Management
University of California
Davis, CA 95616
(*dlwoodruff@ucdavis.edu*)

Peña and Prieto present a new method for robust multivariate estimation of location and shape and identification of multivariate outliers. These problems are intimately connected in that identifying the outliers correctly automatically allows excellent robust estimation results and vice versa.

Some types of outliers are easy to find and some are difficult. In general, the previous literature concludes that problems are more difficult when the fraction of outliers is large. Moreover, widely scattered outliers are relatively easy, whereas

TECHNOMETRICS, AUGUST 2001, VOL. 43, NO. 3

concentrated outliers can be very difficult (Rocke and Woodruff1996). This article is aimed squarely at the case in which the outliers form a single cluster, separated from the main data and of the same shape (but possibly different size) as the main data, an especially difficult case for many outlier-detection methods (Rocke and Woodruff 1996). We believe that it is entirely appropriate that special methods like those in the present article be developed to handle this case, which is often too difficult for general-purpose outlier-detection and robust-estimation methods.

In this discussion, we make some comparisons of the estimators kurtosis1, FAST-MCD, SD, and also certain M and S estimators; then we point out the connection to cluster analysis (Rocke and Woodruff 2001). The MCD, MVE, SD, and S estimators with hard redescending influence functions are known to have maximum breakdown. Kurtosis1 probably also has maximum breakdown, although this article has no formal proof. M estimators are sometimes thought to be of breakdown $1/(p+1)$, but this is actually incorrect. Work by Maronna (1976), Donoho (1982), and Stahel (1981) showed that whenever the amount of contamination exceeds $1/(p+1)$ a root of the estimating equations that can be carried over all bounds exists. In fact, a root may exist that remains bounded. S estimators are a subclass of M estimators, as shown by Lopuhaä (1989), and have maximal breakdown when hard redescending ψ functions are used and the parameters are correctly chosen. This provides an example of a class of M estimators that have maximal breakdown.

M estimators can be highly statistically efficient and are easy to compute by iteratively reweighted least squares but need a high-breakdown initial estimator to avoid converging on the "bad" root (this also applies to S estimators, which are computed by solving the related constrained M estimation problem). MULTOUT (Rocke and Woodruff 1996) combines a robust initial estimator with an M estimator to yield a high-breakdown, statistically efficient methodology. Note also that identifying outliers by distance and reweighting points with weight 0 if the point is declared an outlier and weight 1 otherwise is a type of M estimator with a ψ function that is constant until the outlier rejection cutoff point. This is used by many methods (e.g., FAST-MCD, kurtosis1) as a final step, which improves statistical efficiency.

Thus, the key step in outlier identification and robust estimation is to use an initial procedure that gives a sufficiently good starting place. The gap can be quite large between the theoretical breakdown of an estimator and its "practical breakdown," the amount and type of contamination such that success is unlikely. Consider, for example, the case in Table 6 in which $n = 100$, $p = 20$, $\alpha = .3$, $\lambda = 1$, and $\delta = 100$. The amount of contamination is well below the breakdown point (40%), and the contamination is at a great distance from the main data, but none of the methods used in this study are very successful.

The study reported in Table 6 has some difficulties:

1. The number of data points is held fixed at 100 as the dimension rises, which leads to a very sparse data problem in dimension 20. Many of these methods could perhaps do much better with an adequate amount of data. This is particularly true of FAST-MCD and MULTOUT, which are designed to handle large amounts of data.

2. When $n = 100$ and $p = 20$, the maximum breakdown of any equivariant estimator is .4. Of course, this does not mean that the estimator must break down whatever the configuration of the outliers, but it does show that at $\alpha = .4$ and $p = 20$ and for every dataset generated there is an outlier configuration such that no equivariant estimator will work.

3. The case $\lambda = 5$ is one in which the outliers are actually scattered rather than concentrated. This is especially true when $\delta = 10$. With standard normal data in 20 dimensions, almost all observations lie at a distance less than the .001 point of a χ_{20} distribution, which is 6.73. The outliers are placed at a distance of 10, and the .999 sphere for these has a radius of $(5)(6.73) = 33.65$. Thus, the outliers actually surround the good data, rather than lying on one side. This explains why FAST-MCD gets all of the $\lambda = 5$ cases regardless of other parameter settings.

4. When $n = 100$ and $p = 20$, the preceding computation shows that standard normal data will almost all lie in a sphere of radius 6.73 around the center. If the outliers are displaced by 10 with the same covariance ($\lambda = 1$), these spheres overlap considerably, and in some instances no method can identify all "outliers" because some of them lie within the good data. Shrunken outliers ($\lambda = .1$) do not have this problem since the radius of the outliers is only .67, so a displacement of 10 prevents overlap. Expanded outliers ($\lambda = 5$) are relatively easy because the outliers surround the main data, and the chance of one falling in the relatively small sphere of the good data is small.

The results are better for two and four clusters (Table 8) than for one and would be even better for individually scattered outliers. The important insight is that many methods work well except in one hard case—when the outliers lie in one or two clusters. Methods such as the FAST-MCD and MULTOUT must be supplemented by methods specifically designed to find clustered or concentrated outliers. These clustered outlier methods are not replacements for the other methods because they may perform poorly when there are significant nonclustered outliers.

The remaining question concerns appropriate methods to supplement the more general estimation techniques and allow detection of clustered/concentrated outliers. The methods of Peña and Prieto are aimed at that task, by trying to find the direction in which the clustered outliers lie. The performance of these methods is documented in the article under discussion. Another attempt in this direction was given by Juan and Prieto (2001), who tried to find outliers by looking at the angles subtended by clusters of points projected on an ellipsoid. They showed reasonable performance of this method but provided computations only for very concentrated outliers ($\lambda = .01$).

Rocke and Woodruff (2001) presented another approach. If the most problematic case for methods like FAST-MCD and MULTOUT is when the outliers form clusters, why not apply methods of cluster analysis to identify them? There are many important aspects to this problem that cannot be treated in the limited space here, but we can look at a simple version of our procedure. We search for clusters using a model-based clustering framework and heuristic search optimization methods, then apply an M estimator to the largest identified cluster (for details, see Coleman and Woodruff 2000; Coleman et al. 1999; Reiners 1998; Rocke and Woodruff 2001).

To illustrate our point, we ran a slightly altered version of the simulation study in Table 6 from Peña and Prieto. First, we placed the outliers on a diagonal rather than on a coordinate axis. If $\boldsymbol{u} = (1, 1, \ldots, 1)$, then the mean of the outliers was taken to be $p^{-1/2}\delta\boldsymbol{u}$, which lies at a distance of δ from $\mathbf{0}$. After generating the data matrix $\boldsymbol{X}$ otherwise in the same way as did Peña and Prieto, we centered the data to form a new matrix $\boldsymbol{X}^*$ and then generated "sphered" data in the following way: Let $\boldsymbol{X}^* = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$, the singular value decomposition (SVD) in which $\boldsymbol{D}$ is the diagonal matrix of singular values. Then the matrix $\widetilde{\boldsymbol{X}} = \boldsymbol{X}^*\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}^\top = \boldsymbol{U}\boldsymbol{V}^\top$ is an affine transformed version of the data with observed mean vector $\mathbf{0}$ and observed covariance matrix $\boldsymbol{I}$. The purpose of these manipulations is to ensure that methods that may use nonaffine-equivariant methods do not have an undue advantage. Pushing the outliers on the diagonal avoids giving an advantage to componentwise methods. The SVD method of sphering, unlike the usual Cholesky factor approach, preserves the direction of displacement of the outliers.

We ran all the cases from the simulation of Peña and Prieto. Time constraints prevented us from running the 100 repetitions, but out of the five repetitions we did run, we succeeded in 100% of the cases in identifying all of the outliers, except for the cases in which $p = 20$ and $\lambda = 1$; these required more data for our methods to work. We could identify all of the outliers in 100% of the cases when $n = 500$, for example, instead of $n = 100$. The successful trials included all of the other cases in which no other method reported in Table 6 was very successful.

This insight transforms part of the outlier-identification problem into a cluster-analysis problem. However, the latter is not necessarily easy (Rocke and Woodruff 2001). For example, we tried the same set of simulation trials using the standard clustering methods available in S-PLUS. In this case, we ran 10 trials of each method. The methods used were two hierarchical agglomeration methods, mclust (Banfield and Raftery 1993) and agnes (Kaufman and Rousseeuw 1990; Struyf, Hubert, and Rousseeuw 1997); diana, a divisive hierarchical clustering method; fanny, a fuzzy clustering method; pam, clustering around medoids (Kaufman and Rousseeuw 1990; Struyf et al. 1997); and k-means (Hartigan 1975). First, it should be noted that all of these methods succeed almost all of the time for separated clusters ($\lambda = .1$ or $\lambda = 1$) if the data are not sphered. All of these methods except mclust use Euclidean distance and make no attempt at affine equivariance. Mclust uses an affine equivariant objective function based on a mixture model, but the initialization steps that are crucial to its performance are not equivariant.

Over the 72 cases considered (with sphered data), none of these methods achieved as much as 50% success. The overall success rates were agnes 36%, diana 48%, fanny 41%, k-means 10%, mclust 37%, and pam 12% (compared to MCD 75%, SD 77%, kurtosis1 83%, and our clustering method 89%). For shrunken outliers ($\lambda = .1$), the most successful were fanny 81% and mclust 37%, and the least successful were agnes 1% and k-means 2% (compared to MCD 41%, SD 70%, kurtosis1 88%, and our clustering method 100%). For expanded outliers ($\lambda = 5$), the most successful were agnes 97%, diana 96%, mclust 73%, and fanny 39%, and the least successful was pam 0% (compared to MCD 100%, SD 88%, kurtosis1 100%, and our clustering method 100%). This case is quite easy for robust multivariate estimation methods; FAST-MCD and MULTOUT each get 100% of these. Again, the most difficult case is shift outliers (Rocke and Woodruff 1996; Hawkins 1980, p. 104), in which $\lambda = 1$. The best performance among these clustering methods is diana 34%, with the next best being k-means at 16% (compared to MCD 82%, SD 73%, kurtosis1 62%, and our clustering method 67%).

The poor performance of these clustering methods is probably due to two factors. First, use of the Euclidean metric is devastating when the shape of the clusters is very nonspherical. Since this can certainly occur in practice, such methods should at least not be the only method of attack. Second, some of these problems require more extensive computation than these methods allow, at least in the implementation in S-PLUS. Performance of many descents of the algorithm from a wide variety of starting points, including random ones, can obtain better solutions than a single descent from a single plausible starting point. This is particularly true when the data are extremely sparse, as they are in these simulations. Many of these methods could, of course, be modified to use multiple starting points and might show enormously improved performance.

We confirmed this by using two additional model-based clustering methods that permit control of the amount of computation. The first of these was EMMIX (McLachlan and Basford 1988; McLachlan and Peel 2000). The second was EMSamp (Rocke and Dai 2001). For both programs, the performance was similar to that of our clustering method; success was usual if enough random starting points were used except for shift outliers in dimension 20 in the present $n = 100$ case.

Since different methods are better at different types of outliers and since once the outliers have been identified by any method they can be said to stay identified, an excellent strategy when using real data instead of simulated data (where the structure is known) is to use more than one method. Among the methods tested by Peña and Prieto, the best combination is FAST-MCD plus kurtosis1. Together these can improve the overall rate of identification from 75% and 83%, respectively, to 90%, since FAST-MCD is better for shift outliers and kurtosis1 is better for shrunken outliers. An even better combination is our clustering method plus FAST-MCD, which gets an estimated 95% of the cases.

We can summarize the most important points we have tried to make here as follows:

1. General-purpose robust-estimation and outlier-detection methods work well in the presence of scattered outliers or multiple clusters of outliers.
2. To deal with one or two clusters of outliers in difficult cases, methods specifically meant for this purpose such as those of Peña and Prieto and Juan and Prieto (2001) are needed to supplement the more general methods.

3. An alternative approach for this case is to use clustering methods to supplement the general-purpose robust methods.

4. Choice of clustering method and computational implementation are important determinants of success. We have found that model-based clustering together with heuristic search technology can provide high-quality methods (Coleman and Woodruff 2000; Coleman et al. 1999; Reiners 1998; Rocke and Dai 2001; Rocke and Woodruff 2001).

5. The greatest chance of success comes from use of multiple methods, at least one of which is a general-purpose method such as FAST-MCD and MULTOUT, and at least one of which is meant for clustered outliers, such as kurtosis1, the angle method of Juan and Prieto (2001), or our clustering method (Rocke and Woodruff 2001).

REFERENCES

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

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.

Coleman, D. A., and Woodruff, D. L. (2000), "Cluster Analysis for Large Datasets: Efficient Algorithms for Maximizing the Mixture Likelihood," *Journal of Computational and Graphical Statistics*, 9, 672–688.

Hartigan, J. A. (1975), *Clustering Algorithms*, New York: Wiley.

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

Juan, J., and Prieto, F. J. (2001), "Using Angles to Identify Concentrated Multivariate Outliers," *Technometrics*, 43, 311–322.

Kaufman, L., and Rousseeuw, P. J. (1990), *Finding Groups in Data*, New York: Wiley.

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

McLachlan, G. J., and Basford, K. E. (1988), *Mixture Models: Inference and Application to Clustering*, New York: Marcel Dekker.

McLachlan, G. J., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

Reiners, T. (1998), "Maximum Likelihood Clustering of Data Sets Using a Multilevel, Parallel Heuristic," unpublished Master's thesis, Technische Universität Braunschweig, Germany, Dept. of Economics, Business Computer Science and Information Management.

Rocke, D. M., and Dai, J. (2001), "Sampling and Subsampling for Cluster Analysis in Data Mining With Application to Sky Survey Data," unpublished manuscript.

Rocke, D. M., and Woodruff, D. L. (2001), "A Synthesis of Outlier Detection and Cluster Identification," unpublished manuscript.

Struyf, A., Hubert, M., and Rousseeuw, P. J. (1997), "Integrating Robust Clustering Techniques in S-Plus," *Computational Statistics and Data Analysis*, 26, 17–37.

Discussion

Mia Hubert

Department of Mathematics and Computer Science
University of Antwerp
Antwerp
Belgium
(*mia.hubert@ua.ac.be*)

Peña and Prieto propose a new algorithm to detect multivariate outliers. As a byproduct, the population scatter matrix is estimated by the classical empirical covariance matrix of the remaining data points.

The interest in outlier-detection procedures is growing fast since data mining has become a standard analysis tool in both industry and research. Once information ("data") is gathered, marketing people or researchers are not only interested in the behavior of the regular clients or measurements (the "good data points") but they also want to learn about the anomalous observations (the "outliers").

As pointed out by the authors, many different procedures have already been proposed over the last decades. However, none of them has a superior performance at all kinds of contamination patterns. The high-breakdown MCD covariance estimator of Rousseeuw (1984) is probably the most well known and most respected procedure. There are several reasons for this. First, the MCD has good statistical properties since it is affine equivariant and asymptotically normally distributed. It is also a highly robust estimator, achieving a breakdown value of 50% and a bounded influence function at any elliptical distribution (Croux and Haesbroeck 1999). Another advantage is the availability of a fast and efficient algorithm, called FAST-MCD (Rousseeuw and Van Driessen 1999), which is currently incorporated in S-PLUS and SAS. Therefore, the MCD can be used to robustify many multivariate techniques such as discriminant analysis (Hawkins and McLachlan 1997), principal-component analysis (PCA) (Croux and Haesbroeck 2000), and factor analysis (Pison, Rousseeuw, Filzmoser, and Croux 2000).

Whereas the primary goal of the MCD is to robustly estimate the multivariate location and scatter matrix, it also can be used to detect outliers by looking at the (squared) robust distances $\mathrm{RD}(x_i) = (x_i - T_{\mathrm{MCD}})^t S_{\mathrm{MCD}}^{-1}(x_i - T_{\mathrm{MCD}})$. Here, T_{MCD} and S_{MCD} stand for the MCD location and scatter estimates. One can compare these robust distances with the quantiles of the χ_p^2 distribution. Since this rejection rule often leads to an inflated Type II error, Hardin and Rocke (2000) developed more precise cutoff values to improve the MCD outlier-detection method.

If one is mainly interested in finding the outliers, it is less important to estimate the shape of the good points with great accuracy. As the authors explain, it seems natural to try to find