Partial Mixture Estimation and Outlier Detection in Data and Regression

David W. Scott

Abstract. The covariance matrix is a key component of many multivariate robust procedures, whether or not the data are assumed to be Gaussian. We examine the idea of robustly fitting a mixture of multivariate Gaussian densities in the situation when the number of components estimated is intentionally too few. Using a minimum distance criterion, we show how useful results may be obtained in practice. Application areas are numerous, and examples will be provided.

1. Introduction

The problem of identifying and handling outliers is an important problem in data analysis; see Barnett and Lewis [3] and Huber [11], for example. Finding outliers through interactive graphical exploration of small multivariate sets can be accomplished using Swayne, Cook, and Buja's XGobi system [24]; however, with very large medium-dimensional datasets nonparametric density algorithms such as Scott's ASH [18] can be recommended. Of course, purely graphical and nonparametric approaches to identifying outliers are prone to errors since many smooth densities give rise to data that may appear to have outliers, but do not. The Cauchy distribution is one example of such a density.

Thus the identification of outliers without an explicit probability model should always be viewed as preliminary and exploratory. If a probability model is known, then the tasks of parameter estimation and outlier identification can be more rigorously defined. However, even probability models are usually known only approximately at best, and hence outliers so identified are still subject to certain biases.

The assumption of a multivariate normal model is most common. Often data may be transformed so that normality holds approximately. Successful robust estimation of the mean and covariance allows one to tag outliers more than three

Received by the editors May 1, 2003.

1991 *Mathematics Subject Classification.* Primary 62F35; Secondary 62H12.

Key words and phrases. Mixture models, Covariance estimation, Minimum distance estimation.

This work was supported in part by NSF grant number DMS 02-04723 and NSF Digital Government contract number EIA-9983459.

or four standard units from the mean. Finding the correct shape of the covariance matrix is the more challenging and interesting task. This shape can be sought without assuming normality. The minimum volume ellipse (MVE) has been investigated by Rousseeuw and others [15, 16, 14, 6, 27, 10], for example. Finding the MVE exactly is a challenging combinatorial problem but reasonably good approximate solutions exist. The MVE approach is especially appealing when the number of outliers is nontrivial and/or the outliers themselves form clusters.

Another approach that can be quite successful in this setting was proposed by Aitkin and Wilson [1], who investigated fitting a gaussian mixture model for $\mathbf{x} \in \Re^p$

$$f(\mathbf{x}) = \sum_{k=1}^{K} w_k \phi(\mathbf{x}|\mu_k, \Sigma_k) \tag{1.1}$$

using the expectation-maximization (EM) algorithm [7]. This model is often used for other purposes such as cluster analysis [13]. Here we treat the outliers as members of "nuisance" clusters. Once successful estimation is achieved, we reorder the labels so that the cluster weights are in decreasing magnitude, $w_1 > w_2 > \cdots > w_K$. Then in many situations, we may view the smallest $K-1$ clusters as representing various kinds of outliers and outlying clusters, with (w_1, μ_1, Σ_1) being the parameters of interest. For example, the fraction of outliers would be estimated as being $1 - w_1$.

Getting the EM algorithm to work requires a number of steps. Choosing K is especially tricky since many of the outliers may be singletons. Even more difficult is getting good initial guesses for the mixture parameters, especially when K is much larger or smaller than the "correct" model. Finally, small clusters cannot support estimation of a full covariance matrix, so that special structure may be assumed, such as a diagonal form or pooling of covariance information. Nevertheless, as a conceptual framework for handling large numbers of outliers in a range of challenging situations, the mixture model has great intuitive appeal.

A practical concern with the Aitkin and Wilson approach is that while a normality assumption may make good sense for the "good" data, knowledge of the distribution of the outliers and outlier clusters is more suspect. Since the primary parameters of interest are (w_1, μ_1, Σ_1), a procedure which estimates only that component would be of great interest. In one sense, the MVE approach is capable of this task. In this paper, we propose an alternative estimation approach to EM that can also estimate only a subfraction of the components of the mixture model. Applications are numerous.

2. Mixture Estimation by Minimum Distance

Practical algorithms for estimating parameters in a model $f(x|\theta)$ by minimum distance have been discussed by Beran [5], for example, by minimizing Hellinger distance between the model and a kernel density estimate. Such an indirect approach is not feasible for a parameter-rich model such as the mixture model; see Scott [19]. One distance criterion, integrated squared error (ISE), affords better numerical properties. ISE has been the criterion of choice for nonparametric function estimation [18], and Rudemo [17] showed how to formulate a cross-validated estimate of ISE for histograms using a leave-one-out approach. Terrell [25] first proposed using the ISE as an alternative to maximum likelihood, with extensions by his student Kim [12]. The use of ISE for parametric estimation has been described by Hjort [9] and Scott [19, 23, 26, 20]. A more general divergence criterion (which includes ISE) has been described by Basu, et al. [4]. Mixture estimation by ISE has been discussed in Scott [19].

In the usual presentation, we seek to find the true parameter, θ_0, from the parametric model, $f(x|\theta)$. In our case, our estimate will be of the form $f(x|\hat{\theta})$, but the true density, $g(x)$, will not (necessarily) be from the parametric family, $f(x|\theta)$. Nevertheless, we seek to find the value of θ such that $f(x|\theta)$ is closest to $g(x)$ in the sense of integrated squared error.

$$\begin{aligned}\hat{\theta} &= \arg\min_\theta \left[\int \left[f(x|\theta) - g(x) \right]^2 \, dx \right] \\ &= \arg\min_\theta \left[\int f(x|\theta)^2 dx - 2 \int f(x|\theta) g(x) dx + \int g(x)^2 dx \right] \\ &= \arg\min_\theta \left[\int f(x|\theta)^2 dx - 2E\left[f(X|\theta) \right] \right] ,\end{aligned}$$

since $\int g(x)^2 dx$ is a constant in the second line with respect to choice of θ, and where X is a random variable from the true density, $g(x)$. Now for many models (including the normal mixture model), the first integral can be computed in closed form for any value of the parameter vector, θ. Given a random sample, an unbiased estimate of $E[f(X|\theta)]$ is the mean. Thus a completely data-based version of the ISE criterion is given by

$$\hat{\theta} = \arg\min_\theta \left[\int f(x|\theta)^2 dx - \frac{2}{n} \sum_{i=1}^{n} f(x_i|\theta) \right] . \tag{2.1}$$

In practice, this nonlinear optimization problem falls in the well-studied class of M-estimators [9, 4, 20], and $\hat{\theta}$ is often asymptotically normal. Scott [20] calls the value of $\hat{\theta}$ which minimizes Equation 2.1 the *L2E* estimator, since integrated squared error and the L_2 distance are equivalent.

A careful examination of the argument leading up to Equation 2.1 reveals that the density model, $f(x|\theta)$, need not be a density function, whereas the fact that $g(x)$ is a density was critical in estimating $E[f(X|\theta)]$. Thus we may consider an incomplete mixture model for $f(x|\theta)$. The simplest such model is the multivariate partial density component (MPDC) given by

$$f(\mathbf{x}|\theta) = w\,\phi(\mathbf{x}|\mu, \Sigma) \, , \tag{2.2}$$

where $\theta = (w, \mu, \Sigma)$. Equation 2.1 for the MPDC is given explicitly by

$$\hat{\theta} = \arg\min_{\theta} \left[w^2\, \phi(0|0, 2\Sigma) - \frac{2w}{n} \sum_{i=1}^{n} \phi(\mathbf{x}_i|\mu, \Sigma) \right] . \tag{2.3}$$

Splus code that finds θ is available from the author.

3. Bivariate Examples

A thorough simulation study of robust properties often requires a book; see the Princeton robustness study [2], for example. Here we show a sample of bivariate examples of normal mixtures where the number of components, K, ranges from two to five. We try to avoid any hidden assumptions, such as symmetry of outliers. There are no singleton outliers. In fact, most outliers are in clusters with at least 50 or 100 points. The main central cluster has 500 or 1000 points.

In the first five and eighth examples, the "true" mixture component is located at the origin, $\mu = (0, 0)^T$, with $\Sigma = I_2$, the identity matrix. In the bivariate case, the parameter estimated is six-dimensional:

$$\theta = (w, \mu_x, \mu_y, \sigma_{xx}, \sigma_{xy}, \sigma_{yy})^T .$$

The iteration was started with the maximum likelihood estimates of the parameters

$$\theta^{(0)} = (1, \bar{x}, \bar{y}, s_x^2, r s_x s_y, s_y^2)^T .$$

Both initial and final estimates of θ are displayed graphically in each figure as one-sigma ellipses.

Here are the particulars of the mixture samples for the four examples displayed in Figure 1:

(i) $K = 2$, $n_1 = 500$, $n_2 = 100$, $\mu_2 = (5, 5)^T$, $\Sigma_2 = I_2$;
(ii) $K = 3$, $n_1 = 500$, $n_2 = n_3 = 100$, $\mu_2 = (5, 5)^T$, $\mu_3 = -\mu_2$, $\Sigma_2 = \Sigma_3 = I_2$;
(iii) $K = 2$, $n_1 = 500$, $n_2 = 100$, $\mu_2 = (0, 0)^T$, $\Sigma_2 = 25\, I_2$;
(iv) $K = 2$, $n_1 = 500$, $n_2 = 100$, $\mu_2 = (7, 7)^T$, $\Sigma_2 = 9\, I_2$.

In the first frame of Figure 1, 100 points are centered at $(5, 5)^T$. The estimated value of $\hat{w} = 81.5\%$, which is very close to 500/600. Notice the lack of any visual correlation in the final estimate of $\hat{\Sigma}_1$.

In the second frame of Figure 1, two groups of 100 outlying points are centered at $(5, 5)^T$ and $(-5, -5)^T$. The estimated value of $\hat{w} = 68.3\%$, which is very close to 500/700. Again, the apparent correlation in the initial maximum likelihood estimate of Σ is correctly missing in the $L2E$ estimate.

In the third frame of Figure 1, the 100 outliers share the same center but their scale is 5 times greater. Given the overlap of "outliers" and "inliers," $\hat{w}$ is biased upwards. However, observe how $\hat{\Sigma}_1$ is much closer to I_2.

In the fourth frame of Figure 1, which is a hybrid of frames one and three, the 100 outlying points are centered at $(7,7)^T$ and their scale is $9\,I_2$. The estimated value of $\hat{w} = 84.1\%$, which is very close to 500/600. Notice the lack of any correlation in the final estimate of $\hat{\Sigma}_1$.

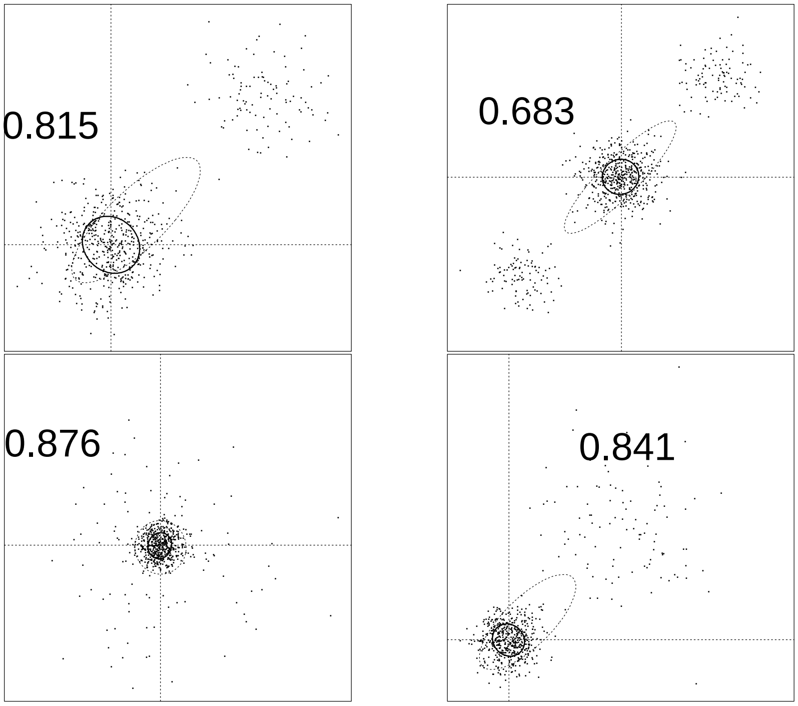

FIGURE 1. Multivariate partial density component estimation examples. The maximum likelihood one-sigma ellipse is shown as a dotted line, while the $L2E$ ellipse is shown as a thick solid line. $\hat{w}$ is also given. The origin is at the intersection of the two axes. See text for true parameter values.

Our second set of examples are displayed in Figure 2. In two of these, the true correlation coefficient of the main cluster is $\rho = 0.7$. Here are the particulars of the mixture samples for the four examples displayed in Figure 2:

(v) $K = 4$, $n_1 = 1000$, $n_2 = n_3 = n_4 = 100$, $\mu_2 = (5,5)^T$, $\mu_3 = -\mu_2$, $\mu_3 = (5,-5)^T$, $\Sigma_2 = \Sigma_3 = \Sigma_4 = I_2$;

(vi) $K = 2$, $n_1 = 500$, $\rho_1 = 0.7$, $n_2 = 100$, $\mu_2 = (5,5)^T$, $\Sigma_2 = I_2$;

(vii) $K = 2$, $n_1 = 500$, $\rho_1 = 0.7$, $n_2 = 100$, $\mu_2 = (5,-5)^T$, $\Sigma_2 = I_2$;

(viii) Multiple starting values with $K = 5$, $n_1 = 1000$, $n_2 = 50$, $n_3 = 100$, $n_4 = 150$, $n_5 = 200$, $\mu_2 = (5,5)^T$, $\mu_3 = (-5,-5)^T$, $\mu_4 = (5,-5)^T$, $\mu_5 = (-5,5)^T$, $\Sigma_2 = \Sigma_3 = \Sigma_4 = \Sigma_5 = I_2$.

In the first frame of Figure 2, three outlying clusters of 100 points each surround 1000 points at the origin. The estimated value of $\hat{w} = 74.9\%$, which is very close to 1000/1300.

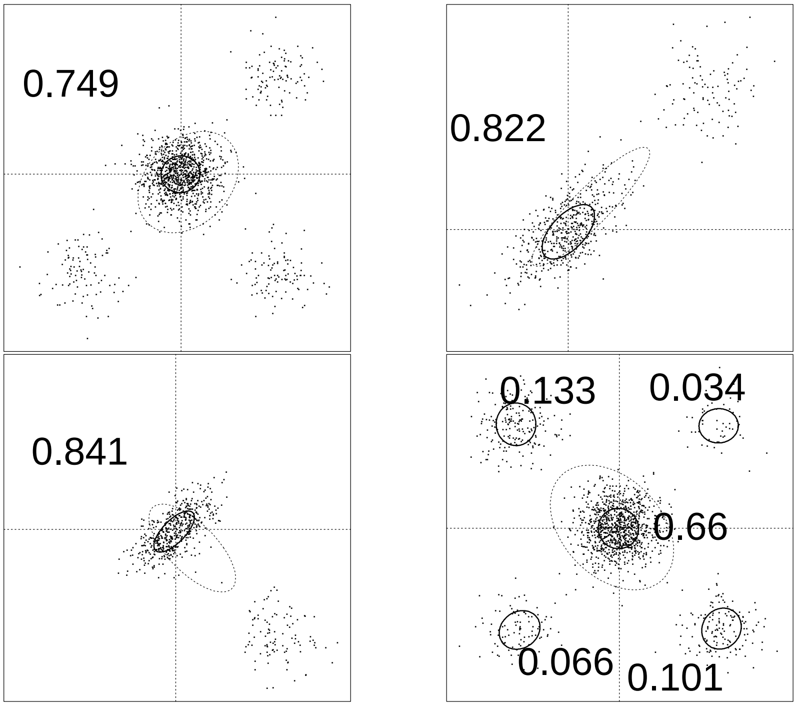

Figure 2. Second set of Multivariate partial density component estimate examples. In the final frame, solutions for five different initial guesses of θ are displayed. See text for true θ values.

In the second frame of Figure 2, the 500 center points have a correlation of $\rho = 0.7$. The group of 100 outlying points are centered at $(5, 5)^T$, which is in line with the main axis of the covariance matrix. Here, the apparent correlation in the initial maximum likelihood estimate of Σ is correctly retained in the $L2E$ estimate, and the center is correctly estimated, too.

In the third frame of Figure 2, the main cluster is identical, but now the 100 outliers are centered at $(5, -5)^T$. Thus the initial covariance estimate shows a strong, but negative, correlation, when the true correlation is strong, but positive. The $L2E$ estimates properly reorient the covariance matrix.

The fourth frame of Figure 2 is similar to the first, but with four outlying clusters (of different size 50, 100, 150, and 200). Starting from the maximum likelihood estimates, a $L2E$ MPDC component of size $\hat{w} = 0.66$ is found (actual value is 1000/1500).

Of special interest are the four other $L2E$ MPDC estimates displayed. These were found using different starting values for θ that were closer to the true parameters of each outlying cluster. This confirms the earlier suggestion that the "size" of a component is not as important as how separated the component is from the remaining data. The smallest cluster found (in one of five runs) has $\hat{w} = 0.034$, which is very close to 50/1500.

We have run a number of limited experiments in higher dimensions. The algorithm initially failed to converge on the Fisher Iris data. Closer examination revealed the reason to be that the data only had one decimal point. Adding a small amount of random noise (blurring) fixed the convergence problem.

Examples similar to those in Figure 1 were run in dimensions up to seven, with full covariance estimation. In seven dimensions, that amounts to $1 + 7 + 28 = 36$ quantities in the parameter vector, θ. An unfortunate aspect of the choice of integrated squared error is that ISE is not dimensionless (as is Hellinger distance, L_1 distance, maximum likelihood). The practical implication of this fact is that numerical optimization does not behave well due to scaling issues in these dimensions. Gaining a better understanding of this phenomenon and extending the range of dimensions where $L2E$ may be applied should be a valuable area of research.

4. Regression

The extension of $L2E$ to regression

$$y_i = \mathbf{x}_i^T \beta + \epsilon_i$$

is described and illustrated in Scott [20]. The algorithm is driven by a normal assumption on the residuals, $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. In our experience, we often see clusters of outliers in the regression plot, or even multiple regression curves mixed together. Here we briefly describe the extension of the MPDC approach to regression. The beauty of the regression setting is that the error distribution is univariate even with p predictor variables. Thus the MPDC is given by this assumption for the error variable:

$$\epsilon_i \sim w \cdot N(0, \sigma_\epsilon^2)\,.$$

The parameter vector estimated is $\theta = (w, \beta, \sigma_\epsilon)^T$, which is of length $p + 3$, assuming the parameter vector β includes an intercept term, β_0.

We illustrate an application to a well-known set of data first described by Harrison and Rubinfeld [8] on the median house value in census tracts in Boston, following the usual transformation of the $p = 13$ variables. The residuals from the least-squares fit were smoothed using a kernel estimate, as shown in the first frame of Figure 3. Also shown is a $N(0, \sigma_\epsilon^2)$ curve, with residual variance estimated in the usual manner. The residual distributions are not too close, but only a few outliers are apparent.

We next fit the $L2E$ model with noise distribution $w \cdot N(0, \sigma_\epsilon^2)$. The kernel smooth of the residuals and the estimated MPDC are both drawn in the right frame of Figure 3. Here, $\hat{w} = 84.5\%$. Interpreting and diagnosing the fit is much easier with the $L2E$ fit. For example, while there are some outliers on the low end, almost 14% of the outliers appear on the high side. In a more complete analysis where the outlying census tracts are interactively displayed on maps, we can learn that these tracts fall in certain regions (some along the Charles River) and not at random; see Scott and Christian [21].

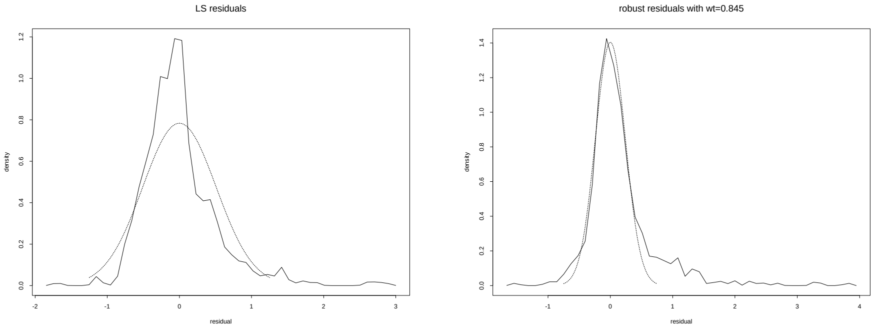

FIGURE 3. Kernel estimate of estimated residuals and fitted normal error density for the least squares fit (left frame) and the $L2E$ fit (right frame), for the Boston housing dataset.

5. Discussion

The PMDC approach is useful for outlier detection in many situations, and for clustering in particular; see Scott and Szewczyk [22]. The concept of breakdown is more complicated in this setting, as the algorithm is local. That is, the algorithm converges to local normal clusters for certain ranges of initial parameter settings. We have carefully examined the attractiveness of a single univariate component as a function of the parameters of an adjacent component. Generally, if the components overlap to a significant amount, the algorithm may not converge to either of the separate components, but rather to a single large component.

We have found a way to overcome this limitation by optimizing over a subset of the MPDC parameters. In Scott and Szewczyk [22], numerous random initial guesses are used and the collection of parameter estimates, θ, is clustered to find the most common solutions. We take these to be cluster locations.

Models other than normal may be chosen, however, the closed form expression of the $L2E$ criterion is very convenient for optimization. If the main cluster is not approximately normal, the $L2E$ solution may find the cluster, but the estimated parameters will vary depending upon the degree of non-normality.

There are numerous other extensions which we only touch upon here. For example, the MPDC may be formulated with more than one mixture component, or sequentially with some fixed mixture components, but then getting good initial estimates for θ becomes more difficult. Other regression applications may be formulated, including image processing tasks. We describe these separately.

References

[1] M. Aitkin and G.T. Wilson, *Mixture models, outliers, and the EM algorithm,* Technometrics, **22** (1980), 325–331.

[2] D.F. Andrews, P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers, and J.W. Tukey, *Robust Estimates of Location: Survey and Advances*, Princeton University Press, Princeton, 1972.

[3] V. Barnett and T. Lewis *Outliers in Statistical Data*, John Wiley & Sons, 1994.

[4] A. Basu, I.R. Harris, H.L. Hjort, and M.C.Jones, *Robust and Efficient Estimation by Minimising a Density Power Divergence.* Biometrika, **85** (1998), 549–560.

[5] R. Beran, *Robust Location Estimates,* The Annals of Statistics, **5** (1977), 431–444.

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

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm,* Journal of the Royal Statistical Society, Series B, **39** (1977), 1–22.

[8] D. Harrison and D.L. Rubinfeld, *Hedonic Housing Prices and the Demand for Clean Air,* Journal of Environmental Economics and Management, **5** (1978), 81–102.

[9] H.L. Hjort, *Minimum L2 and Robust Kullback-Leibler Estimation,* Proceedings of the 12th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, eds. P. Lachout and J.Á. Víšek, Prague Academy of Sciences of the Czech Republic, (1994) pp. 102–105.

[10] O. Hössjer, P.J. Rousseeuw, and C. Croux, *Asymptotics of an estimator of a robust spread functional.* Statistica Sinica, **6** (1996), 375–388.

[11] P.J. Huber, *Robust Statistics.* John Wiley & Sons, 1981.

[12] D. Kim, *Least Squares Mixture Decomposition Estimation,* Unpublished doctoral dissertation, Department of Statistics, Virginia Tech, Blacksburg, VA, (1995).

[13] G.J. McLachlan and D. Peel, *Finite mixture models*, John Wiley & Sons, 2001.

[14] W.L. Poston, E.J. Wegman, C.E. Priebe, and J.L. Solka, *A deterministic method for robust estimation of multivariate location and shape*, Journal of Computational and Graphical Statistics, **6** (1997), 300–313.

[15] P.J. Rousseeuw, *Multivariate estimation with high breakdown point.* In: W. Grossmann et al., editors, Mathematical Statistics and Applications, Vol. B, (1985) pp. 283–297, Akadémiai Kiadó: Budapest.

[16] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, New York, 1987.

[17] M. Rudemo, *Empirical Choice of Histogram and Kernel Density Estimators,* Scandinavian Journal of Statistics, **9** (1982), 65–78.

[18] D.W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization,* John Wiley, New York, 1992.

[19] D.W. Scott, *Remarks on Fitting and Interpreting Mixture Models,* Computing Science and Statistics, K. Berk and M. Pourahmadi, Eds., **31** (1999), 104–109.

[20] D.W. Scott, *Parametric Statistical Modeling by Minimum Integrated Square Error,* Technometrics, **43** (2001), 274–285.

[21] D.W. Scott and J.B. Christian, *Finding Outliers in Models of Spatial Data,* Proceedings of the Third National Conference on Digital Government, E Hovy, Ed., Digital Government Research Center, Boston, (2003) www.dgrc.org/dgo2003/start.html.

[22] D.W. Scott and W.F. Szewczyk, *The Stochastic Mode Tree and Clustering*, Journal of Computational and Graphical Statistics, **12** (2004), in press.

[23] D.W. Scott and W.F. Szewczyk, *From Kernels to Mixtures,* Technometrics, **43** (2001) 323–335.

[24] D.F. Swayne, D. Cook, and A. Buja, *XGobi: Interactive dynamic data visualization in the X Window System*, Journal of Computational and Graphical Statistics, **7** (1998), 113–130.

[25] G.R. Terrell, *Linear Density Estimates,* Proceedings of the Statistical Computing Section, American Statistical Association, (1990), 297–302.

[26] W.C. Wojciechowski and D.W. Scott, *Robust Location Estimation with L2 Distance,* Computing Science and Statistics, K. Berk and M. Pourahmadi, Eds., **31** (1999), 292–295.

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

Acknowledgment

The author would also like to thank William Szewczyk and the Department of Defense for support and collaboration.

Department of Statistics, MS-138, Rice University, P.O. Box 1892, Houston, Texas 77251-1892, USA
E-mail address: `scottdw@rice.edu` `http://www.stat.rice.edu/~scottdw`