OUTLIER DETECTION AND CLUSTERING BY PARTIAL MIXTURE MODELING

David W. Scott

Key words: Minimum distance estimation, robust estimation, exploratory data analysis..

COMPSTAT 2004 section: Statistical software.

Abstract: Clustering algorithms based upon nonparametric or semiparametric density estimation are of more theoretical interest than some of the distance-based hierarchical or ad hoc algorithmic procedures. However density estimation is subject to the curse of dimensionality so that care must be exercised. Clustering algorithms are sometimes described as biased since solutions may be highly influenced by initial configurations. Clusters may be associated with modes of a nonparametric density estimator or with components of a (normal) mixture estimator. Mode-finding algorithms are related to but different than gaussian mixture models. In this paper, we describe a hybrid algorithm which finds modes by fitting incomplete mixture models, or partial mixture component models. Problems with bias are reduced since the partial mixture model is fitted many times using carefully chosen random starting guesses. Many of these partial fits offer unique diagnostic information about the structure and features hidden in the data. We describe the algorithms and present some case studies.

1 Introduction

In this paper, we consider the problem of finding outliers and/or clusters through the use of the normal mixture model

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

Mixture models afford a very general family of densities. If the number of components, K, is quite large, then almost any density may be well-approximated by this model. Aitkin and Wilson (1980) first suggested using the mixture model as a way of handling data with multiple outliers, especially when some of the outliers group into clumps. They used the EM algorithm to fit the mixture model. Assuming that the "good" data are in one cluster and make up at least fifty percent of the total data, then it is easy to see that we have introduced a number of "nuisance parameters" into the problem (to model the outliers).

Implementing this idea in practice is challenging. If there are just a few "clusters" of outliers, then the number of nuisance parameters should not pose too much difficulty. However, as the dimension increases, the total number of parameters grows quite rapidly, especially if a completely general covariance matrix, Σ_k, is used for each component. The most directly challenging problem is finding an appropriate choice of the number of components, K, and initial guesses for the many parameters. An obvious first choice is to use a clustering algorithm such as k-means (MacQueen, 1967) as an approach to find an initial partition, and then compute the relative size, means, and covariances of each group to use as initial guesses for the EM algorithm.

It is abundantly clear that for many of our fits, we will in fact be using the wrong value of K. Furthermore, even if we happen to be using the appropriate value for K, there may be a number of different solutions, depending upon the specific initialization of the parameters. Starting with a large number of initial configurations is helpful, but as the dimension and sample size increase, the number of possibilities quickly exceeds our capabilities.

However, the least discussed and least understood problem arises because so little is generally known about the statistical distributions of the clusters representing the outliers. It certainly seems more reasonable to know something about the distribution of the "good" data; however, one is on much less firm ground trying to claim the same knowledge about the distributions of the several non-informative clusters. Even in the situation where the "good" data are in more than one cluster, sometimes little is known about the distribution in one or more of those "good" clusters.

In this paper, we discuss how an alternative to the EM algorithm can provide surprisingly useful estimates and diagnostics, even when K is incorrect. Such technology is especially interesting when K is too small, since in this situation the number of parameters to be estimated may be a small fraction of the number in the full, correct model. Furthermore, this technology is of special interest in the situation where little is known about the correct distribution of many of the clusters. This latter capability is of growing importance and interest in the analysis of massive datasets typically encountered in data mining applications.

2 Mixture Fits With Too Few Components

We examine some empirical results to reinforce these ideas. One well-known trimodal density in two dimensions is the lagged Old Faithful Geyser duration data, $\{(x_{t-1}, x_t),\ t = 2, \ldots, 298\}$; see Azzalini and Bowman (1990) and Weisberg (1985). Successive eruptions were observed and the duration of each eruption, $\{x_t,\, t = 1, \ldots, 299\}$, recorded to the nearest second. A quick count shows that 23, 2, and 53 of the original 299 values occurred exactly at $x_t = 2$, 3, and 4 minutes, respectively. Examining the original time sequence suggests that those measurements are clumped; perhaps accurate measurements were not taken after dark. We modified the data as follows: the 105 values that were only recorded to the nearest minute were blurred by adding uniform noise of 30 seconds in duration. Then all of the data were blurred by adding uniform noise, $U(-.5, .5)$, seconds, and then converted back into minutes.

In Figure 1, maximum likelihood estimates (MLE) of a bivariate normal and three two-component bivariate normal mixture fits are shown. Each bivariate normal density is represented by 3 elliptical contours at the 1, 2, and 3-σ levels. Figure 1 provides some examples of different solutions, depending upon the value of K selected and the starting values for the parameters chosen. In two dimensions, your eye can tell you what is wrong with these fits. In higher dimensions, diagnostics indicating a lack of fit leave unclear if a component should be split into two, or if the assumed shaped of the component is not correct.

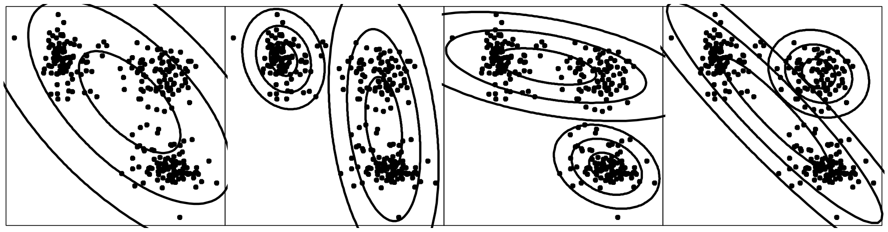

Figure 1: Maximum likelihood bivariate normal mixture fits to the lagged Old Faithful geyser eruption data with $K = 1$ and $K = 2$. The weights in each frame from L to R are (1.0), (.350, .650), (.645, .355), and (.728, .272). Each bivariate normal component is represented by 3 contours at the 1, 2, and 3-σ levels.

3 The L2E Criterion

Minimum distance estimation for parametric modeling of $f_\theta(x) = f(x|\theta)$ is a well-known alternative to maximum likelihood; see Beran (1984). In practice, several authors have suggested modeling the data with a nonparametric estimator (such as the histogram or kernel method), and then numerically finding the values of the parameters in the parametric model that minimize the distance between f_θ and the curve; see Beran (1977) and Brown and Hwang (1993), who considered Hellinger and L2 distances, respectively. Using a nonparametric curve as a target introduces some choices, such as the smoothing parameter, but also severely limits the dimension of the data and the number of parameters that can be modeled. (Precise numerical integration is quite expensive even in two dimensions. Numerical optimization algorithms require very good accuracy in order to numerically estimate the gradient vectors.)

Several authors have discovered an alternative criterion for parametric estimation in the case of L2 or integrated squared error (ISE); see Terrell (1990), Hjort (1994), Basu et al (1998), Scott (1998, 1999, 2001), for example.

(This idea follows from the pioneering work of Rudemo (1982) and Bowman (1984) on cross-validation of smoothing parameters in nonparametric density estimates.) In particular, Scott (1998,1999) considered estimation of mixture models by this technique. Given a true density, $g(x)$, and a model, $f_\theta(x)$, the goal is to find a fully data-based estimate of the L2 distance between g and f, which is then minimized with respect to θ. Expanding the L2 criterion

$$d(\hat{f}_\theta, g) = \int \left[\hat{f}_\theta(x) - g(x)\right]^2 dx \,, \tag{2}$$

we obtain the three integrals

$$d(\hat{f}_\theta, g) = \int \hat{f}_\theta(x)^2 dx - 2\int \hat{f}_\theta(x)\, g(x)\, dx + \int g(x)^2 dx \,. \tag{3}$$

The third integral is unknown but is constant with respect to θ and therefore may be ignored. The first integral is often available as a closed form expression that may be evaluated for any posited value of θ. Additionally, we must add an assumption on the model that this integral is always finite, i.e. $f_\theta \in L_2$. The second integral is simply the average height of the density estimate, given by $-2\,\mathrm{E}[\hat{f}_\theta(X)]$, where $X \sim g(x)$, and which may be estimated in an unbiased fashion by $-2n^{-1}\sum_{i=1}^n \hat{f}_\theta(x_i)$. Combining, the L2E criterion for parametric estimation is given by

$$\hat{\theta} = \arg\min_\theta \left[\int \hat{f}_\theta(x)^2 dx - \frac{2}{n}\sum_{i=1}^n \hat{f}_\theta(x_i)\right] \,. \tag{4}$$

For the multivariate normal mixture model in Equation 1,

$$\int_{\Re^d} \hat{f}_\theta(x)^2 dx = \sum_{k=1}^K \sum_{\ell=1}^K w_k\, w_\ell\, \phi(0\,|\,\mu_k - \mu_\ell, \Sigma_k + \Sigma_\ell). \tag{5}$$

Since this is a computationally feasible closed-form expression, estimation of the normal mixture model by the L2E procedure may be performed by use of any standard nonlinear optimization code; see Scott (1998, 1999). In particular, we used the *nlmin* routine in the Splus library for the examples in this paper.

Next, we return to the Old Faithful geyser example. Using the same starting values as in Figure 1, we computed the corresponding L2E estimates, which are displayed in Figure 2. Clearly, both algorithms are attracted to the same (local) estimates, which combine various clusters into one (since $K < 3$). However, there are interesting differences. First we compare the estimated weights: in Figure 1, the MLE weight of the larger component in each frame is 1, 0.65, 0.65, and 0.73, respectively, while in Figure 2 the corresponding L2E weights are 1, 0.74, 0.72, and 0.71. Of more interest, the L2E covariance matrices are either tighter or smaller. Since the (explicit) goal of L2E is to find the most normal fit (locally), observe that a number of points in the smaller clusters fall outside the 3-σ contours in frames 2 and 3 of Figure 2. The MLE covariance estimate is not robust and is inflated by those (slight) outliers. These differences are likely due to the inherent robustness properties of any minimum distance criterion; see Donoho and Liu (1988). Increasing the covariance matrix to "cover" a few outliers results in a large increase in the integrated squared or L2 error, and hence those points are largely ignored.

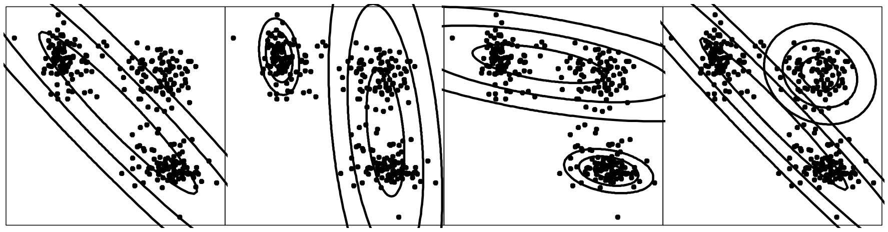

Figure 2: Several L2E mixture fits to the lagged Old Faithful geyser eruption data with $K = 1$ and $K = 2$; see text. The weights in each frame are (1.0), (.258, .742), (.714, .286), and (.711, .289).

4 Partial Mixture Modeling

The two-component L2E estimates above were computed with the constraint that $w_1 + w_2 = 1$. Is this constraint necessary? Can the weights w_1 and w_2 be treated as unconstrained variables? Certainly, when using EM or maximum likelihood, increasing the weights increases the likelihood without bound, so that the constraint is necessary (and active). However, *the L2E criterion does not require that the model $\hat{f}_\theta$ be a density.* The second integral in Equation 3 measures the average height of the density model, but a careful review of the argument leading to Equation 4 confirms the fact that only $g(x)$ is required to be a density, not $\hat{f}_\theta(x)$; see Scott (2001).

With this understanding, when we fit a L2E mixture model with $K = 2$, we are only assuming that the true mixture has at least 2 components. That is, we explicitly use our model for the local components of "good" data (local in the sense of our initial parameter guesses), but make no explicit assumption about the (unknown) distribution of the remaining data, no matter how many or few clusters they clump into. Our algorithm is entirely local. Different starting values may lead to quite different estimates.

Thus, we re-coded our L2E algorithm treating all of the weights in Equation 5 as *unconstrained* variables. In Figure 3, we display some of the "unconstrainted" L2E mixture estimates, using the same starting values as in Figure 2. These estimates are qualitatively quite similar to those in Figure

2, with some interesting differences. Comparing the first frames in Figures 2 and 3, the covariance matrix has narrowed as the weight decreased to .783. The sums of the (unconstrained) weights in the final three frames of Figure 3 are 0.947. 0.966, and 1.048. In the first two cases, the total probability modeled is less than unity, suggesting a small fraction of the data are being treated/labeled as outliers with respect to the fitted normal mixture model. The fact that the third total probability exceeds unity is consistent with our previous observation that the best fitting curve in the L2 or ISE sense often integrates to more than 1, when there is a gap in the middle of the data.

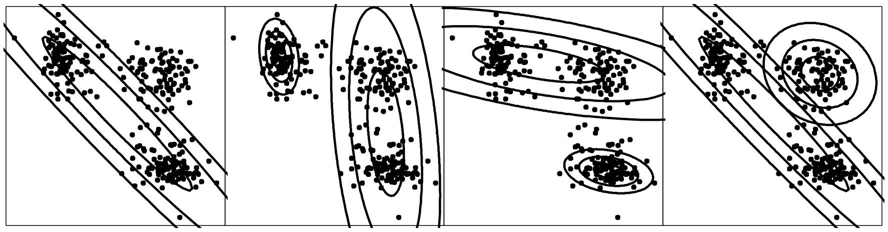

Figure 3: Several L2E partial mixture fits to the lagged Old Faithful geyser eruption data with $K = 1$ and $K = 2$, but without any constraints on the weights; see text. The weights in each frame are (.783), (.253, .694), (.683, .283), and (.751, .297).

Since there are potentially many more local solutions, we display four more L2E solutions in Figure 4. Some of these estimates are quite unexpected and deserve careful examination. The first frame is a variation of a $K = 1$ component which captures 2 clusters. However, the $K = 2$ estimates in the last 3 frames each capture two individual clusters, while completely ignoring the third. Comparing the contours in the last three frames of Figure 4, we see that exactly the same estimates appear in different pairs. Looking at the weights in Figures 3 and 4, we see that the smaller isolated components are almost exactly reproduced while entirely ignoring the third cluster. This feature of L2E is quite novel and we conclude that many of the local L2E results hold valuable diagnostic information as well as quite useful estimates of the local structure of the data.

Finally, in Figure 5, we conclude this investigation of the geyser data by checking a number of $K = 1$ unconstrained L2E solutions. In this case, the three individual components are found one at a time, depending upon the initial parameter values. Notice that the weights are identical to those in the previous figure. Furthermore, these weights are less than 50%, which is the usual breakdown point of robust algorithms; see Rousseeuw and Leroy (1987). However, the L2E algorithm is local and different ideas of breakdown apply.

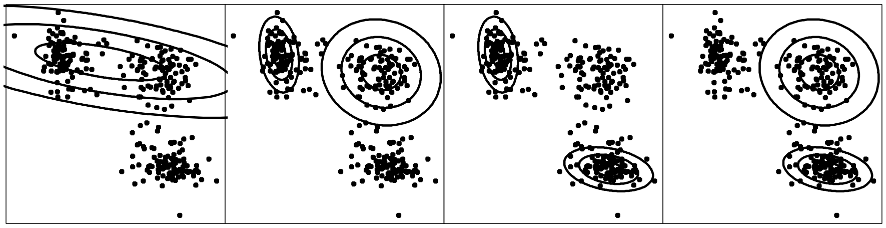

Figure 4: Same as Figure 3 but different starting values; see text. The weights in each frame are (.683), (.253, .316), (.253, .283), and (.316, .283).

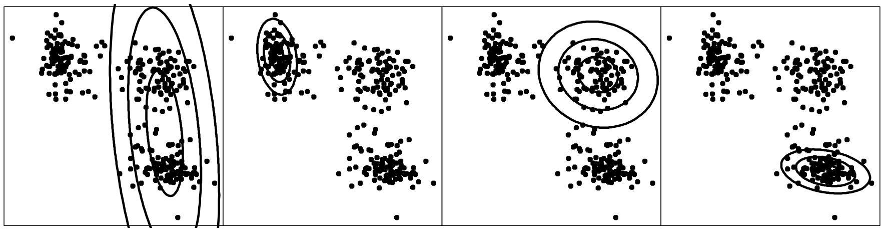

Figure 5: Four more $K = 1$ partial mixture fits to the geyser data; see text. The weights in each frame are (.694), (.253), (.316), and (.283).

5 Other Examples

5.1 Star Data

Another well-studied bivariate dataset was discussed by Rousseeuw and Leroy (1987). The data are measurements of the temperature and light intensity of 47 stars in the direction of Cygnus. For our analysis, the data were blurred by uniform $U(-.005, .005)$ noise. Four giant stars exert enough influence to distort the correlation of a least-squares or maximum likelihood estimate; see the first frame in Figure 7. In the second frame, a $K = 2$ MLE normal mixture is displayed. Notice the four giant stars are represented by one of the two mixture components and has a nearly singular covariance matrix. The third frame shows a $K = 1$ partial component mixture fit by L2E, with $\hat{w} = 0.937$. The shape of the two covariance matrices of the "good" data is somewhat different in these three frames. In particular, the correlation coefficients are -0.21, 0.61, and 0.73, respectively.

These data were recently re-analyzed by Wang and Raftery (2002) with nearest-neighbor variance estimator (NNVE), an extension of the NNBR estimator (Byers and Raftery, 1998). They compared their covariance estimates to the minimum volume ellipsoid (MVE) of Rousseeuw and Leroy (1987) as

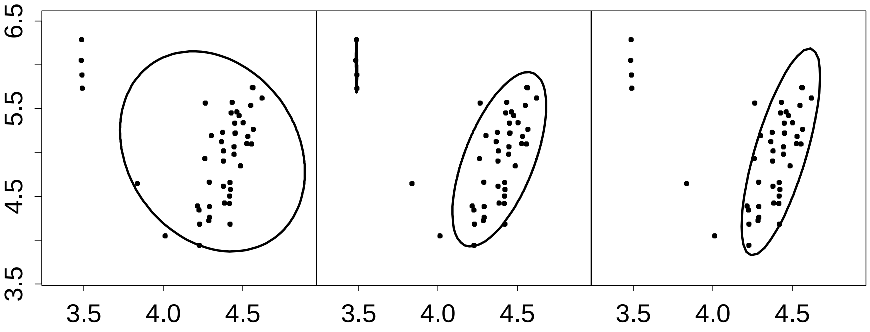

Figure 6: Two-σ contours of MLE ($K = 1$), MLE mixture ($K = 2$), and partial L2E mixture ($K = 1$) fits to the blurred star data.

well as the (non-robust) MLE. In Figure 7, I have overlaid these 4 covariance matrices (at the 1-σ contour level) with that of the partial density component (PDC) estimate obtained by L2E shown in the third frame of Figure 6. For convenience, I have centered these ellipses on the origin. The NNVE and NNBR ellipses are virtually identical, while the MVE ellipse is slightly rotated and narrower. These three are surrounded by the slightly elongated L2E PDC ellipse. Of course, the MLE has the wrong (non-robust) orientation. The correlation coefficients for NNVE and NNBR are 0.65 versus 0.73 for MVE and L2E. Observe that L2E does not explicitly require a search for the good data. The other three algorithms require extensive search and/or calibration of an auxiliary parameter. L2E is driven by the choice of the shape of the mixing distribution. One might choose instead to use t_ν components, as suggested by McLachlan and Peel (2001), although the degrees of freedom must be specified. In either case, L2E provides useful diagnostic information as a byproduct of the estimation, rather than as a follow-on step of analysis.

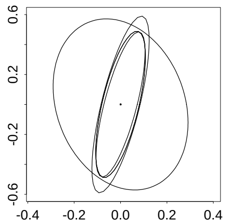

Figure 7: Ellipses representing the 2-σ contours of five estimates of the covariance matrix of the star data; see text.

5.2 Australian Athlete Data

For our final example, we consider four variables from the AIS data on Australian Athletes (Cook and Weisberg, 1994). These data are available in the R package with the command `data(ais,package='sn')`. Following Wang and Raftery (2002), we selected the variables body fat (BFAT), body mass index (BMI), red cell count (RCC), and lean body mass (LBM). (Wang and Raftery also included ferritin in their analysis.) We blurred the data then standardized each variable.

We fit a $K = 1$ L2E starting with the maximum likelihood estimate. The result was $\hat{w}_1 = 0.98$. A pairwise scatterdiagram of the 202 points is shown in Figure 8, together with contours of the fitted 4-dimensional ellipse. A careful examination of this plots suggests some clusters. In fact, the first 100 measurements are of female athletes and the last 102 measurements are of male athletes.

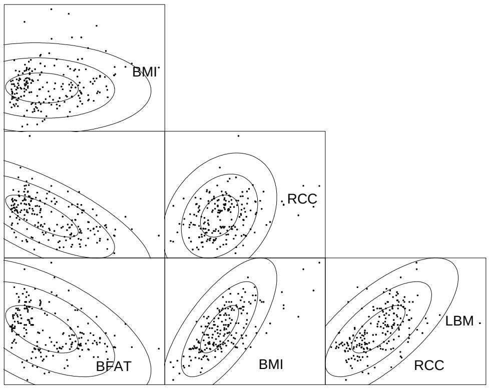

Figure 8: Ellipses representing the (1,2,3)-σ contours of a L2E partial mixture estimate of the Australian athlete data; see text.

Starting with the MLE values for the female athletes, we re-fit a $K = 1$ L2E. Now $\hat{w}_1 = 0.41$ (somewhat less than the 49.5% female population). The contours of the fitted 4-dimensional ellipse are superimposed upon the scatter matrix in Figure 9. The L2E is clearly modeling a large fraction of the female athletes.

Finally, we started the L2E with the male values. However, L2E found a smaller subset of the data lying in a subspace. (L2E is just as susceptible at MLE at being attracted to singular mixture components, depending upon initial guesses. That is why blurring was applied in all our examples to remove trivial singularities due to rounding.) Further experimentation would be interesting.

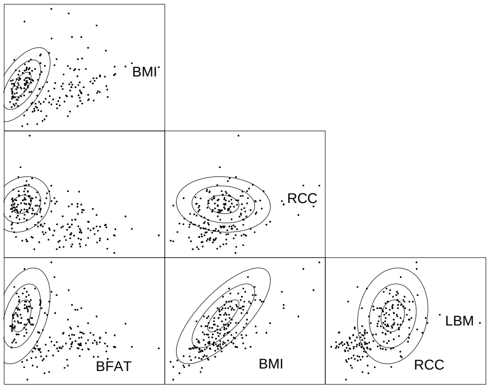

Figure 9: Ellipses representing the (1,2,3)-σ contours of a second L2E partial mixture estimate of the Australian athlete data; see text.

6 Discussion

We have shown how a minimum distance criterion and a mixture model with only one or two partial components can provide useful estimates and diagnostics. In particular, the value of $\hat{w}_1 + \hat{w}_2$ provides an indication of the fraction of the data being modeled by a $K = 2$ mixture. In our experience, the proportion of solutions that are interesting when $K = 2$ and the parameters are initialized by some random process is quite small. Further research on this question is open. However, many of the $K = 1$ solutions following random initialization are quite useful. The systematic use of these ideas for clustering is explored further in Scott and Szewczyk (2001).

Alternatively, Banfield and Raftery (1993) allow a number of outliers to be modeled as a spatial Poisson process. It would be interesting to apply that model with $K = 2$ to these data, where the noise is not Poisson, and to compare the parameter estimates.

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 general topic of outlier detection is discussed in Barnett and Lewis (1994). Robust estimation is described by Huber (1981). Coupled with a good exploratory such as XGobi (Swayne et al., 1998), the L2E PDC has much potential for helping unlock information in complex data.

References

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

[2] Azzalini, A. and Bowman, A.W. (1990), *A Look at Some Data on the Old Faithful Geyser,* Applied Statistics, **39**, 357 − 365.

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

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

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

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

[7] Beran, R. (1984), *Minimum Distance Procedures,* In Handbook of Statistics Volume 4: Nonparametric Methods, pp. 741 − 754.

[8] Bowman, A.W. (1984), *An alternative method of cross-validation for the smoothing of density estimates,* Biometrika, **71**, 353 − 360.

[9] Brown, L.D. and Hwang, J.T.G. (1993), *How To Approximate a Histogram By a Normal Density,* The American Statistician, **47**, 251 − 255.

[10] Byers, S. and Raftery, A.E. (1998), *Nearest-neighbor clutter removal for estimating features in spatial point processes,* Journal of the American Statistical Association, **93**, 577 − 584.

[11] Cook,R.D. and Weisberg, S. (1994), *An Introduction to Regression Graphics*, Wiley, New York.

[12] Donoho, D.L. and Liu, R.C. (1988), *The 'Automatic' Robustness of Minimum Distance Functional,* The Annals of Statistics, **16**, 552 − 586.

[13] Hjort, H.L. (1994), *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. Vı̃šek, Prague Academy of Sciences of the Czech Republic, pp. 102 − 105.

[14] Huber, P.J. (1981), *Robust Statistics.* John Wiley & Sons, New York.

[15] MacQueen, J.B. (1967), *Some Methods for Classification and Analysis of Multivariate Observations,* Proc. Symp. Math. Statist. Prob 5th Symposium, **1**, 281 − 297, Berkeley, CA.

[16] McLachlan, G.J. and Peel, D. (2001), *Finite mixture models*, John Wiley & Sons, New York.

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

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

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

[20] Scott, D.W. (1998), *On Fitting and Adapting of Density Estimates,* Computing Science and Statistics, S. Weisberg, Ed., **30**, 124 − 133.

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

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

[23] Scott, D.W. and Szewczyk, W.F. (2001), *The Stochastic Mode Tree and Clustering*, Journal of Computational and Graphical Statistics, under revision.

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

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

[26] Wang, N. and Raftery, A.E. (2002), *Nearest-neighbor variance estimation: Robust covariance estimation via nearest-neighbor cleaning,* Journal of the American Statistical Association, **97**, 994 − 1019.

[27] Weisberg, S. (1985), *Applied Linear Regression*, John Wiley, New York.

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

Address: Rice University, Department of Statistics, MS-138, POBox 1892, Houston, TX 77251-1892 USA

E-mail: `scottdw@rice.edu`