Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Parametric Statistical Modeling by Minimum Integrated Square Error

David W Scott

Parametric Statistical Modeling by Minimum Integrated Square Error

David W. Scott

Department of Statistics
Rice University
Houston, TX 77005-1892
(*scottdw@rice.edu*)

The likelihood function plays a central role in parametric and Bayesian estimation, as well as in nonparametric function estimation via local polynomial modeling. However, integrated square error has enjoyed a long tradition as the goodness-of-fit criterion of choice in nonparametric density estimation. In this article, I investigate the use of integrated square error, or L_2 distance, as a theoretical and practical estimation tool for a variety of *parametric* statistical models. I show that the asymptotic inefficiency of the parameters estimated by minimizing the integrated square error or L_2 estimation (L_2E) criterion versus the maximum likelihood estimator is roughly that of the median versus the mean. I demonstrate by example the well-known result that minimum distance estimators, including L_2E, are inherently robust; however, L_2E does not require specification of any tuning factors found in robust likelihood algorithms. L_2E is particularly appropriate for analyzing massive datasets in which data cleaning is impractical and statistical efficiency is a secondary concern. Setting up the L_2E criterion is relatively simple even with some very complex model specifications. Specific problems studied in this article include univariate density estimation, mixture density estimation, multivariate regression estimation, and robust estimation of the mean and covariance.

John Tukey had a pivotal role in both nonparametric and robust estimation. This article is dedicated to his memory.

KEY WORDS: Cross-validation; Minimum distance estimation; Normal mixture models; Outlier detection; Regression; Robust estimation.

Some of the practical deficiencies of maximum likelihood estimation are the lack of resistance to outliers and the general nonrobustness with respect to model misspecification. On the other hand, the class of minimum distance estimators has been shown to have excellent robustness properties (Beran 1977; Donoho and Liu 1988). Hellinger distance, $\int [\hat{f}(x)^{1/2} - f(x)^{1/2}]^2 dx$, and L_1 error, $\int |\hat{f}(x) - f(x)|\, dx$, have special attraction since they are dimensionless. However, these distance measures are not immediately available in practice, and certain approximations are typically encountered. For example, Beran (1977) suggested finding the parameter value, $\theta = \hat{\theta}$, that minimizes the Hellinger distance between $f(x|\theta)$ and a kernel density estimate, $\hat{f}_h(x)$. Brown and Hwang (1993) made a similar proposal but with a histogram. Since $\hat{\theta}$ varies with the choice of kernel or histogram smoothing parameter, a rule for determining h must be specified.

Parametric and nonparametric estimators seldom employ the same estimation criteria. Parametric algorithms typically rely on maximum likelihood while nonparametric algorithms favor the L_2 or integrated square error (ISE) criterion. However, the use of *local* likelihood and *local* least squares in nonparametric estimation is growing in popularity (Fan and Gijbels 1996). The present study arose from a desire to understand the applicability of the nonparametric criterion, ISE, to parametric problems.

Consider parametric estimation of the uniform density, $U(0, \theta)$, given a random sample $x_1, \ldots, x_n$. The maximum likelihood estimator (MLE) is $\hat{\theta} = x_{(n)}$, the largest order statistic. If alternative estimators of the form $\hat{\theta} = c \cdot x_{(n)}$ are entertained, then $c = (n+1)/n$ makes $\hat{\theta}$ unbiased, while $c = (n+2)/(n+1)$ minimizes mean squared error. On the other hand, Scott (1992) showed that $c = 2^{1/(n-1)}$ minimizes the average ISE or mean integrated square error (MISE), which is defined for a parametric estimator with true parameter $\theta = \theta_0$ as $\text{MISE}(\hat{\theta}) = E_{\hat{\theta}} \int [f(x|\hat{\theta}) - f(x|\theta_0)]^2 dx$; note that $2^{1/(n-1)} \approx 1 + (n-1)^{-1} \log 2$. All three estimators are slightly larger than the MLE. This simple example highlights the most important advantage of the MLE—namely, its constructive nature. The other criteria were applied only in a one-dimensional neighborhood of the MLE. Furthermore, in (other) regular cases, MLE's generally enjoy asymptotic optimality properties.

In this article, a fully constructive parametric estimation algorithm is devised based on the ISE criterion. The ISE- or L_2-minimizing estimate is abbreviated as L_2E. Its robustness behavior is demonstrated by an example and through the induced M estimator. Finally, it is shown how the basic density estimation framework may be extended to estimation in general statistical models.

1. MOTIVATION

The process of building useful models invokes a sequence of steps involving problem definition, estimation, criticism, reformulation, and corrective actions. Parametric models

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

approximate truth to varying degrees, complicated by any data contamination. Tukey organized a careful study of location estimators with symmetric contamination (Andrews et al. 1972). Of more general application are algorithms that control the influence of bad data. Such robust algorithms (Hampel 1974; Huber 1981) bound the influence of any datum. Maximizing the likelihood, $\sum_{i=1}^{n} \log f(x_i|\theta)$, means solving the equation $\sum_{i=1}^{n} \psi(x_i, \theta) = 0$, where $\psi = f'/f$. Robust M estimators have the same form of the estimating equation but use different choices for the influence function, ψ. For example, Tukey (Beaton and Tukey 1974) proposed the popular biweight $\psi(x) = x(r^2 - x^2)^2$ for $|x| < r$ and 0 elsewhere, where r is a scale parameter. Although many other specific forms have been proposed for the shape of ψ, each has a scale parameter whose choice is critical for success. This scale parameter may be determined by a prior robust estimate or by iteration.

In contrast, the minimum distance estimators described here implicitly define the shape and scale of the influence function as a byproduct of an *explicit* parametric assumption of the underlying density. Since the scale is estimated simultaneously with the model parameters, the fitted model is often much easier to properly evaluate when a large contamination exists (as will be illustrated). Intuitively, the minimum distance estimator tries to find the largest portion of the data that "matches" the model. In many instances, the analyst can easily identify the data not well fitted and take effective corrective actions.

2. THE PARAMETRIC MINIMUM L_2 ERROR CRITERION

The motivation for parametric L_2E originates in the derivation of the nonparametric least squares cross-validation algorithm for choosing the bin width, h, of a histogram (Rudemo 1982; Bowman 1984). The role played by the parameter h may be viewed quite generally. Let the estimate of an integral be denoted by placing a hat above the integral sign, $\widehat{\int} g(x)\, dx$. Consider minimizing an estimate of ISE with respect to h:

$$\hat{h} = \arg\min_h \widehat{\int} \left[\hat{f}_h(x) - f(x)\right]^2 dx \tag{1}$$

$$= \arg\min_h \left[\widehat{\int} \hat{f}_h(x)^2 dx - 2\widehat{\int} \hat{f}_h(x) f(x) dx + \widehat{\int} f(x)^2 dx\right] \tag{2}$$

$$= \arg\min_h \left[\int \hat{f}_h(x)^2 dx - 2\widehat{E}\left[\hat{f}_h(X)\right]\right], \tag{3}$$

since the minimizing value of h is not changed by eliminating $\int f(x)^2 dx$, an (unknown) constant, from Equation (2). Furthermore, the first integral in (2) can be evaluated exactly for any value of h (and hence does not require estimation). The remaining term in equation (2) is the average height of the histogram with bin width h. Rudemo (1982) proposed an unbiased estimate by partitioning the sample into $n-1$ points for estimation and 1 point for evaluation, $\hat{f}_{h,-i}(x_i)$, and then cycling over all n points and averaging so that

$$\hat{h} = \arg\min_h \left[\int \hat{f}_h(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(x_i)\right]$$

$$= \arg\min_h \left[\frac{2}{(n-1)h} - \frac{n+1}{n^2(n-1)h}\sum_k \nu_k^2\right], \tag{4}$$

where ν_k is the bin count of $B_k = (x_0 + kh, x_0 + (k+1)h]$ and x_0 is the bin origin.

In the parametric setting with model $f(x|\theta)$, Equation (1) may be rewritten with θ replacing h as the unknown parameter:

$$\hat{\theta} = \arg\min_\theta \widehat{\int} [f(x|\theta) - f(x|\theta_0)]^2\, dx, \tag{5}$$

where the true parameter θ_0 is unknown. (Hence, $\hat{\theta} = \theta_0$ is not available as the optimal estimator.) Once again, the expected height of the density, $\int f(x|\theta) f(x|\theta_0)\, dx$, is the key quantity to estimate. Data partitioning is not required in the parametric setting, since the entire random sample is available to estimate the average height of $f(x|\theta)$. Thus, the proposed estimator minimizing the parametric ISE criterion is

$$\hat{\theta}_{L_2E} = \arg\min_\theta \left[\int f(x|\theta)^2 dx - \frac{2}{n}\sum_{i=1}^{n} f(x_i|\theta)\right]. \tag{6}$$

Here I have assumed the correct parametric family; however, Equation (6) may also be applied in the case in which the assumed parametric form is known to be incorrect to achieve robustness.

Next, I introduce several examples of L_2E functionals. Some related simulations are presented in Section 6. The univariate and multivariate normal density will be denoted by $\phi(x|\mu, \sigma^2)$ and $\phi(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$, respectively.

Example 2.1 (Normal Density). Suppose $X \sim \mathrm{N}(\mu, 1)$. Then

$$\hat{\mu}_{\mathrm{MLE}} = \arg\max_\mu \sum_{i=1}^{n} \log \phi(x_i|\mu, 1)$$

$$\hat{\mu}_{L_2E} = \arg\min_\mu \left[\frac{1}{2\sqrt{\pi}} - \frac{2}{n}\sum_{i=1}^{n} \phi(x_i|\mu, 1)\right].$$

Observe that L_2E maximizes the *sum* of the densities while MLE maximizes the *product* of the densities. The robustness of $\hat{\mu}_{L_2E}$ can be easily demonstrated empirically for this problem in an interesting setting. Consider a sample of size 100 from $\mathrm{N}(0, 1)$ with up to 25 additional samples from a contamination density, $\mathrm{N}(5, 1)$. In Figure 1, I plot the log-likelihood and L_2E criteria for $n = 100, 105, \ldots, 125$, always adding five new samples from the contamination density while retaining the previous samples. Of course, a closed-form solution $\hat{\mu}_{MLE} = \bar{x}$ is available so that plotting the likelihood is not necessary. The upper right frame in Figure 1 illustrates how resistant L_2E is to the contaminated data. However, the lower frames, which are plotted over a wider interval, reveal a fuller story. The likelihood shows little sign of the contaminated data, but the L_2E curves show a local minimum near $\mu = 5$. The existence of the local minimum is appropriate since the contaminated data also come from the assumed parametric model, $\mathrm{N}(\mu, 1)$. Studying how the amount of con-

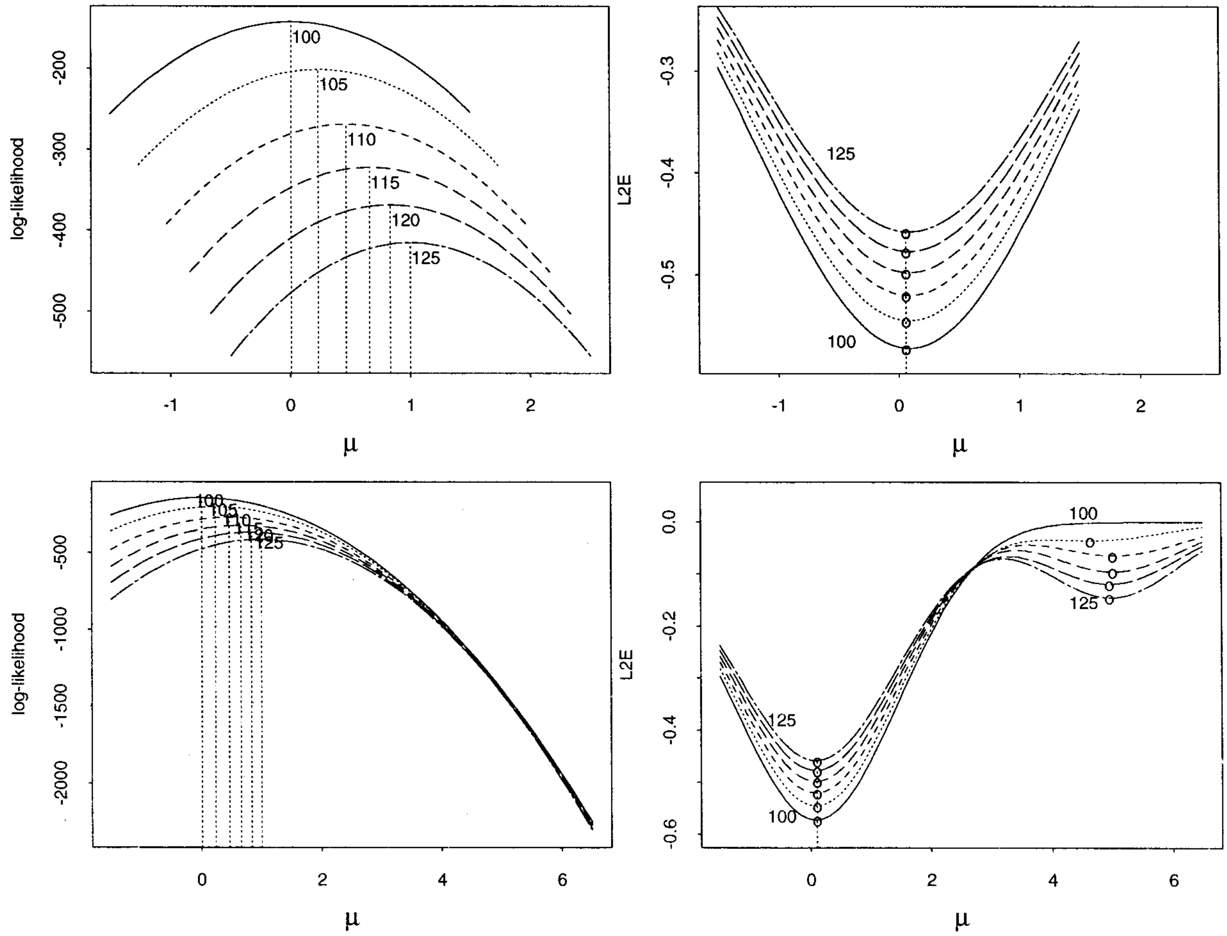

Figure 1. Upper Left: Log-likelihood Profiles for Contaminated Data (n = 100, 105, 110, 115, 120, 125). Best values are indicated by circles and/or dotted lines. Upper right: L_2E profiles for same data. Bottom left: MLE profiles on a larger interval. Bottom right: L_2E profiles on a larger interval. (Note that the line type corresponding to each sample size is the same in all four frames.)

tamination affects the level of the curves, I judge from the lower right frame that the values at the two minima would be approximately equal when $n = 200$. This and other empirical properties of the L_2E may be exploited in practice.

Example 2.2 (Normal Influence Functions). Recall that the influence function for *MLE* is given by $\psi = f'/f$. Thus, for the normal model at the particular values $(\mu, \sigma) = (0, 1)$, $\psi(x) = x$ for μ, and $\psi(x) = x^2 - 1$ for σ. Both are unbounded. Now the L_2E criterion for the two-parameter normal is

$$L_2E(\mu, \sigma) = \frac{1}{2\sqrt{\pi}\sigma} - \frac{2}{n}\sum_{i=1}^{n}\phi(x_i|\mu, \sigma^2)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2\sqrt{\pi}\sigma} - 2\phi(x_i|\mu, \sigma^2)\right]. \quad (7)$$

Thus the ψ function for L_2E is the derivative of the bracketed quantity, leading to $\psi(x) \propto x\exp(-x^2/2)$ for μ and $\psi(x) \propto (2\sqrt{2})^{-1} + (x^2 - 1)\exp(-x^2/2)$ for σ. The shapes of the influence functions for MLE and L_2E are similar for small values of the data. Interestingly, ψ is a "redescending" function for μ, but ψ is a "bounded" function for σ. Of practical importance is the fact that robust scaling issues for the ψ function are automatic and do not require iteration. Specification of the functional form for $f(x|\theta)$ in L_2E obviates the need for specification of the shape and scale of a ψ function and any iteration.

Example 2.3 (Multivariate Normal Density). Suppose $X \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$. Then

$$L_2E(\boldsymbol{\mu}, \Sigma) = \frac{1}{2^d\pi^{d/2}|\Sigma|^{1/2}} - \frac{2}{n}\sum_{i=1}^{n}\phi(\boldsymbol{x}_i|\boldsymbol{\mu}, \Sigma).$$

This example provides a simple demonstration of the multivariate extension of L_2E.

Example 2.4 (Uniform Density) Suppose $X \sim \mathrm{U}(0, \theta)$. Then

$$L_2E(\theta) = \frac{1}{\theta} - \frac{2}{n\theta}\sum_{i=1}^{n}I[x_i \le \theta].$$

For most samples, the L_2E will turn out to equal the MLE estimate, $x_{(n)}$. Recall that the MISE estimator of θ is slightly larger than $x_{(n)}$, but the data-based L_2E estimator is not. However, if the ratio of the adjacent order statistics is sufficiently large, then $x_{(n)}$ will not be the L_2E minimizer; see an example in Figure 2. In particular, if $x_{(n)}/x_{(n-1)} > n/(n-2)$, then $\hat{\theta}_{L_2E} \neq x_{(n)}$; that is, $\hat{\theta}_{L_2E} < x_{(n)}$ if $x_{(n)}$ is an "outlier." The ratio for other order statistics can be derived similarly.

Example 2.5 (Discrete Random Variables). The loss function analogous to (5) for discrete random variables is

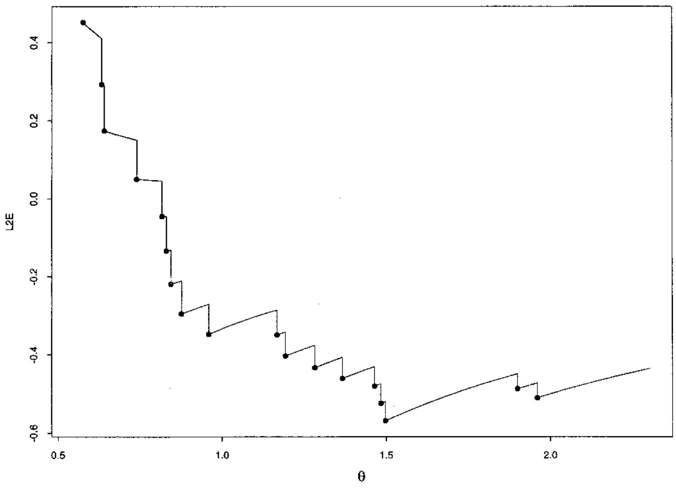

Figure 2. The L_2E Function When $f(x|\theta)$ Is $U(0,\theta)$ for a Sample of 25 Points From $U(0,1.5)$ With Two Contaminated Points Added at 1.90 and 1.96. $\theta_{L_2E} = 1.498$ in this case.

$\sum_x [f(x|\theta) - f(x|\theta_0)]^2$, which leads to the criterion

$$L_2E(\theta) = \sum_x f(x|\theta)^2 - \frac{2}{n}\sum_{i=1}^{n} f(x_i|\theta). \tag{8}$$

The first sum is over all values of the discrete random variable.

Example 2.6 (Poisson Density). Barnett and Lewis (1994) summarized tests to decide if the one or two largest data points from a Poisson sample are outliers. I generated a sample of size 100 from a Poisson density with mean 5. In this sample, only the values $0, 1, \ldots, 12$ were observed, and occurred with frequency $(1, 6, 9, 15, 19, 11, 15, 6, 10, 6, 0, 1, 1)$, totaling 100. Now the mean is 4.89 and the L_2E minimizer is 4.80. However, when eight outliers were inserted at $15, 20, 25, \ldots, 50$, the mean increased to 6.94 while the L_2E minimizer only increased to 4.85. Since the L_2E criterion is a continuous function of λ, finding $\hat{\lambda}$ is very easy by graphical or numerical techniques, with or without outliers.

An interesting open problem is handling densities that are mixtures of continuous and discrete components. Perhaps a weighted average of the criteria in (6) and (8) would work.

3. ORIGINS AND ASYMPTOTICS

The first suggestion of replacing the likelihood function with L_2E was given by Terrell (1990), who proposed an alternative to nonparametric penalized-likelihood estimators. The L_2E criterion for parametric problems was rediscovered by Hjort (1994) and later by Scott (1998). In an inspired article, Basu, Harris, Hjort, and Jones (1998) included MLE and L_2E within a general family of minimum-divergence estimators, indexed by a metaparameter $\alpha > 0$, given by

$$\hat{\theta}_\alpha = \arg\min_\theta \left[\int f(x|\theta)^{1+\alpha} dx - \frac{1+\alpha}{n\,\alpha} \sum_{i=1}^{n} f(x_i|\theta)^\alpha \right]. \tag{9}$$

L_2E corresponds to $\alpha = 1$. MLE corresponds to $\alpha \to 0$.

Hjort (1994) and Scott (1998) demonstrated the consistency and asymptotic normality of the L_2E parameters, summarized in the following proposition. The more general result of Basu et al. (1998) closely follows Lehmann's (1983, theorem 6.4.1) conditions for the MLE. Less restrictive assumptions are required for L_2E when $\alpha = 1$; the interested reader is directed to Jurečková and Sen (1996). In practice, the L_2E functional may not be strictly convex so that consistency is understood to hold in a neighborhood of θ_0. For complex models, generating random starting guesses to try to avoid local minima is suggested.

Proposition 3.1 (Asymptotic Normality). If $\boldsymbol{\theta}$ is a vector of parameters, then under mild conditions, the L_2E parameters are consistent and asymptotically normal:

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \text{AN}\left(\mathbf{0}, H^{-1}\left[G_2 - G_1 G_1^T\right] H^{-1}\right),$$

where

$$G_1 = \int_{\Re^p} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}|\boldsymbol{\theta}_0) f(\boldsymbol{x}|\boldsymbol{\theta}_0)\, d\boldsymbol{x},$$

$$G_2 = \int_{\Re^p} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}|\boldsymbol{\theta}_0) \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}|\boldsymbol{\theta}_0)^T f(\boldsymbol{x}|\boldsymbol{\theta}_0)\, d\boldsymbol{x},$$

and

$$H = \int_{\Re^p} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}|\boldsymbol{\theta}_0) \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}|\boldsymbol{\theta}_0)^T\, d\boldsymbol{x}.$$

Example 3.1 (Normal Density). If $X \sim \text{N}(\mu, \sigma^2)$, then Proposition 3.1 gives

$$\sqrt{n}\begin{pmatrix} \hat{\mu} - \mu \\ \hat{\sigma} - \sigma \end{pmatrix} = \text{AN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \text{diag}\left(\frac{8\sigma^2}{3\sqrt{3}}, \frac{4(16\sqrt{3} - 9)\sigma^2}{81} \right) \right).$$

In this example, the L_2E parameters are asymptotically uncorrelated. The standard deviation of $\hat{\mu}$ is 24.1% greater than that of $\bar{x}$ (the standard error of the median is 25.3% greater). The standard deviation of $\hat{\sigma}$ is 36.0% greater than that of the sample standard deviation (the standard deviation of the interquartile range/1.349 is 65.0% greater); see Kendall and Stuart (1977, vol. I, sec. 10.12).

4. FITTING GAUSSIAN MIXTURE DENSITIES

A powerful parametric density model is the mixture model (Titterington, Smith, and Makov 1985)

$$f(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k\, \phi(x|\mu_k, \sigma_k^2).$$

In practice, mixture fitting is often a difficult task. Among the many maximum likelihood solutions are Dirac spikes ("infinite" likelihood) so that a local solution is desired. Estimation is facilitated by knowing the correct number of components. The EM algorithm is generally favored, although Ripley (1996) recommended directly optimizing the likelihood with Newton methods. In my experience, the EM algorithm is preferred for very hard problems (overparameterized, high-dimensional).

The L_2E criterion is particularly easy to apply with the use of the following identity:

$$\int_{-\infty}^{\infty} \phi(x|\mu_1, \sigma_1^2)\, \phi(x|\mu_2, \sigma_2^2)\, dx = \phi(\mu_1 - \mu_2|0, \sigma_1^2 + \sigma_2^2),$$

one of many useful formulas given by Wand and Jones (1995).

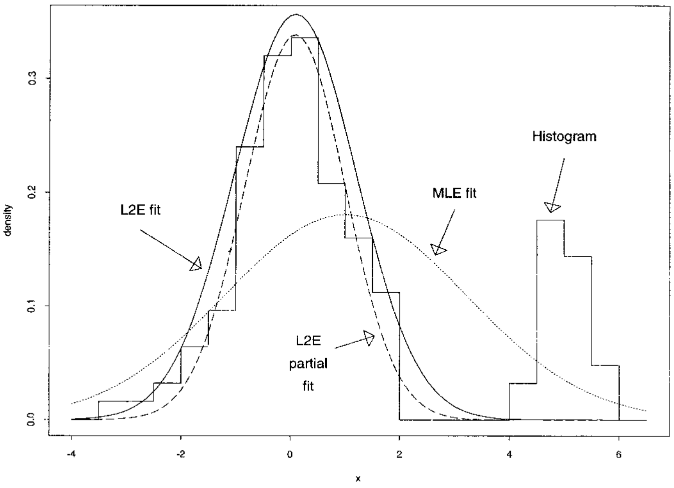

Figure 3. The Parametric L_2E Result With $f(x|\boldsymbol{\theta}) = N(\mu,\sigma^2)$ for the $n = 125$ Points From the Mixture $.8N(0,1) + .2N(5,1)$ Shown in Figure 1. The MLE is displayed as well. The last curve is a partial L_2E fit with the three-parameter model $w \cdot N(\mu,\sigma^2)$.

For example, when $K = 2$,

$$
\begin{aligned}
L_2E(w_1,\mu_1,\mu_2,\sigma_1,\sigma_2)
&= \frac{w_1^2}{2\sqrt{\pi}\sigma_1} + \frac{(1-w_1)^2}{2\sqrt{\pi}\sigma_2} + 2w_1(1-w_1)\phi\big(\mu_1-\mu_2|0,\sigma_1^2+\sigma_2^2\big) \\
&\quad - \frac{2}{n}\sum_{i=1}^{n}[w_1\phi(x_i|\mu_1,\sigma_1^2) + (1-w_1)\phi(x_i|\mu_2,\sigma_2^2)]. \qquad (10)
\end{aligned}
$$

Using logarithmic transformations for the variances and a logistic-like transformation for the weight, w_1, leads to an unconstrained optimization problem that can be solved with standard quasi-Newton method algorithms such as *nlmin* in S-PLUS.

I begin by reexamining an important practical property of L_2E mixture fits; namely, if the number of components fitted is less than required and if the components are sufficiently well separated, then the L_2E solution tends to find the largest components. We illustrate this by refitting the data used in Figure 1 with a single two-parameter Normal fit using Equation (7). (Recall that $\sigma = 1$ was assumed to be known before.) With the two-parameter model, there is a unique L_2E minimizer, which is shown in Figure 3 along with the data histogram and the MLE fit. The L_2E parameters are (.09, 1.12) compared to the sample moments of (1.00, 2.20). [For the 100 "good" data points, the respective estimates are (.08, .94) and (.01, 1.01).] This L_2E behavior may be valuable in practice. In particular, one can hope to exclude not only one or two bad data points but entire groups of outliers in many circumstances. I shall also find this property useful when I turn to regression problems.

Clearly, the variance of the fitted normal in Figure 3 is inflated. I digress for a moment to study that property. Using *nlmin* in S-PLUS, I numerically find the best value of σ so that the ISE between the model $\mathrm{N}(0,\sigma^2)$ and the isolated component $w \cdot \mathrm{N}(0,1)$, $0 < w < 1$, is minimized. Some of the density pairs are shown in Figure 4. When $w = .8$, this model predicts that the standard deviation will be inflated by 17.6%. The actual inflation in my example is 19.5%. I find that when $w \le (2\sqrt{2})^{-1} = .3536$, the "best" value of σ is infinite so that $f(x) = 0$ is the best L_2 fit. This result suggests that isolated components with less than 40 or 50% of the data will not be fitted separately but rather will be ignored or combined with other data. One promising line of research that will be pursued elsewhere is the additional simultaneous estimation of a weight $w \in [0,1]$ in the expanded three-parameter model, $w \cdot \mathrm{N}(\mu,\sigma^2)$. For my example, the estimates are .780, .080, and $.919^2$, respectively; see the long-dashed line in Figure 3.

Next, I consider a more challenging mixture problem of a sample of the net incomes of 7,428 British households in 1982 (Family Expenditure Survey 1968–1983; Härdle 1990). About 20 apparent outliers lie outside the plotting range in Figure 5, in which the $K =$ 1-, 2-, and 3-component L_2E Gaussian mixture fits are displayed. The weights, means, and standard deviations of the two-component fit from left to right are (36.1%, 63.9%), (1.81, 2.23), and (.172, .170), respectively. An MLE solution (not shown) found by the EM algorithm is close by (McLachlan 1992). A second EM solution is located at (12.9%, 87.1%), (1.69, 2.14), and (.105, .236).

Fitting the three-component model was more interesting. The weights, means, and standard deviations of the three L_2E components from left to right are (18.1%, 12.4%, 69.5%), (1.70, 1.90, 2.21), and (.103, .083, .177), respectively. Using these components as an initial guess, EM converged to essentially the first two-component EM solution with a small third diffuse component that covers the outliers—$.0181\,\phi(x|2.12, .565^2)$. When the 20 outliers were removed, the three-component EM solution was essentially the same as the L_2E solution on all the data shown in Figure 5. Finally, a four-component EM solution handled the outliers with the fourth component. This example illustrates the somewhat unpredictable influence outliers can have on MLE fits, as well as the robustness of the L_2E criterion.

Kim (1995) studied the normal mixture problem extensively. Together with Terrell (1990), he devised a penalized L_2E mixture algorithm whose solution is a quadratic program. Specialized algorithms for solving quadratic programs can be significantly faster than Newton's method.

5. LINEAR REGRESSION

Regression and prediction problems are among the most important in statistics. At first glance, there does not appear to be an obvious role for L_2E in regression problems. But if I focus on the distributional assumption of the residuals, then I see that the regression coefficients may be obtained indirectly by using the L_2E criterion to model the distribution of the estimated residuals.

Consider the simple linear regression model $Y = a + bx + \epsilon$, where $\epsilon \sim \mathrm{N}(0,\sigma_\epsilon^2)$. The L_2E criterion directly employs the parametric model of the residual density, $f(\epsilon) = \phi(\epsilon|0,\sigma_\epsilon^2)$. Invoking Equation (6), I have

$$
(\hat{a},\hat{b},\hat{\sigma}_\epsilon) = \arg\min_{a,b,\sigma_\epsilon}\left[\frac{1}{2\sqrt{\pi}\sigma_\epsilon} - \frac{2}{n}\sum_{i=1}^{n}\phi(\epsilon_i|0,\sigma_\epsilon^2)\right],
$$

where $\epsilon_i = y_i - a - bx_i$. Note that all three parameters are estimated simultaneously. L_2E tries to find the model with the most Normal set (or subset) of residuals available.

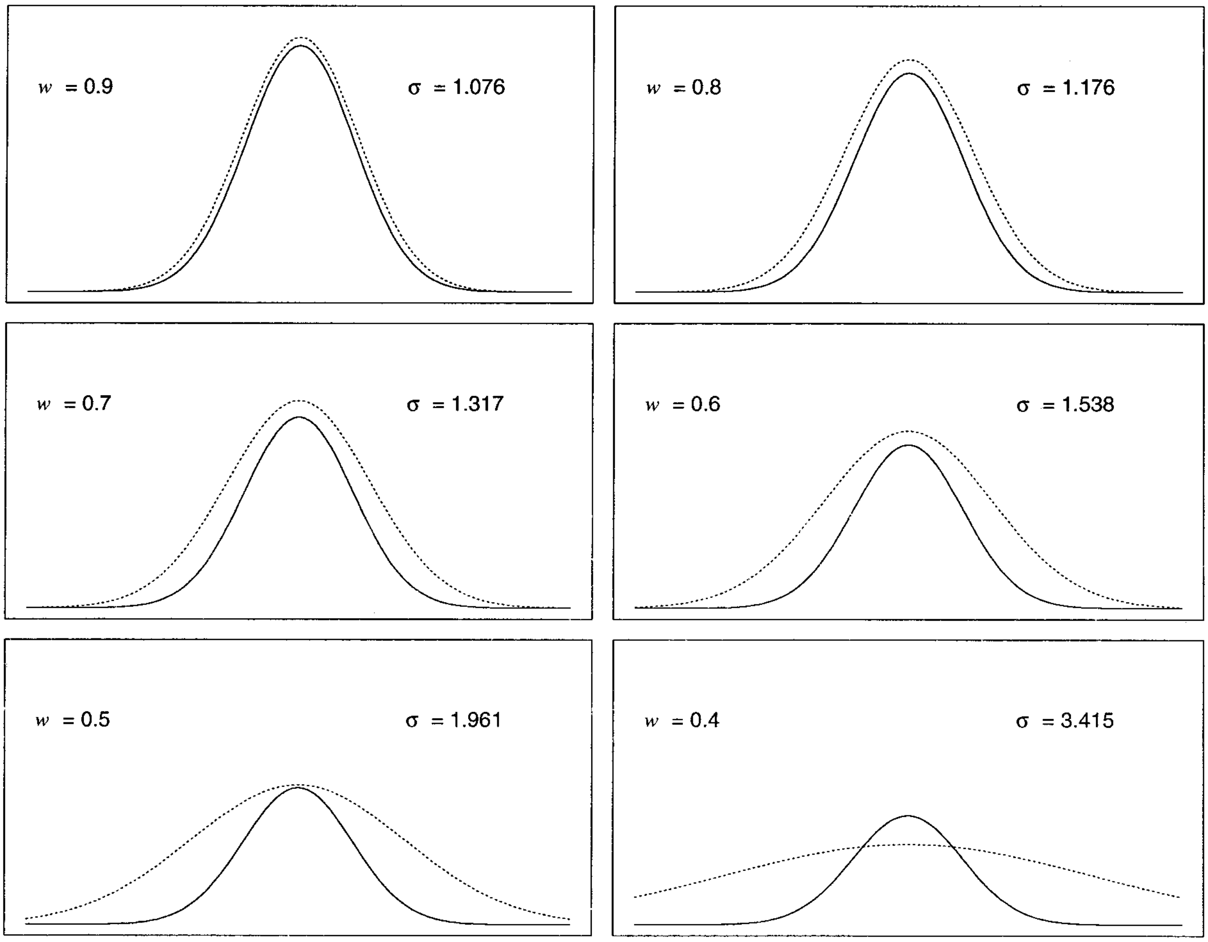

Figure 4. Plots of the Component $w\ N(0,1)$ (solid line) and the Best Fitting Model $N(0,\sigma^2)$ (dashed line) for Several Values of w. For $w \leq .3536$, $\sigma = \infty$; see text.

On the other hand, the method of least squares does not require any parametric assumption. Any prior assumption of Normality must be verified after fitting, using a variety of tests and graphical diagnostic plots of the residuals.

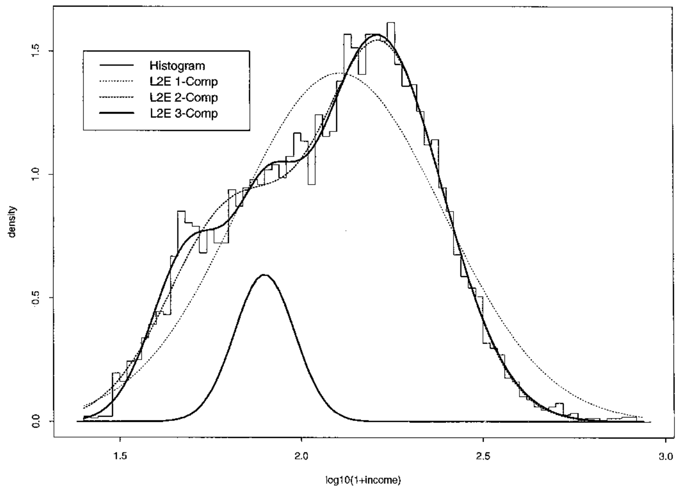

Figure 5. The One-, Two-, and Three-Component L_2E Gaussian Estimates of the British Household Income Data. The middle component of the three-component fit is also shown. The other components are apparent from the left and right edges. (Note the expanded horizontal scale.)

I compare these ideas on a simulated dataset of 250 points, 200 from the model $y = x + \epsilon$ and 50 from $y = \epsilon$, where $\epsilon \sim \mathrm{N}(0, 1)$. For clarity, the x design was chosen so that the points are in three clusters; see Figure 6. To facilitate comparison and display between the L_2E and MLE criteria, I have used the true value of $\sigma_\epsilon = 1$ in L_2E. With this knowledge, L_2E finds two plausible regression lines that explain different subsets of these data. The second row of Figure 6, shows the location and shape of the residual histograms around the point $\epsilon = 0$. In particular, the L_2E plots show clearly the outlying cluster of points, while the MLE residual histogram is much less clear. In other words, the diagnostic step is easier for L_2E fits. Note that the mean of the L_2E residuals is not necessarily equal to 0. Without prior knowledge of σ_ϵ, there is only one L_2E solution—$\hat{a} = -.04$, $\hat{b} = .98$, and $\hat{\sigma}_\epsilon = 1.35$—which should be compared to the values given in the residual plots in Figure 6. Only the estimate of σ_ϵ is inflated.

Next, I revisit an example discussed by Rousseeuw and Leroy (1987). In Figure 7, the least squares (LS) regression line is heavily influenced by four giant stars in the Hertzsprung–Russell diagram of the star cluster CYG OB1. Rousseeuw and Leroy derive the least median squares (LMS) of residuals estimator, $\hat{y} = -12.3 + 3.90x$. The equation of my L_2E fit for these data is quite similar, $\hat{y} = -8.77 + 3.11x$, with $\hat{\sigma}_\epsilon = .414$. Kernel estimates of the residuals for the LS

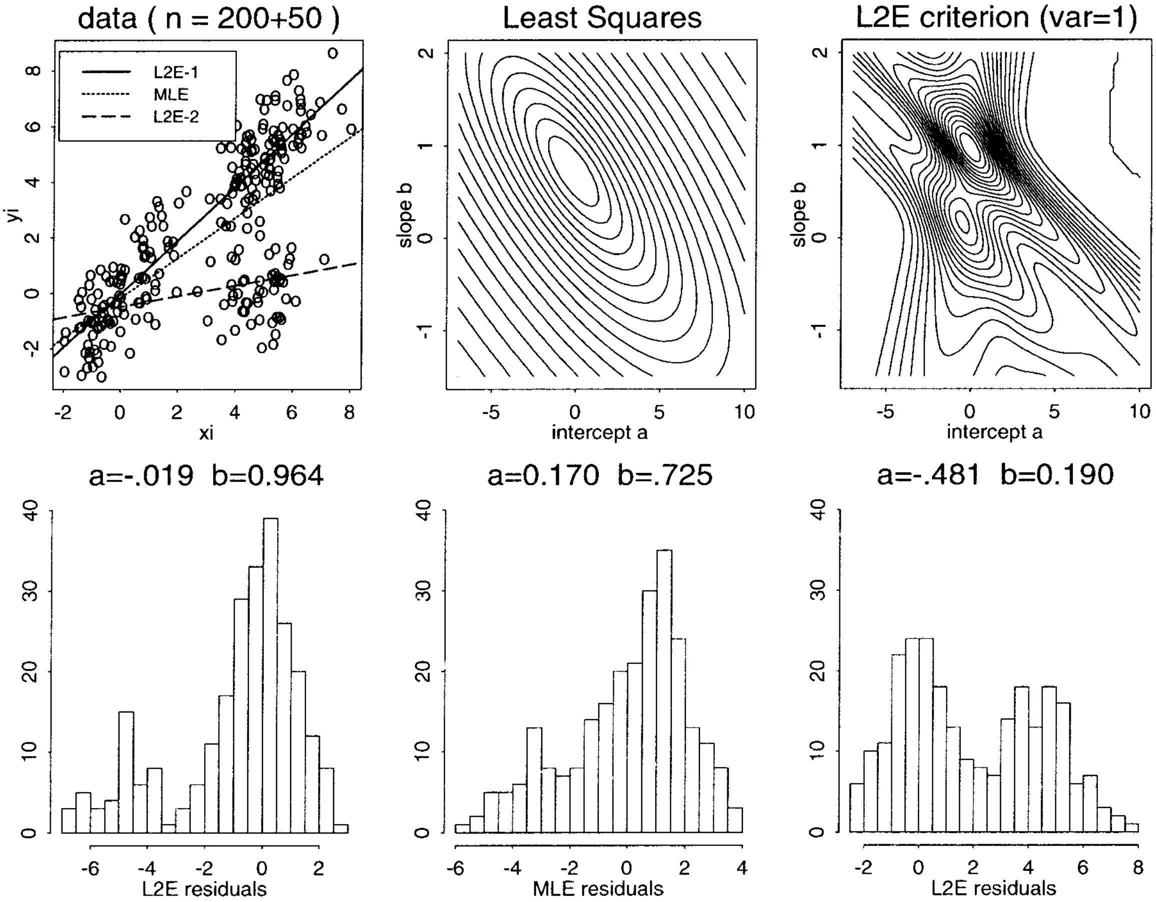

Figure 6. MLE and L_2E Regression Criteria With Data and Estimates Shown in the Upper Left Frame. Residual plots for the single MLE curve and the two L_2E curves are shown in the second row. Note the location of the origin ($\epsilon = 0$) in each.

and L_2E fits are plotted in Figure 8, together with the fitted normal model of the residuals. Again, the LMS and L_2E residuals are quite similar.

My final example revisits the Brownlee (1960) stack-loss data. This small 21-sample 4-variable set is interesting for the variety of findings. Dodge (1996) cataloged 26 distinct published sets of detected outliers. In Figure 9, I display kernel estimates of the residuals computed by LS, by *rreg* (a function in S-PLUS using iteratively reweighted least squares), and by L_2E. In this case, the goal of finding a model in which a large fraction of the residuals are normal is achieved, highlighting five cases as outliers—1, 3, 4, 13, and 21. Case 13 is a borderline outlier. The relative diagnostic value of the L_2E residual plot seems clear.

6. SIMULATION STUDIES

In this section, I report on simulations comparing MLE, L_2E, and minimum Hellinger distance (MHD) parameter estimates in some of the previous settings. In each case, the MHD estimator was fitted to a Gaussian kernel density estimator using the exact best MISE bandwidth (Marron and Wand 1992).

I first consider fitting a two-parameter normal model to $n = 100$ N(0, 1) data, repeated 1,000 times. Histograms of the L_2E and MHD estimates of μ and σ are shown in Figure 10. The optimal bandwidth for the MHD target kernel density was $h^* = .445$. Note that $\sigma(\bar{X}) = \sigma/\sqrt{n} = .01$, a value matched by MHD, while L_2E is 25.0% greater than this (in close agreement to the theoretical figure of 24.1% given at the end of Sec. 3). For the standard deviation, I have that $\sigma(S) = \sigma/\sqrt{2n} = .071$ for the MLE. From my simulation, the L_2E value is 38.0% greater than that of the MLE (compared to the 36.0% figure predicted by using Proposition 3.1). Observe that $\hat{\sigma}$ for MHD is biased upward by 7.9%. However, the standard deviation of a kernel density estimate is larger than the sample standard deviation, s_x, and is given by $(s_x^2 + h^2\sigma_K^2)^{1/2}$ (Scott 1992, p. 193); therefore, the predicted increase is 9.5% in this case since the kernel variance $\sigma_K^2 = 1$. This figure nearly matches the observed 7.9% increase. On the other hand, the standard deviation of $\hat{\sigma}_{\text{MHD}}$ is quite small. I computed the root mean squared error of $\hat{\sigma}_{L_2E}$ and $\hat{\sigma}_{\text{MHD}}$ as .0984 and .1009, respectively, so that the accuracy of each is essentially equivalent. Again, the inefficiency of these estimators must be balanced against other properties in practice.

In the case of MHD estimation, note that the use of bandwidths other than h^* has little effect on the estimate of location; however, the estimated standard deviation increases with the increasing bandwidth. The MHD estimation algorithm was implemented in S-PLUS using the built-in functions *nlmin* and *integrate*, which perform quasi-Newton optimization and numerical integration, respectively. This approach gives approximately six significant digits. [Such accuracy increases estimation time and may be more than the

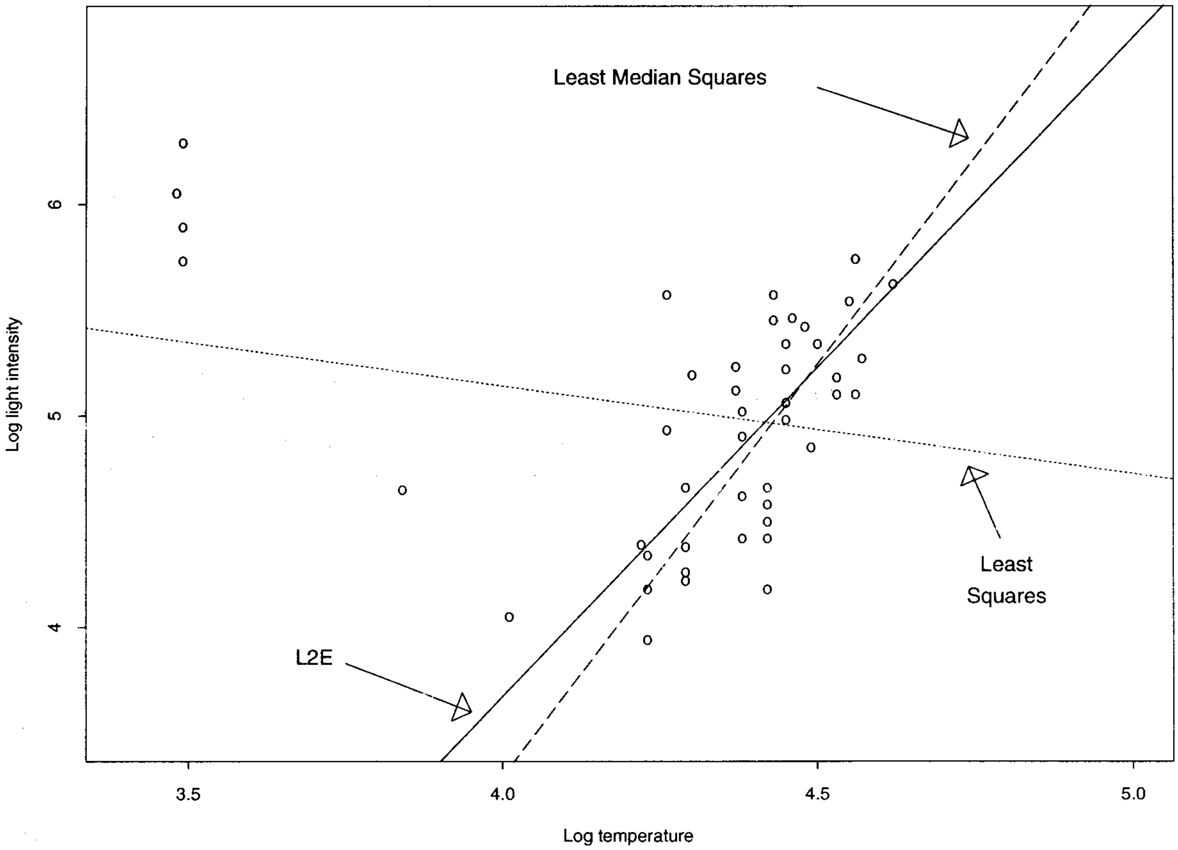

Figure 7. The Least Squares, L_2E, and Least Medium Squares Estimators for the Hertzsprung–Russell Diagram of the Star Cluster CYG OB1.

required time in practice. Woodward, Whitney, and Eslinger (1995), for example, used a golden section search to find MHD parameters, computing the criterion using Simpson's rule with a mesh of 201 points. They also chose to use the biweight rather than Gaussian kernel.] Estimation in our simulations was started using the values $(\bar{x}, s)$, although several other starting values were examined without observing any other solutions in this simple setting. Each individual sample of the simulation required an average of .016, .175, and 20.74 seconds for MLE, L_2E, and MHD, respectively, on a Sun Ultra 1 computer. These times include the overhead of generating the data and storing and plotting the results. The number of iterations required for *nlmin* to converge for the L_2E and MHD estimates was usually no more than 10. Obviously, these numbers would change if other starting values were chosen.

Our second set of simulations extends our study of the set of contaminated data presented in Figure 3. To each sample of 100 N(0, 1) points, 25 N(c, 1) points were added, for $c =$ .0, .5, 1.0, ..., 10.0. I expect c to be sufficiently large that any minimum distance estimator will eventually ignore the 25 "bad" data points. In Figures 11 and 12, boxplots of $\hat{\mu}_{L_2E}$,

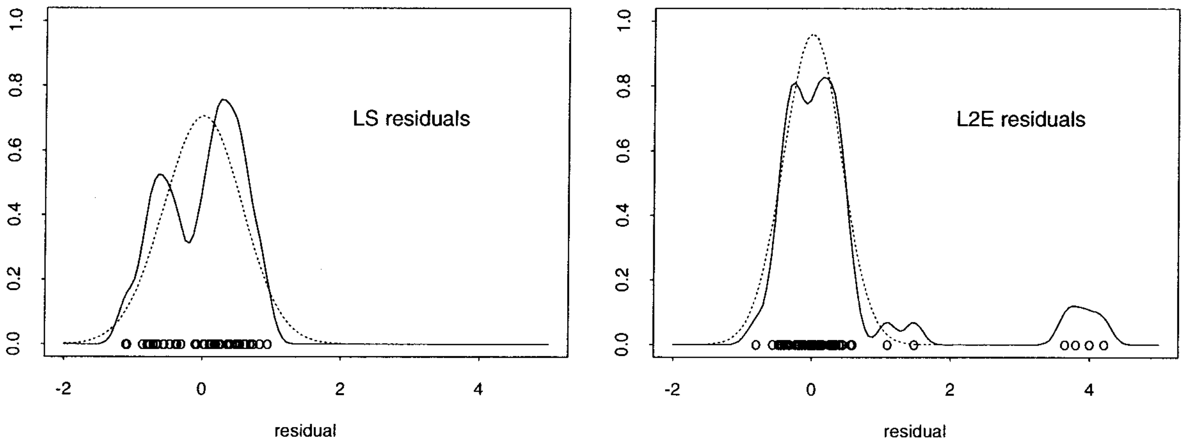

Figure 8. Gaussian Kernel Density Estimates of the Estimated Residuals From the Least Squares and L_2E Fits, Together With the Estimated Normal Models for the Residuals (dotted line).

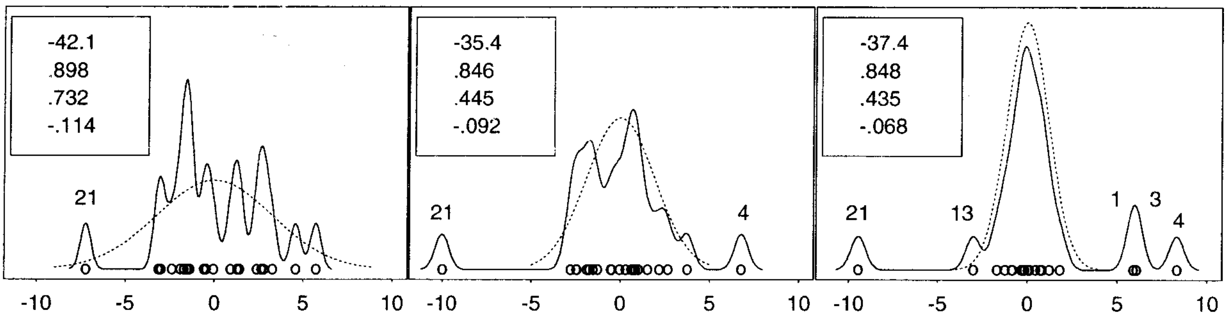

Figure 9. Kernel Density Estimates of the Stack-Loss Residuals From the Least Squares, rreg Robust Regression, and L_2E Fits, Together With a Normal Model. The intercept and weights on the variables (1) air flow, (2) water temperature, and (3) acid concentration are displayed, as well as the case numbers of outliers.

$\hat{\mu}_{\mathrm{MHD}}$, $\hat{\sigma}_{L_2E}$, and $\hat{\sigma}_{\mathrm{MHD}}$ for 256 simulations for each value of c are displayed. Note that all the parameters initially track the increasing contamination location until the separation is apparent to the algorithm. In all of the MHD optimizations, I used $h^* = .445$, which is appropriate for 100 N(0, 1) samples. There were no observed local minima in the L_2E samples, but local minima (large $\hat{\sigma}_{\mathrm{MHD}}$) were observed in an increasing number of the MHD samples with c greater than 5.0. I have recorded the better solution. Finally, note that, although $\hat{\mu}_{L_2E}$ returns to the value of 0, $\hat{\sigma}_{L_2E}$ does not return all the way to the value of 1.0 but to the value 1.166; see Figure 4, in which the predicted value is 1.176 for an 80% component. These graphs confirm my hypothesis, although the MHD estimator is slower to ignore the contaminated data than initially expected.

My final simulation uses the five-parameter two-component normal mixture model $w_1\,\phi(x|\mu_1,\sigma_1^2)+(1-w_1)\,\phi(x|\mu_2,\sigma_2^2)$. The same S-PLUS functions mentioned previously were used to estimate these parameters by MLE, L_2E, and MHD for 100 samples of size $n = 400$ from our favorite "hard" mixture

$$\frac{3}{4}\mathrm{N}(0,1)+\frac{1}{4}\mathrm{N}(3,3^{-2}).$$

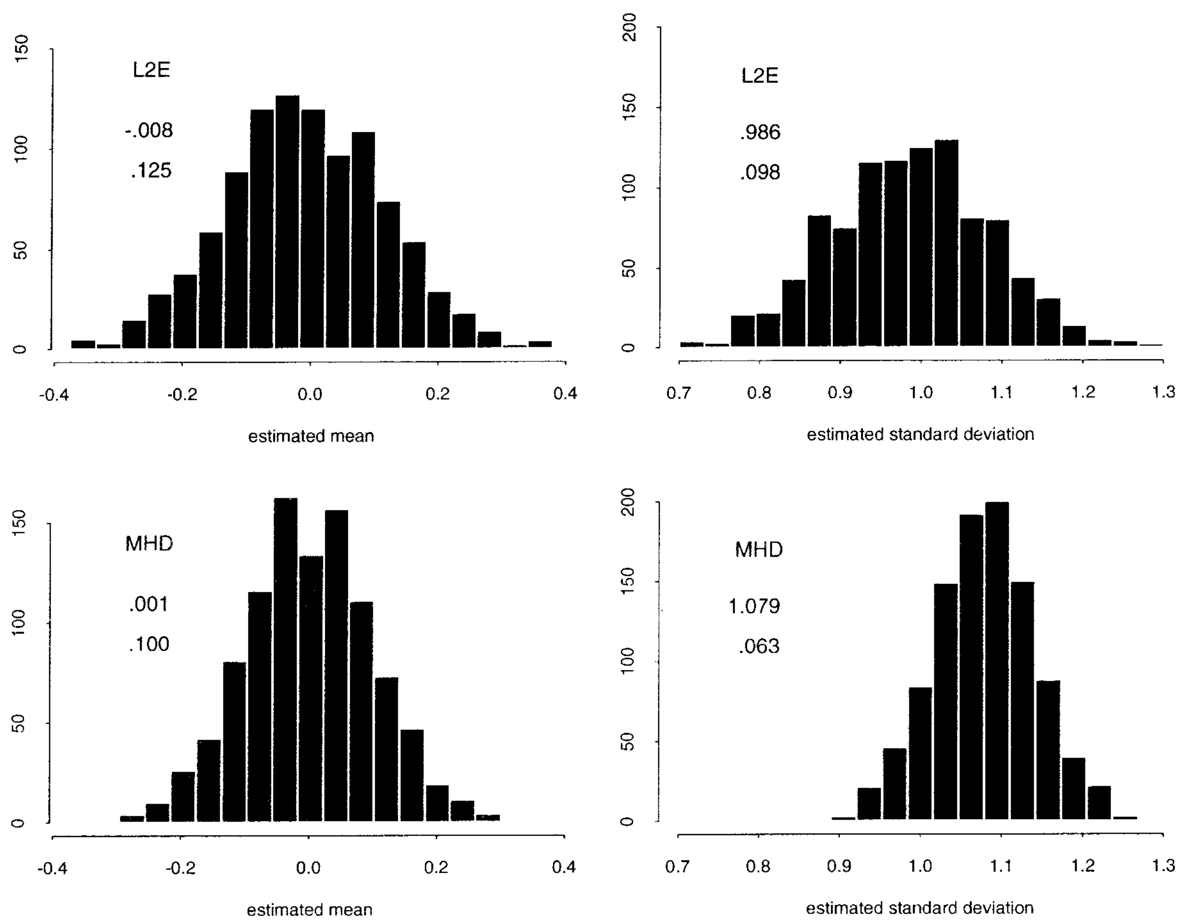

Figure 10. Frequency Curves of Estimated Parameters ($\hat{\mu}$,$\hat{\sigma}$) From 1,000 N(0,1) Samples Using L_2E and MHD Algorithms. The sample mean and standard deviation for each statistic is given on the figure.

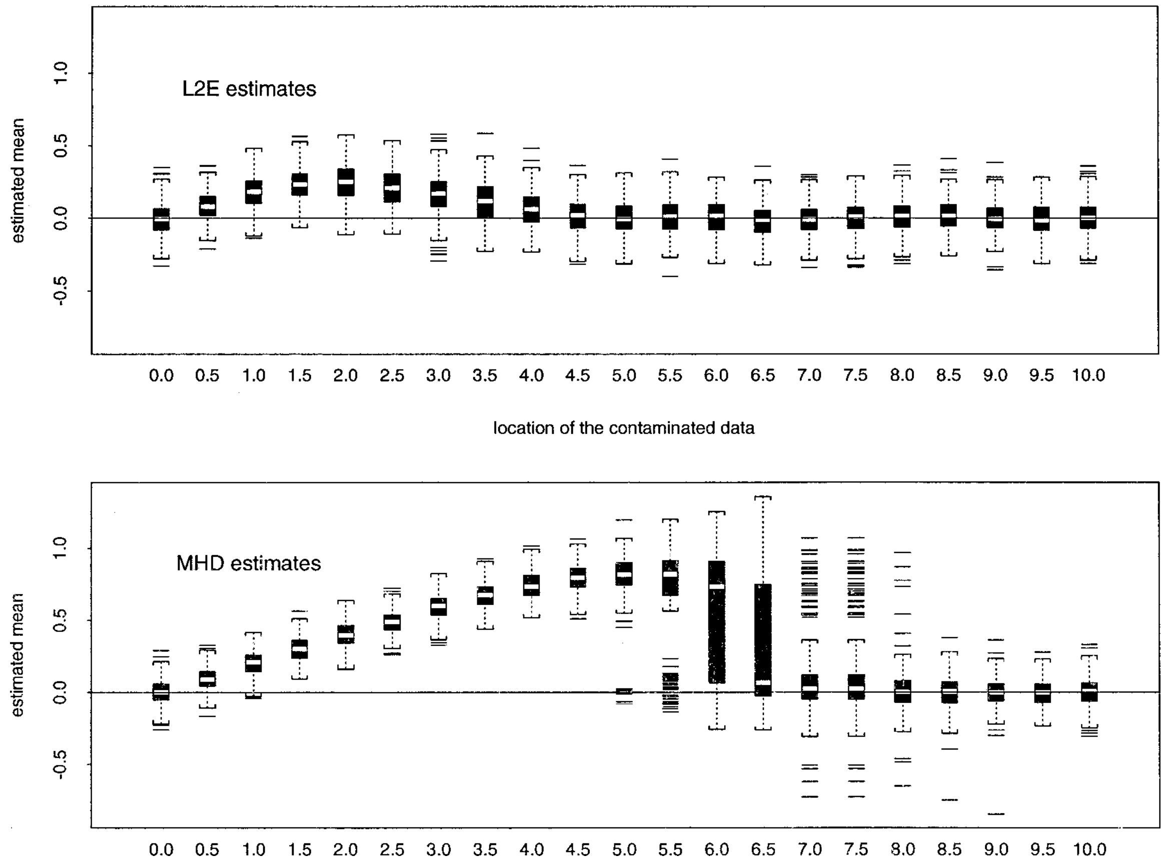

Figure 11. In This Figure, 256 Samples of Size 125 Were Generated for 13 Different Contamination Locations, Which Are Indicated Along the Horizontal Axis. One hundred standard normal samples were combined with 25 N(c,1) points, for c = .0, .5, 1.0, ... , 10.0. Boxplots of the estimated location parameter are shown for L_2E and MHD algorithms that attempted to fit the partially correct model, $N(\mu,\sigma^2)$. (See Fig. 12 for $\hat{\sigma}$.)

Boxplots for all five parameters and three methods are shown in Figure 13. The starting values were the true parameter values. The optimal bandwidth, $h^* = .206$, was used in the MHD algorithm. The results are similar. The L_2E parameters are somewhat more variable, while the estimated MHD standard deviations are inflated. As is generally the case, more complex models are more susceptible to poor structural fits if the sample size is too small. Nevertheless, all the algorithms converged in this study. In a few MHD cases, the numerical integration algorithm was sufficiently noisy that the S-PLUS optimization code returned a warning that the convergence was "false." I checked these solutions and believe the numerical derivatives could not be computed with sufficient accuracy to confirm convergence, although the numerical solution had in fact been reached.

In more extensive simulations of location and scale problems, Wojciechowski (2001) compared 15 robust estimators, including L_2E, M estimators, and minimum distance estimators. L_2E often came out on top, particularly with heavy-tailed, asymmetric contamination.

7. DISCUSSION

In this article, I have demonstrated how an often-used nonparametric estimation criterion can be applied to a variety of parametric problems. In particular, my implementation is fully constructive in the same sense as maximum likelihood. Numerical optimization is required in almost every case, but very complicated models can be implemented quickly. The parameters are relatively inefficient compared to maximum likelihood theory at the correct model. But as some nonparametric workers like to argue, the more interesting and challenging situations are away from the model. In my applications, I believe that a known parametric model represents a significant fraction of the data. The ability to successfully handle a nontrivial fraction of bad data should be of extreme value with high-dimensional problems and, more importantly, with massive datasets (for which careful data preparation is not feasible). In other words, I build models as good approximations and not as representing absolute truth. With large samples, standard testing almost always rejects my models. L_2E is better suited to treating models as approximations, handling outliers and underspecified models in a useful manner. A comparison of L_2E and MLE fits almost always provides an informative diagnostic. However, comparing the residuals in Figures 8 and 9 emphasizes the added benefit of criticizing the model using L_2E versus MLE residuals. Since MLE must account for all the data, the fits often blur the distinction between good and bad data. The difficulty grows with dimension.

Tukey's study of robust estimators (Andrews et al. 1972) was an important step to modern data analysis. Tukey's models of contamination focused on symmetric, heavy-tailed data. By making explicit parametric assumptions, the L_2E approach can

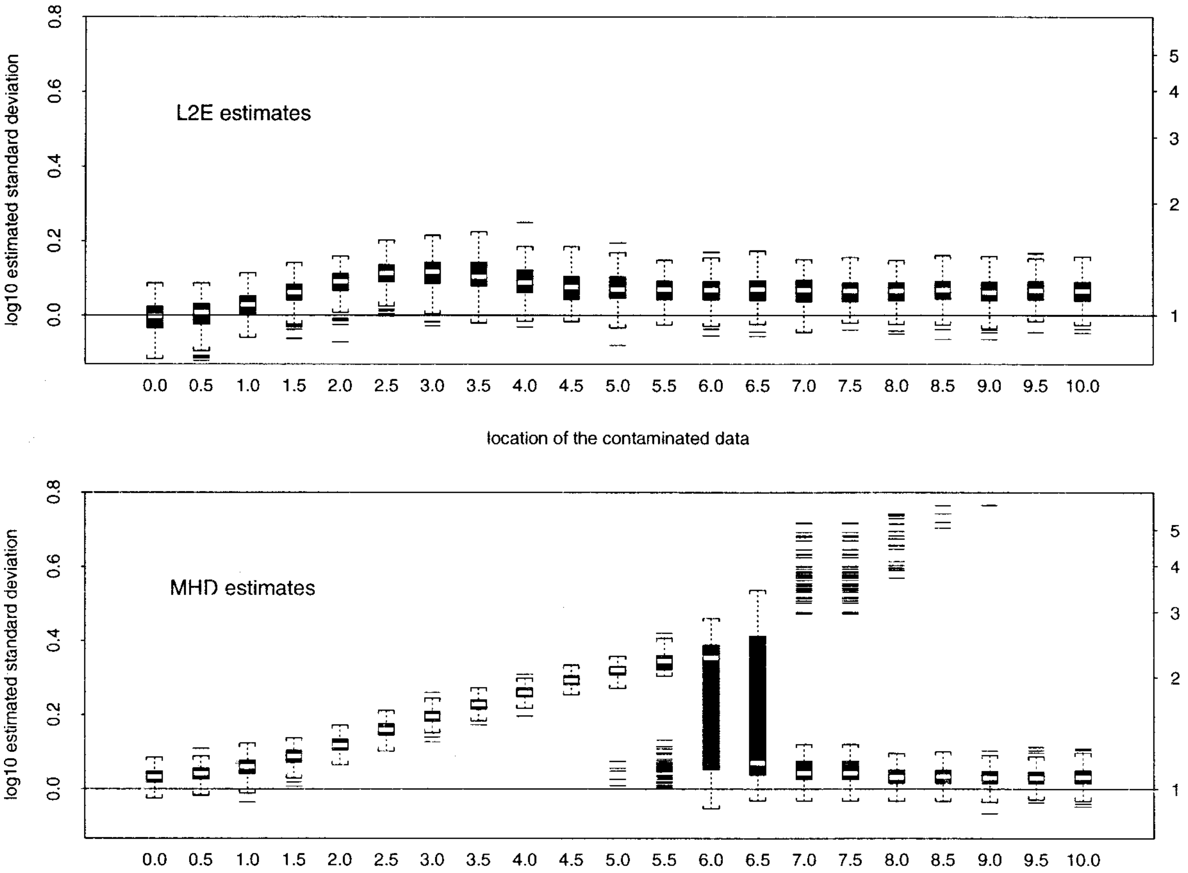

Figure 12. Continuing the Summary of Data Depicted in Figure 11, Boxplots of the Estimated Standard Deviation Parameter Are Shown for L_2E and MHD Algorithms That Attempted to Fit the Partially Correct Model, $N(\mu,\sigma^2)$.

handle asymmetric error distributions as well. The shape of the influence function is implicitly determined and in some sense is best suited for the task, given the choice of ISE.

This line of research can trace its origins to our longstanding interest in bandwidth selection and to a series of lectures by George Terrell in the RIMS summer program at Rice University from 1996–1998. I was originally curious to see how these nonparametric ideas would work with parametric models. The inefficiency of L_2E relative to other choices of α was noted by Basu et al. (1998). There is no general procedure for choosing α; however, they suggested that α's less than .25 are sufficiently robust. On the other hand, I have taken the point of view that the wealth of practical experience and success in the nonparametric world lends credence to the idea that L_2E is a special class of robust parametric estimators that, like median-based estimators, sacrifice some asymptotic efficiency for clear benefits in difficult problems faced by practicing statisticians. Furthermore, in practice for multivariate mixture problems and partial mixture modeling, the availability of a closed-form minimum-distance criterion is critical, compared to numerical integration required by minimum divergence. Basu et al. also gave further details

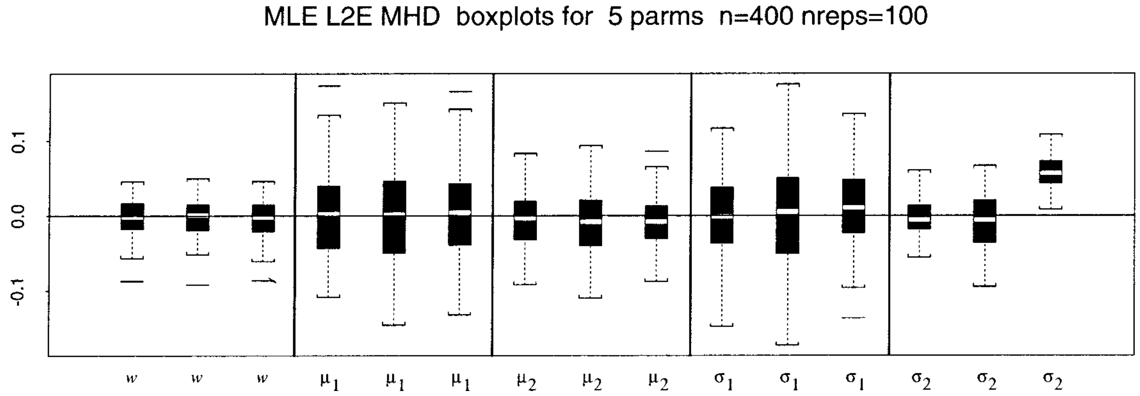

Figure 13. Boxplots Of Estimated MLE, L_2E, and MHD (left, middle, and right) Parameters From 100 Simulations of a Five-Parameter Normal Mixture. The boxplots have been centered by subtracting the true parameter values, $w = .75$, $\mu_1 = 0$, $\mu_2 = 3$, $\sigma_1 = 1$, and $\sigma_2 = 1/3$.

about hypothesis-testing possibilities and breakdown points. Clearly, the parametric bootstrap has application for obtaining confidence intervals.

My simulation studies also reinforce the benefit of a closed-form expression for a criterion in any numerical optimization. The numerical integration required for computing Hellinger distance is a practical limitation—not only in computation time but also for obtaining sufficient accuracy to perform quasi-Newton optimization [especially for multivariate models, $f(x|\theta)$]. L_2E enjoys one such advantage over the other cases of divergence measure considered by Basu et al.—namely, fitting normal mixtures. The integral $\int f^{1+\alpha}$ in Equation (9) does not have a closed form for $0 < \alpha < 2$ except at L_2E; see Equation (10). For fitting multivariate normal mixture models, existence of a closed-form criterion is of great practical importance.

I have begun to investigate a number of other applications. One involving estimation of an economic stochastic frontier function will be reported elsewhere (Scott, Simar, and Wilson 2001). Multivariate regression and time series, especially with massive datasets, are particularly interesting. Other statistical algorithms that could benefit from robustness are excellent candidates for minimum distance procedures. I hope to explore those applications soon.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation grant DMS 96-26187 and 99-71797 and the Department of Defense contract MDA 904-95-C-2203. I thank E. Parzen, Changkon Hong, Savas Papadopoulos, Leopold Simar, Paul Wilson, and the editor for helpful comments.

[Received December 2000. Revised March 2001.]

REFERENCES

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

Barnett, V., and Lewis, T. (1994), *Outliers in Statistical Data*, Chichester, U.K.: Wiley.

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

Beaton, A. E., and Tukey, J. W. (1974), "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data," *Technometrics*, 16, 147–185.

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

Bowman, A. W. (1984), "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates," *Biometrika*, 71, 353–360.

Brown, L. D., and Hwang, J. T. G. (1993), "How to Approximate a Histogram by a Normal Density," *The American Statistician*, 47, 251–255.

Brownlee, K. A. (1960), *Statistical Theory and Methodology in Science and Engineering*, New York: Wiley.

Dodge, Y. (1996), "The Guinea Pig of Multiple Regression," in *Robust Statistics, Data Analysis, and Computer Intensive Methods*, ed. H. Rieder, New York: Springer-Verlag, pp. 91–117

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

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman & Hall.

Hampel, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383–393.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge, U.K.: Cambridge University Press.

Hjort, N. L. (1994), "Minimum $L2$ and Robust Kullback–Leibler Estimation," in *Proceedings of the 12th Prague Conference*, eds. P. Lachout and J. Visek, Prague: Academy of Sciences of the Czech Republic, pp. 102–105.

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

Jurečková, J., and Sen, P. K. (1996), *Robust Statistical Procedures*, New York: Wiley.

Kendall, M., and Stuart, A. (1977), *The Advanced Theory of Statistics* (Vol. 1), New York: Macmillan.

Kim, D. (1995), "Least Squares Mixture Decomposition Estimation," unpublished doctoral dissertation, Virginia Polytechnic and State University, Dept. of Statistics.

Lehmann, E. L. (1983), *Theory of Point Estimation*, New York: Wiley.

Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20, 712–736.

McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley.

Ripley, B. D. (1996), *Pattern Recognition and Neural Network*, Cambridge, U.K.: Cambridge University Press.

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

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

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

——— (1998), "Parametric Modeling by Minimum L_2 Error," Technical Report 98-3, Rice University, Dept. of Statistics.

Scott, D. W., Simar, L., and Wilson, P. W. (2001), "A New Model and Criterion for Estimating the Stochastic Frontier Error," unpublished manuscript.

Terrell, G. R. (1990), "Linear Density Estimates," in *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 297–302.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

Wand, M. P., and Jones, M. C. (1995), *Kernel Smoothing*, London: Chapman and Hall.

Wojciechowski, W. (2001), "Robust Modeling," unpublished doctoral dissertation, Rice University, Dept. of Statistics.

Woodward, W. A., Whitney, P., and Eslinger, P. W. (1995), "Minimum Hellinger Distance Estimation of Mixture Proportions," *Journal of Statistical Planning and Inference*, 48, 303–319.