Parametric Modeling by Minimum L_2 Error[1]

David W. Scott[2]

February 10, 1999

ABSTRACT: The likelihood function has played a central role in parametric estimation, Bayesian procedures, and even nonparametric function estimation via local polynomial modeling. However, the integrated squared error criterion has enjoyed a long tradition as the goodness-of-fit criterion of choice in nonparametric density estimation. In this paper, we investigate the use of integrated squared error, or L_2 distance, as a theoretical and practical estimation tool for a variety of *parametric* statistical models. We show that the asymptotic inefficiency of the parameters estimated by minimizing the integrated squared error or L_2 estimation (L_2E) criterion versus the MLE is roughly that of the median versus the mean. We demonstrate by example the well-known result that minimum distance estimators, including L_2E, are inherently robust, without requiring specification of any tuning factors found in robust likelihood algorithms. L_2E is particularly appropriate for analyzing massive data sets where 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 paper include univariate density estimation, mixture density estimation, regression estimation, and robust estimation of the mean and covariance.

KEY WORDS: Robust estimation; Minimum distance estimators; Outlier detection; Normal mixture models; Cross-validation; Integrated squared error.

---

[1]Technical Report 98-3, Department of Statistics, Rice University

[2]David W. Scott is Professor, Department of Statistics (MS–138), Rice University, 6100 Main Street, Houston, TX 77005-1892. This research was supported in part by the National Science Foundation grant DMS 96-26187 and the Department of Defense contract MDA 904-95-C-2203. The author would like to thank Professors E. Parzen, Changkon Hong, Savas Papadopoulos, Leopold Simar, and Paul Wilson for helpful comments.

1. INTRODUCTION

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, $\int \left[\hat{f}(x)^{1/2} - f(x)^{1/2}\right]^2 dx$, has special attraction since it is dimensionless. Another dimensionless distance measure is L_1 error, $\int |\hat{f}(x) - f(x)|\, dx$. However, these distance measures are not immediately available in practice, and certain approximations and bounds 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, $\hat{f}_h(x)$. Clearly, $\hat{\theta}$ will vary with the choice of smoothing parameter in the kernel estimator so that a rule for determining h must be specified.

Parametric and nonparametric estimators seldom employ the same estimation criterion. Parametric algorithms typically rely on maximum likelihood while nonparametric algorithms favor L_p error measures, with integrated squared error (*ISE*, $p = 2$) the favorite. However, using *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 criteria, *ISE*, to parametric problems.

Consider parametric estimation of the simple 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 consideration of $E(\hat{\theta})$ gives $c = (n+1)/n$ (unbiasedness), while mean squared error gives $c = (n+2)/(n+1)$. 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

$$MISE(\hat{\theta}) = E_{\hat{\theta}} \int \left[f(x|\hat{\theta}) - f(x|\theta_0)\right]^2 dx\,;$$

note that $2^{1/(n-1)} \approx 1 + (n-1)^{-1} \log 2$. All three bias the *MLE* upward ($c > 1$). This simple example highlights the most important advantage of the *MLE*, namely its constructive nature. The other criteria were only applied in a one-dimensional neighborhood of the

MLE. Furthermore, in (other) regular cases, maximum likelihood estimates generally enjoy asymptotic optimality properties.

In this paper, a fully constructive parametric estimation algorithm is devised based upon the integrated squared error criterion. Its robustness behavior is demonstrated by an example and it is shown how the basic density estimation framework may be extended to estimation in general statistical models.

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 (Rudemo, 1982; Bowman, 1984) for choosing the bin width, h, of a histogram. The role played by the parameter h will 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:

$$\begin{aligned}
\hat{h} &= \arg\min_h \widehat{\int} \left[\hat{f}_h(x) - f(x)\right]^2 dx & (1)\\
&= \arg\min_h \left[\widehat{\int} \hat{f}_h(x)^2 dx - 2\widehat{\int} \hat{f}_h(x) f(x) dx + \widehat{\int} f(x)^2 dx\right] & (2)\\
&= \arg\min_h \left[\int \hat{f}_h(x)^2 dx - 2\widehat{E}\left[\hat{f}_h(X)\right]\right], & (3)
\end{aligned}$$

since the minimizing value of h is not changed by eliminating the (unknown) constant, $\int f(x)^2 dx$, from equation (2). Furthermore, the first integral in (2) can be evaluated exactly for any value of h (and hence does not require further estimation). The middle term remaining in equation (2) requires an estimate of the average height of the histogram with bin width h. Rudemo (1982) proposed an unbiased estimator 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

$$\begin{aligned}
\hat{h} &= \arg\min_h \left[\int \hat{f}_h(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(x_i)\right] & (4)\\
&= \arg\min_h \left[\frac{2}{(n-1)h} - \frac{n+1}{n^2(n-1)h}\sum_k \nu_k^2\right],
\end{aligned}$$

where ν_k is the number of points in the k-th bin, $B_k = (x_0 + kh, x_0 + (k+1)h]$, and x_0 is the bin origin.

Replace h with any parameter θ and rewrite equation (1) for a parametric problem as

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

where the true parameter θ_0 is unknown. Thus, $\hat{\theta} = \theta_0$ is not available as the optimal estimator. As in Rudemo's derivation, 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 we can use the entire random sample to estimate the average height of $f(x|\theta)$. Thus, the proposed estimator minimizing the parametric integrated squared error criterion is

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

Compare equations (6) and (4). Here we have assumed the correct parametric family; however, we will also be interested in the case when the assumed parametric form is incorrect.

Next, we introduce several examples of L_2E functionals. Some related simulations are presented in Section 6. The normal density will be denoted by $\phi(x|\mu, \sigma^2)$.

Example 2.1 (Normal Density) *Suppose $X \sim N(\theta, 1)$. Then*

$$\begin{aligned}
\hat{\theta}_{MLE} &= \arg\max_{\theta} \sum_{i=1}^{n} \log \phi(x_i|\theta, 1) \\
\hat{\theta}_{L_2E} &= \arg\min_{\theta} \left[\frac{1}{2\sqrt{\pi}} - \frac{2}{n}\sum_{i=1}^{n} \phi(x_i|\theta, 1)\right].
\end{aligned}$$

Interestingly, L_2E maximizes the *sum* of the densities while *MLE* maximizes the *product* of the densities. The robustness of $\hat{\theta}_{L_2E}$ can be easily demonstrated for this problem in an interesting setting. Consider a sample of size 100 from $N(0, 1)$ with up to an additional 25 samples from a contamination density, $N(5, 1)$, in the same "true" parametric family. In Figure 1, we plot the log-likelihood and L_2E criteria for $n = 100, 105, \ldots, 125$, always adding 5 new samples from the contamination density while retaining the previous samples. Of course, a closed form solution $\hat{\theta}_{MLE} = \bar{x}$ is available, so that plotting the likelihood is not necessary. The upper right frame in Figure 1 illustrates how resistant the L_2E is to the contaminated data. However, the lower frames, which are plotted over a wider interval, reveal the true story. The likelihood shows little sign of the contaminated data, while the L_2E curves show a local minimum near $\theta = 5$. The existence of the local minimum is appropriate since the contaminated data also come from the assumed parametric model, $N(\theta, 1)$. Study how the amount of contaminated data affects the level of the curves. In particular, extrapolating to 100 contaminated samples, we may judge from the lower right frame that the values at the two minima would be approximately equal. This and other empirical properties of the L_2E may be exploited in practice.

Example 2.2 (Multivariate Normal Density) *Suppose $X \sim N(\boldsymbol{\mu}, \Sigma)$. Then*

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

In Figure 2, we examine a sample of 250 bivariate normal points with $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ $= (-3, -2, 2, 1, 0.75)$. The parameters fitted by L_2E are $(-2.96, -1.97, 1.79, 0.89, 0.70)$, respectively. The parameters fitted by *MLE* are $(-2.93, -2.00, 1.83, 0.97, 0.73)$. The *nlmin* function in Splus was used to minimize L_2E and converged in 27 iterations starting from the true values. This example simply demonstrates the multivariate extension, which will be considered more carefully in a separate paper.

Example 2.3 (Uniform Density) *Suppose $X \sim U(0, \theta)$. Then*

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

For most samples, ironically, 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. In addition, if the ratio of the adjacent order statistics is too big, then the larger order statistic cannot be the L_2E minimizer; see an example in Figure 3. 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." We leave it as an exercise for the reader to find the ratio for other order statistics.

Example 2.4 (Discrete Random Variables) *The loss function analogous to (5) for discrete random variables is $\sum_x \left[f(x|\theta) - f(x|\theta_0)\right]^2$, which leads to the criterion*

$$L_2E(\theta) \quad = \quad \sum_x f(x|\theta)^2 - \frac{2}{n} \sum_{i=1}^{n} f(x_i|\theta)\,. \tag{7}$$

The first sum is over all values of the discrete random variable, and the second sum is over the observed sample. We do not provide an example here. An interesting open problem is handling densities that are a mixture of continuous and discrete components. Perhaps a weighted average of the criteria in (6) and (7) would work.

3. ASYMPTOTICS

In this section, we demonstrate the consistency and asymptotic normality of the L_2E parameters. These results are special cases in a class of divergence measures considered by Basu, et al. (1998); see Section 7 below. The proofs are somewhat simplier in this less general form. In particular, we only consider the case where estimation is in the correct parametric family. The specific assumptions necessary for these results are given by Basu, et al., and closely follow Lehmann's (1983, Theorem 6.4.1) conditions for the *MLE*. 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 (Univariate L_2E Asymptotics) *Under mild assumptions, the L_2E parameter is consistent and asymptotically normal:*

$$\sqrt{n}\left(\hat\theta-\theta_0\right)=AN\left(0,\frac{\int\left[\frac{\partial f(x|\theta_0)}{\partial\theta}\right]^2 f(x|\theta_0)\,dx-\left[\int\frac{\partial f(x|\theta_0)}{\partial\theta}f(x|\theta_0)\,dx\right]^2}{\left[\int\left[\frac{\partial f(x|\theta_0)}{\partial\theta}\right]^2dx\right]^2}\right).$$

Proof: The L_2E parameter, $\hat\theta$, is a root of the equation

$$L_2E'(\theta)=2\int\frac{\partial f(x|\theta)}{\partial\theta}f(x|\theta)\,dx-\frac{2}{n}\sum_{i=1}^{n}\frac{\partial f(x_i|\theta)}{\partial\theta}\,; \tag{8}$$

hence; $L_2E'(\hat\theta)=0$. Taking a Taylor's series about the true parameter θ_0, we have

$$0=L_2E'(\hat\theta)=L_2E'(\theta_0)+(\hat\theta-\theta_0)L_2E''(\theta_0)+\frac{1}{2}(\hat\theta-\theta_0)^2L_2E'''(\tilde\theta)\,, \tag{9}$$

where $\tilde\theta\in(\theta_0,\hat\theta)$ and

$$L_2E''(\theta)=2\int\left[\frac{\partial f(x|\theta)}{\partial\theta}\right]^2dx+2\int\frac{\partial^2 f(x|\theta)}{\partial\theta^2}f(x|\theta)\,dx-\frac{2}{n}\sum_{i=1}^{n}\frac{\partial^2 f(x_i|\theta)}{\partial\theta^2}\,. \tag{10}$$

In both equations (8) and (10), the expected value of the term with the summation sign exactly cancels the integral immediately preceding it, so that

$$E\left[L_2E'(\theta_0)\right] = 0 \qquad \text{and} \tag{11}$$

$$E\left[L_2E''(\theta_0)\right] = 2\int\left[\frac{\partial f(x|\theta_0)}{\partial\theta}\right]^2 dx\,. \tag{12}$$

Consistency follows by standard arguments from equation (9) since both $L_2E'(\theta_0)$ and $L_2E''(\theta_0)$ converge to their mean values given in equations (11) and (12), and assuming that $L_2E'''(\tilde{\theta})$ is bounded. The asymptotic normality of $\sqrt{n}\,L_2E'(\theta_0)$ implies that $\sqrt{n}(\hat{\theta}-\theta_0)$ is asymptotically normal with mean zero and variance

$$Var\left[\sqrt{n}\,L_2E'(\theta_0)\right]\Big/\left\{E\left[L_2E''(\theta_0)\right]\right\}^2 .$$

From (8), the variance of $\sqrt{n}\,L_2E'(\theta_0)$ is easily seen to be

$$Var\left[\sqrt{n}\,L_2E'(\theta_0)\right] = 4\int\left[\frac{\partial f(x|\theta_0)}{\partial\theta}\right]^2 f(x|\theta_0)\,dx - 4\left[\int\frac{\partial f(x|\theta_0)}{\partial\theta}f(x|\theta_0)\,dx\right]^2 . \quad \Box$$

This completes the proof of the proposition. We remark that in equation (9), the exact Taylor's form could be inserted in the factor $L_2E''(\tilde{\theta})$ to avoid introduction of the third derivative; however, we believe inclusion of the extra term results in an easier proof to follow.

Next, the corresponding multivariate case is stated without further proof.

Proposition 3.2 (Multivariate Parameter Vector) *If $\boldsymbol{\theta}$ is a vector of parameters, then*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) = AN\left(\mathbf{0}, H^{-1}\left[G_2 - G_1G_1^T\right]H^{-1}\right),$$

where

$$\begin{aligned} G_1 &= \int_{\Re^p}\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}|\boldsymbol{\theta}_0) f(\boldsymbol{x}|\boldsymbol{\theta}_0)\,d\boldsymbol{x} \\ G_2 &= \int_{\Re^p}\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}|\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}|\boldsymbol{\theta}_0)^T f(\boldsymbol{x}|\boldsymbol{\theta}_0)\,d\boldsymbol{x} \\ \textit{and} \quad H &= \int_{\Re^p}\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}|\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}|\boldsymbol{\theta}_0)^T\,d\boldsymbol{x}\,. \end{aligned}$$

Example 3.1 (Normal Density) *If $X \sim N(\mu,\sigma^2)$, then Proposition 3.2 gives*

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

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 IQR/1.349 is 65.0% greater); see Kendall and Stuart (1977, Volume I, Section 10.12).

4. FITTING GAUSSIAN MIXTURE DENSITIES

A powerful parametric density model is the mixture model (Titterington, et al., 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) has recommended directly optimizing the likelihood with Newton methods. In our 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 formulae found in Wand and Jones (1995). For example, when $K = 2$,

$$\begin{aligned} L_2E(w_1, \mu_1, \mu_2, \sigma_1, \sigma_2) \;\; &= \;\; \frac{w_1^2}{2\sqrt{\pi}\sigma_1} + \frac{(1-w_1)^2}{2\sqrt{\pi}\sigma_2} + 2w_1(1-w_1)\phi\left(\mu_1 - \mu_2|0, \sigma_1^2 + \sigma_2^2\right) \\ & \quad - \frac{2}{n} \sum_{i=1}^{n} \left[w_1 \phi(x_i|\mu_1, \sigma_1^2) + (1-w_1)\phi(x_i|\mu_2, \sigma_2^2) \right] \, . \end{aligned} \tag{13}$$

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 Splus.

We begin by re-examining 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 (recall that we assumed known $\sigma = 1$ before). With the two-parameter model, there is a unique L_2E minimizer (aside from singular Dirac solutions if n is very small or there are many ties), which is shown in Figure 4 along with the data histogram and *MLE* fit. This L_2E behavior will be proved valuable in practice again and again. In particular, we can hope to exclude groups of outliers in many circumstances. We will also find this property useful when we turn to regression problems below.

Clearly, the variance of the fitted normal in Figure 4 is inflated. We digress for a moment to study that property. Again using *nlmin*, we numerically found the best value of σ so that the *ISE* between the model $N(0, \sigma^2)$ and the isolated component $w \cdot N(0, 1)$, $0 < w < 1$, is minimized; see Figure 5. We may show that when $w \leq (2\sqrt{2})^{-1} = 0.3536$, the "best" value of σ is infinite! This theory 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. Some of the density pairs are shown in Figure 6. When $w = 0.8$, this model predicts the standard deviation will be inflated by 17.6%. The actual inflation in our example is 19.5%; see Section 6 below. 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 3-parameter model, $w \cdot N(\mu, \sigma^2)$. For our example, the estimates are 0.780, 0.080, and 0.919^2, respectively; see the big-dashed line in Figure 4.

Next, we consider a more challenging mixture problem, which is displayed in Figure 7. A sample of the incomes of 7,203 British households in 1982 (Härdle, 1990; Family Expenditure Survey (1968–1983)) has about 20 apparent outliers, although not enough to affect the MLE fit in Figure 7. In Figure 8, 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. A *MLE* solution 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 $(.105, .236)$. However, using this solution as a starting value for L_2E converged to the previous curve in Figure 8.

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 as an initial guess, EM converged to essentially the first two-component EM solution with a small third diffuse component which covers the outliers — $0.0181\,\phi(x|2.12, .565^2)$. 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 8. 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.

Alois Kneip has investigated a longer time series of these data, many of which show three components. A time-parameterized three-component Gaussian mixture fitted by L_2E will be discussed in a separate paper by several of us.

Kim (1995) has studied the normal mixture problem extensively. Together with Terrell, they have 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. Maximum likelihood gives rise to least squares when residuals are assumed to be Gaussian, but least squares is used much more generally. At first glance, there does not appear to be a role for L_2E in regression problems. But if we focus on the distributional assumption of the residuals, then we 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

$$Y = a + bx + \epsilon \qquad \text{where} \quad \epsilon \sim N(0, \sigma_\epsilon^2)\,.$$

The L_2E criterion directly employs the parametric model of the residual density, $f(\epsilon) = \phi(\epsilon|0, \sigma_\epsilon^2)$. Invoking equation (6), we 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.

The maximum likelihood or least squares solution is

$$(\hat{a}, \hat{b}) = \arg\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2 .$$

The assumption of Normality must be verified after the fact using a variety of tests and graphical diagnostic plots of the residuals.

We compare these ideas on a simulated data set of 250 points, 50 of which are different from the true model $y = x + \epsilon$, $\epsilon \sim N(0,1)$. For clarity, the x-design was chosen so that the points are in three clusters; see Figure 9. In order to compare and display how the *L2E* and *MLE* criteria vary over (a, b), we have used the true value of $\sigma_\epsilon = 1$ in *L2E*. With this knowledge, *L2E* finds two plausible regression lines which explain different subsets of these data. Look closely at the location and shape of the residual density relative to the point $\epsilon = 0$ in the second row of Figure 9. In particular, the *L2E* plots clearly show the outlying cluster of points, while the *MLE* residual histogram is much less clear. Note that the mean of the *L2E* residuals is not necessarily equal to zero.

The indirect estimation of parameters in the *L2E* approach is reminiscent of the linear programming solution to the least absolute deviation criterion, $\sum |\epsilon_i|$. Decomposing the residuals into their positive and negative parts ($\epsilon_i = \epsilon_i^+ - \epsilon_i^-$), the objective function can be written as $\sum(\epsilon_i^+ + \epsilon_i^-)$, with constraints $\epsilon_i^+ \geq 0$, $\epsilon_i^- \geq 0$, and $y_i - a - bx_i = \epsilon_i^+ - \epsilon_i^-$. Thus the parameters a and b are in fact contained in the constraint matrix.

Finally, we revisit a real example discussed by Rousseeuw and Leroy (1987). In Figure 10, the least-squares regression line is heavily influenced by the data of four giant stars in the Hertzsprung-Russell diagram of the star cluster CYG OB1. The authors derive the least median squares (*LMS*) of residuals estimator, $\hat{y} = -12.3 + 3.90x$. The equation of our *L2E* fit for these data is quite similar, $\hat{y} = -8.77 + 3.11x$, with $\hat{\sigma}_\epsilon = 0.414$. Kernel estimates of the residuals for the LS and *L2E* fits are plotted in Figure 11, together with the fitted normal model of the residuals. Again, the *LMS* and *L2E* residuals are quite similar.

6. SIMULATION STUDIES

In this section, we report on simulations comparing *MLE*, *L2E*, and Minimum Hellinger Distance (*MHD*) parameter estimates in some of the settings above. In each case, the *MHD* estimator was fitted to a Gaussian kernel density estimator using the exact best bandwidth (Marron and Wand, 1992). By the conclusion of this section, we hope to obtain a basic understanding of the performance of these algorithms.

We first consider fitting a two-parameter normal model to $n = 100$ $N(0,1)$ data, repeated 1,000 times. Histograms of the *L2E* and *MHD* estimates of μ and σ are shown in Figure 12. The optimal bandwidth for the *MHD* target kernel density was $h^* = 0.44547$. Note that $\sigma(\bar{X}) = \sigma/\sqrt{n} = 0.01$, a value matched by *MHD*, while *L2E* is 25.0% greater than this (in close agreement to the theoretical figure of 24.1% given at the end of Section 3). For the standard deviation, we have that $\sigma(S) = \sigma/\sqrt{2n} = 0.071$ for the MLE. From our simulation, the *L2E* value is 38.0% greater than the *MLE* (compared to the 36.0% figure predicted by using Proposition 3.1). Observe that $\hat{\sigma}$ for *MHD* is biased upwards 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 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 $\hat{\sigma}_{MHD}$ is quite small. We computed the root mean squared error of $\hat{\sigma}_{L2E}$ and $\hat{\sigma}_{MHD}$ as 0.0984 and 0.1009, 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, we note that the use of bandwidths other than h^* has little affect on the estimate of location; however, the estimated standard deviation increases with the increasing bandwidth. Both the *L2E* and *MHD* estimation algorithms were implemented in Splus using the built-in functions *nlmin*, *dnorm*, and *integrate*, which perform quasi-Newton optimization, compute the normal density, and perform numerical integration. This approach gives approximately six significant digits. (Such accuracy increases estimation time and may be more than the required time in practice. Woodward, Whitney, and Eslinger (1995), for example, used 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 $(\bar{x}, s_x)$, although a number of other starting values were examined without observing any other solutions in this simple setting. Each individual sample of the simulation required an average of 0.016, 0.175, 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, 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 expands our study of the performance of these algorithms when given a set of contaminated points are present as in Figure 4. To each sample of 100 $N(0,1)$ points, 25 $N(c,1)$ points were combined, for $c = 0.0, 0.5, 1.0, \ldots, 10.0$. We expected for c sufficiently large, that any minimum distance estimator would eventually ignore the 25 "bad" data points. In Figures 13 and 14, boxplots of $\hat{\mu}_{L_2E}$, $\hat{\mu}_{MHD}$, $\hat{\sigma}_{L_2E}$, and $\hat{\sigma}_{MHD}$ for 256 simulations for each value of c are displayed. We note that all the parameters initially track the increasing contamination location, until the separation is apparent to the algorithm. We used the same $h^* = 0.44547$ appropriate for 100 $N(0,1)$ samples in the *MHD* optimizations. There were no observed local minima in the L_2E samples, but local minima (large $\hat{\sigma}_{MHD}$) were observed in an increasing number of the *MHD* samples with c greater than 5.0. We have recorded the better solution. Finally, we note that while $\hat{\mu}_{L_2E}$ 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 5, in which the predicted value is 1.176 for an 80% component. These figures confirm our hypothesis, although the *MHD* estimator is slower to ignore the contaminated data than we initially expected.

Our final simulation uses the 5-parameter 2-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 Splus functions mentioned above were used to estimate these parameters by *MLE*, L_2E, and *MHD* for 100 samples of size $n = 400$ from our favorite "hard" mixture

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

Boxplots for all 5 parameters and 3 methods are shown in Figure 15. The starting values were the true parameter values. The optimal bandwidth, $h^* = 0.2062$, was used in the *MHD* algorithm. We see similar results. 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 Splus optimization code returned a warning that the convergence was "false." We 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.

7. DISCUSSION

In this paper, we have demonstrated how a well-used nonparametric estimation criterion can be applied to a variety of parametric problems. In particular, our 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 our applications, we believe that a significant fraction of the data is well represented by a known parametric model. 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 data sets (for which careful data preparation is not feasible). In other words, we build models as good approximations and not as representing absolute truth. With large samples, standard testing almost always rejects our 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.

This line of research can trace its origins to our long-standing interest in bandwidth selection and to a series of lectures by George Terrell in the RIMS summer program at Rice University. For some time, Terrell has argued that penalized likelihood is not the most natural way to do nonparametric estimation, but rather to replace the likelihood with L_2E. For a variety of choices of penalty functionals, Terrell (1990) was able to derive well-known estimators (such as the kernel estimator), and then generalize to include a whole class of new estimators solving special problems (such as boundary estimators). I was originally curious to see how these would work with parametric models, and I dropped the penalty. Kim and Terrell have also examined a class of penalized normal mixture models, which leads to a quadratic programming problem. That could have great potential computational advantage compared to the pure L_2E version I have presented here.

The inefficiency of L_2E was noted in a recent technical report pointed out to me by Manny Parzen. Basu et al. (1998) have proposed a family of minimum divergence estimators, which range from maximum likelihood to L_2E and beyond. Many of the problems considered here are also considered by the authors. Their procedure gives a family of parameter estimates, indexed by a metaparameter, $\alpha \geq 0$, that essentially changes the weights in the score function. The general form of the minimum divergence criterion is

$$\int f(x|\theta)^{1+\alpha}dx - \frac{1+\alpha}{\alpha} \cdot \frac{1}{n}\sum_{i=1}^{n} f(x_i|\theta)^{\alpha} . \tag{14}$$

L_2E is the special case $\alpha = 1$, and *MLE* is the special case $\alpha \to 0$. At this time, the authors do not believe there is any general advice about which metaparameter to choose, short of choosing some particular level of inefficiency (α closer to zero is recommended, for example). 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, sacrifices some asymptotic efficiency for clear benefits in difficult problems faced by practicing statisticians. Basu, et al. also give details about hypothesis testing possibilities and breakdown points. Clearly, the parametric bootstrap will be applicable as well. A deeper understanding of the practical differences, if any, between distance and divergence estimators would be interesting.

Our 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 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 Base, et al, namely, fitting normal mixtures. The integral $\int f^{1+\alpha}$ in Equation (14) does not have a closed form for $0 < \alpha < 2$ except at L_2E; see equation (13). For fitting multivariate normal models, existence of a closed form criterion is of great practical importance.

We have begun to investigate a large number of other applications. One involving estimation of an economic stochastic frontier function will be reported elsewhere (Scott, Simar, and Wilson, 1998). Multivariate regression and time series, especially with massive data sets, are particularly interesting. Other statistical algorithms which could benefit from robustness are excellent candidates for minimum distance procedures. We hope to explore those soon.

REFERENCES

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.

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.

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*, Chapman & Hall, London.

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

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

Kim, Donggeon, (1995), "Least Squares Mixture Decomposition Estimation," Unpublished doctoral dissertation, Virginia Tech, Blacksburg, VA.

Marron, J.S. and Wand, M.P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, **20**, 712–36.

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

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

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

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*, John Wiley & Sons, New York.

Scott, D.W., Simar, L., Wilson, P.W. (1999), "A New Model and Criterion for Estimating the Stochastic Frontier Error," in preparation.

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

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

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

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

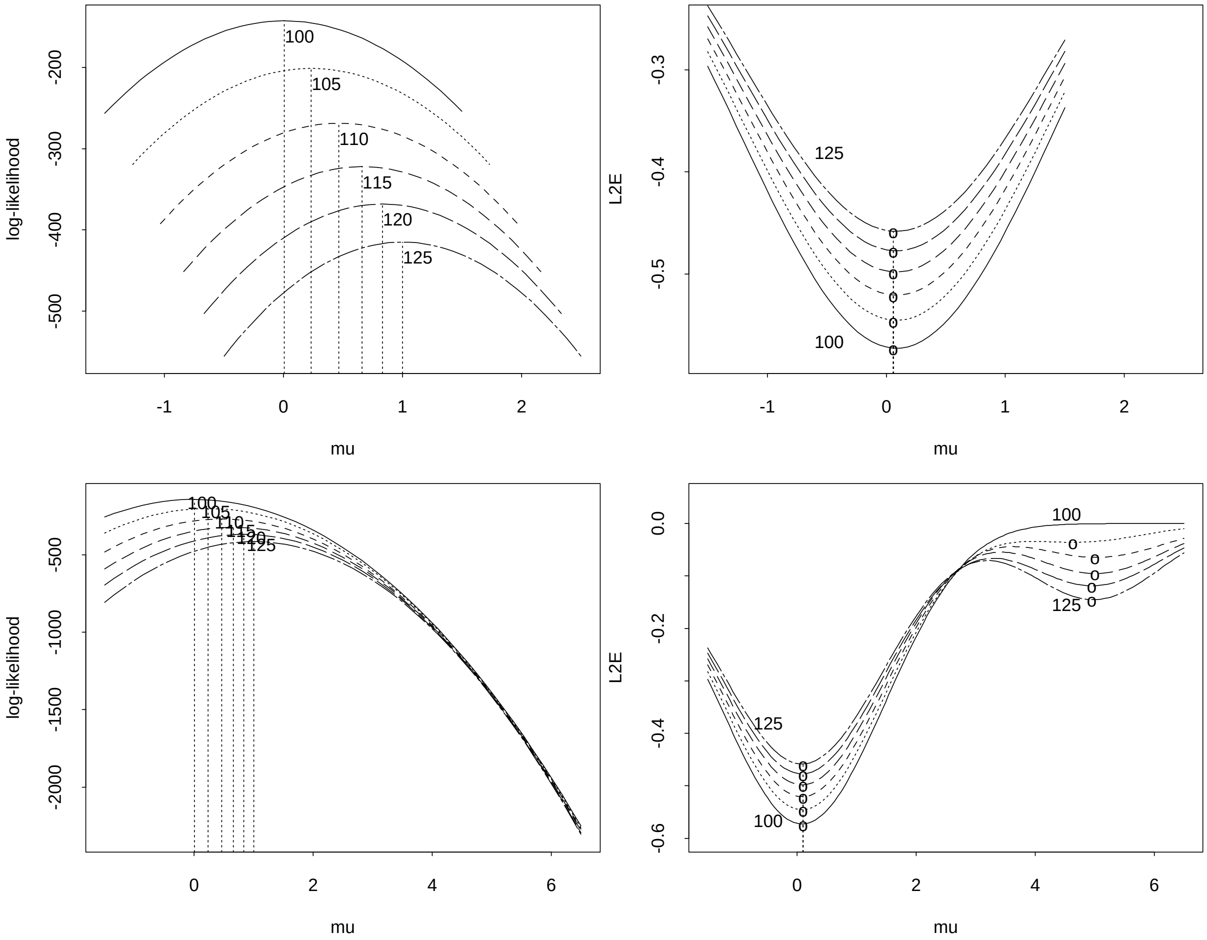

Figure 1: (Upper left) Log-likelihood profiles for contaminated data (n = 100, 105, 110, 115, 120, 125). Best values are indicated by circles and/or dotted lines. (Upper right) L_2E profiles for same data. (Bottom left) MLE profiles on a larger interval. (Bottom right) L_2E profiles on a larger interval.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

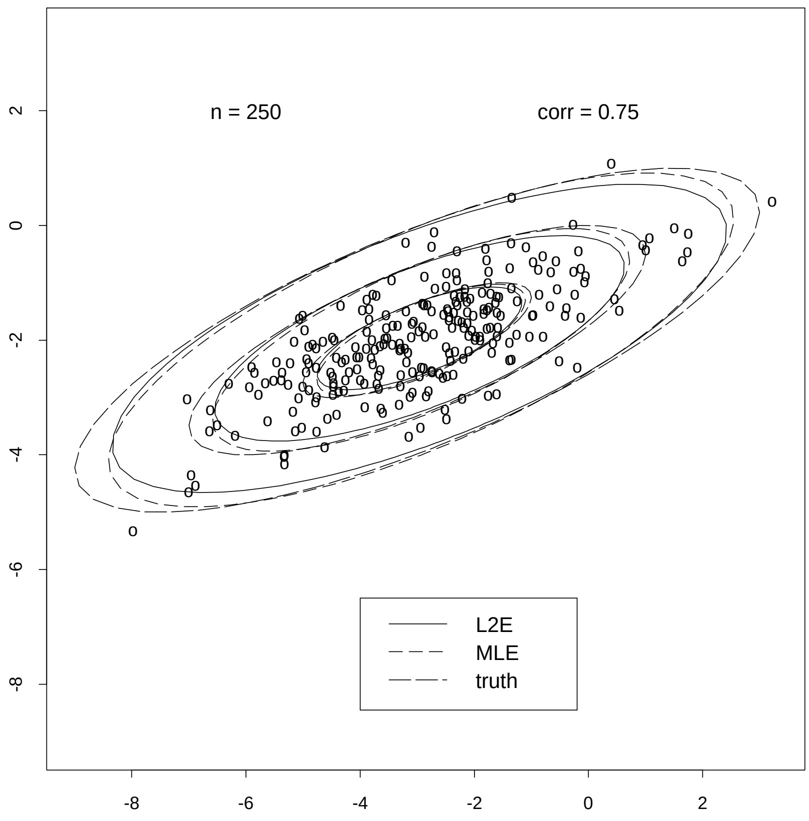

Figure 2: Fitted bivariate normal (5 parameters) using L_2E. The estimated correlation coefficients by L_2E and MLE were 0.697 and 0.725 versus the true 0.750.

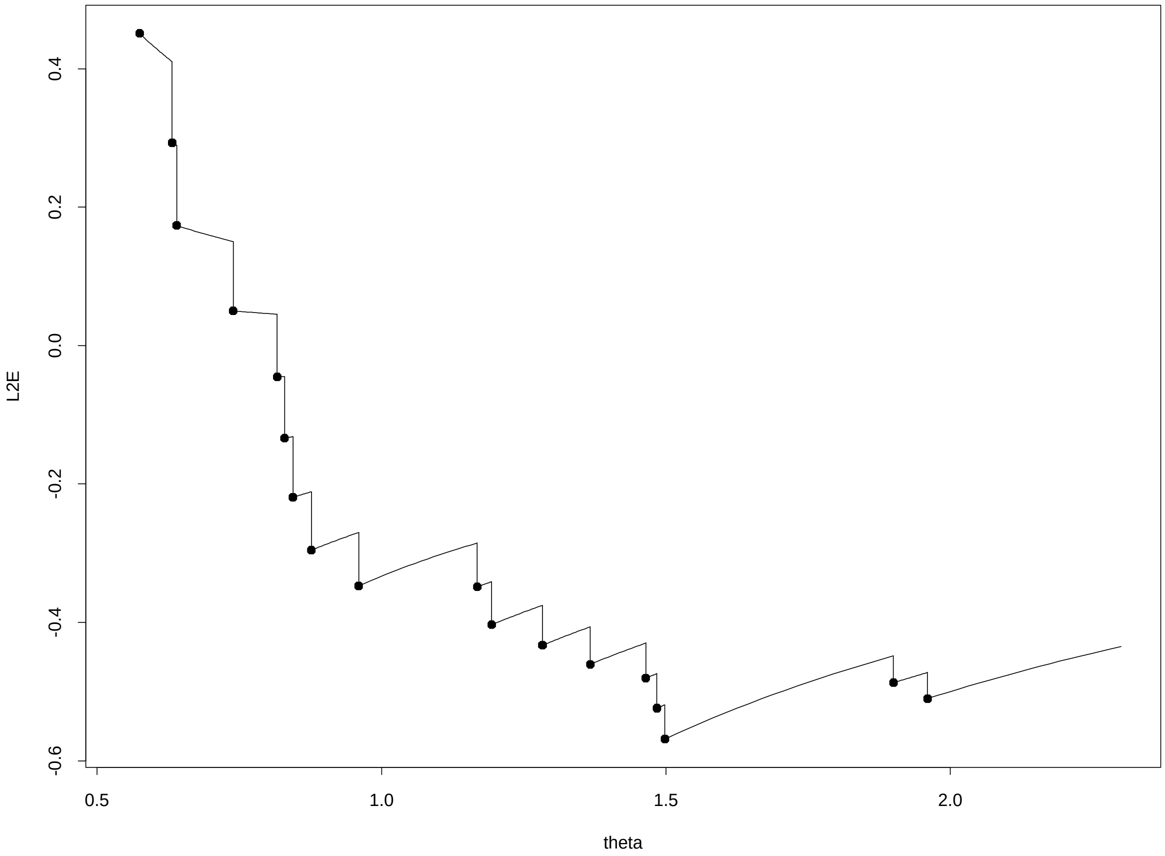

Figure 3: The L_2E function when $f(x|\theta)$ is $U(0,\theta)$ for a sample of 25 points from $U(0,1.5)$ with two contaminated points added at 1.90 and 1.96. $\hat{\theta}_{L_2E} = 1.498$ in this case.

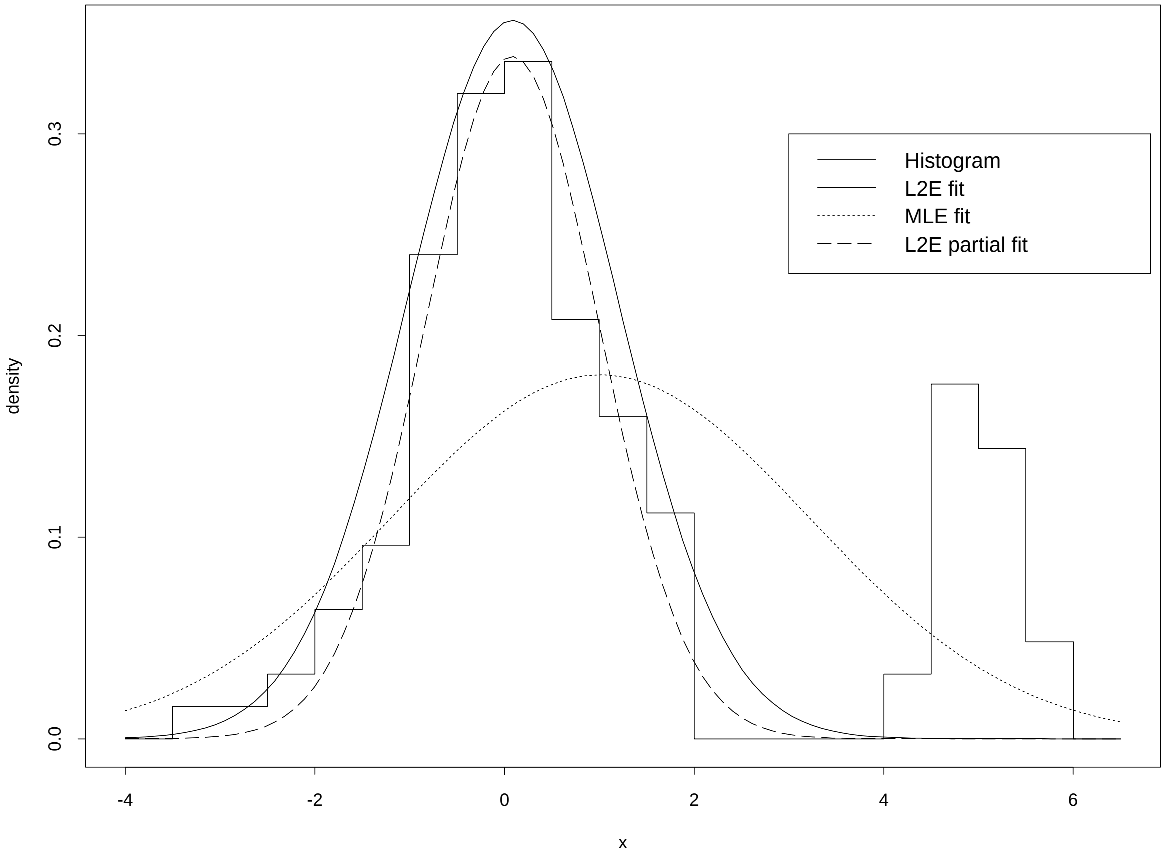

Figure 4: The parametric L_2E result with $f(x|\boldsymbol{\theta}) = N(\mu, \sigma^2)$ for the $n = 125$ points from the mixture $0.8N(0,1) + 0.2N(5,1)$ shown in Figure 1. The *MLE* is displayed as well. The last curve is a partial L_2E fit with the 3-parameter model $w \cdot N(\mu, \sigma^2)$.

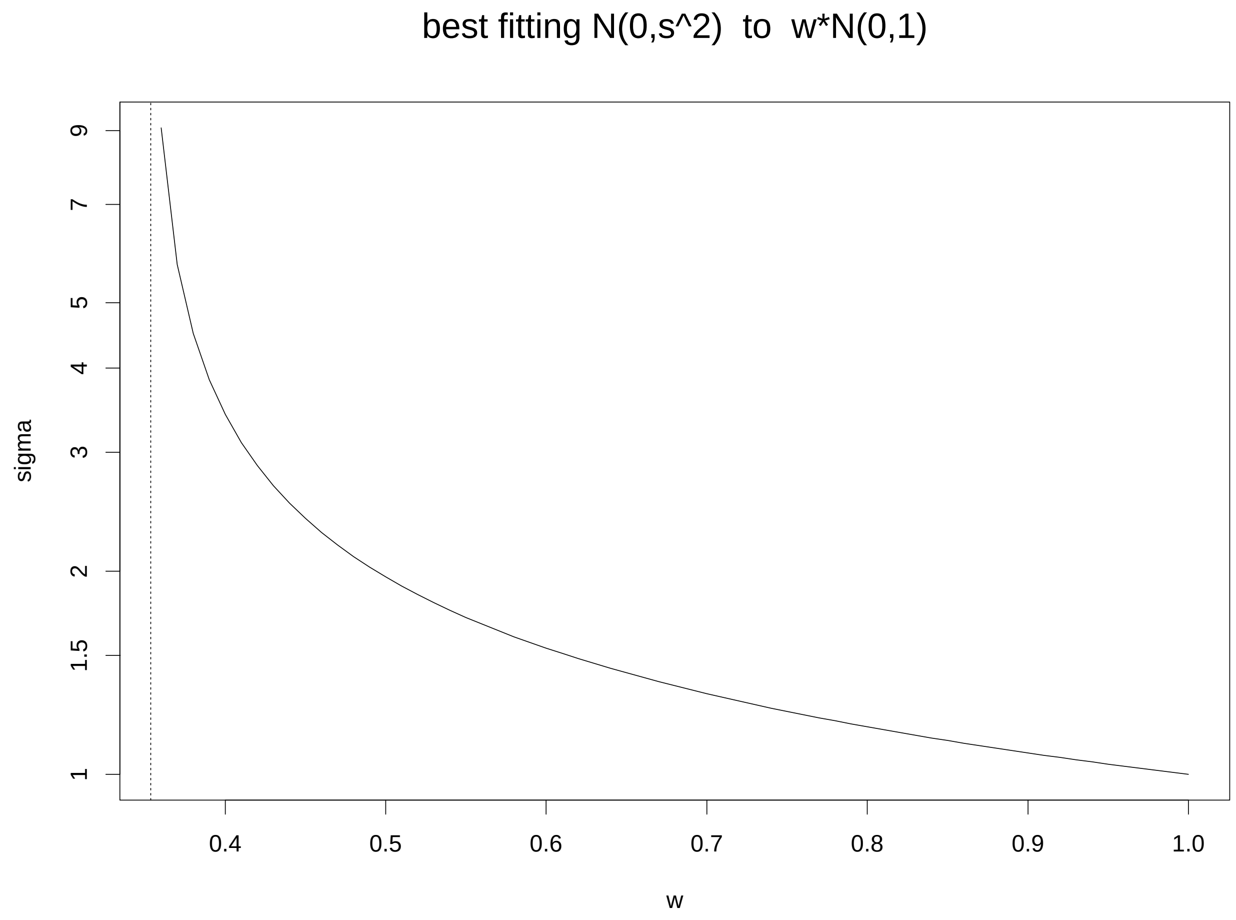

Figure 5: Suppose the true density has the component, $w\,\phi(0,1)$, with the remaining $1-w$ mass far away, and that this component is to be fitted by the model $N(0,\sigma^2)$ minimizing L_2 error. This figure contains the optimal pairs (w,σ), for $0 < w < 1$.

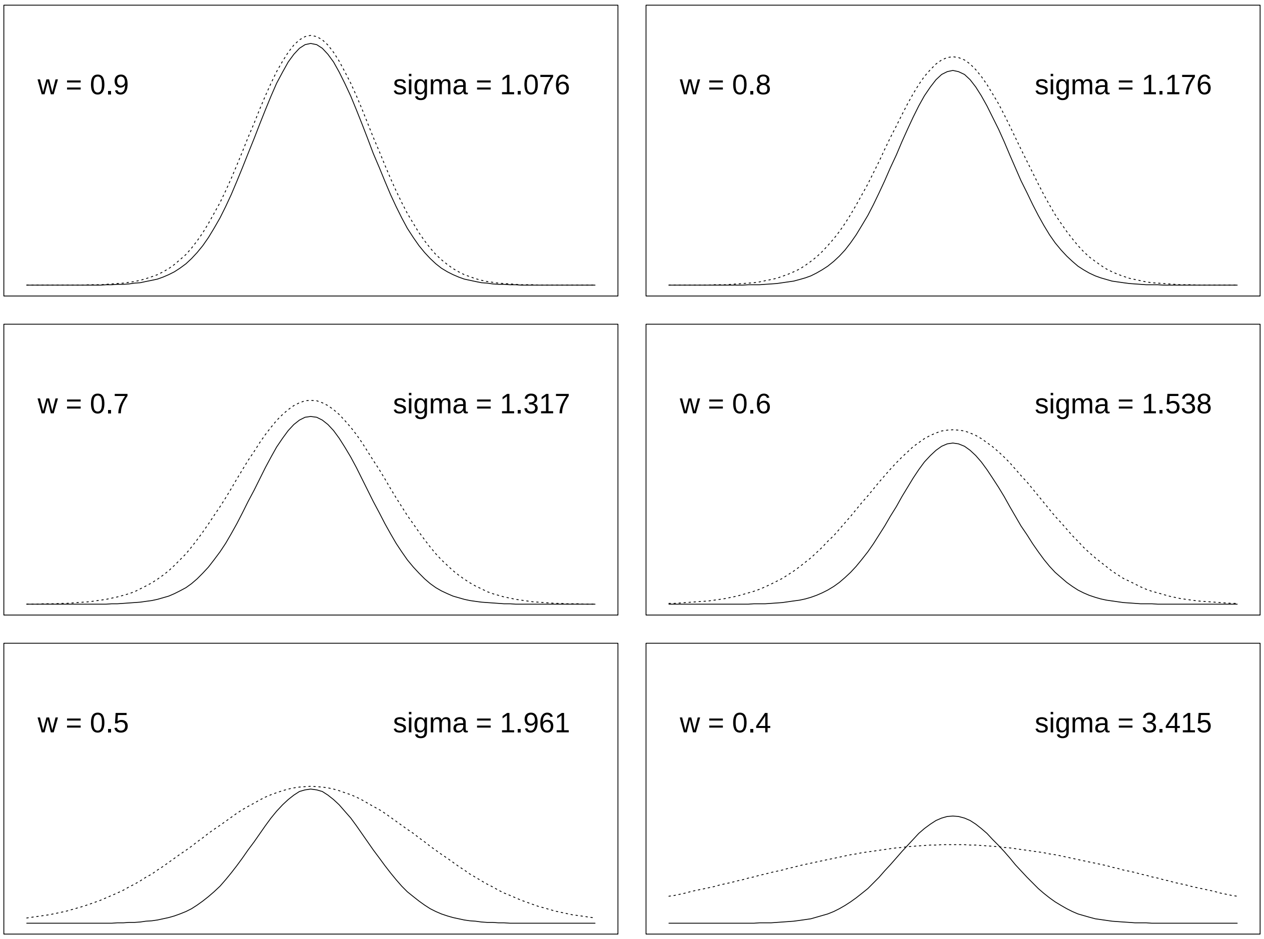

Figure 6: Plots of the component, $w\,N(0,1)$, (solid line) and the best fitting model, $N(0,\sigma^2)$, (dashed line) for several values of w. For $w \leq 0.3536$, $\sigma = \infty$; see text.

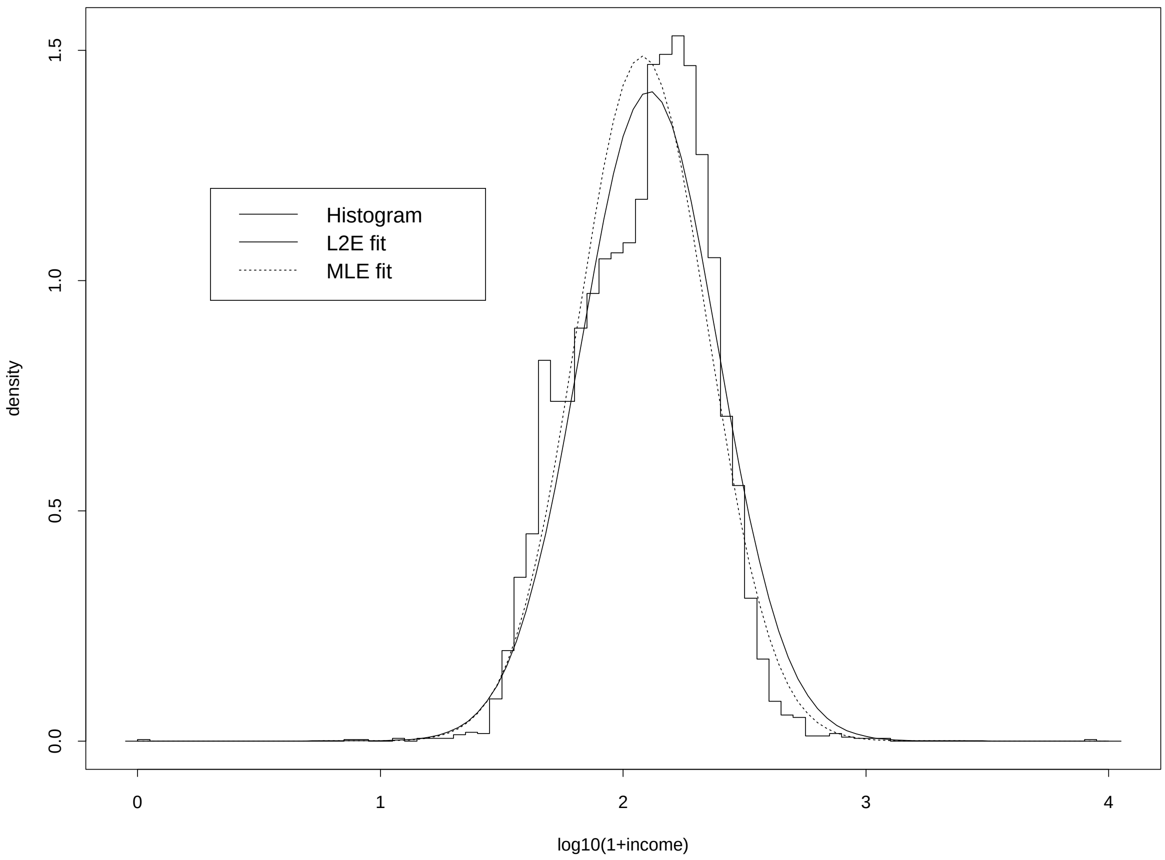

Figure 7: L_2E and MLE Gaussian estimates of the British household income data.

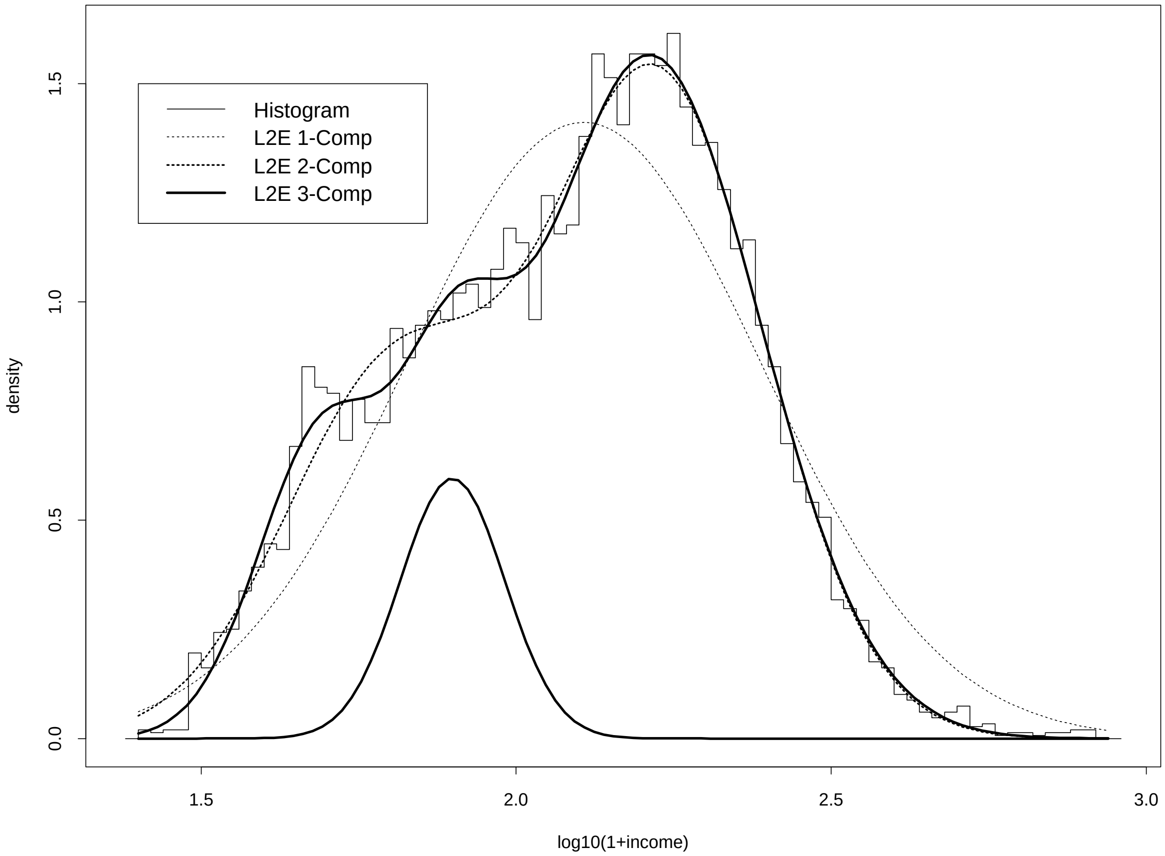

Figure 8: The one-, two-, and three-component L_2E Gaussian estimates of the British household income data. The middle component of the three-component fit is also shown. The other components are apparent from the left and right edges. (Note the expanded horizontal scale.)

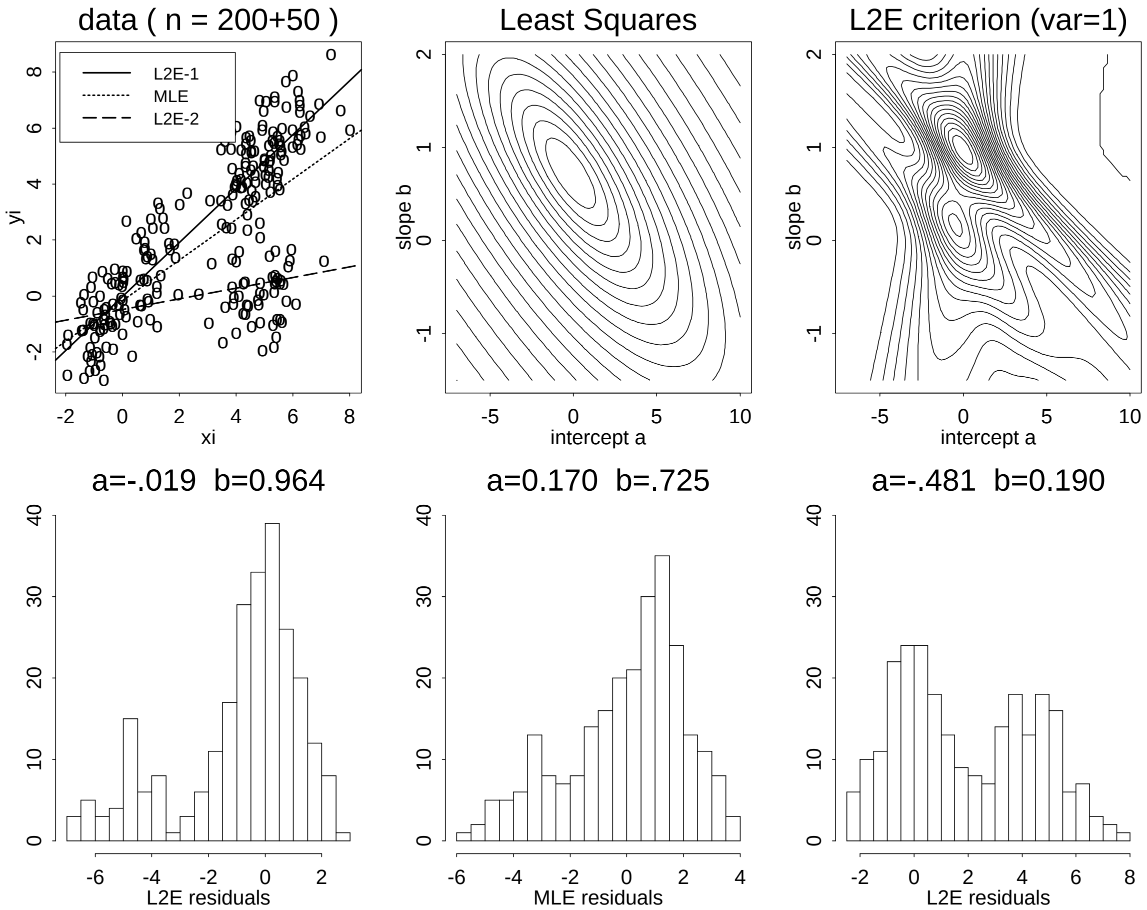

Figure 9: *MLE* and L_2E regression criteria with data and estimates shown in the upper left frame. Residual plots for the single *MLE* curve and the two L_2E curves are shown in the second row. Note the location of the origin ($\epsilon = 0$) in each.

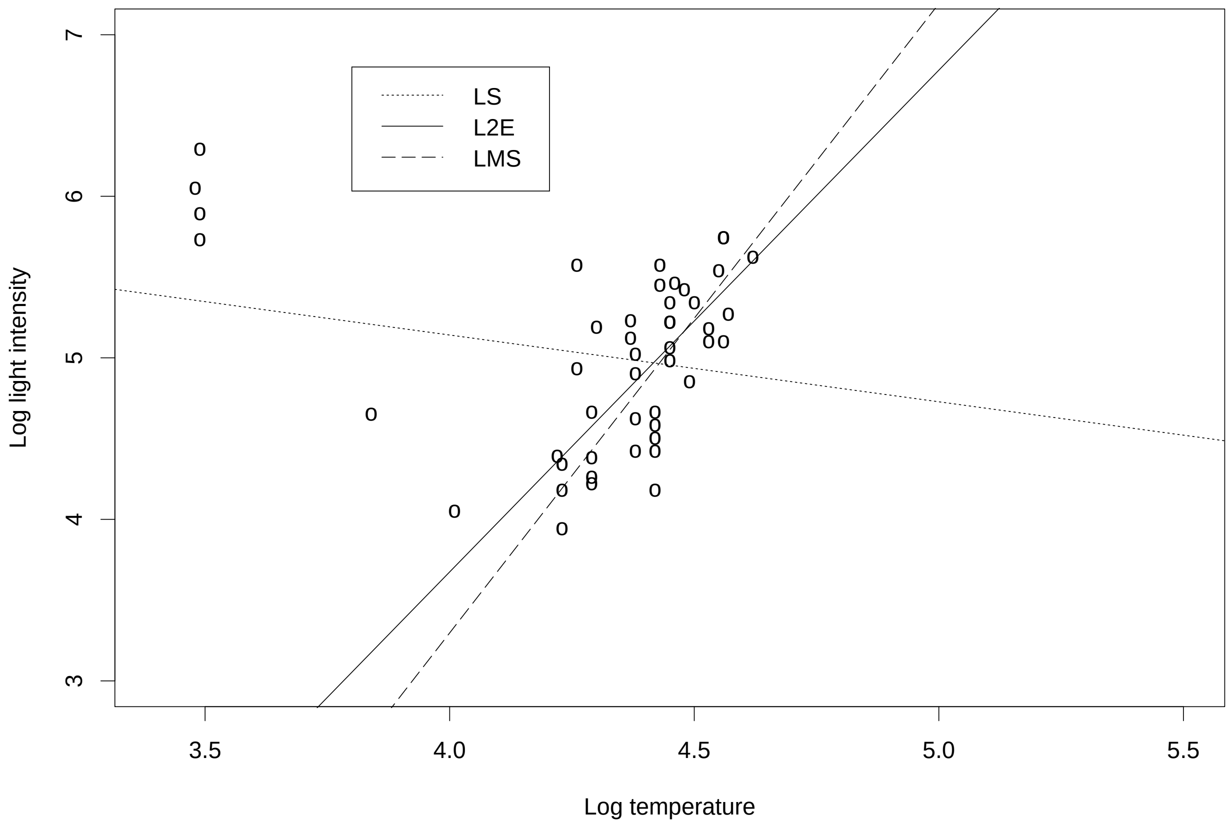

Figure 10: The least squares, L_2E, and least medium squares estimators for the Hertzsprung-Russell diagram of the star cluster CYG OB1.

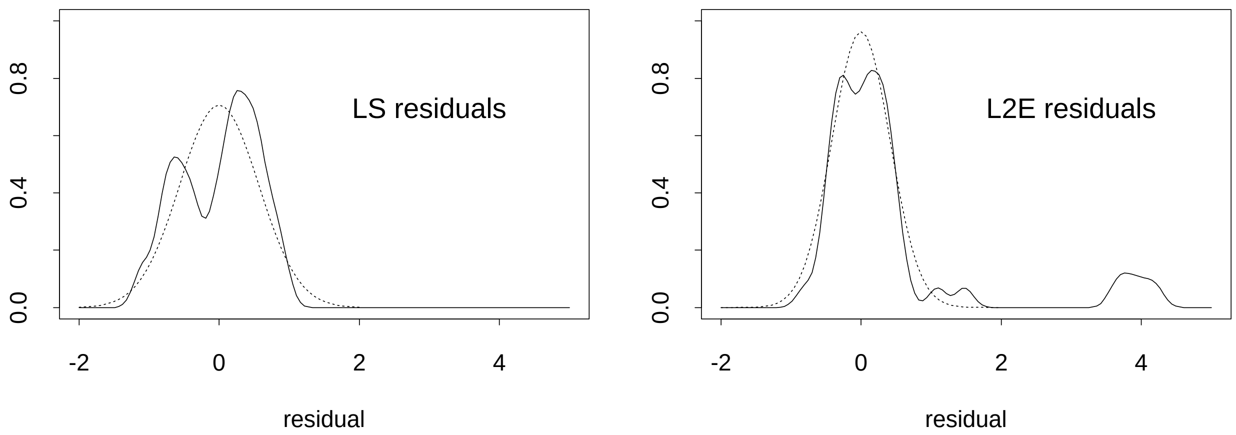

Figure 11: Gaussian kernel density estimates of the estimated residuals from the least squares and L_2E fits, together with the estimated normal models for the residuals (dotted line).

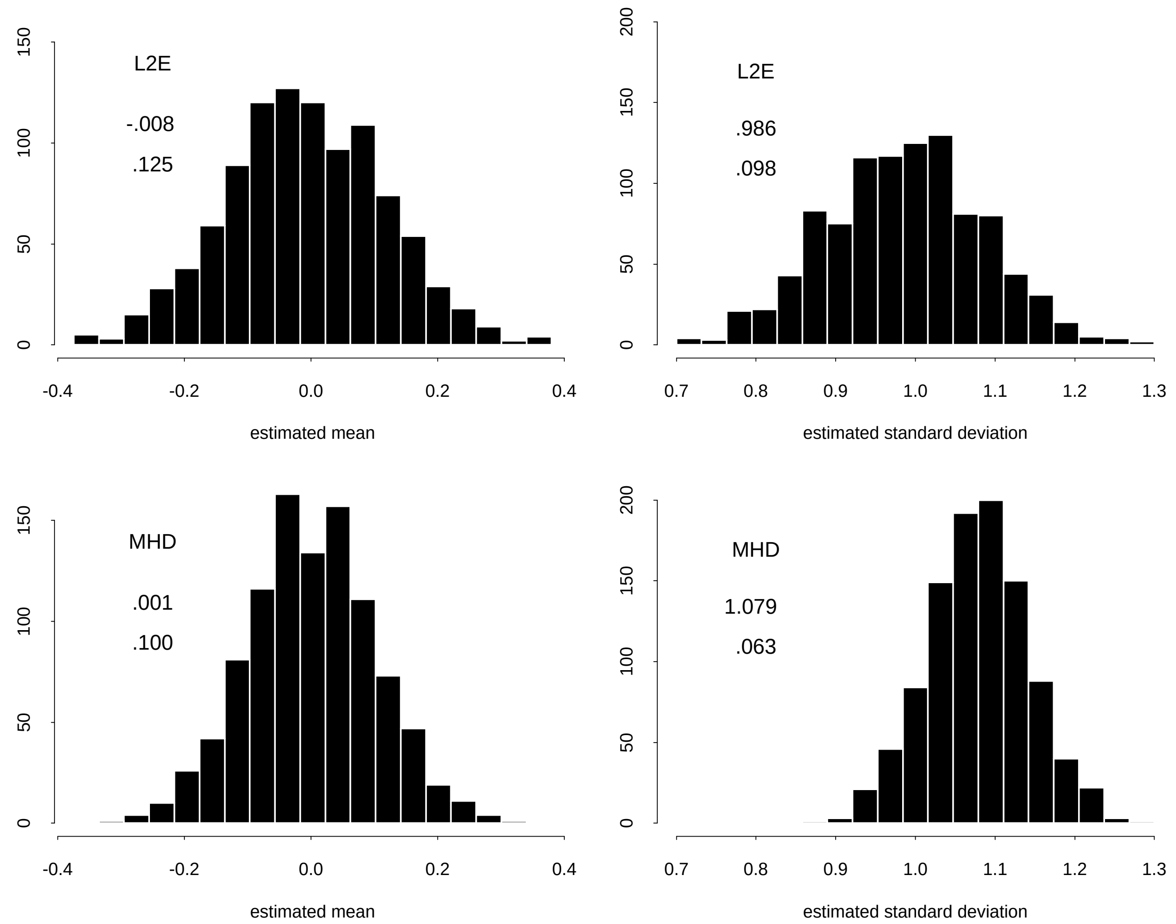

Figure 12: Frequency curves of estimated parameters $(\hat{\mu}, \hat{\sigma})$ from 1,000 $N(0,1)$ samples using L_2E and MHD algorithms. The sample mean and standard deviation for each statistic is given on the figure.

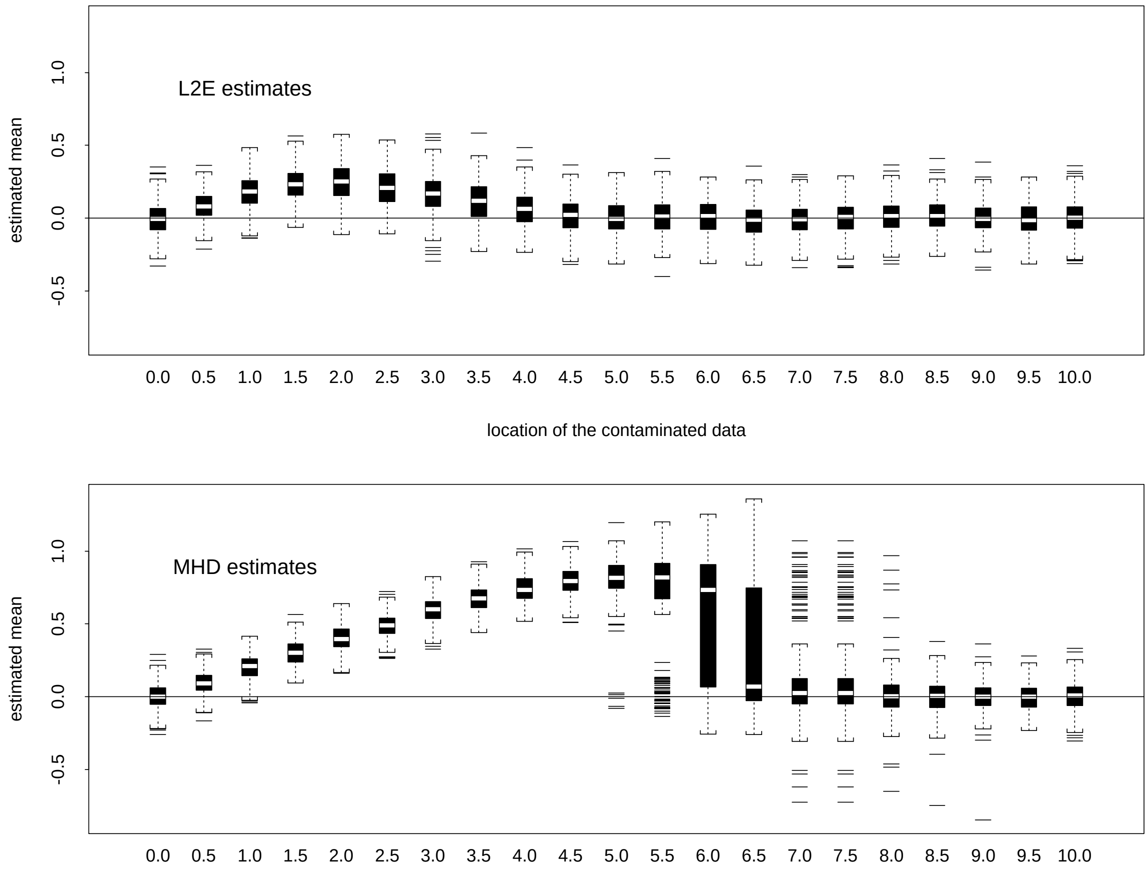

Figure 13: In this figure, 256 samples of size 125 were generated for 13 different contamination locations, which are indicated along the horizontal axis. 100 standard normal samples were combined with 25 $N(c, 1)$ points, for $c = 0.0$, 0.5, $1.0, \ldots, 10.0$. Boxplots of the estimated location parameter are shown for L_2E and MHD algorithms which attempted to fit the partially correct model, $N(\mu, \sigma^2)$. (See Figure 14 for $\hat{\sigma}$.)

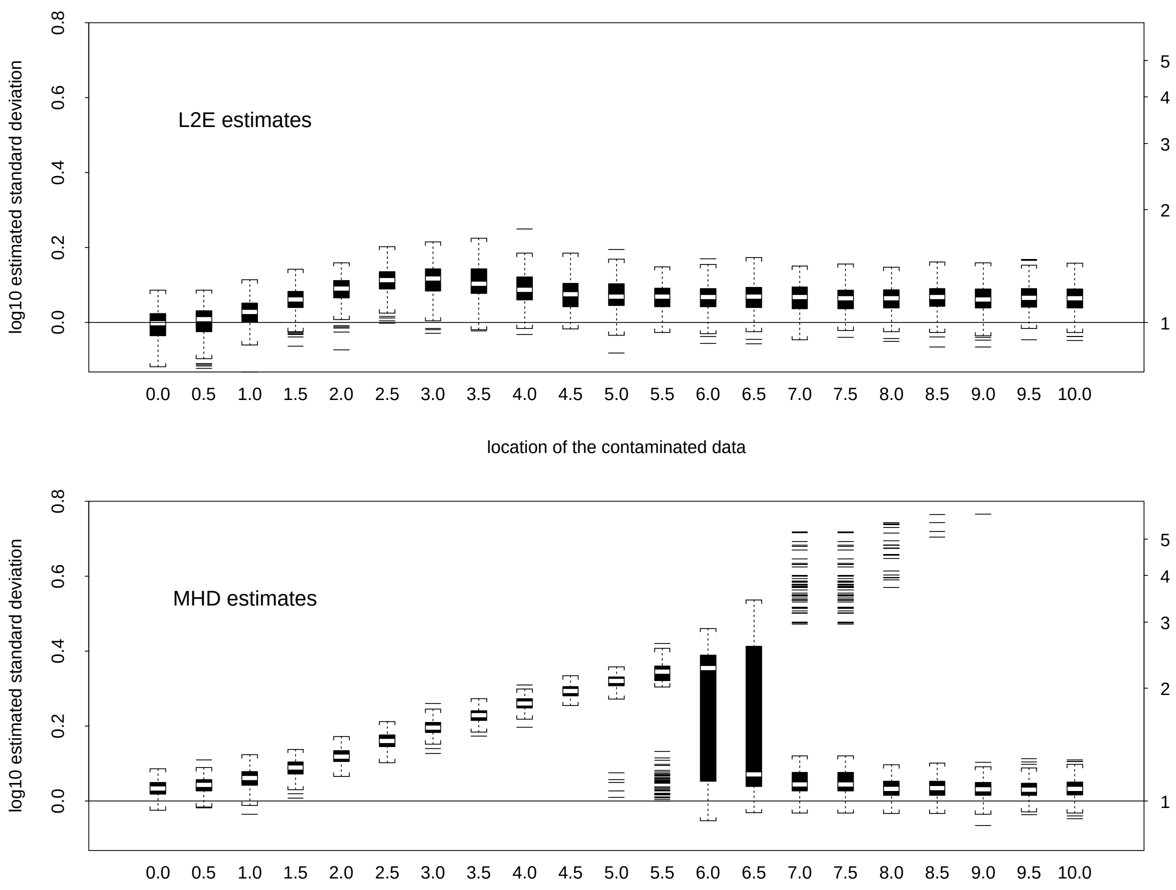

Figure 14: Continuing the summary of data depicted in Figure 13, boxplots of the estimated standard deviation parameter are shown for L_2E and MHD algorithms which attempted to fit the partially correct model, $N(\mu, \sigma^2)$.

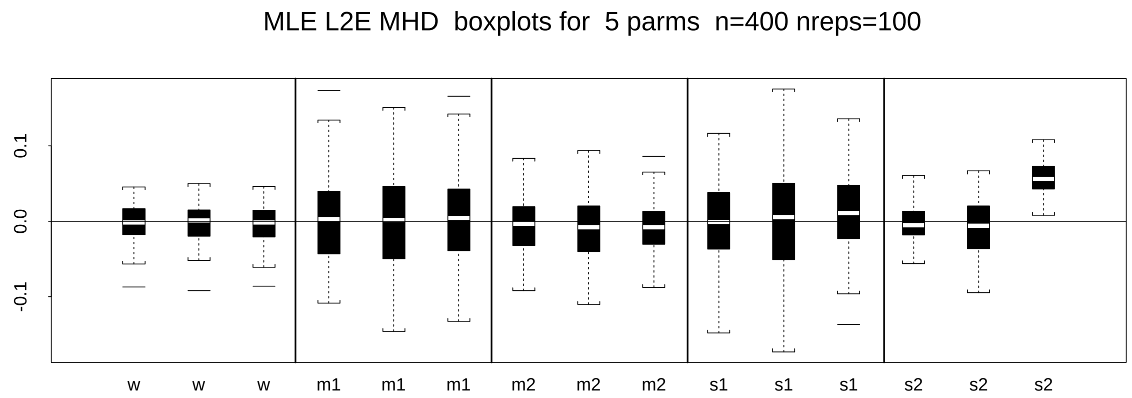

Figure 15: Boxplots of estimated *MLE*, *L_2E*, and *MHD* (left-to-right) parameters from 100 simulations of a 5-parameter normal mixture. The boxplots have been centered by subtracting the true parameter values, $w = 0.75$, $\mu_1 = 0$, $\mu_2 = 3$, $\sigma_1 = 1$, and $\sigma_2 = 1/3$.