Parametric Modeling by Minimum L_2 Error[1]

David W. Scott[2]

November 16, 1998

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 estimation. Parametric alternatives to maximum likelihood estimation have received theoretical treatment primarily in the form of Hellinger distance, which seems to be the most popular in the class of minimum distance estimators. In this paper, we investigate the integrated squared error, or L_2 distance, as a theoretical and practical estimation tool for a variety of simple and complex parametric statistical models. We demonstrate by example the well-known result that minimum distance estimators are inherently robust, without requiring specification of any tuning factors found in robust likelihood algorithms. We show that the asymptotic inefficiency of the parameters estimated by minimizing the integrated squared error or L_2 estimation (L_2E) criterion is roughly that of the median versus the mean. As with maximum likelihood estimation, specification of the L_2E criterion requires no additional metaparameters, such as penalty weights or influence function thresholds. Coupled with its robustness property, L_2E is particularly appropriate for analyzing massive data sets where data cleaning is impractical and statistical efficiency is a secondary concern. We believe that the choice of L_2E as a method of data fitting is more consistent with the philosophy that our models are only approximations to truth. Setting up the L_2E criterion is relatively simple even

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

[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.

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; Resistant estimation; Minimum distance; Normal mixture models; Cross-validation; Integrated squared error.

1. INTRODUCTION

Some of the major deficiencies of maximum likelihood estimation are the lack of resistance to outliers and the general nonrobustness to with respect to model misspecification. However, other estimation criteria such as minimum distance techniques have been shown to have excellent robustness properties without any modification (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. In nonparametric density estimation, another dimensionless quantity, the L_1 error, is of special interest since it is directly related to classification error. However, these distance measures are not fully tractable in practice, and certain approximations and bounds are typically encountered. For example, one suggestion is to find the parameter value, $\hat{\theta}$, that minimizes the Hellinger distance between $f(x|\hat{\theta})$ and a kernel density estimate (Beran, 1977). Clearly, $\hat{\theta}$ will depend upon the choice of smoothing parameter in the kernel estimator so that certain approximations remain.

Not only are parametric and nonparametric estimators worlds apart in philosophy, but they seldom share estimation criteria. Parametric algorithms typically rely on maximum likelihood while nonparametric algorithms favor L_p error measures, with integrated squared error (*ISE*, $p = 2$) a favorite. Using likelihood and least squares in nonparametric estimation is growing in popularity. This study arose from a desire to understand the applicability of nonparametric criteria to parametric problems.

Consider the simple parametric estimation of the 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. Consider alternative estimators of the form $\hat{\theta} = c\, x_{(n)}$, for some constant c. Consideration of $E\hat{\theta}$ gives $c = (n+1)/n$ (unbiasedness), while mean squared error gives $c = (n+2)/(n+1)$ (and variance with the surprising $O(n^{-2})$ rate). At the other extreme, Scott (1992) showed that $c = 2^{1/(n-1)}$ minimizes the average or mean integrated square error (*MISE*); note $c \approx 1 + (n-1)^{-1} \log 2$. All three bias the *MLE* upward ($c > 1$), a sensible thing to do. This simple example, however, highlights the most important advantage of the *MLE*, namely the constructive nature of the estimator. The other criteria were only applied in a one-dimensional neighborhood of the *MLE*. Furthermore, maximum likelihood estimates generally enjoy asymptotic optimality properties.

In this paper, we devise a fully constructive parametric algorithm based upon the integrated squared error criterion. We demonstrate its robustness behavior (which is predicted by general results of minimum distance procedures). We show how the basic density estimation framework may be extended to general statistical model estimation.

2. THE PARAMETRIC MINIMUM L_2 ERROR CRITERION

Mats Rudemo (1982) and Adrian Bowman (1984) gave a fresh viewpoint to the problem of estimating the bin width of a histogram. 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*:

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

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. 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

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

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 shift.

Let us rewrite equation (1) for our parametric problem as

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

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 level of the density is the key quantity to estimate as in equation (3). The data partitioning trick is not required in the parametric setting, since we have a random sample of size n to estimate the average height. Thus the proposed estimator minimizing the parametric integrated squared error (L_2E) criterion is

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

Compare equations (6) and (4). Here we have assumed that we know the correct parametric family; however, we will also be interested in the case when the assumed parametric form $f(x|\theta)$ is in fact incorrect.

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

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

Interestingly, L_2E maximizes the sum of the density while MLE maximizes the product of the density. The robustness of $\hat{\theta}_{L_2E}$ can be easily demonstrated in this example. Consider a sample of size 100 from $N(0,1)$ with up to 25 additional samples from a contaminated density, $N(5,1)$. In Figure 1, we plot the log-likelihood and L_2E functions for $n = 100, 105, \ldots, 125$, always adding 5 new samples from the contamination density while retaining 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 no sign of the contamination, while the L_2E shows local minima in the vicinity of contamination. The local minimum is appropriate to consider since the contaminated data also come from the assumed form of the density. Observe how the contaminated data affect the level of these sets of curves. In particular, extrapolating to 100 contaminated samples, we see from the lower right frame that the L_2E breakdown point would be reached.

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

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

In Figure 2, we examine a sample of 250 bivariate normal points with $\boldsymbol{\mu} = (-3, -2)$, $\boldsymbol{\sigma}^2 = (4, 1)$, and $\rho = 0.75$. The parameters fitted by L_2E are $\hat{\boldsymbol{\mu}} = (-2.96, -1.97)$, $\hat{\boldsymbol{\sigma}}^2 = (3.22, 0.80)$, and $\hat{\rho} = 0.70$. The parameters fitted by *MLE* are $\hat{\boldsymbol{\mu}} = (-2.93, -2.00)$, $\hat{\boldsymbol{\sigma}}^2 = (3.34, 0.94)$, and $\hat{\rho} = 0.73$. The *nlmin* function in Splus converged in 27 iterations starting from the true values. This example simply demonstrates the multivariate extension, which will be considered more fully in a separate paper.

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

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

It is easy to check that for most samples, the L_2E and *MLE* estimates are both $x_{(n)}$. However, if the ratio of the adjacent order statistics is too big, then the larger order statistic cannot be the L_2E minimizer; see Figure 3. It is easy to see, for example, that $\hat{\theta} \neq x_{(n)}$ if $x_{(n)}/x_{(n-1)} > n/(n-2)$; that is, if $x_{(n)}$ is an "outlier."

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*

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

where the first sum is over all possible values of the random variable, and the second sum is over the observed samples.

We do not provide any examples here, but many applications rely upon models of discrete random variables. An interesting open problem is handling densities that are a mixture of continuous and discrete components, which would be helpful for censored data. 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. In an example with the Normal density, the relative efficiency compared to maximum likelihood is seen to be similar to the median versus the mean and the interquartile range versus the standard deviation. These results are a special case in a class of divergence measures considered by Basu, et al. (1998); see discussion.

Theorem 3.1 (Univariate L_2E Asymptotics) *Under similar assumptions as for maximum likelihood estimation, the L_2E parameter is consistent and asymptotically normal:*

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

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

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

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

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

In both equations (8) and (10), the expected value of the random sum exactly cancels the integral immediately preceding the sum, 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

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

This completes the proof of the theorem.

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

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

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)=AN\left(\mathbf{0},\boldsymbol{H}^{-1}\left[\boldsymbol{G}_2-\boldsymbol{G}_1\boldsymbol{G}_1^T\right]\boldsymbol{H}^{-1}\right),$$

where

$$\begin{aligned}\boldsymbol{G}_1 &= \int_{\Re^p}\nabla_{\boldsymbol{\theta}}f(\boldsymbol{x}|\boldsymbol{\theta}_0)f(\boldsymbol{x}|\boldsymbol{\theta}_0)\,d\boldsymbol{x}\\ \boldsymbol{G}_2 &= \int_{\Re^p}\nabla_{\boldsymbol{\theta}}f(\boldsymbol{x}|\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}}f(\boldsymbol{x}|\boldsymbol{\theta}_0)^T f(\boldsymbol{x}|\boldsymbol{\theta}_0)\,d\boldsymbol{x}\\ \textit{and}\quad \boldsymbol{H} &= \int_{\Re^p}\nabla_{\boldsymbol{\theta}}f(\boldsymbol{x}|\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}}f(\boldsymbol{x}|\boldsymbol{\theta}_0)^T\,d\boldsymbol{x}\,.\end{aligned}$$

Example 3.1 (Normal Density) *If $X\sim N(\mu,\sigma^2)$, then Theorem 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, Section 10.12, Volume I).

4. FITTING GAUSSIAN MIXTURE DENSITIES

A powerful parametric density model is the mixture model (Titterington, et al., 1985):

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

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 getting the number of components correct. 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:

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

Wand and Jones (1995) present a similar formula in the multivariate setting. Using logarithmic transformations for the variances and logistic-like transformations for the weights leads to an unconstrained optimization problem that can be solved with standard Newton method algorithms such as *nlmin* in Splus.

We begin by examining a very practical property of L_2E mixture fitting; 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 $\sigma = 1$ was known before). With the two-parameter model, there is a unique L_2E minimizer (aside from singular Dirac solutions), which is shown in Figure 4 along with the data histogram and *MLE* fit.

This L_2E property, which is actually robustness in another guise, will be proved valuable again and again. In particular, we can find groups of outliers in many circumstances. We will also find this property useful when we turn to regression problems below.

A more challenging mixture problem is displayed in Figure 5. A sample of the incomes of 7,203 British households in 1982 (Härdle, 1990; Family Expenditure Survey (1968–1983)) has a number of apparent outliers, although not enough to affect either of ($K = 1$) Gaussian fits displayed. In Figure 6, the $K = 1-3$ 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). Yet another 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 led to the original curve shown in Figure 6.

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 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 6. Alternatively, a four-component EM solution could handle the outliers with the fourth component (and several thousand EM iterations). This example illustrates the unpredictable influence outliers can have on the EM algorithm, 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 this 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 fitting the L_2E criterion to the distribution of the estimated residuals.

Consider the simple linear regression model

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

The maximum likelihood solution is

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

Note that knowledge of $(\hat{a}, \hat{b})$ is required to estimate σ_ϵ^2, but not vice versa, and that the assumption of Normality must be verified after the fact using a variety of tests and graphical diagnostic plots of the residuals.

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

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

where $\epsilon_i = y_i - a - bx_i$. Note that all three parameters are estimated simultaneously in the numerical optimization. Including σ_ϵ in this step allows L_2E to find the most Normal set (or subset) of residuals available.

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 groups; see Figure 7. In order to compare and display both the L_2E and *MLE* criteria, we have used the true value of $\sigma_\epsilon = 1$. With this knowledge, L_2E has no problem finding two plausible regressions which explain different portions of these data. Look closely at the size of the Gaussian bump around $\epsilon = 0$ in the residual plots in the second row of Figure 7. In particular, the L_2E plots clearly show the outlying cluster of points, while the *MLE* residual histogram is much less clear. Note that the mean of the L_2E residuals does not necessarily equal zero.

The indirect estimation of parameters in the L_2E approach is reminiscent of the linear programming solution to the least absolute deviation problem given by

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

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 found by linear programming are in fact hidden in the constraint matrix.

Finally, we present a real example that has structure similar to that in Figure 7. We have taken the 107 consecutive durations of the Old Faithful Geyser data set (Weisberg, 1980), and created 106 points of the form (x_t, x_{t+1}). A linear fit to these data is essentially a first-order autoregressive model. Our L_2E fit finds all three parameters, as the residual variance is unknown. Once again, there is a unique L_2E fit in this example. The lowess fit is included for comparison with its default settings.

6. DISCUSSION

In this paper, we have demonstrated how a 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 reliably and quickly. These parameters are inefficient compared to maximum likelihood theory at the correct model. But as nonparametric estimators 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. *L2E* is better suited to treating models as approximations, handling outliers and underspecified models in a useful manner. A comparison of *L2E* and *MLE* fits is almost always informative.

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 to 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 *L2E*. For a variety of choices of penalty functionals, Terrell (1990) was able to derive well-known estimators (such as the kernel estimator), but then generalize to include a whole class of 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 *L2E* version I have presented here.

The inefficiency of *L2E* 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 *L2E* 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. *MLE* and *L2E* correspond to $\alpha = 0$ and $\alpha = 1$, respectively. 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. 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. 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.

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.

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.

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.

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. (1998), "A New Model and Criterion for the Stochastic Frontier Error Structure," 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.

Weisberg, S. (1980), *Applied Linear Regression*, John Wiley & Sons, New York.

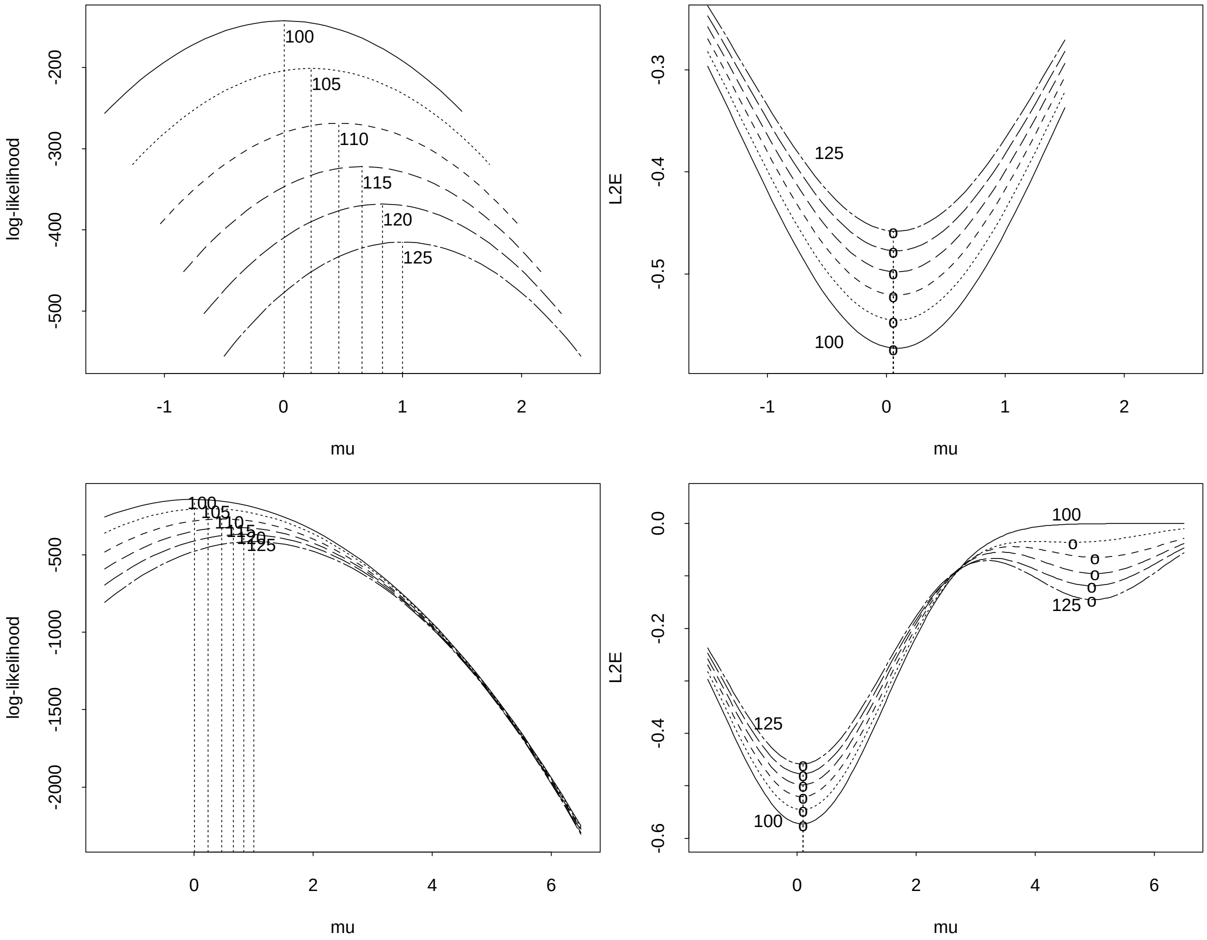

Figure 1: (Upper left) Log-likelihood profiles for contaminated data (n = 100, 105, 110, 115, 120, 125). Best values are indicated by circles and/or dotted lines. (Upper right) L_2E profiles for same data. (Bottom left) *MLE* profiles on larger interval. (Bottom left) L_2E profiles on larger interval.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

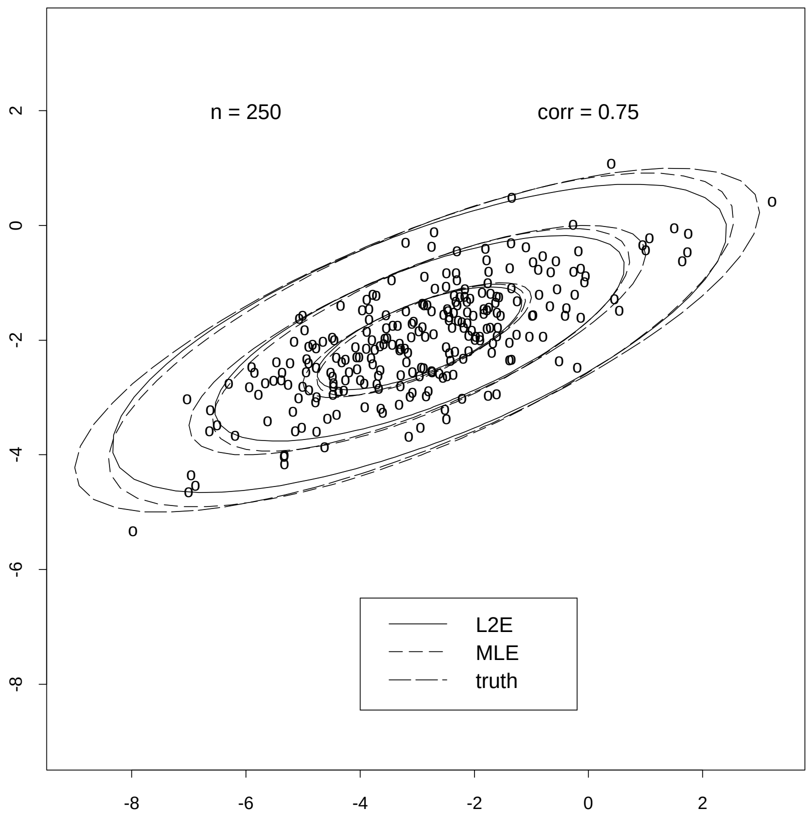

Figure 2: Fitted bivariate normal (5 parameters) using L_2E. The estimated correlation coefficients by L_2E and MLE were 0.697 and 0.725 versus the true 0.750.

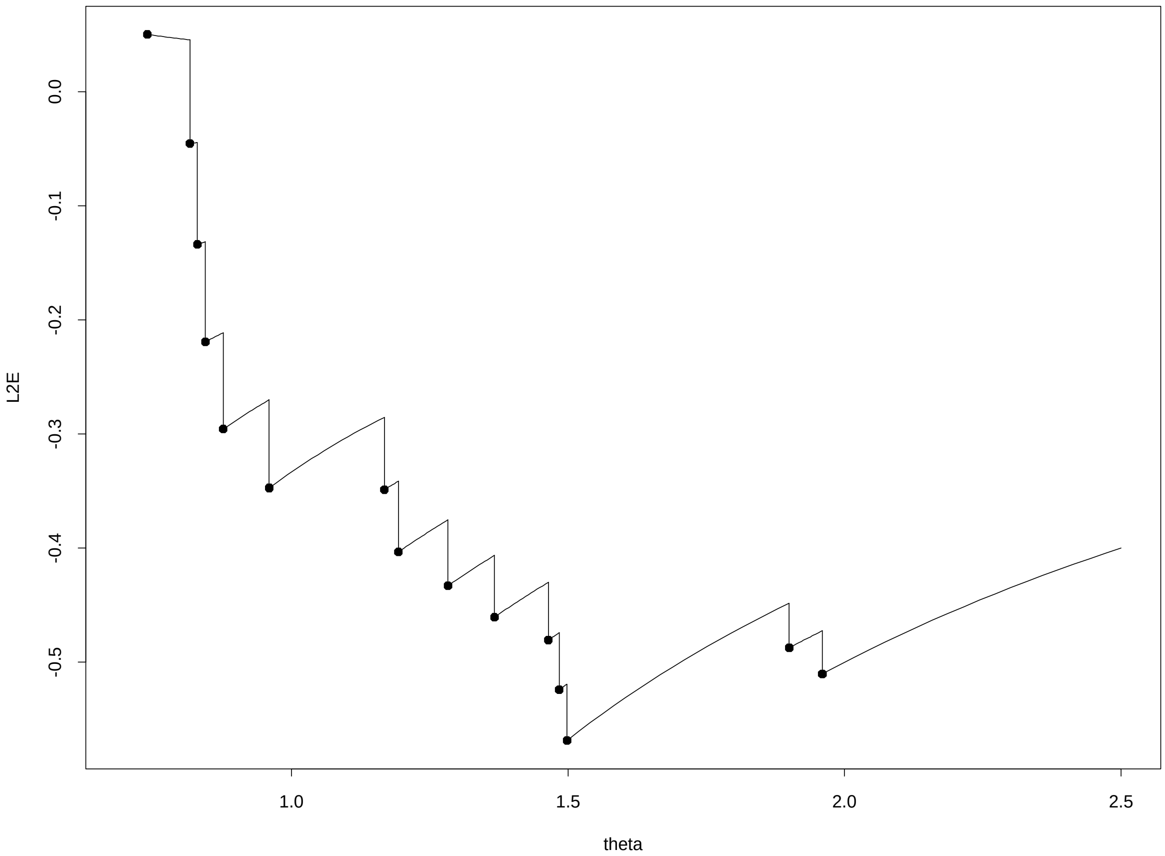

Figure 3: The L_2E function when $f(x|\theta)$ is $U(0,\theta)$ for a sample of 25 points from $U(0,1.5)$ with two contaminated points added at 1.90 and 1.96. Notice that $\hat{\theta}_{L_2E} = 1.498$ in this case.

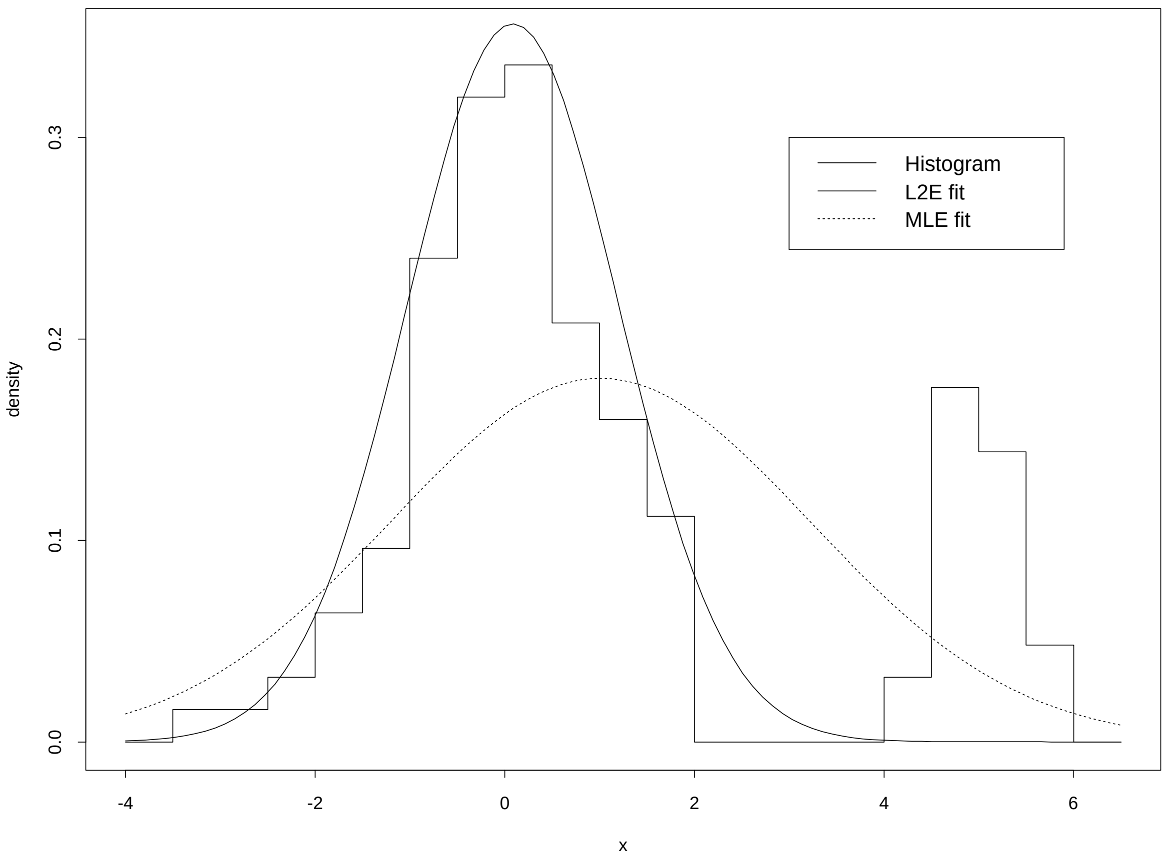

Figure 4: The parametric L_2E result with $f(x|\boldsymbol{\theta}) = N(\mu, \sigma^2)$ for the $n = 125$ points from the mixture $0.8N(0,1) + 0.2N(5,1)$ shown in Figure 1. The *MLE* is displayed as well.

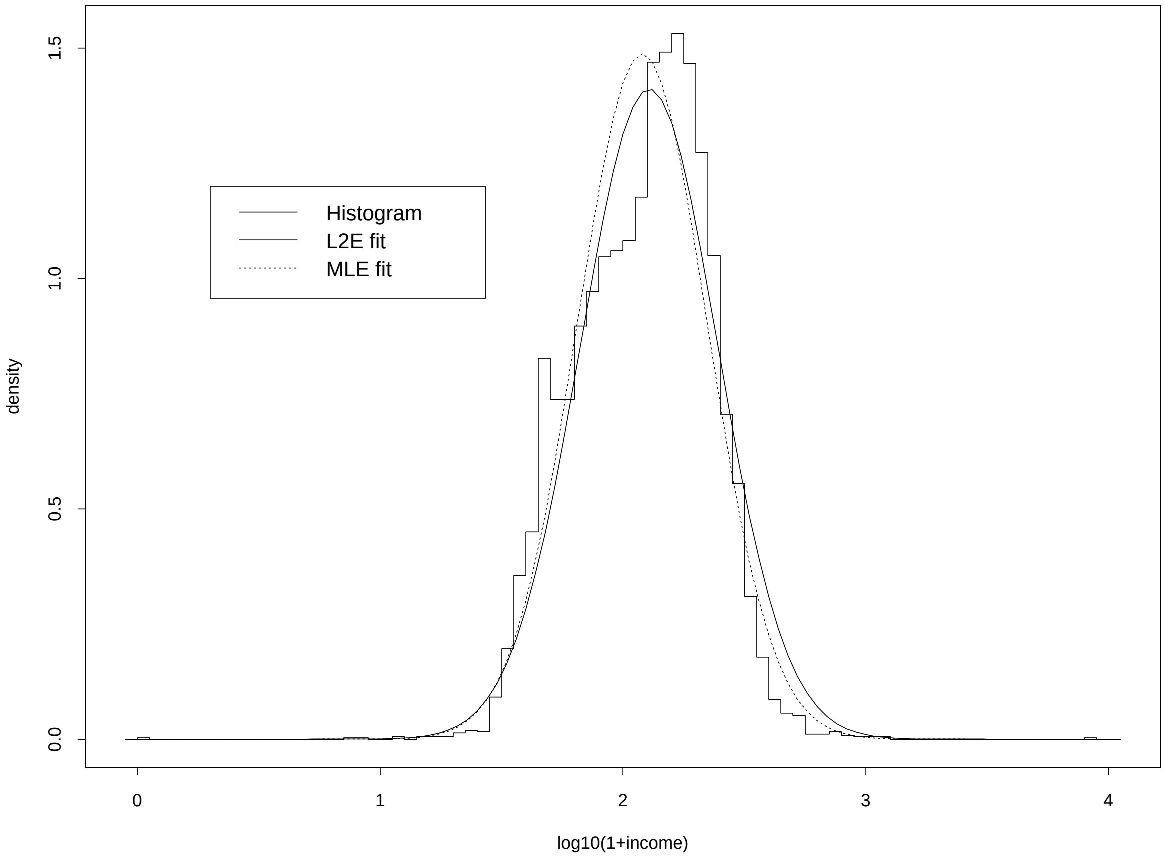

Figure 5: L_2E and MLE Gaussian estimates of the British household income data.

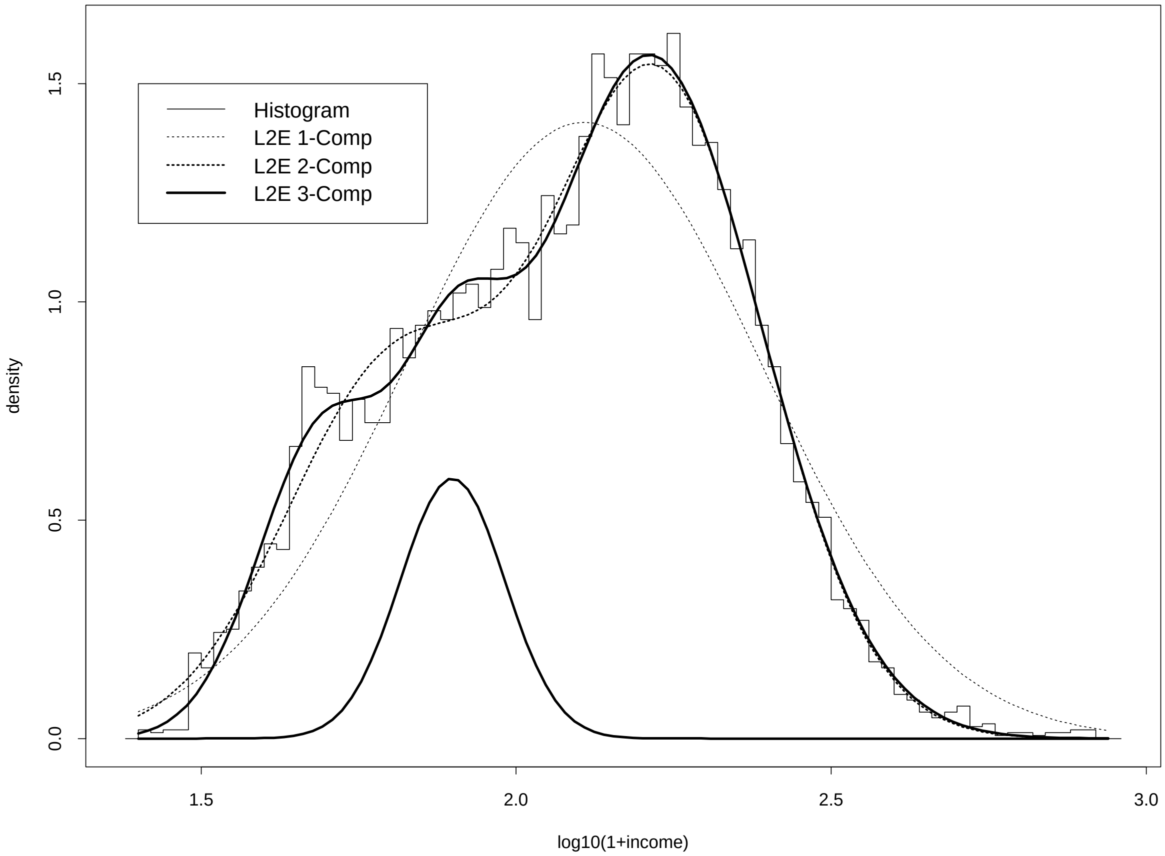

Figure 6: The one-, two-, and three-component L_2E Gaussian estimates of the British household income data. The middle component of the three-component fit is also shown. The other components are apparent from the left and right edges. (Note the expanded horizontal scale.)

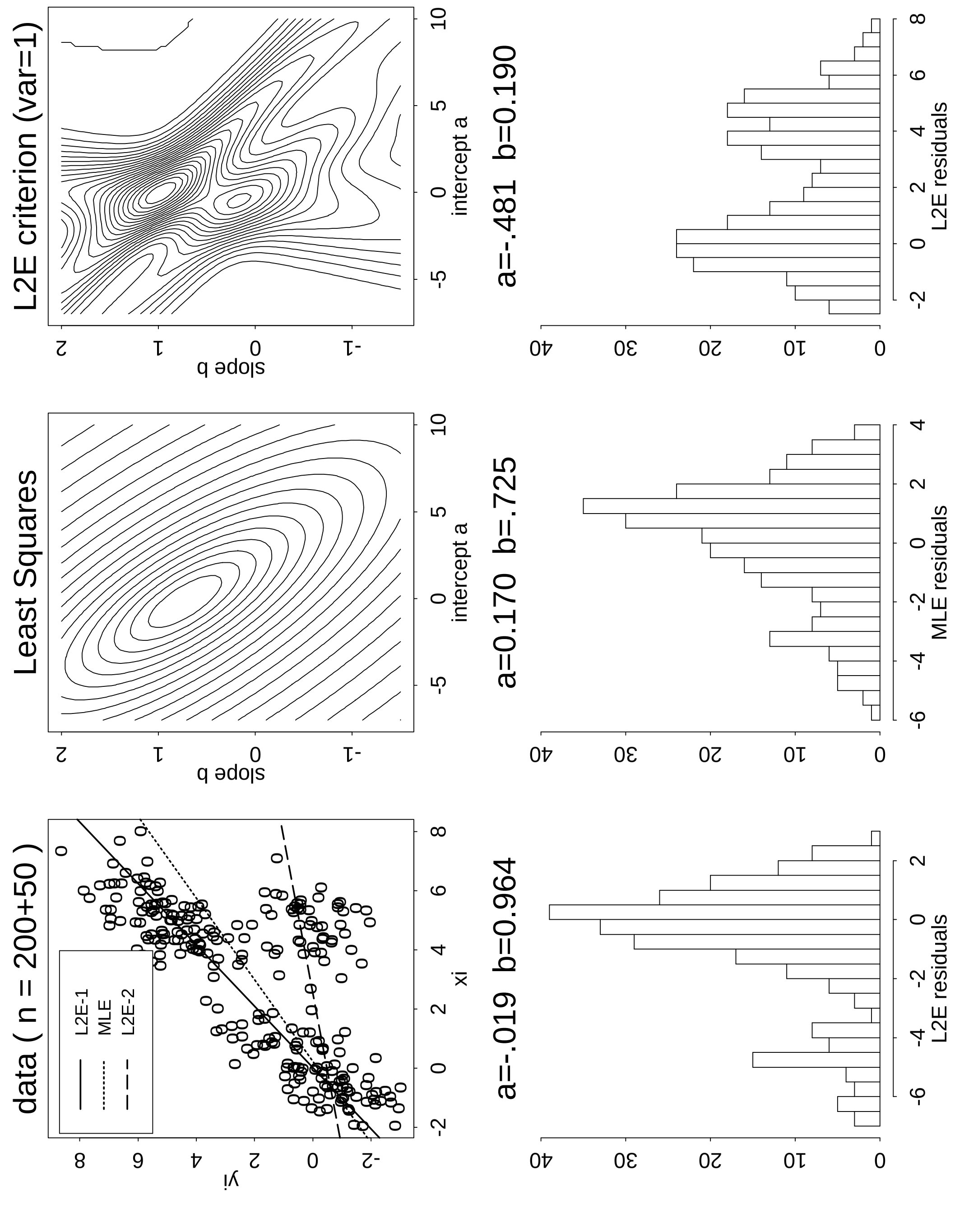

Figure 7: *MLE* and L_2E regression criteria with data and estimates shown in the upper left frame. Residual plots for the single *MLE* curve and the two L_2E curves are shown in the second row. Note the location of the origin ($\epsilon = 0$) in each.

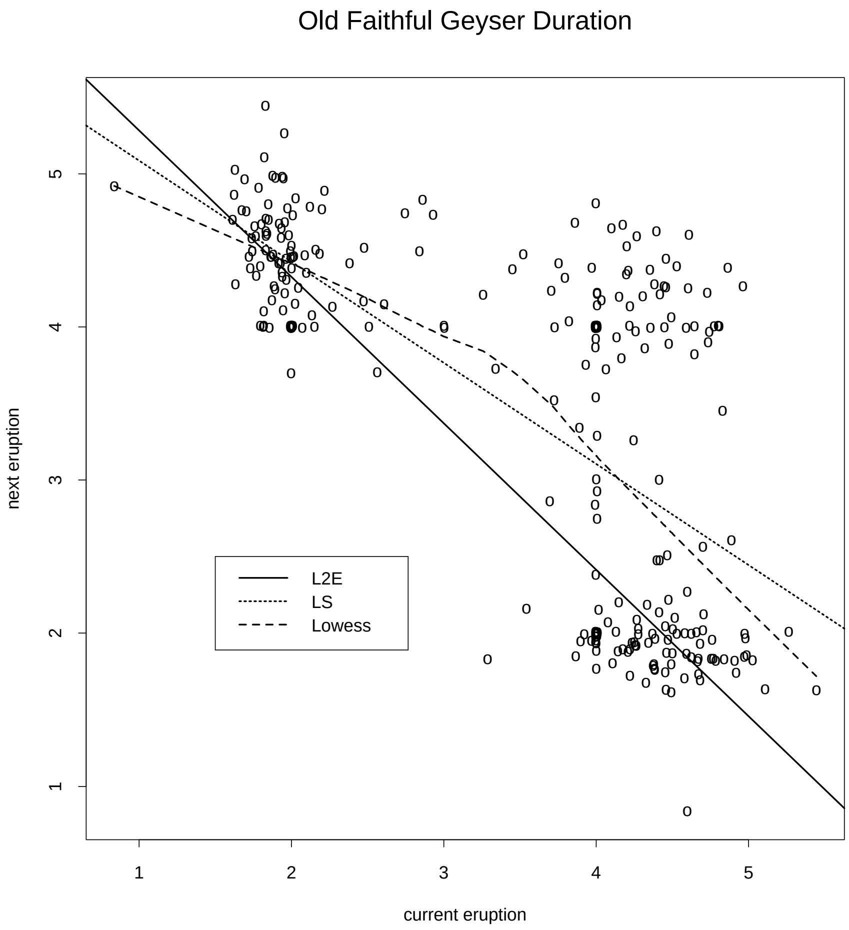

Figure 8: The least squares, L_2E, and lowess estimators for the lagged Old Faithful Geyser duration data.