Fitting Mixtures of Regression Models by L2E

David W. Scott * William F. Szewczyk †

July 18, 2000

Abstract

The generality afforded by modeling the density of a random sample by a mixture of K normals may be extended to regression in a natural manner. Each of K regression models appears unlabeled in the data with an unknown frequency. The statistical challenge is to simultaneously estimate the parameters of each regression and their frequency.

In this paper we demonstrate how to accomplish this task using an integrated squared error criterion. Minimum distance criteria are inherently robust and generally fall into the category of the well-studied M-estimators. M-estimators modify maximum-likelihood procedures by combining a robust estimate of scale with a choice of influence functional to control the effects of outliers and other "influential" points. What distinguishes the use of the ISE criterion is that by making an explicit distributional assumption for the errors, the robust scale estimate may be obtained simultaneously with the model parameters. Furthermore, the influence functional is implicitly (rather than explicitly) defined.

The $\mathrm{L_2E}$ method for estimating a mixture of normals may be extended to estimate a mixture of regressions (MORE), each with normal error. The algorithm for fitting the MORE model is explicitly given and illustrated.

Keywords: Integrated squared error, robust estimation.

1. Introduction

Given a set of data $\{x_1, \ldots, x_n\}$ and a parametric model, $f(x|\theta)$, an estimator of $\hat{\theta}$ may be obtained by maximum likelihood or by a robust M-estimation procedure. A particular M-estimator is afforded by choosing $\hat{\theta}$ as the minimizer of the $\mathrm{L_2E}$ criterion

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

*David W. Scott (*scottdw@rice.edu*) is Professor of Statistics, MS-138, Rice University, PO Box 1892 Houston, TX 77005-1892. This research was supported in part by the National Science Foundation grant #99-71797 and the National Security Agency. This work was completed while the author was on sabbatical at the National Security Agency.

†William F. Szewczyk (*wfszewc@afterlife.ncsc.mil*) is a Senior Mathematician in the Mathematics Research Office, R&E Building Suite 6515, National Security Agency, Fort George G. Meade, MD 20755-6515.

which is a fully data-based attempt to minimize the integrated squared error (ISE) or L_2-error ($\mathrm{L_2E}$) criterion

$$\mathrm{ISE}(\theta) = \int \left[f(x|\theta) - g(x)\right]^2 dx, \tag{2}$$

where $g(x)$ is the true but unknown density of the data; see Scott (1998), Basu, Harris, Hjort, and Jones (1998), Hjort (1994), and Terrell (1990). The ISE is the sum of the three terms, $\int f(x|\theta)^2 dx$, $-2E_g f(X|\theta)$, and $\int g(x)^2 dx$. The $\mathrm{L_2E}$ criterion includes an unbiased estimator of $-2E_g f(X|\theta)$, while excluding the unknown quantity $\int g(x)^2 dx$, which is constant with respect to θ in any case.

For example, if the model for $f(x|\theta)$ is the two-parameter normal density, $\phi(x|\mu,\sigma^2)$, then

$$\mathrm{L_2E}(\mu,\sigma) = \frac{1}{2\sqrt{\pi}\sigma} - \frac{2}{n}\sum_{i=1}^{n} \phi(x_i|\mu,\sigma^2),$$

which must be minimized numerically using, for example, the *nlmin* procedure in Splus. The asymptotic normality of $(\hat{\mu},\hat{\sigma})$ follows from the M-estimation theory.

Virtually any square-integrable model $f(x|\theta)$ may be fitted by $\mathrm{L_2E}$, but particularly in the multivariate case ($\mathbf{x} \in \Re^d$ and $\boldsymbol{\theta} \in \Re^p$), the computational feasibility is greatly enhanced if the term $\int f(\mathbf{x}|\boldsymbol{\theta})^2 d\mathbf{x}$ can be evaluated in closed form. While excellent numerical integration routines exist, the level of accuracy demanded as part of a quasi-Newton minimization algorithm severely limits both the complexity of $f(\mathbf{x}|\boldsymbol{\theta})$ and the dimensions of $\mathbf{x}$ and $\boldsymbol{\theta}$. These remarks apply even to simple models such as mixtures when using Hellinger distance to find the closest model to a kernel estimate (Beran 1977) or the minimum divergence criterion of Basu, Harris, Hjort, and Jones (1998), both of which require the use of numerical integration more often than $\mathrm{L_2E}$.

2. Fitting Mixtures of Normals

When the explicit parametric form of $f(x|\theta)$ is unknown, a nonparametric approach such as the normal kernel estimator may be employed:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \phi(x|x_i,h^2)\,,$$

where h is the smoothing parameter; see, for example, Scott (1992), Wand and Jones (1995), or Simonoff (1996). A somewhat simpler and often more interpretable approach is the mixture of K normals (MON) model

$$\hat{f}_K(x) = \sum_{i=1}^{K} w_k\, \phi(x|\mu_k,\sigma_k^2)\,, \tag{3}$$

where $\sum w_k = 1$; see Titterington, Smith, and Makov (1985).

Estimating the parameters of a MON model can be difficult, as singular solutions exist. But given reasonably good initial estimates, MLE values may be obtained by the EM algorithm (Dempster, Laird, and Rubin 1977). Alternatively, using the identity

$$\int \phi(x|\mu_1, \sigma_1^2)\, \phi(x|\mu_2, \sigma_2^2)\, dx = \phi(0|\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

in Equation (1), L$_2$E parameters may be obtained by minimizing

$$\sum_{k=1}^{K}\sum_{\ell=1}^{K} w_k w_\ell \phi(0|\mu_k - \mu_\ell, \sigma_k^2 + \sigma_\ell^2) - \frac{2}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w_k \phi(x_i|\mu_k, \sigma_k^2).$$

This approach has been introduced and evaluated by Scott (1998), Scott (1999). and Wojciechowski and Scott (1999).

3. Parametric Regression by L2E

Given data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, a parametric regression model

$$y_i = m_{\boldsymbol{\beta}}(\mathbf{x}_i) + \epsilon_i\,, \tag{4}$$

the parameters $\boldsymbol{\beta}$ may be fit by (nonlinear) least squares. If, in addition, a parametric model for the noise density $f(\epsilon|\boldsymbol{\theta})$ is specified, then the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ may be estimated by L$_2$E:

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) = \arg\min_{(\boldsymbol{\beta},\boldsymbol{\theta})} \int f(\epsilon|\boldsymbol{\theta})^2 d\epsilon - \frac{2}{n}\sum_{i=1}^{n} f(y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i)|\boldsymbol{\theta})\,, \tag{5}$$

since $\epsilon_i = y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i)$. If the error distribution also depends on $\mathbf{x}$, then the first term may be replaced by

$$\frac{1}{n}\sum_{i=1}^{n} \int f(\epsilon_i|\boldsymbol{\theta}, \mathbf{x}_i)^2 d\epsilon_i \tag{6}$$

see Basu, Harris, Hjort, and Jones (1998). The performance and robust behavior of the L$_2$E parameters when the model $m_{\boldsymbol{\beta}}(x)$ is linear has been examined by Scott (1998).

4. Mixtures of Regressions

Consider the simulated data in Figure 1. The 300 data points are a 1:2 mixture of two regression lines: $y = .25 + \epsilon_1$ and $y = .875 + .5x - .5x^2 + \epsilon_2$, where $\epsilon_i \sim N(0, \sigma_i^2)$, $\sigma_1 = .15$, $\sigma_2 = .05$, and $X \sim U(-.5, .5)$. If primary interest is in the quadratic portion of the data with the linear portion thought of as some contamination, then L$_2$E has a reasonable chance of providing good estimates by limiting the influence of the "noise." In fact, from all reasonable starting values, the same quadratic fit is obtained; see Figure 1a. The quadratic parameter vector $\hat{\boldsymbol{\beta}} = c(.873, .494, -.490)^T$ and $\hat{\sigma_2} = .0705$, which are all quite accurate save the 40% inflation of σ_2. When the linear equation is fitted (with 1000 random starts), the solutions are all identical; see Figure 1b, where $\hat{\boldsymbol{\beta}} = c(.841, .466)^T$ and $\sigma_1 = .0811$. The $\mathrm{L_2E}$ solution tends to converge to the regression line with more points and smaller variance, depending on (1) the degree of overlap of the two regressions; (2) the relative sample sizes; and (3) the ratio of σ_1/σ_2. With well-separated regressions, two local $\mathrm{L_2E}$ solutions may exist, but no easy rule may be cited (Scott 1998).

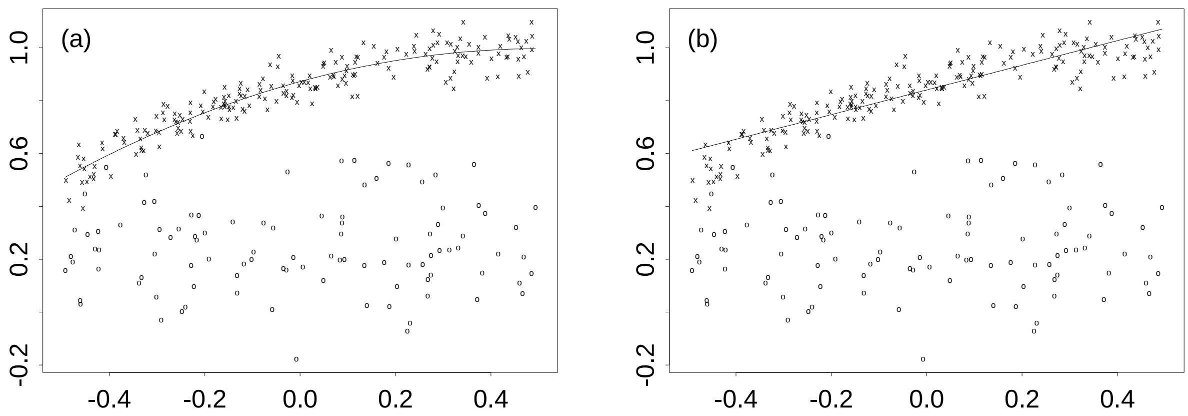

Figure 1: Best quadratic (left frame) and linear (right frame) $\mathrm{L_2E}$ fits to a simulated two-regression mixture dataset.

Yet in many modern powerful data analytic settings (give references), the proper model of the data is as a mixture (combination or overlay) of K parametric regressions. To be specific, each data point (x_i, y_i) comes, with probability w_k, from the k-th regression model

$$y = m_{\boldsymbol{\beta}_k}(\mathbf{x}) + \epsilon_k \,, \tag{7}$$

for $k = 1, \ldots, K$. If the data are labeled according to k, then the $\mathrm{L_2E}$ regression procedure may be applied K times. Here we assume no labeling is available, and that the probabilities $w_1, \ldots, w_K$ are also unknown.

Let us assume $\epsilon_k \sim N(0, \sigma_k^2)$, as the generalization to other densities is straightforward. If several values of y are available for the same $\mathbf{x}$, then the y's could be modeled as a mixture and the approach in Section 2 would apply. In fact, this observations indicates how we should proceed by recasting the mixture of regressions model as a mixture of normals and using formula (6). Model (7) and the assumption $\epsilon_k \sim N(0, \sigma_k^2)$ may be stated more concisely as

$$Y|\mathbf{x} \sim N(m_{\boldsymbol{\beta}_k}(\mathbf{x}), \sigma_k^2) \qquad \text{w/prob } w_k$$

or

$$Y|\mathbf{x} \sim \sum_{k=1}^{K} w_k \, N(m_{\boldsymbol{\beta}_k}(\mathbf{x}), \sigma_k^2) \qquad \text{w/prob } w_k \,. \tag{8}$$

The parameters $\{(w_k, \boldsymbol{\beta}_k, \sigma_k),\ k = 1, \ldots, K\}$ may be estimated by minimizing

$$\frac{1}{n} \sum_{i=1}^{n} \int f(y_i|\mathbf{x}_i)^2 dy_i - \frac{2}{n} \sum_{i=1}^{n} f(y_i|\mathbf{x}_i) \,. \tag{9}$$

In practice, simple transformations of the parameters are employed so that a standard unconstrained numerical quasi-Newton algorithm will suffice.

The computational feasibility rests on the choice of the parametric regression models and the computability of $\int f(y|\mathbf{x})^2 dy$. For our example, suppose

$$\begin{aligned} m_{\boldsymbol{\beta}_1}(x) &= a_1 + b_1 x \\ m_{\boldsymbol{\beta}_2}(x) &= a_2 + b_2 x + c_2 x^2 \end{aligned}$$

and $\sigma_1 \neq \sigma_2$. Then for each x_i,

$$f(y_i|x_i) = w_1 \phi(y_i | m_{\boldsymbol{\beta}_1}(x_i), \sigma_1^2) + w_2 \phi(y_i | m_{\boldsymbol{\beta}_2}(x_i), \sigma_2^2)$$

and

$$\int f(y_i|x_i)^2 dy = \frac{w_1^2}{2\sqrt{\pi}\sigma_1} + \frac{w_2^2}{2\sqrt{\pi}\sigma_2} + 2 w_1 w_2 \phi(0 | m_{\boldsymbol{\beta}_1}(x_i) - m_{\boldsymbol{\beta}_2}(x_i), \sigma_1^2 + \sigma_2^2).$$

For our example, good starting guesses yield the following estimates: $\hat{\boldsymbol{w}} = c(.316, .674)$, $\hat{\boldsymbol{\sigma}} = c(.1450, .0514)$, $\hat{\boldsymbol{\beta}}_1 = (.225, -.023)$ and $\hat{\boldsymbol{\beta}}_2 = (.871, .490, . - .446)$, which are all quite good. The data and estimates are displayed in Figure 2(a).

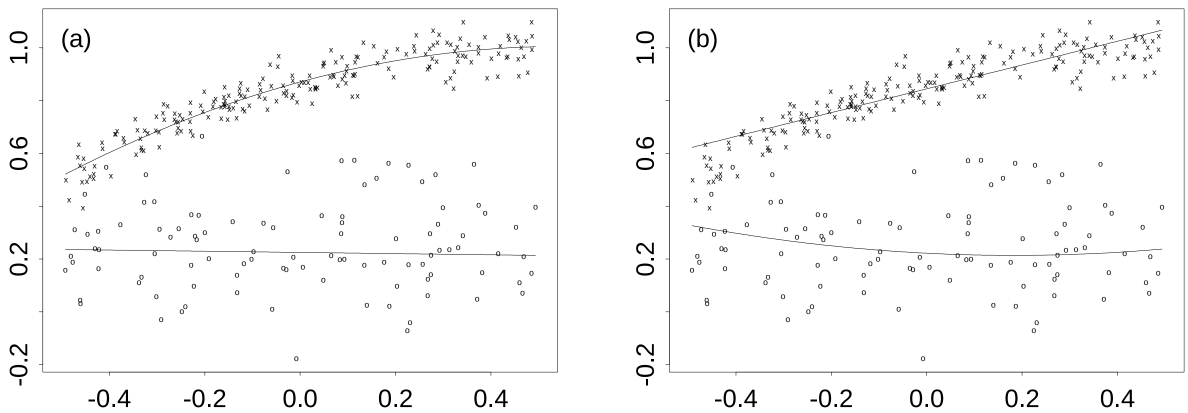

Figure 2: (a) Best mixture of linear and quadratic L$_2$E fits to the regression mixture dataset. (b) Specifications reversed, giving a local optimum L$_2$E fit.

When the initial guesses have the correct polynomial degrees switched, the result is as in Figure 2(b). The estimates are: $\hat{\boldsymbol{w}} = c(.628, .347)$, $\hat{\boldsymbol{\sigma}} = c(.0519, .1606)$, $\hat{\boldsymbol{\beta}}_1 = (.845, .450)$ and $\hat{\boldsymbol{\beta}}_2 = (.222, -.090, .248)$. The L$_2$E score is -2.701 and -2.375 for these models, respectively. Thus the correct model has a better (more negative) score. When 1600 bootstrap samples were fit with the correct model, only 59 had a score greater than -2.375, for an empirical p-value of 3.7%; see Figure 3.

The ability of L$_2$E to handle a misspecified model can be illustrated by adding 100 random points on $(-.75, .75) \times (-.25, 1.25)$ as background noise to our data. Using a good

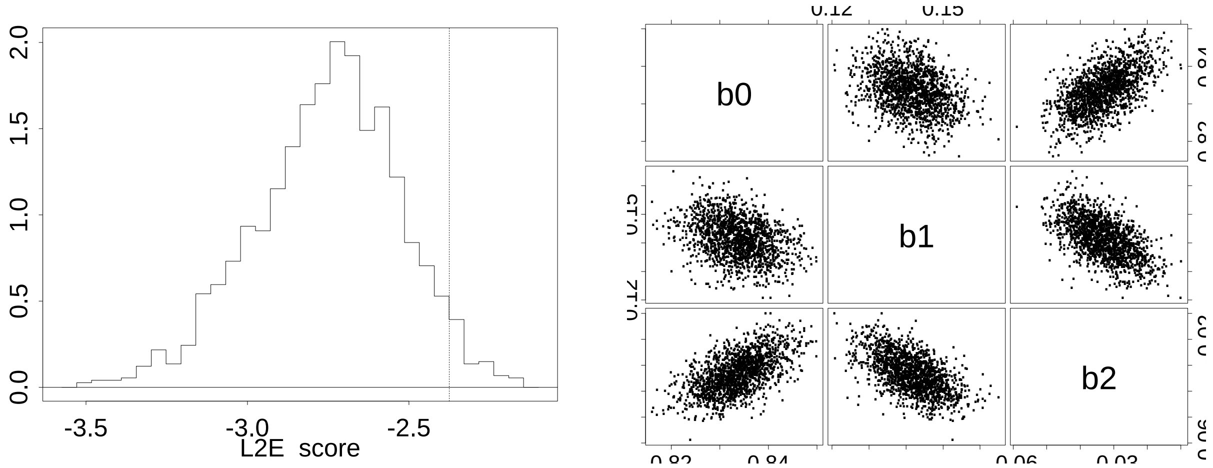

Figure 3: (a) Histogram of the L_2E score of 1600 bootstrap samples. The verticle dotted line coresponds to a score of -2.375. (b) Scatterplot of the quadratic bootstrap fits expressed in terms of an orthonormal polynomial basis on $[-.5, .5]$.

starting value, we obtain the result in Figure 4(a). The estimates are $\hat{\boldsymbol{w}} = c(.38, .54)$, $\hat{\boldsymbol{\sigma}} = c(.181, .053)$, $\hat{\boldsymbol{\beta}}_1 = (.22, .10)$ and $\hat{\boldsymbol{\beta}}_2 = (.87, .51, -.40)$. The L_2E score is -1.83. The fit to the quadratic model is quite accurate. Note that 50% of the points are from this model, but some of the background points also fall there by chance. A greater fraction of the background noise is mixed with the linear model. Only 7% of the data are estimated to not belong to either of the two regression models. By way of contrast, if the standardized residuals from the quadratic fit in Figure 1 are computed, 96 of them are less than -3. Recall that 100 of the 300 points were from the linear model.

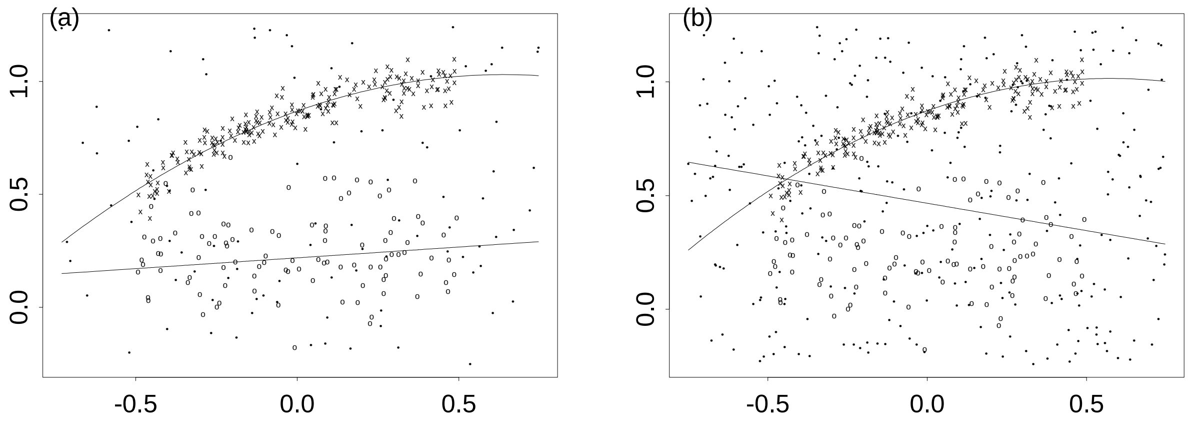

Figure 4: L_2E fits with (a) 100 background noise points added; (b) 300 noise points added.

This conclusion is strengthened by observing the estimates when 300 background points are mixed with the 300 original points; see Figure 4(b). The estimates are $\hat{\boldsymbol{w}} = c(.73, .30)$, $\hat{\boldsymbol{\sigma}} = c(.481, .045)$, $\hat{\boldsymbol{\beta}}_1 = (.47, -.24)$ and $\hat{\boldsymbol{\beta}}_2 = (.87, .49, -.43)$. The L_2E score is -1.14.

ADD REAL EXAMPLE??? COMPARE TO MLE??? DO SOME BOOTSTRAP ESTIMATES TO GET STANDARD ERRORS??? SHOW RESIDUAL HISTOGRAMS???

5. Discussion

The simplicity of performing M-estimation with L$_2$E is apparent. In particular, the algorithm eliminates the need to make prior estimates of scale and to choose an influence function. By making an explicit specification for the noise distribution, normal for example, such choices are automatically incorporated into the numerical estimation and solution.

Early Technometrics work used extra components to model the outliers. Recent EM software assembled by Rolf Turner.

References

Basu, A., I. R. Harris, N. L. Hjort, and M. C. Jones (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.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–37.

Hjort, N. L. (1994). Minimum L2 and robust Kullback-Leibler estimation. In P. Lackout and J. A. V. sek (Eds.), *Proceedings of the 12th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, Prague: Academy of Sciences of the Czech Republic, pp. 102–105.

Scott, D. W. (1992). *Multivariate Density Estimation.* New York: John Wiley & Sons, Inc.

Scott, D. W. (1998). Parametric modeling by minimum L_2 error. Technical Report 98-3, Department of Statistics, Rice University.

Scott, D. W. (1999). Remarks on fitting and interpreting mixture models. In K. Berk and M. Pourahmadi (Eds.), *Computing and Statistics*, Proceedings of the 31st Symposium on the Interface, pp. 104–109.

Simonoff, J. S. (1996). *Smoothing methods in statistics.* New York: Springer-Verlag.

Terrell, G. R. (1990). Linear density estimates. In *Proceedings of the Statistical Computing Section*, American Statistical Association, pp. 297–302.

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

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

Wojciechowski, W. C. and D. W. Scott (1999). Robust location estimation with l2 distance. In K. Berk and M. Pourahmadi (Eds.), *Computing Science and Statistics*, Proceedings of the 31st Symposium on the Interface, pp. 292–295.