Remarks on Fitting and Interpreting Mixture Models

David W. Scott
Department of Statistics, MS–138
Rice University
Houston, TX 77005

Abstract

The problem of fitting mixtures has attracted much attention since the work of Pearson. Mixture models can be a parsimonious representation for data. Mixture parameters are highly interpretable. Maximum likelihood estimates are available through the EM algorithm, which although simple is often rather slow to converge. Furthermore, outliers can significantly affect the quality and interpretability of a mixture model. This paper describes a new algorithm for fitting of mixtures based on an integrated squared error criterion. As a member of the family of minimum distance estimators, the integrated squared error criterion is by nature robust and, hence, less influenced by outliers. Finally, integrated squared error is the most widely applied criterion in nonparametric modeling, and thus we bridge a gap between parametric and nonparametrics methodology by applying the criterion in the parametric framework.

1 Introduction

The mixture model is widely applied due to its ease of interpretation. A model with K Gaussian components may be written as

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

where the weights, $\{w_k\}$, sum to one. Parameters of the model are easily interpreted in exploratory analysis by viewing each fitted component as a distinct cluster in the data. This view works well if K is not too large, and the components are somewhat distinct.

Mixture densities lie somewhere between parametric and nonparametric models. Indeed, a Gaussian kernel estimator,

$$\hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^{n} \phi\left((x - x_i)/h\right),$$

is a particular mixture model with $w_i = 1/n$, $\mu_i = x_i$, and $\sigma_i = h$, the smoothing parameter. Of course, the value $K = n$ is not a parsimonious choice, and thus the individual components are not interpretable.

Nonparametric and parametric density models were once thought to be totally different, but in recent years, local likelihood modeling has blurred the distinction. Densities estimated by local likelihood are modeled by local polynomials and are fully nonparametric. However, local likelihood retains the usual parametric criterion of maximum likelihood, whereas the usual goodness-of-fit criterion for nonparametric estimates is integrated squared error (*ISE*) defined as

$$ISE = \int \left[\hat{f}(x) - f(x)\right]^2 dx.$$

ISE is a member of the minimum distance family of estimation criteria. Consideration of *ISE* usually leads to the mean *ISE* or *MISE*, and finally to the asymptotic *MISE* or *AMISE*. Smoothing parameters (such as the histogram bin width or kernel bandwidth) may be estimated by cross-validation or plug-in approaches based upon estimates of these criteria.

In the parametric world, maximum-likelihood is not a minimum distance criterion but rather an example of a minimum divergence criterion. Parametric minimum distance criteria do exist, for example, Hellinger distance, which measures the L_2 distance between the square roots of the density functions (Beran, 1977). Curiously, Beran suggested that the parameters could be estimated by minimizing the distance to a particular kernel estimate! Minimum distance estimators are well-known to possess general robustness properties (Donoho and Liu, 1988).

The idea of using *ISE* or *MISE* for parametric estimates was discussed by Scott (1992). However, those estimates are based upon perturbations of maximum likelihood estimates. In other words, no fully data-based algorithm with the generality of MLE was available.

In the following sections, we introduce a data-based parametric version of the *ISE* criterion. We consider specifically its application to the classical mixture density estimation problem. We demonstrate how *subsets* of the components may be estimated separately, and provide some historical background for these ideas.

2 L_2 Estimators

Consider a parametric model, $f(x|\theta)$ for the unknown density, $f(x)$. The true density may be in the same parametric family, in which case $f(x) = f(x|\theta_0)$. However, we shall also be interested in L_2 approximations when $f(x)$ is not exactly a member of the estimating family. An algorithm or estimator which attempts to minimize integrated squared error or L_2 error shall be denoted by *L2E*. The class of algorithms will be referred to as *L2E* methods.

Thus we seek a data-based algorithm that solves the optimization problem

$$\begin{aligned}\hat{\theta}_{L2E} &= \arg\min_\theta \int [f(x;\theta) - f(x)]^2 dx \\ &= \arg\min_\theta \left[\int f(x;\theta)^2 dx - 2\int f(x;\theta) f(x)\, dx \right],\end{aligned}$$

since the remaining integral, $\int f(x)^2\, dx$, does not depend on θ and does not enter in to the determination of the minimizing value of θ. The first integral may be evaluated exactly (or numerically) for each posited value for θ. However, the second integral involves the unknown density, $f(x)$, and must be estimated. In a similar but nonparametric setting, Rudemo (1982) observed that

$$\int f(x;\theta) f(x)\, dx = E_f[f(X;\theta)],$$

which only requires an estimate of the mean height of the density estimate for each posited value of θ. An unbiased estimator is readily available in the form of $n^{-1}\sum f(x_i|\theta)$. Thus we propose to investigate the data-based parametric estimator based upon a level shift of the integrated squared error criterion:

$$\hat{\theta}_{L2E} = \arg\min_\theta \left[\int f(x;\theta)^2 dx - \frac{2}{n}\sum_{i=1}^{n} f(x_i;\theta) \right]. \quad (1)$$

The minimizing value of θ is typically found by using a numeric optimization routine. In all our examples below, we utilized the routine *nlmin* found in the Splus package. This routine only requires programming a function to evaluate Equation (1), given a random sample and particular value of θ supplied by *nlmin*. The general approach with *nlmin* works whether θ is a scalar parameter or a vector of parameters, as long as the density function is differentiable.

For example, if we model our data by the two-parameter Normal density, $\phi(x|\mu,\sigma)$, then we simply note that

$$\int \phi(x;\mu,\sigma)^2 dx = \frac{1}{2\sigma\sqrt{\pi}}.$$

Thus the vector of *L2E* parameters solves

$$\hat{\theta}_{L2E} = \arg\min_\theta \left[\frac{1}{2\sigma\sqrt{\pi}} - \frac{2}{n}\phi(x_i;\mu,\sigma) \right]. \quad (2)$$

In Figure 1, we plot the maximum likelihood and *L2E* density estimates for a contaminated sample of $N(0,1)$ data described in the caption of the figure. A careful study of this example hints at the potential practical utility of *L2E* methods. The *ISE* criterion rejects possible values of the parameters such as the *MLE* since $\hat{f}(x)$ is too far away from the true density curve! Thus the *L2E* parameters are picked to model the larger normal component. These estimates are not perfect: the estimated standard deviation is inflated by 10% due to the missing probability mass near $x = 4$. However, the mean is well-estimated and hints at the general excellent breakdown properties of minimum distance estimators.

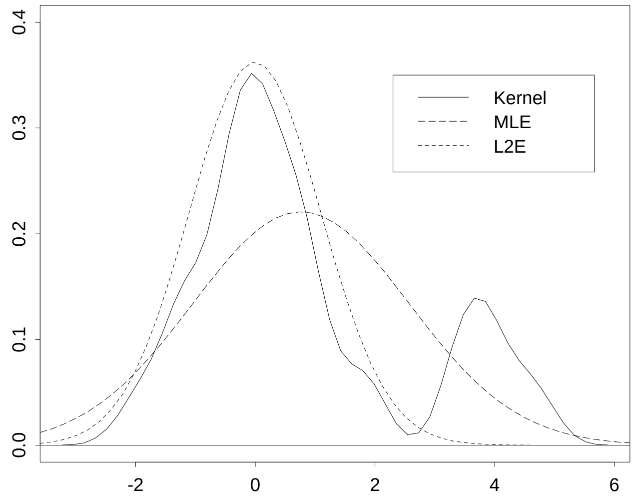

Figure 1: *A sample combining* 100 $N(0,1)$ *and* 25 $N(4, 0.5^2)$ *points, but modeled by a single* $N(\mu,\sigma^2)$ *component. The kernel estimate displays the nature of the true underlying density. Taking the parametric model to be* $N(\mu,\sigma^2)$*, the maximum likelihood estimates of* μ *and* σ *are* 0.763 *and* 1.808*, respectively, and the L2E parameters are* $-.005$ *and* 1.101*, respectively.*

We conclude this example by observing that if the value of σ is assumed known, then the *L2E* problem given in Equation (2) may be reduced to maximizing $\sum f(x_i|\theta)$. Recall that MLE maximizes the product rather than the sum. Use of the sum greatly reduces the influence of low-probability events in the optimization functional. This observation generalizes to any density supported on the entire real line, where θ is a location parameter, that is, $f(x|\theta) = f(x - \theta)$.

The asymptotic normality and efficiency of the general *L2E* parameters is described in Scott (1998) and Basu, et al. (1998). Other simulations, including parametric estimation using Hellinger distance, are described by Scott (see ftp.stat.rice.edu/pub/scottdw/l2e.r1.ps). Basu gives formulae describing the breakdown properties of these and related estimators. We defer further comment to the discussion below.

3 Normal Mixture Estimation

We turn to the particular class of parametric models given by mixtures of normals. These have been considered widely; see Titterington, et al. (1985) and McLachlan (1992). Let us begin by considering the two-component mixture model:

$$\hat{f}(x) = w\,N(\mu_1, \sigma_1^2) + (1 - w)\,N(\mu_2, \sigma_2^2). \qquad (3)$$

The *L2E* parameters are obtained by performing the 5–D numerical minimization indicated in Equation (1) over the parameter vector $\theta = (w, \mu_1, \mu_2, \sigma_1, \sigma_2)'$. In practice, w and σ_i are transformed in the usual manner to the real line so that the optimization is unconstrained.

We consider again the data used in Figure 1, but modeled with the five-parameter, two-component mixture model given in Equation (3). The *L2E* and *MLE* estimates of θ were obtained rapidly by use of routine *nlmin*, and the results are displayed in Figure 2. The estimated parameters are given by

$$\begin{aligned} \hat{\theta}_{L2E} &= (.7975, -.0046, 3.796, .8980, .4860)' \qquad (4) \\ \hat{\theta}_{MLE} &= (.8003, -.0261, 3.928, .9338, .4857)' \end{aligned}$$

We remark that both *L2E* and *MLE* are susceptible to singularities in the numerical optimization. Thus we seek the best local optimum. Problems typically arise if σ_1 is initialized near 0 and μ_1 is close to a data value. Simulation results given in Scott (1998) show that *MLE* and *L2E* estimates of the weight w are about equally variable, but *L2E* estimates of the mean and standard deviation are more variable (as predicted by theory). This extra variability is compensated by the robustness property. Unlike robustified *MLE* approaches, there are no design parameters to specify in the *L2E* approach. Real data, especially massive data sets, are prone to isolated outliers or groups of outliers, and maximum likelihood estimates are adversely affected as in Figure 1.

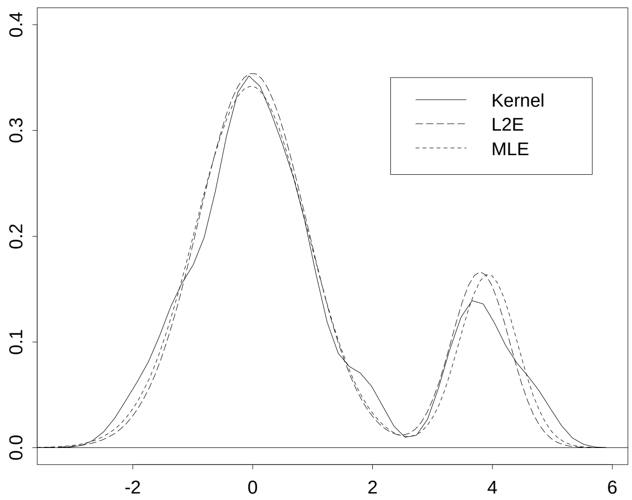

Figure 2: With the same data as in Figure 2, two-component mixtures estimated by *L2E* and *MLE* are displayed. The values for $\theta = (w, \mu_1, \mu_2, \sigma_1, \sigma_2)'$ are given in the text, but are approximately equal.

We conclude this section with a 3-component fit to the annual Buffalo snowfall data from 1910–1972. In Figure 3, we display the kernel, *L2E*, and *MLE* density estimates. The kernel estimate shows a mild outlier at $x = 25$, which occurred in the 1919/20 winter. Compare the parameter estimates with this outlier in mind. The three components are labeled 1, 2, and 3 from left-to-right. The estimated weight vectors are given by

$$\begin{aligned} \hat{w}_{L2E} &= (0.2610, 0.4619, 0.2771)' \\ \hat{w}_{MLE} &= (0.3237, 0.4055, 0.2708)'. \end{aligned}$$

The estimated mean vectors are given by

$$\begin{aligned} \hat{\mu}_{L2E} &= (52.62, 80.27, 110.33)' \\ \hat{\mu}_{MLE} &= (54.39, 81.00, 110.20)'. \end{aligned}$$

The estimated standard deviation vectors are given by

$$\begin{aligned} \hat{\sigma}_{L2E} &= (\ 8.55, 7.83, 10.36)' \\ \hat{\sigma}_{MLE} &= (12.54, 6.92,\ 9.38)'. \end{aligned}$$

A few observations are worth emphasizing. First, in spite of the outlier in the first bump, the estimated *MLE* mean is greater than the corresponding *L2E* mean. How can that be? The answer is that the outlier inflates

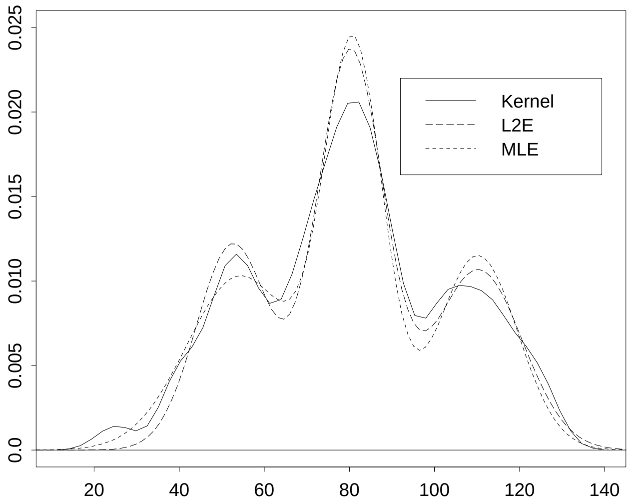

Figure 3: For the Buffalo snowfall data, three-component normal mixtures estimated by *L2E* and *MLE* are superimposed on a kernel estimate.

the *MLE* standard deviation $\hat{\sigma}_1$ by 32% compared to *L2E*. Thus, the *MLE* density estimate at the outlier value $x = 25$ is many times greater than the *L2E* level there. But the influence of the outlier in the *MLE* estimates extends beyond the first component. The inflated variance in the first component results in an increased weight there (w_1) and a decreased weight in the second component (w_2). This simple example serves to remind us how fragile the interpretability of the parameters of a mixture model are in the face of even mildly difficult data. With multiple outliers in large datasets, the problem persists. (See a forthcoming paper on income distributions in Great Britain by Scott, Simar, and Wilson.) The problem only grows in higher dimensions. We believe that the minimum distance approach attempts to find the estimates which appropriately model the largest fraction of data possible. This diagnostic behavior at the time of estimation is in sharp contrast to the usual graphical diagnostics after the estimation.

4 Partial Component Estimation

In the course of our using *L2E* to estimate mixtures with varying numbers of components, we noticed that the criterion allows for density models that are "incomplete." For example, we may take as our model,

$$\hat{f}(x) = w\,\phi(x|\mu,\sigma).$$

Re-examining the steps leading to Equation (2), we see that the modified criterion is simply

$$\hat{\theta}_{L2E} = \arg\min_{(w,\mu,\sigma)} \left[\frac{w^2}{2\sigma\sqrt{\pi}} - \frac{2\,w}{n}\phi(x_i;\mu,\sigma) \right].$$

In practice, this optimization is more likely to find singular solutions, so that multiple starting guesses are recommended.

We revisited the data in Figure 1, and found two interesting one-component solutions. These are displayed in Figure 4. The estimated parameters for the two components (which I must repeat were estimated individually and independently) are from left to right:

$$\begin{aligned}
(\hat{w}_1, \hat{\mu}_1, \hat{\sigma}_1)_{L2E} &= (0.7837, -0.0024, 0.8899) \\
(\hat{w}_2, \hat{\mu}_2, \hat{\sigma}_2)_{L2E} &= (0.1947, \quad 3.7853, 0.4728)
\end{aligned}$$

Observe that the two estimated weights do not add to 1, for example. However, comparing these estimates with the two-component mixture results in Equation (4) shows very close agreement. The application to clustering seems clear and promising.

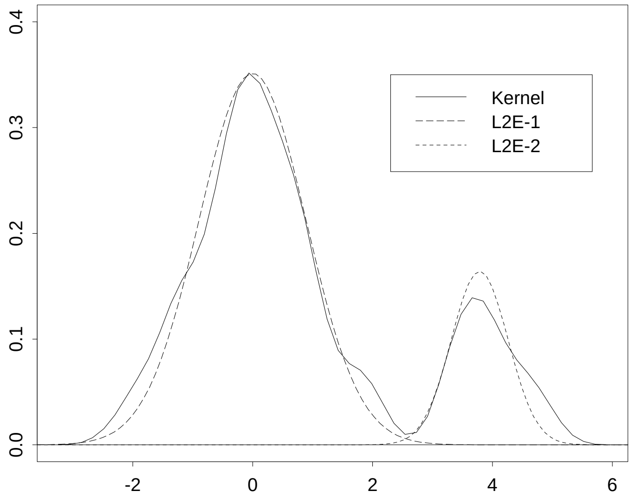

Figure 4: For the data in Figure 1, partial component normal fits were computed for different starting values. The two interesting solutions are displayed, together with the kernel estimate.

5 Multivariate Mixtures

When $f(x|\theta)$ is a mixture of multivariate normals,

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

L2E is feasible because of the closed form expression

$$\int \phi(x|\mu_1, \Sigma_1)\phi(x|\mu_2, \Sigma_2) = \phi(\mu_1 - \mu_2|0, \Sigma_1 + \Sigma_2).$$

Examples are given in Scott (1998).

6 Summary and Closing Remarks

We have demonstrated a new algorithm for fitting normal mixtures. The algorithm may be shown to be robust, since all minimum distance estimators possess that trait. Of special interest is the novel ability to estimate isolated components of a mixture (or other subsets) separately, without a full mixture model. This property and behavior may be especially useful to exploit in higher dimensions.

The *L2E* approach is applicable to a wide set of statistical problems. Mixtures are of course an especially challenging class of problem. The close relationship to clustering algorithms can and should be exploitable. We hope to pursue this in detail later.

The genesis of the approach may be traced to the nonparametric work of Rudemo (1982), who attempted to use the *ISE* criterion to estimate the histogram bin width and the kernel bandwidth successfully. However, it was the innovative work of Terrell (1990) who described a penalized version of the *L2E* criterion for use in nonparametric curve estimation. After listening to George lecture on this so-called linear estimator for several summers in our Rice RIMS program, I thought to drop the penalty and apply the algorithm to parametric density estimation. I was certainly motivated by the invasion of parametric approaches into nonparametric problems, and wished to turn the tables.

I find it interesting to try to understand the deeper implications of the use of a divergence criterion such as maximum likelihood as compared to the several minimum distance criteria that now exist. In a related work, Basu et al. (1998) have devised a whole continuum of divergence estimators that begin at *MLE* and interpolate to *L2E* (and beyond). The key quantity in their data-based class is $\int \hat{f}(x)^{1+\alpha} dx$, where $\alpha \geq 0$. *MLE* and *L2E* correspond to $\alpha = 0$ and $\alpha = 1$, respectively. The authors suggest that an appropriate range in practice for α is in the interval $[0, 1/4]$, as the parameters in this range are fairly robust and the loss of efficiency more modest. However, we note that for the mixture problem, this key integral cannot be computed in closed form except at values 0 and 1. We have examined using numerical integration for the other cases, and indeed this can work for a few mixtures in one dimension. However, the lack of closed form basically renders values of α other than 0 or 1 useless in practice. Similar computational burdens exist with Hellinger distance; see Woodward, et al. (1995). Hjort (1994) described the *L2E* briefly in a paper focusing on information criteria. The asymptotics given in this manuscript were folded into the later Basu, et al. paper. The relative inefficiency of squared error was the primary focus. The computational and partial component benefits provide strong counterpoints.

A number of more complicated models that contain mixture models deep within have been posed. These can rather successfully be handled in the *L2E* framework due to its simplicity and fitting properties.

Acknowledgments

This research was performed in part while the author was a visiting scholar at the Institute of Statistics, Louvain-la-Neuve in Belgium during 1998. The author wishes to acknowledge the gracious hospitality of Leopold Simar, Irene Gijbels, and Alois Kneip. This research was also supported in part by the National Science Foundation grant DMS 96-26187. A fuller discussion of the *L2E* approach and other applications are available on the web at *ftp://ftp.stat.rice.edu/pub/scottdw/l2e.r1.ps.*

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.

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

Hjort, N.L. (1994). "Minimum L2 and Robust Kullback-Leibler Estimation," *Proceedings of the 12th Prague Conference.*

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*, 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. (1998), "Parametric Modeling by Minimum L_2 Error," TR 98-3, Department of Statistics, Rice University.

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.

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.