Adaptive Mixtures

Carey E. PRIEBE*

The estimation of a probability density function based on a sample $\{\zeta_i\}_{i=1}^n$ of independent identically distributed observations is essential in a wide range of applications. In particular, a sequence of estimates $\hat{\alpha}_n$ that converges in some sense to the true density α_0 can yield asymptotically optimal performance in classification and discrimination problems. In this article an estimation technique called "adaptive mixtures" is developed from the related methods of kernel estimation and finite mixture models. Asymptotic properties of adaptive mixtures are obtained via the so-called method of sieves, yielding almost sure L_1 convergence. Monte Carlo simulations indicate the performance of the method, and an experimental study based on a typical discrimination problem is performed, indicating the scope of applicability.

KEY WORDS: Finite mixture models; Kernel estimators; Method of sieves; Nonparametric estimation; Probability density estimation.

1. INTRODUCTION AND SUMMARY

This article discusses nonparametric, or distribution-free, maximum likelihood density estimators. Modern engineering practice has exploited nonparametric density estimates in a wide variety of settings. One application of density estimation is to pattern recognition problems. Indeed, under appropriate conditions, consistent density estimates yield asymptotically optimal discriminant procedures. The need for nonparametric techniques stems from a wide range of applications in which the experimenter is unwilling to assume a parametric family for the true underlying probability density function. In these cases it is necessary to consider estimation in an infinite-dimensional parameter space A. Furthermore, recursive procedures are often required due to the nature of the application. In recursive estimation of a probability density function, the estimate based on $n+1$ observations $\{\zeta_1, \ldots, \zeta_{n+1}\}$ is a function of the $n+1$st observation ζ_{n+1} and the estimate based on the n previous observations $\{\zeta_1, \ldots, \zeta_n\}$; that is, $\hat{f}_{n+1}(x; \zeta_1, \ldots, \zeta_{n+1}) = Z(\hat{f}_n, \zeta_{n+1})$. Such a procedure obviates the need to store all the incoming observations, thus allowing for high data rates.

A consideration in nonparametric probability density function estimation is the smoothness of the estimator. Consider the standard kernel estimation approach where $\hat{f}_n(x; \zeta_1, \ldots, \zeta_n) = (nh)^{-1} \sum_{i=1}^n K((x-\zeta_i)/h)$ for certain choices of the kernel function $K(\cdot)$. The smoothness of $\hat{f}$ is directly influenced by the choice of the window-width parameter, h. Small values of h produce rough estimates with spikes at the individual observations. Conversely, a large h yields an oversmoothed estimate. Although certain basic conditions are placed on h by asymptotic considerations, the choice of the smoothing parameter is nevertheless an art, guided by rules-of-thumb, heuristic techniques such as cross-validation, or user interaction.

The study of data-driven smoothing (i.e., estimators that develop their smoothing properties stochastically based on the observations) is of significant current interest. In this article a particular adaptive nonparametric maximum likelihood technique termed "adaptive mixtures" is developed and justified. The theory of maximum likelihood estimation with data-driven smoothing is discussed from the viewpoint of the method of sieves. A simulation analysis is performed and experimental results are presented to indicate the use of adaptive mixtures in discriminant analysis.

The conclusions to be drawn from this work stem from the need for more robust estimators. Robust statistics has been defined as the study of situations in which simple parametric assumptions fail to allow for an adequate model of the data. In this context, consider the computational statistics agenda presented by Wegman (1988). To motivate the study of nontraditional probability and statistics techniques (including recursive processing and nonparametric estimation), Wegman argued that current and future problems have inherent difficulties requiring expanded capabilities. Driven primarily by modern data collection capabilities, it is no longer sufficient to make simple parametric assumptions concerning the character of the data being provided. The central limit theorem, unfortunately, is true, but often irrelevant. It is necessary to consider robust methods, such as the work presented herein, to deal with ever less-conventional data.

Given a probability density function α_0 belonging to some class $\mathcal{F}$, the adaptive mixtures procedure produces a sequence of estimators $\{\hat{\alpha}_n\}$ that is consistent. The kernel estimator is consistent for a large class of functions $\mathcal{F}'$, but the computational complexity of the method is often prohibitive. The finite mixture model approach to density estimation, on the other hand, has very appealing properties but can be consistent only when the number of terms in the model has been correctly chosen. The adaptive mixtures technique yields the appealing properties of both of these approaches: consistency for a large class of functions $\mathcal{F}$ with the low computational complexity associated with finite mixture models. The method automatically determines the number of terms in the model based on the data.

Section 2 develops the adaptive mixtures nonparametric maximum likelihood technique. The drawbacks that make neither kernel estimation nor finite mixture models an acceptable solution to the problems of computational statistics

* Carey E. Priebe is an Assistant Professor, Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD 21218. This work was partially supported by the Naval Surface Warfare Center's Dahlgren Division Independent Research program and by the Office of Naval Research (R&T #4424314). The author thanks Edward J. Wegman, David J. Marchette, Jeffrey L. Solka, George W. Rogers, and an anonymous referee for many helpful comments and suggestions.

are presented, and the two methods are merged to produce an estimation technique with the desirable properties of both. Section 3 outlines a theoretical analysis of adaptive mixtures using the method of sieves. The previous work of Wald (1949), Grenander (1981), Geman and Hwang (1982), and others is applied to yield asymptotic results for a particular class of sieves encompassing adaptive mixtures. Section 4 is devoted to simulation analysis of the adaptive mixtures technique. The goal is to develop a feel for the performance of the algorithm. Section 5 indicates the relationship of these results to discriminant analysis by relating the results of previous sections to an important pattern recognition measure, the probability of misclassification, and an illustrative experimental study is presented. Section 6 concludes with a discussion of the relevance of the results presented.

2. A SOLUTION PROCEDURE

2.1 Kernel Estimation

The method of kernel (or Parzen) estimation was introduced by Rosenblatt (1956) and Parzen (1962), studied by many authors, and summarized in detail by Silverman (1986) and Scott (1992). For the univariate case, $\hat{\alpha}_n(x; \zeta_1, \ldots, \zeta_n) = (nh)^{-1} \sum_{i=1}^{n} K((x - \zeta_i)/h)$, where $K(\cdot)$ is the kernel function. The asymptotic properties of kernel estimators, in particular the strong convergence in L_1 under very weak conditions as given in Devroye and Gyorfi (1985), Chapter 3, are very appealing. A probability density function estimation method with such power is, on the surface, hard to dismiss. However there are computational disadvantages to this method. To evaluate $\hat{\alpha}_n$ all n observations $\zeta_1, \ldots, \zeta_n$ must be available and n evaluations of $K(\cdot)$ are necessary. For many computational statistics problems of interest (i.e., those with very large data sets) such a constraint is unacceptable. Although there are recursive versions of the kernel estimator and even Fourier domain results, these approaches do not fully address the computational complexity issues outlined previously.

2.2 Finite Mixture Models

Assume for the moment that the true but unknown density is of the form $\alpha_0(x) = \sum_{i=1}^{N} \pi_i \phi(x; \mu_i, \sigma_i)$, where $N < \infty$ is known, the nonnegative mixing coefficients π_i sum to unity, and $\phi_i(x) = \phi(x; \mu_i, \sigma_i) = ((2\pi)^{1/2} \cdot \sigma_i)^{-1} \exp[-.5((\mu_i - x)/\sigma_i)^2]$ is the normal probability density function with mean μ_i and standard deviation σ_i. The goal is to estimate the parameter vector θ, which consists of $3N$ components; $\theta = \{\pi_1, \mu_1, \sigma_1, \ldots, \pi_N, \mu_N, \sigma_N\}$. (Actually, because $\sum_{i=1}^{N} \pi_i = 1$, θ may be reduced to a $3N - 1$ vector.) Finite mixture models have been discussed at length by Everitt and Hand (1981), Titterington, Smith, and Makov (1985), and McLachlan and Basford (1988).

A standard technique for estimating the parameter vector θ based on observations $Z_n = \{\zeta_1, \ldots, \zeta_n\}$ is to maximize the (log)likelihood $l_{Z_n}(f)$ over the family of N mixtures $\mathcal{F}$. Denote an estimate for $\alpha_0(x)$ with estimated parameter vector $\hat{\theta}$ as $\hat{\alpha}(x) = \hat{\alpha}(x; \hat{\theta})$. The iterative expectation-maximization (EM) algorithm given by Dempster, Laird, and Rubin (1977) and Redner and Walker (1984) is a method for maximum likelihood estimation of these parameters that can yield, under appropriate circumstances, strong L_1 convergence. (EM is an alternative numerical algorithm to, say, Newton–Raphson.)

To approach this estimation problem recursively requires further development, however. Following Titterington (1984), let $\mathbf{S}(x, \theta)$ denote the vector of scores. That is, for each component θ_i of the parameter vector θ, let $S_i(x, \theta_i) = (\partial/\partial\theta_i)\log(\hat{\alpha}(x; \hat{\theta}))$. Consider now the recursive update formula $\hat{\theta}_{n+1} = \hat{\theta}_n + k_n \mathbf{S}(\zeta_{n+1}; \hat{\theta}_n)$, with k_n a sequence converging to 0. This equation can be interpreted as a gradient ascent on the log-likelihood surface. With the proper choice of the sequence k_n, this stochastic approximation procedure can be made consistent. An example of this kind of approximation formula, which will be used herein, is the following set of recursive update equations for normal components from Titterington (1984) and Titterington et al. (1985, chap. 6):

$$\rho_{n+1}^{(i)} = \pi_n^{(i)} \frac{\phi_i(\zeta_{n+1})}{\hat{\alpha}_n(\zeta_{n+1})}, \tag{1}$$

$$\pi_{n+1}^{(i)} = \pi_n^{(i)} + \beta_n^{(i)}(\rho_{n+1}^{(i)} - \pi_n^{(i)}), \tag{2}$$

$$\mu_{n+1}^{(i)} = \mu_n^{(i)} + (\pi_n^{(i)})^{-1}\beta_n^{(i)}\rho_{n+1}^{(i)}(\zeta_{n+1} - \mu_n^{(i)}), \tag{3}$$

and

$$\sigma_{n+1}^{2(i)} = \sigma_n^{2(i)} + (\pi_n^{(i)})^{-1}\beta_n^{(i)}\rho_{n+1}^{(i)} \times ((\zeta_{n+1} - \mu_n^{(i)})(\zeta_{n+1} - \mu_n^{(i)}) - \sigma_n^{2(i)}), \tag{4}$$

with

$$\beta_n^{(i)} = n^{-1}. \tag{5}$$

This set of equations (1)–(5) will be called the "update rule" $\mathcal{U}_n(\zeta_{n+1}; \hat{\theta}_n)$. The superscript (i) used on the dummy variable ρ and the three parameters π, μ, and σ in these equations indicates the ith term in the mixture estimate. The idea behind this update rule is to distribute the effect of the new observation to all the terms in proportion to their respective likelihoods. The mean, variance, and mixing coefficient are then updated by this proportion. In the case of a single term, where $\rho = 1$ and $\pi = 1$, these update rules are just recursive versions of the sample mean and sample variance calculations.

Convergence results regarding the recursive update rule $\mathcal{U}(\cdot)$ have been given by Titterington (1984). Once again, this procedure can be made strongly convergent in the L_1 sense.

Even if the true probability density function α_0 is not known to be a mixture of normals, one might still wish to use the foregoing formulation to find an approximation to the density by a mixture. The kernel estimator is an extreme example of such an approximation, with $N = n$ and no maximum likelihood updating involved. Thus one could choose N "large enough," start the estimate with some initial $\hat{\theta}_0$, and then recursively update the estimate using $\mathcal{U}(\cdot)$. Assuming that the density is well approximated by such a mixture (which is the case if N is large enough) and a reasonable initial estimate is used, this procedure will result in a good estimate of the density.

To obtain consistency, very strong assumptions must be placed on the underlying density and on the initial state of the estimator. In particular, the underlying density must be a mixture of the same type as the estimator.

2.3 Adaptive Mixtures

If an approximation of the density α by a finite mixture is used, then the number of terms N and an initial estimate must be chosen. It would be helpful if an algorithm could choose N from the data in a recursive manner. The approach taken by the adaptive mixtures estimator is to recursively adapt not only the parameters, as in $\mathcal{U}(\cdot)$, but also the number of terms needed to fit the data. If the number of terms is allowed to grow indefinitely, then the requirement that the true density be a mixture of normals can be relaxed. An extreme case of this is the kernel estimator, which has been shown to be consistent under very weak conditions on the underlying density. The adaptive mixtures approach is designed to allow the number of terms to grow, but at a much slower rate than that of a kernel estimator. The adaptive mixtures approach is much less computationally and memory intensive in practice, can produce a more useful small-sample estimator, and allows general consistency results.

It is well understood that one cannot perform maximum likelihood estimation in an infinite-dimensional manifold if one attempts unconstrained maximization. The likelihood can be made arbitrarily large, for example, by taking $f(x)$ as $f(x; \zeta_1, \ldots, \zeta_n) = (n \cdot (2\pi)^{1/2} \cdot \sigma)^{-1} \sum_{i=1}^{n} \exp[-.5((x - \zeta_i)/\sigma)^2]$. In this case as $\sigma \to 0$, $l_{Z_n}(f) \to \infty$. Thus Dirac δ functions obtain, and maximum likelihood density estimation fails. Thus care must be taken to properly constrain the growth of the number of terms in the adaptive mixture model.

To extend finite mixtures to nonparametric estimation with a variable number of terms, consider using the stochastic approximation procedure

$$\hat{\theta}_{n+1} = [1 - \mathcal{P}_n(\zeta_{n+1}; \hat{\theta}_n)]\mathcal{U}_n(\zeta_{n+1}; \hat{\theta}_n) + \mathcal{P}_n(\zeta_{n+1}; \hat{\theta}_n)\mathcal{C}_n(\zeta_{n+1}; \hat{\theta}_n) \quad (6)$$

to recursively update the density. $\mathcal{P}_n(\cdot)$ represents a (possibly stochastic) create decision and takes on value 0 or 1. $\mathcal{U}_n(\cdot)$ updates the current parameters as in recursive maximum likelihood estimation, whereas $\mathcal{C}_n(\cdot)$ adds a new term to the model analogous to a kernel estimation approach.

2.3.1 Update Rule. The update function $\mathcal{U}(\cdot)$ guides a traversal of the estimate of the likelihood surface provided by the observations $\{\zeta_i\}_{i=1}^{n}$ and based on the likelihood equations. This recursive maximum likelihood technique converges to the desired resultant estimator when properly constrained. For completeness, note that an alternative iterative version of adaptive mixtures can be easily defined in which all of the data are stored and the update rule is the iterative EM algorithm.

2.3.2 Create Rule. Assuming that the system has decided to add a term ($\mathcal{P}(\cdot) = 1$), a create rule $\mathcal{C}(\cdot)$ can be derived from the fact that the kernel estimator based on $n + 1$ observations is closely related to the kernel estimator based on n observations. The differences are a new kernel, or term, centered at the newest observation ζ_{n+1}; updated proportionality constants for each term, from n^{-1} to $(n + 1)^{-1}$; and possibly different variances. This analogy is captured by the create rule $\mathcal{C}(\cdot)$ defined by equations (7)–(11):

$$\mu_{n+1}^{(N+1)} = \zeta_{n+1}, \quad (7)$$

$$\sigma_{n+1}^{(N+1)} = \sigma_n^0, \quad (8)$$

$$\pi_{n+1}^{(i)} = \pi_n^{(i)}(1 - \beta_{n+1}) \qquad (i = 1, \ldots, N), \quad (9)$$

$$\pi_{n+1}^{(N+1)} = \beta_{n+1}, \quad (10)$$

and

$$N = N + 1. \quad (11)$$

Thus the new term is centered at the newest observation and given a small mixing coefficient and an initial standard deviation σ_n^0, which may be user defined or derived from the terms in the neighborhood of the observation. All of the other mixing coefficients must be updated so that they sum to unity. Otherwise, the other terms are unaffected.

$\mathcal{C}_n(\cdot)$ adds new parameters to $\hat{\theta}$, changing the character and dimensionality of the likelihood surface. The fact that $\mathcal{P}_n(\cdot)$ depends on ζ_{n+1} implies that this change can be data driven. The create rule $\mathcal{C}_n(\cdot)$ is chosen so that the proportion and variance of the new term decrease with the number of terms.

If $\mathcal{P}_n(\zeta_{n+1}; \hat{\theta}_n) \equiv 1$ for $n \leq T$ and $\mathcal{P}_n(\zeta_{n+1}; \hat{\theta}_n) \equiv 0$ for $n > T$, then the algorithm will fit T terms to the data. Alternately, one could start with T terms chosen using some a priori knowledge and then let $\mathcal{P}_n(\zeta_{n+1}; \hat{\theta}_n) \equiv 0$ for all n. In particular, if $K(\cdot)$ is the normal distribution, then this yields a normal mixture model as described earlier. On the other hand, if $\mathcal{P}_n(\zeta_{n+1}; \hat{\theta}_n) \equiv 1$, then the algorithm always creates a new term, centered at the new data point, and the estimate then becomes

$$\hat{\alpha}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - \zeta_i}{h_i}\right).$$

This $\hat{\alpha}(x)$ is the recursive kernel estimator considered by Wolverton and Wagner (1969), Devroye (1979), and Wegman and Davies (1979). Its consistency is easily established. Thus in this extreme case of $\mathcal{P}(\cdot) \equiv 1$, the estimator (6) with $\mathcal{C}(\cdot)$ as described in equations (7)–(11) is consistent. Therefore, it is reasonable that because the update rule $\mathcal{U}(\cdot)$ is a recursive maximum likelihood estimator, it improves the estimate between the addition of new terms, and that if $\mathcal{P}(\cdot)$, the decision to add a term, is properly chosen, then (6) can be made consistent. The performance of the estimator obtained by using recursive updates, as opposed to merely always adding another term as in kernel estimation, is important. The reduction in the number of terms required in the estimate (investigated as "model complexity" in Sec. 4) yields a storage and computational advantage. Of interest here is the "reduced" kernel estimator of Fukunaga and Hayes (1989), which is an iterative technique with the sole goal of reducing the number of terms in a kernel-like estimator.

Note that the recursive kernel estimator has no practical advantage over a kernel estimator unless the estimate is desired at only a finite number of predetermined points. Thus the idea of adaptive mixtures is to reduce the number of terms so that the computational requirements necessary to compute the estimate for arbitrary x at any time is lessened and the estimate is more efficiently represented.

2.3.3 Decision Rule. Observations for which $\mathcal{P}_n(\zeta_{n+1}; \hat{\theta}_n) = 1$ in (6) correspond to jumps in the likelihood surface. This is a so-called "dynamic dimensionality," which can be useful for guiding the estimator toward a good solution and away from local maxima corresponding to poor solutions. The consistency of adaptive mixtures hinges on the fact that these jumps do in fact propel the recursive EM-type algorithm into a high-quality estimate.

The decision to add a term $\mathcal{P}(\cdot)$, a kind of clustering criterion, can be made in a number of ways. The simplest is to check the Mahalanobis distance from the observation to each of the terms; if the minimum of these exceeds a threshold (called the create threshold, T_C) then the point is in some sense too far away from the existing terms, and a new term should be created. Recall that the square of the Mahalanobis distance between a point x and a term with mean $\mu^{(i)}$ and standard deviation $\sigma^{(i)}$ is defined by $M^{(i)}(x) = ((x - \mu^{(i)})/\sigma^{(i)})^2$. Thus if the create threshold is T_C, then a new term is created at the point ζ_{n+1} if and only if $M(\zeta_{n+1}) = \min_i(M^{(i)}(\zeta_{n+1})) > T_C$.

$T_C = 1$ implies creation of a new term for any observation that is at least one standard deviation away from the mean of each term. Similarly, $T_C = 4$ implies creating a new term for any observation that is at least two standard deviations away from the mean of each term. The former will yield a faster rate of increase in the number terms in the model than the latter. Considering $\Lambda(x) = \exp(-\frac{1}{2}M(x))$, $\mathcal{P}(\cdot)$ may be defined as $\mathcal{P}(\zeta_{n+1}) = 1$ if and only if $\Lambda(\zeta_{n+1}) < T_C$ and $\mathcal{P}(\zeta_{n+1}) = 0$ otherwise. This version of $\mathcal{P}(\cdot)$ is used in the sequel. In this case $T_C = \exp(-.5) \approx .6065$ translates into a threshold of one standard deviation; $T_C = \exp(-2.) \approx .1353$; into a threshold of two standard deviations.

Other approaches would be to create stochastically with probability inversely proportional to $M(\zeta_{n+1})$ (scaled appropriately so that the values lie in the range [0, 1]) analogous to a simulated annealing technique or use the estimated density directly rather than the individual terms.

2.3.4 Relationship to Other Methods. The adaptive mixture model given by (6) is designed to allow for both data-driven smoothing and data-driven increases in complexity. Although such an approach is relatively novel, it is instructive to consider the relationship with maximum penalized likelihood estimation and semiparametric estimation.

The decision rule $\mathcal{P}$ described earlier is analogous to the penalty function used in maximum penalized likelihood estimation in that it can be chosen to force a smoother estimate by allowing fewer terms in the estimate. But the theory of maximum penalized likelihood has no explicit method for data-driven complexity (for example, Tapia and Thompson 1978).

Roeder (1990), (1992) and Lindsay and Roeder (1992) considered an estimator superficially similar to adaptive mixtures but with a focus on mixtures with a single smoothing parameter. Motivation for pursuing adaptive mixtures can be gained from Roeder (1992), who echoed the words of Geman and Hwang (1982) in stating that no satisfactory data-based technique exists by which to choose the number of terms in a mixture and that for the method of sieves, no data-based method exists by which to choose the smoothing parameter(s). The adaptive mixtures technique was formulated specifically to address these two difficulties.

Another version of semiparametric estimation of interest was given by Olkin and Spiegelman (1987), who considered a probabilistic combination of parametric and nonparametric models. This approach does not address the computational statistics questions raised earlier, however, as a full-kernel estimator is inherent in the model.

2.3.5 Example. Consider a simple example of (6), where the procedure is rewritten as

$$\hat{\alpha}_n(x) = \sum_{j=1}^{m_n} \pi_j \phi(x; \mu_j, \sigma_j^2) \tag{12}$$

and

$$\hat{\alpha}_{n+1} = [1 - \mathcal{P}_n(\zeta_{n+1}; \hat{\alpha}_n)]\mathcal{U}_n(\zeta_{n+1}; \hat{\alpha}_n) + \mathcal{P}_n(\zeta_{n+1}; \hat{\alpha}_n)\mathcal{C}_n(\zeta_{n+1}; \hat{\alpha}_n) \tag{13}$$

for clarity. For illustrative purposes, the case in which $\alpha_0 = .5N(-2, 1) + .5N(2, 1)$, $\sigma_0 = 1$, and $T_c = .606$ is investigated for a particular random sample of 100 observations. The purpose is simply to walk through the approach. In this example $\zeta_1 = 1.71$, and thus $\hat{\alpha}_1 = N(1.71, 1)$. For the second observation, the decision rule $\mathcal{P}(\cdot)$ tests $\zeta_2 = 1.57$ against the model $\hat{\alpha}_1$. ζ_2 is close to $\hat{\alpha}_1$ in Mahalanobis distance and hence is within the threshold. Therefore, $\mathcal{P}(\cdot) = 0$ and the update rule rather than the create rule is used. $\zeta_3 = -2.72$ is far away in Mahalanobis distance from the single term in $\hat{\alpha}_2$. Thus $\mathcal{P}(\cdot) = 1$, and a new term is created based on this third observation. The estimate is now a mixture of two terms. Continuing, new terms are created at $\zeta_1, \zeta_3, \zeta_4, \zeta_8, \zeta_{17}, \ldots$. The rest of the observations yield updates. Figure 1 shows the model after the third observation. By the time that 100 observations have been processed, the estimate is definitely taking on the character of the true distribution (see $\hat{\alpha}_{100}$ in Fig. 2).

Table 1 gives the model depicted in Figure 2 after 100 observations, consisting of nine terms. Investigation of this table indicates that there are two major terms in $\hat{\alpha}_{100}$. That is, there are two terms with $\pi \geq .1$. These two terms have means in the vicinity of -2 and 2, the means of the true distribution. This lends credence to the conjecture that the adaptive mixture model is going toward a reasonable estimate.

Simulations and applications discussed in Sections 4 and 5 indicate that the approximation (12)–(13) has desirable properties for the recursive estimation of unknown densities. In particular, it seems to converge quickly to a good estimate of the density for a large class of densities. Nevertheless, be-

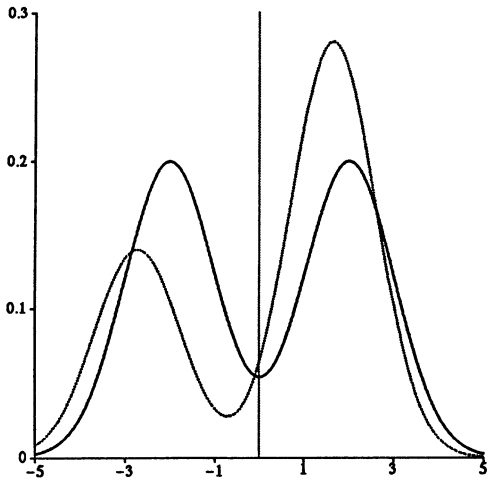

Figure 1. Adaptive Mixtures Example after Three Observations. True distribution $\alpha_0 = .5N(-2, 1) + .5N(2, 1)$ (solid line) versus adaptive mixtures estimate $\hat{\alpha}_3$ (dashed line). After three observations, the estimate has two terms.

cause adaptive mixtures have been designed for use in recursive, nonparametric applications, the traditional small-sample analysis is not the main issue. Asymptotic results will allow a more complete understanding of the algorithm and will be of use in evaluating the utility of (12)–(13) for particular applications.

3. CONVERGENCE PROPERTIES

3.1 The Method of Sieves

Letting $BC = BC(R)$ be the set of bounded, continuous functions $f: R \to R$, consider the parameter space $A = \{\alpha \mid \alpha \in BC, \int \alpha(x) = 1, \alpha(x) \geq 0 \text{ for all } x\}$. That is, A is the set of univariate, bounded, continuous probability density functions. Consider also the associated metric space (A, d), where $d(\cdot)$ is the L_1 metric, and the true and unknown parameter to be estimated is $\alpha_0 \in A$. The method of sieves (Geman and Hwang 1982; Grenander 1981) is a scheme by

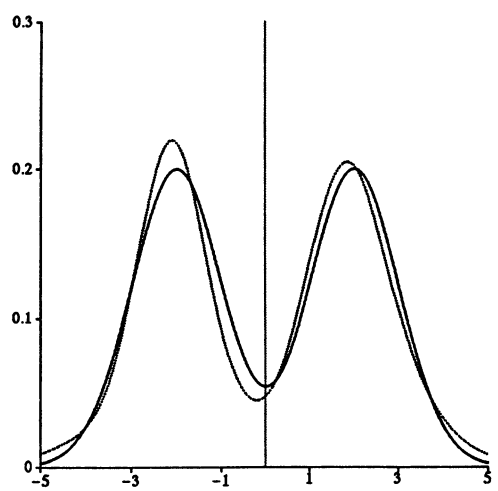

Figure 2. Adaptive Mixtures Example After 100 Observations. True distribution $\alpha_0 = .5N(-2, 1) + .5N(2, 1)$ (solid line) versus adaptive mixtures estimate $\hat{\alpha}_{100}$ (dashed line). After 100 observations, the estimate has 9 terms (see Table 1).

Table 1. Model α_{100} After 100 Observations in the Adaptive Mixtures Example

Term	Mean	Variance	Proportion
1	1.845	.801	.410
2	−2.098	.589	.367
3	3.085	.874	.057
4	−3.545	.958	.039
5	−1.461	.800	.043
6	.992	.809	.035
7	−.569	.816	.026
8	4.153	.911	.012
9	−4.603	.972	.010

which the set of admissible estimators in A is constrained, with the constraint relaxed as the number of observations increases in an effort to ensure that the estimator remains in A. This constrained mathematical optimization is important for nonparametric maximum likelihood estimation in that unconstrained maximum likelihood methods can yield a discrete estimate with a spike at each observation. Recall the example of Dirac δ spikes encountered previously. Such an estimate is not absolutely continuous with respect to Lebesgue measure and normally is not considered an admissible estimator.

A sieve for A is a sequence of subsets $\{S_m\}$ of A, usually requiring the constrains (a) $\cup S_m$ is dense in A, (b) $S_m \subset S_{m+1}$, and (c) S_m is compact for each m. The idea is to approximate the density α_0 by a succession of densities from S_m, with m increasing slowly compared to n. Examples of sieves include the conventional histogram where the number of bins is increased slowly compared to the number of observations, as well as the more sophisticated data-adaptive estimate discussed by Wegman (1975).

Let $\zeta_1, \zeta_2, \ldots$ be independent identically distributed observations drawn from the probability density $\alpha_0 \in A$. Consider the following sieve of mixtures contained in A:

$$S_m = \Bigg\{ \alpha_m : \alpha_m(x) = \sum_{i=1}^{m} \pi_i \phi(x; \mu_i, \sigma_i^2); \quad \sum_{i=1}^{m} \pi_i = 1; \varepsilon_m \leq \pi_i \leq 1 - \varepsilon_m; \quad \delta_m \leq \sigma_i \leq \gamma_m; |\mu_i| \leq \tau_m \Bigg\}, \tag{14}$$

$m = 1, 2, \ldots$, where ϕ is the standard normal probability density function, $\varepsilon_m > 0$ for $m > 1$, $0 < \delta_m < \gamma_m < \infty$, and $0 < \tau_m < \infty$ for all m. Under appropriate conditions on the true density α_0 and the sequence $\{m_n\} = \{m\}$, this is a consistent sieve.

Let the entropy be defined as $H(\alpha, \beta) = \int \alpha(x) \ln \beta(x)\, dx$. Consider $A' \subset A$, where $A' = \{\alpha \in A \mid H(\alpha, \alpha) < \infty \text{ and } \alpha \in C_0\}$. $C_0 \subset BC$ is defined as the set of functions that vanish at ∞; that is, A' is the restriction of A to densities with finite self-entropy and sufficiently regular tail behavior. A function f that vanishes at infinity is one in which for every $\varepsilon > 0$, the set $\{x \mid |f(x)| \geq \varepsilon\}$ is compact. This restricts the tail behavior of f.

Further restrict attention to the parameter space $A'' \subset A' \subset A$, where $A'' = \{\alpha \in A' \mid \text{supp}(\alpha) \subset [-k, k] \text{ for some } k$

$< \infty\}$; that is, A'' is the restriction of A' to densities with compact (but possibly unknown) support. Letting BC_c be the set of bounded, continuous functions with compact support, we have $BC_c \subset BC_0 \subset BC$.

For notational purposes, let $\mathcal{L}_n(\alpha)$ be the likelihood function, $\mathcal{L}_n(\alpha) = \prod_{i=1}^n \alpha(x_i)$. Let $M_m^n = M_m^n(\omega)$ be the set of all maximum likelihood estimators in S_m given a sample size n. Thus $M_m^n = \{\alpha \in S_m\colon \mathcal{L}_n(\alpha) = \sup_{\beta \in S_m} \mathcal{L}_n(\beta)\}$. Let $qM_m^n = \{\alpha \in S_m\colon \mathcal{L}_n(\alpha) \geq q \sup_{\beta \in S_m} \mathcal{L}_n(\beta)\}$ be the subset of S_m for which the likelihood is within a constant multiplier $0 < q \leq 1$ of the maximum likelihood. Using qM_m^n and applying Wald's theorem 2 (Wald 1949), S_m can be shown to be a consistent sieve.

Theorem 1. For $\alpha_0 \in A'$ there exists a sequence $m = m_n$, and for $\alpha_0 \in A''$ this sequence can be specified, such that $qM_m^n \to \alpha_0$ in L_1 a.s. for S_m.

Sieve (14) is a generalization of the sieve considered by Geman and Hwang (1982) allowing separate smoothing parameters for each term in the mixture. It is this more complex model of data-driven smoothing that gives the adaptive mixtures procedure its power and utility, that causes trouble in maximum likelihood estimation, and that is investigated next.

3.2 Adaptive Mixtures Asymptotics

Consider now the adaptive mixtures procedure from the standpoint of the method of sieves. The procedure described by (12) and (13) can easily be seen to be an instantiation of sieve (14). The sieve parameter m_n is the number of terms in the mixture after n observations, and the decision to move to the next higher sieve parameter is governed by $\mathcal{P}(\cdot)$. Thus Theorem 1 can be applied to the adaptive mixtures procedure. It remains to ensure that the number of terms in (13) grows slowly enough with respect to the number of observation (as identified in Theorem 1) and that the estimate $\hat{\alpha}_n$ is eventually in qM_m^n.

Technical details require the consideration of a convergent baseline sequence in S_m and small neighborhoods about this sequence. For $\delta > 0$, let $D_m = \{\alpha \in S_m\colon H(\alpha_0, \alpha) \leq H(\alpha_0, \alpha_m) - \delta\}$, where $\{\alpha_m\} = \{mh_m^{-1} \sum_{j=1}^m \phi(x; \xi_j, h_m) \in S_m\}$ is a subsample kernel estimator based on the data. Using α_m as a baseline, D_m is the set of all estimators in S_m that are at least δ worse than α_m (in entropy). Given $\{\mathcal{O}_\kappa\}_{\kappa=1}^{\lambda_m}$, with each set $\mathcal{O}_\kappa \subset S_m$, let $\psi(x; \mathcal{O}_\kappa) = \sup_{\beta \in \mathcal{O}_\kappa} \beta(x)$. Let $\rho_m = \max_\kappa \inf_{t \geq 0} \int \alpha_0(x) \exp[t \ln\{\psi(x; \mathcal{O}_\kappa)/\alpha_m(x)\}]\, dx$. Note that $\rho_m = \rho_{m_n}$ and $\lambda_m = \lambda_{m_n}$ are dependent on n as $\{m\} = \{m_n\}$. For the following analysis consider the conditions (A) $D_m \subset U\mathcal{O}_\kappa$ and (B) $\Sigma_m \lambda_m (\rho_m)^n < \infty$. Condition (A) couples the magnitude of λ_m (i.e., the number of covering sets) together with ρ_m (i.e., how well they cover) through the size of the covering sets $\mathcal{O}_\kappa$. Once the covering sets are chosen to satisfy (A), condition (B) gives the condition for consistency (as in Geman and Hwang 1982, thm. 2). The sequence discussed in Theorem 1 can thus be specified.

To indicate the procedure of adding new terms to the model, write $\hat{\alpha}_n \in S_m$ for $N_m \leq n \leq N_{m+1} - 1$. A new term is added (the mth) at observation N_m; that is, $\mathcal{P}_{N_m - 1}(\zeta_{N_m}; \hat{\alpha}_{N_m - 1}) = 1$. In this notation, (13) becomes $\hat{\alpha}_n = \mathcal{U}(\hat{\alpha}_{n-1}; \zeta_n)$ for $N_j + 1 \leq n \leq N_{j+1} - 1$, $j \geq 1$ and $\hat{\alpha}_n = \mathcal{C}(\hat{\alpha}_{n-1}; \zeta_n)$ for $n \in \{N_1, N_2, \ldots\}$.

Assume that based on the first n' observations, $m' = m_{n'}$ terms have been created, where $m' \ll n'$. (Thus $N_{m'} = n'$.) Otherwise, there are minimal constraints on this early creating process, with the provision that $N_2 \geq 3$. Now, in accordance with condition (B), consider

$$\sum_{n=1}^{\infty} \lambda_m \rho_m^n = \sum_{m=1}^{\infty} \sum_{n=N_m}^{N_{m+1}-1} \lambda_m \rho_m^n$$

$$= \sum_{m=1}^{m'-1} \sum_{n=N_m}^{N_{m+1}-1} \lambda_m \rho_m^n + \sum_{m=m'}^{\infty} \sum_{n=N_m}^{N_{m+1}-1} \lambda_m \rho_m^n.$$

After observation, $N_{m'}$, $\lambda_{m'+1}$, and $\rho_{m'+1}$ can be calculated, for by letting $\xi_1 = \zeta_1$, $\xi_2 = \zeta_2$, $\ldots$, $\xi_{m+1} = \zeta_{N_m}$ for $m \geq 2$ and requiring that $N_2 \geq 3$, $\{\xi_i\}_{i=1}^{\infty} = \zeta_1, \zeta_2, \{\zeta_{N_m}\}_{m=2}^{\infty}$ is the subsequence required for the kernel estimator α_m described earlier. At this point, the covering sets can be identified and $N_{m'+1}$ can be chosen to conform with the requirements of conditions (A) and (B) by choosing $N_{m'+1}$ such that $\sum_{n=N_{m'}}^{N_{m'+1}-1} \lambda_{m'} \rho_{m'}^n < \varepsilon_{m'} = 2^{-m'}$. More generally, after creating term k on observation N_k, λ_k, ρ_k, and N_{k+1} can be calculated to conform with conditions (A) and (B) and ensure consistency.

The adaptive mixtures procedure thus has an approximation theorem derived from the method of sieves. In particular, the adaptive mixtures procedure is strongly consistent in L_1 for $\alpha_0 \in A'$ (and also for A'', because $A'' \subset A'$).

Theorem 2. If $\alpha_0 \in A'$ (resp. A''), then the sequence of estimates $\{\hat{\alpha}_n\}$ produced by the adaptive mixtures procedure, under the conditions detailed later on $\mathcal{P}(\cdot)$ and $\mathcal{C}(\cdot)$, is strongly consistent. That is, $\hat{\alpha}_n \to \alpha_0$ in L_1 a.s.

The assumptions placed on the decision rule $\mathcal{P}(\cdot)$ are that one "waits long enough" between creations—that is, m increases slowly enough with respect to n. It is also required that when there are local maxima of the likelihood surface, $\mathcal{C}(\cdot)$ must propel $\hat{\alpha}_n$ into a sufficiently small neighborhood of a sufficiently good maxima. This can entail alteration of the step size [β in Eq. (5)] in the recursive update procedure $\mathcal{U}(\cdot)$, because "sufficiently good maxima" is in terms of qM_m^n and, given such a maxima, "sufficiently small neighborhood" is in terms of the recursive step size. Thus this second assumption assures one of eventually being near a good (possibly local) maximum, whereas the first stipulation is necessary to allow the estimate to conform to the requirements of Theorem 1. The geometry of these likelihood surfaces was discussed by Lindsay (1983a), (1983b).

In light of these assumptions, given a true density α_0 from which $Z_n = \{\zeta_1, \ldots, \zeta_n\}$ is drawn, define $\Lambda_m^n = \{\alpha \in S_m | \alpha$ as the location of a (possibly local) maxima of $\mathcal{L}_n(\cdot)\}$. Note that Λ_m^n depends on α_0 through the dependence of the likelihood function $\mathcal{L}_n(\cdot)$ on Z_n. It is necessary to argue that for some M and all $m > M$, $\hat{\alpha}_{N_{m+1}-1} \in qM_m^{N_{m+1}-1}$.

With $\mathcal{P}(\cdot) = 0$ and $\mathcal{U}(\cdot)$ in effect, the procedure is a recursive version of the EM algorithm (Titterington 1984, thm. 2). In this case, for any $\alpha \in (q + \varepsilon)M_m^{N_{m+1}-1}$, $0 < \varepsilon < 1 - q$, there exists a step size β and a sufficiently small neigh-

borhood Ω_α such that

$$\hat{\alpha}_{N_m} \in \Omega_\alpha \Rightarrow \hat{\alpha}_{N_{m+1}-1} \in qM_m^{N_{m+1}-1}. \quad (15)$$

Thus any $\mathcal{C}(\cdot)$ that propels $\hat{\alpha}_{N_m}$ into such an Ω_α will suffice.

Consider the iterative version of adaptive mixtures, $\hat{\alpha}_{N_{m+1}-1} = \mathrm{EM}(\hat{\alpha}_{N_m}, Z_{N_{m+1}-1})$, with a step size β small enough so the estimate does not jump out of the convex neighborhood of the likelihood surface defined by $Z_{N_{m+1}-1}$ in which $\hat{\alpha}_{N_m}$ resides. Then for each $\alpha \in \Lambda_m^{N_{m+1}-1} \cap (q+\varepsilon)M_m^{N_{m+1}-1}$, the convex neighborhood Ω_α of α has the property (15) required for Theorem 2.

Returning now to the recursive procedure, the identification of appropriate neighborhoods Ω_α is more difficult. In this case Theorem 2 requires $\hat{\alpha}_{N_{m+1}-1} = \mathrm{REM}(\hat{\alpha}_{N_m}, z_{N_{m+1}-1}) \in qM_m^{N_{m+1}-1}$. Thus Ω_α can depend on the data ordering. Nevertheless, for β sufficiently small, each $\alpha \in \Lambda_m^{N_{m+1}-1} \cap (q+\varepsilon)M_m^{N_{m+1}-1}$ has a nonempty neighborhood Ω_α satisfying (15).

Furthermore, for $\alpha_0 \in B'$ (resp. B'') where B' (resp. B'') $= \{\alpha \in A'$ (resp. A''), there exists a fixed $0 < q < 1$ and M such that $m > M \Rightarrow$ there exists (a.s.) $N'_m \geq N_m$ such that $\Lambda_m^{N'_m} \subset (q+\varepsilon)M_m^{N'_m}$, where $\mathcal{L}_{N'_m}(\cdot)$ (and hence $\Lambda_m^{N'_m}$ and $(q+\varepsilon)M_m^{N'_m}$) depend on α as earlier, the identification of $\mathcal{C}(\cdot)$ is trivial; $S_m = \Omega_\alpha$.

For $\alpha_0 \in B'$ (and B''), Theorem 2 not only yields strong L_1 consistency for recursive adaptive mixtures, but also the create function $\mathcal{C}(\cdot)$ and the decision function $\mathcal{P}(\cdot)$ are easily identified via Theorem 1 for A' (and A''). In particular, $\mathcal{C}(\cdot)$ is (almost) arbitrary ($S_m = \Omega_\alpha$), and $\mathcal{P}(\cdot)$ is driven by the N'_m. Otherwise, for $\alpha_0 \in A'$ (and A''), $\mathcal{C}(\cdot)$ must propel $\hat{\alpha}$ into a sufficiently small neighborhood of some density $\alpha \in \Lambda_m^n \cap (q+\varepsilon)M_m^n$.

4. SIMULATION RESULTS

The following simulation examples, using the adaptive mixtures algorithm described in Section 2, focus on large-sample properties and computational complexity for adaptive mixtures. To analyze these simulation results from the viewpoint of conventional estimators, consider three simulations that, taken together, indicate the utility of the adaptive mixtures technique. Simulation 1 considers data drawn from a normal pdf, $\alpha_0 = \phi(0, 1)$. For simulation 2, a simple normal mixture is considered with $\alpha_0 = 1/2\phi(-2, 1/2) + 1/2\phi(1/2, 3/2)$. Simulation 3 considers the much more difficult situation of a log-normal pdf, $\alpha_0 = x^{-1}(2\pi)^{-1/2} \exp(-1/2(\ln x)^2) - 5$ for $0 < x < \infty$. Each simulation example consists of 20 Monte Carlo replications, with the number of observations (i.e., the sample size) going from 100 to 1,000 in increments of 100 for simulations 1 and 2 and from 1,000 to 10,000 in increments of 1,000 for simulation 3. The computer run time required for the adaptive mixtures procedure is minimal, and the number of Monte Carlo replications could have been increased dramatically; however, 20 replications is sufficient to get a feel for the performance of the estimator.

For the normal case, the adaptive mixtures procedure converges quite quickly. Specifically, both the mean L_1 and L_2 errors and the variance of the estimator under both norms appear to be decreasing toward 0. For the purposes of preliminary quantitative comparisons, an estimated rate of convergence for the adaptive mixtures based on a regression of the error curves to the model $O(n^{-\gamma})$ is considered. In this case, $\gamma(L_1) = .49$ and $\gamma(L_2) = .91$. The relevant numbers for comparison in L_2 (see Silverman 1986) are 1.0, .8, and .5. That is, a convergence rate of $O(n^{-1})$ is the best that one can expect even with a parametric estimator, $O(n^{-.8})$ with an optimal kernel estimator, and $O(n^{-.5})$ with a simple function approach. Thus adaptive mixtures perform quite well. Although the procedure is nonparametric, this particular implementation is inherently based on the normal model, and hence this performance may not be completely unexpected.

The computational complexity of the model is defined as the number of terms used in the data-driven adaptive mixtures development. Here this complexity grows quite slowly with n, with an average of less than 8 terms used for 1,000 observations. Recall that this complexity increase is the sieve parameter m_n in Equation (13). This is compared to the kernel estimator, which requires a separate term for each of the 1,000 observations. A qualitative comparison indicates that the adaptive mixtures estimator is significantly more parsimonious than the kernel estimator. Comparing the adaptive mixtures with a conventional normal approximation would obviously indicate degraded performance, as the average of 8 terms used in the adaptive mixtures is compared with a single term in the normal approximation. But for a situation in which the true distribution is not known to be normal, one simply desires a good estimator while at the same time keeping the model complexity as small as possible. The aforementioned error rates, together with the complexity indicated, would appear to meet these goals. If the true distribution is known to be normal, then the normal approximation is obviously a superior choice.

Simulation 2 indicates a similarly impressive, although slower, rate of convergence for the adaptive mixtures procedure in the normal mixture case. Again, both the mean L_1 and L_2 errors and the variance of the estimator under both norms appear to be decreasing toward 0. The estimated rate of convergence for the adaptive mixtures based on a regression of the error curves to the model $O(n^{-\gamma})$ yields $\gamma(L_1) = .37$ and $\gamma(L_2) = .69$. Thus the adaptive mixtures perform better than the simple functions and nearly as well as the kernel estimator for this mixture case.

The computational complexity again grows quite slowly with n with an average of 9.5 terms used for 1,000 observations. Considering this complexity, note that if one attempted to perform the estimation with a conventional normal approximation, convergence would be impossible. If knowledge of the specific nature of the true distribution is given, then one can estimate the α_0 with a two-term normal mixture. But lacking such information, the nonparametric adaptive mixtures approach allows estimation of this mixture α_0 without prior knowledge of its two-term character. The same algorithm that exhibited quality performance in the normal case (simulation 1) succeeds when the true distribution is nonnormal. Furthermore, the increase in computational complexity from an average of 8 terms in simulation 1 to an average of 9.5 terms here is quite acceptable.

Simulation 3 indicates the performance of the system on the much more difficult problem of a log-normal distribution. Because the procedure is based on a sieve of normal mixtures, the added difficulty of this third estimation problem is in some very real sense a different kind of difficulty, as opposed to just the difference in degree of difficulty between the first two examples. The convergence of the adaptive mixtures procedure is not nearly as clear in this example. Here simulations of 10,000 observations (compared to 1,000 in Simulations 1 and 2) are used, and the convergence is much slower. This is of course to be expected. The estimated rate of convergence for the adaptive mixtures based on a regression of the error curves to the model $O(n^{-\gamma})$ yields $\gamma(L_1) = .12$ and $\gamma(L_2) = .09$. Again, this deterioration of performance from the normal and mixture cases is expected.

The computational complexity shown in Figure 3 grows to an average of 28.75 terms for 10,000 observations (and less than 47 terms for 100,000 observations), which for performance like that depicted in Figure 4 (a single example of the recursive estimate produced after 100,000 observations with 46 terms) seems outstanding. Again, it is noteworthy that the same algorithm that successfully fitted an average of 8 terms to a normal distribution in simulation 1 and an average of 9.5 terms to a mixture distribution in simulation 2 has successfully used a relatively parsimonious average of 28.75 terms to estimate this decidedly nonnormal distribution. The error results indicate that the performance is not as good in L_1 or L_2 error, but in terms of the difficulty of the problem, the results presented for simulation 3 may indeed be more impressive. A nonparametric estimator was required, and the adaptive mixtures procedure automatically allocated a reasonable number of terms for the problem.

Figure 5 depicts the largest terms in a preliminary model (after 1,000 observations) for the estimate shown in Figure 4, which is based on 100,000 observations. The three terms shown in Figure 5 indicate a typical instantiation of the application of the adaptive mixtures procedure to the log-normal distribution based on a sample size of 1,000 observations. Here one sees the aforementioned data-driven smoothing. In the region of support where the true density spikes, the terms in the model have relatively small variances. Conversely, in the broad tail of the support of the true density, the model gravitates toward terms with a large variance. Table 2 shows this phenomenon numerically for the terms depicted in Figure 5. The individual variances in the adaptive mixture model, allowed under sieve S_m [Equation (14)], yield the ability to fit the local smoothness of the true density.

These local smoothing results should be compared to the transformed kernel estimator approach of Wand, Marron, and Ruppert (1991). Their approach involves using a transformation to normality and then performing a standard kernel estimator in the transformation space. The inverse transformation then yields an estimator with nonuniform smoothing parameters superficially similar to the results in Figure 5. But it should be noted that their approach is not

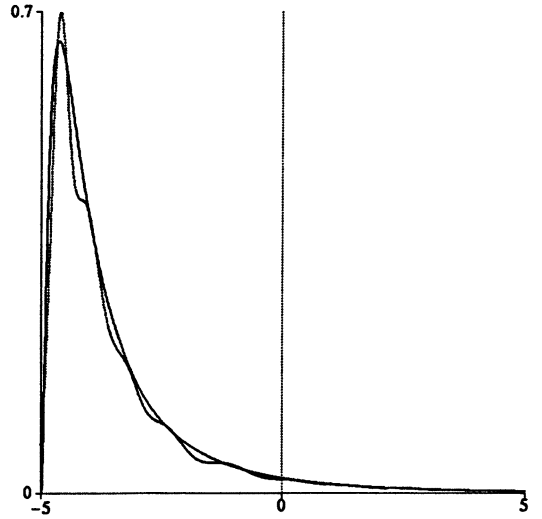

Figure 4. Log-Normal Simulation: Example Estimate $\hat{\alpha}_{100,000}$ (Dashed Line) Versus True Distribution α_0 (Solid Line). This recursive model has 46 terms.

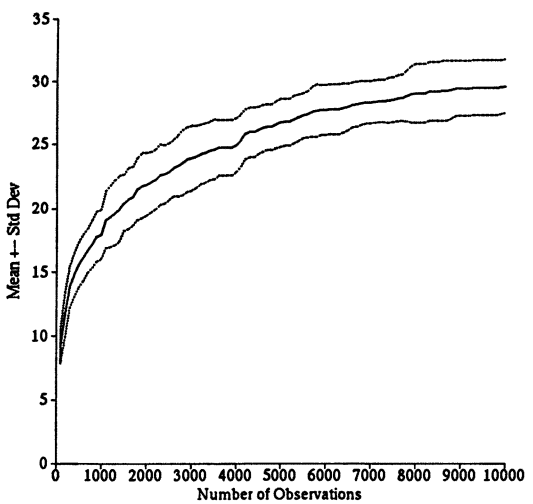

Figure 3. Log-Normal Simulation: Model Complexity (number of terms) and Standard Deviation Versus Number of Observations for the Adaptive Mixtures Estimator, Based on 20 Monte Carlo Replications.

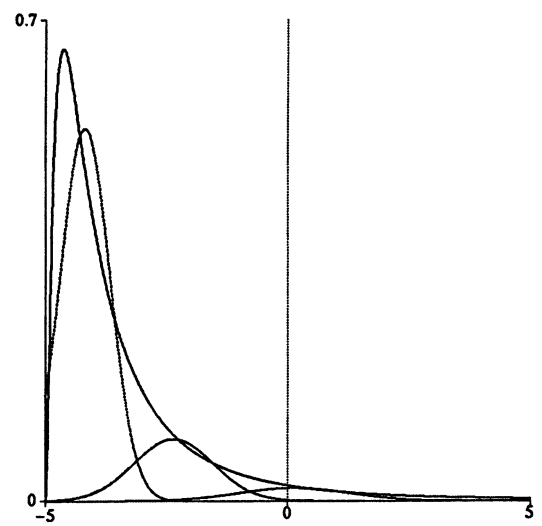

Figure 5. Log-Normal Simulation: Three Major Terms for an Example Estimate $\hat{\alpha}_{1,000}$ (Dashed Line) Versus True Distribution α_0 (Solid Line). This model indicates data-driven smoothing (see Table 2).

Table 2. Simulation 3: Log-Normal Distribution

Term 3	$\mu = -4.184744$
	$\sigma^2 = .259182$
	$\pi = .689485$
Term 4	$\mu = -2.358960$
	$\sigma^2 = .658321$
	$\pi = .179715$
Term 5	$\mu = .128446$
	$\sigma^2 = 1.260401$
	$\pi = .051398$

NOTE: Terms in the adaptive mixture model for an example estimate $\hat{\alpha}$ based on 1,000 observations. Only the terms with $\pi \geq .02$ are shown.

designed to reduce the computational complexity of kernel estimation and hence does not directly address the concerns considered here.

Although these simulation results are based on qualitative analysis, they nevertheless lend credence to the conclusions developed in Section 3—that adaptive mixtures density estimation has desirable complexity and convergence properties. A theoretical analysis of rates of convergence will require coupling the results available for the recursive EM algorithm with an understanding of the impact of the creation process. Convergence rates for adaptive mixtures would not be expected to be superior to a recursive implementation of finite mixtures when the true distribution is a mixture of normals and the number of terms has been correctly specified.

5. APPLICATION TO DISCRIMINANT ANALYSIS

5.1 Discriminant Analysis

As an application of this density estimation technique, consider its relationship to discriminant analysis. In the two-class discrimination problem, the observations are assumed to be independent identically distributed random variables with a probability density of the overall distribution of $\alpha(x) = \sum_{i=1}^{2} \pi_i \alpha_i(x)$, where π_i are the prior probabilities for the individual classes and α_i are the probability density functions for the individual classes. A formal motivation for using density estimation in discriminant analysis is obtained via consideration of the asymptotics of the discriminant procedure. When using the Bayesian discriminant function, L_1 convergence of density estimates to the true (though unknown) class-conditional densities implies convergence of the discriminant procedure to Bayes-optimal in the minimum probability of misclassification sense (see Devroye and Gyorfi 1985, chap. 10). Thus Theorem 2 immediately implies the following.

Theorem 3. If α_1 and α_2 are elements of A' (resp. A''), then, under the conditions of Theorem 2, the probability of misclassification produced by the adaptive mixtures technique converges to the Bayes optimal.

Although Theorem 3 does not give the universality of kernel discriminant analysis, it does allow for significantly more robust estimation than the finite mixture models. One need not know the structure of the model.

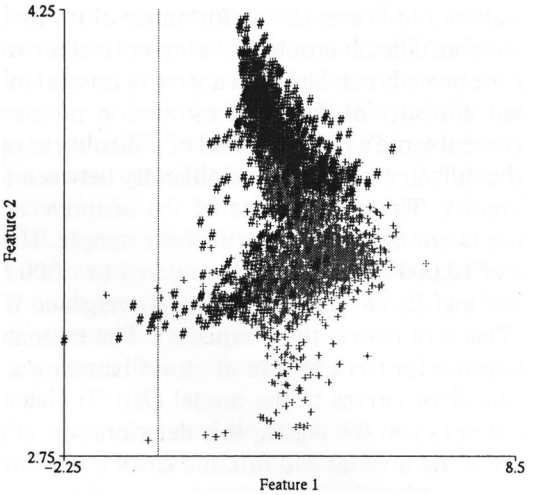

Figure 6. Two-Class Scatterplot of Experimental Data: 1,000 Observations From Class 0 (#) and 2,000 Observations From Class 1 (+).

5.2 Experimental Results

This experiment is based on feature vector observations drawn from a gray-scale image using power law theory. Solka, Priebe, and Rogers (1992) discussed these texture-based observations, which are local in nature in that they depend only on a given pixel and a small neighborhood about that pixel, and provided references to the literature. The problem considered here is to discriminate between two classes of objects (man-made vs. natural) based on these observations.

The data set consists of 2,000 observations from class 1 (natural) and 1,000 observations from class 0 (man-made) chosen at random (see Fig. 6). Performance is considered from the standpoint of the trade-off between increases in probability of correct classification and decreases in probability of false alarm.

Numerical performance results are given in a hypothesis testing framework in which the null hypothesis is "H_0: observation ζ is drawn from class 0," the alternative hypothesis is "H_1: observation ζ is drawn from class 1," and the likelihood ratio test statistic is used. The P(CC) values represent the probability of correctly classifying a class 0 observation as class 0; the P(FA) values represent the probability of incorrectly classifying a class 1 observation as class 0. The results for the estimators are given based on leave-one-out cross-validation.

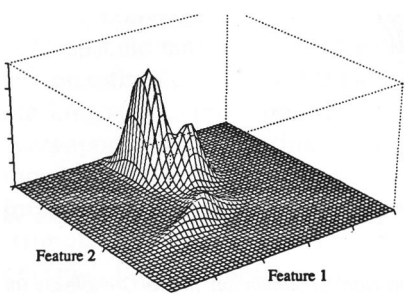

Figure 7. Adaptive Mixtures Density Estimate for Class 0.

The adaptive mixture models presented use 5 terms per class, as opposed to the 3,000 total terms required by the kernel estimator. That is, the complexity of the adaptive mixture model shown in Figure 7 is based on models with significantly fewer terms.

Figure 6 presents a scatterplot of the 3,000 observations used in this example. For illustrative purposes, Figure 7 presents the adaptive mixtures density plot for class 0 in the bivariate feature space. For this sample at least, the structure of the estimates is quite nonnormal. Figure 8 shows the class 0 observations overlaid on the adaptive mixtures–based classifier. The purpose of Figure 8 is to illustrate the significant increase in faithfulness to the data that a mixture-based approach can have as compared to an approach predicted on the normal assumption. The close adherence of the adaptive mixtures approach to the structure of the data indicates that this nonparametric approach is more powerful than an approach with strict parametric assumptions. In addition, this adaptive mixtures approach uses significantly fewer terms than a kernel estimator, making the approach more applicable. In fact, adaptive mixtures is the superior estimator for this problem, as is shown in Figure 9, which compares adaptive mixtures to linear and quadratic classifiers as well as kernel estimation. This can be explained by the fact that although the problem at hand requires more complex discriminant surfaces than the linear and quadratic classifiers can yield, five-term adaptive mixture models are sufficiently complex. Furthermore, the sample is perhaps too small to support the extreme complexity afforded by the kernel estimator. The adaptive mixtures procedure fills a niche in the middle ground between the normal assumption and the kernel estimator.

6. CONCLUSIONS

Preliminary work in developing adaptive mixtures has been done by Priebe and Marchette (1991, 1993), and an original formulation was given by Marchette and Priebe (1990). The technique combines the appealing properties of both kernel estimators and finite mixture models in that it converges for a large class of probability density functions $\mathcal{F}$ while maintaining the low computational complexity associated with finite mixture models. The automatic determination of the number of terms in the model based on the data, as well as the ability to determine separate smoothing parameters for the separate terms, gives the technique powerful new capabilities that directly address the issues inherent in the field of computational statistics. Preliminary work by Geman and Hwang (1982) considered the theoretical properties of a restricted version of adaptive mixtures. The theory has been extended to allow for data-driven smoothing, and explicit mechanisms have been developed by which the complexity of the model adapts to the data. The utility of the procedure for finite sample problems is in its ability to automatically determine the number of terms and their placement for a finite mixture model.

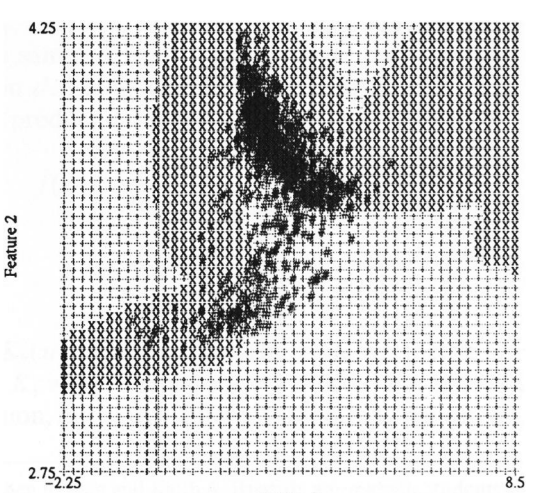

Figure 8. Adaptive Mixtures Discriminant Regions Overlaid with Class 0 Data: The Region for Class 0 (××××) Versus the Region for Class 1 (++++).

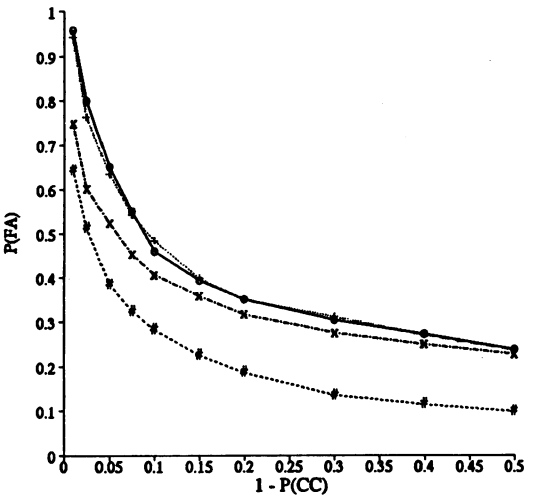

Figure 9. Discriminant Results for Four Classifiers: #, Adaptive Mixtures; ×, Kernel Estimator; +, Quadratic Classifier; and ○, Linear Classifier.

[Received April 1993. Revised September 1993.]

REFERENCES

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society,* Ser. B, 39, 1–38.

Devroye, L. (1979), "On the Pointwise and the Integral Convergence of Recursive Kernel Estimates of Probability Densities," *Utilitas Mathematicas,* 15, 113–128.

Devroye, L., and Gyorfi, L. (1985), *Nonparametric Density Estimation: The L_1 View,* New York: John Wiley.

Everitt, B. S., and Hand, D. J. (1981), *Finite Mixture Distributions,* New York: Chapman and Hall.

Fukunaga, K., and Hayes, R. R. (1989), "The Reduced Parzen Classifier," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2, 423–425.

Geman, S., and Hwang, C.-R. (1982), "Nonparametric Maximum Likelihood Estimation by the Method of Sieves," *The Annals of Statistics,* 10, 401–414.

Grenander, U. (1981), *Abstract Inference,* New York: John Wiley.

Lindsay, B. G. (1983a), "The Geometry of Mixing Likelihoods: A General Theory," *The Annals of Statistics,* 11, 86–94.

——— (1983b), "The Geometry of Mixing Likelihoods, Part II: The Exponential Family," *The Annals of Statistics,* 11, 783–792.

Lindsay, B. G., and Roeder, K. (1992), "Residual Diagnostics for Mixture Models," *Journal of the American Statistical Association,* 87, 785–794.

Marchette, D. J., and Priebe, C. E. (1990), "The Adaptive Kernel Neural Network," *Mathematical and Computer Modelling,* 14, 328–333.

McLachland, G. J., and Basford, K. E. (1988), *Mixture Models,* New York: Marcel Dekker.

Olkin, I., and Spiegelman, C. H. (1987), "A Semiparametric Approach to Density Estimation," *Journal of the American Statistical Association,* 82, 858–865.

Parzen, E. (1962), "On the Estimation of a Probability Density and Mode," *The Annals of Mathematical Statistics,* 33, 1065–1076.

Priebe, C. E., and Marchette, D. J. (1991), "Adaptive Mixtures: Recursive Nonparametric Pattern Recognition," *Pattern Recognition,* 24, 1197–1209.

——— (1993), "Adaptive Mixture Density Estimation," *Pattern Recognition,* 26, 771–785.

Redner, R. A., and Walker, H. F. (1984), "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review,* 26, 195–239.

Roeder, K. (1990), "Density Estimation With Confidence Sets Exemplified by Superclusters and Voids in the Galaxies," *Journal of the American Statistical Association,* 85, 617–624.

——— (1992), "Semiparametric Estimation of Normal Mixture Densities," *The Annals of Statistics,* 20, 929–943.

Rosenblatt, M. (1956), "Remarks on Some Nonparametric Estimates of a Density Function," *The Annals of Mathematical Statistics,* 27, 832–837.

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

Silverman, B. W. (1986), *Density Estimation,* London: Chapman and Hall.

Solka, J. L., Priebe, C. E., and Rogers, G. W. (1992), "An Initial Assessment of Discriminant Surface Complexity for Power Law Features," *Simulation,* 58, 311–318.

Tapia, R. A., and Thompson, J. R. (1978), *Nonparametric Probability Density Estimation,* Baltimore: Johns Hopkins University Press.

Titterington, D. M. (1984), "Recursive Parameter Estimation Using Incomplete Data," *Journal of the Royal Statistical Society,* Ser. B, 46, 257–267.

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

Wald, A. (1949), "Note on the Consistency of the Maximum Likelihood Estimate," *The Annals of Mathematical Statistics,* 20, 595–601.

Wand, M. P., Marron, J. S., and Ruppert, D. (1991), "Transformation in Density Estimation," *Journal of the American Statistical Association,* 86, 343–361.

Wegman, E. J. (1975), "Maximum Likelihood Estimation of a Probability Density Function," *Sankhya,* Ser. A, 37, 211–224.

——— (1988), "Computational Statistics: A New Agenda for Statistical Theory and Practice," *Journal of the Washington Academy of Science,* 78, 310–322.

Wegman, E. J., and Davies, H. I. (1979), "Remarks on Some Recursive Estimators of a Probability Density," *The Annals of Statistics,* 7, 316–327.

Wolverton, C. T., and Wagner, T. J. (1969), "Asymptotically Optimal Discriminant Functions for Pattern Classification," *IEEE Transactions on Information Theory,* 15, 258–265.