Pattern Recognition, Vol. 26, No. 5, pp. 771-785, 1993
Printed in Great Britain

0031-3203/93 $6.00 + .00
Pergamon Press Ltd
Pattern Recognition Society

ADAPTIVE MIXTURE DENSITY ESTIMATION

CAREY E. PRIEBE†§ and DAVID J. MARCHETTE‡
† Naval Surface Warfare Center, Code B10, Systems Research & Technology Department, Dahlgren, VA 22448-5000, U.S.A.
‡ Naval Ocean Systems Center, Code 421, San Diego, CA 92152-5000, U.S.A.

(*Received* 26 *December* 1990; *in revised form* 4 *May* 1992; *received for publication* 13 *October* 1992)

Abstract—A recursive, nonparametric method is developed for performing density estimation derived from mixture models, kernel estimation and stochastic approximation. The asymptotic performance of the method, dubbed "adaptive mixtures" (Priebe and Marchette, *Pattern Recognition* **24**, 1197–1209 (1991)) for its data-driven development of a mixture model approximation to the true density, is investigated using the method of sieves. Simulations are included indicating convergence properties for some simple examples.

Density estimation Kernel estimator Mixture model Stochastic approximation
Recursive estimation Nonparametric estimation Method of sieves Maximum likelihood
EM algorithm

1. INTRODUCTION

A fundamental problem in statistics is estimating the density from which a set of observations is drawn. Among the applications for such an estimate are discriminant analysis[1] and unsupervised learning.[2,3] The existence of a (stationary) stochastic process X, from which the independent identically distributed observations $\{x_i\}$ are drawn, yields the necessity for a stochastic approximation procedure. In this paper we address the issue of recursive[4,5] and nonparametric[6-8] density estimators. The necessity of nonparametric techniques stems from the wide range of applications in which no parametric family can be assumed for the probability density function corresponding to X. Recursive procedures are often required due to the nature of the (classification or discrimination) application and the quantity or rate of observations.

By virtue of addressing the types of applications that can be termed recursive and nonparametric, we have at once made the problem more difficult and more interesting. The recursive assumption eliminates the possibility of using iterative techniques. It is necessary, by hypothesis, to develop our estimate at time t only from our previous estimate and the newest observation. The nonparametric assumption implies that we cannot make any but the simplest assumptions about our data. Realistic restrictions on processing and memory, as might be imposed in automatic target recognition, remote sensing, and automatic control applications, in conjunction with high data rates, make such applications, and the procedure discussed herein, an important subject in pattern recognition.

Adaptive mixtures are a nonparametric statistical pattern recognition technique developed in reference (9) from the methods of kernel estimation[10,11] and finite mixture modelling.[4,12,13] Similarities exist between adaptive mixtures and potential functions,[14] reduced kernel estimators,[15] and maximum penalized likelihood methods.[6,16] The scheme employs stochastic approximation methods to recursively develop density estimates for use in classification and discriminant analysis. As with conventional Robbins–Monro methods,[17,18] the asymptotic convergence of adaptive mixture methods is an important issue.

In Section 2 we discuss the development of adaptive mixtures. In Section 3 we discuss the method of sieves and use it to prove some asymptotic theorems related to adaptive mixtures. Section 4 gives three simulations indicating the performance of adaptive mixtures relative to conventional procedures.

2. ADAPTIVE MIXTURES

The idea behind the adaptive mixtures approach is to approximate the probability density function of a stochastic process with a finite mixture of known densities. For example, in many clustering problems, a mixture of Gaussians may be used, with the intention of associating to each component of the mixture a group or cluster within the data. Most techniques of this sort are parametric, and require a fixed number of components for the approximating mixture. Various stochastic estimation techniques have been proposed to estimate the parameters of such a mixture. The approach taken by the adaptive mixture estimator is to adapt recursively the number of components to fit the data.

One of the useful properties that might be desired of an estimator is consistency. The idea is that the estimate should approach the true density as the number of data points goes to infinity, and that the variance of the estimator should likewise go to zero. Thus we require that our estimator be asymptotically unbiased, with the variance of the estimator approach-

§ Author to whom all correspondence should be addressed.

ing zero as the number of data points increases. In order for a finite mixture estimator (with a fixed number of components) to be consistent, 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. If the number of components in the estimator is allowed to grow indefinitely, however, these requirements can be relaxed. An extreme case of this is the kernel estimator, which is consistent under very weak conditions on the underlying density. The adaptive mixtures approach is to allow the number of components to grow, but at a much slower rate than a kernel estimator. This makes the adaptive mixtures approach much less computationally and memory intensive in practice, and hopefully produces a more useful small-sample estimator as well.

If the number of components of a mixture is fixed, then the mixture can be parameterized with a fixed-length vector of parameters, θ. Thus, if we write our estimate as

$$\alpha^*(x;\theta) = \sum_{i=1}^{m} \pi_i K(x;\Gamma_i) \tag{1}$$

where $K(x;\Gamma)$ is some fixed density parameterized by Γ, then

$$\theta = (\pi_1, \ldots, \pi_{m-1}, \Gamma_1, \ldots, \pi_m, \Gamma_m).$$

(We can assume for much of what follows that $K(\cdot)$ is taken to be the normal distribution, in which case Γ_i becomes $\{\mu_i, \sigma_i^2\}$.)

The basic stochastic approximation approach is to update recursively the estimate $\hat{\theta}$ of the true parameters θ_0 based on the latest estimate $\hat{\theta}_t$ and the newest observation x_{t+1}. That is

$$\hat{\theta}_{t+1} = \hat{\theta}_t + \Theta_t(x_{t+1}; \hat{\theta}_t) \tag{2}$$

for some update function Θ_t. This approach is usually used in situations where it is known that the true distribution of the data is of the form (1). However, one can certainly approach the problem from the perspective of fitting the data to a given model, where one finds the best fit of the form (1) to the data.

The specific form of equation (2) that will be used in this work is the one suggested by Titterington,[19] Nevel'son and Has'minskii[18] and others. If we let $I(\theta)$ be the Fisher information matrix, then the version of the recursive update formula we will use is

$$\hat{\theta}_{t+1} = \hat{\theta}_t + (nI(\hat{\theta}_t))^{-1} \frac{\partial}{\partial\theta} \log(\alpha^*(x_{t+1}; \hat{\theta}_t)) \tag{3}$$

where the derivative represents the vector of partial derivatives with respect to the components of θ.

This update function is discussed in, for example, reference (19). Basically, it guides a traversal of the estimate of the likelihood surface provided by the observations $\{x_i\}_{i=1}^t$ and based on the likelihood equations. This recursive maximum likelihood technique converges to the desired resultant estimator when properly constrained. The relationship of this estimator to the EM algorithm[20,21] is noted in reference (19).

In order to extend this approach to a nonparametric one with a variable number of components, the algorithm (3) must be extended to allow the addition of new components. The stochastic approximation procedure

$$\hat{\theta}_{t+1} = \hat{\theta}_t + [1 - P_t(x_{t+1}; \hat{\theta}_t)] U_t(x_{t+1}; \hat{\theta}_t) + P_t(x_{t+1}; \hat{\theta}_t) C_t(x_{t+1}; \hat{\theta}_t, t) \tag{4}$$

is used to update recursively the density. $P_t(\cdot)$ represents a (possibly stochastic) create decision and takes on values 0 or 1. This "penalty function" serves to constrain the addition of new terms, similar to maximum penalized likelihood, and will allow for the application of the constrained optimization technique described in Section 3 below. $U_t(\cdot)$ updates the current parameters using formula (3), while $C_t(\cdot)$ adds a new component to the model (1) analogous to a kernel estimation approach.

The addition of new components due to $C_t(\cdot)$ adds new parameters to θ (increments m in equation (1)), and the character of the likelihood surface is changed. The fact that $P_t(\cdot)$ may depend on x_{t+1} implies that this change can be data driven. Intuitively, this allows (4) to add a component if necessary, and allows for more efficient maximum likelihood estimation. The creation rule $C_t(\cdot)$ is chosen so that the proportion and variance of the new component decreases with the number of components. If the system always creates a component, never updating, the form of the estimator is (5), below.

If $P_t(x_{t+1}; \hat{\theta}_t) \equiv 1$ for $t < T$ and $P_t(x_{t+1}; \hat{\theta}_t) \equiv 0$ for $t \geq T$, the algorithm will fit T terms to the data (alternately, one could start with T components chosen using some a priori knowledge, and then let $P_t(x_{t+1}; \hat{\theta}_t) \equiv 0$ for all t). In particular, if $K(\cdot)$ is the normal distribution, then we have a normal mixture model. On the other hand, if $P_t(x_{t+1}; \hat{\theta}_t) \equiv 1$, the algorithm always creates a new component, centered at the new data point, and the estimate then becomes

$$\alpha^*(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - x_i}{h_i}\right). \tag{5}$$

This estimator is consistent under conditions similar to those required for the kernel estimator. (Much has been written about these kernel-type estimators. See, for instance, references (22–24).) Also, the process of fitting a mixture to a density can be made consistent as the number of components goes to infinity. Thus, it is reasonable to conjecture that the combination of the two approaches would be consistent, under reasonable assumptions about the create decision $P_t(\cdot)$.

Note that although (5) is a recursive estimator, in the sense that the estimator at time $n+1$ requires only the estimator at time n and the new point, it 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 components, so that the computational requirements necessary to compute the estimate at any time is lessened, and the estimate is

more compactly represented. One could simply use fewer points in the kernel estimator, throwing the extra points away, but this is clearly unacceptable. We would like to use all the points to improve the estimate.

Observations for which $P_t(x_{t+1};\hat{\theta})=1$ in (4) correspond to jumps in the likelihood surface, 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. Theorem 4 below hinges on the fact that these jumps do in fact propel our recursive EM-type algorithm into a high-quality estimate. As we move into the infinite-dimensional space (as (4) does under this dynamic dimensionality), consideration shifts from the parameters θ to the parameter α taking its values in the infinite-dimensional parameter space A (a restriction of the set of all densities, discussed below).

Simulations and applications indicate that the approximation (4) has desirable properties for the recursive estimation of known densities. In particular, it seems to converge quickly to a good estimate of the density for a large class of densities. Nevertheless, because 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 the approximation (4) for particular applications.

3. RESULTS

Throughout the following, we will denote by A the set of density functions containing the true density α_0. We will assume that A is a metric space with associated metric d, and will denote this by (A,d). The method of sieves[25-27] is a scheme by which the parameter space A is constrained, with the constraint slowly relaxed as the number of observations increases, in an effort to insure that the estimator remains in some desirable subspace of A. This is important, notably, for nonparametric maximum likelihood (ML) estimation (of which adaptive mixtures are an example) in that ML methods can yield a discrete estimator with a spike at each observation. Examples of sieves include the conventional histogram as a special case, as well as the more sophisticated estimate discussed in reference (28).

A sieve for A is a sequence of subsets $\{S_m\}$ of A. We often require one or more of the following constraints:

(i) $\cup S_m$ is dense in A;
(ii) $S_{m+1} \supseteq S_m$;
(iii) S_m is compact for each m.

The idea behind the method of sieves is to approximate the density α_0 by a succession of densities from S_m, with m increasing slowly compared to the number of observations n. A familiar sieve is the histogram, where we increase the number of bins slowly compared to the number of data points.

Consider the parameter space

$$A=\{\alpha|\alpha\in\mathscr{B}\mathscr{C},\int\alpha(x)=1,\alpha(x)\geq 0 \text{ for all } x\}.$$

That is, A is the set of univariate, bounded, continuous probability density functions. Let $x_1,x_2,\ldots$ be independent, identically distributed points drawn from the distribution $\alpha_0\in A$. Consider the following set of "sieves of mixtures" contained in A:

$$S_m^1=\left\{\alpha_m:\alpha_m(x)=m^{-1}\sum_{i=1}^{m}\phi(x;\mu_i,\sigma^2);\ \delta_m\leq\sigma^2\leq\gamma_m;|\mu_i|\leq\tau_m\right\} \quad (6)$$

$$S_m^2=\left\{\alpha_m:\alpha_m(x)=\sum_{i=1}^{m}\pi_i\phi(x;\mu_i,\sigma^2);\ \sum_{i=1}^{m}\pi_i=1;\varepsilon_m\leq\pi_i\leq 1-\varepsilon_m(i=1,\ldots,m);\ \delta_m\leq\sigma^2\leq\gamma_m;|\mu_i|\leq\tau_m\right\} \quad (7)$$

$$S_m^3=\left\{\alpha_m:\alpha_m(x)=m^{-1}\sum_{i=1}^{m}\phi(x;\mu_i,\sigma_i^2);\ \delta_m\leq\sigma_i^2\leq\gamma_m;|\mu_i|\leq\tau_m\right\} \quad (8)$$

$$S_m^4=\left\{\alpha_m:\alpha_m(x)=\sum_{i=1}^{m}\pi_i\phi(x;\mu_i,\sigma_i^2);\ \sum_{i=1}^{m}\pi_i=1;\varepsilon_m\leq\pi_i\leq 1-\varepsilon_m(i=1,\ldots,m);\ \delta_m\leq\sigma_i^2\leq\gamma_m;|\mu_i|\leq\tau_m\right\} \quad (9)$$

$m=1,2,\ldots$, where ϕ is the standard normal probability density function, $\varepsilon_m>0$ for $m>1$, $0<\delta_m<\gamma_m<\infty$ and $0<\tau_m<\infty\ \forall m$. The sequence S_m^1, S_m^2, S_m^4 can be seen to be progressively more general sieves.

We wish to discuss maximum likelihood estimation in terms of these sieves. Following Geman and Hwang,[26] Theorem 1, we show that under appropriate conditions on the true density α_0 and the sequence $\{m_n\}=\{m\}$, these are "consistent sieves".

Sieves (6) and (7) have a single smoothing parameter σ^2, while (8) and (9) have a separate parameter σ_i^2 for each term in the mixture. It is this more complex model of data-driven smoothing that causes trouble in maximum likelihood estimation, and that we shall investigate herein. Note that $\alpha\in S_m^i, i\in\{1,2,3,4\},\Rightarrow\alpha$ is a probability density function. For $\alpha\in S_m$, note that the Radon–Nikodym derivative of the probability measure induced by α with respect to Lebesgue measure is just $\alpha(x)$. Let $L_n(\alpha)$ be the likelihood function

$$L_n(\alpha)=\prod_{i=1}^{n}\alpha(x_i).$$

Let $M_m^n=M_m^n(\omega)$ be the set of all maximum likelihood estimators in $S_m=S_m^i$, $i\in\{1,2,3,4\}$, given a sample

size n

$$M_m^n = \{\alpha \in S_m : L_n(\alpha) = \sup_{\beta \in S_m} L_n(\beta)\}.$$

Let the entropy

$$H(\alpha, \beta) = \int \alpha(x) \ln \beta(x)\, dx.$$

Then the Kullback–Leibler information $J(\alpha, \beta) = H(\alpha, \alpha) - H(\alpha, \beta)$. Let A_m be the set of all maximum entropy estimators in $S_m = S_m^i$, $i \in \{1, 2, 3, 4\}$

$$A_m = \{\alpha \in S_m : H(\alpha_0, \alpha) = \sup_{\beta \in S_m} H(\alpha_0, \beta)\}.$$

Let $B_m(\alpha, \varepsilon) = \{\beta \in S_m : d(\alpha, \beta) < \varepsilon\}$ for $\alpha \in S_m$, where $d(\cdot)$ is the L_1 metric. Let $\psi(x; m, \alpha, \varepsilon) = \sup_{\beta \in B_m(\alpha, \varepsilon)} \beta(x)$. Notationally, we say that a sequence of sets $C_m \to \alpha$, if $\sup_{\beta \in C_m} d(\alpha, \beta) \to 0$.

Lemma 1. For $\alpha_0 \in A$, if $H(\alpha_0, \alpha_0) < \infty$, then for $S_m = S_m^i$, $i \in \{1, 2, 3, 4\}$

$$A_m \to \alpha_0 \text{ in } L_1 \text{ as } m \to \infty.$$

Proof. As noted in reference (26), it suffices to show (i) $H(\alpha_0, \alpha_n) \to H(\alpha_0, \alpha_0)$ implies $\alpha_n \to \alpha_0$ in L_1, and (ii) $\exists \{\alpha_m \in S_m\} \ni H(\alpha_0, \alpha_m) \to H(\alpha_0, \alpha_0)$. Statement (i) follows from the fact that we are operating in basically the same parameter space as Geman and Hwang,[26] and the Kullback–Leibler information $J(\alpha, \beta) = 0 \leftrightarrow \alpha = \beta$. See Proposition 2 of reference (27), for proof. For (ii), we consider

$$\{\alpha_m\} = \left\{ m h_m^{-1} \sum_{j=1}^{m} \phi(x; \xi_j, h_m) \in S_m \right\}$$

where the ξ_j $(j = 1, \ldots, m)$ are a random subsample of the observations $x_i (i = 1, \ldots, n)$. That (ii) holds for these standard, nonrecursive kernel estimators is well known (see, for example, references (29, 30)).

In light of the above lemma, we consider $A' \subset A$, where

$$A' = \{\alpha \in A | H(\alpha, \alpha) < \infty \text{ and } \alpha \in \mathscr{C}_0\}.$$

That is, A' is the restriction of A to densities with finite formal entropy and sufficiently regular tail behavior. ($\mathscr{C}_0$ denotes functions that "vanish at ∞".)

Theorem 1. If $\alpha_0 \in A'$, then for $S_m = S_m^i$, $i \in \{1, 2, 3, 4\}$, $\exists$ sequence $m \xrightarrow{n} \infty$ such that $M_m^n \to \alpha_0$ in L_1 a.s.

Proof.

(1) S_m is compact for each m.
(2) $\forall m, \alpha \in S_m, \varepsilon > 0$, $\psi(x; m, \alpha, \varepsilon)$ is measurable in x.
(3) $\forall m, \alpha \in S_m, \exists \varepsilon > 0 \ni$

$$\int_{-\infty}^{\infty} \alpha_0(x) \ln \psi(x)\, dx < \infty.$$

(1)–(3), together within Lemma 1, Theorem 1 of Geman and Hwang[26] and Proposition 2 of Geman,[27] imply M_m^n is almost surely nonempty and $M_m^n \to \alpha_0$ in L_1 a.s. for m increasing slowly enough with respect to n, as desired.

Similarly, letting $qM_m^n = \{\alpha \in S_m : L_n(\alpha) \geq q \sup_{\beta \in S_m} L_n(\beta)\}$ and using Theorem 2 of Wald,[31] we have:

Corollary 1. If $\alpha_0 \in A'$, $0 < q \leq 1$, $\exists\, m \xrightarrow{n} \infty$ such that $qM_m^n \to \alpha_0$ in L_1 a.s.

It is in this form that the theorem becomes most useful, as in practice we are sometimes able to meet these relaxed conditions.

Theorem 2. If $\alpha_0 = \sum_{i=1}^{N} \pi_i \phi(x; \mu_i, \sigma_i^2)$ for some $N < \infty$ (that is, α_0 is a finite mixture model), then, for sieve S_m^4, with m increasing slowly enough with respect to n, $M_m^n \to \alpha_n'$ and $n^{(1/2)}(\alpha_n' - \alpha_0)$ is asymptotically normal with optimal variance.

Proof. Lemma 1 and Theorem 1 hold and thus, as noted on p. 406 of reference (26), our result follows since $\alpha_0 \in S_\lambda \forall \lambda > m$. (See, for example, Chapters 31–33 of reference (32).)

In Theorems 1 and 2, the conditions on m are that it "increase slowly enough w.r.t. n". Theorem 3 allows us to be a bit more specific. We consider the parameter space $A'' \subset A' \subset A$, where

$$A'' = \{\alpha \in A' | \operatorname{supp}(\alpha) \subset [-k, k] \text{ for some } k < \infty\}.$$

That is, A'' is the restriction of A' to densities with compact (but possibly unknown) support. For $\delta > 0$, let $D_m = \{\alpha \in S_m : H(\alpha_0, \alpha) \leq H(\alpha_0, \alpha_m) - \delta\}$, where $\{\alpha_m\} = \left\{ m h_m^{-1} \sum_{j=1}^{m} \phi(x; \xi_j, h_m) \in S_m \right\}$ is the kernel estimator used above in Lemma 1. Using α_m as a baseline, D_m is the set of all estimators in S_m which are at least δ-worse than α_m (in entropy). Given $\{\mathscr{O}_\kappa\}_{\kappa=1}^{\lambda_m}$, for each set $S_m \supseteq \mathscr{O}_\kappa$, let $\psi(x; \mathscr{O}_\kappa) = \sup_{\beta \in \mathscr{O}_\kappa} \beta(x)$. Let

$$\rho_m = \max_{\kappa} \inf_{t \geq 0} \int \alpha_0(x) \exp[t \ln \{\psi(x; \mathscr{O}_\kappa)/\alpha_m(x)\}]\, dx.$$

Note $\rho_m = \rho_{m_n}$ and $\lambda_m = \lambda_{m_n}$ are dependent on n as $\{m\} = \{m_n\}$. For Theorem 3 below we consider conditions (A) $\cup_\kappa \mathscr{O}_\kappa \supseteq D_m$, and (B) $\Sigma_n \lambda_m (\rho_m)^n < \infty$.

Theorem 3. For $\alpha_0 \in A''$, a sequence $m = m_n$ can be specified such that

$$M_m^n \to \alpha_0 \text{ in } L_1 \text{ a.s. for } S_m = S_m^i,\ i \in \{1, 2, 3, 4\}.$$

Proof. Using Lemma 1 and Theorem 1 above, it suffices to describe an appropriate sequence $\{m\} = \{m_n\}$ and corresponding sequence of sets $\{\{\mathscr{O}\}_{\kappa=1}^{\lambda_m}\}_m$ satisfying (A) and (B) and apply Theorem 2 of Geman and Hwang.[26] For S_m^1, $m_n = O(n^{1/5-\varepsilon})$ for $\varepsilon > 0$.[26,27] A similar argument applies for the other mixture sieves.

We will proceed for S_m^4, letting $\delta_m = 1/m$, $\gamma_m = m$, $\tau_m = m$.

We begin by describing appropriate sets $\{\{\mathscr{O}_\kappa\}_{\kappa=1}^{\lambda_m}\}_{m=1}^{\infty}$, such that (A) holds. Given a positive integer a_m, let $\aleph = \{\alpha \in S_m^4 : \alpha(x) = \sum \pi_i \phi(x : \mu_j, \sigma_k)\}_{1 \leq i,j,k \leq a_m}$. $\aleph$ is then a finite set, with $|\aleph| = a_m^3$. Associate with each $\alpha \in \aleph$ all β such that $L_1(\alpha, \beta) < \Delta_m$. Call each set $\hat{\mathscr{O}}_\alpha$. The set of such $\hat{\mathscr{O}}_\alpha$ is also finite. Now, choose Δ_m large

enough s.t. $\cup_\alpha \hat{\mathcal{O}}_\alpha \supseteq S_m$. Finally, let $\mathcal{O}_\alpha = \hat{\mathcal{O}}_\alpha \cap D_m$. Then $\{\mathcal{O}_\alpha\}$ is a finite set (of size, say, $a_m^3 = \lambda_m$) and (A) is satisfied. It remains to determine a choice of $a_m \Rightarrow \Delta_m \Rightarrow \lambda_m$ such that (B) holds. (Note that, by choosing a_m large enough, we can make Δ_m as small as we wish, at the expense of larger λ_m, since all $\alpha \in S_m^4$ are well behaved.) This is analogous to the "small ball" argument used in Geman.[27] If $\mathcal{O}$ is small enough in L_1, and $\exists\, \alpha \in \mathcal{O}$, then $\psi(x;\mathcal{O}) = \sup_{\beta \in \mathcal{O}} \beta(x)$ will behave much like α. That is,

$\exists \Delta_m > 0$ s.t. for $\alpha \in \mathcal{O}$

$$|\alpha(x) - \psi(x)| < \varepsilon_\Delta \quad \forall x$$

$$\Rightarrow \int \alpha_0 \ln \psi/\alpha_m < \int \alpha_0 \ln (\alpha + \varepsilon_\Delta)/\alpha_m$$

$$< (\varepsilon'_\Delta) \int \alpha_0 \ln \alpha/\alpha_m$$

$$= (\varepsilon'_\Delta)(-\delta).$$

Since ε_Δ is arbitrary, we can make $\varepsilon'_\Delta \leq 1/2$, and hence $\int \alpha_0 \ln \psi/\alpha_m < -\delta/2$. Thus

$$H(\alpha_0, \alpha) - H(\alpha_0, \alpha_m) < -\delta \Rightarrow$$

$$H(\alpha_0, \psi) - H(\alpha_0, \alpha_m) < -\delta/2.$$

But then $\rho_m < 1$ ($\int \alpha_0 \ln \psi/\alpha_m < -\delta/2 \Rightarrow \int \alpha_0 \ln \psi/\alpha_m < 0$), and hence there exists a sequence $\{m\} = \{m_n\}$ such that (B) is satisfied. The maximum rate of increase of m, in terms of n, for which (B) is convergent, gives us our sequence.

Corollary 2. For $\alpha_0 \in A''$, a sequence $m = m_n$ can be specified such that

$$qM_m^n \to \alpha_0 \text{ in } L_1 \text{ a.s. for } S_m = S_m^i,\ i \in \{1,2,3,4\}.$$

Once the consistency of S_m^4 is established, it is necessary to investigate the relationship of the adaptive mixture process to these sieves. It is a simple matter to constrain the adaptive mixture so that it lies in S_m^4, and we assume that this is done. Thus, writing (4) in terms of sieve S_m^4 we have

$$\alpha_n^*(x) = \sum_{j=1}^{m_n} \pi_j \phi(x; \mu_j, \sigma_j^2) \tag{10}$$

$$\alpha_{n+1}^* = \alpha_n^* + [1 - P_n(x_{n+1}; \alpha_n^*)]\, U_n(x_{n+1}; \alpha_n^*) + P_n(x_{n+1}; \alpha_n^*) C_n(x_{n+1}; \alpha_n^*, n). \tag{11}$$

The sieve parameter m is the number of terms in the mixture, and the decision to move to the next higher sieve parameter is governed by $P(\cdot)$. The above theorems will be applied to the adaptive mixtures procedure. If we knew that the adaptive mixture attained a maximum likelihood estimate between creations of new components, then Theorems 1 and 3 would apply and say that the adaptive mixture is consistent. In fact, we only need to come within a fixed proportion of the maximum likelihood estimate, as noted in the corollaries above.

The adaptive mixtures procedure thus has an approximation theorem derived from the method of sieves, and the dynamics of the create decision $P(\cdot)$ and the create rule $C(\cdot)$ in equations (10) and (11) make the complexity of the model data-driven, while the recursive maximum likelihood estimation of the individual variances, inherent in $U(\cdot)$, yields the data-driven smoothing described at the outset. Specifically, $C(\cdot)$ must perform as a recursive kernel estimator, while $U(\cdot)$ performs as recursive maximum likelihood estimation. We recall, in the case $P_t(x_{t+1}; \hat{\theta}_t) \equiv 0\ \forall t \geq \tau$, we have a recursive maximum likelihood EM algorithm with known convergence properties, and if $P_t(x_{t+1}; \hat{\theta}_t) \equiv 1$ (that is, if we always create and never update), the recursive kernel estimator (5) is a consistent estimator.

The assumptions placed on the create rule $P(\cdot)$ are that one "waits long enough" between creations—that is, m increases slowly enough with respect to n—and that, when there are local maxima of the likelihood surface, $P(\cdot)$ propels α_n^* into a sufficiently small neighborhood of a sufficiently good maxima. The second assumption assures one of eventually being near a good (possibly local) maximum, while the first stipulation is necessary in order to allow the estimate to approach its asymptotic value at each step in the sieve and is the subject of Corollaries 1 and 2 above.

Theorem 4. If $\alpha_0 \in A'$ [resp. A''], then the sequence of estimates $\{\alpha_n^*\}$ produced by the adaptive mixture procedure ((10),(11)), under the conditions described above for $P(\cdot)$, is consistent. That is, $\alpha_n^* \to \alpha_0$ in L_1 a.s.

Proof. In light of Corollaries 1 and 2 above, it suffices to argue that, for some M and all $m > M$, there exists N_m such that $n > N_m \Rightarrow \alpha_n^* \in qM_m^n$ a.s. With $P(\cdot) = 0$ and $U(\cdot)$ in effect, α_n^* is a recursive version of the EM algorithm.[19,21] In this case we are assured that, for any $\alpha \in qM_m^n$, there exists a sufficiently small neighborhood Ω_α such that α_n^* in Ω_α implies the existence of such an N_m. (See Theorems 1 and 2 of Titterington[19] and Sections 3 and 4 of Redner and Walker.[21]) Thus any $P(\cdot)$ which propels α_n^* into such an Ω_α will suffice.

Note 1: Grenander[25] contains an alternate formulation of much of the relevant work found in Geman and Hwang[26] and Wald[31] used above.

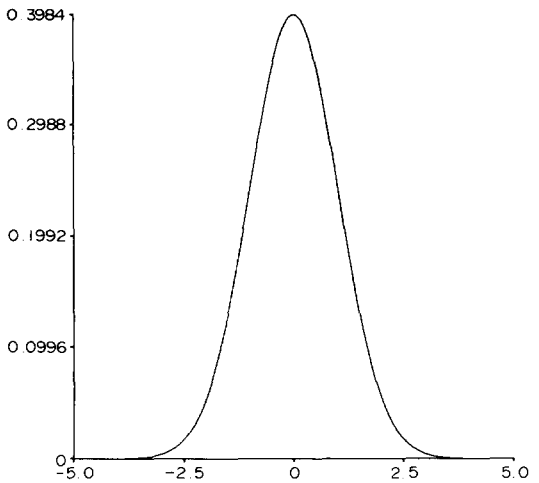
Note 2: No explicit formulation of $P(\cdot)$ is given here, only the existence of such a procedure. However, for $\alpha_0 \in B'$ [resp. B''], where

B'[resp. B''] = $\{\alpha \in A'$[resp. A'']:

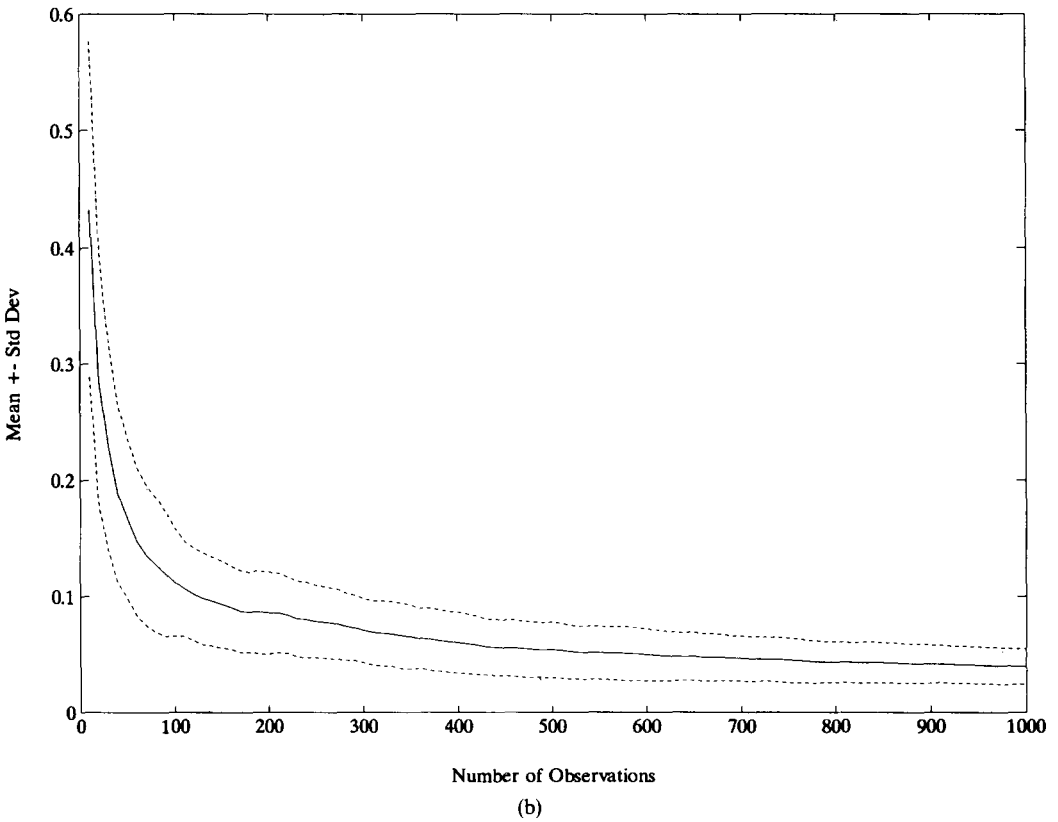
there exists a fixed $0 < q < 1$ and M such that $m > M \Rightarrow$ there exists N_m such that $n > N_m \Rightarrow$ {all local maxima in S_m} $\subset qM_m^n$ a.s.$\}$

the identification of $P(\cdot)$ is trivial; $S_m = \Omega_\alpha$.

The use of the sieve (9), and the set of maximum likelihood estimators M_m^n, in the consideration of the large-sample behavior of (10),(11) is appealing. As a recursive maximum likelihood estimator, the adaptive mixture procedure traverses the estimate of the likelihood surface provided by the sample in much the same way as the EM algorithm. When (11) adds a new term to its estimate ($P_t(x_{t+1}; \hat{\theta}_t) = 1$), the sieve index m in (9) is

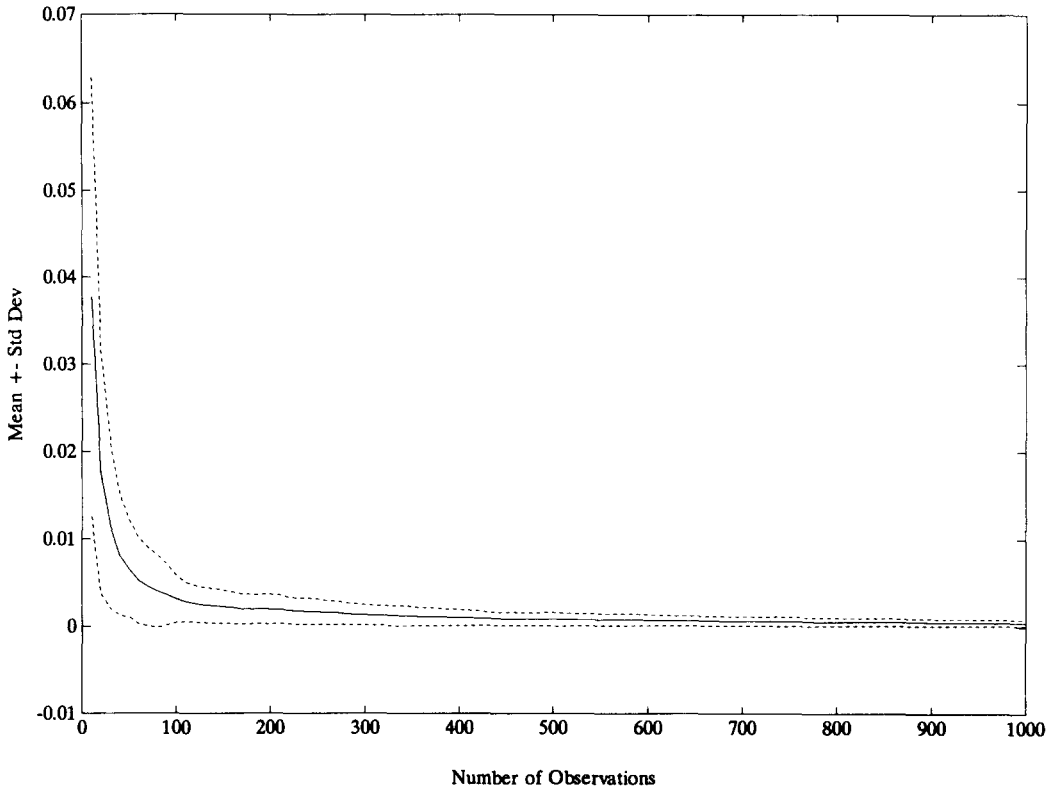


(a)

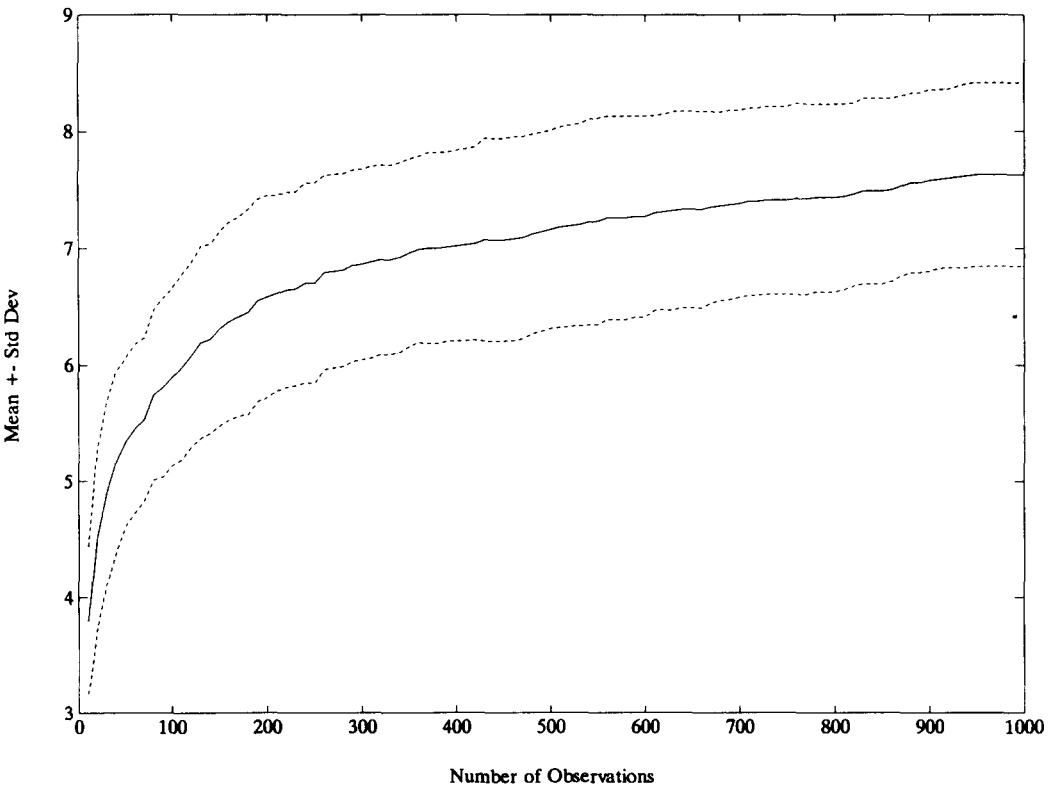


(b)

Fig. 1. $\alpha_0 = \phi(0, 1)$. (a) The true normal distribution α_0 from which the observations are drawn for Simulation 1. (b) L_1 convergence in mean and standard deviation based on 100 runs of 1000 observations. (c) L_2 convergence in mean and standard deviation based on 100 runs of 1000 observations. (d) Computational complexity in mean and standard deviation. The plot is number of terms in the model vs. number of observations based on 100 runs of 1000 observations. (e) An example of the model produced after 1000 observations. The solid curve is the true distribution $\alpha_0 = \phi(0, 1)$, the dashed line is the estimate α^*. To put this particular model in the perspective of the curves shown in (b)–(d), we note that this model has seven terms, an ISE of 0.000281, and an IAE of 0.037565.

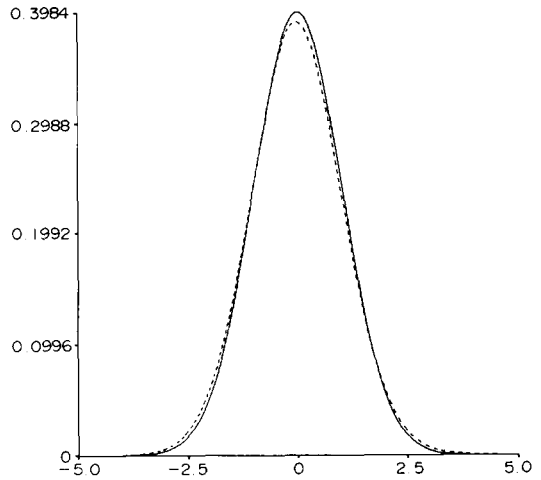


(c)



(d)

Fig. 1. (*Continued.*)



(e)

Fig. 1. (*Continued.*)

incremented. This increases the dimensionality of the likelihood surface, allowing the estimate to improve its likelihood, to "jump out" of local maxima of the likelihood surface. As the number of components increases, the local maxima into which the estimator may fall gets closer and closer to the true maximum, allowing the estimator to converge to α_0.

These results (most notably Theorems 3 and 4) lend asymptotic validity to the use of the recursive formula (4), or (10), (11), for density estimation problems. While analytical rates of convergence are not provided here (Bahadur[33] is of interest in this respect), it is noted that asymptotic results are, by hypothesis, our highest concern. Nevertheless, simulations indicate that the estimator is good enough for practical requirements in a wide range of problems. In practice, the assumptions placed on $P(\cdot)$ in Theorem 4 do not appear overly restrictive.

4. SIMULATION ANALYSIS

The following simulation examples, using the adaptive mixtures algorithm described above and in reference (9) indicate convergence properties and computational complexity for adaptive mixtures. To analyze these simulation results from the viewpoint of conventional estimator convergence rates, we consider three simulations, each of which points out different qualities of the estimator and its relation to standard techniques.

Simulation 1. Normal p.d.f.: $\alpha_0 = \phi(0, 1)$.

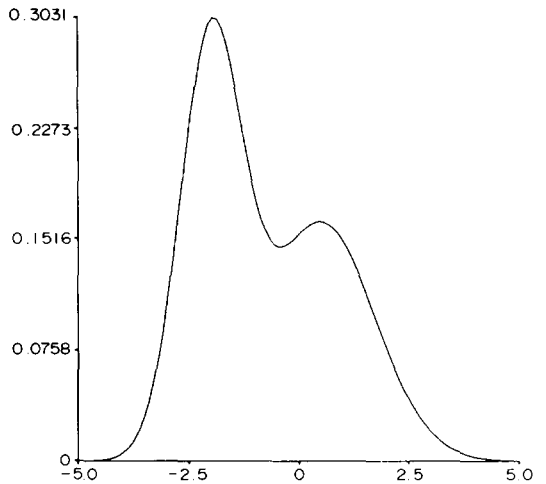
Simulation 2. Simple normal mixture: $\alpha_0 = 1/2\,\phi(-2, 1/2) + 1/2\,\phi(1/2, 3/2)$.

Simulation 3. Log-normal p.d.f.: $\alpha_0 = x^{-1}(2\pi)^{-1/2} \times \exp(-1/2(\ln x)^2) - 5$ for $0 < x < \infty$.

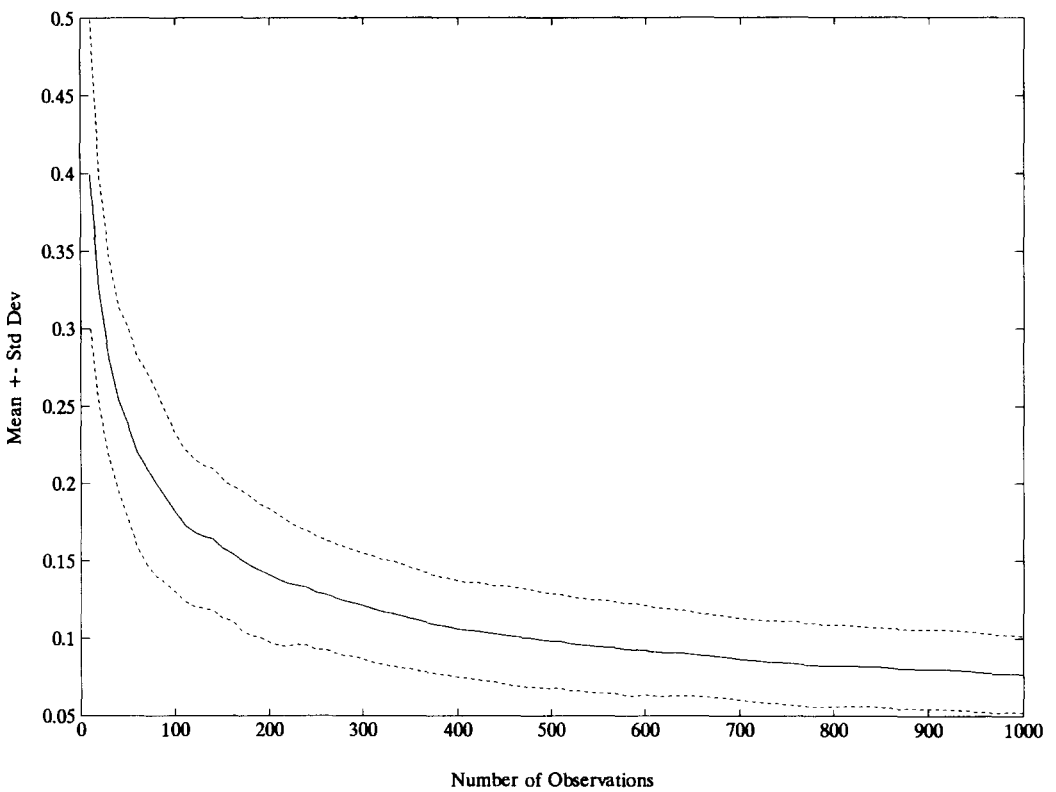
For each of these three simulation examples, we show the true distribution α_0, curves for L_1 and L_2 convergence vs. number of observations, a curve for number of terms used in the model vs. number of observations (as a measure of computational complexity), and an example of the model produced by the procedure.

Simulation 1 indicates quite fast convergence of the adaptive mixtures procedure for the normal case (Fig. 1(a)) as one might expect. Specifically, both the mean L_1 and L_2 errors and the variance of the estimator under both norms appear to be decreasing to zero (Figs 1(b) and (c)), indicating consistency. For the purposes of preliminary comparisons, we consider an estimated rate of convergence for the adaptive mixtures based on a regression of the curves in Figs 1(b) and (c) to the model $O(n^{-\gamma})$. We obtain $\gamma(L_1) = 0.4856$ and $\gamma(L_2) = 0.9144$. The relevant numbers for L_2 comparison (see references (11, 34, 35)) are 1.0, 0.8, and 0.5. That is, a convergence rate of $O(n^{-1})$ is the best one can expect even with a parametric estimator, $O(n^{-0.8})$ with an optimal kernel estimator, and $O(n^{-0.5})$ with a simple function approach. Thus the adaptive mixture performs quite well. While the procedure is nonparametric, this particular implementation is inherently based on the normal model, and this performance may not be completely unexpected. Figure 1(d) indicates the computational complexity of the model in number of terms used in the data-driven adaptive mixtures development. Here we see that this complexity grows quite slowly with n, with an average of less than eight terms used for 1000 observations. This is compared to the kernel estimator, which requires a separate term for each observation. (Recall that this complexity increase is the sieve parameter in equation (9)). Figure 1(e) is a single example of the estimate produced after 1000 observations using seven terms.

Simulation 2 indicates a similarly impressive, although slower, rate of convergence for the adaptive mixtures procedure in the normal mixture case (Fig. 2(a)). Again, both the mean L_1 and L_2 errors and the var-

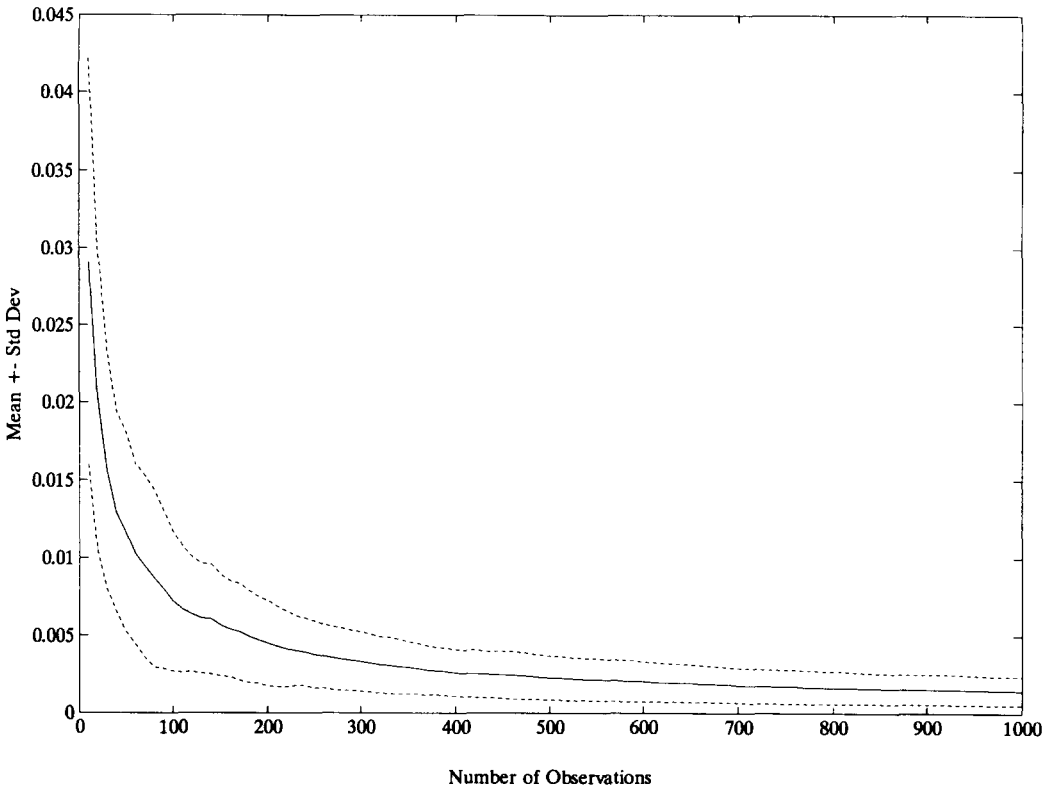


(a)

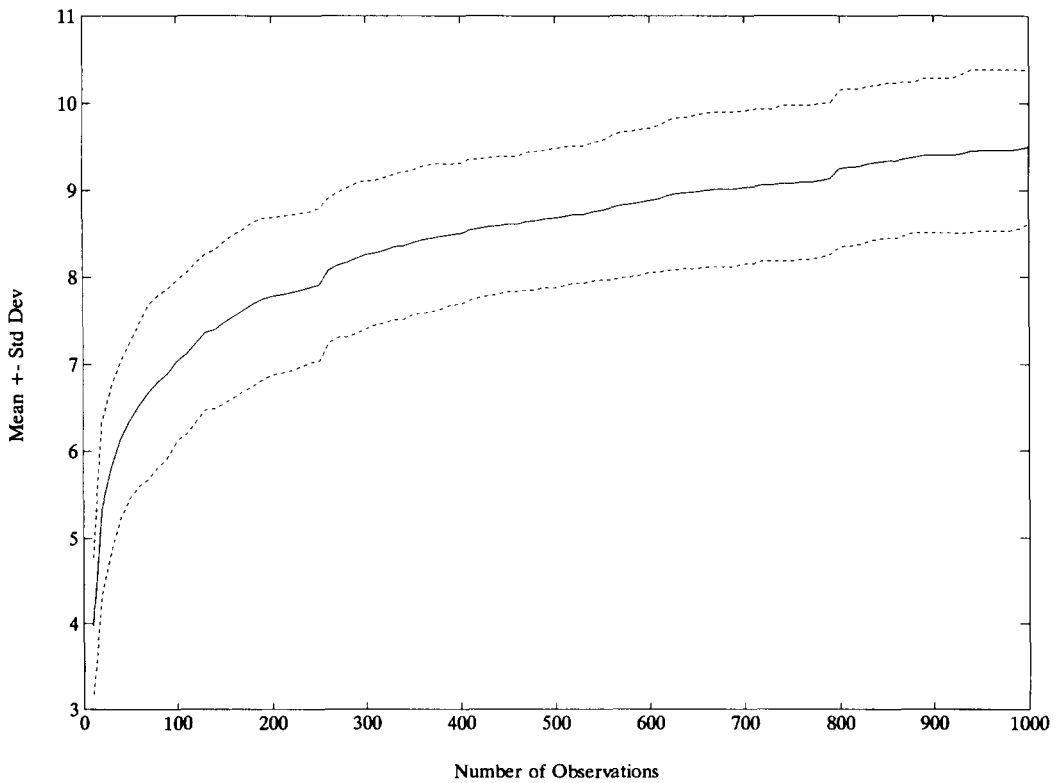


(b)

Fig. 2. $\alpha_0 = 1/2\,\phi(-2, 1/2) + 1/2\,\phi(1/2, 3/2)$. (a) The true finite mixture model distribution α_0 from which the observations are drawn for Simulation 2. (b) L_1 convergence in mean and standard deviation based on 100 runs of 1000 observations. (c) L_2 convergence in mean and standard deviation based on 100 runs of 1000 observations. (d) Computational complexity in mean and standard deviation. The plot is number of terms in the model vs. number of observations based on 100 runs of 1000 observations. (e) An example of the model produced after 1000 observations. The solid curve is the true distribution $\alpha_0 = 1/2\,\phi(-2, 1/2) + 1/2\,\phi(1/2, 3/2)$, the dashed line is the estimate α^*. To put this particular model in the perspective of the curves shown in (b)–(d), we note that this model has nine terms, an ISE of 0.002200, and an IAE of 0.091290.

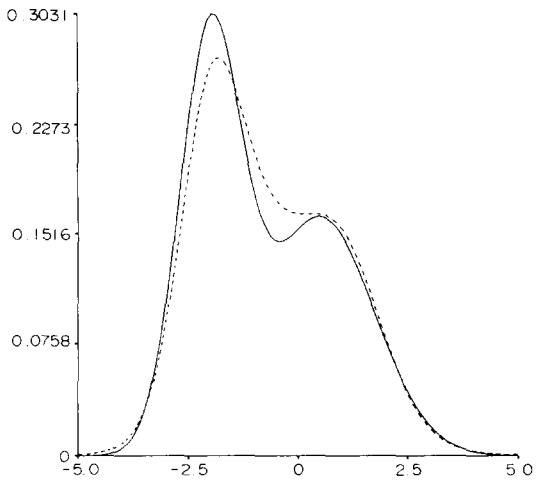


(c)



(d)

Fig. 2. (*Continued.*)



(e)

Fig. 2. (*Continued.*)

iance of the estimator under both norms appear to be decreasing to zero (Figs 2(b) and (c)), indicating consistency. The estimated rate of convergence for the adaptive mixtures based on a regression of the curves in Figs 2(b) and (c) to the model $O(n^{-\gamma})$ yields $\gamma(L_1) = 0.3734$ and $\gamma(L_2) = 0.6926$. Thus, the adaptive mixture performs better than the simple function's $O(n^{-0.5})$ and nearly as well as the optimal kernel estimator's $O(n^{-0.8})$ for this mixture case. The computational complexity (Fig. 2(d)) again grows quite slowly with n, with an average of 9.5 terms used for 1000 observations. Figure 2(e) depicts a single example of the estimate produced after 1000 observations with nine terms. This estimate is by no means the best produced and is in fact about one standard deviation worse, in both L_1 and L_2, than the mean.

Simulation 3 indicates the performance of the system on a much more difficult problem: that of a log-normal distribution (Fig. 3(a)). The convergence of the adaptive mixtures procedure is not nearly as clear in this example. Here we use simulations of 10,000 observations (as compared to 1000 in Simulations 1 and 2), and the convergence is much slower. This is, of course, to be expected since the procedure is a sieve of normal mixtures. The estimated rate of convergence for the adaptive mixtures based on a regression of the curves in Figs 3(b) and (c) to the model $O(n^{-\gamma})$ yields $\gamma(L_1) = 0.1158$ and $\gamma(L_2) = 0.0919$. Again, this deterioration of performance from the normal and mixture cases is expected. The computational complexity (Fig. 3(d)) 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 Fig. 3(e) (a single example of the estimate produced after 100,000 observations with 46 terms) seems outstanding.

Figure 3(f) depicts the three largest terms in a preliminary model of Simulation 3, based on only 1000 observations. Here we see the data-driven smoothing about which we have spoken. In the region of support where the true density spikes, the terms in our model have relatively small variances. Conversely, in the broad tail of the support of the true density, our model gravitates toward terms with a large variance. The individual variances in our adaptive mixture model, allowed under sieve S_m^4 (equation (9)), give us the ability to fit the local smoothness of the true density.

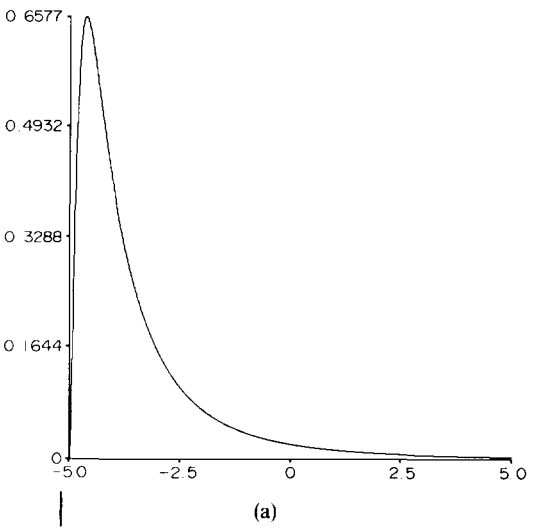
These results compare favorably to the "transformed kernel estimator" approach of Wand *et al.*[36] Their approach is to use a transformation to normality and then perform a standard kernel estimator in the transformation space. The inverse transformation then yields an estimator with non-uniform smoothing parameters superficially similar to the results in Fig. 3(f). However, it should be noted that their approach is not designed to reduce the computational complexity of kernel estimation, and hence does not directly address our concerns.

While these simulation results are based upon qualitative analysis, they nevertheless lend credence to the conclusions developed in Section 3: that adaptive mixtures density estimation has desirable large-sample properties.

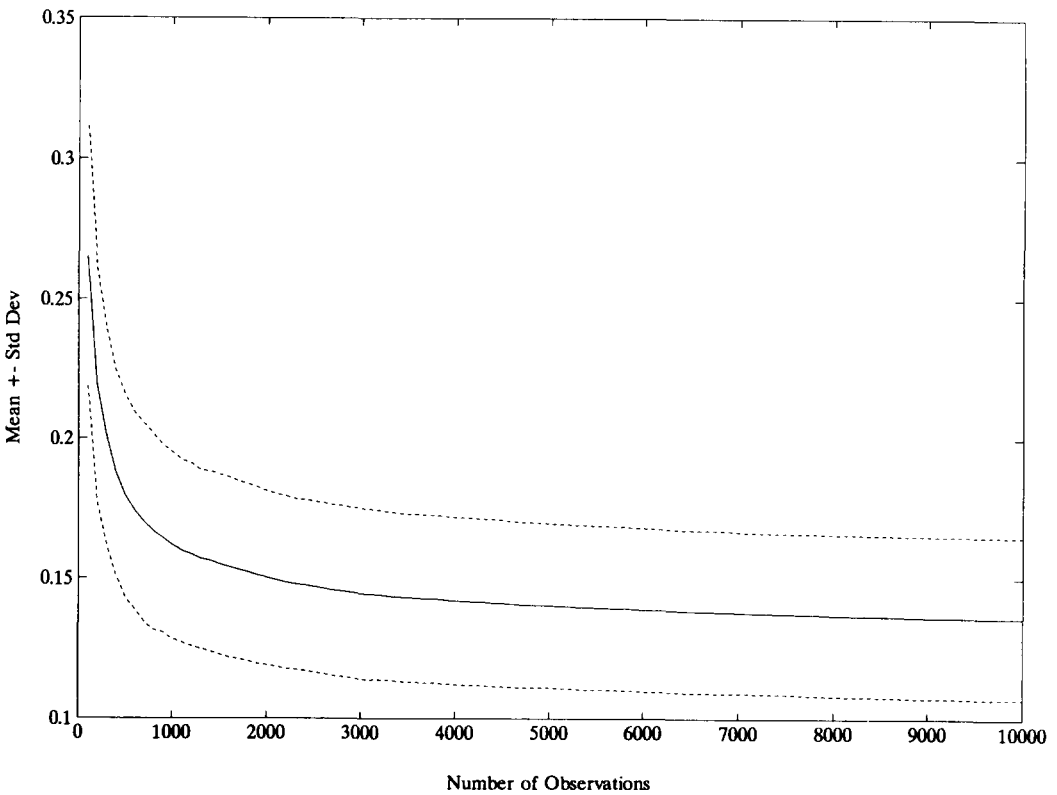
Experimental studies of the performance of the technique for situations in which the data are decidedly non-normal, and in which the feature space is two- and three-dimensional, is ongoing. Preliminary results have been published,[37] and a more detailed study has been concluded indicating performance comparable to the kernel estimator with a significant savings in computation complexity.[38]

5. DISCUSSION

The method of sieves, applied to the adaptive mixture density estimation procedure, indicates that the procedure is consistent under appropriate conditions. The algorithm gives a useful method of recur-

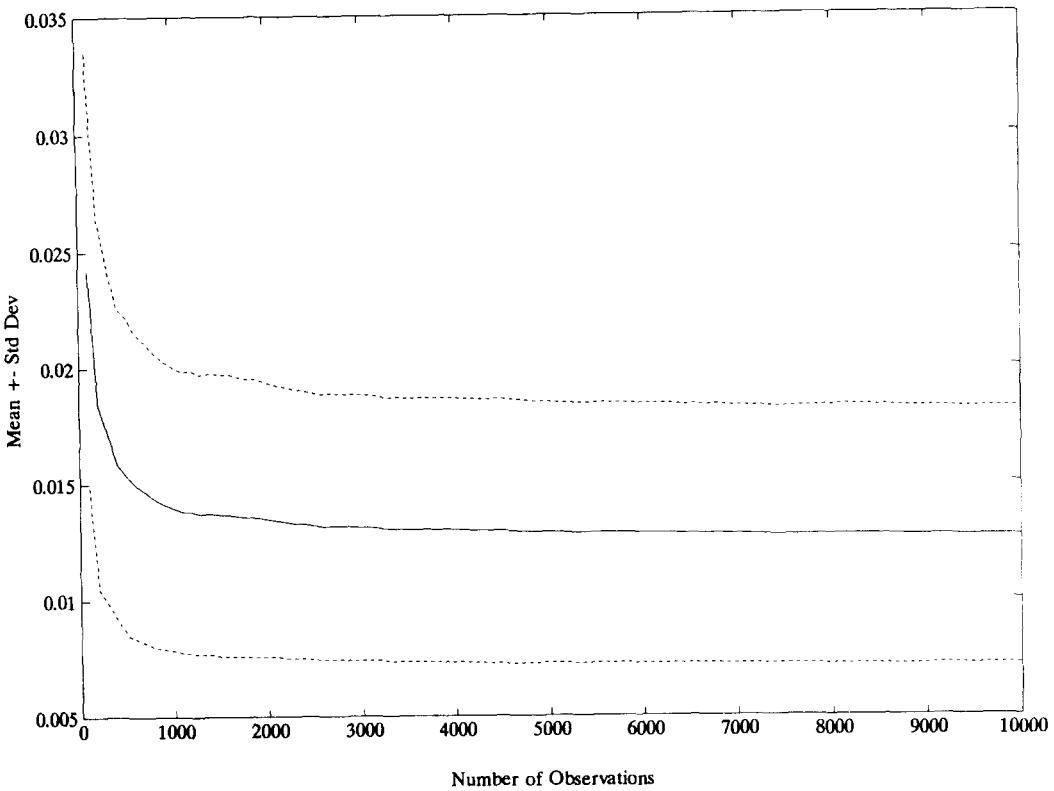


(a)

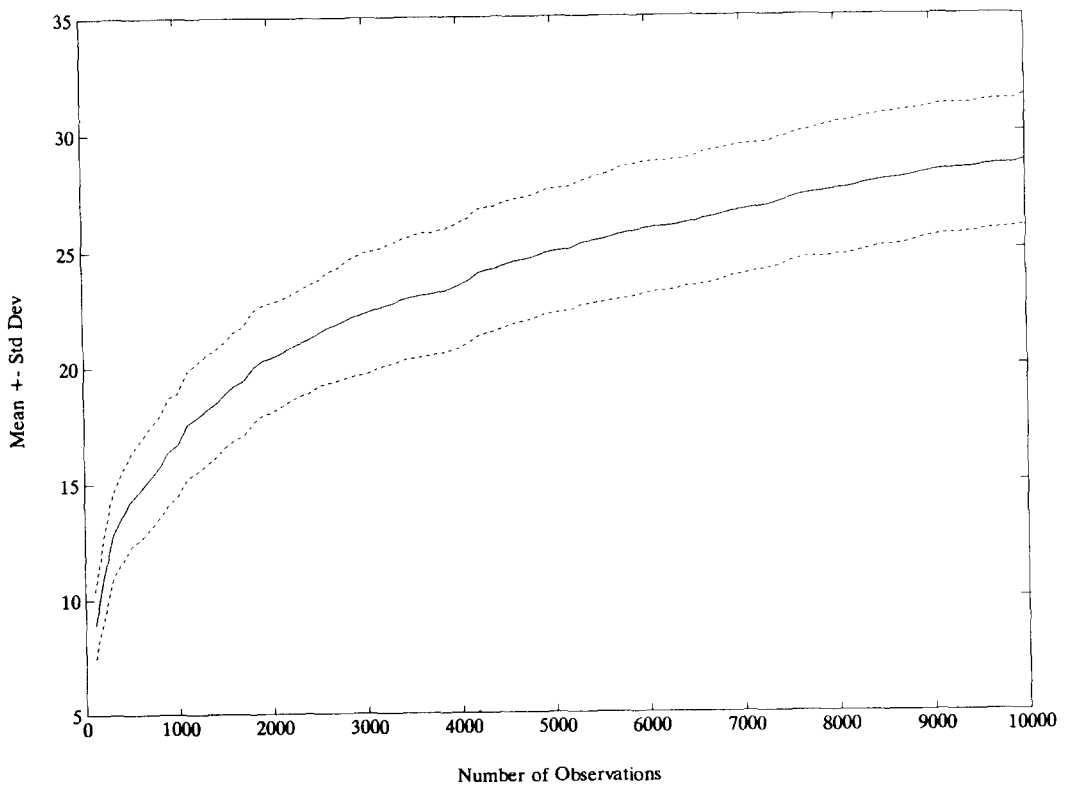


(b)

Fig. 3. $\alpha_0 = x^{-1}(2\pi)^{-1/2}\exp(-1/2(\ln x)^2) - 5$ for $0 < x < \infty$. (a) The true log-normal distribution α_0 from which the observations are drawn for Simulation 3. (b) L_1 convergence in mean and standard deviation based on 20 runs of 10,000 observations. The mean and standard deviation at $n = 10{,}000$ are $\mu = 0.1359$ and $\sigma = 0.0288$. To consider an extension of this simulation, we note that, based on 20 runs of 100,000 observations, we obtain $\mu = 0.1197$ and $\sigma = 0.0238$. (c) L_2 convergence in mean and standard deviation based on 20 runs of 10,000 observations. The mean and standard deviation at $n = 10{,}000$ are $\mu = 0.0127$ and $\sigma = 0.0055$. To consider an extension of this simulation, we note that, based on 20 runs of 100,000 observations, we obtain $\mu = 0.0107$ and $\sigma = 0.0047$. (d) Computational complexity in mean and standard deviation. The plot is number of terms in the model vs. number of observations based on 20 runs of 10,000 observations. The mean and standard deviation at $n = 10{,}000$ are $\mu = 28.75$ and $\sigma = 2.8011$. To consider an extension of this simulation, we note that, based on 20 runs of 100,000 observations, we obtain $\mu = 46.9$ and $\sigma = 3.3071$. (e) An example of the model produced after 100,000 observations. The solid curve is the true distribution $\alpha_0 = x^{-1}(2\pi)^{-1/2}\exp(-1/2(\ln x)^2) - 5$ for $0 < x < \infty$, the dashed line is the estimate α^*. To put this particular model in the perspective of the curves shown in (b)–(d), we note that this model has 46 terms, an ISE of 0.005811, and an IAE of 0.093628. (f) A preliminary example of the data-driven smoothing, produced after 1000 observations. The solid curve is the true distribution $\alpha_0 = x^{-1}(2\pi)^{-1/2}\exp(-1/2 \times (\ln x)^2) - 5$ for $0 < x < \infty$, the dashed lines are the three major terms in the estimate α^*. Note the correspondence between the variances in the model terms and the character of the true distribution near the mean of the terms.

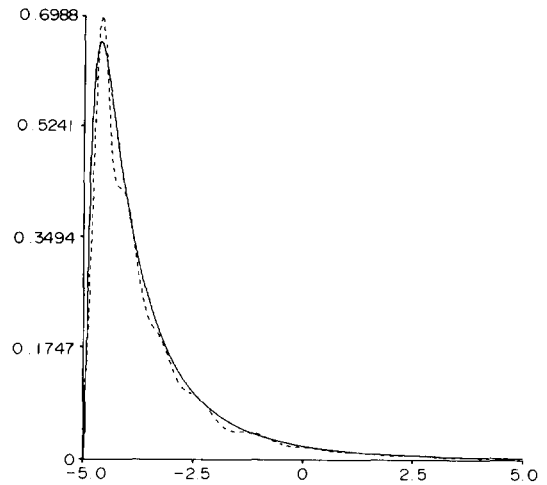


(c)

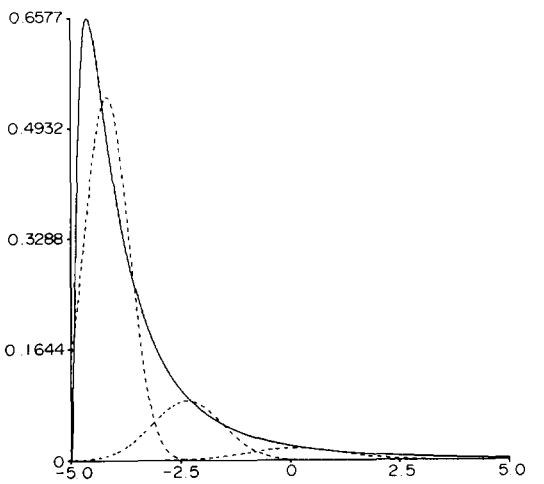


(d)

Fig. 3. (*Continued.*)



(e)



(f)

Fig. 3. (*Continued.*)

sively developing the estimate. When considered as a recursive instantiation of the method of sieves, adaptive mixtures provides a data-driven relaxation of the sieve constraints. One might wish to view this procedure in conjunction with a sieve approach described in Wegman[28] in which the sieve depends on the random sample itself.

Considered as a combination of sieves and an extension of a recursive version of the EM algorithm, adaptive mixtures as a maximum likelihood estimator shows potential for both small- and large-sample estimation applications. The main applications of the procedure, however, should come in large-sample problems for which iterative techniques are not feasible. In view of the discussion in Section 6 of Geman and Hwang,[26] an advantage of the adaptive mixture procedure is that the individual masses π_i and variances σ_i^2 yield more flexibility. In some real sense, we have a more powerful sieve.

Similarities between adaptive mixtures and maximum penalized likelihood are clear. The slow addition of mixture terms corresponds to a desire to keep the estimator simple (and in some sense smooth). That is, $P(\cdot)$ acts as a penalty function.

Regardless of asymptotic performance, any recursive estimation procedure will face a susceptibility to data-ordering problems for small-sample applications. While it must be remembered that adaptive mixtures are developed as an approach to problems in which a recursive solution is desired, there is an obvious iterative procedure which can be formulated in an analogous manner.

It is argued that the ability to model a rich class of densities, inherent in the adaptive mixture procedure, provides more powerful pattern recognition capabilities than simple parametric approaches. Modeling complicated probability density functions can translate into sophisticated discriminant procedures. Preliminary work in this area has appeared previously.[37] Modeling complex discriminant surfaces is a major focus of the neural network field.[39] The relationship between the techniques described above and neural networks is easily seen. In particular, kernel estimators[40] and finite mixture models[41] have been discussed from the perspective of artificial neural implementation, as has adaptive mixtures.[42] The method of adding terms to our model corresponds to the "automatic node creation" receiving attention in neural networks.

Among the ongoing work with adaptive mixtures, we have the extension of the method to the multi-class discrimination problem, unsupervised learning problems and situations in which the stochastic process X is not assumed to be stationary.

REFERENCES

1. C. T. Wolverton and T. J. Wagner, Asymptotically optimal discriminant functions for pattern classification, *IEEE Trans. Inf. Theory* **IT-15**, 258–265 (1969).
2. S. Yakowitz, A consistent estimator for the identification of finite mixtures, *Ann. Math. Statist.* **40**, 1728–1735 (1969).
3. S. Yakowitz, Unsupervised learning and the identification of finite mixtures, *IEEE Trans. Inf. Theory* **16**, 330–338 (1970).
4. D. M. Titterington, A. F. M. Smith and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York (1985).
5. E. J. Wegman and H. I. Davies, Remarks on some recursive estimators of a probability density, *Ann. Statist.* **7**, 316–327 (1979).
6. R. A. Tapia and J. R. Thompson, *Nonparametric Probability Density Estimation*. Johns Hopkins University Press, Baltimore (1978).
7. B. L. S. Prakasa Rao, *Nonparametric Functional Estimation*. Academic Press, Orlando (1983).
8. L. Devroye and L. Gyorfi, *Nonparametric Density Estimation: The L_1 View*. Wiley, New York (1985).

9. C. E. Priebe and D. J. Marchette, Adaptive mixtures: recursive nonparametric pattern recognition, *Pattern Recognition* **24**, 1197–1209 (1991).
10. E. Parzen, On the estimation of a probability density and mode, *Ann. Math. Statist.* **33**, 1065–1076 (1962).
11. B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London (1986).
12. B. S. Everitt and D. J. Hand, *Finite Mixture Distributions*. Chapman & Hall, London (1981).
13. G. J. McLachland and K. E. Basford, *Mixture Models*. Marcel Dekker, New York (1988).
14. M. A. Aizerman, E. M. Braverman and L. T. Rozonoer, The probability problem of pattern recognition learning and the method of potential functions, *Automn Remote Control* **26**, 1175–1190 (1966).
15. K. Fukunaga and R. R. Hayes, The reduced parzen classifier, *IEEE Trans. Pattern Analysis Mach. Intell.* **2**, 423–425 (1989).
16. I. J. Good and R. A. Gaskins, Nonparametric roughness penalties for probability densities, *Biometrika* **58**, 255–277 (1971).
17. T. Y. Young and T. W. Calvert, *Classification, Estimation and Pattern Recognition*. Elsevier, New York (1974).
18. M. B. Nevel'son and R. Z. Has'minskii, *Stochastic Approximation and Recursive Estimation*, Translations of Mathematical Monographs, Vol. 47. American Mathematical Society, Rhode Island (1973).
19. D. M. Titterington, Recursive parameter estimation using incomplete data, *J. R. Statist. Soc. Ser. B* **46**, 257–267 (1984).
20. A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. R. Statist. Soc. Ser. B* **39**, 1–38 (1977).
21. R. A. Redner and H. F. Walker, Mixture densities, maximum likelihood and the EM algorithm, *SIAM Rev.* **26** (1984).
22. J. Van Ryzin, On the strong consistency of density estimates, *Ann. Math. Statist.* **40**, 1765–1772 (1969).
23. H. Yamato, Sequential estimation of a continuous probability density function and the mode, *Bull. Math. Statist.* **14**, 1–12 (1971).
24. L. P. Devroye, On the pointwise and the integral convergence of recursive kernel estimates of probability densities, *Utilitas Math.* **15**, 113–128 (1979).
25. U. Grenander, *Abstract Inference*. Wiley, New York (1981).
26. S. Geman and C. Hwang, Nonparametric maximum likelihood estimation by the method of sieves, *Ann. Statist.* **10**, 401–414 (1982).
27. S. Geman, Sieve for nonparametric estimation of densities and regressions, Reports in Pattern Analysis, No. 99, D.A.M., Brown University (1981).
28. E. J. Wegman, Maximum likelihood estimation of a probability density function, *Sankhyā, Ser. A* **37**, 211–224 (1975).
29. T. Y. Young and G. Coraluppi, Stochastic estimation of a mixture of normal density functions using an information criterion, *IEEE Trans. Inf. Theory* **16**, 258–263 (1970).
30. P. Hall, On Kullback–Leibler loss and density estimation, *Ann. Statist.* **15**, 1491–1519 (1987).
31. A. Wald, Note on the consistency of the maximum likelihood estimate, *Ann. Math. Statist.* **20**, 595–601 (1949).
32. H. Cramèr, *Mathematical Methods of Statistics*. Princeton University Press, Princeton, New Jersey (1966).
33. R. R. Bahadur, Rates of convergence of estimates and test statistics, *Ann. Math. Statist.* **38**, 303–324 (1967).
34. C. J. Stone, Optimal rates of convergence for nonparametric estimators, *Ann. Statist.* **8**, 1348–1360 (1980).
35. G. Wahba, Data-based optimal smoothing of orthogonal series density estimates, *Ann. Statist.* **9**, 146–156 (1981).
36. M. P. Wand, J. S. Marron and D. Ruppert, Transformation in density estimation, *J. Am. Statist. Assoc.* **86**, 343–361 (1991).
37. J. L. Solka, C. E. Priebe and G. W. Rogers, An initial assessment of discriminant surface complexity for power law features, *Simulation*, May (1992).
38. C. E. Priebe, J. L. Solka, J. B. Ellis and G. W. Rogers, On discriminant surface complexity for power law features (in preparation).
39. P. A. Shoemaker, M. J. Carlin, R. L. Shimabukuro and C. E. Priebe, Least-squares learning and approximation of posterior probabilities on classification problems by neural networks, *Proc. 2nd Wkshop Neural Networks*, SPIE Vol. 1515 (1991).
40. D. F. Specht, Probabilistic neural networks, *Neural Networks* **3**, 109–118 (1990).
41. L. I. Perlovsky and M. M. McManus, Maximum likelihood neural networks for sensor fusion and adaptive classification, *Neural Networks* **4**, 89–102 (1991).
42. D. J. Marchette and C. E. Priebe, The adaptive kernel neural network, *Math. Comput. Modelling* **14**, 328–333 (1990).

About the Author—CAREY E. PRIEBE received his B.S. degree in mathematics from Purdue University in 1984, his M.S. degree in computer science from San Diego State University in 1988 and his Ph.D. in information technology (computational statistics) at George Mason University in 1993. Since 1985, Mr. Priebe has been working in adaptive systems and recursive estimators, first for the Naval Ocean Systems Center, San Diego, California, and since April 1991, with the Naval Surface Warfare Center, Dahlgren, Virginia.

About the Author—DAVID J. MARCHETTE received his B.S. and M.S. degrees in mathematics from the University of California, San Diego, in 1980 and 1982, respectively. Since 1985, Mr Marchette has been working for the Naval Ocean Systems Center, San Diego, in the field of pattern recognition.