The Annals of Statistics
2001, Vol. 29, No. 5, 1281–1296

CONSISTENT ESTIMATION OF MIXTURE COMPLEXITY[1]

BY LANCELOT F. JAMES, CAREY E. PRIEBE AND DAVID J. MARCHETTE

Johns Hopkins University and Naval Surface Warfare Center

The consistent estimation of mixture complexity is of fundamental importance in many applications of finite mixture models. An enormous body of literature exists regarding the application, computational issues and theoretical aspects of mixture models when the number of components is known, but estimating the unknown number of components remains an area of intense research effort. This article presents a semiparametric methodology yielding almost sure convergence of the estimated number of components to the true but unknown number of components. The scope of application is vast, as mixture models are routinely employed across the entire diverse application range of statistics, including nearly all of the social and experimental sciences.

1. Introduction. Let $\phi(x;\theta)$ be the normal density function with parameters $\theta = (\mu, \sigma^2) \in \Theta = \Re \times (0, \infty)$ and consider the generalized mixture

$$f(x;\eta) = \int \phi(x;\theta)\, d\eta(\theta),$$

where η is the mixing distribution. If the mixing distribution is finite, then f is a finite mixture model

$$f(x;\eta) = \sum_{t=1}^{m} \pi_t \phi(x;\theta_t)$$

and $m < \infty$ is the mixture complexity. Given a random sample $\mathbf{X}_n = (X_1, \ldots, X_n)$ drawn from f, this article proposes a semiparametric density estimator of the form

$$\hat{f}(x;\mathbf{X}_n) = \sum_{t=1}^{\hat{m}} \hat{\pi}_t \phi(x;\hat{\theta}_t)$$

with the property that

$$\hat{m} \to m \quad \text{a.s. as } n \to \infty.$$

That is, our methodology results in a consistent estimate of mixture complexity. If f is indeed a finite mixture of m normals ($f \in \mathscr{F}_m$), then the resultant estimate converges to the correct mixture representation (up to relabeling of the components). If, on the other hand, f is not an element of $\mathscr{F}_m$ for any m, then $\hat{m} \to \infty$ but $\hat{f} \to f$ nonetheless.

Received December 1999; revised May 2001.
[1]Supported in part by Office of Naval Research Grant N00014-95-1-0777.
AMS 2000 *subject classifications*. Primary 62G05; secondary 62G07.
Key words and phrases. Finite mixture model, number of components, semiparametric.

Methodologies for consistent estimation of a mixing distribution are well known; for example, the nonparametric maximum likelihood estimate of η is consistent [see, e.g., Kiefer and Wolfowitz (1956); Pfanzagl (1988); Leroux (1992)]. However, the literature regarding consistent estimation of the mixture complexity m is sparse but burgeoning: see Henna (1985); Chen and Kalbfleisch (1996); Dacunha-Castelle and Gassiat (1997, 1999); Keribin (2000); and Priebe and Marchette (2000). Efforts to address the related problem of testing hypotheses about m have met with mixed results; the limiting distribution for the likelihood ratio test statistic was until recently unavailable [Dacunha-Castelle and Gassiat (1999)], and so bootstrap testing methodologies have been developed [see, e.g., McLachlan (1987)]. Finally, in nonparametric Bayesian density estimation, posterior consistency for the number of components can be established for mixtures of normals using Dirichlet process priors [see, e.g., Escobar and West (1995)] or by the method of Roeder and Wasserman (1997).

Some preliminaries regarding mixture models and kernel density estimation are given in Section 2. In Section 3 we develop our alternating kernel and mixture semiparametric density estimation procedure. In particular, equation (3) gives our estimator $\hat{m}_n$ for mixture complexity. The main result, consistency of $\hat{m}_n$, is presented in Section 4. We conclude, in Section 5, with simulation experiments and an example of application.

2. Preliminaries. Let

$$\mathscr{F} = \bigcup_{m=1}^{\infty} \mathscr{F}_m, \mathscr{F}_m \subset \mathscr{F}_{m+1} \ \forall\, m$$

denote the family of normal mixtures. That is, for each fixed $m < \infty$,

$$\mathscr{F}_m = \{f(\cdot\,; m, \pi, \theta)\colon (\pi, \theta) \in \Pi_m \times \Theta_m\}$$

where

$$f(x; m, \pi, \theta) = \sum_{t=1}^{m} \pi_t \phi(x; \theta_t),$$

$$\Pi_m = \left\{\pi_t, t = 1, \ldots, m\colon \sum_{t=1}^{m} \pi_t = 1,\ \pi_t \geq 0\right\}$$

and

$$\Theta_m = \{(\mu_t, \sigma_t^2),\ t = 1, \ldots, m\colon (\mu_t, \sigma_t^2) \in \Theta = \Re \times (0, \infty)\}.$$

Note that $\mathscr{F}_1$ corresponds to the family of univariate normal densities. We hereafter denote the vector of $(3m - 1)$ parameters as

$$v_m = (\pi_1, \ldots, \pi_{m-1}, \mu_1, \sigma_1^2, \ldots, \mu_m, \sigma_m^2)$$

and write, for elements of $\mathscr{F}_m$,

$$f(x; v_m) = \sum_{t=1}^{m} \pi_t \phi_{\sigma_t}(x - \mu_t),$$

where

$$\phi_\sigma(z) = \sigma^{-1}\phi(\sigma^{-1}z),$$

in which $\phi(z)$ denotes the standard normal density, and

$$\pi_m = 1 - \sum_{t=1}^{m-1} \pi_t.$$

Note that the nonparametric normal kernel density estimator with bandwidth $h > 0$, defined as

$$\tilde{f}_h(x) = (1/n)\sum_{i=1}^{n} \phi_h(x - X_i),$$

is an element of $\mathscr{F}_n$.

For each fixed $m < \infty$ we say that two parameters v_m and $\tilde{v}_m$ are equivalent in the sense of the quotient topology [see Redner (1981), Redner and Walker (1984)] if they define the same density; that is, if v_m and $\tilde{v}_m$ are both elements of a set

$$T(v) = \{v' \colon f(x; v') = f(x; v) \quad \forall\, x\}.$$

Given an arbitrary density g, we define the *index of the economical representation of* g, relative to the family of normal mixtures, as

$$m(g) = \min\{m \colon g \in \mathscr{F}_m\}.$$

Thus if g is a finite mixture of normals then $m(g)$ is finite and denotes the mixture complexity. That is, if $m(g) = m$, then there exists a vector v_m such that

$$g(x) = f(x; v_m),$$

but there does not exist v_j for any $j < m$ such that

$$g(x) = f(x; v_j);$$

clearly for all $k > m$ there exist v_k such that

$$g(x) = f(x; v_k).$$

Thus $m(g)$ represents the index of the most parsimonious normal mixture model representation for g. Naturally, if g is not a finite normal mixture then $m(g) = \infty$.

3. Estimation procedure. Given $\mathbf{X}_n = (X_1, \ldots, X_n)$ independent and identically distributed from an unknown density g_0, we devise an iterative estimation scheme to estimate $m_0 = m(g_0)$. Let

$$\mathrm{KL}(g, f) = \int g(x) \ln\left(\frac{g(x)}{f(x)}\right) dx$$

denote the Kullback–Leibler distance between two densities g and f. Let $\tilde{f}_h$ be the normal kernel density estimator as above and define for an integer $m > 0$,

$$\hat{g}^m = \arg \min_{f \in \mathscr{F}_m} \mathrm{KL}(\tilde{f}_h, \phi_h * f) \tag{1}$$

and

$$g_0^m = \arg \min_{f \in \mathscr{F}_m} \mathrm{KL}(\phi_h * g_0, \phi_h * f), \tag{2}$$

where g_0 denotes the true underlying density and $*$ denotes the convolution operator. Note that for $f \in \mathscr{F}_m$, the density $\phi_h * f$ is a normal mixture model with variance components $\sigma_t^2 + h^2$, $t = 1, \ldots, m$. Our initial motivation for convolving elements of $\mathscr{F}_m$ with ϕ_h is simple; the kernel estimator $\tilde{f}_h$ is an unbiased estimator for $\phi_h * g_0$, rather than for g_0. Moreover, if one considers the case of $m = 1$—that is, the class of univariate normal densities with mean μ and variance σ^2—then

$$\hat{g}^1(x) = \phi(x; (\hat{\mu}, \hat{\sigma}^2)),$$

where $(\hat{\mu}, \hat{\sigma}^2)$ are the well-known maximum likelihood estimators. Thus $(\hat{\mu}, \hat{\sigma}^2)$ are root n-consistent provided that g_0 has a finite second moment, and efficient if $m_0 = 1$. In contrast, if one were to use

$$\tilde{g}^1 = \arg \min_{f \in \mathscr{F}_m} \mathrm{KL}(\tilde{f}_h, f),$$

then

$$\tilde{g}^1(x) = \phi(x; (\hat{\mu}, \hat{\sigma}^2 + h^2)),$$

which, when $m_0 = 1$, produces an asymptotically biased estimator for σ^2 with bias of order h^2. The estimator (1) is similar to the method of Beran (1977), specialized to a finite mixture model. However Beran's (1997) method is based on the Hellinger distance. Other work which specifically considers the case of a fixed finite mixture model using a Hellinger or other mimimum distance method includes Cutler and Cordero-Braña (1996), Cordero-Braña and Cutler (2001), Cao, Cuevas and Fraiman (1995), Cao and Devroye (1996), Clarke and Heathcote (1994), Tamura and Boos (1986).

We now present a brief comparison showing the relationship between our convolution approach and penalized maximum likelihood. Notice that, for $f \in \mathcal{F}_m$,

$$\int \tilde{f}_h(x) \ln \phi_h * f(x; v_m)\, dx$$

$$= \sum_{i=1}^{n} \frac{1}{n} \int \phi(z) \ln\left(\sum_{t=1}^{m} \pi_t \phi(X_i + hz; (\mu_t, \sigma_t^2 + h^2)) \right) dz.$$

A Taylor expansion about $z = 0$ gives

$$\int \tilde{f}_h(x) \ln \phi_h * f(x; v_m)\, dx$$

$$\cong \sum_{i=1}^{n} \frac{1}{n} \ln\left(\sum_{t=1}^{m} \pi_t \phi(X_i; (\mu_t, \sigma_t^2 + h^2)) \right) + \frac{1}{2} h^2 G_n \Psi_h,$$

where G_n is the empirical measure on the X_i's and

$$\Psi_h = \frac{\phi_h * f''}{\phi_h * f} - \left[\frac{\phi_h * f'}{\phi_h * f} \right]^2.$$

It follows that our maximization procedure may be viewed as a variant of a penalized m.l.e. (with a random penalty function). [See Bickel, Klaassen, Ritov and Wellner (1993) for some references on this subject.]

Our estimator $\hat{m}_n$ for m_0 is defined as

$$\hat{m}_n = \min\{m\colon \mathrm{KL}(\tilde{f}_h, \phi_h * \hat{g}^m) \le \mathrm{KL}(\tilde{f}_h, \phi_h * \hat{g}^{m+1}) + a_{n,m+1}\}, \tag{3}$$

where $\{a_{n,j}\colon j \ge 1\}$ are positive sequences chosen such that they converge to zero as n increases to ∞. We take $a_{n,m} = 3/n$ in practice.

One sees that our choice of $\hat{m}_n$ is natural by noting that m_0 can be expressed as

$$\begin{aligned} m_0 &= \min\{m\colon \mathrm{KL}(\phi_h * g_0, \phi_h * g_0^m) \le \mathrm{KL}(\phi_h * g_0, \phi_h * g_0^{m+1})\} \\ &= \min\{m\colon \mathrm{KL}(\phi_h * g_0, \phi_h * g_0^m) = 0\} \\ &= \min\{m\colon \mathrm{KL}(g_0, g_0^m) = 0\}. \end{aligned}$$

Our estimator $\hat{m}_n$ also bears some similarities to one developed in Section 3 of Ritov and Bickel (1990), although their estimator appears in a somewhat different context. In practice we may wish to choose h based upon the optimal bandwidth for the estimated $\hat{g}^m$; as such, the scalar h is replaced by a random h_m at each stage. In Section 5 we will demonstrate the performance of this random bandwidth approach.

4. Main result. In this section we show that consistency of our method essentially follows from consistency of the nonparametric kernel density estimator. To support our main results, for bandwidths which may be random, we present the following lemma (obtained under minimal conditions) which is deduced from an application of Nolan and Marron (1989).

LEMMA 1. *Let h be a bandwidth satisfying $h \to 0$ and $nh/\log n \to \infty$. Then for g_0 such that $\mathrm{KL}(\tilde{f}_h, \phi_h * g_0) < \infty$,*

$$\mathrm{KL}(\tilde{f}_h, \phi_h * g_0) \to 0 \quad a.s.$$

In addition, if h is replaced by a random bandwidth $h(x;n)$, possibly depending on x, such that there exists positive sequences $\alpha_n \le \beta_n$ satisfying $\beta_n \ge h(x;n) \ge \alpha_n$ for all x, eventually almost surely, $\beta_n \to 0$ and $n\alpha_n/\log n \to \infty$, then

$$\mathrm{KL}(\tilde{f}_{h(x;n)}, \phi_{h(x;n)} * g_0) \to 0 \quad a.s.$$

PROOF. Since the class of translates of the standard normal kernel, $\{h\phi_h(x-\cdot): x \in \Re, h > 0\}$, constitutes a Euclidean class, we have by Theorem 1 of Nolan and Marron (1989) [see also Pollard (1984), page 35] that

$$\sup_x \left| \frac{\tilde{f}_h(x)}{\phi_h * g_0(x)} - 1 \right| \to 0 \quad \text{a.s.}$$

The analogous result for $h(x;n)$ follows from Corollary 2 of Nolan and Marron (1989). These results, coupled with the fact that

$$\begin{aligned} \mathrm{KL}(\tilde{f}_h, \phi_h * g_0) &= \int \tilde{f}_h(x) \ln\left(\frac{\tilde{f}_h(x)}{\phi_h * g_0(x)} \right) dx \\ &\le \int \tilde{f}_h(x) \left[\frac{\tilde{f}_h(x)}{\phi_h * g_0(x)} - 1 \right] dx, \end{aligned}$$

conclude the proof. □

THEOREM 1. *Suppose that n and h satisfy the conditions in Lemma 1. Then for any sequence $a_{n,m} \to 0$,*

$$\hat{m}_n \to m_0 \quad a.s.$$

PROOF. The proof follows by showing that for each $m > 0$,

$$\begin{aligned} &\mathrm{KL}(\tilde{f}_h, \phi_h * \hat{g}^m) - \mathrm{KL}(\tilde{f}_h, \phi_h * \hat{g}^{m+1}) \\ &\quad = \int \tilde{f}_h(x) \ln\left(\frac{\phi_h * \hat{g}^{m+1}(x)}{\phi_h * \hat{g}^m(x)} \right) dx \to c_m \quad \text{a.s.}, \end{aligned}$$

where $c_m > 0$ for $m < m_0$ and $c_m = 0$ for $m \ge m_0$. This is established in Lemmas 2, 3, 4 below. □

LEMMA 2. *For $m > 0$ an arbitrary integer,*

$$\int \tilde{f}_h(x) \ln\left(\frac{\phi_h * \hat{g}^m(x)}{\phi_h * g_0^m(x)} \right) dx \to 0 \quad a.s.$$

as $n \to \infty$.

PROOF. From the definitions of $\hat{g}^m$ and g_0^m in (1) and (2) it follows that

$$\mathrm{KL}(\tilde{f}_h, \phi_h * g_0^m) - \mathrm{KL}(\tilde{f}_h, \phi_h * \hat{g}^m) = \int \tilde{f}_h(x) \ln\left(\frac{\phi_h * \hat{g}^m(x)}{\phi_h * g_0^m(x)}\right) dx \geq 0$$

and similarly

$$\begin{aligned}\mathrm{KL}(\phi_h * g_0, \phi_h * g_0^m) - \mathrm{KL}(\phi_h * g_0, \phi_h * \hat{g}^m) \\ = \int \phi_h * g_0(x) \ln\left(\frac{\phi_h * \hat{g}^m(x)}{\phi_h * g_0^m(x)}\right) dx \leq 0.\end{aligned}$$

However, from the results in Lemma 1, it follows that the two quantities are asymptotically equal almost surely and hence they must both almost surely converge to zero as $n \to \infty$. □

LEMMA 3. *If $m \geq m_0$, then*

$$\int \tilde{f}_h(x) \ln\left(\frac{\phi_h * \hat{g}^{m+1}(x)}{\phi_h * \hat{g}^m(x)}\right) dx \to 0 \quad a.s.$$

as $n \to \infty$.

PROOF. Notice that

$$\begin{aligned}\int \tilde{f}_h(x) \ln\left(\frac{\phi_h * \hat{g}^{m+1}(x)}{\phi_h * \hat{g}^m(x)}\right) dx = \int \tilde{f}_h(x) \ln\left(\frac{\phi_h * \hat{g}^{m+1}(x)}{\phi_h * g_0(x)}\right) dx \\ - \int \tilde{f}_h(x) \ln\left(\frac{\phi_h * \hat{g}^m(x)}{\phi_h * g_0(x)}\right) dx.\end{aligned}$$

Recall that $g_0 \in \mathscr{F}_{m_0}$ implies $g_0 \in \mathscr{F}_j$ for $j \geq m_0$ and hence using (1) it follows that

$$\begin{aligned}\int \tilde{f}_h(x) \ln\left(\frac{\phi_h * \hat{g}^j(x)}{\phi_h * g_0(x)}\right) dx \\ = \mathrm{KL}(\tilde{f}_h, \phi_h * g_0) - \mathrm{KL}(\tilde{f}_h, \phi_h * \hat{g}^j) \geq 0.\end{aligned}$$

Thus since $\mathrm{KL}(\tilde{f}_h, \phi_h * \hat{g}^j) \geq 0$, it follows that

$$0 \leq \mathrm{KL}(\tilde{f}_h, \phi_h * \hat{g}^j) \leq \mathrm{KL}(\tilde{f}_h, \phi_h * g_0)$$

and hence by Lemma 1, $\mathrm{KL}(\tilde{f}_h, \phi_h * \hat{g}^j) \to 0$ a.s. as $n \to \infty$. □

LEMMA 4. *If $m < m_0$, then for some positive constant c_m,*

$$\int \tilde{f}_h(x) \ln\left(\frac{\phi_h * \hat{g}^{m+1}(x)}{\phi_h * \hat{g}^m(x)}\right) dx \to c_m \quad a.s.$$

PROOF. Notice that

$$\int \tilde{f}_h(x) \ln\left(\frac{\phi_h * \hat{g}^{m+1}(x)}{\phi_h * \hat{g}^m(x)}\right) dx = \int \tilde{f}_h(x) \ln\left(\frac{\phi_h * \hat{g}^{m+1}(x)}{\phi_h * g_0^{m+1}(x)}\right) dx + \int \tilde{f}_h(x) \ln\left(\frac{\phi_h * g_0^m(x)}{\phi_h * \hat{g}^m(x)}\right) dx + \int \tilde{f}_h(x) \ln\left(\frac{\phi_h * g_0^{m+1}(x)}{\phi_h * g_0^m(x)}\right) dx \geq 0.$$

By Lemma 2, the first two terms on the r.h.s. of the equality converge almost surely to zero, which implies that for large enough n,

$$\int \tilde{f}_h(x) \ln\left(\frac{\phi_h * g^{m+1}(x)}{\phi_h * g_0^m(x)}\right) dx \geq 0 \quad \text{a.s.}$$

Moreover, it follows from Lemma 1 that

$$\left| \int [\tilde{f}_h(x) - \phi_h * g_0(x)] \ln\left(\frac{\phi_h * g_0^{m+1}(x)}{\phi_h * g_0^m(x)}\right) dx \right| \to 0 \quad \text{a.s.}$$

An application of Leroux [(1992), Lemma 3] gives

$$\int \phi_h * g_0(x) \ln\left(\frac{\phi_h * g_0^{m+1}(x)}{\phi_h * g_0^m(x)}\right) dx \to c_m$$

for some $c_m > 0$, completing the proof. □

5. Computational experiments. The development above gives rise to an iterative algorithm described as follows. First one fits a nonparametric estimator $\tilde{f}$ to the data, and computes the KL distance between $\tilde{f}$ and a single normal. A component is added (yielding a mixture of two normal components at this first stage) and the mixture ($\hat{g}^2$) is updated in such a way as to satisfy equation (1). The change in KL distance is computed as in equation (3) and compared with the threshold $a_{n,2}$. This process repeats, adding more components to the mixture, until the change is less than a, at which time the procedure terminates.

We present three examples of this algorithm in action. The first is a Monte Carlo simulation demonstrating the performance on a given target density over a variety of sample sizes. The second is a Monte Carlo simulation for a fixed sample size on a variety of target densities taken from Marron and Wand (1992). The final example involves the income data set from Marron and Schmitz (1992).

Among the numerous implementation details which must be considered in the course of realizing the above algorithm, the precise nature of the nonparametric estimator $\tilde{f}$ is an important issue. One choice for $\tilde{f}$ is a kernel estimator with the bandwidth chosen initially (as a function of n) and fixed throughout. Alternatively, the bandwidth can be updated at each iteration of the algorithm based on the current best-fit mixture, using the mixture to improve the fit of

the kernel estimator. In the simulations below we explore these two choices. In one case we use the normal reference rule described in Silverman (1986); we denote this estimator the NKE (for "normal reference rule kernel estimator"). We compare the performance of the algorithm based on the NKE with that of the "mixture reference rule kernel estimator" MKE. That is, we choose the bandwidth for the kernel estimator to be "optimal" (in the approximate mean integrated squared error sense) for the current mixture estimate, rather than a fixed density.

A slight modification is in order for the MKE. Intuitively, we want to compare the best m component mixture with the best $m+1$ component mixture. We modify equations (1) and (3) as follows:

$$\hat{g}^{m+1} = \arg \min_{f \in \mathscr{F}_{m+1}} \mathrm{KL}(\tilde{f}_{h_m}, \phi_{h_m} * f) \tag{4}$$

and

$$\tilde{g}^{m} = \arg \min_{f \in \mathscr{F}_{m}} \mathrm{KL}(\tilde{f}_{h_m}, \phi_{h_m} * f). \tag{5}$$

The estimate of mixture complexity is now given by

$$\hat{m}_n = \min\{m\colon \mathrm{KL}(\tilde{f}_h, \phi_h * \tilde{g}^m) \le \mathrm{KL}(\tilde{f}_h, \phi_h * \hat{g}^{m+1}) + a_{n,m+1}\}. \tag{6}$$

Thus, at each iteration, $\hat{g}^m$ is used to obtain a new bandwidth h_m by utilizing the "optimal" bandwidth for the mixture $\hat{g}^m$, which in turn defines a new kernel estimator. Several methods for obtaining this bandwidth are possible; we choose a simple one based on minimizing the approximate mean integrated squared error. We then fit two new mixtures to this kernel estimator, one with m components and one with $m+1$ components, as in equations (4) and (5). Finally, we determine whether the change in the estimators is small, using equation (6). This modification has no effect on the theory presented earlier, but it does seem to improve the performance of the algorithm in simulation.

The choice of the threshold a is critical to the functioning of these algorithms. One basis for this model selection criterion is the minimum description length (MDL) penalty of Rissanen (1978); this choice is used throughout this section, and leads to the threshold $a_{n,m} = 3/n$ [see, e.g., Barron and Cover (1991)].

5.1. *Simulation experiment.* The target density for this simulation is the three component mixture

$$\begin{aligned} f(x) = (1/2)\phi(x; 0, 10) + (1/4)\phi(x; -0.3, 0.05) \\ +(1/4)\phi(x; 0.3, 0.05), \end{aligned} \tag{7}$$

a large variance normal (the first component) with two small variance components giving rise to dramatic modes. This bimodal mixture is chosen to be difficult for the normal reference rule kernel estimator; accurately detecting the two modes at $x = \pm 0.3$ comes at the cost of undersmoothing the tails.

This is also true for the MKE; however, changing the bandwidth according to the mixture allows the MKE to improve its fit.

We investigate the performance of the two algorithm implementations on sample sizes $n = 50, 250, 500, 1000$. For 100 Monte Carlo replications at each sample size, we tally the number of components chosen by each algorithm for each replication. Results are given in Table 1. For comparison, we provide results obtained via the bootstrapping procedure of McLachlan (1987) (*Bootstrap*), the CDF method of Henna (1985) (*Henna*), and the Bayesian methodology proposed by Roeder and Wasserman (1997) (*R&W*).

The simulation demonstrates that incorporating mixture reference rule bandwidth selection yields superior performance in correctly identifying the mixture. The NKE gives a relatively poor estimate of the mixture complexity

TABLE 1

Mixture complexity esimation results for Monte Carlo simulation [*Target mixture*, *equation* (7), *has three components*]

	Estimated number of components							
	1	2	3	4	5	6	7	8
$n = 50$								
NKE	44	56						
MKE	44	53	3					
R&W	22	7	59	10	1	1		
Bootstrap	0	96	4					
Henna	25	68	6	1				
$n = 250$								
NKE	0	99	1					
MKE	0	87	11	1	1			
R&W	0	0	60	22	18			
Bootstrap	0	83	16	1				
Henna	0	90	10					
$n = 500$								
NKE	0	97	3					
MKE	0	58	34	6	2			
R&W	0	0	22	12	61	5		
Boostrap	0	74	20	6				
Henna	0	85	15					
$n = 1000$								
NKE	0	86	14					
MKE	0	18	63	10	2	3	1	3
R&W	0	0	0	1	89	10		
Bootstrap	0	79	15	4	2			
Henna	0	78	15	5	1	0	1	

for these sample sizes, while the MKE does significantly better. For a sample size of $n = 50$, neither approach is accurate, while the *R&W* algorithm does quite well for $n = 50, 250$. For larger sample sizes, the MKE correctly identifies the complexity a substantial and increasing percentage of the time. (The MKE does occasionally overestimate the number of components; at $n = 1000$, 19 of 100 replicates yield $\hat{m} > 3$ while 18 of 100 replicates yield $\hat{m} < 3$.) The other two algorithms, *Bootstrap* and *Henna*, perform roughly comparably to the NKE version.

It is possible that the estimated number of components is a misleading measure of performance and does not accurately represent the quality of the estimator. For example, an estimate with two components in the tails and one for the two modes at $x = \pm 0.3$ should not be considered accurate despite having (by coincidence) the correct number of components. Figure 1 depicts the 63 three-component estimates obtained via MKE for $n = 1000$ and demonstrates that these estimates do indeed correctly identify the bimodal structure nearly every time. That is, the tabular results represent accurate estimation of mixture complexity.

5.2. *Marron and Wand mixtures.* Marron and Wand (1992) presented a set of 15 normal mixture densities which have come to be a standard set for comparison of mixture estimators. In this section we consider a comparison of the different algorithms on mixtures 2–10. The sample size for the study is $n = 1000$.

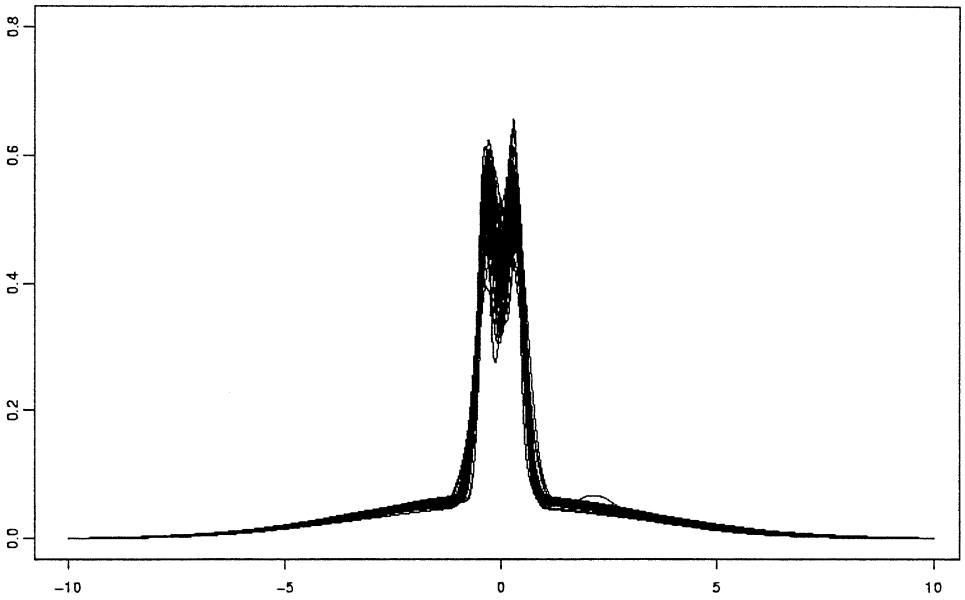

FIG. 1. *Simulation results indicate accurate estimation of mixture complexity for target density* (7). *Depicted are the* 63 *three-component estimates obtained via MKE for* $n = 1000$. *This figure demonstrates that these estimates do indeed correctly identify the bimodal structure.*

We first need to set the parameters of the algorithms. This is done by reference to the standard normal density (mixture 1 in Marron and Wand's set). The parameters were set so that in an experiment consisting of 100 replicates of size $n = 1000$ from a normal distribution, the algorithm returns an estimate of 1 for the number of components 95 times. This was then the choice of parameters used throughout the study.

The densities are depicted in Marron and Wand [(1992), pages 717 and 718]. These densities show a range of unimodal, skewed and multimodal densities that provide a good exercise for any algorithm. We drew 100 replicates of size 1000 from each density, and the results of running the algorithms on these samples is depicted in Table 2. The true model complexity is indicated by an asterisk for each model.

As may be expected, no algorithm outperforms all the others on all the densities, and each algorithm has at least one mixture for which its performance is as good or better than all the others. Our proposed method is not dominated by any of the alternatives and, from a practical point of view, is therefore a valuable addition to the practitioner's toolbox.

All the algorithms considered except the bootstrap have similar, and acceptable, computational requirements. The bootstrap, while it does a very good job on these examples, is quite computationally complex. This is due in part to the use of the EM algorithm, which in this implementation is run several times with different initial values in order to avoid local minima. For some of the models (in particular mixture 10), this resulted in an unacceptably long run time (several days on a dedicated machine to run the 100 simulations). This could be alleviated by using a less expensive estimation algorithm within the bootstrap. The drawback might be a less accurate estimator. We did not pursue this in this work.

5.3. *Example application: UK income data.* We now consider the application of our algorithm to the income data of Marron and Schmitz (1992). The data are from the ESCR Data Archive at the University of Essex: Family Expenditure Survey, covering the year 1975. The data are household incomes normalized by the arithmetic mean for the year. There are $n = 7201$ observations in this data set. More details, and the results of analysis performed on several years' data, can be found in the reference.

Figure 2 depicts the MKE estimate for this income data, along with an undersmoothed kernel estimator for comparison. We display the undersmoothed kernel estimator so that the fine structure of the data can be discerned. The NKE chooses a five-component estimate, while the complexity estimate for the depicted MKE model is $\hat{m} = 6$. (The NKE estimate was deemed inferior based on visual inspection.)

While our algorithm is a consistent estimator of mixture complexity, it is not designed as a density estimator and does not necessarily produce the best estimate of the mixture itself. Once the mixture complexity is determined, it is reasonable to estimate the mixture parameters by maximum likelihood, for instance. The MKE estimate depicted in Figure 2 is the result of

TABLE 2
Mixture complexity estimation results for the Marron and Wand densities, 2–10[a]

	Estimated number of components									
	1	2	3	4	5	6	7	8	9	10
Mixture 2										
NKE	0	99	1*							
MKE	0	99	1*							
R&W	3	96	1*							
Henna	0	100	*							
Boot	0	89	11*							
Mixture 3										
NKE	0	0	96	4				*		
MKE	0	1	54	37	8			*		
R&W	0	0	0	8	38	25	20	7*	2	
Henna	0	0	26	74				*		
Boot	0	0	0	17	59	21	2	1*		
Mixture 4										
NKE	0	99*	1							
MKE	0	91*	6	3						
R&W	0	0*	0	0	75	18	5	2		
Henna	0	88*	12							
Boot	0	95*	5							
Mixture 5										
NKE	0	96*	4							
MKE	0	91*	8	1						
R&W	0	55*	45							
Henna	1	97*	1	0	0	0	0	0	0	1
Boot	0	95*	5							
Mixture 6										
NKE	0	100*								
MKE	0	98*	2							
R&W	0	100*								
Henna		97*	3							
Boot	0	95*	5							
Mixture 7										
NKE	0	100*								
MKE	0	96*	4							
R&W	0	100*								
Henna	0	96*	4							
Boot	0	93*	6	1						

TABLE 2
(*Continued*)

	Estimated number of components									
	1	2	3	4	5	6	7	8	9	10
Mixture 8										
NKE	0	100*								
MKE	0	97*	3							
R&W	0	80*	20							
Henna	0	99*	1							
Boot	0	93*	7							
Mixture 9										
NKE	0	94	6*							
MKE	0	38	59*	2						
R&W	0	91	9*	1						
Henna	0	82	18*							
Boot	0	13	75*	12						
Mixture 10										
NKE	33	51	15	1		*				
MKE	33	13	3	6	1	42*	2			
R&W	15	0	0	0	0	0*	39	28	17	1
Henna	0	0	5	8	15	33*	14	9	10	6
Boot	5	28	15	21	11	11*	5	4		

*A indicates the correct number of components for the mixture.

an EM algorithm search for a maximum likelihood estimate based on the $n = 7201$ observations, using the mixture obtained via the "consistent estimation of mixture complexity" algorithm as the starting point for the EM algorithm.

6. Conclusions. We have described a method for the estimation of mixture complexity and showed its consistency. The method relies on comparing a nonparametric estimator with the best parametric fit of a given complexity. As shown by the simulations and example, this estimator is competitive with other existing techniques and is therefore a valuable addition to the practitioner's toolbox.

The simulations imply that the performance of the estimator is dependent on the quality of the nonparametric model. Thus, the performance is best when the nonparametric model is allowed to adapt its bandwidth using the parametric model. This suggests that the performance may be further improved by using a multiple bandwidth estimator such as the filtered kernel estimator [Marchette, Priebe, Rogers and Solka (1996); Priebe and Marchette (2000)] or variable kernel estimators. This is an area for further research.

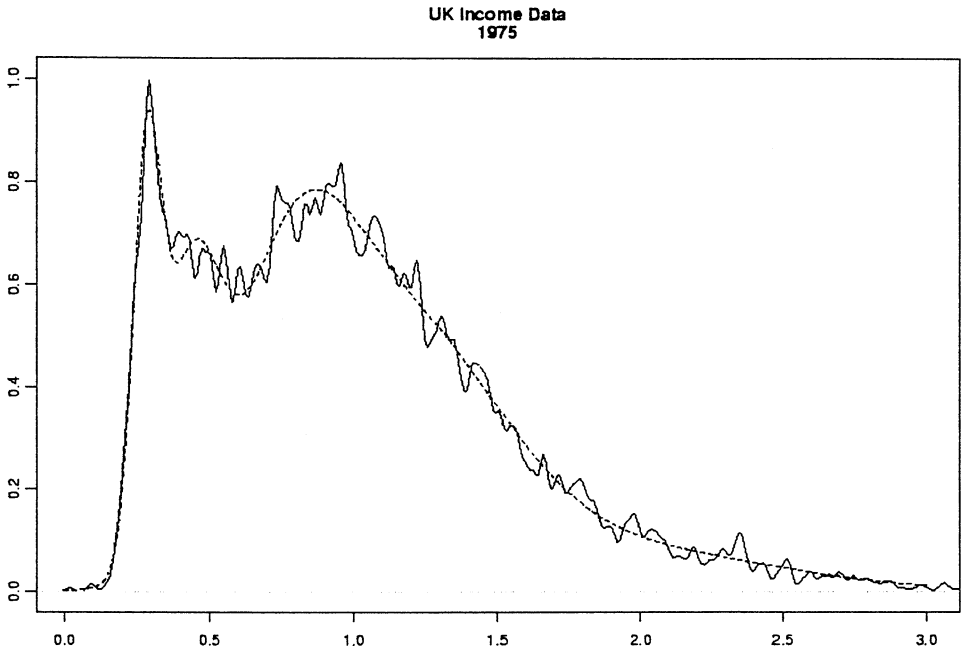

FIG. 2. *Experimental results indicate accurate estimation of mixture complexity for U.K. income data. Depicted are the six component MKE estimate (after application of the EM algorithm) and an undersmoothed kernel estimator for comparison.*

Acknowledgments. Constructive criticism provided by the Editor, Associate Editor and two referees improved the presentation of this material. The authors kindly thank Vince LaRiccia and Paul Eggermont for a helpful conversation.

REFERENCES

BARRON, A. R. and COVER, T. M. (1991). Minimum Hellinger distance estimates for parametric models. *IEEE Trans. Inform Theory* **37** 1034–1054.

BERAN, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5** 445–463.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press.

CAO, R., CUEVAS, A. and FRAIMAN, R. (1995). Minimum distance density-based estimation. *Comput. Statist. Data Anal.* **20** 611–631.

CAO, R. and DEVROYE, L. (1996). The consistency of a smoothed minimum distance estimate. *Scand. J. Statist.* **23** 405–418.

CHEN, J. and KALBFLEISCH. J. D. (1996). Penalized minimum distance estimates in finite mixture models. *Canad. J. Statist.* **24** 167–175.

CLARKE, B. R. and HEATHCOTE, C. R. (1994). Robust estimation of k-component univariate normal mixtures. *Ann. Inst. Statist. Math.* **46** 83–93.

CORDERO-BRAÑA, O. I and CUTLER, A. (2001). On the asymptotic properties of the minimum Hellinger estimation in the case of a mixture model. Research Report 7/01/104, Dept. Mathematics and Statistics, Utah State Univ.

CUTLER, A. and CORDERO-BRAÑA, O. I. (1996). Minimum Hellinger distance estimation for finite mixture models. *J. Amer. Statist. Assoc.* **91** 1716–1723.

DACUNHA-CASTELLE, D. and GASSIAT, E. (1997). The estimation of the order of a mixture model. *Bernoulli* **3** 279–299.

DACUNHA-CASTELLE, D. and GASSIAT, E. (1999). Testing the order of a model using locally conic parameterization: population mixtures and stationary ARMA processes. *Ann. Statist.* **27** 1178–1209.

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588.

HENNA, J. (1985). On estimating of the number of constituents of a finite mixture of continuous distributions. *Ann. Inst. Statist. Math.* **37** 235–240.

KERIBIN, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A* **62** 49–62.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906.

LEROUX, B. G. (1992). Consistent estimation of a mixing distribution. *Ann. Statist.* **20** 1350–1360.

MARCHETTE, D. J., PRIEBE, C. E., ROGERS, G. W. and SOLKA, J. L. (1996). The filtered kernel estimator. *Comp. Statist.* **11** 95–112.

MARRON, J. S. and SCHMITZ, H.-P. (1992). Simultaneous density estimation of several income distributions. *Econometric Theory* **8** 476–488.

MARRON, J. S. and WAND, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20** 712–736.

MCLACHLAN, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Statist.* **36** 318–324.

NOLAN, D. and MARRON, J. S. (1989). Uniform consistency of automatic and location adaptive delta sequence estimators. *Probab. Theory Related Fields* **80** 619–632.

PFANZAGL, J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures. *J. Statist. Plann. Inference* **19** 137–158.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

PRIEBE, C. E. and MARCHETTE, D. J. (2000). Alternating kernel and mixture density estimates. *Comput. Statist. Data Anal.* **35** 43–65.

REDNER, R. A. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann. Statist.* **9** 225–228.

REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239.

RISSANEN, J. (1978). Modeling by shortest data description. *Automatica* **14** 465–471.

RITOV, Y. and BICKEL, P. J. (1990). Achieving information bounds in non- and semiparametric models. *Ann. Statist.* **18** 925–938.

ROEDER, K. and WASSERMAN, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92** 894–902.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

TAMURA, R. N. and BOOS, D. D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *J. Amer. Statist. Assoc.* **81** 223–229.

L. F. JAMES
DEPARTMENT OF MATHEMATICAL SCIENCES
JOHNS HOPKINS UNIVERSITY
BALTIMORE, MARYLAND 21218-2682
E-MAIL: james@mts.jhu.edu

C. E. PRIEBE
DEPARTMENT OF MATHEMATICAL SCIENCES
JOHNS HOPKINS UNIVERSITY
BALTIMORE, MARYLAND 21218-2682
E-MAIL: cep@jhu.edu

D. J. MARCHETTE
NAVAL SURFACE WARFARE CENTER
B10
DAHLGREN, VIRGINIA 22448
E-MAIL: marchettedj@nswc.navy.mil