ELSEVIER

Computational Statistics & Data Analysis 35 (2000) 43–65

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Alternating kernel and mixture density estimates

Carey E. Priebe[a, *], David J. Marchette[b]

[a] *Department of Mathematical Sciences, Johns Hopkins University, Baltimore, MD 21218, USA*
[b] *Naval Surface Warfare Center, Dahlgren, VA 22448, USA*

Received 1 February 1999; received in revised form 1 December 1999

Abstract

We describe and investigate a data-driven procedure for obtaining parsimonious mixture model estimates or, conversely, kernel estimates with data-driven local smoothing properties. The main idea is to obtain a semiparametric estimate by alternating between the parametric and nonparametric viewpoints. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Mixture model; Kernel estimator; Semiparametric

1. Introduction

A ubiquitous and practical problem in statistics, data analysis, and many engineering disciplines is that of estimating the common probability density function f_0 for n identically distributed random variables $\boldsymbol{X} = [X_1, \ldots, X_n]'$. While technical issues abound when consideration shifts to multivariate or dependent samples, there are nonetheless numerous difficulties associated with this practice even for independent univariate samples. Of particular interest in this article is the parametric/nonparametric quandary: nonparametric estimates are often asymptotically 'safer', while parametric estimators can perform better 'when they work' and can offer advantages in terms of model interpretability.

* Corresponding author. Tel.: +1-410-516-7200; fax: +1-410-516-7459.
E-mail address: cep@jhu.edu (C.E. Priebe).

PII: S 0167-9473(00)00003-7

In this article we address specifically semiparametric density estimation, in the sense of estimating the complexity of the parametric estimator one should employ. We present an iterative procedure, alternating between parametric and nonparametric estimates, which ultimately yields a parametric model with data-driven complexity and a nonparametric estimate with data-driven smoothing. This approach can be considered a frequentist competitor for the Bayesian procedure described by Roeder and Wasserman (1997).

The fundamental building blocks of our approach are the dual probability density estimation techniques of the parametric finite mixture estimator and the nonparametric kernel estimator. It is the interplay between these two methodologies that we exploit in this article.

1.1. Finite mixture models

The m component finite mixture model (FMM) is given by $f(x;m,\theta) = \sum_{t=1}^{m} \pi_t \varphi(x;\psi_t)$ where the mixing coefficients π_t are nonnegative and sum to unity (see, e.g., Titterington et al., 1985). The probability density function f is a mixture of elements of $\Phi = \{\varphi(x;\psi) \mid \psi \in \Psi\}$, the family of probability density functions parameterized by ψ; for instance, $\varphi(x;\psi)$ is commonly taken to be the normal density, with $\psi = [\mu, \sigma^2]'$ where $\mu \in \Re$ and $\sigma^2 \in (0,\infty)$. The number of components m represents the *complexity* of the mixture. Given m, the overall parameter vector for the mixture is $\theta=[\pi_1,\ldots,\pi_{m-1},\psi_1,\ldots,\psi_m]'$. For example, given m the maximum likelihood finite mixture estimator based on an i.i.d. sample $\boldsymbol{X}$ of size n is obtained by maximizing the likelihood function $L_X(f)=L_X(m,\theta)=\prod_{i=1}^{n} f(X_i;m,\theta)$ with respect to $\theta \in \Theta$; $\hat{\theta} = \operatorname{argmax}_{\theta\in\Theta} L_X(m,\theta)$. Then $f(x;m,\hat{\theta}) = \sum_{t=1}^{m} \hat{\pi}_t \varphi(x;\hat{\psi}_t)$. When the assumed model is correct this estimate converges in mean integrated squared error at the parametric rate $\mathrm{O}(n^{-1})$ and can allow for interpretation of the model through the assignment of physical meaning to the individual components $\varphi(x;\psi_t)$ and their 'priors' π_t (Table 2.1.3 of Titterington et al. (1985) presents examples for which the physical interpretability of the mixture model is of interest.) Misspecification of the parametric model, of course, can result in misinterpretation as well as a lack of consistency.

While mixtures of normals are dense in the space of well-behaved probability density functions under various distance functions and therefore comprise a rich class of estimators, a major practical problem which must be addressed when employing mixture modelling as a density estimation technique is determining the complexity of the model — estimating the number of components m to use in the mixture. This problem has been addressed in the literature by numerous authors, including but not limited to Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1988), Henna (1988), and, more recently, Chen and Kalbfleisch (1996), Roeder and Wasserman (1997), Dacunha-Castelle and Gassiat (1997), and Solka et al. (1998). One result of the methodology presented herein is a finite mixture model with data-driven complexity estimate $\hat{m}$.

1.2. Kernel estimators

The standard kernel estimator (KE) based on $\boldsymbol{X}$ is given by $\hat{\mathscr{g}}(x;h) = (nh)^{-1}\sum_{i=1}^{n}\varphi((x-X_i)/h)$ (see, e.g. Silverman, 1986). Again, the kernel function φ is commonly taken to be the normal density with standard deviation (bandwidth) h. Convergence for this nonparametric procedure is slower than for the parametric FMM method – $\mathrm{O}(n^{-4/5})$ in mean integrated squared error – but consistency for KE holds much more generally than for FMM.

A major practical problem which must be addressed when employing KE as a density estimation technique is determining the smoothing parameter h. This problem has been addressed in the literature by numerous authors, including but not limited to Abramson (1982), Silverman (1986), Wand et al. (1991), Scott (1992), Sheather (1992), Terrell and Scott (1992), and Wand and Jones (1995). This KE literature strongly suggests that a single bandwidth provides insufficient flexibility in many situations. A second interpretation of the methodology presented herein is a multiple-bandwidth kernel estimator with data-driven smoothing.

1.3. Filtered kernel estimators

Fundamental to combining kernel and mixture estimates in an alternating refinement fashion is the idea of the filtered kernel estimator (FKE) (Marchette 1996; Marchette et al., 1996), one version of which uses a pilot (normal) ‘filtering mixture’ estimate with estimated complexity $\hat{m}$ to construct a multiple-bandwidth KE. Given $f(x;\hat{m},\hat{\theta}) = \sum_{t=1}^{\hat{m}}\hat{\pi}_t\varphi(x;\hat{\mu}_t,\hat{\sigma}_t^2)$, the filtered kernel estimator is defined to be

$$\hat{g}(x;h,\hat{m},\hat{\theta}) = n^{-1}\sum_{i=1}^{n}\sum_{t=1}^{\hat{m}}\frac{\hat{\pi}_t\varphi(X_i;\hat{\mu}_t,\hat{\sigma}_t^2)}{h\hat{\sigma}_t f(X_i;\hat{m},\hat{\theta})}\varphi((x-X_i)/(h\hat{\sigma}_t)).$$

Note from the above equation that the filtered kernel estimator encompasses the standard kernel estimator, with $\hat{\mathscr{g}}(x;h) = \hat{g}(x;h,1,[0,1]')$.

For the FKE the selection of the bandwidth h is no longer as crucial as for KE, since the posterior ‘filter functions’ $\rho_t(x) = \hat{\pi}_t\varphi(x;\hat{\mu}_t,\hat{\sigma}_t^2)/f(x;\hat{m},\hat{\theta})$ and the local standard deviations $\hat{\sigma}_t$ provide data-driven weighting and smoothing according to the individual components of the filtering mixture. This single bandwidth h can be chosen to minimize the asymptotic mean integrated squared error (AMISE) of $\hat{g}(x;h,\hat{m},\hat{\theta})$ under the assumption that the filtering mixture $f(x;\hat{m},\hat{\theta})$ is true; $h_{opt} = \operatorname{argmin}_h \mathit{AMISE}(\hat{g}(x;h,\hat{m},\hat{\theta}) \mid f_0 = f(x;\hat{m},\hat{\theta}))$. See Marchette et al. (1996) for details. This choice for the single bandwidth parameter in the FKE will be used throughout. Therefore the FKE takes the data $\boldsymbol{X}$ and the filtering mixture $f(x;\hat{m},\hat{\theta})$ and, with no parameter settings required, produces the estimate $\hat{g}(x;h,\hat{m},\hat{\theta})$.

1.4. Alternating kernel and mixture estimators

The motivation behind the proposed algorithm is that the nonparametric estimator will suggest (potential) structure which is not yet accounted for in the parametric model. One manifestation of this additional structure is mismatch between the filtered kernel estimate and the m component mixture. Mismatch suggests adding a component to the mixture model to account for the discrepancy. The significance of the change between the original m component mixture estimate and the new $m+1$ component mixture is tested, with the null hypothesis being that the simpler mixture is preferred. Rejection suggests performing another iteration, beginning with the $m+1$ component mixture. Thus the nonparametric portion of the alternating kernel and mixture procedure serves as a 'feature detector' while the parametric portion of the algorithm acts as 'Occam's Razor' to disregard those features which are not supported by sufficient evidence. Related efforts involving the interplay between parametric and nonparametric estimators include Hjort and Glad (1995), Rudzkis and Radavicius (1995), Cao et al. (1995), Cao and Devroye (1996), and Chen and Kalbfleisch (1996).

In Section 2 we present the alternating kernel and mixture (AKM) algorithm for developing a hybrid semiparametric density estimate by iteratively increasing the complexity of the mixture toward eliminating the mismatch between the mixture and the filtered kernel estimator. Section 3 presents simulation and experimental results for the AKM. As an aid to assessing the performance of AKM the simulation analysis includes comparisons with the Bayesian procedure described by Roeder and Wasserman (1997). Experimental analysis includes one application in which the desired output of the AKM is the multiple-bandwidth filtered kernel estimator with data-driven smoothing and one application in which it is the finite mixture model estimate with data-driven complexity that we seek. We conclude in Section 4 with a discussion of our results and their ramifications.

2. AKM algorithm

Given $\boldsymbol{X}$, the idea of the AKM algorithm is to alternate between parametric estimates $\hat{f}^m$ and nonparametric estimates $\hat{g}^m$, basing each on the other in turn. At each iteration m, the m component finite mixture estimate $\hat{f}^m$ is selected to minimize the mismatch between $\hat{f}^m$ and the filtered kernel estimate $\hat{g}^{m-1}$ based on the filtering mixture $\hat{f}^{m-1}$. Then $\hat{g}^m$ is defined as the filtered kernel estimator using the parameter estimates $\hat{\theta}^m$ from the filtering mixture $\hat{f}^m$.

Let $d(\cdot,\cdot)$ represent a distance function defined on the space of probability density functions, and let $\mathscr{F}_m$ denote the family of m component normal mixtures with lower bound l_m and upper bound u_m on term variances; $\sigma_t^2 \in [l_m, u_m]$. Let $\hat{\theta}^1 \equiv [\bar{X}, S^2]', \hat{f}^1(x) \equiv \varphi(x;\hat{\theta}^1)$ be the standard normal density estimate, and $\hat{g}^1(x) \equiv \hat{g}(x; h_{opt}, 1, \hat{\theta}^1)$ be the standard kernel density estimate with bandwidth chosen via the normal reference rule. For $m=2,\dots,n$ define $\hat{f}^m \equiv \operatorname{argmin}_{\mathscr{F}_m} d(f, \hat{g}^{m-1})$, and define $\hat{f}^{n+1} \equiv \hat{f}^n$. The algorithm described above can now be presented in pseudocode as follows:

```
Algorithm AKM($\boldsymbol{X}$)
    $\hat{f}^1(x) \equiv \varphi(x; [\bar{X}, S^2]')$
    $\hat{g}^1(x) \equiv \hat{g}(x; h_{opt}, 1, \hat{\theta}^1)$
    $\hat{f}^2 \equiv \mathrm{argmin}_{\mathscr{F}_2}\, d(f, \hat{g}^1)$
    $m \leftarrow 1$
    While $d(\hat{f}^m, \hat{f}^{m+1}) \geq c > 0$
        $m \leftarrow m + 1$
        $\hat{g}^m(x) \equiv \hat{g}(x; h_{opt}, 1, \hat{\theta}^m)$
        $\hat{f}^{m+1} \equiv \mathrm{argmin}_{\mathscr{F}_{m+1}} d(f, \hat{g}^m)$
    EndWhile
    Return $\hat{m}_n \leftarrow m$
EndAlgorithm
```

The algorithm returns a mixture complexity estimate $\hat{m}_n \leq n$. The resultant parametric mixture estimate is $\hat{f}_n^{\hat{m}_n} \in \mathscr{F}_{\hat{m}_n}$, and $\hat{g}^{\hat{m}_n}$ is the nonparametric filtered kernel estimate which uses $\hat{f}_n^{\hat{m}_n}$ as the filtering mixture.

The iteration termination criterion requires a choice of the constant c; a large value of c will result in relatively fewer components in the resultant mixture. Convergence results require $c_n \to 0$ slowly enough with respect to n. We also require that $\{\mathscr{F}_m\}_{m=1}^{\infty}$ be a sieve dense in the class of continuous density functions, so $l_m \to 0$ and $u_m \to \infty$ as $m \to \infty$.

Theoretical properties of the sequence of mixture estimators $\hat{f}_n^{\hat{m}_n} = f(x; \hat{m}_n, \hat{\theta}_n)$ produced by the AKM algorithm are established in the following two theorems, where $d(f,g) \equiv \int (f-g)^2$ is the integrated squared error.

Theorem 1. *Let $f_0 \in C$, the class of continuous densities on $\Re$.*

Then $d(f_0, \hat{f}_n^{\hat{m}_n}) \to 0$ a.s.

Theorem 2. *Let $f_0 = f(x; m_0, \theta_0) = \sum_{t=1}^{m_0} \pi_t \varphi(x; \mu_t, \sigma_t^2)$ such that φ represents the normal density, $m_0 < \infty$ represents the number of components in the mixture, the mixing coefficients satisfy $0 \leq \pi_t \leq 1$ and $\sum_{t=1}^{m_0} \pi_t = 1$, the $\mu_t \in \Re$ are component means, the $\sigma_t^2 \in [l_{m_0}, u_{m_0}]$ are component variances for some specified bounds l_{m_0} and u_{m_0} on the component variances allowed in an m_0 component mixture, and $\theta_0 = [\pi_1, \ldots, \pi_{m_0-1}, \mu_1, \ldots, \mu_{m_0}, \sigma_1^2, \ldots, \sigma_{m_0}^2]'$. Then $\hat{m}_n \to m_0$ a.s. and $\hat{\theta}_n \to \theta_0$ a.s.*

That is, the AKM mixture estimator is consistent. Moreover, if the target density is a finite mixture of normals then the algorithm successfully estimates the true mixture complexity and parameters. The proofs are given in the appendix.

The theoretical properties established for the AKM estimator in Theorems 1 and 2 hold when the FKE is replaced by the KE. It is our experience, however, that this simplified version of the algorithm does not perform well in practice. This is due to the fact that a single bandwidth kernel estimator often provides a poor estimate of the local structure of the density for a given sample. Even allowing the KE to adjust its single bandwidth with reference to the current mixture model often does not provide sufficient improvement in the KE. Clearly, a variable kernel estimator of some type is preferable. The coupling of FMM and KE inherent in FKE improves the AKM estimator obtained via iteration between the parametric and the nonparametric approaches, as compared with an analogous AKM estimator which uses a single bandwidth KE rather than the FKE. Selecting a FMM produces a new FKE (with the new smoothing parameters) which then, in turn, drives the selection of a new FMM. As long as the FMM is at least somewhat successful in capturing the local smoothness properties inherent in the data, the FKE will improve from iteration to iteration. In the end, the AKM is attempting to fit a mixture to a more appropriate nonparametric estimator.

Given the filtered kernel estimate $\hat{g}^m$, the problem of identifying $\hat{f}^{m+1} \equiv \operatorname{argmin}_{\mathscr{F}_{m+1}} d(f, \hat{g}^m)$ represents a difficult nonlinear optimization task (see, e.g., Bertsekas, 1995; McLachlan and Krishnan, 1997). Stochastic optimization, multiple optimization attempts using different starting points, or a procedure for identifying a good starting point for the optimization are required. We present an approach to providing a 'smart start' from which to begin the optimization through the consideration of '*excess mass*' (see, e.g., Muller and Sawitzki, 1991).

To obtain the starting mixture $\tilde{f}^{m+1}$ for the optimization at iteration $m+1$ we consider the mismatch e^m between the FKE and the FMM at iteration m, $e^m(x) = \hat{g}^m(x) - \hat{f}^m(x)$. Then $\tilde{f}^{m+1}$ is obtained by adding a new component to $\hat{f}^m$ in the region with the greatest excess mass $e^m_+(x) = \chi_{\{e^m>0\}} e^m(x)$. (Here χ_S represents the indicator function on the set S.) Letting $R = \{R_1, \ldots, R_K\} (1 \leq K < \infty)$ be the set of maximal (open) intervals of positive excess mass and $R^* = \operatorname{argmax}_{R_k \in R} \int_{R_k} e^m_+$ be the region with maximum excess mass (see Fig. 1) we define $\tilde{f}^{m+1} = (1-w)\hat{f}^m(x) + w\varphi(x; \mu, v)$ where $w = \int_{R^*} e^m$, $\mu = \int_{R^*} xe^m/w$, and $v = \int_{R^*}(x-\mu)^2 e^m/w$. That is, the starting mixture $\tilde{f}^{m+1}$ updates $\hat{f}^m$ with the mixture component appropriate for matching the size w, shape v, and location μ of the region with maximum excess mass R^*. The optimization to find a (possibly local) minimum of $d(f, \hat{g}^m)$ over $f \in \mathscr{F}_{m+1}$ proceeds from $\tilde{f}^{m+1}$ in a straightforward manner. (This approach to providing a 'smart start' is, of course, not the only possible choice. This is the approach used in the implementation from which the results presented in Section 3 are obtained.)

The iteration termination decision is an exercise in model selection (see, e.g., George and Foster, 1997). The value of $d(\hat{f}^m, \hat{f}^{m+1})$ is employed to evaluate the significance of the improvement realized by employing the $m+1$ component mixture; $d(\hat{f}^m, \hat{f}^{m+1}) < c$ indicates that the simpler m component model is to be preferred and the iteration terminated with an estimated complexity $\hat{m}_n = m$, while $d(\hat{f}^m, \hat{f}^{m+1}) \geq c$ indicates that the more complex model is appropriate and the iteration should continue. Thus c acts as a critical value in the test of significance. The consistency result of Theorem 1 requires only that $c = c_n \to 0$ as $n \to \infty$ and therefore places a constraint on the powers of the sequence of tests: for fixed m the sequence of probabilities of rejection attains unity as $n \to \infty$ whenever the true target density is

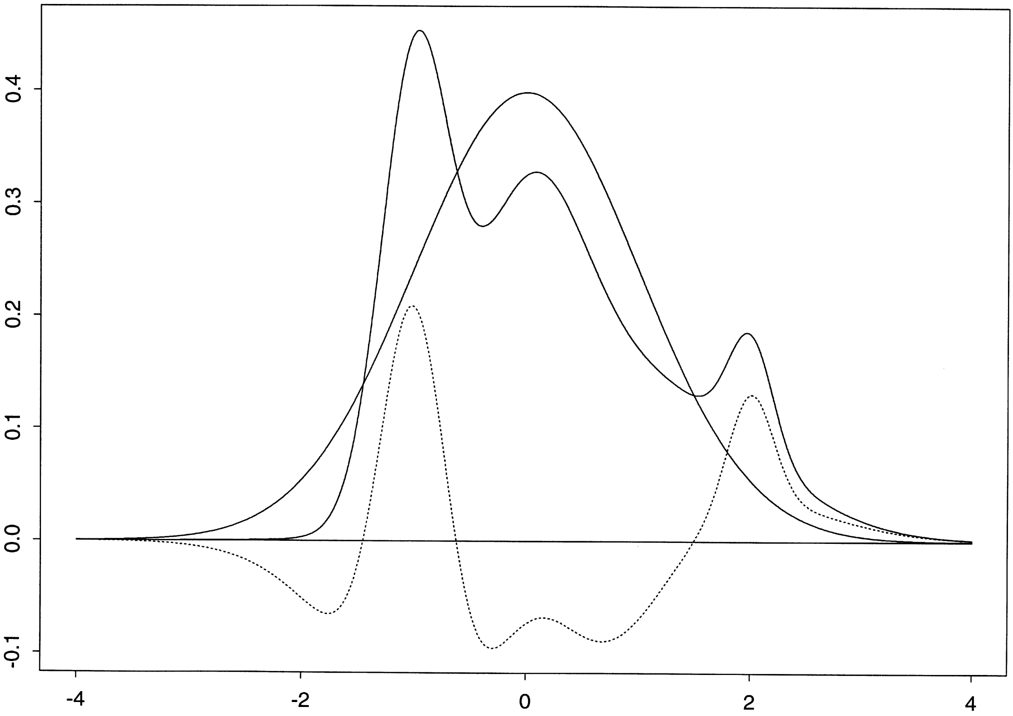


Fig. 1. Excess mass. The optimization involved in the AKM procedure is based on the location and size of the region of greatest excess mass of the finite mixture estimate subtracted from the filtered kernel estimate.

not an m component mixture. Theorem 2 requires, essentially, that $c_n \to 0$ slowly enough so that the sequence of significance levels goes to zero sufficiently fast to ensure that the probability of type I error (adding an $(m+1)$th component when the true target density is an m component mixture) vanishes.

In practice, we can consider $c = c_{n,m}$ provided $c_{n,m}$ satisfies the requirements for each fixed m. For a finite sample of size n, knowledge of the null distribution of $T = d(\hat{f}^m, \hat{f}^{m+1})$ for each m (obtained analytically or via simulation) would allow the critical values in the algorithm to be tailored to satisfy desired power and significance probabilities. In particular, the bootstrap likelihood ratio test for the number of components (see, e.g., McLachlan, 1987) could be used. Unfortunately, the statistic T is not distribution-free. Thus, for simplicity, we employ the following altered form of the algorithm using a simplistic model selection methodology. (The approach described here is used for the simulations presented in Section 3.) The decision on continuing the iteration uses an AIC penalized likelihood criterion (Akaike, 1974): the iteration continues provided

$$d_{\mathrm{AIC}}(\hat{f}^m, \hat{f}^{m+1}) = \log L_X(\hat{f}^{m+1}) - \log L_X(\hat{f}^m) \geq 3.$$

(The value of d_{AIC} can be negative, and is therefore not actually a distance.) Given an initial sample of size n_*, the result of the AIC-based algorithm is a complexity estimate $\hat{m}' \leq n_*$. We then choose c_{n_*} such that $\hat{m}_{n_*} = \hat{m}'$ to begin a sequence $c_n \to 0$

slowly enough as required by Theorem 2; that is, we let the complexity estimate $\hat{m}_{n_*}$ obtained by running the original algorithm be determined by AIC model selection.

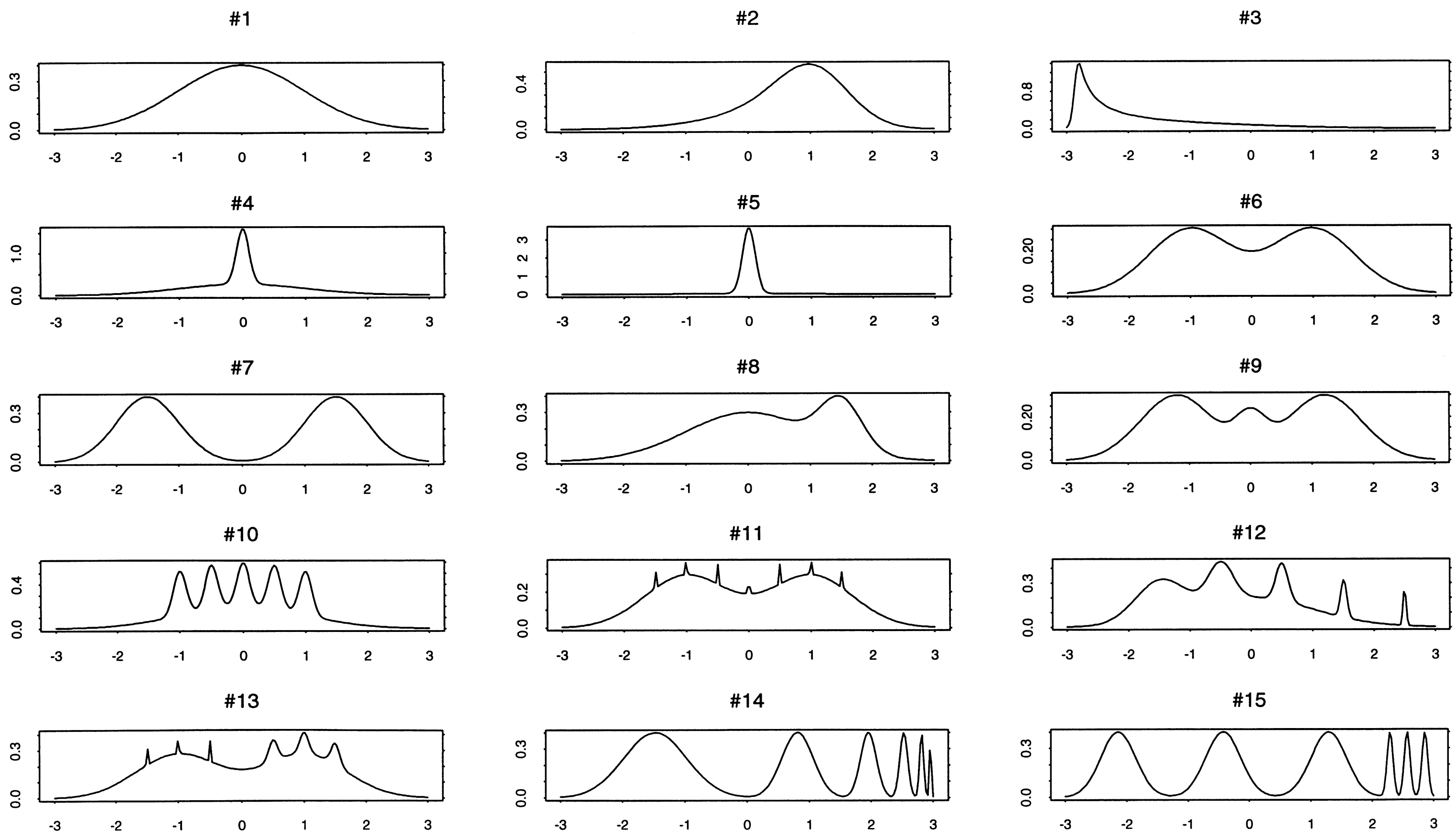
It is noteworthy that the AKM estimate $\hat{f}_n^{\hat{m}_n}$ is not a maximum likelihood estimate. The data enters into the estimate through the filtered kernel estimate only. Given an estimate $\hat{m}_n$ of the complexity, one might be tempted to consider using as the final estimate an $\hat{m}_n$ component maximum likelihood estimate. However, the consistency of such an estimator requires more severe constraints on the rate of growth of $\hat{m}_n$ as a function of n (see, e.g., Geman and Hwang, 1982; Roeder and Wasserman, 1997). The constraints on the rate of increase of $\hat{m}_n$ for the AKM algorithm are implicit and probabilistic. The requirement in Theorem 2 that $c_n \to 0$ slowly constraints the allowable rate of increase for $\hat{m}_n$ with respect to n for this result.

3. Simulation and experimental results

In this section we present an investigation of the performance of the AKM estimators $\hat{f}^{\hat{m}_n}$ and $\hat{g}^{\hat{m}_n}$ involving Monte Carlo simulation and experimental analysis. As an aid to understanding the reported performance numbers we also provide detailed comparative results with the Bayesian competitor described in Roeder and Wasserman (1997) and denoted as R&W. Section 3.1 investigates, via Monte Carlo simulation, the performance of AKM and R&W on normal mixture target densities. Section 3.2 presents a similar simulation in which the target density is not a mixture of normals (lognormal, in this case). In Sections 3.3 and 3.4 the two procedures are compared experimentally on the UK income data and a digital mammography texture data set, respectively. For these last two examples we consider bootstrap resamples from a kernel estimator to allow for quantitative performance comparisons.

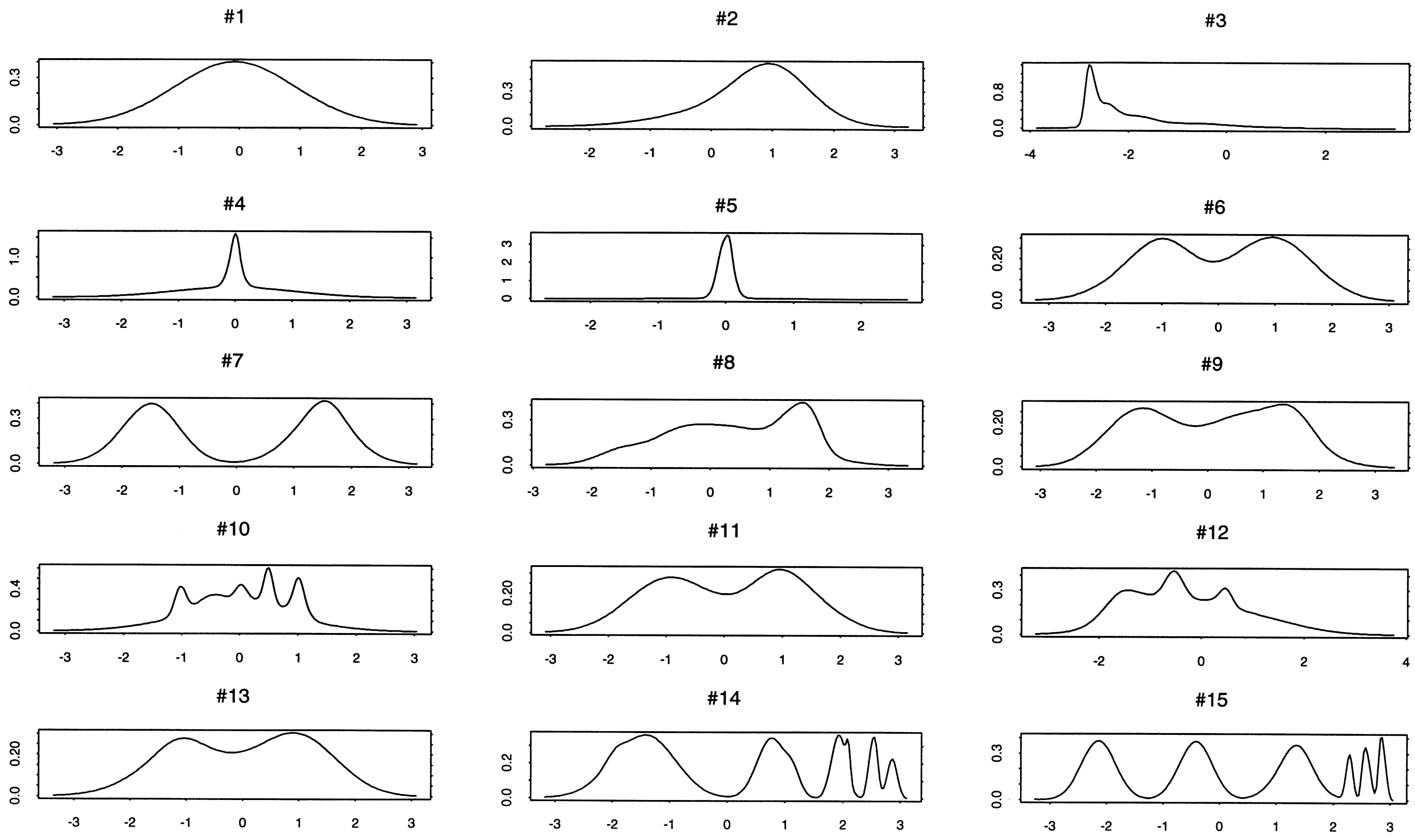
3.1. Monte Carlo simulation: normal mixture target densities

Fig. 2a depicts the Marron and Wand (1992) test suite of 15 normal mixture target densities as M&W#1 through M&W#15. Figs. 2b and c present representative mixture estimation results for the M&W target densities for the AKM and R&W algorithms, respectively. For each estimator, the representative result is the estimate with the median integrated squared error (*ISE*) among 1000 Monte Carlo replications for a sample size of $n = 1000$. (For $n = 1000$ the results for $\hat{g}^{\hat{m}_n}$ are indistinguishable from those presented in Fig. 2b for $\hat{f}^{\hat{m}_n}$.) Tables 1–3 present results from Monte Carlo simulations comparing the performance of AKM and R&W on the test suite for three sample sizes, $n = 50, 250, 1000$. Each result is based on 1000 Monte Carlo replications. In all simulations the maximum value allowed for the estimated mixture model complexity – the number of terms $\hat{m}$ – is 10 for both $\hat{f}^{\hat{m}_n}$ and $\hat{f}_{\mathrm{R\&W}}$. Monte Carlo estimates for the relative mean integrated squared error ($RMISE(\hat{f}) = MISE(\hat{f})/MISE(\hat{f}_{\mathrm{KE}})$) and standard error thereof are reported for $\hat{f}^{\hat{m}_n}, \hat{g}^{\hat{m}_n}, \hat{f}_{\mathrm{R\&W}}$, and $\hat{f}_{\mathrm{Parametric}}$ in Tables 1a, 2a and 3a. The $\hat{f}_{\mathrm{KE}}$ estimate used in the definition of *RMISE* is the optimal (normal theory) kernel estimator, and $\hat{f}_{\mathrm{Parametric}}$



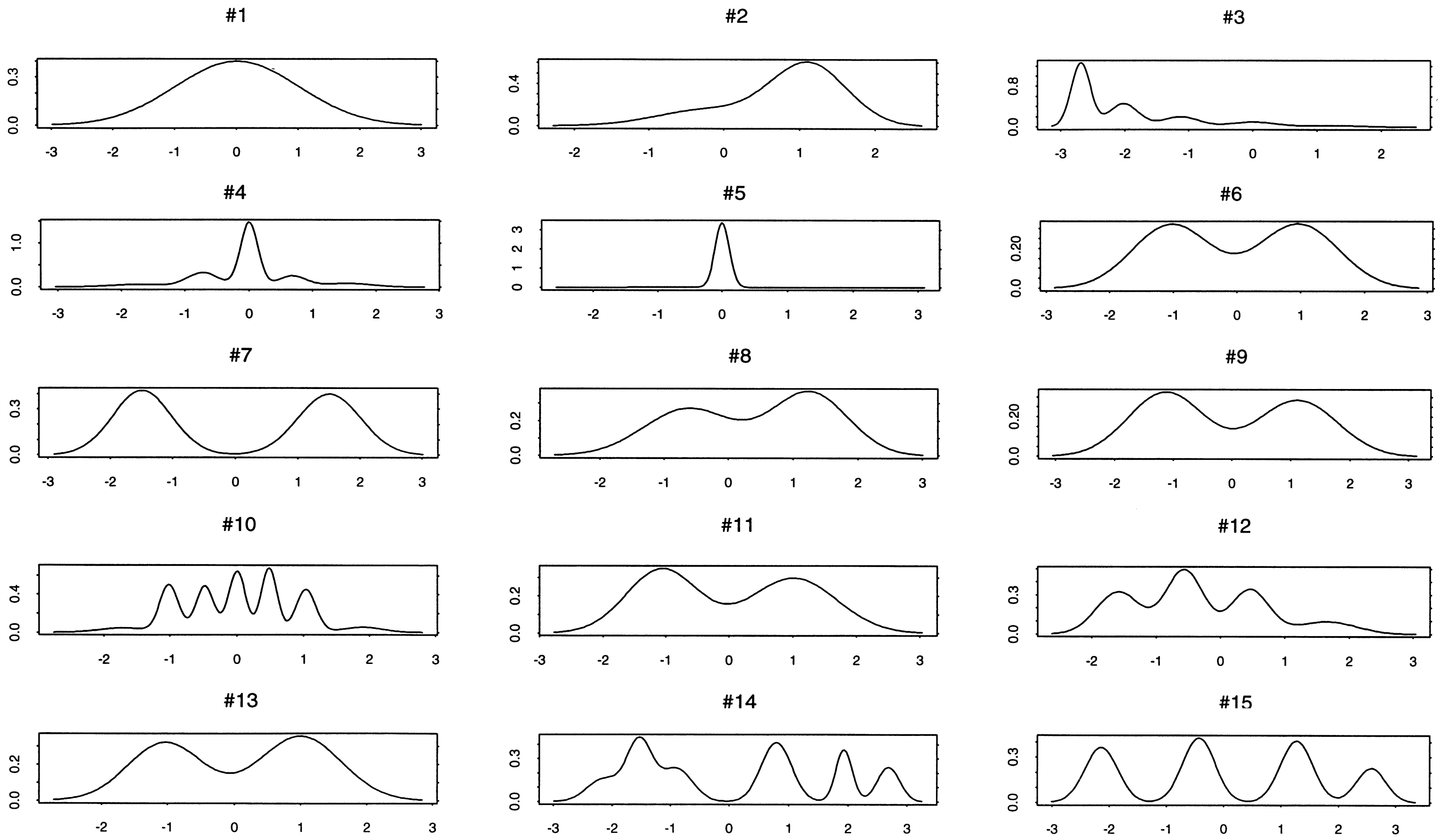
(a)

Figure 2a. (a) The 15 M&W target densities provide a comprehensive suite of finite mixture estimation challenges. (b) Representative AKM estimation results on the 15 M&W target densities. (c) Representative R&W estimation results on the 15 M&W target densities.



(b)

Figure 2. *Continued.*



(c)

Figure 2. *Continued.*

Table 1a
RMISE[a] performance of AKM for $n = 50$

M&W	m	$\hat{f}^{\hat{m}_n}$	$\hat{g}^{\hat{m}_n}$	$\hat{f}_{\text{R\&W}}$	$\hat{f}_{\text{Parametric}}$
#1	1	0.531 (0.016)	1.011 (0.004)	0.537 (0.018)	0.508 (0.015)
#2	3	1.099 (0.020)	1.033 (0.008)	1.342 (0.03)	0.573 (0.018)
#3	8	0.593 (0.007)	0.455 (0.006)	0.841 (0.008)	0.263 (0.007)
#4	2	0.921 (0.011)	0.734 (0.010)	1.102 (0.011)	0.119 (0.004)
#5	2	0.325 (0.007)	0.270 (0.007)	0.298 (0.011)	0.121 (0.004)
#6	2	1.769 (0.027)	1.064 (0.007)	2.121 (0.030)	0.846 (0.014)
#7	2	0.419 (0.004)	0.413 (0.006)	0.300 (0.007)	0.307 (0.007)
#8	2	1.494 (0.019)	1.028 (0.005)	1.751 (0.026)	0.536 (0.011)
#9	3	1.627 (0.024)	1.033 (0.006)	2.035 (0.03)	0.936 (0.016)
#10	6	1.028 (0.005)	1.008 (0.001)	1.043 (0.006)	0.582 (0.008)
#11	9	1.708 (0.025)	1.065 (0.008)	2.065 (0.029)	0.941 (0.017)
#12	6	1.103 (0.006)	1.013 (0.002)	1.157 (0.012)	0.828 (0.054)
#13	8	1.565 (0.019)	1.031 (0.006)	1.821 (0.021)	0.782 (0.015)
#14	6	0.885 (0.003)	0.779 (0.003)	0.695 (0.007)	0.890 (0.196)
#15	6	1.026 (0.005)	0.896 (0.004)	0.513 (0.006)	0.540 (0.010)

[a] Reported are $RMISE(\hat{f})$ (standard error) based on 1000 Monte Carlo replications.

Table 1b
Model complexity[a] performance of AKM for $n = 50$

		$\hat{f}^{\hat{m}_n}$				$\hat{f}_{\text{R\&W}}$			
M&W	m	$P=$	$P<$	$P>$	$P(\vert\vert) \leq 1$	$P=$	$P<$	$P>$	$P(\vert\vert) \leq 1$
#1	1	0.987	0	0.013	1	0.997	0	0.003	1
#2	3	0.005	0.994	0.001	0.265	0.002	0.998	0	0.123
#3	8	0	1	0	0.001	0	1	0	0
#4	2	0.237	0.435	0.328	0.829	0.039	0.751	0.210	0.961
#5	2	0.190	0.036	0.774	0.781	0.492	0.059	0.449	0.943
#6	2	0.303	0.667	0.030	0.996	0.277	0.721	0.002	1
#7	2	0	0	1	0.618	0.996	0	0.004	1
#8	2	0.234	0.742	0.024	0.999	0.148	0.846	0.006	1
#9	3	0.070	0.929	0.001	0.493	0.012	0.988	0	0.414
#10	6	0	1	0	0	0	1	0	0
#11	9	0	1	0	0	0	1	0	0
#12	6	0	1	0	0	0	1	0	0.002
#13	8	0	1	0	0	0	1	0	0
#14	6	0.007	0.991	0.002	0.019	0.001	0.999	0	0.010
#15	6	0.035	0.948	0.017	0.097	0	1	0	0.007

[a] Reported are estimated probabilities that estimated model complexity $\hat{m}$ equals ($P=$), is less than ($P<$), is greater than ($P>$) and is within one of ($P(||) \leq 1$) true model complexity m.

is the parametric estimate assuming the number of terms to be known and obtained by starting the iterative EM algorithm at the true parameters. Note that an exact result is available for $MISE(\hat{f}_{\text{KE}})$ (Marron and Wand, 1992) and thus the *ISE*s for $\hat{f}^{\hat{m}_n}, \hat{g}^{\hat{m}_n}, \hat{f}_{\text{R\&W}}$, and $\hat{f}_{\text{Parametric}}$ have been obtained in closed form rather than

Table 2a
RMISE[a] performance of AKM for $n = 250$

M&W	m	$\hat{f}^{\hat{m}_n}$	$\hat{g}^{\hat{m}_n}$	$\hat{f}_{\mathrm{R\&W}}$	$\hat{f}_{\mathrm{Parametric}}$
#1	1	0.359 (0.014)	1.021 (0.004)	0.324 (0.010)	0.313 (0.010)
#2	3	0.853 (0.011)	0.998 (0.009)	2.723 (0.046)	0.339 (0.009)
#3	8	0.142 (0.002)	0.121 (0.002)	0.620 (0.005)	0.066 (0.001)
#4	2	0.195 (0.004)	0.142 (0.003)	0.504 (0.009)	0.031 (0.001)
#5	2	0.142 (0.003)	0.149 (0.004)	0.164 (0.005)	0.041 (0.002)
#6	2	1.011 (0.008)	1.019 (0.010)	1.232 (0.032)	0.453 (0.010)
#7	2	0.234 (0.003)	0.275 (0.005)	0.130 (0.003)	0.123 (0.003)
#8	2	0.877 (0.007)	0.740 (0.007)	2.651 (0.027)	0.242 (0.006)
#9	3	0.943 (0.006)	0.860 (0.006)	1.250 (0.021)	0.404 (0.007)
#10	6	1.073 (0.003)	0.964 (0.002)	1.132 (0.002)	0.125 (0.002)
#11	9	1.014 (0.008)	1.005 (0.006)	1.205 (0.027)	0.495 (0.008)
#12	6	1.121 (0.003)	1.012 (0.002)	1.395 (0.006)	0.219 (0.004)
#13	8	1.003 (0.005)	0.961 (0.005)	0.989 (0.016)	0.324 (0.006)
#14	6	0.385 (0.004)	0.347 (0.004)	0.515 (0.002)	0.178 (0.003)
#15	6	0.270 (0.002)	0.259 (0.002)	0.268 (0.001)	0.127 (0.002)

[a] Reported are *RMISE*($\hat{f}$) (standard error) based on 1000 Monte Carlo replications.

Table 2b
Model complexity performance[a] of AKM for $n = 250$

		$\hat{f}^{\hat{m}_n}$				$\hat{f}_{\mathrm{R\&W}}$			
M&W	m	$P=$	$P<$	$P>$	$P(\|\|) \leq 1$	$P=$	$P<$	$P>$	$P(\|\|) \leq 1$
#1	1	0.969	0	0.031	0.999	1	0	0	1
#2	3	0.053	0.946	0.001	0.955	0.014	0.986	0	0.375
#3	8	0.001	0.999	0	0.014	0	1	0	0.006
#4	2	0	0	1	0.025	0	0.033	0.967	0.594
#5	2	0.001	0	0.999	0.743	0.280	0	0.720	0.926
#6	2	0.642	0.001	0.357	0.988	0.949	0.051	0	1
#7	2	0	0	1	0.510	1	0	0	1
#8	2	0.638	0.004	0.358	0.981	0.575	0.389	0.036	1
#9	3	0.653	0.302	0.045	0.995	0.016	0.984	0	0.993
#10	6	0.017	0.974	0.009	0.070	0	1	0	0
#11	9	0	1	0	0	0	1	0	0
#12	6	0.002	0.998	0	0.007	0	1	0	0.012
#13	8	0	1	0	0	0	1	0	0
#14	6	0.098	0.142	0.760	0.294	0.040	0.959	0.001	0.155
#15	6	0.027	0	0.973	0.306	0	1	0	0.001

[a] Reported are estimated probabilities that estimated model complexity $\hat{m}$ equals ($P=$), is less than ($P<$), is greater than ($P>$) and is within one of ($P(||) \leq 1$) true model complexity m.

calculated numerically. Investigation of these *RMISE* tables indicates that each of the procedures in the class $\mathscr{C} = \{\hat{f}^{\hat{m}_n}, \hat{g}^{\hat{m}_n}, \hat{f}_{\mathrm{R\&W}}\}$ appears to be admissible (relative to $\mathscr{C}$) in that each procedure is significantly better than the other two for at least one of the target mixtures. Tables 1b, 2b and 3b present the results of the two mixture estimates $\hat{f}^{\hat{m}_n}$ and $\hat{f}_{\mathrm{R\&W}}$ in terms of the complexity of the estimated mixtures.

Table 3a
RMISE performance of AKM for $n = 1000$[a]

M&W	m	$\hat{f}^{\hat{m}_n}$	$\hat{g}^{\hat{m}_n}$	$\hat{f}_{\mathrm{R\&W}}$	$\hat{f}_{\mathrm{Parametric}}$
#1	1	0.259 (0.010)	1.027 (0.004)	0.201 (0.006)	0.193 (0.006)
#2	3	0.730 (0.008)	0.990 (0.009)	5.671 (0.105)	0.183 (0.006)
#3	8	0.055 (0.001)	0.052 (0.001)	0.430 (0.004)	0.019 (0.000)
#4	2	0.072 (0.001)	0.060 (0.001)	0.197 (0.002)	0.009 (0.000)
#5	2	0.083 (0.002)	0.105 (0.002)	0.140 (0.003)	0.019 (0.001)
#6	2	0.693 (0.008)	0.910 (0.008)	1.091 (0.027)	0.242 (0.007)
#7	2	0.136 (0.002)	0.186 (0.002)	0.059 (0.001)	0.055 (0.001)
#8	2	0.488 (0.006)	0.536 (0.007)	3.405 (0.039)	0.113 (0.003)
#9	3	0.776 (0.007)	0.679 (0.005)	1.903 (0.021)	0.196 (0.005)
#10	6	0.451 (0.006)	0.375 (0.005)	0.412 (0.014)	0.034 (0.001)
#11	9	0.845 (0.004)	0.945 (0.004)	1.064 (0.012)	0.208 (0.004)
#12	6	0.931 (0.007)	0.809 (0.006)	1.113 (0.011)	0.065 (0.001)
#13	8	0.904 (0.002)	0.855 (0.002)	0.981 (0.006)	0.106 (0.002)
#14	6	0.129 (0.002)	0.111 (0.001)	0.430 (0.001)	0.049 (0.001)
#15	6	0.105 (0.002)	0.085 (0.001)	0.297 (0.000)	0.038 (0.001)

[a] Reported are $RMISE(\hat{f})$ (standard error) based on 1000 Monte Carlo replications.

Table 3b
Model complexity performance of AKM for $n = 1000$[a]

		$\hat{f}^{\hat{m}_n}$				$\hat{f}_{\mathrm{R\&W}}$			
M&W	m	$P=$	$P<$	$P>$	$P(\|\|) \leq 1$	$P=$	$P<$	$P>$	$P(\|\|) \leq 1$
#1	1	0.946	0	0.054	0.997	1	0	0	1
#2	3	0.230	0.750	0.020	1	0.016	0.984	0	0.826
#3	8	0.010	0.987	0.003	0.183	0.072	0.905	0.023	0.297
#4	2	0	0	1	0	0	0	1	0.002
#5	2	0	0	1	0.658	0.072	0	0.928	0.9
#6	2	0.059	0	0.941	0.948	1	0	0	1
#7	2	0	0	1	0.538	1	0	0	1
#8	2	0.034	0	0.966	0.907	0.831	0.004	0.165	0.998
#9	3	0.443	0	0.557	0.923	0.060	0.940	0	1
#10	6	0.032	0.037	0.931	0.092	0	0.184	0.816	0.334
#11	9	0	1	0	0	0	1	0	0
#12	6	0.068	0.527	0.405	0.299	0.125	0.855	0.020	0.575
#13	8	0.001	0.999	0	0.002	0	1	0	0
#14	6	0.007	0	0.993	0.008	0.829	0.125	0.046	0.984
#15	6	0.001	0	0.999	0.001	0	1	0	0

[a] Reported are estimated probabilities that estimated model complexity $\hat{m}$ equals ($P=$), is less than ($P<$), is greater than ($P>$) and is within one of ($P(||) \leq 1$) true model complexity m.

Target mixtures M&W#11 through M&W#15 are difficult to estimate, even for $\hat{f}_{\mathrm{parametric}}$ where the mixture complexity is assumed known, given the sample sizes under consideration in the simulations. The results we present for these five challenging standard finite mixture densities indicate the (reasonable) limitations of estimation with small-to-moderate sample sizes when the number of components is unknown.

Table 4
Significance testing for H$_0$: $Median[ISE(\hat{f}_{\mathrm{R\&W}}) - ISE(\hat{f}^{\hat{m}_n})] \leq 0$[a]

M&W	$n = 50$	$n = 250$	$n = 1000$
#1	0.004	0.253	0.012
#2	0.000	0.000	0.000
#3	0.000	0.000	0.000
#4	0.000	0.000	0.000
#5	1.000	0.726	0.000
#6	0.000	0.015	0.000
#7	1.000	1.000	1.000
#8	0.004	0.000	0.000
#9	0.000	0.000	0.000
#10	0.180	0.000	1.000
#11	0.000	0.020	0.000
#12	0.747	0.000	0.000
#13	0.000	1.000	0.000
#14	1.000	0.000	0.000
#15	1.000	0.985	0.000

[a] The table of p-values for the sign test, based on 1000 paired Monte Carlo samples, indicates admissibility for both estimator.

Table 4 presents the results of testing H$_0$: $Median[ISE(\hat{f}_{\mathrm{R\&W}}) - ISE(\hat{f}^{\hat{m}_n})] \leq 0$ based on the 1000 paired samples and allows a quick and quantitative analysis of the significance of the *RMISE* results discussed above. A small p-value indicates superiority of the AKM procedure. Results of the one-sided sign test are presented. Investigation of Table 4 indicates that neither procedure uniformly outperforms the other. At $n = 50$ there are 6 of 15 mixtures for which $\hat{f}_{\mathrm{R\&W}}$ outperforms $\hat{f}^{\hat{m}_n}$ under this criterion, at $n = 250$ the ratio drops to 5/15, and for $n = 1000$ there are 2 of the 15 mixtures for which one would choose $\hat{f}_{\mathrm{R\&W}}$ over $\hat{f}^{\hat{m}_n}$.

A discrepancy exists between the results reported in Tables 1 and 2 for R&W and those originally reported in Table 1 of Roeder and Wasserman (1997). The results presented here for R&W are obtained using their code (personal communication, KR). The discrepancy is due to a bookkeeping error in Table 1 of Roeder and Wasserman (1997) (personal communication, LW).

3.2. Monte Carlo simulation: non-normal mixture target density

Table 5 presents results from Monte Carlo simulations comparing the performance of AKM and R&W on the standard lognormal target density ($E[\log X] = 0$; $Var[\log X] = 1$) at sample sizes $n = 100$ and $n = 1000$. The results are based on 100 Monte Carlo replicates. We see that (a) AKM allocates more terms, and (b) AKM performs better in terms of *MISE*. The sign test of H$_0$: $Median[ISE(\hat{f}_{\mathrm{R\&W}}) - ISE(\hat{f}^{\hat{m}_n})] \leq 0$ based on the 100 paired samples yields $p = 0.382$ for $n = 100$ and $p = 0.000$ for $n = 1000$, indicating strong evidence that the AKM is superior for this particular estimation problem, at least for large sample sizes. Fig. 3 shows representative example estimates.

Table 5
Lognormal investigation

	$\hat{g}^{\hat{m}_n}$	$\hat{f}^{\hat{m}_n}$		$\hat{f}_{\mathrm{R\&W}}$	
n	*MISE*	$\hat{m}$	*MISE*	$\hat{m}$	*MISE*
100	0.016 (0.001)	4.36 (0.064)	0.021 (0.001)	2.89 (0.069)	0.060 (0.002)
1000	0.003 (0.000)	7.31 (0.081)	0.003 (0.000)	5.55 (0.153)	0.039 (0.002)[a]

[a] Estimates are based on 100 Monte Carlo replications.

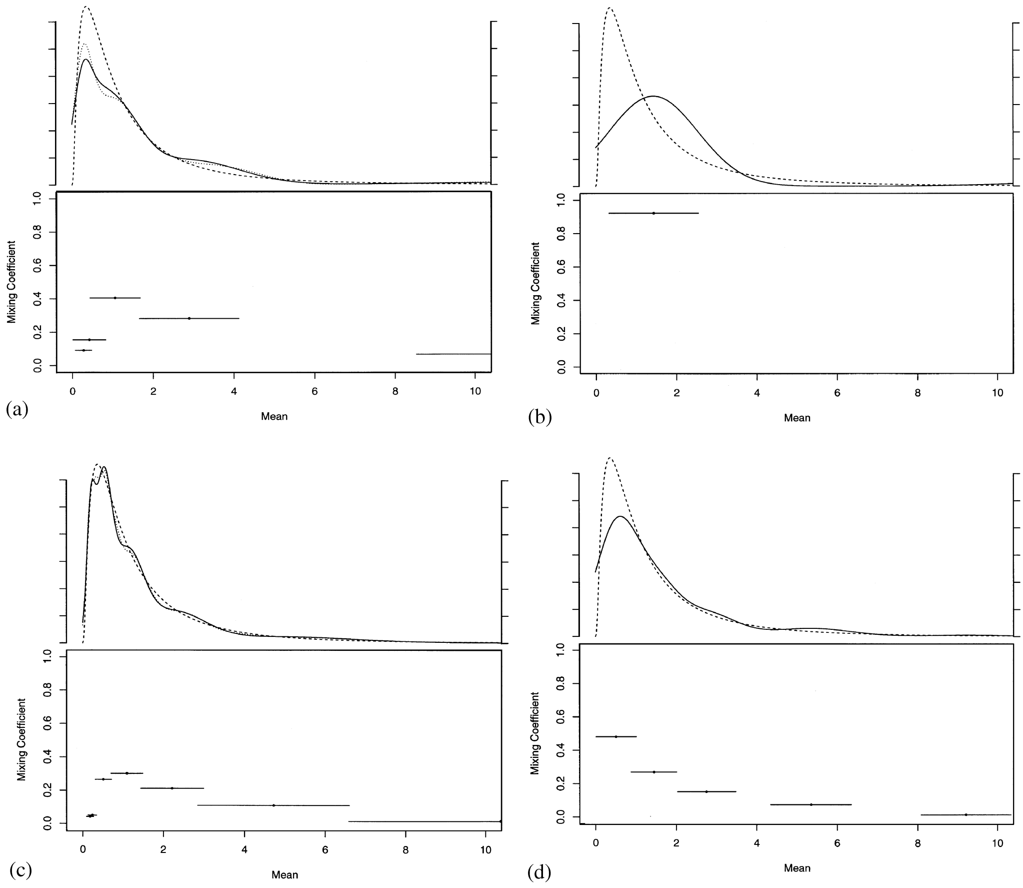


Fig. 3. Lognormal investigation. Presented are representative $\hat{g}^{\hat{m}_n}$ and $\hat{f}^{\hat{m}_n}$ (left) and $\hat{f}_{\mathrm{R\&W}}$ (right) estimates for $n = 100$ (top) and $n = 1000$ (bottom) randomly generated observations. Included in each figure is the true density (dashed line), for comparison. The AKM estimates $\hat{g}^{\hat{m}_n}$ (dotted line) and $\hat{f}^{\hat{m}_n}$ (solid line) are nearly indistinguishable for $n = 1000$.

Each plot in Figs. 3–5 gives two views of the mixture density. On the top of the plot is the curve associated with the density. Below is a representation of the mixture model itself. Each component is plotted as a (*point, line segment*) pair. The means of each term are represented on the x-axis, with the y-axis corresponding to the mixing coefficient for that term. The line segments represent one standard deviation

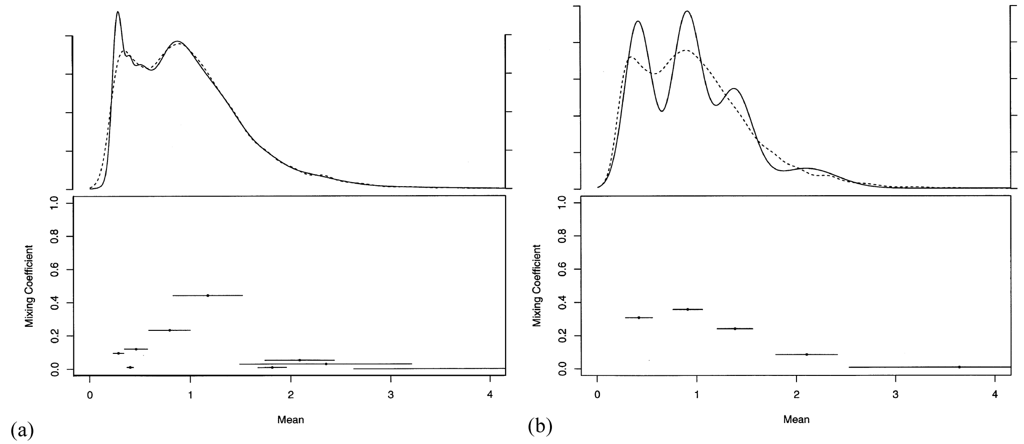


Fig. 4. Investigation of UK Income Data. Presented are the AKM estimate $\hat{f}^{\hat{m}_n}$ (left) and the $\hat{f}_{\mathrm{R\&W}}$ estimate (right) of the $n = 7201$ net income observations. Included in each figure is a standard kernel estimator (dotted line), for comparison. The AKM estimates $\hat{g}^{\hat{m}_n}$ and $\hat{f}^{\hat{m}_n}$ are indistinguishable.

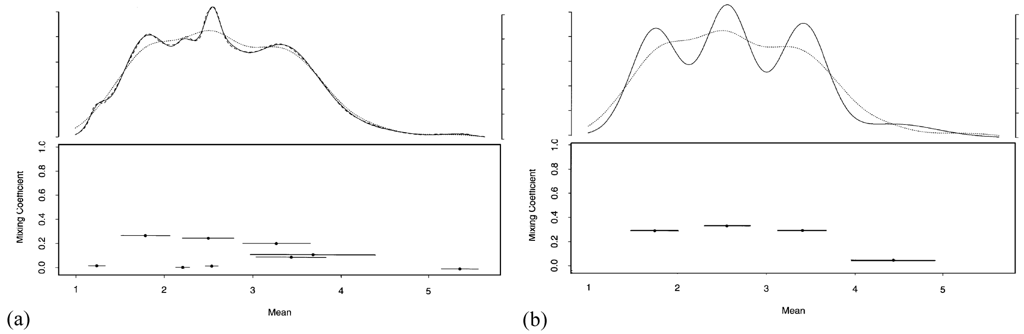


Fig. 5. Investigation of Digital Mammography texture data. Presented are the AKM finite mixture estimate $\hat{f}^{\hat{m}_n}$ (left) and the $\hat{f}_{\mathrm{R\&W}}$ estimate (right) of the $n = 2031$ local coefficient of variation texture observations. Included in each figure is a standard kernel estimator (dotted line), for comparison. The AKM estimates $\hat{g}^{\hat{m}_n}$ (dashed line) and $\hat{f}^{\hat{m}_n}$ (solid line) are nearly indistinguishable.

on either side of the mean. This provides a convenient graphic for understanding the underlying mixture.

3.3. *Experimental analysis: UK income data*

Fig. 4 and Table 6 present results from an investigation of the UK Income data set of normalized net income observations from the UK Family Expenditure Survey for 1975 (Park and Marron, 1990; Marron and Schmitz, 1992). The plots in Fig. 4 represent $\hat{f}^{\hat{m}_n}$ and $\hat{f}_{\mathrm{R\&W}}$ models obtained based on the $n = 7201$ observations. Table 6 presents results based on 100 bootstrap resamples from an undersmoothed kernel estimator. From the table we see that AKM tends to produce a slightly more complex mixture than R&W and a statistically significant improvement in *MISE*.

Table 6
Investigation of UK Income data[a]

	$\hat{m}$	*MISE*
$\hat{g}^{\hat{m}_n}$		0.002 (0.000)
$\hat{f}^{\hat{m}_n}$	6.75 (0.716)	0.015 (0.001)
$\hat{f}_{\mathrm{R\&W}}$	5.03 (0.937)	0.039 (0.003)

[a] Estimates are based on 100 smoothed bootstrap resamples.

Table 7
Investigation of Digital Mammography texture data[a]

	$\hat{m}$	*MISE*
$\hat{g}^{\hat{m}_n}$		0.003 (0.002)
$\hat{f}^{\hat{m}_n}$	6.96 (1.537)	0.003 (0.001)
$\hat{f}_{\mathrm{R\&W}}$	3.95 (0.500)	0.017 (0.003)

[a] Estimates are based on 100 smoothed bootstrap resamples.

The sign test based on the 100 paired bootstrap resamples yields $p = 0.000$ for H_0: $Median[ISE(\hat{f}_{\mathrm{R\&W}}) - ISE(\hat{f}^{\hat{m}_n})] \leq 0$. We caution that while Table 6 suggests that $\hat{f}^{\hat{m}_n}$ and $\hat{g}^{\hat{m}_n}$ may be better estimators than $\hat{f}_{\mathrm{R\&W}}$ for samples of size $n = 7201$ drawn from a kernel estimator obtained from the original UK Income data, such an inference concerning the original experimental data should be regarded as tentative at best. See, e.g., Efron and Tibshirani (1993) or Shao and Tu (1995) for a discussion of the practice and utility of smoothed bootstrap estimates for quantities such as *MISE*.

Qualitative examination of the density estimates in Fig. 4 strongly suggests that the AKM estimate better captures the modal structure in the data. Indeed, the bimodal structure provided by $\hat{f}^{\hat{m}_n}$ is supported by the analysis in Park and Marron (1990) and, more rigorously, Marron and Schmitz (1992). The interest in analyzing this data set with a multiple-bandwidth filtered kernel estimator with data-driven smoothing stems from the desire to determine the modal structure, and we claim that investigation of $\hat{f}^{\hat{m}_n}$ is a useful approach to modal structure analysis.

3.4. Experimental analysis: digital mammography texture data

Fig. 5 and Table 7 present results for a data set of digital mammography texture observations. These data were previously investigated in Priebe et al. (1997a,b), and a similar data set was considered in Priebe (1996). The observations are estimates of the local coefficient of variation – a measure of local ‘roughness’ or texture – for a biopsy-proven malignant tumor region of tissue in a digitized mammogram. The coefficient of variation at a pixel location z, $\kappa_z = \sigma_z/\mu_z$, is estimated as $\hat{\kappa}_z = s_z/\bar{x}_z$ where the sample statistics are obtained based on the gray-level pixel observations in the ball $B(z,r)$ of radius $r > 0$ centered at z. Here $r = 3$. While the coefficient of variation observations are not independent, finite mixture estimates of the marginal texture density for known tumorous regions may be valuable as an aid in the detection of tumors in undiagnosed mammograms. Indeed, the interest in a finite mixture model estimate with data-driven complexity for this data stems from the belief that local image texture features can be useful in discriminating tumorous tissue from healthy tissue (see, e.g., Miller and Astley 1992; Priebe et al. 1994) and the fact that finite mixture models can allow for improved discrimination in spatial scan analysis applications such as this one through 'borrowed strength' density estimation (Priebe, 1996, Priebe et al., 1997a, Priebe and Chen, 2000).

Analogous to the presentation in Section 3.3, the plots in Fig. 5 represent $\hat{f}^{\hat{m}_n}$ and $\hat{f}_{\mathrm{R\&W}}$ models obtained based on the $n = 2031$ observations, and Table 7 presents results based on 100 smoothed bootstrap resamples. Again, the sign test based on the 100 paired bootstrap resamples yields $p = 0.000$ for H_0: $Median[ISE(\hat{f}_{\mathrm{R\&W}}) - ISE(\hat{f}^{\hat{m}_n})] \leq 0$. (In fact, we observe $ISE(\hat{f}_{\mathrm{R\&W}}) > ISE(\hat{f}^{\hat{m}_n})$ for all 100 bootstrap resamples.) The conclusions, tentative though they may be, are the same for this mammography example as for the UK Income data; $\hat{f}^{\hat{m}_n}$ should be our choice for this data set.

4. Discussion and conclusions

The applied statistician using finite mixture models to obtain a probability density estimate for a given data set must first choose the number of terms to use in the mixture, a nuisance parameter in the density estimation problem. Current practice involves exploratory data analysis – choose $\hat{m}$ (perhaps based on visual examination of a kernel estimator), estimate a finite mixture model of order $\hat{m}$, and compare (implicitly) with the kernel estimator. This process is repeated, using a new $\hat{m}$ until a satisfactory mixture model is obtained.

The AKM presented herein can fairly be said to be a 'no-parameter' algorithm in that the procedure implements the above iterative exploratory data analysis methodology in a completely automated fashion in which the user need supply no parameter settings. The careful and comprehensive Monte Carlo analysis presented here indicates that AKM provides acceptable mixture density estimates.

Furthermore, the computational burden imposed by AKM is minimal. For example, for a lognormal target density and a sample size of $n = 1000$, Fig. 3 and Table 5 indicate that the AKM procedure provides a reasonably good mixture estimate with an average complexity of just over seven terms per model. These estimates take approximately twenty seconds computing time on a single processor Silicon Graphics Origin2000 workstation. For much larger data sets, it may be possible to employ binning to allow for efficient computation with minimal loss in estimation performance (Scott, 1992; Rogers et al., 1997). The simulation results presented in Section 3 do not use binning, and the details are not pursued here.

As for the method of comparison employed herein, Monte Carlo investigations are necessary. The theoretical difficulty of finite mixture estimation when the mixture complexity is unknown is notorious; theoretical results for finite sample admissibility will be a continuing challenge. The availability of estimators which are basically 'no-parameter' algorithms (AKM and R&W) allows the foregoing Monte Carlo admissibility study and should provide, in addition to algorithms for the data analyst, fodder for the study of alternate mixture estimation schemes.

We conclude by claiming that the AKM procedure should be added to the applied statistician's estimation toolbox. The practical requirement for a procedure which can easily produce mixture estimates with data-driven complexity is obvious. That AKM meets this requirement has been demonstrated.

Acknowledgements

This work is partially supported by Office of Naval Research Grant N00014-95-1-0777. Editorial comments by anonymous reviewers helped improve the presentation of this material.

Appendix

Some technical definitions will be required.

For a random sample of size n, the result of the AKM algorithm is denoted by $\hat{f}_n^{\hat{m}_n}$; the use of $\hat{m}_n$ indicates that the complexity of the mixture is an estimate, and is a random variable.

We present the proof for integrated squared error: $d^2(f,g) \equiv \int (f-g)^2$. The results can be shown to hold for other distances.

The class of continuous densities on $\mathfrak{R}$ is denoted by C.

The family of m component normal mixtures with lower bound l_m and upper bound u_m on term variances is denoted by $\mathscr{F}_m$.

We denote by $\hat{g}_m$ the standard kernel estimator on the first m observations with bandwidth $h_m \geq \varepsilon_m$, and $\hat{g}_n$ denotes a filtered kernel estimator on n observations.

The mth iteration of the AKM algorithm requires $\hat{f}_n^{\hat{m}} \equiv \operatorname{argmin}_{\mathscr{F}_m} d(f, \hat{g}_n)$. The filtering mixture for the filtered kernel estimator $\hat{g}_n$ used here is $\hat{f}_n^{m-1}$. Notice that the mixture complexity here is denoted by m rather than $\hat{m}$, indicating that the complexity is fixed and is *not* estimated, that is, m is *not* a random variable. Similarly, in the notation $d(\hat{f}_n^{\hat{m}}, \hat{g}_n)$ the filtering mixture for the filtered kernel estimator $\hat{g}_n$ is assumed to be $\hat{f}_n^{m-1}$.

We first state a proposition concerning filtered kernel estimator convergence, established as a straightforward modification of Theorem 1 of Marchette et al. (1996). This result requires only that the filtered kernel bandwidth h_n satisfy $\max_t h_n \hat{\sigma}_t \to 0$ and $n \min_t h_n \hat{\sigma}_t \to \infty$, which in turn constrains the rate at which $l_{\hat{m}_n}$ and $u_{\hat{m}_n}$ go to zero and infinity, respectively.

Proposition A.1. *$d(f_0, \hat{g}_n) \to 0$ a.s. for $f_0 \in C$.*

To prove Theorem 1 we will require two lemmas.

Lemma A.1. *Let* $f_0 \in C$. *Then*
(a) $f_0 \notin \bigcup_{m=1}^{\infty} \mathscr{F}_m \Rightarrow \hat{m}_n \to \infty$ *a.s.*
(b) $f_0 \in \mathscr{F}_{m_0} \Rightarrow \hat{m}_n \to m_* \geq m_0$ *a.s.*

Proof. (a) Assume $\hat{m}_n \to m_* < \infty$. From the definitions of $\hat{f}_n^{m_*}$ and $\hat{f}_n^{m_*+1}$ as minimizers against $\hat{g}_n$ over their respective mixture classes, we have $\lim_{n\to\infty} d(\hat{f}_n^{m_*}, \hat{f}_n^{m_*+1}) \equiv \delta$ a.s. since $\hat{g}_n$ converges. Clearly $\delta \geq 0$; strict monotonicity ($\delta > 0$ a.s.) can be established by noting that the mixture parameters $\hat{\theta}_n$ are continuous in the observations x_i and that $\delta = 0$ requires that $\hat{f}_n^{m_*+1}$ be a degenerate $m_* + 1$ component mixture. Since $c_n \to 0$, eventually $d(\hat{f}_n^{m_*}, \hat{f}_n^{m_*+1}) \geq c_n$, resulting in the addition of an $(m_* + 1)$th component. This contradiction implies $\hat{m}_n \to \infty$ a.s. for $f_0 \notin \bigcup_{m=1}^{\infty} \mathscr{F}_m$. The proof of (b) follows precisely the same argument until $m_* = m_0$, and thus leaves us with the conclusion $m_0 \leq m_* \leq \infty$ a.s for $f_0 \in \mathscr{F}_{m_0}$. □

The proof of Theorem 1 proceeds as follows. First note that $d(\hat{f}_n^{\hat{m}_n}, \hat{g}_n) \leq d(\hat{\mathscr{g}}_{\hat{m}_n}, \hat{g}_n)$, where $\hat{\mathscr{g}}_m$ is defined as the standard kernel estimator on the first m observations with bandwidth $l_m \leq h_m \leq u_m$. This follows from the definition of $\hat{f}_n^{\hat{m}_n}$ as a minimizer of $d(f, \hat{g}_n)$ over $f \in \mathscr{F}_{\hat{m}_n}$ and the fact that $\hat{\mathscr{g}}_{\hat{m}_n} \in \mathscr{F}_{\hat{m}_n}$ by construction. Next, recall that $d(\hat{\mathscr{g}}_{\hat{m}_n}, f_0) \to 0$ a.s. provided $\hat{m}_n \to \infty$ and the sequence of bandwidths h_m satisfies $h_m \to 0$ and $mh_m \to \infty$. (These requirements on h_m in turn imply that the mixture component variance constraints be chosen so that $l_m \to 0$.) The desired result is established for $f_0 \notin \bigcup_{m=1}^{\infty} \mathscr{F}_m$ by invoking Lemma A.1(a). By Lemma A.1(b) it remains to establish consistency for $f_0 \in \mathscr{F}_{m_0}$ if $m_0 \leq m_* < \infty$. In this case, in an argument analogous to the proof of Lemma A.1(a), we see that $d(\hat{f}_n^{m_*}, \hat{g}_n) \to 0$ a.s., for otherwise the algorithm would eventually add another component. The consistency of $\hat{g}_n$ (Proposition A.1) completes the proof. □

To prove Theorem 2 we first establish that $\hat{m}_n \to m_0$ almost surely when $f_0 = f(x; m_0, \theta_0) \in \mathscr{F}_{m_0}$. Once this has been accomplished, the almost sure convergence of $\hat{f}_n^{\hat{m}_n} \to f_0$ together with the identifiability of normal mixtures will imply $\hat{\theta}_n \to \theta_0$ a.s., completing the proof.

We begin by noting that since both $\hat{f}_n^{m_0}$ and f_0 are elements of $\mathscr{F}_{m_0}$, $d(\hat{f}_n^{m_0}, \hat{g}_n) \leq d(f_0, \hat{g}_n)$ and thus $d(\hat{f}_n^{m_0}, \hat{g}_n) \to 0$ by Proposition 1. Furthermore, $d(\hat{f}_n^{m_0+1}, \hat{g}_n) \leq d(\hat{f}_n^{m_0}, \hat{g}_n)$. We require the existence of n' such that

$$d(\hat{f}_n^{m_0}, \hat{g}_n) \leq c_n/2 \quad \text{for all } n \geq n', \tag{A.1}$$

from which we can conclude that $d(\hat{f}_n^{m_0}, \hat{f}_n^{m_0+1}) < c_n$ for $n \geq n'$ and thus the algorithm will not add an $m_0 + 1^{st}$ component after $n = n'$. Since $d(f_0, \hat{g}_n)$ converges to zero at a rate of $O(n^{-4/5})$ for all $f_0 \in \bigcup_{m=1}^{\infty} \mathscr{F}_m$, a choice of $c_n = n^{-4/5+\beta}$ for $0 < \beta < 4/5$ implies the almost sure existence of an n' satisfying (A.1). Combining this result with Lemma A.1(b) establishes $\hat{m}_n \to m_0$ a.s. as desired. □

References

Abramson, I.S., 1982. On bandwidth variation in kernel estimates – a square root law. Ann. Statist. 10, 1217–1223.

Akaike, H., 1974. A new look at statistical model identification. IEEE Trans. Automat. Control 19, 716–723.

Bertsekas, D.P., 1995. Nonlinear Programming. Athena Scientific, Belmont, MA.

Cao, R., Cuevas, A., Fraiman, R., 1995. Minimum distance density-based estimation. Comput. Statist. Data Anal. 20, 611–631.

Cao, R., Devroye, L., 1996. The consistency of a smoothed minimum distance estimate. Scand. J. Statist. 23, 405–418.

Chen, J., Kalbfleisch, J.D., 1996. Penalized minimum-distance estimates in finite mixture models. Canad. J. Statist. 24, 167–175.

Dacunha-Castelle, D., Gassiat, E., 1997. The estimation of the order of a mixture model. Bernoulli 3, 279–299.

Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Everitt, B.S., Hand, D.J., 1981. Finite Mixture Distributions. Chapman & Hall, London.

Geman, S., Hwang, C.-R., 1982. Nonparametric maximum likelihood estimation by the method of sieves. Ann. Statist. 10, 401–414.

George, E.I., Foster, D.P., 1997. Calibration and empirical Bayes variable selection. Unpublished technical manuscript, University of Texas at Austin.

Henna, J., 1988. An estimator for the number of components of finite mixtures and its applications. J. Jpn. Statist. Soc. 18, 51–64.

Hjort, N.L., Glad, I.K., 1995. Nonparametric density estimation with a parametric start. Ann. Statist. 23, 882–904.

Marchette, D.J., 1996. The filtered kernel density estimator. Ph.D. Dissertation, George Mason University.

Marchette, D.J., Priebe, C.E., Rogers, G.W., Solka, J.L., 1996. Filtered kernel density estimation. Comput. Statist. 11, 95–112.

Marron, J.S., Schmitz, H.P., 1992. Simultaneous estimation of several size distributions of income. Econometric Theory 8, 476–488.

Marron, J.S., 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.

McLachlan, G.J., Basford, K.E., 1988. Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York.

McLachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley, New York.

Miller, P., Astley, S., 1992. Classification of breast tissue by texture analysis. Image Vision Comput. 10, 277–282.

Muller, D.W., Sawitzki, G., 1991. Excess mass estimates and tests for multimodality. J. Amer. Statist. Assoc. 86, 738–746.

Park, B.U., Marron, J.S., 1990. Comparison of data-driven bandwidth selectors. J. Amer. Statist. Assoc. 85, 66–72.

Priebe, C.E., 1996. Nonhomogeneity analysis using borrowed strength. J. Amer. Statist. Assoc. 91, 1497–1503.

Priebe, C.E., Chen, D., 2000. Borrowed strength density estimates. Technometrics, to appear.

Priebe, C.E., Marchette, D.J., Rogers, G.W., 1997a. Segmentation of random fields via borrowed strength density estimation. IEEE Trans. Pattern Anal. Mach. Intell. 19, 494–499.

Priebe, C.E., Marchette, D.J., Rogers, G.W., 1997b. Segmentation of random fields via borrowed strength density estimation. Technical Report No. 546, Department of Mathematical Sciences, Johns Hopkins University.

Priebe, C.E., Solka, J.L., Lorey, R.A., Rogers, G., Poston, W., Kallergi, M., Qian, W., Clarke, L.P., Clark, R.A., 1994. The application of fractal analysis to mammographic tissue classification. Cancer Lett. 77, 183–189.

Roeder, K., Wasserman, L., 1997. Practical Bayesian density estimation using mixtures of normals. J. Amer. Statist. Assoc. 92, 894–902.

Rogers, G.W., Wallet, B.C., Wegman, E.J., 1997. A mixed measure formulation of the EM algorithm for huge data set applications. Comput. Sci. Statist. 28, 492–497.

Rudzkis, R., Radavicius, M., 1995. Statistical estimation of a mixture of Gaussian distributions. Acta Appl. Math. 38, 37–54.

Scott, D.W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley, New York.

Shao, J., Tu, D., 1995. The Jackknife and Bootstrap. Springer, New York.

Sheather, S.J., 1992. The performance of six popular bandwidth selection methods on some real data sets. Comput. Statist. 7, 225–250.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, New York.

Solka, J.L., Wegman, E.J., Priebe, C.E., Poston, W.L., Rogers, G.W., 1998. Mixture structure analysis using the Akaike information criterion and the bootstrap. Statist. Comput. 8, 177–188.

Terrell, G.R., Scott, D.W., 1992. Variable kernel density estimation. Ann. Statist. 20, 1236–1265.

Titterington, D.M., Smith, A.F.M., Makov, U.E., 1985. Statistics Analysis of Finite Mixture Distributions. Wiley, New York.

Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman & Hall, London.

Wand, M.P., Marron, J.S., Ruppert, D., 1991. Transformations in density estimation. J. Amer. Statist. Assoc. 86, 343–361.