Bayesian Model Selection in Finite Mixtures by Marginal Density Decompositions

Hemant ISHWARAN, Lancelot F. JAMES, and Jiayang SUN

We consider the problem of estimating the number of components d and the unknown mixing distribution in a finite mixture model, in which d is bounded by some fixed finite number N. Our approach relies on the use of a prior over *the space of mixing distributions* with at most N components. By decomposing the resulting marginal density under this prior, we discover a weighted Bayes factor method for consistently estimating d that can be implemented by an iid generalized weighted Chinese restaurant (GWCR) Monte Carlo algorithm. We also discuss a Gibbs sampling method (the blocked Gibbs sampler) for estimating d and also the mixing distribution. We show that our resulting posterior is consistent and achieves the frequentist optimal $O_p(n^{-1/4})$ rate of estimation. We compare the performance of the new GWCR model selection procedure with that of the Akaike information criterion and the Bayes information criterion implemented through an EM algorithm. Applications of our methods to five real datasets and simulations are considered.

KEY WORDS: Blocked Gibbs sampler; Dirichlet prior; Generalized weighted Chinese restaurant; Identification; Partition; Uniformly exponentially consistent test; Weighted Bayes factor.

1. INTRODUCTION

Consider the finite mixture problem where we wish to estimate Q_0, an unknown finite mixing distribution with d atoms. We assume that $d < \infty$ is unknown but that a finite upper bound N is known for it: $1 \le d \le N < \infty$. Our inference for Q_0 and its dimension d is to be based on n iid observations $\mathbf{X} = (X_1, \ldots, X_n)$ from the *true* distribution P_0 with density

$$f_0(x) = \int_y f(x|y)\, dQ_0(y) = \sum_{k=1}^{d} W_{k,0}\, f(x|Y_{k,0}), \qquad x \in \mathcal{X} \subseteq \Re,$$

with respect to a σ-finite measure λ, where the unknown Q_0 is written as

$$Q_0(\cdot) = \sum_{k=1}^{d} W_{k,0}\, \delta_{Y_{k,0}}(\cdot),$$

where $0 < W_{k,0} < 1$ are weights summing to 1 ($\sum_{k=1}^{d} W_{k,0} = 1$), $\delta(\cdot)$ is the standard indicator function, and $Y_{k,0} \in \mathcal{Y}$ are the distinct atoms of Q_0, where $\mathcal{Y} \subseteq \Re$.

How tight the bound N is for d often depends very much on the context of the problem. For example, Roeder (1994) analyzed the sodium-lithium countertransport (SLC) measurements from $n = 190$ individuals using a finite normal mixture model with an upper bound of $N = 3$. There it was argued that the upper bound of three components was a natural choice because of the genetics underlying the problem. However, one cannot always expect such a tight bound for d in all problems. For example, Izenman and Sommer (1988), in reanalyzing Wilson's (1983) data of the thickness of the 1872–1874 Hidalgo postage stamps of Mexico, found evidence of seven modes using Silverman's (1981) critical bandwidth test. Although Silverman's test can be viewed as a nonparametric approach, Izenman and Sommer (1988) presented compelling evidence showing that the data could be derived from a finite normal mixture model with at least $d = 7$ components. Such a conclusion was also partially shared by Minnotte and Scott (1993), who, using a mode tree analysis, found the original seven modes of Izenman and Sommer (1988) in addition to three new modes. Efron and Tibshirani (1993, chap. 6) looked at the same data using a bootstrap approach and concluded that there was any where from two to seven modes. Thus the foregoing analyses, combined with Izenman and Sommer's (1988) detailed arguments, suggest that the Hidalgo stamp data can be safely analyzed using a finite normal mixture model. However, it is not clear how this information can be used to arrive at a sharp upper bound for d. To be safe, we could, for example, choose a value for N that is clearly larger than any value of d estimated from the previous analyses (to be conservative, say $N = 15$), but it may not be so easy to improve on this strategy. (See, however, Basford, McLachlan, and York 1997 for likelihood ratio methods for bounding N.)

The foregoing discussion makes it clear that one can encounter both tight and conservative bounds for d, and thus it is important to have an approach that will work well over a broad range of values of N. As we show, one of the nice features of our Bayesian approach that we present is that it not only works well for small values of N, but also works well with very large N, and in fact has a natural nonparametric limit as $N \to \infty$. This important feature is a direct consequence of our choice of prior.

1.1 Defining the Parameter Space

There are essentially two basic approaches to modeling the parameter space for the finite mixture model. The first approach is more along the lines of a parametric method and considers the parameter space as

$$\mathcal{Y}^N \times \left\{ (W_1, \ldots, W_N) : 0 \le W_k \le 1, \sum_{k=1}^{N} W_k = 1 \right\},$$

Hemant Ishwaran is Associate Staff, Department of Biostatistics and Epidemiology/Wb4, Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, OH 44195 (E-mail: *ishwaran@bio.ri.ccf.org*). Lancelot James is Assistant Professor, Department of Mathematical Sciences, Johns Hopkins University, Baltimore, MD 21218 (E-mail: *james@brutus.mts.jhu.edu*). Jiayang Sun is Associate Professor, Department of Statistics, Case Western Reserve University, Cleveland, OH 44106 (E-mail: *jiayang@sun.cwru.edu*).
The authors thank the associate editor for a careful review of the paper. Thanks are also extended to Jiahua Chen for helpful discussion related to Theorem 2. The first two authors' work was partially supported by the Acheson J. Duncan Fund for the Advancement of Research in Statistics, Awards 00-1 and 01-3, Department of Mathematical Sciences, Johns Hopkins University. The last author's work was supported in part by the National Science Foundation.

corresponding to the atoms and weights for the mixing distribution. This is perhaps the more common approach used in Bayesian analyses (see, e.g., Aitkin and Rubin 1985; Chib 1995; Diebolt and Robert 1994; Raftery 1996; Richardson and Green 1997; Roeder and Wasserman 1997). McLachlan and Peel (2000, chap. 4) have provided more discussion and other related references. The second method, which we adopt here, considers the parameter space to be a space of mixing distributions. Thus if $\mathcal{Q}_j$ denotes the space of finite mixtures Q over $\mathcal{Y}$ with exactly j atoms, then the parameter space for Q_0 is

$$\mathcal{Q}(N) = \bigcup_{j=1}^{N} \mathcal{Q}_j,$$

the space of finite mixtures with at most N atoms. This approach is more in line with those used in non-Bayesian settings (see, e.g., Chen 1995; Jewell 1982; Lambert and Tierney 1984; Leroux 1992; Lindsay 1983; Pfanzagl 1988; Simar 1976; Teicher 1960; van de Geer 1996; Zhang 1990).

There are both computational and theoretical advantages to working with $\mathcal{Q}(N)$ as our parameter space. From a conceptual/theoretical perspective, it puts us more on par with non-Bayesian methods, thus allowing us to exploit some key concepts used in these approaches to derive analogous results for our Bayesian procedure. For example, exploiting the concept of *strong identification* used by Chen (1995), we are able to show that our posterior estimates Q_0 at a $O_p(n^{-1/4})$ rate (see Theorem 2). As shown by Chen (1995), this is the optimal (frequentist) rate of estimation for Q_0.

The form of identification assumed in the finite mixture model is intimately tied to the asymptotic behavior of the posterior. Under a different type of identification, which we call *$\mathcal{F}$-identification*, we show how to use a weighted Bayes factor (BF) to consistently estimate d (see Theorem 1). This method again relies on considering the mixing distribution Q as our parameter and is of benefit computationally. The weighted BF approach is suggested by our decomposition of the marginal density for the data, $m_N(\mathbf{X})$, as

$$m_N(\mathbf{X}) = \sum_{k=1}^{N} m_{k,N}(\mathbf{X}),$$

where each piece $m_{k,N}(\mathbf{X})$ corresponds roughly to the contribution from a prior over $\mathcal{Q}_k$. This suggests the use of the ratio, a weighted BF,

$$\Delta(k, k') = \frac{m_{k,N}(\mathbf{X})}{m_{k',N}(\mathbf{X})}, \tag{1}$$

for selecting the dimension d. (A more precise argument is given in Sec. 2.) Moreover, this approach allows us to apply an iid Monte Carlo procedure, the generalized weighted Chinese restaurant (GWCR) algorithm (Ishwaran and James 2000a; see also Brunner, Chan, James, and Lo 2001; Lo, Brunner, and Chan 1996) in computing (1). This connection arises from the identity

$$m_{k,N}(\mathbf{X}) = \sum_{\{\mathbf{p}:\, n(\mathbf{p})=k\}} \Lambda(\mathbf{p}) q(\mathbf{p}),$$

where the sum is over all partitions $\mathbf{p}$ of the set of integers $\{1, \ldots, n\}$, where $n(\mathbf{p})$ equals the number of sets in $\mathbf{p}$, $q(\mathbf{p})$ is the GWCR density and $\Lambda(\mathbf{p})$ is a known function of $\mathbf{p}$ (see Secs. 3.1 and 3.2 for further details and a specific example). Thus, by drawing iid values from $q(\mathbf{p})$, we can approximate (1). In addition to the GWCR algorithm, we also use a Gibbs sampling procedure called the blocked Gibbs sampler (Ishwaran and James 2001; Ishwaran and Zarepour 2000a) for inference. The blocked Gibbs sampler also relies on the use of $\mathcal{Q}(N)$ as our parameter space and gives us a method for drawing values directly from the posterior of the mixing distribution Q, and hence a method for estimating Q_0 as well as d.

1.2 Random Measures

Our approach to the problem relies on the alternate representation of the mixture model as a hierarchical model involving hidden variables Y_i:

$$\begin{aligned} &(X_i|Y_i) \overset{\text{ind}}{\sim} f(X_i|Y_i), \qquad i = 1, \ldots, n \\ &(Y_i|Q_0) \overset{\text{iid}}{\sim} Q_0. \end{aligned} \tag{2}$$

To estimate Q_0, we place a prior on Q using a random measure. Thus, in analogy to (2), we base inference for Q_0 on the posterior for the following hierarchical model:

$$\begin{aligned} &(X_i|Y_i) \overset{\text{ind}}{\sim} f(X_i|Y_i), \qquad i = 1, \ldots, n \\ &(Y_i|Q) \overset{\text{iid}}{\sim} Q \\ &Q \sim \mathcal{P}_N(\cdot) = \sum_{k=1}^{N} W_k \delta_{Z_k}(\cdot), \end{aligned} \tag{3}$$

where

$$\mathbf{W} = (W_1, \ldots, W_N) \sim \text{Dirichlet}_N(\alpha/N, \ldots, \alpha/N) \tag{4}$$

is independent of Z_k, which are iid H, where we assume that H is *nonatomic* [e.g., if $\mathcal{Y} = \Re$, then H is usually taken as a flat $\text{N}(0, A)$ distribution with large variance A; see Sec. 3.2 for illustration] and α is some positive value. Thus, analogous to (2), we assume that Y_i are iid Q, but where Q is now a random mixing distribution drawn from $\mathcal{P}_N$, a distribution over $\mathcal{Q}(N)$.

The selection of shape parameters α/N in the Dirichlet distribution for $\mathbf{W}$ is critical to the performance of $\mathcal{P}_N$. As shown by Ishwaran and Zarepour (2000b), this choice of parameters implies the weak convergence result

$$\mathcal{P}_N \overset{d}{\Rightarrow} \text{DP}(\alpha H), \text{ as } N \to \infty,$$

where $\text{DP}(\alpha H)$ is the Ferguson Dirichlet process with finite measure parameter αH (Ferguson 1973, 1974). Thus $\mathcal{P}_N$ has an appropriate nonparametric limit, which is the reason for its good performance for large N. However, we demonstrate that it also has good properties for small values of N.

1.3 Organization

The article is organized as follows. Section 2 presents the decomposition of the marginal density for the data. This sug-

gests a method for consistently estimating d using a weighted BF (Sec. 2.2). The small N and large N properties for $\mathcal{P}_N$ are investigated in Section 2.3, while the asymptotic consistency of the BF approach is given in Theorem 1 in Section 2.4.

The proofs for Theorem 1 and the optimal rate result of Theorem 2 (appearing later in Sec. 5) both rely on the notion of a *uniformly exponentially consistent* (UEC) test and its implication for posterior consistency under various metrics. The proofs build on work of Schwartz (1965), Barron (1988, 1989), and Clarke and Barron (1990). For ease of presentation, we have placed these proofs in the Appendix.

The iid GWCR algorithm is discussed in Section 3, and the blocked Gibbs sampling algorithm is covered in Section 4. The performance of these algorithms when applied to a range of well-known datasets as well as simulated data is studied in Section 6. In some cases we compare these results to the EM algorithm. It is well known that the performance of the EM algorithm depends very much on the initial values used. We find that our new model selection procedures are relatively insensitive to the initial values, and thus they avoid the sometimes laborious work needed for the EM algorithm to find good initial values with each candidate d model. More details are given in Section 6.

2. MODEL SELECTION BY MARGINAL DENSITY DECOMPOSITIONS

The marginal density is a key ingredient in computation of the BF, a widely used method for model selection (see, e.g., Bernardo and Smith 1993, chap. 6). In the context of the finite mixture model, suppose that Π_1 and Π_2 are priors for Q and let P_Q be the distribution for the mixed density

$$f_Q(x) = \int_y f(x|y)\, dQ(y).$$

Then the BF for comparing the model induced under Π_1 to that induced under Π_2 is the ratio of posterior odds to prior odds or, equivalently, the ratio of marginal densities,

$$\mathrm{BF} = \frac{\int \prod_{i=1}^n f_Q(X_i)\, \Pi_1(dQ)}{\int \prod_{i=1}^n f_Q(X_i)\, \Pi_2(dQ)}.$$

A large value of BF is evidence for favoring model 1 over model 2.

Such an approach typically requires specification of a prior Π that distributes mass over the space of distributions $\mathcal{Q}(N)$ for Q. A standard approach is to use a mixture such as

$$\Pi(\cdot) = \sum_{k=1}^N p_k \Pi_k(\cdot),$$

where $p_1 + \cdots + p_N = 1$ and Π_k is a prior over $\mathcal{Q}_k$, the space of distributions with exactly k atoms. The prior Π, although conceptually appealing, has several hidden difficulties with its use. First, there are no well-known default choices for the weights p_k, and arbitrary selection of these values could lead to undesirable properties. A good lesson of what can go wrong when choosing a prior for mixture models occurs with the use of a uniform prior in place of $\mathcal{P}_N$ in the model (3). Although a uniform prior may seem an intuitive choice, we show that it can have poor properties even when N is fairly small and becomes inconsistent as $N \to \infty$. Computations with Π are another concern. Fitting a model based on Π requires an algorithm that must traverse over parameter spaces of different dimensions, which can create difficulties for Monte Carlo procedures such as Gibbs sampling. One solution to this problem was presented by Richardson and Green (1997), who discussed a reversible jump Markov chain method for fitting mixture models.

2.1 The Density for $\mathbf{X}$

Our approach is different. Rather than working with a prior like Π, we see by careful decomposition of the marginal density that the prior $\mathcal{P}_N$ naturally decomposes the mixture problem into models based on mixing distributions of different dimension, thus effectively taking the problem and carving it out into smaller ones as a natural byproduct. This decomposition naturally suggests a method for selecting models using an approximate BF and thus avoids the problems associated with priors such as Π. (For some more rationale and intuition for marginal decompositions, see Lo 1984, who looked at kernel density decompositions under the Dirichlet process prior.)

Let $\mathbf{p} = \{C_j\colon\ j = 1, \ldots, n(\mathbf{p})\}$ be a partition of the set $\{1, \ldots, n\}$, where C_j is the jth cell of the partition, e_j is the number of elements in a cell C_j, and $n(\mathbf{p})$ is the number of cells in the partition. From (3), the density for $\mathbf{X}$ must be

$$\begin{aligned} m_N(\mathbf{X}) &= \int \prod_{i=1}^n f_Q(X_i)\, \mathcal{P}_N(dQ) \\ &= \int \left(\int \prod_{i=1}^n f(X_i|Y_i)\, Q(dY_i) \right) \mathcal{P}_N(dQ) \\ &= \int\int\int \prod_{i=1}^n f(X_i|Y_i) \prod_{i=1}^n \left(\sum_{k=1}^N W_k\, \delta_{Z_k}(dY_i) \right) \\ &\quad \times d\pi_{\mathbf{W}}(\mathbf{W})\, dH^N(\mathbf{Z}) \\ &= \sum_{\mathbf{p}} \pi_N(\mathbf{p}) \prod_{j=1}^{n(\mathbf{p})} \int \prod_{i \in C_j} f(X_i|y)\, H(dy) \\ &= \sum_{\mathbf{p}} \pi_N(\mathbf{p}) f(\mathbf{X}|\mathbf{p}), \end{aligned} \tag{5}$$

where the sum is over all partitions $\mathbf{p}$, $\pi_{\mathbf{W}}$ is the symmetric Dirichlet distribution (4) for $\mathbf{W}, \mathbf{Z} = (Z_1, \ldots, Z_N)$ and

$$\begin{aligned} \pi_N(\mathbf{p}) &= \mathcal{P}_N\{\mathbf{P} = \mathbf{p}\} \\ &= \sum_{\{i_1 \neq \cdots \neq i_k\}} E\left(W_{i_1}^{e_1} \cdots W_{i_k}^{e_k} \right) \\ &= \frac{(\alpha/N)^k N!}{\alpha^{(n)} (N-k)!} \prod_{j=1}^k \left(1 + \frac{\alpha}{N} \right)^{(e_j - 1)}, \qquad k = n(\mathbf{p}), \end{aligned}$$

where $a^{(j)} = a(a+1)\cdots(a+j-1)$ for each $a > 0$ and positive integer $j \geq 1$. (Note that $a^{(0)} = 1$.)

The expression $\pi(\mathbf{p})$ is the *exchangeable partition probability function* (EPPF), which characterizes the prediction rule (conditional distribution) for $\mathcal{P}_N$ (Pitman 1995, 1996). (For more on its derivation, see Ishwaran and Zarepour 2000b;

Pitman 1995; Watterson 1976.) It is interesting to note that

$$\pi_N(\mathbf{p}) \to \frac{\alpha^k}{\alpha^{(n)}} \prod_{j=1}^{k} (e_j - 1)! \quad \text{as } N \to \infty,$$

where the right side is the EPPF for a DP(αH) measure. This provides some intuition for why $\mathcal{P}_N$ converges to a Dirichlet process (see Ishwaran and Zarepour 2000b, c for further details).

2.2 Decomposing the Marginal Density

We can rewrite the marginal density as $m_N(\mathbf{X}) = \sum_{k=1}^{N} m_{k,N}(\mathbf{X})$, where

$$m_{k,N}(\mathbf{X}) = \sum_{\{\mathbf{p}:\, n(\mathbf{p})=k\}} \pi_N(\mathbf{p}) f(\mathbf{X}|\mathbf{p}).$$

Each of the terms $m_{k,N}(\mathbf{X})$ represents a measure of the relative posterior mass for a k-component mixture. This follows because the posterior probability that Q has k components can be measured by

$$\mathbb{P}\{n(\mathbf{p}) = k|\mathbf{X}\} = \frac{\sum_{\mathbf{p}} I\{n(\mathbf{p}) = k\}\pi_N(\mathbf{p}) f(\mathbf{X}|\mathbf{p})}{\sum_{\mathbf{p}} \pi_N(\mathbf{p}) f(\mathbf{X}|\mathbf{p})} = \frac{m_{k,N}(\mathbf{X})}{m_N(\mathbf{X})}.$$

This suggests a natural method for selecting models based on $m_{k,N}$. Thus models can be selected based on $\Delta(k, k') = m_{k,N}/m_{k',N}$, the posterior odds of a partition of size k versus one of size k'.

As alluded to in Section 1, the use of the ratio $\Delta(k, k')$ for model selection can also be motivated by recognizing that it is an approximate weighted BF and thus should inherit some of the asymptotic properties typically enjoyed by BFs (see Theorem 1 for details). This connection to the BF is revealed by showing that the $m_{k,N}$ are near marginal densities. A little bit of work shows that $m_{k,N}$ approximates the marginal density (up to a proportionality constant) for a mixture model based on a prior over $\mathcal{Q}_k$. Thus, effectively, m_N has a natural decomposition into the N terms $m_{k,N}$, with each term essentially representing the contribution to the posterior from a mixture model based on k components.

By considering the decomposition (5), notice that

$$m_{k,N}(\mathbf{X}) = r_{k,N} \int \left(\int_{\mathcal{Y}_k} \prod_{i=1}^{n} f(X_i|Y_i)\, Q(dY_i) \right) \mathcal{P}_k(dQ), \qquad k = 1, \ldots, N,$$

where $\mathcal{Y}_k = \{\mathbf{Y} \in \mathcal{Y}^n : n(\mathbf{Y}) = k\}$ consist of those $\mathbf{Y}$ in $\mathcal{Y}^n$ with exactly k distinct coordinate values,

$$r_{k,N} = \binom{N}{k} \frac{(k\alpha/N)^{(n)}}{\alpha^{(n)}}$$

and

$$\mathcal{P}_k(\cdot) = \sum_{j=1}^{k} W_{j,k}\, \delta_{Z_j}(\cdot)$$

is a random probability measure with random weights

$$(W_{1,k}, \ldots, W_{k,k}) \sim \text{Dirichlet}_k(\alpha/N, \ldots, \alpha/N).$$

Under $\mathcal{P}_k$, the Y_i values are effectively the constraining set $\mathcal{Y}_k$ involved in the function $m_{k,N}$. From this observation, it becomes clear that $m_{k,N}$ (except for a constant) approximates the marginal density

$$m_k(\mathbf{X}) = \int \left(\int \prod_{i=1}^{n} f(X_i|Y_i)\, Q(dY_i) \right) \mathcal{P}_k(dQ)$$

derived from $\mathcal{P}_k$, a prior over the space of k-component mixtures, $\mathcal{Q}_k$. The following lemma shows that this approximation is exponentially accurate.

Lemma 1. Let $r^*_{k,N} = r_{k,N}\mathcal{P}_k(\mathcal{Y}_k)$. Define the density $m^*_{k,N} = m_{k,N}/r^*_{k,N}$. Then

$$\int_{\mathcal{X}^n} |m^*_{k,N}(\mathbf{X}) - m_k(\mathbf{X})|\, d\lambda^n(\mathbf{X}) \le 2k(1-1/k)^n \approx 2k\exp(-n/k).$$

Proof. By using a decomposition similar to (5), we have that

$$\begin{aligned} &\int_{\mathcal{X}^n} |m^*_{k,N}(\mathbf{X}) - m_k(\mathbf{X})|\, d\lambda^n(\mathbf{X}) \\ &\quad \le \sum_{\{\mathbf{p}:\, n(\mathbf{p})=k\}} \left| \frac{1}{\mathcal{P}_k(\mathcal{Y}_k)} - 1 \right| \mathcal{P}_k(\mathbf{p}) + \sum_{\{\mathbf{p}:\, n(\mathbf{p})<k\}} \mathcal{P}_k(\mathbf{p}) \\ &\quad = 2(1 - \mathcal{P}_k(\mathcal{Y}_k)). \end{aligned}$$

But $\mathcal{P}_k(\mathcal{Y}_k^c) = \mathcal{P}_k\{Y_i \ne Z_j \text{ some } j\}$, which is bounded by $k(1-1/k)^n$.

2.3 Weighted Bayes Factors

Lemma 1 reveals that $m_{k,N}$ is approximated to a high degree of accuracy by $m_k \times r^*_{k,N}$, and thus can be seen to be made up of two pieces: the marginal density m_k and the prior weight for a model of dimension k, $r^*_{k,N}$. Thus, another way to motivate the ratio $\Delta(k, k')$ is that it corresponds approximately (in n) to the weighted BF,

$$\frac{m_k(\mathbf{X})}{m_{k'}(\mathbf{X})} \times \frac{r^*_{k,N}}{r^*_{k',N}} \approx \frac{m_k(\mathbf{X})}{m_{k'}(\mathbf{X})} \times \frac{r_{k,N}}{r_{k',N}}. \tag{6}$$

The expression on the right side of (6) reveals that the effect of the finite-dimensional Dirichlet prior (3) on $\Delta(k, k')$ is essentially captured by the value of $r_{k,N}$, and thus a careful study of its value should indicate how well the prior will work in our model selection procedure. For example, as indicated earlier, the choice of shape parameters α/N in our $\mathcal{P}_N$ prior ensures that it has a limiting Dirichlet process distribution, thus hinting that $\mathcal{P}_N$ should work well in mixture models, at least for the case when N is large. But how exactly does this prior stack up against other choices for different values of N, as measured by $r_{k,N}$?

As a competitor, consider the prior with parameters $\alpha/N = 1$ (i.e., $\alpha = N$) corresponding to weights

$$\mathbf{W} = (W_1, \ldots, W_N) \sim \text{Dirichlet}_N(1, \ldots, 1).$$

This represents a "uniform prior," and intuitively we might expect it to act like a noninformative prior. This is in fact exactly the opposite of what happens, at least when N is large.

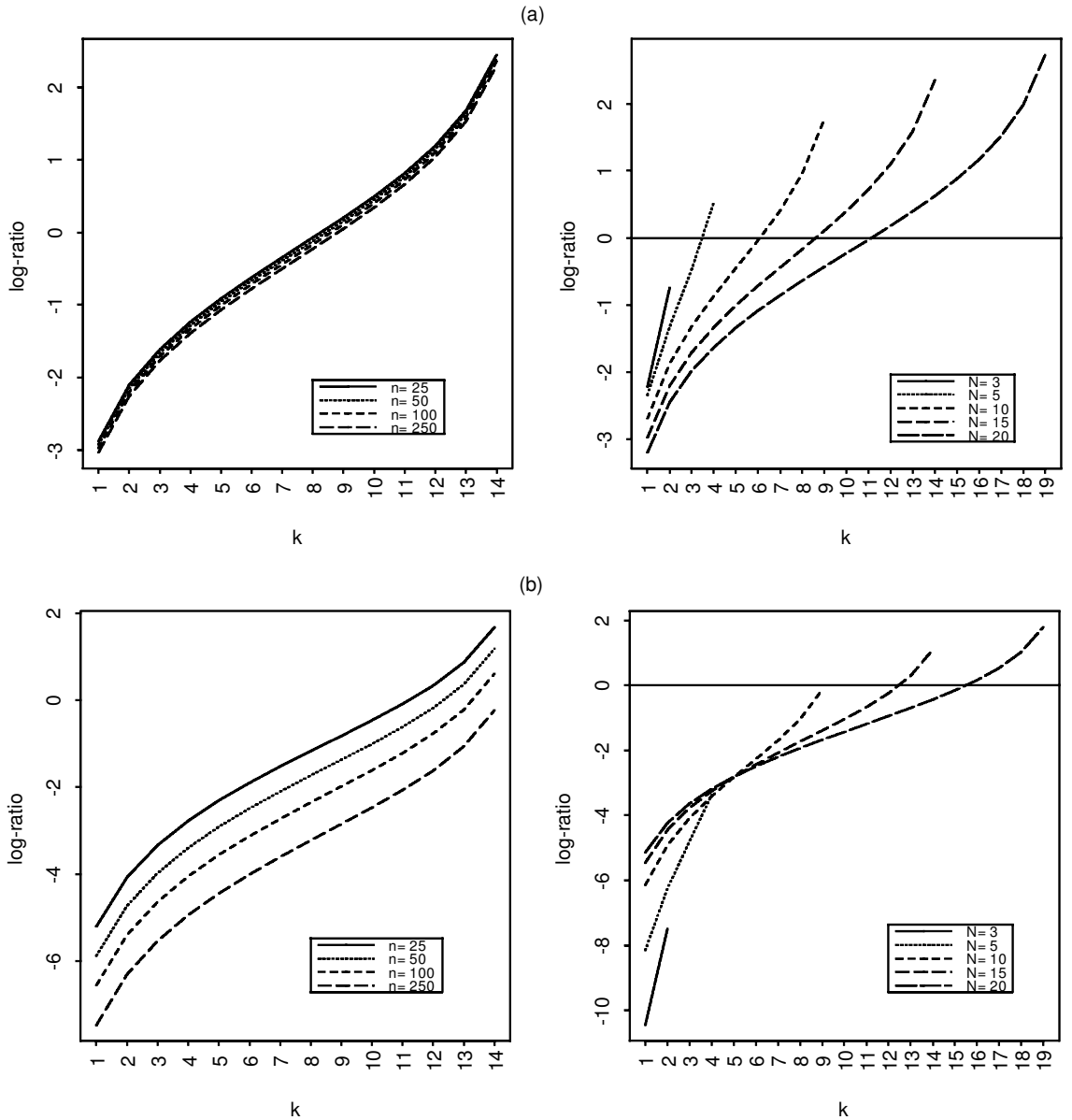

Figure 1. Plots of $\log(r_{k,N}/r_{k+1,N})$ for Various Choices of k,N and n. (a) The $\mathcal{P}_N$ prior with $\alpha = 1$. (b) The uniform prior ($\alpha = N$). Left-side plots show values for different sample size values n with $N = 15$. Right-side plots use different values of N with $n = 100$. The horizontal line at 0 is the break-even point where $r_{k,N} = r_{k+1,N}$, at which point the prior starts to favor smaller models as k increases.

Asymptotically the prior has the almost-sure limit H, the law of our Z_k random variables, and thus is completely informative. This is exactly the wrong prior to use in nonparametric mixture problems, where it has been shown to be inconsistent (Ishwaran and Zarepour 2000c).

Although this tells us that the uniform prior behaves poorly for very large N, how does it behave for finite N, and how does its performance compare to the approximate Dirichlet process? To answer these questions, we plotted the log of the ratios $r_{k,N}/r_{k+1,N}$ under each prior for various values for n and N. The results appear in Figure 1. The top row of the figure reveals the behavior for the approximate Dirichlet process prior with parameter $\alpha = 1$. The news seems quite good here. The left-side plot shows that the ratio is almost independent of the sample size n, while the right-side plot shows that the prior neither overpenalizes smaller models nor overly encourages larger models, even as N increases. (Note that a value of 0 corresponds to the break-even k value where $r_{k,N} = r_{k+1,N}$ at which point the prior starts to favor smaller models as k increases.) This is much different than the bottom row plots for the uniform prior. The left-side plots show a clear dependence on sample size, while the right-side plots reveal that the prior heavily penalizes smaller models as most of the values for the log-ratio value are below 0. This last property is what causes this prior to be inconsistent in nonparametric problems. Clearly, the approximate Dirichlet process is the more appropriate choice for both small and large values of N, almost independently of the sample size.

2.4 Asymptotic Behavior for $\Delta(k, k')$

The connection between $\Delta(k, k')$ and the weighted BF (6) suggests that $\Delta(k, k')$ should inherit some of the asymptotic

properties for marginal densities that BFs typically benefit by. Indeed, Theorem 1 establishes fairly mild assumptions under which $\Delta(k, k')$ has good asymptotic behavior. A key requirement is a form of model identification that is slightly stronger than the usual notion of identification. Recall that a mixture model is said to be identified over $\mathcal{Q}(N)$ if for $Q_1, Q_2 \in \mathcal{Q}(N)$ we have $f_{Q_1}(x) = f_{Q_2}(x)$ for almost all $x[\lambda]$; then $Q_1 = Q_2$. For example, mixtures of Poisson distributions, binomial distributions (under various constraints), and scale and location mixtures of gamma distributions and normal distributions are known to be identified (Teicher 1963). We require a stronger form of identification, which we call *$\mathcal{F}$-identification*:

$\mathcal{F}$-Identification. We say that the mixture model is $\mathcal{F}$-identified over $\mathcal{Q}(N)$ if there exists a countable measurable partition $\mathcal{F}$ for $\mathcal{X}$ with the property that if $P_{Q_1}(A) = P_{Q_2}(A)$ for each $A \in \mathcal{F}$, where $Q_1, Q_2 \in \mathcal{Q}(N)$, then $Q_1 = Q_2$.

By selecting $\mathcal{F} = \mathcal{X}$, we can see immediately that identified mixtures with discrete sample spaces are $\mathcal{F}$-identified. Thus, for example, mixtures of Poisson distributions are $\mathcal{F}$-identified by choosing $\mathcal{F} = \{0, 1, 2, \dots\}$. It also follows that continuous mixtures based on exponential family densities of the form

$$f(x|y) = \exp(ys(x) + t(x) + b(y)) \tag{7}$$

are $\mathcal{F}$-identified under fairly simple conditions. Consider the following proposition (see the Appendix for a proof).

Proposition 1. Suppose that $f(x|y)$ is a density of the form (7) with respect to Lebesgue measure. If there exists an x_0 either in $\mathcal{X}$ or its closure so that $s(x) \to +\infty$ or $s(x) \to -\infty$ as $x \to x_0$, then the corresponding mixture model is $\mathcal{F}$-identified over $\mathcal{Q}(N)$.

Proposition 1 implies $\mathcal{F}$-identification for several important mixtures. For example, with $x_0 = +\infty$ and $s(x) = x$, it follows that location mixtures of normals are $\mathcal{F}$-identified. As another example, consider scale mixtures of gamma distributions, which are $\mathcal{F}$-identified using $x_0 = +\infty$ and $s(x) = -x$.

Another key condition needed in our theorem is the notion of *information denseness* for the prior, a concept related to relative entropy. Recall that the relative entropy (Kullback–Leibler distance) for two probability measures $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$D(\mathbb{P}||\mathbb{Q}) = \mathbb{P}\log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) = \int \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}.$$

The prior $\mathcal{P}_d$ is said to be information dense at Q_0 if for each $\epsilon > 0$,

$$\mathcal{P}_d\{Q : D(P_0||P_Q) < \epsilon\} > 0.$$

In words, $\mathcal{P}_d$ concentrates on each Kullback–Leibler neighborhood of the true model.

Information denseness can be shown to hold under various smoothness conditions for the underlying density $f(x|y)$, where smoothness is in terms of y. However, we prefer to keep the conditions of our theorem straightforward, leaving these to be checked on a case-by-case basis. For example, it is easy to show that information denseness holds for scale and location mixtures of normals, as well as for mixtures of Poisson distributions.

Theorem 1. If $\mathcal{P}_d$ is information dense at Q_0 and the mixture model is $\mathcal{F}$-identified over $\mathcal{Q}(d)$, then there exists an $\epsilon > 0$ such that

$$\Delta(d, k) = \frac{m_{d,N}(\mathbf{X})}{m_{k,N}(\mathbf{X})} \geq \exp(n\epsilon) \quad \text{almost surely } \mathcal{P}_0^\infty,$$

for each $k = 1, \dots, d-1$.

Thus, under mild assumptions, $\Delta(d, k)$ is asymptotically exponentially large whenever $k < d$. Therefore, it discrimates well against (incorrect) smaller models. Although the case of $d < k \leq N$ is not addressed here, Theorem 2 (in Sec. 5) establishes a more general result showing that the posterior achieves the optimal $O_p(n^{-1/4})$ rate of estimation for Q_0. (This requires more stringent assumptions, however.) This should help alleviate any concerns about Theorem 1 and the performance of $\Delta(k, k')$. Proof of Theorem 1 is given in the Appendix.

Remark 1. There are several parallel results to Theorem 1 for the case where the dimension d is unknown. That is, when Q_0 is a finite mixture distribution with d components but d is completely unknown, $1 \leq d < \infty$. Such a problem was considered by Leroux (1992), who showed that the use of a maximum penalized likelihood method was consistent for the dimension in that it was not underestimated, although the result did not say whether penalization would overestimate d. Recently, Keribin (2000) resolved this issue, showing that a maximum penalized likelihood method was almost surely consistent for the complexity d.

3. GENERALIZED WEIGHTED CHINESE RESTAURANT ALGORITHMS

The approach based on the weighted BF $\Delta(k, k')$ lends itself nicely to computations using the iid GWCR algorithm (Ishwaran and James 2000a; see also Brunner et al. 2001; Lo et al. 1996). The GWCR algorithm is a sequential importance sampling technique that draws values from a density $q(\mathbf{p})$ over the space of partitions, where $q(\mathbf{p})$ (the GWCR density) acts as an importance function for approximating $\pi_N(\mathbf{p})f(\mathbf{X}|\mathbf{p})$. That is,

$$\pi_N(\mathbf{p})f(\mathbf{X}|\mathbf{p}) = \Lambda(\mathbf{p})q(\mathbf{p}),$$

where $\Lambda(\mathbf{p})$ are the importance weights. In particular, observe that this implies

$$m_{k,N}(\mathbf{X}) = \sum_{\{\mathbf{p}:\, n(\mathbf{p})=k\}} \pi_N(\mathbf{p})f(\mathbf{X}|\mathbf{p}) = \sum_{\{\mathbf{p}:\, n(\mathbf{p})=k\}} \Lambda(\mathbf{p})q(\mathbf{p}).$$

Thus, by drawing values from $q(\mathbf{p})$ we can devise an effective iid method for approximating $\Delta(k, k')$. To decide between two models k and k', draw B iid partitions $\mathbf{p}^1, \dots, \mathbf{p}^B$ from q and use the approximation

$$\Delta(k, k') \approx \frac{\sum_{j=1}^B I\{n(\mathbf{p}^j) = k\}\Lambda(\mathbf{p}^j)}{\sum_{j=1}^B I\{n(\mathbf{p}^j) = k'\}\Lambda(\mathbf{p}^j)}.$$

The validity of this approximation is guaranteed by the strong law of large numbers. Note that the same technique can also

be used for the Monte Carlo estimate

$$\mathbb{P}\{n(\mathbf{p}) = k|\mathbf{X}\} \approx \frac{\sum_{j=1}^{B} I\{n(\mathbf{p}^j) = k\}\Lambda(\mathbf{p}^j)}{\sum_{j=1}^{B} \Lambda(\mathbf{p}^j)}.$$

3.1 The Generalized Weighted Chinese Restaurant Algorithm for Finite Mixtures

The GWCR algorithm works by building up a sequence of nested partitions $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$ by assigning labels $\{1, \ldots, n\}$ into sets using a posterior *partition rule*. The nth partition, $\mathbf{p}_n$, is our eventual draw for $\mathbf{p}$ from q. To run the algorithm, we first assign $\mathbf{p}_1$ to be the set containing label 1, after which partitions are created sequentially so that at the rth stage, $\mathbf{p}_r$ contains $n(\mathbf{p}_r)$ sets derived from the labels $\{1, \ldots, r\}$. The partition rule for producing $\mathbf{p}_r$ depends on the configuration for the previous partition $\mathbf{p}_{r-1}$. Thus, if $C_{1,r-1}, \ldots, C_{m,r-1}$ are the $m = n(\mathbf{p}_{r-1})$ sets in $\mathbf{p}_{r-1}$ (derived from $\{1, \ldots, r-1\}$), then $\mathbf{p}_r$ is formed by assigning label r to one of the previous sets $C_{j,r-1}$ with probability

$$\rho_r(j) = \frac{e_{j,r-1} + \alpha/N}{\lambda(r)} \times \frac{\int_y f(X_r|y) \prod_{i \in C_{j,r-1}} f(X_i|y)\, H(dy)}{\int_y \prod_{i \in C_{j,r-1}} f(X_i|y)\, H(dy)}, \quad (8)$$

where $e_{j,r-1}$ are the number of labels in $C_{j,r-1}$, or by assigning label r to a new set with probability

$$\rho_r = \frac{\alpha(1 - m/N)}{\lambda(r)} \times \int_y f(X_r|y)\, H(dy), \quad (9)$$

where $\lambda(r)$ is the normalizing constant ensuring that $\rho_r + \sum_{j=1}^{m} \rho_r(j) = 1$.

Thus, producing a draw $\mathbf{p}$ from the GWCR density involves the following steps:

1. Set $\mathbf{p}_1 = \{1\}$. Generate partitions $\mathbf{p}_2, \ldots, \mathbf{p}_n$ *sequentially* using the partition rule defined by (8) and (9).
2. Set $\mathbf{p} = \mathbf{p}_n$. This is our draw from q with importance weight

$$\Lambda(\mathbf{p}) = \lambda(1) \times \cdots \times \lambda(n),$$

where $\lambda(1) = \int_y f(X_1|y)\, H(dy)$.

Remark 2. Note that the draws for $\mathbf{p}$ depend on the order of the data $\mathbf{X}$. For small sample sizes this sometimes can be important, and thus we always apply a *shuffle step* at each iteration by randomizing the order of the data. We applied this technique in all of our examples of Section 6.

3.2 Location Mixtures of Normals

Working out the assignment probabilities (8) and (9) is straightforward in conjugate models. We give the computations here for the case of finite location mixtures of normals. Later in Section 6 we study the performance of our algorithms applied to this model. In the case of nonconjugacy, however, it generally is not possible to work out the required integrals in closed form. In this case, one can apply an approximate GWCR algorithm (the N-GWCR algorithm), which avoids the need to compute such integrals (see Ishwaran and James 2000a for details.)

We consider finite normal mixture densities of the form

$$f_Q(x) = \int_{\mathcal{R}} \frac{1}{\sigma} \phi\left(\frac{x - y}{\sigma}\right) dQ(y), \quad (10)$$

where ϕ is the standard normal density, σ is a common standard deviation, and Q is a finite mixture distribution over $\mathcal{Y} = \Re$. For the moment, we assume that σ is a fixed known value. In the following section we consider the case where σ is unknown.

To take advantage of conjugacy, we take H to be a $\mathrm{N}(0, A)$ distribution in which A is taken to be a large number relative to the scale of the data to ensure noninformativeness (in our computations in Section 6 we use $A = 1000$). Elementary calculations show that

$$\rho_r(j) = \frac{e_{j,r-1} + \alpha/N}{\lambda(r)} \times \sqrt{\frac{\sigma^2 + A e_{j,r-1}}{2\pi\sigma^2\left[\sigma^2 + A(e_{j,r-1} + 1)\right]}} \times \exp\left[-\frac{1}{2\sigma^2}\left(X_r^2 - \frac{A\left(\sum_{i \in C_{j,r-1}} X_i + X_r\right)^2}{\sigma^2 + A(e_{j,r-1} + 1)} + \frac{A e_{j,r-1}^2 \overline{X}_{j,r-1}^2}{\sigma^2 + A e_{j,r-1}}\right)\right]$$

and that

$$\rho_r = \frac{\alpha(1 - m_{r-1}/N)}{\lambda(r)} \times \frac{1}{\sqrt{2\pi(\sigma^2 + A)}} \exp\left(-\frac{X_r^2}{2(\sigma^2 + A)}\right),$$

where $\overline{X}_{j,r-1}$ is the average of the data $\{X_i : i \in C_{j,r-1}\}$.

3.3 Location Mixtures of Normals With Unknown Variance

Often the value for σ will be unknown in (10). This is the case considered in Section 6, and thus we outline a simple modification to the foregoing algorithm to handle this scenario. This idea was discussed by Ishwaran, James, and Lo (2001) as a general method for updating finite-dimensional parameters within the GWCR algorithm for semiparametric models.

The general principle underlying the modification is that once we are given the partition structure $\mathbf{p}$, we are told how the data X_i are clustered by means, and thus our problem reduces to a collection of parametric normal models. Thus, given $\mathbf{p}$, it should be clear how to estimate σ using standard parametric arguments. This idea is used sequentially within the GWCR, with the value for σ updated after the generation of each partition $\mathbf{p}_r$. As worked out by Ishwaran et al. (2001), step 1 of the GWCR algorithm is thus modified so that a partition $\mathbf{p}_r$ is now generated using the partition rules defined by (8) and (9), but with σ replaced by its current estimate σ_{r-1}. After generating $\mathbf{p}_r$, replace σ^2 by the estimate

$$\sigma_r^2 = \frac{1}{r} \sum_{j=1}^{n(\mathbf{p}_r)} \sum_{i \in C_{j,r}} (X_i - \overline{X}_{j,r})^2.$$

Observe that σ_r^2 is the maximum likelihood estimator.

Remark 3. To avoid numerical problems in the early stages of the algorithm, we use a sampled value for σ. (In our example we generated σ_0^2, the starting value for σ^2, from a uniform $[0, 3]$ distribution.) After a fixed number of iterations (we used 10), we then begin updating σ using the above method.

Remark 4. The GWCR can also be applied straightforwardly in location-scale mixtures of normals. Although here we have focused only on the case where $\mathcal{Y} \subseteq \Re$, the GWCR algorithm in fact applies to much more general spaces (Brunner et al. 2001; Ishwaran and James 2000a). Letting $y = (u, s)$, we could also apply our GWCR model selection procedure to location-scale mixtures of normals of the form

$$f_Q(x) = \int_{\Re\times\Re^+} \frac{1}{s}\phi\Big(\frac{x-u}{s}\Big)\, dQ(u, s). \tag{11}$$

In this case the use of a conjugate normal-inverse gamma distribution for H leads to a simple closed-form expression for the partition update rule. Note that with in this approach, one does not need to estimate σ as in location mixtures of normals.

4. BLOCKED GIBBS SAMPLING

Another method for estimating the mixture dimension d as well as the mixing distribution Q_0 can be based on the blocked Gibbs sampler (Ishwaran and James 2001). In this approach, one uses a Gibbs sampler to draw values from the distribution of $(\mathbf{K}, \mathbf{Z}, \mathbf{W}|\mathbf{X})$, where $\mathbf{K} = (K_1, \ldots, K_n)$ and K_i are classification variables with the property that $Y_i = Z_{K_i}$ [thus $(Y_1, \ldots, Y_n)$ is equivalent to $(\mathbf{K}, \mathbf{Z})$] and chosen so that K_i are iid multinomial$(N, \mathbf{W})$. That is,

$$(K_i|\mathbf{W}) \overset{\text{iid}}{\sim} \sum_{k=1}^{N} W_k\, \delta_k(\cdot).$$

The blocked Gibbs sampler works by iteratively drawing values from the conditional distributions

$$(\mathbf{K}|\mathbf{Z}, \mathbf{W}, \mathbf{X}),$$
$$(\mathbf{Z}|\mathbf{K}, \mathbf{X}),$$

and

$$(\mathbf{W}|\mathbf{K}).$$

Each iteration generates a value $(\mathbf{K}^*, \mathbf{Z}^*, \mathbf{W}^*)$, which produces a draw,

$$\mathcal{P}_N^*(\cdot) = \sum_{k=1}^{N} W_k^* \delta_{Z_k^*}(\cdot),$$

from the posterior of $\mathcal{P}_N$ in (3) and thus can be used to directly estimate Q_0. (See Ishwaran and James 2001 and Ishwaran and Zarepour 2000a for more details.)

Remark 5. Each of the foregoing conditional distributions can be drawn exactly, including the draw from $\mathbf{Z}$ if H is a conjugate prior. In particular,

$$(\mathbf{W}|\mathbf{K}) \sim \text{Dirichlet}_N(\alpha/N + n_1, \ldots, \alpha/N + n_N),$$

where n_k is the number of K_i's that equal k.

Remark 6. The method can also be extended to include further parameters. For example, in the location mixture model with unknown variance (10), we can include σ as a parameter. Its conditional distribution $(\sigma|\mathbf{K}, \mathbf{Z}, \mathbf{X})$ is then updated in the blocked Gibbs sampler. A convenient prior for σ is

$$\sigma^{-2} \sim \text{gamma}(s_1, s_2).$$

This is the method used in Section 6. Note that by conjugacy, the required draw from $(\sigma|\mathbf{K}, \mathbf{Z}, \mathbf{X})$ is straightforward.

4.1 Penalized Stochastic Maximum Likelihood Estimation

A nice feature of the blocked Gibbs sampler is that the draws $\mathcal{P}_N^*$ obtained from the sampler can be used to produce an estimate for the dimension of the model as well as a point estimate for Q_0 using a penalized approach. This gives us a Bayesian penalized maximum likelihood estimation procedure for mixture models somewhat analogous to non-Bayesian penalization methods (see, e.g., Leroux 1992).

In this approach, we first replace $\mathcal{P}_N^*$ with a random measure $\widehat{\mathcal{P}}_N^*$ that more correctly reflects its *effective dimension*. Many of the probability weights W_k^* in $\mathcal{P}_N^*$ can be near 0, and thus we propose replacing $\mathcal{P}_N^*$ with

$$\widehat{\mathcal{P}}_N^*(\cdot) = \sum_{k=1}^{N} \frac{I\{n_k^* > 0\} W_k^*}{\sum_{k=1}^{N} I\{n_k^* > 0\} W_k^*}\, \delta_{Z_k^*}(\cdot),$$

where (as before) $n_k^* = \#\{i\colon K_i^* = k\}$. Thus we propose using a measure that retains only those weights that are nonnegligible, that is, only those weights whose values correspond to an atom Z_k^* that has been selected by some Y_i. This replaces $\mathcal{P}_N^*$ with its N components, with $\widehat{\mathcal{P}}_N^*$ having $n(\mathbf{K}^*)$ components and an effective number of parameters of $2n(\mathbf{K}^*) - 1$, where $n(\mathbf{K}^*)$ is the number of distinct values of $\mathbf{K}^*$ (i.e., the number of distinct clusters in $Y_1^*, \ldots, Y_n^*$).

The optimal $\widehat{\mathcal{P}}_N^*$ is that draw (over a large number of draws) with the largest value,

$$l_n(\widehat{\mathcal{P}}_N^*) - a_n(\widehat{\mathcal{P}}_N^*), \tag{12}$$

where $l_n(Q) = \sum_{i=1}^{n} \log f_Q(X_i)$ is the log-likelihood evaluated at a mixing distribution Q and $a_n(\widehat{\mathcal{P}}_N^*)$ is the penalty term for $\widehat{\mathcal{P}}_N^*$. Many penalty terms can be considered. One that we look at is Schwartz's Bayes information criterion (BIC) (Schwartz 1978), which corresponds to the penalty

$$a_n(\widehat{\mathcal{P}}_N^*) = \frac{1}{2}\log n \times \dim(\widehat{\mathcal{P}}_N^*) = \log n \times \Big(n(\mathbf{K}^*) - \frac{1}{2}\Big).$$

Another penalty term that we consider is the Akaike information criterion (AIC) (Akaike 1973), which corresponds to

$$a_n(\widehat{\mathcal{P}}_N^*) = \dim(\widehat{\mathcal{P}}_N^*) = 2n(\mathbf{K}^*) - 1.$$

We also consider a minimum distance penalty term proposed by Chen and Kalbfleisch (1996) corresponding to

$$a_n(\widehat{\mathcal{P}}_N^*) = -\sum_{k=1}^{N} I\{n_k^* > 0\} \log(W_k^*).$$

This penalizes small weights W_k^* and thus indirectly discourages large dimensions (see Chen and Kalbfleisch 1996 for more discussion.)

The optimal $\widehat{\mathcal{P}}_N^*$ that maximizes (12), produces our estimate for Q_0 as well as our estimate $\hat{d} = n(\mathbf{K}^*)$ for the model dimension d. In Section 6 we study the performance of this procedure in detail.

5. RATES OF ESTIMATION

The optimal rate of estimation for Q_0 when d is unknown but bounded by a finite N is $O_p(n^{-1/4})$. This result is from Chen (1995), who established $n^{-1/4}$ as the lower rate of estimation in this setting and then showed how the rate could be achieved using a minimum distance estimator. Although this result was derived from a purely frequentist perspective, it stands to reason that the same rate should continue to hold for any well-designed Bayesian method. Theorem 2 shows this to be the case for the posterior of (3), which achieves the $O_p(n^{-1/4})$ rate for Q_0 under conditions analogous to those used by Chen (1995). Key among these conditions is the notion of *strong identification* for the mixture model. Let

$$F(x|y) = \int_{-\infty}^{x} f(u|y)\, d\lambda(u)$$

denote the distribution function for $f(\cdot|y)$ and let $F_Q(x) = \int F(x|y)\, dQ(y)$ denote the distribution function for the mixed density f_Q.

Strong Identification (Chen 1995). We say that the mixture model is strongly identified if $F(\cdot|y)$ is twice differentiable in y and for any m distinct values $y_1, \ldots, y_m$ in $\mathcal{Y}$, the equality

$$\sup_x \left| \sum_{j=1}^{m} [\alpha_j F(x|y_j) + \beta_j F'(x|y_j) + \gamma_j F''(x|y_j)] \right| = 0$$

implies that $\alpha_j = \beta_j = \gamma_j = 0$.

As noted by Chen (1995), strong identification implies that the mixture model is identified over $\mathcal{Q}(N)$. Strong identification is known to hold for several important models. For example, both location and scale mixtures of normal distributions are strongly identified, as are mixtures of Poisson distributions (see Chen 1995 for details).

In addition to the assumption of strong identification, we also need a Lipschitz-type condition on $f'(\cdot|y)$. This and an assumption of compactness allow us to appeal to lemma 2 of Chen (1995), which makes it possible to bound the $\mathcal{L}_1$-squared distance between two mixing distributions by the Kolmogorov–Smirnov distance between the corresponding mixed distribution functions. Recall that the Kolmogorov-Smirnov distance between two distribution functions F_{Q_1} and F_{Q_2} is defined by

$$D_K(F_{Q_1}, F_{Q_2}) = \sup_x |F_{Q_1}(x) - F_{Q_2}(x)|.$$

This square-root distance relationship implied by Chen's lemma is what converts an $n^{-1/2}$ rate for estimating F_{Q_0} into the optimal $n^{-1/4}$ rate for estimating Q_0.

Smoothness Condition. We say that the mixture model satisfies a smoothness condition if the following conditions hold:

(a) Each atom $Y_{k,0}$ of Q_0 is an interior point of $\mathcal{Y}$.
(b) For each $x \in \mathcal{X}$, the first and second derivatives of $\log(f(x|y))$ exist for y. Moreover, for each interior point $y \in \mathcal{Y}$, the derivatives are continuous in some open neighborhood of y and can be bounded in absolute value by a square integrable function of $\mathcal{L}^2(P_0)$.
(c) For each $x \in \mathcal{X}$, there exists a fixed constant $C > 0$ and a value $\eta > 0$ such that

$$|f'(x|y_1) - f'(x|y_2)| \le C|y_1 - y_2|^{\eta}$$

for all $y_1, y_2 \in \mathcal{Y}$.

Condition (c) represents our Lipschitz condition, whereas (a) and (b) are the key conditions needed for establishing the $n^{-1/2}$ rate for F_{Q_0}. (These are usually straightforward to check for exponential families and hold for the examples mentioned earlier). Let

$$v(Q_1, Q_2) = \sup_{B \in \mathcal{B}} |Q_1(B) - Q_2(B)| \tag{13}$$

equal the total variation distance between two distributions, Q_1 and Q_2, over $(\mathcal{Y}, \mathcal{B})$.

Theorem 2. Suppose that the mixture model satisfies the smoothness conditions (a) and (b). Assume also that H has a density that is positive and continuous over $\mathcal{Y}$. Let K_n be any positive increasing sequence such that $K_n \to \infty$. Then, if $Q_0 \in \mathcal{Q}(N)$,

$$\Pi_n\{D_K(F_Q, F_{Q_0}) \ge n^{-1/2} K_n\} \to 0 \quad \text{in } P_0^{\infty} \text{ probability}, \tag{14}$$

where $\Pi_n(\cdot) = \mathbb{P}\{\cdot|X_1, \ldots, X_n\}$ is the posterior for (3). In addition, if the mixture model is strongly identified and satisfies the smoothness condition (c), and if $\mathcal{Y}$ is compact, then

$$\Pi_n\{v(Q, Q_0) \ge n^{-1/4} K_n\} \to 0 \quad \text{in } P_0^{\infty} \text{ probability}. \tag{15}$$

Note that (15) implies the optimal $O_p(n^{-1/4})$ rate for Q_0.

Remark 7. One might wonder what rates of estimation are possible in the finite mixture model in which the dimension d is unbounded: $1 \le d < \infty$. Rates for Q_0 appear to be an open question in this setting, although recent results for density estimation in normal mixture models (Genovese and Wasserman 2000) hint at rates for the mixed density f_{Q_0} that are slower than the rate implied by (14). Consider, for example, finite location normal mixture models with $Q_0 \in \mathcal{Q}(N)$. From (14), the posterior concentrates on $n^{-1/2}$ neighborhoods of the distribution function F_{Q_0} in the Kolmogorov distance, which in turn implies that the posterior is $O_p(n^{-1/2})$ consistent for the density f_{Q_0} in an $\mathcal{L}_1$ sense. This follows from the bound between the Kolmogorov distance and the $\mathcal{L}_1$ distance for finite normal mixtures (Cao and Devroye 1996),

$$\int_{\Re} |f_Q(x) - f_{Q_0}(x)| dx \le C_N \times D_K(F_Q, F_{Q_0}),$$

where C_N is a finite constant depending only on N. We suspect that the foregoing $\sqrt{n}$-parametric rate for the density is

faster than $\mathcal{L}_1$ rates for finite normal mixtures with unbounded complexity. For example, Genovese and Wasserman (2000), using a sieve of dimension of order $O(\sqrt{n/\log n})$, showed that the density of compact mixtures of location-scale normals (11) could be estimated at an $\mathcal{L}_1$ rate of $O_p(n^{-1/4}\log n)$. Although this result was derived by considering the more general class of mixing distributions with compact support (which covers the unbounded d case), and hence the much slower rate of estimation, we believe that such results lend strength to the conjecture that rates should naturally be slower when d is unbounded.

6. EXAMPLES

We begin by investigating the performance of the GWCR algorithm and the blocked Gibbs sampler of Sections 3 and 4 applied to a collection of well-known datasets. Each of these five examples were analyzed using a finite location mixture model (10) with unknown standard deviation as discussed in Sections 3.2 and 3.3. We used $A = 1000$ for the variance in the $\mathrm{N}(0, A)$ prior for H in both the GWCR algorithm and blocked Gibbs sampler and $s_1 = s_2 = .01$ for the shape and scale parameters used in the inverse gamma prior for σ^2 in the blocked Gibbs sampler. For convenience, a simple description as well as further references to these data are given next. Most of these datasets can be retrieved from Geoff McLachlan's web page (www.maths.uq.edu.au/~gjm).

Galaxy Data. This dataset, from Roeder (1990), comprises 82 observations of relative velocities in thousands of kilometers per second of galaxies from six well-separated conic sections of space. The original data have been divided by 1000 here (thus velocity is recorded as kilometers per second). As discussed by Roeder (1990), there is strong evidence to believe that modes in the data correspond to clumped galaxies and that observed velocities are values derived from a finite location mixture of normals.

Sodium-lithium countertransport (SLC) data. These data consist of red blood cell SLC activity measured on 190 individuals. As argued by Roeder (1994), it is believed that the SLC measurements are derived from one of two competing genetic models, corresponding to either a two-point or three-point normal mixture model. The original data have been multiplied by 10 here.

Hidalgo Stamp Data. This dataset comprises 485 observations of the stamp thickness (in millimeters) of the 1872–1874 Hidalgo postage stamps of Mexico. A detailed analysis by Izenman and Sommer (1988) argues that the data are derived from a finite mixture of normals, thus providing evidence that the stamps were issued on several different types of paper. The original data have been multiplied by 100 here.

Acidity Data. This dataset is an acidity index measured on 155 lakes in north-central Wisconsin (see Richardson and Green 1997). Data are measured on the log scale.

Enzyme Data. This dataset comprises measurements of enzymatic blood activity in 245 individuals. The focus here is in identifying subgroups of slow or fast metabolizers as a marker of genetic polymorphism via a finite normal mixture model (see Richardson and Green 1997). The original data have been multiplied by 10 here.

The results from the GWCR algorithm are presented in Table 1 and results from the blocked Gibbs sampler appear in Tables 2, 3, and 4 under BIC, minimum distance (MD), and AIC penalty (see Sec. 4). Estimates $\hat{d}$ for the dimension d from the GWCR agree with the blocked Gibbs sampler for three of the datasets under BIC and four of the datasets under AIC and MD. However, we note that the various point estimates for Q_0 tend to agree closely. The discrepancy in the estimates of $\hat{d}$ under the different penalties is generally due to the appearance of atoms with small probabilities. Nevertheless,

Table 1. GWCR Algorithm Based on 150,000 iid Values Under the Dirichlet Prior With $\alpha = 1$ and $N = 15$. Values Reported Are Means Plus or Minus Standard Deviations Over 20 Blocks of Values Each of Size 7,500 for the Weighted Bayes Factors $\Delta(k,k^*)$, Where $m_{k^*,N}$ is the Largest $m_{k,N}$ Value Within a Block (thus $\Delta(k,k^*) \le 1$ within each block).

k	Galaxy	slc	Stamp	Acidity	Enzyme
1	0	0	0	0	0
2	0	$4.8^{-1} \pm 2^{-1}$	0	$9.9^{-1} \pm 2^{-1}$	0
3	$2.8^{-1} \pm 1^{-1}$	$9.9^{-1} \pm 7^{-2}$	0	$4.1^{-1} \pm 2^{-1}$	0
4	$4.4^{-1} \pm 1^{-1}$	$2.5^{-1} \pm 1^{-1}$	0	$9.1^{-2} \pm 6^{-2}$	0
5	$3.5^{-1} \pm 2^{-1}$	$2.3^{-2} \pm 4^{-2}$	0	$2.1^{-2} \pm 2^{-3}$	0
6	$9.9^{-1} \pm 6^{-2}$	0	0	0	$9.5^{-2} \pm 1^{-1}$
7	$4.3^{-1} \pm 2^{-1}$	0	$5.3^{-2} \pm 2^{-1}$	0	$4.1^{-1} \pm 3^{-1}$
8	$8.2^{-2} \pm 1^{-1}$	0	$7.2^{-1} \pm 4^{-1}$	0	$7.7^{-1} \pm 3^{-1}$
9	$5.2^{-3} \pm 1^{-2}$	0	$3.9^{-1} \pm 5^{-1}$	0	$4.8^{-1} \pm 4^{-1}$
10	0	0	$1.8^{-2} \pm 6^{-2}$	0	$2.3^{-1} \pm 3^{-1}$
11	0	0	0	0	$1.2^{-2} \pm 2^{-2}$
12	0	0	0	0	0
13	0	0	0	0	0
14	0	0	0	0	0
15	0	0	0	0	0
n	82	190	485	155	245
$\hat{d}$	6	3	8	2	8

NOTE: Superscripts are used to indicate values raised to the power 10 (thus $a^{-b} = a \times 10^{-b}$). Entries have been set to zero for mean values less than 10^{-4}. The value $\hat{d}$ is the estimate for d and represents the largest mean value for $\Delta(k,k^*)$

Table 2. Results From the Blocked Gibbs Sampler using a 2,000 Iteration Burn-in Followed by 25,000 Sampled Iterations. Prior Based on Dirichlet Parameters With $\alpha = 1$ and $N = 15$. Each Column Displayed Contains the Probabilities and Atoms for the Top Model $\widehat{Q}$ Subject to a BIC Penalty

Galaxy		slc		Stamp		Acidity		Enzyme	
pr	*atoms*	*pr*	*atoms*	*pr*	*atoms*	*pr*	*atoms*	*pr*	*atoms*
0.44	19.87	0.87	2.36	0.37	7.92	0.63	4.38	0.61	1.98
0.36	22.96	0.13	4.46	0.26	7.19	0.37	6.33	0.16	9.48
0.09	9.78	—	—	0.13	10.01	—	—	0.13	13.07
0.05	26.20	—	—	0.10	10.92	—	—	0.07	17.46
0.04	33.11	—	—	0.08	9.06	—	—	0.03	23.68
0.02	16.14	—	—	0.04	11.97	—	—	0.003	29.31
—	—	—	—	0.02	12.92	—	—	—	—
—	—	—	—	0.004	6.22	—	—	—	—

the conclusion here seems to be that GWCR tends to agree somewhat more closely with AIC and MD than with BIC. In the following section we further assess the accuracy of GWCR through the use of simulations.

Remark 8. One could also fit the foregoing data using a location-scale mixture model (11). As mentioned earlier, the use of a conjugate normal-inverse gamma prior in this model would lead to a simple update rule for the GWCR algorithm, although now several hyperparameters would need to be carefully chosen. For example, in selecting the hypervariance for the mean we would use a large value to encourage noninformativeness. Just as in selecting A here, this value should be chosen so that it is large relative to the scale of the data. The blocked Gibbs sampler can also be applied to location-scale mixture models (11) with appropriate modifications. Ishwaran and James (2000b) have given computational details as well as a discussion on selecting priors and hyperparameters.

6.1 Simulations

To further assess our procedures, we ran 10 different simulations in which simulated data were drawn from a finite location normal mixture with unknown variance (10). Experiments 1–6 used a sample size of $n = 100$, while Experiments 7–10 were based on $n = 400$. Each simulation experiment was repeated 500 times, with each sample drawn independently from a location mixture of normal densities of the form

$$f(x|\pi, \mu, \sigma) = \sum_{k=1}^{d} \frac{\pi_k}{\sigma} \phi\Big(\frac{x - \mu_k}{\sigma}\Big),$$

where $\pi = (\pi_1, \ldots, \pi_d)$ is a positive weight vector summing to one, σ is chosen to be one, and $\mu = (\mu_1, \ldots, \mu_d)$ are prespecified means. Figure 2 gives the exact specifications. Note that two components in a mixture model appear graphically as separate modes (see Fig. 2) if and only if their mean difference is larger than $2\sigma = 2$.

The samples were carefully simulated so that the change from one simulation experiment to the next reflects only the change of parameters π and μ and not the random fluctuation from changing random seeds. Each experiment was analyzed by the EM algorithm under AIC and BIC penalties. Tables 5 and 6 present these results, as well as results from the GWCR algorithm. Note that in constructing the tables, the estimate $\hat{d}$ for the dimension d corresponded to the largest AIC and BIC values for EM, and for the GWCR procedure it was the value k with the largest estimated posterior probability $\mathbb{P}\{n(\mathbf{p}) = k|\mathbf{X}\}$.

It is worth noting that good initial values for the EM algorithm are important. For producing initial values, in each experiment we randomly selected one of our 500 samples, then plotted three histograms of these data under three different bandwidth values. Based on these histograms, for each candidate dimension $k = 1, \ldots, 15$, we guessed reasonable initial values of π, μ, and σ. Because of the laborious nature of this work, we used these same initial values for each of the 500 samples in an experiment. We found that the final estimate $\hat{d}$ based on either AIC or BIC was less sensitive to the initial value of σ than to that of μ or π. Thus we simply used $\sigma = 1$ as the initial value. To make this a fair comparison, we also started the GWCR algorithm with an initial value

Table 3. Similar Analysis as in Table 2 Using the Blocked Gibbs Sampler but with Models Chosen Subject to a Minimum Distance Penalty (note that analysis is based on a new set of Gibbs sampled values)

Galaxy		slc		Stamp		Acidity		Enzyme	
pr	*atoms*	*pr*	*atoms*	*pr*	*atoms*	*pr*	*atoms*	*pr*	*atoms*
0.44	19.86	0.75	2.23	0.36	7.92	0.57	4.32	0.61	1.88
0.30	22.78	0.22	3.76	0.26	7.19	0.26	6.53	0.13	9.79
0.09	9.75	0.03	5.66	0.12	10.03	0.17	5.72	0.12	12.95
0.09	25.44	—	—	0.10	10.98	—	—	0.05	18.56
0.05	32.99	—	—	0.09	9.09	—	—	0.03	16.04
0.03	16.28	—	—	0.04	11.98	—	—	0.03	23.88
—	—	—	—	0.02	12.92	—	—	0.03	6.86
—	—	—	—	0.01	6.36	—	—	0.004	28.63

Table 4. Similar Analysis as in Table 2 Using the Blocked Gibbs Sampler but with Models Chosen Subject to an AIC Penalty (analysis is based on sampled values different than Table 2 and Table 3)

Galaxy		*slc*		*Stamp*		*Acidity*		*Enzyme*	
pr	*atoms*	*pr*	*atoms*	*pr*	*atoms*	*pr*	*atoms*	*pr*	*atoms*
0.44	19.91	0.76	2.22	0.38	7.94	0.52	4.26	0.63	2.00
0.36	23.05	0.21	3.73	0.26	7.20	0.20	6.67	0.13	9.59
0.10	9.81	0.03	5.67	0.13	9.99	0.16	5.98	0.12	12.85
0.04	33.07	—	—	0.09	10.96	0.12	5.08	0.05	15.81
0.04	26.23	—	—	0.08	9.05	0.005	3.04	0.03	18.77
0.02	15.90	—	—	0.03	12.02	—	—	0.02	23.85
—	—	—	—	0.02	12.90	—	—	0.02	6.64
—	—	—	—	0.01	6.34	—	—	0.003	28.65

of $\sigma = 1$. Updates for σ within the GWCR were then carried out as in our examples in the previous section, as discussed in Section 3.3.

From these two tables, it is clear that the GWCR procedure was the winner in experiments 1 and 2; in all other experiments, GWCR and EM-AIC generally performed better than EM-BIC. Specifically, in experiment 3, consisting of only one apparent mode, all procedures tended to favor a one-component model, although the EM-AIC procedure did uncover the true two-component model 26.4% of the time. In experiment 4, with four close modes, EM-AIC was the winner and GWCR did reasonably well. In experiment 5, both GWCR and EM-AIC usually recognized that the flat mode must be a mixture of distributions although GWCR picked this up more frequently. In experiment 6, with two apparent modes, all procedures favored a two-component model. In experiment 7, EM-AIC was the winner and GWCR did better than the EM-BIC procedure. In experiments 8, 9, and 10, all procedures tried to defragment flat mode(s) into mixtures; EM-AIC and GWCR did a better job than EM-BIC. Thus overall, GWCR and EM-AIC outperformed each other in some circumstances, and both procedures were better than EM-BIC in general. However, we emphasize once more that the GWCR is a more automatic procedure than the EM-AIC in the sense that GWCR is less sensitive to initial values, which are crucial in EM-based procedures.

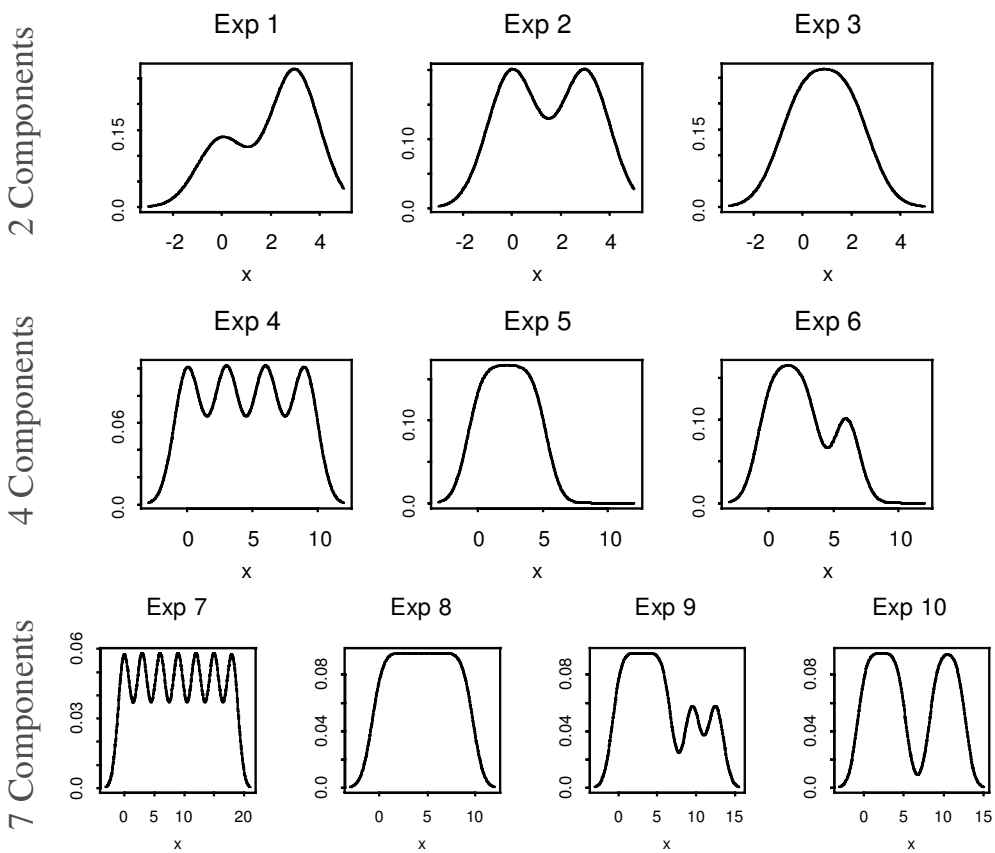

Figure 2. True Mixture Densities used in Simulation Experiments 1-10. In experiment 1, $\pi = (1/3,2/3)$ while in all other experiments π is a uniform weight vector. In experiments 1 – 3, $d = 2$ and $\mu = (0,3),(0,3),(0,1.8)$, respectively. In experiments 4 – 6, $d = 4$ and $\mu = (0,3,6,9),(0,1.5,3,4.5),(0,1.5,3,6)$, respectively. In experiments 7 – 10, $d = 7$ and $\mu = (0,3,6,9,12,15,18),(0,1.5,3,4.5,6,7.5,9),(0,1.5,3,4.5,6,9.5,12.5),(0,1.5,3,4.5,9,10.5,12)$ respectively.

Table 5. Simulations 1–6: Entries in the Last Three Columns are Percentages of Times Out of the 500 Samples for Which $\hat{d}$ (estimate for d) Equals a Candidate Dimension Value k. Entries for AIC and BIC are Derived From the EM Algorithm, While Last Column is Via the GWCR Algorithm and is Based on 2500 iid Values with $\alpha = 1$ and $N = 15$. Percentages Highlighted by Boxes Indicate Highest Value and Thus Represent the Best Model for a Specific Procedure

Simulation	n	d	#Modes	k	AIC	BIC	GWCR
				1	0.018	0.150	0.018
				2	0.896	0.838	0.920
1	100	2	2	3	0.062	0.012	0.058
				4	0.024	0.000	0.004
				5	0.000	0.000	0.000
				1	0.022	0.212	0.030
				2	0.900	0.780	0.916
2	100	2	2	3	0.050	0.006	0.054
				4	0.028	0.002	0.000
				5	0.000	0.000	0.000
				1	0.702	0.968	0.868
				2	0.264	0.030	0.130
3	100	2	1	3	0.024	0.002	0.002
				4	0.010	0.000	0.000
				5	0.000	0.000	0.000
				1	0.000	0.110	0.000
				2	0.178	0.596	0.102
4	100	4	4	3	0.110	0.110	0.554
				4	0.674	0.182	0.306
				5	0.038	0.002	0.038
				1	0.244	0.748	0.144
				2	0.556	0.246	0.818
5	100	4	1	3	0.142	0.004	0.032
				4	0.044	0.002	0.006
				5	0.014	0.000	0.000
				1	0.016	0.188	0.000
				2	0.474	0.698	0.612
				3	0.392	0.106	0.368
6	100	4	2	4	0.102	0.008	0.020
				5	0.014	0.000	0.000
				6	0.000	0.000	0.000
				7	0.002	0.000	0.000

APPENDIX: PROOFS

Proof of Proposition 1

It is clear that we can construct $\mathcal{F}$ so it contains the singleton sets $\{x_m\}$ where $x_m \in \mathcal{X}$ are chosen so that $x_m \to x_0$ as $m \to \infty$. Now assume that P_{Q_1} and P_{Q_2} agree over $\mathcal{F}$. Then $f_{Q_1}(x_m) = f_{Q_2}(x_m)$ for each m. That is, for each m,

$$\sum_{i=1}^{k} p_i f(x_m|y_i) = \sum_{j=1}^{k'} p'_j f(x_m|y'_j) \tag{A.1}$$

for some probability weights $0 < p_i, p'_j < 1$, where $\sum_i p_i = \sum_j p'_j = 1$ and y_i and y'_j are atoms in $\mathcal{Y}$.

We can assume without loss of generality that $s(x) \to +\infty$ as $x \to x_0$ and that y_i and y'_j are ordered so that $y_1 > y_2 > \cdots > y_k$ and $y'_1 > y'_2 > \cdots > y'_{k'}$. Furthermore, we can also assume that $y'_1 \le y_1$. First, consider the case where y'_1 is strictly smaller than y_1. If we divide the left and right sides of (A.1) by $f(x_m|y_1)$ and let $m \to \infty$, then the left side converges to p_1 and the right side converges to 0. Thus it must be that $y_1 = y'_1$. Letting m converge to infinity again now shows that $p_1 = p'_1$. Thus $y_1 = y'_1$ and $p_1 = p'_1$, which allows us to cancel the terms corresponding to $i = 1$ on the left side and $j = 1$ on the right side of (A.1). Repeat the foregoing argument a finite number of times to deduce that $Q_1 = Q_2$.

Proof of Theorem 1

By bounding $m_{k,N}$ above by $m_k \times r_{k,N}$, we have

$$\begin{aligned}\frac{m_{d,N}(\mathbf{X})}{m_{k,N}(\mathbf{X})} &\ge \frac{m^*_{d,N}(\mathbf{X})}{m_k(\mathbf{X})} \times \frac{r^*_{d,N}}{r_{k,N}} \\ &= \left[\frac{m_d(\mathbf{X})/f_0^n(\mathbf{X})}{m_k(\mathbf{X})/f_0^n(\mathbf{X})} + \frac{(m^*_{d,N} - m_d)(\mathbf{X})/f_0^n(\mathbf{X})}{m_k(\mathbf{X})/f_0^n(\mathbf{X})}\right] \times \frac{r^*_{d,N}}{r_{k,N}}. \end{aligned} \tag{A.2}$$

Because $k < d$, deduce that

$$\frac{r^*_{d,N}}{r_{k,N}} = \mathcal{P}_d(y_d)\binom{N}{d}\binom{N}{k}^{-1}\frac{(d\alpha/N)^{(n)}}{(k\alpha/N)^{(n)}} \ge \mathcal{P}_d(y_d)\binom{N}{d}\binom{N}{k}^{-1},$$

which is strictly bounded away from 0. Thus we need only consider the two terms in square brackets in (A.2).

We start with the first term. For each $\epsilon > 0$ let $\mathcal{N}_\epsilon = \{Q\colon D(P_0||P_Q) < \epsilon\}$. By restricting the range of integration in m_d to

Table 6. Simulations 7–10: Format and Methods Used are Similar to that Described in Table 5

Simulation	n	d	#Modes	k	AIC	BIC	GWCR
				1	0.004	0.816	0.000
				2	0.000	0.000	0.000
				3	0.000	0.000	0.010
				4	0.302	0.168	0.188
7	400	7	7	5	0.212	0.016	0.424
				6	0.098	0.000	0.178
				7	0.326	0.000	0.114
				8	0.036	0.000	0.056
				9	0.022	0.000	0.030
				1	0.030	0.538	0.000
				2	0.684	0.462	0.078
				3	0.000	0.000	0.590
8	400	7	1 flat region	4	0.248	0.000	0.272
				5	0.000	0.000	0.048
				6	0.012	0.000	0.008
				7	0.024	0.000	0.004
				8	0.002	0.000	0.000
				1	0.002	0.458	0.000
				2	0.000	0.000	0.002
				3	0.144	0.398	0.120
9	400	7	2+1 flat region	4	0.460	0.138	0.408
				5	0.308	0.006	0.312
				6	0.048	0.000	0.128
				7	0.016	0.000	0.024
				8	0.022	0.000	0.006
				1	0.000	0.000	0.000
				2	0.496	0.992	0.020
				3	0.000	0.000	0.370
10	400	7	1+1 flat region	4	0.302	0.006	0.466
				5	0.118	0.002	0.128
				6	0.064	0.000	0.010
				7	0.016	0.000	0.006
				8	0.004	0.000	0.000

the set $\mathcal{N}_{\epsilon/2}$, we have

$$\frac{m_d(\mathbf{X})/f_0^n(\mathbf{X})}{m_k(\mathbf{X})/f_0^n(\mathbf{X})} \geq \frac{\int_{\mathcal{N}_{\epsilon/2}} \exp\bigl(-nD_n(P_0^n||P_Q^n)\bigr)\mathcal{P}_d(dQ)}{m_k(\mathbf{X})/f_0^n(\mathbf{X})}, \qquad \text{(A.3)}$$

where

$$D_n(P_0^n||P_Q^n) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f_0(X_i)}{f_Q(X_i)}\right).$$

We deal with the numerator and denominator of (A.3) separately. For the numerator, we use the following argument from Barron (1988) (see also Verdinelli and Wasserman 1998, p. 1232). By the strong law of large numbers, $D_n(P_0^n||P_Q^n)$ converges almost surely to $D(P_0||P_Q) < \epsilon/2$ for each Q in $\mathcal{N}_{\epsilon/2}$. Therefore, a combination of Fubini's theorem and Fatou's lemma implies that

$$\begin{aligned}\liminf_{n\to\infty} \exp(n\epsilon)\int_{\mathcal{N}_{\epsilon/2}} \exp\bigl(-nD_n(P_0^n||P_Q^n)\bigr)\mathcal{P}_d(dQ)\\ \geq \int_{\mathcal{N}_{\epsilon/2}} \liminf_{n\to\infty} \exp\bigl(n\bigl[\epsilon - D_n(P_0^n||P_Q^n)\bigr]\bigr)\mathcal{P}_d(dQ)\\ = +\infty, \qquad \text{almost surely.}\end{aligned}$$

Note that we have used the assumption of information denseness: $\mathcal{P}_d(\mathcal{N}_{\epsilon/2}) > 0$. Thus the numerator in (A.3) is bounded below by $\exp(-n\epsilon)$ almost surely.

For the denominator in (A.3), use Markov's inequality to get

$$\begin{aligned}P_0^n\left\{\frac{m_k}{f_0^n} > \exp(-2n\epsilon)\right\} &\leq \exp(n\epsilon)P_0^n\left(\sqrt{\frac{m_k}{f_0^n}}\right)\\ &\leq \exp(n\epsilon)\bigl[1 - v(P_0^n, M_k)^2\bigr]^{1/2}, \qquad \text{(A.4)}\end{aligned}$$

where M_k is the distribution for m_k and our last bound is a well-known inequality relating the Hellinger distance to the total variation distance (13).

Now we use the assumption of $\mathcal{F}$-identification. By lemma 2 of Barron (1989), for each $\delta > 0$ there exists a set A_n and an $r > 0$ so that

$$P_0^n(A_n) \geq 1 - \exp(-nr)$$

and

$$P_Q^n(A_n) \leq \exp(-nr), \qquad \text{for each } Q \in U_\delta^c,$$

where $U_\delta = \{Q\colon d_{\mathcal{F}}(P_Q, P_0) < \delta\}$ and

$$d_{\mathcal{F}}(P_Q, P_0) = \sum_{A\in\mathcal{F}} |P_Q(A) - P_0(A)|.$$

A set such as A_n is often called a UEC test because it discriminates uniformly and exponentially well between a simple hypothesis (here Q_0) and a class of alternative hypotheses (our set U_δ^c). Schwartz (1965) has provided background on UEC tests.

By $\mathcal{F}$-identification we can choose a small enough $\delta > 0$ so that $\mathcal{Q}(k) \subseteq U_\delta^c$ (recall that $\mathcal{Q}(k) = \bigcup_{l=1}^{k} \mathcal{Q}_l$). If this were not the case,

then we could find a sequence $Q_j \in \mathcal{Q}(k)$ with limit Q^*, a distribution with at most $k < d$ atoms, such that $d_{\mathcal{F}}(P_{Q_j}, P_0) \to d_{\mathcal{F}}(P_{Q^*}, P_0) = 0$. However, this would contradict the assumption of $\mathcal{F}$-identification over $\mathcal{Q}(d)$. Therefore, $\mathcal{Q}(k) \subseteq U_\delta^c$ for a small enough $\delta > 0$, and hence

$$\begin{aligned} v(P_0^n, M_k) &\geq P_0^n(A_n) - M_k(A_n) \\ &\geq 1 - \exp(-nr) - \int_{\mathcal{Q}(k)} P_Q^n(A_n)\mathcal{P}_k(dQ) \\ &\geq 1 - 2\exp(-nr). \end{aligned}$$

Therefore, if ϵ is small enough, then the right side of (A.4) is exponentially small. Thus by the Borel–Cantelli lemma, $m_k/f_0^n \leq \exp(-2n\epsilon)$ almost surely for all large enough n. Therefore, deduce that (A.3) is almost surely bounded below by $\exp(n\epsilon)$ for a small enough $\epsilon > 0$.

Now we need to consider the second term in (A.2)

$$\frac{(m_{d,N}^* - m_d)/f_0^n}{m_k/f_0^n}. \tag{A.5}$$

By Markov's inequality,

$$\begin{aligned} P_0^n\left\{\frac{|m_{d,N}^* - m_d|}{f_0^n} > \exp(-4n\epsilon)\right\} &\leq \exp(4n\epsilon)\int_{\mathcal{X}^n} |m_{d,N}^*(\mathbf{X}) \\ &\quad - m_d(\mathbf{X})|d\lambda^n(\mathbf{X}) \\ &\approx 2d\exp\bigl(-n(1/d - 4\epsilon)\bigr). \end{aligned}$$

The last inequality is due to our Lemma 1. Thus, by the Borel–Cantelli lemma, the numerator in (A.5) is almost surely bounded in absolute value by $\exp(-4n\epsilon)$ for a small enough $\epsilon > 0$ for all large enough n. To handle the denominator in (A.5), notice that the inequality (A.2) and our arguments dealing with the first term in the square brackets of (A.2) hold if we replace m_k/f_0^n with the upper bound

$$I\{m_k/f_0^n > \exp(-3n\epsilon)\}m_k/f_0^n + \exp(-3n\epsilon)I\{m_k/f_0^n \leq \exp(-3n\epsilon)\}.$$

Thus, without loss of generality, we can assume that $\exp(-3n\epsilon) \leq m_k/f_0^n \leq \exp(-2n\epsilon)$. Deduce that (A.5) is almost surely 0. This takes care of all of the terms in (A.2).

Proof of Theorem 2

If (14) holds, then the optimal rate (15) follows automatically as a consequence of lemma 2 of Chen (1995). More precisely, the assumption of the Lipschitz condition combined with compactness of $\mathcal{Y}$ and strong identification implies the existence of a universal constant $0 < C_0 < \infty$ such that

$$C_0 v(Q, Q_0)^2 \leq D_K(F_Q, F_{Q_0}) < n^{-1/2}K_n$$

holds eventually for all $Q \in \mathcal{Q}(N)$. (We are assuming with no loss of generality that $n^{-1/2}K_n \to 0$.)

Thus we only need to prove (14). To do so, we modify the approach given by Clarke and Barron (1990) for establishing consistency in Bayesian parametric models (see also Barron 1988). Let A_n denote the posterior probability in (14). We show that $P_0^n(A_n) \to 0$.

By Dvoretzky, Kiefer, and Wolfowitz (1956) (see also Massart 1990),

$$P_Q^n\bigl\{D_K(\widehat{F}_n, F_Q) \geq n^{-1/2}K_n\bigr\} \leq \exp(-rK_n^2),$$

where $r > 0$ is a universal constant independent of P_Q and $\widehat{F}_n$ is the empirical distribution function based on P_Q. From this, it follows that there exists a UEC test $0 \leq \xi_n \leq 1$ such that

$$P_0^n\xi_n \leq \exp(-rK_n^2)$$

and

$$P_Q^n(1 - \xi_n) \leq \exp(-rK_n^2) \quad \text{for each } Q \in \mathcal{N}_n^c,$$

where $\mathcal{N}_n = \{Q\colon D_K(F_Q, F_{Q_0}) < n^{-1/2}K_n\}$ (see, e.g., Clarke and Barron 1990, p. 468).

From the foregoing we have $P_0^n(A_n\xi_n) \leq \exp(-rK_n^2)$, and thus

$$\begin{aligned} P_0^n(A_n) &= P_0^n(A_n\xi_n) + P_0^n(A_n(1 - \xi_n)) \\ &\leq \exp(-rK_n^2) + P_0^n(A_n(1 - \xi_n)). \end{aligned} \tag{A.6}$$

Moreover,

$$\begin{aligned} P_0^n(A_n(1 - \xi_n)) &= P_0^n(A_nB_n(1 - \xi_n)) + P_0^n(A_nB_n^c(1 - \xi_n)) \\ &\leq P_0^n(B_n) + P_0^n(A_nB_n^c(1 - \xi_n)), \end{aligned}$$

where

$$B_n = \left\{\frac{\mu_n(\mathbf{X})}{f_0^n(\mathbf{X})} < \exp(-r'K_n^2)\right\}$$

for some $r' > 0$ and

$$\mu_n(\mathbf{X}) = \int_{U_n}\prod_{i=1}^n f_Q(X_i)\mathcal{P}_N(dQ),$$

where $U_n = \{Q\colon D(P_0||P_Q) \leq \epsilon_n\}$ and ϵ_n is a positive sequence such that $\epsilon_n \to 0$.

Notice that

$$P_0^n(B_n) \leq P_0^n\left\{\left|\log\left(\frac{f_0^n}{\mu_n^*}\right)\right| \geq r'K_n^2 + \log(\mathcal{P}_N(U_n))\right\}, \tag{A.7}$$

where $\mu_n^* = \mu_n/\mathcal{P}_N(U_n)$ (observe that this is a density). Smoothness conditions (a) and (b) coupled with the assumption on H implies a type of continuity in $D(P_0||P_Q)$ for mixtures $Q \in \mathcal{Q}_d$ whose atoms and weights are near those of Q_0. In fact, the proof of lemma 4 of Ishwaran (1998) implies that $D(P_0||P_Q)$ has order equal to the squared Euclidean distance between the atoms and weights for Q_0 and a nearby Q. Our choice of prior $\mathcal{P}_N$ puts positive mass on distributions Q that are arbitrarily close to the set $\mathcal{Q}_d$. Thus elementary calculations, coupled with lemma 4 of Ishwaran (1998), show that

$$\log(\mathcal{P}_N(U_n)) \geq C\log(\epsilon_n)$$

for some $C > 0$. Thus we can replace $r'K_n^2 + \log(\mathcal{P}_N(U_n))$ in (A.7) with $r'K_n^2/2$ if we choose ϵ_n so that $\log(\epsilon_n)/K_n^2 \to 0$.

Let $\mathcal{M}_n^*$ denote the distribution for μ_n^*. Using Markov's inequality, bound (A.7) by

$$\begin{aligned} \frac{2}{r'K_n^2}P_0^n\left|\log\left(\frac{f_0^n}{\mu_n^*}\right)\right| &\leq \frac{2}{r'K_n^2}[D(P_0^n||\mathcal{M}_n^*) + 2/e] \\ &\leq \frac{2}{r'K_n^2}\left[\frac{n}{\mathcal{P}_N(U_n)}\int_{U_n} D(P_0||P_Q)\mathcal{P}_N(dQ) + 2/e\right] \\ &= O(n\epsilon_n/K_n^2) + O(1/K_n^2) = o(1), \end{aligned}$$

where the first inequality on the right is due to bounding the negative part of the integrand using $u\log(u) \geq -1/e$, the second inequality follows by Jensen's inequality (see Clarke and Barron 1990, p. 469), and the final inequality holds by our choice for the set U_n and by selecting ϵ_n appropriately small.

To complete the bound for (A.6), we have

$$P_0^n(A_nB_n^c(1-\xi_n)) \le \exp(r'K_n^2)P_0^n\left[\Pi_n(\mathcal{N}_n^c)(1-\xi_n)\frac{\int f_Q^n(\cdot)\mathcal{P}_N(dQ)}{f_0^n}\right]$$

$$= \exp(r'K_n^2)\int_{\mathcal{X}^n}\int_{\mathcal{N}_n^c}(1-\xi_n(\mathbf{X}))f_Q^n(\mathbf{X})$$

$$\times \mathcal{P}_N(dQ)d\lambda^n(\mathbf{X})$$

$$= \exp(r'K_n^2)\int_{\mathcal{N}_n^c}P_Q^n(1-\xi_n)\mathcal{P}_N(dQ),$$

which is bounded by $\exp(-(r-r')K_n^2)$ because of our UEC test. Thus, substituting the various bounds into (A.6), deduce that

$$P_0^n(A_n) \le \exp(-rK_n^2) + o(1) + \exp(-(r-r')K_n^2) = o(1)$$

by choosing $r > r' > 0$.

[Received February 2001. Revised April 2001.]

REFERENCES

Aitkin, M., and Rubin, D. B. (1985), "Estimation and Hypothesis Testing in Finite Mixture Models," *Journal of the Royal Statistical Society*, Ser. B, 47, 67–75.

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, eds. B.N. Petrov and F. Csaki, Budapest: Akademiai Kiado, pp. 267–281.

Barron, A. R. (1988), "The Exponential Convergence of Posterior Probabilities With Implications for Bayes Estimators of Density Functions," Technical Report 7, University of Illinois, Dept. of Statistics.

——— (1989), "Uniformly Powerful Goodness-of-Fit Tests," *The Annals of Statistics*, 17, 107–124.

Basford, K. E., McLachlan, G. J., and York, M. G. (1997), "Modelling the Distribution of Stamp Paper Thickness via Finite Normal Mixtures: The 1872 Hidalgo Stamp Issue of Mexico Revisited," *Journal of Applied Statistics*, 24, 169–179.

Bernardo, J. M., and Smith A. F. M. (1993), *Bayesian Theory*, New York: Wiley.

Brunner, L. J., Chan, A. T., James, L. F., and Lo, A. Y. (2001), "Weighted Chinese Restaurant Processes and Bayesian Mixture Models," *Annals of Statistics*, submitted.

Cao, R., and Devroye, L. (1996), "The Consistency of a Smoothed Minimum Distance Estimate," *Scandinavian Journal of Statistics*, 23, 405–418.

Chen, J. (1995), "Optimal Rate of Convergence for Finite Mixture Models," *The Annals of Statistics*, 23, 221–233.

Chen, J., and Kalbfleisch, J. D. (1996), "Penalized Minimum Distance Estimates in Finite Mixture Models," *Canadian Journal of Statistics*, 2, 167–176.

Chib, S. (1995), "Marginal Likelihood From the Gibbs Output," *Journal of American Statistical Association*, 90, 1313–1321.

Clarke, B. S., and Barron, A. R. (1990), "Information-Theoretic Asymptotics of Bayes Methods," *IEEE Transactions on Information Theory*, 36, 453–471.

Diebolt, J., and Robert, C. P. (1994), "Estimation of Finite Mixture Distributions Through Bayesian Sampling," *Journal of the Royal Statistical Society*, Ser. B, 56, 363–375.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956), "Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator," *The Annals of Mathematical Statistics*, 27, 642–669.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230.

——— (1974), "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 2, 615–629.

Genovese, C. R., and Wasserman, L. (2000), "Rates of Convergence for the Gaussian Mixture Sieve," *The Annals of Statistics*, 28, 1105–1127.

Ishwaran, H. (1998), "Exponential Posterior Consistency via Generalized Pólya Urn Schemes in Finite Semiparametric Mixtures," *The Annals of Statistics*, 26, 2157–2178.

Ishwaran, H., and James, L. F. (2000a), "Generalized Weighted Chinese Restaurant Processes for Species Sampling Mixture Models," *The Annals of Statistics*, submitted.

——— (2000b), "Approximate Dirichlet Process Computing in finite Normal Mixtures: Smoothing and Prior Information," *Journal of Computational and Graphical Statistics*, to appear.

——— (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of American Statistical Association*, 96, 161–173.

Ishwaran, H., James, L. F., and Lo, A. Y. (2001), "Generalized Weighted Chinese Restaurant and SIS Stick-Breaking Algorithms for Semiparametric Models," manuscript.

Ishwaran, H., and Zarepour, M. (2000a), "Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-Parameter Process Hierarchical Models," *Biometrika*, 87, 371–390.

——— (2000b), "Exact and Approximate Sum-Representations for the Dirichlet Process," *Canadian Journal of Statistics*.

——— (2000c), "Dirichlet Prior Sieves in Finite Normal Mixtures," *Statistica Sinica*, ———

Izenman, A. J., and Sommer, C. J. (1988), "Philatelic mixtures and multimodal densities," *Journal of American Statistical Association*, 83, 941–953.

Jewell, N. P. (1982), "Mixtures of Exponential Distributions," *The Annals of Statistics*, 10, 479–484.

Keribin, C. (2000), "Consistent Estimation of the Order of Mixture Models," *Sankhya*, Ser. A, 62, 49–66.

Lambert, D., and Tierney, L. (1984), "Asymptotic Properties of Maximum Likelihood Estimates in the Mixed Poisson Model," *The Annals of Statistics*, 12, 1388–1399.

Leroux, B. G. (1992), "Consistent Estimation of a Mixing Distribution," *The Annals of Statistics*, 20, 1350–1360.

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

Lo, A. Y. (1984), "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates," *The Annals of Statistics*, 12, 351–357.

Lo, A. Y., Brunner, L. J., and Chan, A. T. (1996), "Weighted Chinese Restaurant Processes and Bayesian Mixture Models," Research Report 1, Hong Kong University of Science and Technology.

Massart, P. (1990), "The Tight Constant in the Dvoretzky–Kiefer–Wolfowitz Inequality," *The Annals of Probability*, 18, 1269–1283.

McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

Minnotte, M. C. and Scott, D. W. (1993), "The Mode Tree: A Tool for Visualization of Nonparametric Density Features," *Journal of Computational and Graphical Statistics*, 2, 51–68.

Pfanzagl, J. (1988), "Consistency of Maximum likelihood Estimators for Certain Nonparametric Families, in Particular: Mixtures," *Journal of Statistical Planning and Inference*, 19, 137–158.

Pttman, J. (1995), "Exchangeable and Partially Exchangeable Random Partitions," *Probability Theory and Related Fields*, 102, 145–158.

——— (1996), "Some Developments of the Blackwell–MacQueen Urn Scheme," in *Statistics, Probability and Game Theory*, eds. T. S. Ferguson, L. S. Shapley, and J. B. MacQueen, IMS Lecture Notes—Monograph Series (vol. 30), Hayward, CA: Institute of Mathematical Statistics, pp. 245–267.

Raftery, A. E. (1996), "Hypothesis Testing and Model Selection," in *Markov chain Monte Carlo in Practice*, eds. W. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman and Hall, pp. 163–188.

Richardson, S., and Green, P. J. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components," *Journal of the Royal Statistical Society*, Ser. B, 59, 731–792.

Roeder, K. (1994), "A Graphical Technique for Determining the Number of Components in a Mixture of Normals," *Journal of American Statistical Association*, 89, 487–495.

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

Roeder, K., and Wasserman, L. (1997), "Practical Bayesian Density Estimation Using Mixtures of Normals," *Journal of American Statistical Association*, 92, 894–902.

Schwartz G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Schwartz, L. (1965), "On Bayes Procedures," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 4, 10–26.

Silverman, B. W. (1981), "Using Kernel Density Estimates to Investigate Multimodality," *Journal of the Royal Statistical Society*, Ser. B, 43, 97–99.

Simar, L. (1976), "Maximum Likelihood Estimation of a Compound Poisson Process," *The Annals of Statistics*, 4, 1200–1209.

Teicher, H. (1963), "Identifiability of Finite Mixtures," *The Annals of Mathematical Statistics*, 32, 1265–1269.

Teicher, H. (1960), "On the Mixture of Distributions," *The Annals of Mathematical Statistics*, 31, 55–73.

van de Geer, S. (1996), "Rates of convergence for the Maximum Likelihood Estimator in Mixture Models," *Nonparametric Statistics*, 6, 293–310.

Verdinelli, I., and Wasserman, L. (1998), "Bayesian Goodness-of-Fit Testing Using Infinite-Dimensional Exponential Families," *The Annals of Statistics*, 26, 1215–1241.

Watterson, G. A. (1976), "The Stationary Distribution of the Infinitely-Many Neutral Alleles Diffusion Model," *Journal of Applied Probability*, 13, 639–651.

Wilson, I. G. (1983), "Add a New Dimension to Your Philately," *The American Philatelist*, 97, 342–349.

Zhang, C. H. (1990), "Fourier Methods for Estimating Mixing Densities and Distributions," *The Annals of Statistics*, 18, 806–831.