Statistica Sinica **12**(2002), 941-963

DIRICHLET PRIOR SIEVES IN FINITE NORMAL MIXTURES

Hemant Ishwaran and Mahmoud Zarepour

Cleveland Clinic Foundation and University of Ottawa

Abstract: The use of a finite dimensional Dirichlet prior in the finite normal mixture model has the effect of acting like a Bayesian method of sieves. Posterior consistency is directly related to the dimension of the sieve and the choice of the Dirichlet parameters in the prior. We find that naive use of the popular uniform Dirichlet prior leads to an inconsistent posterior. However, a simple adjustment to the parameters in the prior induces a random probability measure that approximates the Dirichlet process and yields a posterior that is strongly consistent for the density and weakly consistent for the unknown mixing distribution. The dimension of the resulting sieve can be selected easily in practice and a simple and efficient Gibbs sampler can be used to sample the posterior of the mixing distribution.

Key words and phrases: Bose-Einstein distribution, Dirichlet process, identification, method of sieves, random probability measure, relative entropy, weak convergence.

1. Introduction

The finite normal mixture model is a widely applicable model that has received considerable attention in the Bayesian statistical literature. See Escobar (1988, 1994), Diebolt and Robert (1994), Escobar and West (1995), Richardson and Green (1997) and Roeder and Wasserman (1997) for some recent examples. The wide scope of applications has led to many different methods for fitting this model, including Monte Carlo simulation (Lo (1984), Kuo (1986), Ferguson (1983)), and Markov chain Monte Carlo Gibbs sampling (Escobar (1988, 1994), MacEachern (1994), Escobar and West (1995)). The selection of priors used in the normal mixture model is equally varied, with probably the two most popular choices being the Ferguson (1973, 1974) Dirichlet process prior and finite dimensional priors based on Dirichlet random weights (Diebolt and Robert (1994), Chib (1995), Richardson and Green (1997), Roeder and Wasserman (1997), Ishwaran and Zarepour (2000), Neal (2000), Green and Richardson (2001)).

Because of their simplicity, and computational tractability, priors based on Dirichlet random weights have been gaining in popularity and use (see Ishwaran and Zarepour (2002) for more discussion). The focus of this paper will be the study of these priors, which we refer to generally as *Dirichlet priors* or sometimes as *finite dimensional Dirichlet priors* (named so as not to be confused with the

Ferguson Dirichlet process). We show that the use of such priors in the finite normal mixture problem has the effect of acting like a Bayesian finite dimensional sieve procedure with the dimension of the prior N controlling the dimension of the sieve in terms of the sample size n. With a proper selection of Dirichlet parameters the method results in a prior that approximates the Dirichlet process and a sieve procedure that can be used to consistently estimate the normal mixture density (see Theorem 6). Moreover, the method also ensures that the posterior is consistent for the finite mixing distribution (Theorem 7).

The consistency for the mixing distribution is a result that appears to be new to Gaussian sieve procedures, where the use of finite normal mixtures has traditionally been employed as a method to estimate an unknown density (not necessarily a mixture of normals). For example, Roeder and Wasserman (1997) use Gaussian sieves in a Bayesian approach for consistent density estimation while, from a non-Bayesian context, Gaussian sieves have been explored as a method for density estimation by Geman and Hwang (1982) and Priebe (1994). See also Grenander (1981), Shen and Wong (1994) and Wong and Shen (1995) for a general discussion on the use of the method of sieves.

Gaussian sieves have also been used as a method for estimating densities that are explicitly assumed to be a mixture of normals. This was the approach used in Roeder (1992) and Genovese and Wasserman (2000) who consider mixtures of normals under mixing distributions restricted to certain classes (for example mixing distributions with compact support). Density estimation for mixtures of normals has also been considered using non-sieve based approaches. For example, Zhang (1990) uses Fourier techniques for estimating normal mixing densities and mixing distributions. Closer to our work are the papers by Ghosal, Ghosh and Ramamoorthi (1999) and Ghosal and van der Vaart (2001), who each consider density estimation for mixtures of normals using a Bayesian approach with a Dirichlet process prior. Such Dirichlet process approaches are relevant to our method as the underlying prior used here (for an appropriate selection of Dirichlet parameters) will be seen to be a weak limit approximation to the Dirichlet process, and thus, in some sense, our sieve procedure can be seen to be a finite dimensional analogue of such methods. Thus, it should be not too surprising that our method is consistent, akin to the results found in those papers. See Remark 2 of Section 5 for a more thorough comparison of results.

It is important to note that the sieve approach developed here, and the surrounding theory, is more than a study of finite dimensional approximations to existing Dirichlet process methods. For example, an important feature of our approach is that it enables simple and efficient computational procedures for estimating posterior quantities. Such procedures rely on the finite dimensionality of our sieves, specifically exploiting the ability to recast the model in terms of a

finite number of random variables. Such a procedure (the blocked Gibbs sampler) is given in detail in Section 6. Moreover, the theory developed for limits of Dirichlet priors (Theorem 3 of Section 3) should be important in guiding the selection of Dirichlet parameters in the many computational methods based on Dirichlet priors. As we will see, real care needs to be exercised when choosing the Dirichlet parameters, since not all priors will work well. In particular, we will see that naive use of the popular uniform Dirichlet prior leads to an inconsistent posterior. Another important contribution of our research, worth re-emphasizing, is our ability to consistently estimate the mixture distribution. Although Gaussian sieve approaches traditionally focus on density estimation, in the analysis of finite normal mixtures it is often the mixing distribution that is the primary focus of inference, and thus it is important to be able to recover this value consistently. See Lindsay (1995) and McLachlan and Peel (2000, Chapter 1) for more motivation and discussion of this point.

1.1. Hierarchical description of the model

The finite normal mixture arises from data $\mathbf{X} = (X_1, \ldots, X_n)$, where the X_i are conditionally independent normal random variables, with a random mean and a random variance sampled from a finite mixture distribution Q_0. More precisely, the X_i are i.i.d. from the distribution P_0 with the finite normal mixture density

$$f_0(x) = \int \phi(x|\boldsymbol{\mu}(y), \tau(y))\, dQ_0(y) = \sum_{k=1}^{d} p_{k,0}\, \phi(x|\boldsymbol{\mu}_{k,0}, \tau_{k,0}), \tag{1}$$

where $\phi(\cdot|\boldsymbol{\mu}, \tau)$ represents a normal density with a mean of μ and a variance of $\tau > 0$, and where we write $Y = (\boldsymbol{\mu}(Y), \tau(Y))$ for the two-dimensional mean and variance, where $\boldsymbol{\mu}(\cdot)$ extracts the first coordinate of Y (the mean) and $\tau(\cdot)$ extracts the second coordinate (the variance).

Inference for the finite normal mixture is complicated due to the assumption that the underlying mixture distribution Q_0 is completely unspecified except for the assumption that it is a finite distribution, expressible as

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

where $\delta_{Z_{k,0}}(\cdot)$ denotes a discrete measure concentrated at $Z_{k,0} = (\boldsymbol{\mu}_{k,0}, \tau_{k,0})$, and $\tau_{k,0} > 0$ are positive variances, $\boldsymbol{\mu}_{k,0}$ are mean values, while $p_{k,0} > 0$ are positive fixed weights satisfying $\sum_{k=1}^{d} p_{k,0} = 1$. Except for the discrete distributional assumption, Q_0 is left unspecified, and thus in particular, not only is the number of support points $d < \infty$ assumed to be unknown, but so are the atoms of

the distribution $\{(\mu_{k,0}, \tau_{k,0}) : k = 1, \ldots, d\}$ and the weights $p_{k,0}$. It is worth emphasizing here that the assumption that $d < \infty$ is unknown is very much different than the case that $d < \infty$ is unknown but bounded by some fixed finite number $d_0 < \infty$. The latter case has been studied by Chen (1995) from a frequentist perspective for finite mixture models, where it has been shown that the mixing distribution can be estimated at an optimal $O_p(n^{-1/4})$ rate. Recently, in studying the same problem, Ishwaran, James and Sun (2001) devise a Bayesian approach based on the use of a Dirichlet prior and show among other things that the posterior is $\sqrt{n}$-consistent for the density and moreover achieves the $n^{-1/4}$ optimal rate for Q_0. However, such results rest heavily on the assumption that $d < d_0$ and thus apply to mixture problems where such bounds can be naturally deduced from the context of the data. This is very much different than our problem where no such bound d_0 is assumed known.

In hierarchical format, the model derived from (1) can also be expressed as

$$\begin{aligned} (X_i|Y_i) &\overset{\text{ind}}{\sim} \mathrm{N}(\boldsymbol{\mu}(Y_i), \tau(Y_i)), \qquad i = 1, \ldots, n \\ (Y_i|Q_0) &\overset{\text{iid}}{\sim} Q_0(\cdot), \end{aligned} \tag{2}$$

where the Y_i are hidden variables sampled from the unknown mixing distribution Q_0. As mentioned earlier, a Bayesian method for studying this model that is growing in popularity makes use of a finite dimensional Dirichlet prior. In this approach, one introduces latent variables $\mathbf{K} = (K_1, \ldots, K_n)$ which indicate the group membership for the hidden variables. Specifically, (2) is recast as

$$\begin{aligned} (X_i|K_i, \mu, \boldsymbol{\tau}) &\overset{\text{ind}}{\sim} \mathrm{N}(\mu_{K_i}, \tau_{K_i}), \qquad i = 1, \ldots, n, \\ (K_i|\mathbf{p}) &\overset{\text{iid}}{\sim} \text{Multinomial}(\{1, \ldots, N\}, \mathbf{p}) \\ (\mu_k, \tau_k) &\overset{\text{iid}}{\sim} H, \qquad k = 1, \ldots, N, \end{aligned} \tag{3}$$

where N is an integer (the dimension) converging to ∞ as a function of n, and $\mu = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N)$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_N)$, $\mathbf{K}$ and $\mathbf{p} = (p_1, \ldots, p_N)$ are Bayesian parameters that are to be estimated from the posterior, and where

$$\mathbf{p} \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_N) \tag{4}$$

has a finite dimensional Dirichlet prior. Observe that $K_i \in \{1, \ldots, N\}$ are multinomial values such that $\mathbb{P}\{K_i = k|\mathbf{p}\} = p_k$ for $k = 1, \ldots, N$. Each of these K_i values record which mean and variance $(\boldsymbol{\mu}_{K_i}, \tau_{K_i}) \in \{(\boldsymbol{\mu}_k, \tau_k) : k = 1, \ldots, N\}$ are associated with each X_i, and thus provide a clever method for modeling (2) (see McLachlan and Peel (2000, Chapter 4) for more discussion and related references for this latent variable technique).

1.2. Outline of paper

In practice, the prior H for $(\boldsymbol{\mu}_k, \tau_k)$ used in (3) is usually chosen to take advantage of conjugacy in order to simplify computations, with the eventual selection playing a limited role in the behavior of the posterior for a large sample size n. However, the choice for the Dirichlet parameters used in (4) needs to be selected with great care. As Theorem 3 of Section 3 will show, the prior induced by (4) is a random probability measure whose limit is either (a) the parametric prior H, (b) the Ferguson (1973, 1974) Dirichlet process, or (c) a simple zero-one process. In particular, the widely used *uniform Dirichlet prior* with random weights

$$\mathbf{p} \sim \text{Dirichlet}(1, \ldots, 1), \tag{5}$$

is an example of a random measure (a) whose limit is the parametric prior H.

With a prior that acts parametrically in the limit, it may not seem too surprising that the resulting posterior will be inconsistent. In fact, there appears to be a very delicate line for how large the dimension N of the prior (and hence the sieve) can be relative to n before consistency breaks down. In Section 4, Theorem 5, we show that the uniform Dirichlet prior is inconsistent if $n/N \to 0$ (also see Theorem 4). Thus, if N is larger than n, we end up with an inconsistent posterior. However, it seems natural to expect the posterior to eventually become consistent if the value of N is made relatively small compared to n. Indeed Ghosal and van der Vaart (2001, Section 7), in studying mixtures of normal densities with the Dirichlet process prior, observed that the resulting posterior was consistent even if the Dirichlet mass parameter was allowed to vary with the sample size; a sufficient condition being that it remained no larger than $O(\log n)$. We suspect that this same phenomenon occurs here, and that the uniform Dirichlet prior is consistent for values of N that are exponentially smaller than n. However, at some point N will become too large relative to n, signifying the critical lower bound at which consistency breaks down.

Although we have not been able to determine the exact lower bound, we suspect that its value is unlikely to be easy to use in practical application. Moreover, we argue that *work along this line is unnecessary* since a very simple method exists to correct this problem. Section 5 will show that a consistent posterior can be easily obtained by changing the Dirichlet parameters used in the prior for $\mathbf{p}$. By using the prior

$$\mathbf{p} \sim \text{Dirichlet}(\alpha/N, \ldots, \alpha/N), \tag{6}$$

we end up with a random probability measure whose limit is the Ferguson Dirichlet process (see Theorem 3) and a posterior that is $\mathcal{L}_1$ consistent for the density if $\log N/n \to 0$ (Theorem 6) and, under some additional conditions, which is weakly consistent for the mixing distribution (Theorem 7). Hence, consistency

holds over a broad range of values for N, thus resolving the problems seen with (5). See Section 5 for more discussion as well as comparisons to other methods.

In Section 6 we demonstrate that the posterior mixing distribution corresponding to the prior (6) can be sampled using a simple Gibbs sampling algorithm, the blocked Gibbs sampler (Ishwaran and Zarepour (2000) and Ishwaran and James (2001)). The paper begins in Section 2 by describing the relationship between sieves and the use of Dirichlet priors.

2. Dirichlet Priors and Sieves

Unfortunately, the convenience in modeling the finite mixture problem in terms of random variables as in (3) has had the effect of encouraging the notion that the model can be expressed as a parametric problem. The tendency in the literature has been to conceptualize the parameter space for the normal mixture model as the parameter space corresponding to the finite dimensional parameters $(\mu, \boldsymbol{\tau}, \mathbf{K}, \mathbf{p})$, and this seems to have had the adverse effect of hiding the fact that the number of support points $d < \infty$ is unknown, and that the model is nonparametric.

A better way to conceptualize the model is to recast it in terms of a random probability measure. In particular, by recognizing that $Y_i = Z_{K_i}$, where the classification variables K_i can be expressed as

$$(K_i|\mathbf{p}) \overset{\text{iid}}{\sim} \sum_{k=1}^{N} p_k \, \delta_k(\cdot), \tag{7}$$

it follows that (3) can be rewritten as

$$\begin{aligned} (X_i|Y_i) &\overset{\text{ind}}{\sim} \mathrm{N}(\boldsymbol{\mu}(Y_i), \tau(Y_i)), \qquad i = 1, \ldots, n \\ (Y_i|Q) &\overset{\text{iid}}{\sim} Q \\ Q &\sim \mathcal{P}_N, \end{aligned} \tag{8}$$

where $\mathcal{P}_N(\cdot) = \sum_{k=1}^{N} p_k \, \delta_{Z_k}(\cdot)$ is a random probability measure and $Z_k = (\boldsymbol{\mu}_k, \tau_k)$ are i.i.d. with distribution H on $\Re \times \Re^+$ assumed to be independent of $\mathbf{p}$. We assume throughout unless otherwise stated that H is a nonatomic distribution.

Conceptually, the hierarchical model (8) is more easily identified with the original mixture model (2), as it clearly identifies the random probability measure $\mathcal{P}_N$ as a mechanism for modeling the mixture distribution Q_0. Moreover, by rewriting the model in terms of $\mathcal{P}_N$, it reminds us that the parameter space for the normal mixture model is the space of finite distributions on $\Re \times \Re^+$, denoted by $\mathcal{Q}_F = \bigcup_{k=1}^{\infty} \mathcal{Q}_k$, where $\mathcal{Q}_k$ is the space of distributions on $\Re \times \Re^+$ with exactly k atoms.

Note that the random probability measure $\mathcal{P}_N$ is a prior over $\mathcal{Q}_N$, which we now can identify as the parameter space for (8). Consequently, if N is allowed to grow with the sample size n, so does our prior space $\mathcal{Q}_N$. Thus, we can think of $\{\mathcal{Q}_N\}$ as the sieve over which our Bayesian nonparametric approach operates.

2.1. Finite normal mixtures are identified

We note there is another important reason for interpreting our parameter space nonparametrically as a space of discrete distributions $\mathcal{Q}_N$, rather than thinking of it as a parametric space corresponding to $(\mu, \boldsymbol{\tau}, \mathbf{K}, \mathbf{p})$. The latter approach leads to an unidentified model because of the indeterminacy in the coordinates of the parameters. This *unnecessarily complicates matters* since it forces the use of constraints to ensure model identification; such as the common practice of requiring the means to be ordered: $\boldsymbol{\mu}_1 \leq \cdots \leq \boldsymbol{\mu}_N$. From a computational perspective this is also problematic since a constrained parameter space is more difficult to work with. For example, Gibbs sampling based approaches suffer convergence problems in this setting (Celeux, Hurn and Robert (2000)).

However, the approach using (8) produces *a fully identified model without any additional constraints* and resolves the computational problems of working with a constrained space (see Section 6). This identification result is due to Teicher (1963, Proposition 1) which we have slightly generalized in the following theorem:

Theorem 1. [Teicher (1963)] *Suppose that*

$$\int \phi(x|\mu(y), \tau(y))\, dQ_0(y) = \int \phi(x|\mu(y), \tau(y))\, dQ^*(y), \qquad \text{for almost all } x, \quad (9)$$

where $Q^*(\cdot) = \sum_{k=1}^{d^*} p_k^*\, \delta_{(\boldsymbol{\mu}_k^*, \tau_k^*)}(\cdot)$ *is an element of* $\mathcal{Q}_F \bigcup \mathcal{Q}_\infty$, *the space of discrete distributions on* $\Re \times \Re^+$. *That is, suppose that* $\sum_{k=1}^{d} p_{k,0}\, \phi(x|\mu_{k,0}, \tau_{k,0}) = \sum_{k=1}^{d^*} p_k^*\, \phi(x|\mu_k^*, \tau_k^*)$ *for almost all* x, *where* $d^* \leq \infty$. *Then the identity expressed by* (9) *implies that* $Q_0 = Q^*$.

Proof of Theorem 1. Teicher (1963) established the result for $Q^* \in \mathcal{Q}_F$, i.e., for the case when Q^* is a finite distribution with $d^* < \infty$. However, a close inspection of the proof shows that it can be extended to the case when $d^* = \infty$.

In general, the normal mixture model can be made identified by various constraints to the mixing distribution. The following theorem presents an exponential moment condition sufficient for identification of Q_0. It will play an important role in Section 5 when we establish posterior consistency for Q_0. See the Appendix for its proof.

Theorem 2. *Write* $\mathcal{Q}^*$ *for the set of distributions* Q *over* $\Re \times \Re^+$ *satisfying*

$$\int_{\Re \times \Re^+} \exp\left(\frac{\mu^2}{2(\tau^* - \tau)}\right) Q(d\mu, d\tau) < \infty,$$

where $\tau^* = \min\{\tau_{1,0}, \ldots, \tau_{d,0}\}$. *If* $Q \in \mathcal{Q}^*$ *is such that*

$$\sum_{k=1}^{d} p_{k,0}\, \phi(x|\mu_{k,0}, \tau_{k,0}) = \int \phi(x|\mu, \tau)\, Q(d\mu, d\tau) \qquad \textit{for almost all } x, \tag{10}$$

then $Q = Q_0$.

3. Limits for Dirichlet Priors

Good posterior behavior in the mixture model is directly related to our choice of the prior $\mathcal{P}_N$, which is directly related to the choice of parameters in the prior for $\mathbf{p}$. As Theorem 3 below will show, a careful selection for the Dirichlet parameters in (4) is crucial in ensuring a prior rich enough to properly model Q_0.

In what transpires, we write

$$\mathcal{P}_N(\cdot) = \sum_{k=1}^{N} p_{k,N}\, \delta_{Z_k}(\cdot) \tag{11}$$

where Z_k are i.i.d. variables independent of $(p_{1,N}, \ldots, p_{N,N}) \sim \text{Dirichlet}(\alpha_{1,N}, \ldots, \alpha_{N,N})$, where $\alpha_{k,N} > 0$. In order to facilitate proofs of the limits of $\mathcal{P}_N$, and for the purposes of an explicit construction needed for almost sure convergence (see 1(a) of Theorem 3 and also Remark 1 below), we use the well known representation of the Dirichlet distribution in terms of gamma random variables to constructively define our random weights in (11). In particular, we assume that

$$(p_{1,N}, \ldots, p_{N,N}) = \left(\frac{G_{1,N}}{\sum_{k=1}^{N} G_{k,N}}, \ldots, \frac{G_{N,N}}{\sum_{k=1}^{N} G_{k,N}} \right) \tag{12}$$

where $G_{k,N}$ are independent Gamma$(\alpha_{k,N})$ random variables.

In the following theorem we use the mode of convergence indicated by "$\Rightarrow$" to represent convergence of a random probability measure with respect to the weak topology. In particular, if g is a non-negative continuous function with compact support, we write $\mathcal{P}_N \overset{\text{a.s}}{\Rightarrow} \mathcal{P}$ if $\mathcal{P}_N(g) \overset{\text{a.s}}{\to} \mathcal{P}(g)$ for each such g. Furthermore, we write, $\mathcal{P}_N \overset{\text{d}}{\Rightarrow} \mathcal{P}$ if $\mathcal{P}_N(g) \overset{\text{d}}{\to} \mathcal{P}(g)$ for each such g, while $\mathcal{P}_N \overset{\text{p}}{\Rightarrow} \mathcal{P}$ if $\mathcal{P}_N(g) \overset{\text{p}}{\to} \mathcal{P}(g)$ for each such g. See Resnick (1987, Chapter 3.5) for related discussion on vague convergence. See also Billingsley (1968, Chapter 4) for a general discussion of convergence over abstract spaces.

Theorem 3. *Suppose that* $\mathcal{P}_N$ *is the random probability measure defined by* (11) *and* (12)*, where* Z_k *are i.i.d.* H (*here* H *is not necessarily nonatomic*).

1(a). *If* $\alpha_{k,N} = \lambda_k$, *where* $\sum_{k=1}^{\infty} \lambda_k^2/k^2 < \infty$ *and* $\sum_{k=1}^{N} \lambda_k/N \to \lambda_0 > 0$, *then* $\mathcal{P}_N \overset{\text{a.s}}{\Rightarrow} H$.

1(b). *If* $\alpha_{k,N} = \lambda_N$, *where* $N\lambda_N \to \infty$, *then* $\mathcal{P}_N \overset{\text{p}}{\Rightarrow} H$.

2(a). *If* $\alpha_{k,N} = \alpha/N$, *for some* $\alpha > 0$, *then for each real-valued measurable function* g *which is integrable with respect to* H, *we have* $\mathcal{P}_N(g) \overset{\text{d}}{\to} \mathcal{P}_\infty(g)$, *where* $\mathcal{P}_\infty = DP(\alpha H)$ *is the Ferguson* (1973, 1974) *Dirichlet process with finite measure* αH.

2(b). *If* $\sum_{k=1}^{N} \alpha_{k,N} \to \alpha > 0$ *and* $\max(\alpha_{1,N}, \ldots, \alpha_{N,N}) \to 0$, *then* $\mathcal{P}_N \overset{\text{d}}{\Rightarrow} DP(\alpha H)$.

3. *If* $\alpha_{k,N} = \lambda_N$, *where* $N\lambda_N \to 0$, *then* $\mathcal{P}_N \overset{\text{d}}{\Rightarrow} \delta_Z$ *where* Z *has distribution* H.

The proof of the theorem is given in the Appendix. Note that the uniform Dirichlet prior (5) corresponds to case 1(a) by setting $\lambda_k = 1$, and our gamma construction at (12) shows that

$$\mathcal{P}_N(\cdot) = \sum_{k=1}^{N} \frac{E_k}{\sum_{k=1}^{N} E_k} \delta_{Z_k}(\cdot) \overset{\text{a.s}}{\Rightarrow} H(\cdot), \tag{13}$$

where E_k are i.i.d. exp(1) random variables. Note also that, from 2(a) and 2(b), if we select $\alpha_{k,N} = \alpha/N$ we avoid a parametric limit and instead obtain a prior $\mathcal{P}_N$ rich enough that it can approximate the Dirichlet process. In particular, for the choice of Dirichlet parameters in (6), we have from 2(b) that

$$\mathcal{P}_N(\cdot) = \sum_{k=1}^{N} \frac{G_{k,N}}{\sum_{k=1}^{N} G_{k,N}} \delta_{Z_k}(\cdot) \overset{\text{d}}{\Rightarrow} \mathrm{DP}(\alpha H)(\cdot), \tag{14}$$

where $G_{k,N}$ are independent Gamma(α/N) random variables. Notice that the convergence result established in 2(a) is stronger than the result in 2(b), since it tells us that the measure on the left-hand side of (14) can approximate integrable functionals of the Dirichlet process. For related discussion, see Kingman (1975), Muliere and Secchi (1995), Pitman (1996), Ishwaran and Zarepour (2000, 2002), Neal (2000), Ishwaran and James (2001) and Green and Richardson (2001), who have discussed the use of $\mathcal{P}_N$ in different contexts.

Remark 1. In 1(a) the Dirichlet weights are defined by $p_{k,N} = p_k = G_k/(\sum_{k=1}^{N} G_k)$ with G_k independent Gamma(λ_k) random variables. Thus, the sequence of measures $\mathcal{P}_N$ can be defined to live on the same space, which is needed for proper interpretation of our almost sure convergence result.

4. Inconsistency with Uniform Dirichlet Priors

At a superficial level, the selection of a uniform Dirichlet prior (5) for $\mathbf{p}$ is appealing because it represents a flat prior. Unfortunately, as Theorem 3 in the

previous section shows, the random probability measure $\mathcal{P}_N$ associated with this prior (see (13)) converges almost surely to H as $N \to \infty$, and thus the limit for the prior $\mathcal{P}_N$ in the model (8) is the parametric prior H for the mean and variance.

The same behavior also means that the limit of the marginal density for $\mathbf{X}$ from (8),

$$m_N(\mathbf{X}) = \int \left(\prod_{i=1}^{n} \int \phi(X_i|\mu(Y_i), \tau(Y_i))\, dQ(Y_i) \right) \mathcal{P}_N(dQ), \tag{15}$$

is the marginal density based on the prior H

$$m_\infty^*(\mathbf{X}) = \prod_{i=1}^{n} \int \phi(X_i|\mu, \tau)\, dH(\mu, \tau). \tag{16}$$

In particular, the limit of (8), in distribution, is the parametric hierarchical model

$$\begin{aligned} (X_i|\mu_i, \tau_i) &\overset{\text{ind}}{\sim} \mathrm{N}(\mu_i, \tau_i), \qquad i = 1, \ldots, n \\ (\mu_i, \tau_i) &\overset{\text{iid}}{\sim} H. \end{aligned} \tag{17}$$

See Theorem 4, Section 4.2, for a more precise statement. From this, it is not hard to conjecture that a uniform Dirichlet prior produces an inconsistent posterior, as we prove in Theorem 5 of Section 4.3.

4.1. Bose-Einstein distribution

The distribution for the number of distinct Y_i values in (8), under a uniform Dirichlet prior for $\mathbf{p}$, can be described explicitly by exploiting a connection between the distribution of classification variables K_i (as in (7)) and the Bose-Einstein distribution. With a uniform Dirichlet prior, the clustering behavior of K_i is equivalent to the clustering behavior observed when n indistinguishable balls are placed randomly into N distinct urns. This characterization will show that the prior encourages too many distinct Y_i values which will enable us to describe the inconsistent behavior of the posterior. The proof for the following lemma is given in the Appendix.

Lemma 1. *Let D_n be the number of distinct values in the sample $K_1, \ldots, K_n$, where $(K_i|\mathbf{p})$ are i.i.d. from $\sum_{k=1}^{N} p_k\, \delta_k(\cdot)$ for $\mathbf{p} \sim$ Dirichlet$(1, \ldots, 1)$. Then, $\mathbb{P}\{D_n = k\} = \binom{N}{k}\binom{n-1}{k-1}\binom{N+n-1}{N-1}^{-1}$, $k = 1, \ldots, \min(n, N)$.*

4.2. Relative entropy bounds

An immediate application of Lemma 1 identifies the limiting distribution of the marginal density under a uniform prior. Let $D(\mathbb{Q}_1 \| \mathbb{Q}_2) = \int \log(d\mathbb{Q}_1/d\mathbb{Q}_2)\, d\mathbb{Q}_1$

denote the relative entropy (Kullback-Leibler information) between two probability measures $\mathbb{Q}_1$ and $\mathbb{Q}_2$.

Theorem 4. *Suppose that $\mathcal{P}_N$ is the random probability measure in* (13). *If $N \geq n$,*

$$D(M_\infty^* \| M_N) \leq -\log\left(\frac{N!(N-1)!}{(N-n)!(N+n-1)!}\right), \tag{18}$$

where M_N and M_∞^* are the laws for m_N and m_∞^* defined by (15) *and* (16), *respectively. In particular, if $(n-1)^2/N \to 0$,*

$$\int |m_N(\mathbf{X}) - m_\infty^*(\mathbf{X})|\, d\mathbf{X} \to 0.$$

Under a uniform prior the limit in total variation distance of our nonparametric model (8), as $N \to \infty$, is the parametric model (17). This limiting parametric behavior persists as the sample size n increases, as long as $(n-1)^2/N \to 0$.

Proof of Theorem 4. Let $\mathbf{P} = \{C_j : j = 1, \ldots, N(\mathbf{P})\}$ 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. Let π_U denote the uniform Dirichlet distribution (5), and write H^N for the product distribution of $\mathbf{Z} = (Z_1, \ldots, Z_N)$. Integrating over the random measure P in (15), and keeping $\mathbf{Z}$ fixed until the end, we have

$$\begin{aligned}
m_N(\mathbf{X}) &= \int\int\int \prod_{i=1}^n \phi(X_i|\mu(Y_i), \tau(Y_i)) \prod_{i=1}^n \left(\sum_{k=1}^N p_k\, \delta_{Z_k}(dY_i)\right) d\pi_U(\mathbf{p})\, dH^N(\mathbf{Z}) \\
&= \int \left(\sum_{\mathbf{P}} \sum_{\{l_1 \neq \cdots \neq l_m\}} E(p_{l_1}^{e_1} \cdots p_{l_m}^{e_m}) \prod_{j=1}^{m=N(\mathbf{P})} \prod_{i \in C_j} \phi(X_i|\mu(Z_{l_j}), \tau(Z_{l_j}))\right) dH^N(\mathbf{Z}) \\
&= \sum_{\mathbf{P}} f(\mathbf{P}) \prod_{j=1}^{N(\mathbf{P})} \int \prod_{i \in C_j} \phi(X_i|\mu, \tau)\, dH(\mu, \tau),
\end{aligned}$$

where $f(\mathbf{P})$ is the probability that the classification variables $K_1, \ldots, K_n$ in Lemma 1 are made up of $N(\mathbf{P})$ unique values, with K_i all having the same values for $i \in C_j$.

Therefore,

$$D(M_\infty^* \| M_N) = \int \log\left(\frac{\prod_{i=1}^n \int \phi(X_i|\mu, \tau)\, dH(\mu, \tau)}{\sum_{\mathbf{P}} f(\mathbf{P}) \prod_{j=1}^{N(\mathbf{P})} \int \prod_{i \in C_j} \phi(X_i|\mu, \tau)\, dH(\mu, \tau)}\right) dM_\infty^*(\mathbf{X}).$$

Increase the value of the relative entropy by restricting the sum in the denominator to the partition $\mathbf{P}_n$ where $N(\mathbf{P}_n) = n$. This gives the upper bound

$$\int \log\left(\frac{\prod_{i=1}^n \int \phi(X_i|\mu,\tau)\,dH(\mu,\tau)}{\prod_{j=1}^{N(\mathbf{P}_n)} \int \prod_{i\in C_j} \phi(X_i|\mu,\tau)\,dH(\mu,\tau)}\right) dM_\infty^*(\mathbf{X}) - \log\left(f(\mathbf{P}_n)\right).$$

The first term is zero because $N(\mathbf{P}_n) = n$, while the second term is minus the log of the probability that each K_i value is distinct. Thus, by applying Lemma 1 to the second term (with $k = n$) we arrive at the inequality (18).

We have

$$-\log\left(\frac{N!(N-1)!}{(N-n)!(N+n-1)!}\right) = \sum_{j=N-n+1}^{N} \log\left(1+\frac{n-1}{j}\right) \le (n-1)\sum_{j=N-n+1}^{N}\frac{1}{j},$$

which is order $(n-1)^2/N$. If this is order $o(1)$, then using Kemperman's inequality to bound the total variation distance squared by the relative entropy (Kemperman (1969, Theorem 6.1)), deduce that M_N converges to M_∞^* in $\mathcal{L}_1$ distance.

4.3. Limiting posterior behavior

The posterior behavior of the random measure $\mathcal{P}_N$ can be studied by looking at the limit of its functionals. The following theorem shows that the posterior under a uniform prior is inconsistent for the unknown mixing distribution Q_0 if $n/N \to 0$.

Theorem 5. *Suppose that $\mathcal{P}_N$ is the random probability measure in* (13)*. If $n/N \to 0$, $\int Q(A)\,\mathcal{P}_N(dQ|\mathbf{X}) \to H(A)$ almost surely P_0^∞ for each Borel measurable set $A \in \Re \times \Re^+$, where $\mathcal{P}_N(\cdot|\mathbf{X})$ is the posterior of* (8)*.*

Proof. The measure $\mathcal{P}_N$ is a Dirichlet process for a fixed value of $\mathbf{Z}$. In particular, $\mathcal{L}(\mathcal{P}_N|\mathbf{Z}) = \mathrm{DP}(NH_N(\mathbf{Z},\cdot))$, where $H_N(\mathbf{Z},\cdot) = \sum_{k=1}^N \delta_{Z_k}(\cdot)/N$ is the empirical measure based on $\mathbf{Z}$. Thus, by conditioning on $\mathbf{Z}$, we can use Theorem 1 from Lo (1984). Therefore, with probability one,

$$\int Q(A)\,\mathcal{P}_N(dQ|\mathbf{X}) = \int\left(\int Q(A)\,\mathcal{P}_{NH_N+\sum_{i=1}^n \delta_{Y_i}}(dQ|\mathbf{Y},\mathbf{Z})\right) d\pi(\mathbf{Y},\mathbf{Z}|\mathbf{X}), \quad (19)$$

where $\mathbf{Y} = (Y_1,\ldots,Y_n)$, $\mathcal{L}(\mathcal{P}_{NH_N+\sum_{i=1}^n \delta_{Y_i}}|\mathbf{Y},\mathbf{Z}) = \mathrm{DP}(NH_N + \sum_{i=1}^n \delta_{Y_i})$, and

$$d\pi(\mathbf{Y},\mathbf{Z}|\mathbf{X}) = \frac{\prod_{i=1}^n \phi(X_i|\mu(Y_i),\tau(Y_i)) \int\Big(\prod_{i=1}^n Q(dY_i)\,\mathcal{P}_{NH_N}(dQ)\Big) dH^N(\mathbf{Z})}{\sum_{\mathbf{P}} f(\mathbf{P}) \prod_{j=1}^{N(\mathbf{P})} \int \prod_{i\in C_j} \phi(X_i|\mu,\tau)\,dH(\mu,\tau)}. \quad (20)$$

The inner-integral on the right-hand side of (19) is

$$\int Q(A)\,\mathcal{P}_{NH_N+\sum_{i=1}^n \delta_{Y_i}}(dQ|\mathbf{Y},\mathbf{Z}) = \frac{N}{N+n} H_N(\mathbf{Z},A) + \frac{1}{N+n}\sum_{i=1}^n \{Y_i \in A\}.$$

Thus,

$$\int Q(A)\,\mathcal{P}_N(dQ|\mathbf{X}) = \frac{N}{N+n}E\Big(\frac{1}{N}\sum_{k=1}^{N} I\{Z_k \in A\}|\mathbf{X}\Big) + \frac{1}{N+n}E\left(\sum_{i=1}^{n} I\{Y_i \in A\}|\mathbf{X}\right)$$
$$= \frac{N}{N+n}\mathbb{P}\Big(\{Z_1 \in A\}|\mathbf{X}\Big) + o(1), \tag{21}$$

where the second term is order $o(1)$ by the assumption that $n/N \to 0$, expectations taken with respect to (20).

Let $B_{n,1} = I\{Y_i \neq Z_1 : i = 1, \ldots, n\}$. Then, $\mathbb{P}\Big(\{Z_1 \in A\}|\mathbf{X}\Big) = \mathbb{P}\Big(\{Z_1 \in A\} \cap B_{n,1}|\mathbf{X}\Big) + \mathbb{P}\Big(\{Z_1 \in A\} \cap B_{n,1}^c|\mathbf{X}\Big)$. Integrating (20) over $\{Z_1 \in A\} \cap B_{n,1}$, deduce that the first term on the previous right-hand side equals

$$\frac{\mathbb{P}\{Z_1 \in A\}\Big(\sum_{\mathbf{P}} f^*(\mathbf{P}) \prod_{j=1}^{N(\mathbf{P})} \int \prod_{i \in C_j} \phi(X_i|\mu,\tau)\, dH(\mu,\tau)\Big)}{\sum_{\mathbf{P}} f(\mathbf{P}) \prod_{j=1}^{N(\mathbf{P})} \int \prod_{i \in C_j} \phi(X_i|\mu,\tau)\, dH(\mu,\tau)}, \tag{22}$$

where $f^*(\mathbf{P})$ is defined similarly to $f(\mathbf{P})$, but where each of the classification variables $K_1, \ldots, K_n$ must be different from the value 1. For each partition $\mathbf{P}$

$$\frac{f^*(\mathbf{P})}{f(\mathbf{P})} = \frac{(N-1)!/(N-1-N(\mathbf{P}))!}{N!/(N-N(\mathbf{P}))!} = \frac{N-N(\mathbf{P})}{N},$$

which is bounded between $1 - n/N$ and 1 (we can assume that $N \geq n$). Thus, deduce that (22) converges to $H(A)$. Setting $A = \Re \times \Re^+$ now shows that $\mathbb{P}(B_{n,1}^c|\mathbf{X}) \to 0$, and hence that (21) converges to $H(A)$ for each A.

5. Consistency with Dirichlet $(\alpha/N, \ldots, \alpha/N)$ Priors

A consistent posterior can be obtained by working with the Dirichlet prior for $\mathbf{p}$ defined by (6). As discussed in Section 3, Theorem 3, this prior induces a random probability measure $\mathcal{P}_N$ which strongly approximates the Dirichlet process (see (14)). With such a rich prior it is not surprising that it will induce a random density that is information dense at the true density f_0.

Lemma 2. *Let $\mathcal{F}$ be the set of all densities on $\Re$ with respect to Lebesgue measure. Let Π_N be the induced probability measure over $\mathcal{F}$ of the random density $\mathcal{P}_N(\phi(x|\cdot)) = \sum_{k=1}^{N} p_k \phi(x|\mu(Z_k), \tau(Z_k))$, for weights $\mathbf{p}$ defined by (6) and for Z_k which are i.i.d. from H, where H has a density that is positive over a rectangle containing the support for Q_0. Then, for each $\epsilon > 0$, $\liminf_{N\to\infty} \Pi_N\{f \in \mathcal{F} : D(f_0\|f) < \epsilon\} > 0$.*

By Lemma 2, the prior Π_N for the random density induced by (14) puts positive mass on each Kullback-Leibler neighborhood of f_0 (for a proof see the

Appendix). From this, it is fairly straightforward to establish consistency for f_0 by Proposition 2 of Barron (1988). This same method of proof was used by Roeder and Wasserman (1997). For a proof of the following theorem, see the Appendix.

Theorem 6. *Let Π_n^* be the posterior for (8) for the prior defined in Lemma 2, where we assume further that H has a density that is positive over a rectangle containing the support for Q_0 such that for each $\epsilon > 0$, $H\{\tau^{-1/2} \geq n\epsilon\} \leq \exp(-rn)$ for some $r = r(\epsilon) > 0$. If $N \to \infty$ such that $\log N/n \to 0$, then $\Pi_n^*\{f \in \mathcal{F} : \int |f(x) - f_0(x)|\, dx < \epsilon\} \to 1$ almost surely P_0^∞ for each $\epsilon > 0$.*

The conditions for Theorem 6 are easy to satisfy in practice. For example, the left-tail condition for the variance is satisfied if H is selected so that $\tau^{-1/2}$ (the inverse standard deviation) has a gamma distribution. Moreover, one could always choose $N = n$ for small sample sizes or $N = \sqrt{n}$ for large n to satisfy the constraint on N, although of course other automatic methods for selecting N are possible.

Remark 2. Ghosal, Ghosh and Ramamoorthi (1999) note that the Dirichlet process can also be used to consistently estimate the density in a weak sense for mixtures of normals (mixing over the mean and variance as here), in which the true mixing distribution is assumed to have a compact support (see their Remark 1, p.148). In personal correspondence with R.V. Ramamoorthi it was conjectured that this result could be strengthened to $\mathcal{L}_1$ consistency for f_0 under conditions similar to those in Theorem 6. These results are perhaps the closest analogue to Theorem 6 that we are aware of. There is also the work of Genovese and Wasserman (2000) and Ghosal and van der Vaart (2001) which is relevant. Both look at the same scenario as Ghosal, Ghosh and Ramamoorthi (1999), although they consider the more difficult problem of deriving rates of estimation for the density. This naturally requires more stringent assumptions, making direct comparisons of results somewhat difficult. Briefly though, Genovese and Wasserman (2000, Section 4.1) show that the use of a Gaussian sieve of dimension of order $(n \log n)^{2/3}$ yields a rate of estimation for the density arbitrarily close to $(\log n/n)^{1/6}$ when the unknown mixing distribution is assumed to have a compact support. In studying the same problem, Ghosal and van der Vaart (2001, Theorem 5.2) show that the use of a Dirichlet process prior yields the better rate of $(\log n)^\epsilon/\sqrt{n}$ for some $\epsilon > 0$, a near parametric rate, although they assume that the variance is constrained to lie in a fixed compact set.

5.1. Consistency for the mixing distribution

So far our discussion of consistency has centered around the problem of estimation for the density f_0. However, as emphasized in the introduction, it is

often the unknown mixing distribution Q_0 that is the primary focus of analysis in finite normal mixture problems.

To understand some of the difficulties in establishing consistency for Q_0, one should recognize that $\mathcal{L}_1$ consistency for f_0, as proven for example in Theorem 6, does *not* automatically imply consistency for the mixing distribution. The problem is that, although finite normal mixtures are identified (see Theorem 1), the closure of the space is not identified. To convert Theorem 6 into a result for Q_0 we need a stronger form of identification. Theorem 2 gives us such a tool. By requiring that our space of mixing distributions satisfy a uniform moment condition, we can use Theorem 6 to establish consistency for the mixing distribution.

Theorem 7. *Suppose the conditions of Theorem* 6 *hold. Furthermore, suppose the distribution H is permitted to depend upon n so that, for some constant $C > 0$,*

$$H_n\left\{\frac{\mu^2}{2(\tau^* - \tau)} > C\right\} \leq \exp(-nr) \tag{23}$$

for $r = r(C) > 0$, where τ^* is defined in Theorem 2. *Then, $\Pi_n^*(\mathcal{N}(Q_0)) \to 1$ almost surely P_0^∞ for each weak open neighborhood $\mathcal{N}(Q_0)$ containing Q_0.*

Observe that Theorem 7 holds trivially if $\log N/n \to 0$ and if H has a positive density with a compact support which contains the support of Q_0. However, condition (23) also allows for different scenarios. By allowing H to depend upon n, consistency for Q_0 can also be guaranteed if the tails for H_n decrease exponentially with n (notice that for τ it is the values for $\tau < \tau^*$ that we have to be careful with in H_n). For a proof of the theorem see the Appendix.

6. Gibbs Sampling

By conceptualizing the space for the finite normal mixture model as the space of finite distributions $\mathcal{Q}_F$, we have shown that our model is identified (Theorem 1), and, with the use of the weak limit Dirichlet process measure (14), we have outlined a Bayesian sieve approach for consistent density estimation (Theorem 6) and consistent estimation of the mixing distribution (Theorem 7). To complete the story, we outline a method for drawing values directly from the posterior of Q_0 using a simple Gibbs sampling method called the *blocked Gibbs sampler* (Ishwaran and Zarepour (2000) and Ishwaran and James (2001)). The key to its success here implicitly lies in our use of a finite dimensional Dirichlet prior, which ensures that our model is made up of a finite number of variables, thus allowing us to draw values directly from the posterior of Q_0 using a few simple multivariate conditional draws.

The blocked Gibbs sampler works by drawing values from the conditional distribution of $(\mathbf{K}, \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{p}|\mathbf{X})$, which by (3) and (7) has a density proportional to

$$\prod_{i=1}^{n} \phi(X_i|\mu_{K_i}, \tau_{K_i}) \prod_{i=1}^{n} \left(\sum_{k=1}^{N} p_k \, \delta_k(dK_i) \right) d\pi_N(\mathbf{p}) \, dH^N(\boldsymbol{\mu}, \boldsymbol{\tau}),$$

where π_N is the Dirichlet distribution (6). The method works by iteratively draw values from the conditional distributions $(\mathbf{K}|\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{X})$, $(\boldsymbol{\mu}|\boldsymbol{\tau}, \mathbf{K}, \mathbf{X})$, $(\boldsymbol{\tau}|\boldsymbol{\mu}, \mathbf{K}, \mathbf{X})$ and $(\mathbf{p}|\mathbf{K})$. Each cycle of the Gibbs sampler produces a draw $(\mathbf{K}^*, \boldsymbol{\mu}^*, \boldsymbol{\tau}^*, \mathbf{p}^*)$ from the posterior of $(\mathbf{K}, \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{p}|\mathbf{X})$ and a draw $\mathcal{P}_N^*(\cdot) = \sum_{k=1}^{N} p_k^* \, \delta_{(\mu_k^*, \tau_k^*)}(\cdot)$, from the posterior of $\mathcal{P}_N$, and thus can be used to directly estimate Q_0.

Each of the full conditional distributions can be drawn exactly, including the draw from $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$, if H is a conjugate prior. For example, the conditional for $\mathbf{p}$ is simply an updated Dirichlet distribution with parameters $\alpha_k = \alpha/N + n_k$, where n_k is the number of K_i's which equal k. In practice the value for α can be set to some fixed number, often the choice $\alpha = 1$ works well. Alternatively, one can include α as a parameter in the model and update it as part of the Gibbs procedure. See Ishwaran and Zarepour (2000) for more details.

7. Appendix: Proofs

Proof of Theorem 2. The proof is inspired by Theorem 6 in Teicher (1960). Without loss of generality assume that $\tau_{1,0} \geq \cdots \geq \tau_{d,0}$. The equality in distribution (10) implies equality of moment generating functions. Thus,

$$\sum_{k=1}^{d} p_{k,0} \int_{\Re} \exp(tx) \phi(x|\mu_{k,0}, \tau_{k,0}) \, dx = \int \left(\int_{\Re} \exp(tx) \phi(x|\mu, \tau) \, dx \right) Q(d\mu, d\tau)$$

for each t. Hence,

$$\begin{aligned} p_{1,0} + \sum_{k=2}^{d} p_{k,0} \exp \left[t(\mu_{k,0} - \mu_{1,0}) + \frac{1}{2} t^2 (\tau_{k,0} - \tau_{1,0}) \right] \\ = \int \exp \left[t(\mu - \mu_{1,0}) + \frac{1}{2} t^2 (\tau - \tau_{1,0}) \right] Q(d\mu, d\tau). \end{aligned} \tag{24}$$

We can assume without loss of generality that $\mu_{k,0} < \mu_{1,0}$ whenever $\tau_{k,0} = \tau_{1,0}$. Thus, the limit of the left-hand side is $p_{1,0}$ as $t \to +\infty$. The limit on the right-hand side will be $+\infty$ or 0 unless $Q\{\tau > \tau_{1,0}\} = Q\{\mu \neq \mu_{1,0}, \tau = \tau_{1,0}\} = 0$ (use Fatou's Lemma and the Monotone Convergence Theorem). This argument shows that we can restrict attention to the sets $\mathcal{Y}_1 = \{\mu = \mu_{1,0}, \tau = \tau_{1,0}\}$ and $\mathcal{Y}_2 = \{\tau < \tau_{1,0}\}$. Hence, we can rewrite the right-hand side of (24) as

$$Q(\mathcal{Y}_1) + \int_{\mathcal{Y}_2} \exp \left\{ -t\mu_{1,0} - \frac{1}{2}(\tau_{1,0} - \tau) \left[t - \frac{\mu}{(\tau_{1,0} - \tau)} \right]^2 \right\} \psi(\mu, \tau) \, Q(d\mu, d\tau), \tag{25}$$

where $\psi(\mu,\tau) = \exp(\mu^2/[2(\tau_{1,0}-\tau)])$. The moment condition for Q implies that ψ is Q-integrable. Thus, letting $t \to +\infty$, deduce by the Dominated Convergence Theorem that the integral in (25) is zero, so $Q(\mathcal{Y}_1) = p_{1,0}$.

Now subtract $p_{1,0}\exp(t\mu_{1,0}+t^2\tau_{1,0}/2)$ on the left and right-hand sides of (24). Repeat the above argument a finite number of times to obtain $Q = Q_0$.

Proof of Theorem 3. To prove 1(a), we show that, for each positive continuous function g with compact support, $\mathcal{P}_N(g) \overset{\text{a.s}}{\to} E(g(Z))$ where Z has the distribution H. Let $S_N = \mathcal{P}_N(g) = \sum_{k=1}^N p_{k,N}\, g(Z_k)$. Recall from Remark 1 that we can write $p_{k,N}$ as G_k/G, where G_k are independent Gamma(λ_k) random variables and $G = \sum_{k=1}^N G_k$. Now letting $\xi_k = g(Z_k)$, write S_N as

$$\frac{\sum_{k=1}^N [G_k\xi_k - \lambda_k E(g(Z))]/N + E(g(Z))\sum_{k=1}^N \lambda_k/N}{\sum_{k=1}^N (G_k-\lambda_k)/N + \sum_{k=1}^N \lambda_k/N}.$$

To prove our result, we apply the Khintchine-Kolmogorov Convergence Theorem (Chow and Teicher (1978, Chapter 5.1)) to the above numerator and denominator separately to show $S_N \overset{\text{a.s}}{\to} E(g(Z))$. For the numerator, the first sum converges to zero a.s if its second moment converges to zero. From the inequality

$$\sum_{k=1}^N \text{Var}\,(G_k\xi_k)/N^2 \le E(g(Z)^2)\sum_{k=1}^N (\lambda_k+\lambda_k^2)/N^2,$$

it suffices to show that the above right-hand side converges to zero. Note that $\sum_{k=1}^\infty \lambda_k/k^2 \le \sum_{k=1}^\infty (\lambda_k^2+1)/k^2 < \infty$. Thus, by Kronecker's Lemma, the boundedness of g, and our assumptions regarding λ_k, deduce that the above right-hand side converges to zero and that the numerator converges a.s to $E(g(Z))\lambda_0$. A similar argument shows that the denominator converges a.s to λ_0, and therefore $S_N \overset{\text{a.s}}{\to} E(g(Z))$.

To prove 1(b), we show that $S_N = \mathcal{P}_N(g) = \sum_{k=1}^N p_{k,N}\, g(Z_k) \overset{\text{p}}{\to} E(g(Z))$, for $p_{k,N}$ defined by $\alpha_{k,N} = \lambda_N$. Rewrite S_N as

$$\sum_{k=1}^N \frac{G_{k,N}}{G_N}\xi_k = \frac{1}{N\lambda_N}\sum_{k=1}^N G_{k,N}\,\xi_k + \sum_{k=1}^N \frac{G_{k,N}}{G_N}\xi_k\left(1-\frac{G_N}{N\lambda_N}\right), \tag{26}$$

where $G_{k,N}$ are independent Gamma(λ_N) variables, $\xi_k = g(Z_k)$ and $G_N = \sum_{k=1}^N G_{k,N}$.

Call $S_{N,1}$ the first term on the right-hand side of (26). Then, from the assumption that $N\lambda_N \to \infty$ and using the fact that $\xi_k < c$ for some finite c, deduce that

$$V(S_{N,1}) \le \frac{c^2}{(N\lambda_N)^2}\sum_{k=1}^N E(G_{k,n}^2) = \frac{c^2(1+\lambda_N)}{N\lambda_N} = o(1).$$

We have $E(S_{N,1}) = E(g(Z))$, thus from Chebyshev's inequality $S_{N,1} \xrightarrow{\text{p}} E(g(Z))$. Meanwhile, the second term on the right-hand side of (26) can be bounded in absolute value by $c \times |1 - G_N/(N\lambda_N)|$. A similar application of Chebyshev's inequality shows that this term converges in probability to zero, and consequently that $S_N \xrightarrow{\text{p}} E(g(Z))$.

For 2(a) see Ishwaran and Zarepour (2002), while 2(b) can be proven from the results in Kingman (1975, Section 6) or Kingman (1993, Section 9.3) pertaining to the Poisson-Dirichlet distribution. To complete the proof of the theorem we need part 3, which will follow if we can show that $\mathcal{P}_N(A) \xrightarrow{\text{d}} \text{Bernoulli}(H(A))$, for each measurable set A. First notice that $(\mathcal{P}_N(A) \mid \mathcal{K}_N) \sim \text{Beta}(\lambda_N \mathcal{K}_N, \lambda_N(N - \mathcal{K}_N))$, where $\mathcal{K}_N = \#\{k : Z_k \in A\}$, $\#$ is the cardinality of a set. Therefore, integrating over $\mathcal{K}_N$, the characteristic function for $\mathcal{P}_N(A)$ is

$$\psi_N(t) = E \exp\left(it\mathcal{P}_N(A)\right) = 1 + \sum_{j=1}^{\infty} \frac{(it)^j}{j!} E\left(C_{j,N}\right),$$

where

$$C_{j,N} = \frac{(\mathcal{K}_N \lambda_N)^{(j)}}{(N\lambda_N)^{(j)}} = \frac{\mathcal{K}_N}{N} \times \frac{(\mathcal{K}_N \lambda_N + 1)^{(j-1)}}{(N\lambda_N + 1)^{(j-1)}},$$

and where $a^{(0)} = 1$ and $a^{(r)} = a(a+1)\cdots(a+r-1)$ for any real number a and integer $r > 0$.

By the Strong Law of Large Numbers, $C_{j,N} \xrightarrow{\text{a.s}} H(A)$ for each j because $N\lambda_N \to 0$. Each $C_{j,N}$ is bounded by one so, by the Dominated Convergence Theorem, $E\left(C_{j,N}\right) \to H(A)$. One more application of the Dominated Convergence Theorem yields $\psi_N(t) \to 1 + H(A) \sum_{j=1}^{\infty} (it)^j/j! = 1 + H(A)\left(\exp(it) - 1\right)$, which is the characteristic function for a Bernoulli$(H(A))$ distribution.

Proof of Lemma 1. Let $n_k = \#\{i : K_i = k\}$, for $k = 1, \ldots, N$. Observe that $n_1 + \cdots + n_N = n$. The joint density for $(n_1, \ldots, n_N)$ is

$$\begin{aligned} f(n_1, \ldots, n_N) &= \frac{n!}{n_1! \cdots n_N!} E(p_1^{n_1} \cdots p_N^{n_N}) \\ &= \frac{n!}{n_1! \cdots n_N!} \frac{\Gamma(N) \prod_{k=1}^{N} \Gamma(1 + n_k)}{\Gamma(N + n)} = \binom{N + n - 1}{N - 1}^{-1}, \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function. This calculation yields the Bose-Einstein distribution, which is the distribution arising from placing n indistinguishable balls randomly into N distinct urns. In particular, the above probability can also be interpreted as the probability that urns $1, \ldots, N$ contain $n_1, \ldots, n_N$ balls respectively (all configurations are equally likely). From this analogy, it follows

that D_n equals the number of non-empty urns. Therefore,

$$\mathbb{P}\{D_n = k\} = \binom{N}{k}\binom{n-1}{k-1}\binom{N+n-1}{N-1}^{-1},$$

where the first term equals the number of ways of selecting k urns, and the second term equals the number of ways of placing n balls into these k urns so that they are all non-empty.

Proof of Lemma 2. For any $f(\cdot) = \sum_{k=1}^{N} p_k \phi(\cdot|\mu_k, \tau_k)$,

$$D(f_0\|f) \le \int f_0(x) \log\left(\frac{\sum_{k=1}^{d} p_{k,0}\phi(x|\mu_{k,0},\tau_{k,0})}{\sum_{k=1}^{d} p_k \phi(x|\mu_k,\tau_k)}\right) dx. \tag{27}$$

By continuity, the right-hand side can be made smaller than $\epsilon > 0$ if $|p_k - p_{k,0}| \le \eta_0$, $|Z_k - Z_{k,0}| \le \eta_0, k = 1, \ldots, d$, for some $\eta_0 = \eta_0(\epsilon) > 0$, where $Z_{k,0} = (\mu_{k,0}, \tau_{k,0})$ and $Z_k = (\mu_k, \tau_k)$.

Alternatively, the right-hand side of (27) is smaller than $\epsilon > 0$ if

$$|G_k - p_{k,0}| \le \eta, \quad |Z_k - Z_{k,0}| \le \eta, \qquad k = 1, \ldots, d, \tag{28}$$

$$\sum_{k=d+1}^{N} G_k \le \eta \tag{29}$$

for some small $\eta = \eta(\epsilon) > 0$, where $p_k = G_k/G$ and $G = \sum_{k=1}^{N} G_k$ for $G_k > 0$. This follows from the inequality

$$\frac{p_{k,0} - \eta}{1 + \eta(d+1)} \le p_k = \frac{G_k}{\sum_{k=1}^{N} G_k} \le \frac{p_{k,0} + \eta}{1 - \eta d}, \qquad k = 1, \ldots, d.$$

If $\mathbf{p}$ has the Dirichlet distribution (6), then $p_k = G_k/G$ where G_k are i.i.d. Gamma(α/N) random variables. Therefore, (28) and (29) will be satisfied with probability

$$\mathbb{P}\{\sum_{k=d+1}^{N} G_k \le \eta\} \prod_{k=1}^{d} \mathbb{P}\{|Z_k - Z_{k,0}| \le \eta\} \prod_{k=1}^{d} \mathbb{P}\{|G_k - p_{k,0}| \le \eta\}.$$

The first probability remains bounded away from zero, by noting that $\sum_{k=d+1}^{N} G_k$ converges in distribution to a Gamma(α) random variable, while the second term is positive by our assumption of a positive density for H over the support of Q_0. It is easy to verify that the third term is $O(N^{-d})$. For a small enough η, there are $N!/((N-d)!d!)$, or $O(N^d)$, mutually exclusive ways of choosing the coordinates of $\mathbf{p}$ and $\mathbf{Z}$ to satisfy (28) and (29) (for a small enough η, when (29) holds, each $G_{d+1}, \ldots, G_N$ is smaller than any $G_1, \ldots, G_d$). Because all of these sets have the

same probability, deduce that the right-hand side of (27) is smaller than $\epsilon > 0$ with a probability that remains bounded away from zero.

Proof of Theorem 6. Lemma 2 establishes that Π_N is information dense at f_0. This is condition (A) of Proposition 2 in Barron (1988), and our theorem will be proved if we can verify condition (B) of the proposition: for each $\epsilon > 0$,

$$\Pi_N\Big\{f \in \mathcal{F} : \int |f(x) - f^{T_N}(x)|\, dx > \epsilon\Big\} \leq \exp(-r_1 n) \tag{30}$$

for some constant $r_1 = r_1(\epsilon) > 0$, where for each $f \in \mathcal{F}$, the density f^{T_n} is a "theoretical histogram" with bins $\{\ldots, (-1/n, 0], (0, 1/n], \ldots\}$ and heights defined by $f^{T_N}(x) = n\int_{(j-1)/n}^{j/n} f(u)\, du$ for each x, where $j \equiv j(x)$ is the integer satisfying $(j-1)/n < x \leq j/n$.

For any $f(\cdot) = \sum_{k=1}^N p_k \phi(\cdot|\mu_k, \tau_k)$,

$$\begin{aligned}\int |f(x) - f^{T_N}(x)|\, dx &\leq \sum_{k=1}^N p_k \int |\phi(x|\mu_k, \tau_k) - \phi^{T_n}(x|\mu_k, \tau_k)|\, dx \\ &= \sum_{k=1}^N p_k \sum_{j=-\infty}^{\infty} \int_{(j-1)/n}^{j/n} \Big|\phi(x|\mu_k, \tau_k) - n\int_{(j-1)/n}^{j/n} \phi(u|\mu_k, \tau_k)\, du\Big|\, dx.\end{aligned}$$

By the Mean Value Theorem, there exists a u_j, $(j-1)/n < u_j < j/n$, so that for each $(j-1)/n < x \leq j/n$,

$$\begin{aligned}&\Big|\phi(x|\mu_k, \tau_k) - n\int_{(j-1)/n}^{j/n} \phi(u|\mu_k, \tau_k)\, du\Big| = \Big|\phi(x|\mu_k, \tau_k) - \phi(u_j|\mu_k, \tau_k)\Big| \\ &= \Big|\int_{u_j}^x \phi'(u|\mu_k, \tau_k)\, du\Big| \leq \int_{(j-1)/n}^{j/n} |\phi'(u|\mu_k, \tau_k)|\, du.\end{aligned}$$

(This inequality is due to Roeder and Wasserman (1997, p.901)). Therefore,

$$\int |f(x) - f^{T_N}(x)|\, dx \leq \sum_{k=1}^N \frac{p_k}{n} \int |\phi'(u|\mu_k, \tau_k)|\, du = \sqrt{\frac{2}{\pi}} \sum_{k=1}^N \frac{p_k \tau_k^{-1/2}}{n}.$$

By using the previous inequality, we can bound the left-hand side of (30) with

$$\begin{aligned}\mathcal{P}_N\Big\{\sum_{k=1}^N p_k \tau_k^{-1/2} \geq n\epsilon\Big\} &\leq H^N\{\tau_k^{-1/2} \geq n\epsilon, \text{ for some } k = 1, \ldots, N\} \\ &\leq N \times H\{\tau_1^{-1/2} \geq n\epsilon\},\end{aligned}$$

which is bounded by $N\exp(-rn)$ by our assumption on the tail behavior of τ. Now use the constraint on the size of N to verify (30).

Proof of Theorem 7. Let $f_Q(x) = \int \phi(x|\mu,\tau)\, dQ(\mu,\tau)$ denote the normal mixture density for a distribution Q over $\Re \times \Re^+$. Define $\mathcal{F}_* = \{f_Q \in \mathcal{F} : \int \psi(\mu,\tau)\, dQ(\mu,\tau) \leq \exp(C)\}$, where $\psi(\mu,\tau) = \exp\left(\mu^2/[2(\tau^* - \tau)]\right)$. Notice that $\mathcal{F}_*$ is identified by Theorem 2. We have,

$$\Pi_N(\mathcal{F}_*^c) = \mathcal{P}_N\{\sum_{k=1}^{N} p_k \psi(Z_k) > \exp(C)\} \leq N \times H_n\{\psi(Z_k) > \exp(C)\}.$$

Deduce by (23) that $\Pi_N(\mathcal{F}_*^c)$ is exponentially small. Therefore we can replace the set $\mathcal{F}$ in Theorem 6 by $\mathcal{F}_*$ (for example see Proposition 1 of Barron (1988) where $\mathcal{F}_*^c$ acts as the set B_n). Consequently, for each $\epsilon > 0$, the posterior concentrates almost surely on the set of densities $\{f \in \mathcal{F}_* : \int |f(x) - f_0(x)|dx < \epsilon\}$. As $\mathcal{F}_*$ is identified, this implies that the posterior concentrates almost surely on each weak neighborhood of Q_0. If this were not the case, then we could find a sequence $f_{Q_n} \in \mathcal{F}_*$ with limit f_Q where $Q \neq Q_0$ but $f_Q = f_0$. However, by the moment constraint for $\mathcal{F}^*$ we have $\int \psi(\mu,\tau)\, dQ(\mu,\tau) \leq \exp(C)$. This contradicts the identification implied by Theorem 2, which holds for distributions over $\Re \times \Re^+$ and (by inspection of its proof) for distributions over the closure as well.

Acknowledgements

The authors are greatly indebted to Lancelot F. James and Albert Y. Lo for helpful discussion and advice on earlier drafts of this work. The authors also thank the reviewers of the paper, including the editor Yi-Ching Yao, for their many constructive comments.

References

Barron, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report 7, Department of Statistics, University of Illinois, Champaign, IL

Billinglsey, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

Celeux, G., Hurn, M and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Assoc.* **95**, 957-970.

Chen, J. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23**, 221-233.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90**, 1313-1321.

Chow, Y. S. and Teicher, H. (1978). *Probability Theory: Independence, Interchangeability, Martingales.* Springer-Verlag, New York.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. Ser. B* **56**, 363-375.

Escobar, M. D. (1988). Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Unpublished Ph.D. thesis, Department of Statistics, Yale University.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89**, 268-277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**, 577-588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209-230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615-629.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics* (Edited by M. H. Rizvi, J. Rustagi and D. Siegmund), 287-302. Academic Press, New York.

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

Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* 0**28**, 1105-1127.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 143-158.

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29**, 1233-1263.

Green, P., and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scand. J. Statist.* **28**, 355-377.

Grenader, U. (1981). *Abstract Inference.* Wiley, New-York.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96**, 161-173.

Ishwaran H., James, L. F. and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J. Amer. Statist. Assoc.* **96**, 1316-1332.

Ishwaran, H. and Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87**, 371-390.

Ishwaran, H. and Zarepour, M. (2002). Exact and approximate sum-representations for the Dirichlet process. *Canad. J. Statist.* **30**, 1-15.

Kemperman, J. H. B. (1969). On the optimum rate of transmitting information. *Ann. Math. Statist.* **40**, 2156-2177.

Kingman, J. F. C. (1975). Random discrete distributions. *J. Roy. Statist. Soc. Ser. B* **37**, 1-22.

Kingman, J. F. C. (1993). *Poisson Processes.* Oxford University Press, Oxford.

Kuo, L. (1986). Computations of mixtures of Dirichlet processes. *SIAM J. Sci. Statist. Comput.* **7**, 60-71.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications.* NSF-CBMS *Regional Conference Series in Probability and Statistics*, Volume 5. IMS, California.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12**, 351-357.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Comm. Statist. Simulation Comput.* **23**, 727-741.

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

Muliere P. and Secchi, P. (1995). A note on a proper Bayesian bootstrap. Technical report No. 18, Dipartimento di Economia Politica e Metodi Quantitativi, Università degli Sudi di Pavia.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9**, 249-265.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory* (Edited by T. S. Ferguson, L. S. Shapley and J. B. MacQueen), 245-267. IMS Lecture Notes-Monograph series, Vol 30. Hayward CA: Institute of Mathematical Statistics.

Priebe, C. E. (1994). Adaptive mixtures. *J. Amer. Statist. Assoc.* **89**, 796-806.

Resnick, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes.* Springer-Verlag, New York.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59**, 731-792.

Roeder, K. (1992). Semiparametric estimation of normal mixture densities. *Ann. Statist.* **20**, 929-943.

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

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22**, 580-615.

Teicher H. (1960). On the mixture of distributions. *Ann. Math. Statist.* **31**, 55-73.

Teicher H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.* **32**, 1265-1269.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23**, 339-363.

Zhang, C. H. (1990). Fourier methods for estimating mixing densities and distributions. *Ann. Statist.* **18**, 806-831.

Cleveland Clinic Foundation, Department of Biostatistics / Wb4, 9500 Euclid Avenue, Cleveland, OH 44195, U.S.A.

E-mail: ishwaran@bio.ri.ccf.org

University of Ottawa, Department of Math and Statistics, P.O. Box 450, STN A, Ottawa, Ontario, Canada, K1N 6N5, U.S.A.

E-mail: zarepour@expresso.mathstat.uottawa.ca

(Received June 2000; accepted January 2002)