Bayesian Density Estimation and Inference Using Mixtures

By

Michael D. Escobar & Mike West*
Department of Statistics *Institute of Statistics*
University of Toronto *& Decision Sciences*
Toronto, Ontario M5S 1A8 *Duke University*
CANADA *Durham, NC 27708-0251*

ABSTRACT

We describe and illustrate Bayesian inference in models for density estimation using mixtures of Dirichlet processes. These models provide natural settings for density estimation, and are exemplified by special cases where data are modelled as a sample from mixtures of normal distributions. Efficient simulation methods are used to approximate various prior, posterior and predictive distributions. This allows for direct inference on a variety of practical issues, including problems of local versus global smoothing, uncertainty about density estimates, assessment of modality, and the inference on the numbers of components. Also, convergence results are established for a general class of normal mixture models.

Keywords: Kernel estimation; Mixtures of Dirichlet processes; Multimodality; Normal mixtures; Posterior sampling; Smoothing parameter estimation

* Michael D. Escobar is Assistant Professor, Department of Statistics and Department of Preventive Medicine and Biostatistics, University of Toronto, M5S 1A8, Canada. Mike West is Professor and Director, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251. Michael D. Escobar was partially financed by National Cancer Institute #RO1-CA54852-01, a National Research Service Award from NIMH Grant #MH15758 and by the National Science and Engineering Research Council of Canada. Mike West was partially financed by the National Science Foundation under grants DMS-8903842 and DMS-9024793. The authors would also like to thank Hani Doss and Steve MacEachern for helpful discussions.

1. INTRODUCTION

Models for uncertain data distributions based on mixtures of standard components, such as normal mixtures, underly mainstream approaches to density estimation, including kernel techniques (Silverman 1986), nonparametric maximum likelihood (Lindsay 1983), and Bayesian approaches using mixtures of Dirichlet processes (Ferguson 1983). The latter provide theoretical bases for more traditional, nonparametric methods, such as kernel techniques, and hence a modelling framework within which the various practical problems of local versus global smoothing, smoothing parameter estimation, and the assessment of uncertainty about density estimates, may be addressed. In contrast with nonparametric approaches, a formal model allows these problems to be addressed directly via inference about the relevant model parameters. We discuss these issues using data distributions derived as normal mixtures in the framework of mixtures of Dirichlet processes, essentially the framework of Ferguson (1983). West (1990) discusses these models in a special case of the framework studied here. That paper is concerned with developing approximations to predictive distributions based on a clustering algorithm motivated by the model structure, and draws obvious connections with kernel approaches. The current paper develops, in a more general framework, a computational method that allow for the evaluation of posterior distributions for all models parameters and direct evaluation of predictive distributions. As a natural by-product, we develop approaches to inference about the numbers of components and modes in a population distribution.

The computational method developed here is a direct extension of the method in Escobar (1988, 1994) and is another example of a Gibbs sampler or Markov Chains Monte Carlo method which has recently been popularised by Gelfand and Smith (1990). Some of the earlier references on Markov Chain Monte Carlo methods include Geman and Geman (1984), Hasting (1970), Metropolis, et al (1953), and Tanner and Wong (1987). Besag and Green (1993) and Smith and Roberts (1993) recently review Markov Chain Monte Carlo methods.

The basic normal mixture model, similar to that of Ferguson (1983), is described as follows. Suppose data $Y_1, \ldots, Y_n$ are conditionally independent and normally distributed, $(Y_i|\pi_i) \sim N(\mu_i, V_i)$, with means μ_i and variances V_i determining the parameters $\pi_i = (\mu_i, V_i)$, $i = 1, \ldots, n$. Suppose further that the π_i come from some prior distribution on $\Re \times \Re^+$. Having observed data $D_n = \{y_1, \ldots, y_n\}$, with y_i the observed value of Y_i, the distribution of a future case is a mixture of normals; the relevant density function $Y_{n+1} \sim N(\mu_{n+1}, V_{n+1})$ mixed with respect to the posterior predictive distribution for $(\pi_{n+1}|D_n)$. If the common prior distribution for the π_i is uncertain and modelled, in whole or in part, as a Dirichlet process, then the data come from a Dirichlet mixture of normals (Ferguson 1983; Escobar 1988, 1994; West 1990). The important special case in which $V_i = V$ has been studied widely; references appear in West (1990, 1992) who considers the common setup in which the μ_i have a uncertain prior which is modelled as a Dirichlet process with a normal base measure; see also West and Cao (1993). The connections with kernel estimation techniques are explored in these papers, as are some analytic and numerical approximations to the predictive distributions derived from such models. The analysis covers problems of estimating the V_i. Escobar (1988, 1994) considers similar models, differing in the use of a uniform Dirichlet process base

measure, and assuming $V_i = V$, known. Ferguson (1983), using Monte Carlo techniques from Kuo (1986), considers more generally the case of possibly distinct and uncertain V_i. The suitability of this model form for density estimation has been well-argued there, and in the earlier references. With a suitable Dirichlet process prior structure, described below, this model produces predictive distributions qualitatively similar to kernel techniques, but catering for differing degrees of smoothing across the sample space through the use of possibly differing variances V_i. The structure is such that the posterior distribution will strongly support common values of individual parameters π_i and π_j for data points y_i and y_j that are close, thus combining information locally in the sample space to estimate the local structure. We proceed with this general model, noting that similar discussion and analysis applies to the more restricted global smoothing version in which $V_i = V$.

Section 2 completes the model specification and reviews some implications. Section 3 develops the computational technique for Monte Carlo analysis, extending the technique of Escobar (1988, 1994). This improves on the importance sampling based simulation analysis of Ferguson (1983) and Kuo (1986) since it provides for efficient sampling from the posterior distribution of the model parameters π_i. Section 4 discusses prior and posterior inference about the number of components of a discrete mixture, and multi-modality, and this is further developed in application, in Section 5, to a problem in astronomy recently considered by Roeder (1990). Section 6 discusses some advanced techniques related to the smoothing parameter for the Dirichlet process with a further illustration. A summary discussion of the paper is contained in Section 7. In an appendix, we discuss some convergence issues for the Monte Carlo analysis.

2. NORMAL MIXTURE MODELS AND PREDICTION

Suppose the normal means and variances π_i come from some prior distribution $G(.)$ on $\Re \times \Re^+$. If $G(.)$ is uncertain and modelled as a Dirichlet process, then the data come from a Dirichlet mixture of normals (Escobar 1994; Ferguson 1983; West 1990). In particular, we suppose that $G \sim D(\alpha G_0)$, a Dirichlet process defined by α, a positive scalar, and $G_0(.)$, a specified bivariate distribution function over $\Re \times \Re^+$. $G_0(.)$ is the prior expectation of $G(.)$, so that $E\{G(\pi)\} = G_0(\pi)$, for all $\pi \in \Re \times \Re^+$, and α is a precision parameter, determining the concentration of the prior for $G(.)$ about $G_0(.)$. Write $\pi = \{\pi_1, \ldots, \pi_n\}$.

A key feature of the model structure, and of its analysis, relates to the discreteness of $G(.)$ under the Dirichlet process assumption. Details may be found in Ferguson (1973). Briefly, in any sample π of size n from $G(.)$ there is positive probability of coincident values. See this as follows. For any $i = 1, \ldots, n$, let $\pi^{(i)}$ be π without π_i, $\pi^{(i)} = \{\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_n\}$. Then the conditional prior for $(\pi_i | \pi^{(i)})$ is

$$(\pi_i | \pi^{(i)}) \sim \alpha a_{n-1} G_0(\pi_i) + a_{n-1} \sum_{j=1,\, j \neq i}^{n} \delta_{\pi_j}(\pi_i), \tag{1}$$

where $\delta_{\pi_j}(\pi)$ denotes a unit point mass at $\pi = \pi_i$, and $a_r = 1/(\alpha + r)$ for positive integers r.

Similarly, the distribution of $(\pi_{n+1}|\pi)$ is given by

$$(\pi_{n+1}|\pi) \sim \alpha a_n G_0(\pi_{n+1}) + a_n \sum_{i=1}^{n} \delta_{\pi_i}(\pi_{n+1}). \tag{2}$$

Thus, given π, a sample of size n from $G(.)$, the next case π_{n+1} represents a new, distinct value with probability αa_n, and is otherwise drawn uniformly from amongst the first n values. These first n values themselves behave as described by (1), and so with positive probability will reduce to some $k < n$ distinct values. Write the k distinct values amongst the n elements of π as $\pi_j^* = (\mu_j^*, V_j^*)$, $j = 1, \ldots, k$. Suppose there are n_j occurrences of π_j^* and let I_j be the index set for those occurrences; thus $\pi_i = \pi_j^*$ for $i \in I_j$ and $j = 1, \ldots, k$, with $n_1 + \ldots + n_k = n$. Immediately, (2) reduces to the mixture of fewer components

$$(\pi_{n+1}|\pi) \sim \alpha a_n G_0(\pi_{n+1}) + a_n \sum_{j=1}^{k} n_j \delta_{\pi_j^*}(\pi_{n+1}). \tag{3}$$

Theory summarised in Antoniak (1974) gives the prior for k induced by this Dirichlet process model. The prior distribution for k critically depends on α, stochastically increasing with α. For instance, for n moderately large, $E(k|\alpha, n) \approx \alpha \ln(1 + n/\alpha)$. In practical density estimation, suitable values of α will typically be small relative to n; $\alpha = 1$ corresponds to the initial prior $G_0(.)$ for π_{n+1} receiving the weight of one observation in the posterior (2) or (3). Then, for n between 50 and 250, say, the prior for k heavily favours single digit values.

To proceed, we need to specify the prior mean $G_0(.)$ of $G(.)$. A convenient form is the normal/inverse-gamma conjugate to the normal sampling model; thus, under $G_0(.)$, we assume $V_j^{-1} \sim G(s/2, S/2)$, a gamma prior with shape $s/2$ and scale $S/2$, and $(\mu_j|V_j) \sim N(m, \tau V_j)$, for some mean m, and scale factor $\tau > 0$. For the moment assume the prior parameters s, S, m and τ are specified. Generically, let $P(Y|D)$ represent the distribution of any quantity Y given any other D. Then, with respect to predicting Y_{n+1}, it is clear that $P(Y_{n+1}|\pi, D_n) \equiv P(Y_{n+1}|\pi)$, which may be evaluated as $\int P(Y_{n+1}|\pi_{n+1}) dP(\pi_{n+1}|\pi)$. The first component of the integrand is the normal sampling distribution, the second is given in (2), and these imply

$$(Y_{n+1}|\pi) \sim \alpha a_n T_s(m, M) + a_n \sum_{i=1}^{n} N(\mu_i, V_i), \tag{4}$$

where $T_s(m, M)$ is the Student-t distribution with s degrees of freedom, mode m and scale factor $M^{1/2}$, and $M = (1 + \tau)S/s$. Equivalently, using the reduced form (3), we have

$$(Y_{n+1}|\pi) \sim \alpha a_n T_s(m, M) + a_n \sum_{j=1}^{k} n_j N(\mu_j^*, V_j^*). \tag{5}$$

As discussed in Ferguson (1983) there are strong relationships between (4) and standard kernel density estimates (Silverman 1986). The standard kernel density estimator, using a normal kernel, would estimate $(Y_{n+1}|D_n)$ by $(Y_{n+1}|D_n) \sim n^{-1} \sum_{i=1}^{n} N(y_i, H)$, for some window width H.

In addition to obvious data-based estimation of smoothing parameters inducing varying window-widths across the sample space, (4) involves two types of shrinkage: the y_i's are shrunk towards their means, the μ_i's, and the density estimate is shrunk towards the initial prior, $T_s(m, M)$.

The Bayesian prediction, or density estimation problem is solved by summarising the unconditional predictive distribution

$$P(Y_{n+1}|D_n) = \int P(Y_{n+1}|\pi)dP(\pi|D_n). \tag{6}$$

Direct evaluation of (6) is computationally extremely involved for even rather small sample size n due to the inherent complexity of the posterior $P(\pi|D_n)$ (Antoniak 1974; Escobar 1992; Lo 1984; West 1990). Fortunately, Monte Carlo approximation is possible using extensions of the iterative technique in Escobar (1988, 1994), now described.

3. COMPUTATIONS

Recall that, for each i, $\pi^{(i)} = \{\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_n\}$. We note $(\pi_i|D_n)$ has the following conditional structure. For each i, the conditional posterior for $(\pi_i|\pi^{(i)}, D_n)$ is the mixture

$$(\pi_i|\pi^{(i)}, D_n) \sim q_0 G_i(\pi_i) + \sum_{j=1,\, j\neq i}^{n} q_j \delta_{\pi_j}(\pi_i), \tag{7}$$

where:

(a) $G_i(\pi_i)$ is the bivariate normal/inverse gamma distribution whose components are $V_i^{-1} \sim G((1+s)/2, S_i/2)$ with $S_i = S+(y_i-m)^2/(1+\tau)$, and $(\mu_i|V_i) \sim N(x_i, XV_i)$ with $X = \tau/(1+\tau)$ and $x_i = (m + \tau y_i)/(1 + \tau)$; and

(b) the weights q_j are defined as

$$q_0 \propto \alpha c(s)[1 + (y_i - m)^2/(sM)]^{-(1+s)/2}/M^{1/2}$$
$$q_j \propto \exp\{-(y_i - \mu_j)^2/(2V_j)\}(2V_j)^{-1/2}, \qquad j = 1, \ldots, n;\ j \neq i,$$

subject to $q_0 + \ldots + q_{i-1} + q_{i+1} + \ldots + q_n = 1$, with $M = (1 + \tau)S/s$ and $c(s) = \Gamma((1 + s)/2)\Gamma(s/2)^{-1}s^{-1/2}$.

Here $G_i(.)$ is just the posterior distribution of $(\pi_i|y_i)$ under a prior $G_0(.)$ and the weight q_0 is proportional to α times the marginal density of Y_i evaluated at the datum y_i using $G_0(.)$ as the prior for π_i. In our model, therefore, q_0 is proportional to α times the density function of $T_s(m, M)$ evaluated at y_i. The weight q_j is proportional to the likelihood of data y_i being a sample from the normal distribution $(Y_i|\pi_j)$, or just the density function of $N(\mu_j, V_j)$ at the point y_i. The conditional distribution, $(\pi_i|\pi^{(i)}, D_n)$, is a weighted mixture of our best guess of the prior G_0 with single atom distributions on the other values on which we conditioned. The weights are determined according to the relative predictive densities at the data value y_i.

These conditional distributions are easily sampled; given $\pi^{(i)}$, it is straightforward to sample from $(\pi_i|\pi^{(i)}, D_n)$. This fact is important in the iterative resampling process that provides a single, approximate draw from the joint posterior $p(\pi|D_n)$ as follows.

Algorithm I.

(i) Choose a starting value of π; reasonable initial values are samples from the individual conditional posteriors $G_i(.)$ in (7).

(ii) Sequentially sample elements of π by drawing from the distribution of $(\pi_1|\pi^{(1)}, D_n)$, then $(\pi_2|\pi^{(2)}, D_n)$, and so on up to $(\pi_n|\pi^{(n)}, D_n)$, with the relevant elements of the most recently sampled $\pi^{(i)}$ values inserted in the conditioning vectors at each step.

(iii) Return to (ii), and proceed iteratively until convergence.

The sampling process is computationally very straightforward. Note that, in implementation, the required computations are reduced through the fact that each of the mixtures (7) will reduce to typically fewer than the apparent n components due to the clustering of the elements of the elements of $\pi^{(i)}$. Using the earlier superscript $*$ to denote distinct values, suppose that the conditioning quantities $\pi^{(i)}$ in (7) concentrate on $k_i \leq n-1$ distinct values $\pi_j^* = (\mu_j^*, V_j^*)$, with some n_j taking this common value. Then (7) reduces to $(\pi_i|\pi^{(i)}, D_n) \sim q_0 G_i(\pi_i) + \sum_{j=1}^{k_i} q_j^* \delta_{\pi_j^*}(\pi_i)$ where the weights now include the n_j, viz $q_j \propto n_j \exp\{-(y_i - \mu_j^*)^2/(2V_j^*)\}(2V_j^*)^{-1/2}$.

The sampling process results in an approximate draw from $p(\pi|D_n)$. Escobar (1994) discusses theoretical aspects of convergence in the simpler case where V_i is known. Unfortunately, the proof in that simple case does not extend easily to this model, since the q_j can get arbitrarily close to 1. This results in a violation of the equicontinuity condition required in Escobar (1988, 1994), Feller (1971, pp 271-2) and Tanner and Wong (1987). Instead, we use the results from Tierney (in press) which are based on the monograph by Nummelin (1984). The theorem is stated below and the proof and additional discussion of convergence issues is contained in the appendix.

Let $Q_I(\pi(0), A)$ be the probability that, with initial value $\pi(0)$ and after one iteration, Algorithm I produces a sample value which is contained in the measurable set A. Let $Q_I^s(\pi(0), A)$ be the probability that, with initial value $\pi(0)$ and after s iterations, the Algorithm I produces a sample value which is contained in the measurable set A. For the Markov chain implied by Algorithm I, $Q_I(\cdot, \cdot)$ is called the transition kernel for the Markov chain. For an explicit representation of the transition kernel for similar algorithms involving Dirichlet processes, see Escobar (1994). For a fixed value of $\pi(0)$, $Q_I(\pi(0), \cdot)$ and $Q_I^s(\pi(0), \cdot)$ are probability measures, and for a fixed measurable set A, $Q_I^s(\cdot, A)$ and $Q_I^s(\cdot, A)$ are measurable functions. Let the metric $\|\cdot\|$ be the total variation norm as defined by Tierney (in press). Let $P_\pi(\cdot|D_n)$ be the posterior distribution of π. In the theorems which follow, the conditions "almost all" and "almost surely" are with respect to measure generated by the posterior distribution.

Theorem 1. *For almost all starting values of Algorithm I, the probability measure Q_I^s converges in total variation norm to the posterior distribution as s goes to infinity. That is, for almost all $\pi(0)$, $\lim_{s\to\infty} \|Q_I^s(\pi(0), \cdot) - P_\pi(\cdot|D_n)\| = 0$.*

The initial prior variance τ plays a critical role in determining the extent of smoothing in the analysis. For a given k distinct values amongst the elements of π, a larger value of τ leads to increased dispersion amongst the k group means μ_j^* which, for fixed V_j^*, leads to a greater

chance of multi-modality in the resulting predictive distribution. In restricted models with $V_j = V$, choice of τ relates to the choice of window-widths in traditional kernel density estimation. Rather typically, the information content of the data for estimating τ will be small unless the prior for V is reasonably informative. This is relevant in the more general setting here too.

The conditionally conjugate structure built into the model easily allows for an extension of the sampling based analysis to include learning about the prior parameters m and/or τ. Suppose independent priors of the form $m \sim N(a, A)$ and $\tau^{-1} \sim G(w/2, W/2)$, for some specified hyper-parameters a, A, w and W. It follows that

- given τ and π, m is conditionally independent of D_n and normally distributed with moments $E(m|\tau,\pi) = (1-x)a + x\bar{V}\sum(V_j^*)^{-1}\mu_j^*$ and $V(m|\tau,\pi) = x\tau\bar{V}$, where $x = A/(A+\tau\bar{V})$, $\bar{V}^{-1} = \sum(V_j^*)^{-1}$ and all sums are over $j = 1,\ldots,k$; also
- given m and π, τ is conditionally independent of D_n and has the inverse gamma posterior $(\tau^{-1}|m,\pi) \sim G((w+k)/2, (W+K)/2)$ where $K = \sum_{j=1}^{k}(\mu_j^* - m)^2/V_j^*$.

Incorporating m and/or τ into the iterative resampling scheme provides for sampling from the complete joint posterior of $(\pi, m, \tau|D_n)$. Thus steps (i) to (iii) above may be modified as follows:

Algorithm II.

(i) As in (i) above, generate an initial π conditional on a preliminary chosen values of m and τ.
(i') Sample m and τ (in some order) using the relevant distributions as just described.
(ii) As in (ii) above, using the most recently sampled values of m and τ.
(iii) Return to (i'), and proceed iteratively until convergence.

By extending the notation introduced for Algorithm I to Algorithm II, we get the next convergence theorem. The proof involves the straightforward extension of the arguments in the proof of Theorem 1.

Theorem 2. *For almost all starting values of Algorithm II, the probability measure Q_{II}^s converges in total variation norm to the posterior distribution as s goes to infinity. That is, for almost all $(\pi(0), m(0), \tau(0))$, $\lim_{s\to\infty} \|Q_{II}^s((\pi(0), m(0), \tau(0)), \cdot) - P_{\pi,m,\tau}(\cdot|D_n)\| = 0$.*

From specified initial values, we first iterate the sampling procedure to 'burn-in' the process to (approximate) convergence. Following burn-in, successively generated values of π, m and τ are assumed to be drawn from the posterior; denote these values $(\pi(r), m(r), \tau(r))$, for $r = 1,\ldots,N$, where N is the specified simulation sample size required. Approximate predictive inference now follows through the Monte Carlo approximation to (6) given by

$$P(Y_{n+1}|D_n) \approx N^{-1}\sum_{r=1}^{N} P(Y_{n+1}|\pi(r), m(r), \tau(r)), \tag{8}$$

with the summands given by the mixtures in (5), and the notation now explicitly recognizes the dependence on the sampled values of m and τ. Additional information available includes the sampled values of k, $\{k(r), r = 1,\ldots,N\}$, which directly provide a histogram approximation to

$p(k|D_n)$, of interest in assessing the number of components. The posteriors for m and/or τ may also be approximated by mixture of their conditional posteriors noted above, following general principles expounded by Gelfand and Smith (1990); for m, this leads to the mixture of normals $p(m|D_n) \approx N^{-1}\sum p(m|\tau(r),\pi(r))$, and for τ, the mixture of inverse gammas $p(\tau|D_n) \approx N^{-1}\sum p(\tau|m(r),\pi(r))$, the sums being over $r = 1,\ldots,N$ in each case.

Using Theorem 3 of Tierney (in press), it can be shown that the path averages of bounded functions converge almost surely to their posterior expectations. Therefore, estimates of the cumulative distribution functions, estimates of the probability functions of discrete random variables, and histogram estimates of probability density functions all converge almost surely to the posterior expectations. The next three theorems state that the estimates of the following probability density functions also converge almost surely to their posterior expectations.

Theorem 3. *The estimate of the predictive density, evaluated at any fixed point, is strongly consistent, for almost all starting values of the algorithm. That is, for almost all starting values, given any fixed value* $Y_{n+1,0}$,

$$N^{-1}\sum_{r=1}^{N} p(Y_{n+1,0}|\pi(r),m(r),\tau(r)) \xrightarrow{N\to\infty} p(Y_{n+1,0}|D_n) \qquad \text{a.s.}$$

Theorem 4. *The estimate of the posterior density of* τ, *evaluated at the fixed point* τ_0, *is strongly consistent, for almost all starting values of the algorithm. That is, for almost all starting values, given any fixed point* τ_0,

$$N^{-1}\sum_{r=1}^{N} p(\tau_0|\pi(r),m(r)) \xrightarrow{N\to\infty} p(\tau_0|D_n) \qquad \text{a.s.}$$

Theorem 5. *The estimate of the posterior density of* m, *evaluated at the fixed point* m_0, *is strongly consistent, for almost all starting values of the algorithm. That is, for almost all starting values, given any fixed point* m_0,

$$N^{-1}\sum_{r=1}^{N} p(m_0|\pi(r),\tau(r)) \xrightarrow{N\to\infty} p(m_0|D_n) \qquad \text{a.s.}$$

4. MIXTURE DECONVOLUTION

Common, and closely linked, objectives in density estimation are the assessment of the number of components of a discrete mixture, and inference about the number of modes of a population distribution. Roeder (1990), for example, nonparametric inference on the number of modes in a mixture. Various methods exists for inference about the modality of mixtures (Hartigan and Hartigan 1985; Silverman 1981; Roeder 1990), though approaches to direct inference on numbers of components are less well-developed. In our framework, prior and posterior distributions for

the number of components underlying an observed data set are readily derived, as is shown and illustrate here. Inference on modality questions is also derived.

Consider generating a sample π of size n from the model in (1), resulting in predicting an observation using the mixture (5). With knowledge of π, this mixture is the Bayesian estimate of the population distribution. The number of distinct components k from which the n realised observations arise is itself generated in the process of drawing π. The leading component of (5) allows for the observation to come from a further, distinct component. As noted earlier, the Dirichlet structure imposes a prior on k that depends only on n and α. In problems where the number of mixture components is likely to be small relative to n, say, and with moderate sample sizes, then the non-negligible prior probabilities $P(k|\alpha, n)$ do not vary dramatically with n, and decay rapidly as k increases. Table 1 illustrates this for $\alpha = 1$ and sample sizes between one and five hundred, the probabilities computed using results in Antoniak (1974). In such cases, the model may be considered as a proxy for a finite mixture model with fixed but uncertain k. The conditions for this are that α is fairly small, leading to high prior probabilities on small values of k, and that the implied prior for k, for sample size n in the problem of interest, is an acceptable representation of available prior information about the number of components. In Section 5, we assume these conditions in analysis of the astronomical data of Roeder (1990).

Table 1. Prior probabilities $P(k|\alpha = 1, n)$

	k									
n	1	2	3	4	5	6	7	8	9	10
100	0.01	0.05	0.13	0.19	0.21	0.18	0.12	0.07	0.03	0.01
200	0.01	0.03	0.08	0.15	0.19	0.19	0.15	0.10	0.06	0.03
300		0.02	0.06	0.12	0.17	0.19	0.16	0.12	0.07	0.04
400		0.02	0.05	0.11	0.16	0.18	0.17	0.13	0.09	0.05
500		0.01	0.05	0.09	0.15	0.17	0.17	0.14	0.10	0.06

In the computations of posterior and predictive distributions described in Section 3, information is generated that provides a Monte Carlo approximation to the posterior for k based on the observed data. Generating each draw $\{\pi(r), m(r), \tau(r)\}$ also leads, as a by-product, to a value of k, say $k(r)$, from the posterior for k. Thus a histogram approximation to the posterior for k is induced, and may be used to address the question about the number of components.

Issues as to numbers of modes, rather than numbers of components, will often be of secondary consideration from a practical perspective, though remain of some interest since the number of modes provides a conservative estimate, as a lower bound, of the number of components, and one that does not rely as heavily on the normal distributional assumption as the estimate of the number of components.

One particular point of interest concerns the implied prior distribution of the number of modes to be expected in predicting a future observation based on a sample of given sample size.

If this may be calculated in any specified model, then the extent to which the predicted number of modes seems to satisfactorily represent informed prior opinion provides one way of assessing the prior suitability of the model assumptions. This too is explored in the data analysis below. The model does not permit easy analytic calculation of the prior for the number of modes, however, and so we resort to simulation, as follows.

As in (1), and conditional on m and τ, we have, for $i = 2, \ldots, n$, prior distributions

$$(\pi_i|\pi_{i-1}, \ldots, \pi_1) \sim \alpha a_{i-1} G_0(\pi_i) + a_{i-1} \sum_{j=1}^{i-1} \delta_{\pi_j}(\pi_i), \tag{9}$$

with $\pi_1 \sim G_0(.)$. Thus we may trivially sample from the joint prior for (π, m, τ) by drawing m and τ from their joint prior, next generating π_1 from $G_0(.)$ given these values of m and τ, and then using (9) to sequentially sample the remaining elements of π. The density of the prior predictive distribution (5) may then be evaluated over a fine grid and the values searched to count the number of modes. Repeating this procedure provides a random sample from the prior distribution of the number of modes, a histogram estimate of the prior. Similarly, we can calculate the posterior distribution of the number of modes by simply counting and recording the number of modes of the predictive density at each sample point. A simpler version of this this strategy can be employed in cases with m and/or τ specified, and/or with constant variances $V_i = V$, known or unknown, with obvious modification. Finally, note that, whereas the parameter α alone determines the number of mixture components, it is the variance τ that predominates in determining the modality characteristics for any given k; as τ increases, smoothing decreases and the prior favours larger numbers of modes.

5. INITIAL ILLUSTRATION

Roeder (1990) describes data representing measured velocities, relative to our own galaxy, of $n = 82$ identifiable galaxies from six well-separated conic sections of space. Roeder considers the estimation of the density of velocities represented as a finite mixture of normals, and focuses on the effects of uncertainty about density estimates on the assessment of multi-modality, particularly on the hypothesis of unimodality. Of scientific interest is the hypothesis of galaxy clustering consistent with the Big Bang theory. Roeder (1990, p. 617) states, "If the galaxies are clumped, the distributions of velocities would be multimodal, each mode representing a cluster as it moves away at its own speed." However, for the purposes of assessing the scientific issue of clustering, it is appropriate to focus on the number of mixture components rather than on multi-modality. The galaxies may indeed be clustered, or clumped, into several components, but the number of modes cannot exceed the number of components, and may be much lower, the data distribution possibly exhibiting inflection points and skewness induced by distinct, though heavily overlapping, components. Related issues are raised and reviewed in Titterington *et al.* (1985, Sec. 3.3.1 and Sec. 5.5). As a cautionary note, after we calculate the posterior distribution for the number of components, there still remains the inferential leap that each normal component represents a galactic cluster.

The underlying assumption is that each galactic cluster is a normal component. If the distribution of a galactic cluster is skewed or has a very light or heavy tail, then we may use two or more normal components to fit one galactic cluster component. See Titterington *et al.* (1985, Sec. 2.2.9) for more discussion.

We detail the elements and results of a first analysis based on $\alpha = 1$, so that, however many distinct components there may be after sampling $n = 82$ cases, the probability that a further observation is drawn from a new component is $1/83$, small. The prior for k, the number of distinct components, is given in Table 2. Recall that this is determined by n and α alone, and note that this prior differs only very marginally from that with $n = 100$ in Table 1. The prior is appreciable and fairly diffuse over $3 \leq k \leq 7$, though smaller and larger values all have positive probability.

Table 2. Prior probabilities $P(k|\alpha = 1, n = 82)$

k								
1	2	3	4	5	6	7	8	9
0.01	0.06	0.14	0.21	0.21	0.17	0.11	0.06	0.02

To further assist in prior specification, consider the modality issue discussed in Section 4. Let h be the number of modes in the predictive density for a further observation based on a sample of size $n = 82$. For initial simplification, consider standardising the model to $m = 0$ (with no loss of generality) and fixing $V_i = V = 1$ for all i. Then, given a value of τ, simple modification of the discussion of Section 4 provides a way to compute the prior for the number of modes of the predictive density, based on $n = 82$. Under these assumptions, the smoothing parameter τ, critical to the modality issue, is the only unspecified quantity so that the simulation exercise may be performed for various τ to assess its effect on predictions. This simulation exercise was performed for the values of τ, appearing in Table 3. The Monte Carlo sample size in each case in $10,000$, so that the estimated prior probabilities displayed have numerical standard errors less than $(0.25/10,000)^{1/2} = 0.002$. As a cross-check on accuracy, note that the prior for k is also produced in the simulation, and in each of the cases summarised in the table, the Monte Carlo estimates of all prior probabilities for k agree with the exact values in Table 2 to two decimal places. As exemplified in the table, larger values of τ lead to increased chances on larger values of h, and the priors are more sensitive to lower τ values although, from a practical viewpoint, the differences here are small.

Table 3. Prior probabilities $P(h|\alpha = 1, n = 82, \tau)$ in standardised model

	h							
τ	1	2	3	4	5	6	7	8
25	0.09	0.38	0.38	0.13	0.02			
50	0.05	0.24	0.38	0.24	0.08	0.01		
100	0.03	0.15	0.32	0.29	0.16	0.04	0.01	
200	0.02	0.12	0.25	0.28	0.21	0.09	0.03	
300	0.02	0.10	0.23	0.28	0.21	0.11	0.04	0.01

This information helps in choosing parameters of the prior for τ, $\tau^{-1} \sim G(w/2, W/2)$. We choose a rather low value of the shape $w/2$, setting the prior degrees of freedom parameter $w = 1$. This defines a rather imprecise initial prior. Now W/w represents a prior point estimate of τ, in fact the prior harmonic mean $E(\tau^{-1})^{-1} = W/w$. The prior mode is $W/(w+2)$. With $W = 100$, the harmonic mean is 100 and the mode is 33.3, the corresponding prior density appearing as the dashed line in Figure 1(e). This is the very diffuse prior chosen for this analysis; in addition to being suitably diffuse, note the consistency with the predictive assessments of the modality issue above. In fact, analysis is possible under the traditional improper reference prior for τ, proportional to τ^{-1}; it should be noted that the conclusions reported below are essentially unchanged under analysis based on a reference prior. For the prior for the conditional variances $V_j^{-1} \sim G(s/2, S/2)$, we note Roeder (1990, p. 617), citing the original data source of Postman, Huchra and Geller (1986), states, "the error [in observed velocities] is estimated to be less than 50km per second". The uncertainty of the current authors as to the interpretation of this phrase is fairly high, indicating a small initial precision parameter s. We take $s = 4$ here. For location of the prior for the V_j, this 50km per second may variously be interpreted as an estimate of one or more standard deviations; we take it as a baseline standard deviation accounting for experimental error in velocity records. We reflect further that any identifiable cluster of galaxies may be expected to be subject to additional intra-galaxy variation in velocities. As Roeder (1990, p. 617) states, "Given the expansion scenario of the universe, points furthest from our galaxy must be moving at greater velocities". We therefore specify a prior that favours rather larger V_j values, taking $S = 2$ with the degrees of freedom $s = 4$; recall that the $V_j^{1/2}$ are in units of thousands of kilometers per second. The corresponding 95% equal tails interval for each of the $V_j^{1/2}$ is roughly 400-2000km per second.

Before proceeding to analysis, we explore the implied prior for h under this specification and with the prior mean taken (without loss of generality) as $m = E(Y_i) = 20$. Prior simulations described in Section 4 can be carried out in full, sampling the joint prior of $\{\pi, \tau\}$ and hence the prior for the number of modes h. Again based on $10,000$ replications, the prior for h under this model appears in Table 4. Note that, relative to the various cases in Table 3, the prior is rather more diffuse due to the additional uncertainty about τ. Again these probabilities are quoted to two decimal places, being positive, though rapidly decreasing, for smaller and larger values of h.

Table 4. Prior probabilities $P(h|\alpha = 1, n = 82)$ under chosen priors

h							
1	2	3	4	5	6	7	8
0.03	0.13	0.26	0.26	0.18	0.09	0.04	0.01

Posterior and predictive analysis is detailed in this framework, with the additional, and final, assumption of a diffuse or reference prior for m, taking the limiting form of the $N(a, A)$ prior with $A^{-1} \to 0$ as a reference. Clearly other, informative priors might be used here instead. Analysis is based on the techniques in Section 3 with Monte Carlo sample size $N = 10,000$, and the number of iterations used for 'burn-in' to convergence set at 2000. These values are supported through experimentation with different starting values that suggest that an initial 1000 iterations are more than adequate to achieve stability in the estimated posterior distributions. Further analyses with varying sample sizes lead to substantially similar inferences – and see section 6, below. In fact, the 10000 draws used for inference here are based on an actual run of 150 times that number, saving draws only each 150 iterations; this induces posterior samples such that consecutive values in each of the k and h series have negligible autocorrelations. This rather wasteful analysis is reported in order than the approximate posterior probabilities reported in Table 5 be acceptably accurate in the second decimal place; assuming exactly independent draws, 0.01 represents an upper bound on two posterior standard deviations for each of the posterior probabilities reported. Analyses were coded in Risc Fortran running on Ultrix DECstations, and use standard numerical algorithms for random variate generation (e.g. ran1, gasdev, gamdev as in Press *et al* 1992).

The substance of the scientific issue of galaxy clustering is addressed through the posteriors for k and h. The corresponding Monte Carlo approximations from this analysis are given in Table 5. The prior for k (in Table 2) provides heavy support for between 3 and 7 clusters, whilst being reasonably diffuse over a wider range. The posterior supports rather larger values. Since the prior is centered around lower values than the posterior distribution, the likelihood function puts most of its weight on large values of k. Therefore, alternative priors giving more support to larger values of k would produce posteriors shifted upward. As is typical with inference about overlapping mixtures, there is clearly a great deal of uncertainty about the number of components. However, unlike traditional approaches to density estimation, the computations here provide a formal assessment of such uncertainty. The posterior for h heavily favours 5 modes, evident in Figure 1(a). A crude summary of Table 5 would conclude that there is strong support for between 5 and 9 components.

Table 5. Posterior probabilities $P(k|\alpha = 1, D_n)$ and $P(h|\alpha = 1, D_n)$ for Roeder's data

	i										
	1	2	3	4	5	6	7	8	9	10	11
$P(k = i\|D_n)$				0.03	0.11	0.22	0.26	0.20	0.11	0.05	0.02
$P(h = i\|D_n)$			0.04	0.14	0.49	0.29	0.04				

Varying α and repeating the analysis provides insight into just how sensitive the results for k are to α. The sensitivity is marked. For example, repeat analyses with α increasing from 0.5 through $\alpha = 1.0$ to $\alpha = 2.0$ correspondingly shifts the posterior for k from smaller to somewhat larger values, though differences in predictive distribution functions are undetectable. The effects of varying α on inferences about τ are, predictably, that smaller α values shift the posterior for τ to favouring higher values; the effects are not great, however, due to the marked lack of information about τ in such analyses with small numbers of observations. Rather than pursue such sensitivity analyses further, we defer to the next section and formally subsume sensitivity studies in extended analyses incorporating learning for α.

6. LEARNING ABOUT α AND FURTHER ILLUSTRATION

Central to this analysis is the precision parameter α of the underlying Dirichlet process – a critical smoothing parameter for the model. Learning about α from the data may be addressed with a view to incorporating α into the Gibbs sampling analysis. Assume a continuous prior density $p(\alpha)$ (which may depend on the sample size n), hence an implied prior $P(k|n) = E[P(k|\alpha, n)]$ where, using results in Antoniak (1974),

$$P(k|\alpha, n) = c_n(k) n! \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \qquad (k = 1, 2, \ldots, n), \tag{10}$$

and $c_n(k) = P(k|\alpha = 1, n)$, not involving α. If required, the factors $c_n(k)$ are easily computed using recurrence formulæ for Stirling numbers (further details available on request from the second author). This is important, for example, in considering the implications for priors over k of specific choices of priors for α (and *vice-versa*) in the initial prior elicitation process. As an aside, note that there is a great deal of flexibility in representing prior opinions about k through choices of prior for α – this will be elaborated and explored elsewhere is greater generality.

Now suppose we have sampled values of the parameters π_i. By sampling the parameters π_i, we have in fact sampled a value for k, the number of distinct components, and have also sampled a specific configuration of the data D_n into k groups. From our model, the data are initially conditionally independent of α when k, π and the configuration are known, and the parameters π are also conditionally independent of α when k and the configuration are known. We deduce

$$p(\alpha|k, \pi, D_n) = p(\alpha|k) \propto p(\alpha) P(k|\alpha), \tag{11}$$

with likelihood function given in (10) (the sample size n should appear in conditioning, of course, but is omitted for clarity of notation.) Thus the Gibbs sampling analysis can be extended; for given α, we sample parameters π, and hence k, as usual from the conditional posterior $p(\pi|\alpha, D_n)$. Then, at each iteration, we can include α in the analysis by sampling from the conditional posterior (11) based on the previously sampled value of k – no other information is needed. Sampling from (11) may involve using a rejection, or other, method depending on the form of the prior $p(\alpha)$. Alternatively, we may discretise the range of α so that (11) provides a discrete approximation to the posteriors – the so-called 'griddy Gibbs' approach (Ritter and Tanner 1991). More attractively, sampling from the exact, continuous posterior (11) is possible in the Gibbs iterations when the prior $p(\alpha)$ comes from the class of mixtures of gamma distributions. We develop the results here for a single gamma prior, and leave generalisations to mixtures to the reader or refer to West (1992b).

Suppose $\alpha \sim G(a, b)$, a gamma prior with shape $a > 0$ and scale $b > 0$ (which we may extend to include a 'reference' prior (uniform for $log(\alpha)$) by letting $a \to 0$ and $b \to 0$.) In this case, (11) may be expressed as a mixture of two gamma posteriors, and the conditional distribution of the mixing parameter given α and k (and, of course, n) is a simple beta. See this as follows. For $\alpha > 0$, the gamma functions in (10) can be written as

$$\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} = \frac{(\alpha+n)\beta(\alpha+1,n)}{\alpha\Gamma(n)}, \tag{12}$$

where $\beta(.,.)$ is the usual beta function. Then, in (11), and for any $k = 1, 2, \ldots, n$,

$$p(\alpha|k) \propto p(\alpha)\alpha^{k-1}(\alpha+n)\beta(\alpha+1,n) \quad \propto \quad p(\alpha)\alpha^{k-1}(\alpha+n)\int_0^1 x^{\alpha}(1-x)^{n-1}dx,$$

using the definition of the beta function. This implies that $p(\alpha|k)$ is the marginal distribution from a joint for α and a continuous quantity η such that $p(\alpha, \eta|k) \propto p(\alpha)\alpha^{k-1}(\alpha+n)\eta^{\alpha}(1-\eta)^{n-1}$, for $0 < \alpha$ and $0 < \eta < 1$. Hence we have conditional posteriors $p(\alpha|\eta, k)$ and $p(\eta|\alpha, k)$ determined as follows. Firstly, under the $G(a, b)$ prior for α,

$$p(\alpha|\eta,k) \propto \alpha^{a+k-2}(\alpha+n)e^{-\alpha(b-log(\eta))} \quad \propto \quad \alpha^{a+k-1}e^{-\alpha(b-log(\eta))} + n\alpha^{a+k-2}e^{-\alpha(b-log(\eta))}$$

for $\alpha > 0$, which reduces easily to a mixture of two gamma densities, viz

$$(\alpha|\eta,k) \sim \pi_\eta G(a+k, b-log(\eta)) + (1-\pi_\eta)G(a+k-1, b-log(\eta)) \tag{13}$$

with weights π_η defined by $\pi_\eta/(1-\pi_\eta) = (a+k-1)/\{n(b-log(\eta))\}$. Note that these distributions are well defined for all gamma priors, all η in the unit interval and all $k > 1$. Secondly,

$$p(\eta|\alpha,k) \propto \eta^{\alpha}(1-\eta)^{n-1} \qquad (0 < \eta < 1) \tag{14}$$

so that $(\eta|\alpha, k) \sim B(\alpha+1, n)$, a beta distribution with mean $(\alpha+1)/(\alpha+n+1)$.

It is now clear how α can be sampled at each stage of the simulation — at each Gibbs iteration, the currently sampled values of k and α allow us to draw a new value of α by (i) first

sampling an η value from the simple beta distribution (14), conditional on α and k fixed at their most recent values; then (ii) sampling the new α value from the mixture (13) based on the same k and the η value just generated in (i). On completion of the simulation, $p(\alpha|D_n)$ will be estimated by the usual Monte Carlo average of conditional forms (13), viz $p(\alpha|D_n) \approx N^{-1} \sum_{s=1}^{N} p(\alpha|\eta_s, k_s)$ where η_s are the sampled values of η.

One could develop convergence theorems for this new algorithm. The proofs would be straightforward extensions of our results above. For example, to prove new versions of our Theorems 3, 4 and 5, one could bound the expected posterior distributions with the bounds proven in the appendix, since these bounds were constants with respect to α.

We reanalyse the astronomical velocities data with the gamma prior $\alpha \sim G(2, 4)$; this density appears as the dashed line in Figure 1(f). Note that there is a fair degree of support for values near the $\alpha = 1$ used in the previous section. All other assumptions and details of the analysis are as in the previous section. Analysis is summarised graphically in Figure 1. Figure 1(a) displays a histogram of the data, from Table 1 of Roeder (1990), together with a graph of the estimated predictive density function from equation (8). This latter density is very similar to the 'optimal' density estimate of Roeder (1990, Figure 7), but it has five modes rather than her four. To give a qualitative indication of uncertainty, Figures 1(b) display graphs of a random selection of just 100 of the 10,000 sampled predictive densities; the summands of (8). Plots of the corresponding cumulative distribution functions appears in Figures 1(c) and 1(d). A nice way to exhibit uncertainties about density and distribution functions is via 'live' animated graphical display of sequentially sampled functions (Tierney 1991). Restricted to static plots, we prefer displaying sampled curves to bands mapping pointwise interval estimates of the functions since the latter do not define density or distribution functions.

The results summarised here attest to the robustness to α values noted in the previous section in so far as the issues of predictive density estimation is concerned. The predictive distributions and density functions are substantially similar to those obtained under the various analyses noted with α fixed. The estimated posterior $p(\tau|D_n)$ appears as the full line in Figure 1(e), together with the prior $p(\tau)$ (the latter is quite diffuse, having a long tail off to the right of the plotted region). Finally, Figure 1(f) presents the corresponding prior and posterior densities for α. These are typical pictures; the information available in such small datasets about the smoothing parameters τ and α is typically very limited, and this relates to the difficulties of smoothing parameter estimation in traditional approaches.

Again addressing the substantive issue of galaxy clustering through inference about k and h, Table 6 provides the posteriors for k and h now accounting for estimation of α. These are very similar to those at $\alpha = 1$ in Table 5, though rather more diffuse over larger values of k and supporting between 5 and 9 components. There is marked residual uncertainty about the number of components. Inferences about the number of modes h are also rather similar to those based on Table 5.

Table 6. Posterior probabilities $P(k|D_n)$ and $P(h|D_n)$ under $\alpha \sim G(2,4)$

	i											
	1	2	3	4	5	6	7	8	9	10	11	12
$P(k=i\|D_n)$			0.02	0.05	0.14	0.21	0.21	0.16	0.11	0.06	0.03	0.01
$P(h=i\|D_n)$			0.07	0.15	0.47	0.27	0.04					

7. Further Comments

We have described and illustrated Bayesian density estimation and mixture deconvolution in classes of models whose analyses are now routinely implementable using stochastic simulation methods. The key contributions here lie in the development of computational techniques for hierarchical mixture models; though these models have been known for about twenty years, only now can their real utility be realised. Problems of hierarchical prior specification and inference about layers of parameters and hyperparameters – particularly the variance and precision parameters that control and define degrees of local smoothing – have been addressed and may now be (and currently, in various application areas, are) incorporated in routine data analyses using these models. In addition to developing methodology and demonstrating its utility in density estimation and mixture identification, we provide theoretical results proving convergence of the implemented simulation schemes, showing that they provide consistent numerical approximations to the exact Bayesian posterior and predictive distributions of interest.

Current and recent areas of active research on extensions to this paper include generalizations to more elaborate multivariate linear models, and non-linear models. Since this paper was originally written (in 1991) there has been resulting research on refining the basic computational methods; in particular, MacEachern (in press) has introduced important algorithms that improve convergence characteristics; see also West, Müller, and Escobar (1994).

Appendix: Convergence Issues

In the proofs below, it is helpful to use the notion of a configuration defined in West (1990) as follows:

DEFINITION: For each integer k, $1 \leq k \leq n$, let $c = (c_1, \ldots, c_n)$ be any integer $n-$vector whose elements take values between 1 and k, each such value appearing at least once. Define $C_k(c)$ as the configuration of the n elements $\{\pi_i\}$ into exactly k distinct values, $\pi_1^*, \ldots, \pi_k^*$, with $\pi_i = \pi_j^*$ where $c_i = j, i = 1, \ldots, n$. Then $C_k(c)$ is called a $k-$configuration of the $\{\pi_i\}$. Finally, let n_j be the number of the $\{\pi_i\}$ equal to π_j^*, given by $n_j = \#\{c_i = j; i = 1, \ldots, k\}$.

In the above definition, please note that $\pi_i = \pi_j^*$ implies $\mu_i = \mu_j^*$ and $V_i = V_j^*$. With this definition of configurations, the posterior distribution produced from Algorithm I where m and τ are fixed can be written as:

$$P(\pi|D_n, \tau, m) = \sum_{\{C_k(c)\}} P(\pi|D_n, C_k(c), \tau, m) P(C_k(c)|D_n, \tau, m)$$

where the sum is over all the unique configurations and where s and S, the parameters for the prior distribution, are surpressed in the above notation. The posterior distribution for Algorithm II is

$$P(\pi, m, \tau|D_n) = \sum_{\{C_k(c)\}} P(\pi, m, \tau|D_n, C_k(c)) P(C_k(c)|D_n).$$

Again, the parameters for the prior distribution, s, S, a, A, w, and W, are surpressed in the above notion.

Each configuration, $C_k(c)$, is associated with $2k$-dimensional subspace on $\bigotimes_{i=1}^{n}(\Re \times \Re^+)_i$. For example, the configuration $C_1(1, \ldots, 1)$ is associated with the 2 dimensional subspace $\{\pi|\pi_i = \pi_j, \forall i = j\}$. Define $\lambda_{C_k(c)}$ as the lebesgue measure on the subspace associated with $C_k(c)$, and let $\Lambda = \sum_{\{C_k(c)\}} \lambda_{C_k(c)}$. Given the configuration, the posterior and predictive distributions behave like standard hierarchical normal models. Therefore, for example, $P(\pi|D_n, C_k(c))$ and $\lambda_{C_k(c)}$ are mutually absolutely continuous. Therefore, the posterior distribution of π and Λ are mutually absolutely continuous.

Conditioning on the configurations, the model in this paper reduces to the standard normal/inverse gamma hierarchical model. Since the number of configurations is finite, then proofs for the consistency of the Markov chain Monte Carlo estimates of the posterior and predictive densities would be a simple extension of the proofs for the standard normal/inverse gamma hierarchical model. However, the present authors do not know of any such published proofs for the consistency of these estimates for this standard model. Also, since the standard normal/inverse gamma hierarchical model is our model at a fixed configuration, the convergence of the density estimates for the standard model are corollaries to the theorems in this paper.

Proof of Theorem 1 and 2. The arguments for the proof of both theorems are identical, so we will only formally argue the proof of theorem 1. From theorem 1 of Tierney (in press), we need to show that posterior distribution is an invariant distribution for the Markov chain defined by the algorithm and that the Markov chain is aperiodic and irreducible with respect to the posterior distribution. A proof of the invariance of the posterior distribution is similar to the proof of invariance contained in theorem 2 in Escobar (1994). From the construction of the Markov chain, we can see that for any set A such that $\Lambda(A) > 0$ then for all starting points $\pi(0)$, $Q_I(\pi(0), A) > 0$. So Q_I is Λ-irreducible and by mutually absolute continuity Q_I is also irreducible with respect to the posterior distribution. Also, since $\Lambda(A) > 0$ implies $Q(\pi(0), A) > 0$, then Q_I is aperiodic. $\diamondsuit$

Proof of Theorem 3. In order to show convergence of path averages, we use theorem 3 of Tierney (in press) which requires that the transition kernel of the Markov chain converge (in total

variation norm) to the posterior distribution, that the chain be Harris recurrent, and that the posterior expectations of $p(Y_{n+1}|\pi(r), m(r), \tau(r))$ be bounded and equal to $p(Y_{n+1}|D_n)$. By our theorem 2, we know that the Markov chain converges. It is straightforward to show that the path averages have the right expectation.

We will make our recurrent Markov chain Harris recurrent by throwing away a set of starting values which have measure zero under the posterior distribution. See this as follows. First of all, we know that our Markov chain is positive recurrent by Theorem 1 of Tierney (in press). Theorems 9.0.1 and 9.1.5 of Meyn and Tweedie (1993) state that the state space can be divided into two disjoint sets H and T, where the set T is a transient, null set, and where the set H is an absorbing set with the property that our Markov chain restricted to this set is Harris recurrent. Therefore, if we do not use starting values in T, then we start our chain in the absorbing set H and in this state space our recurrent chain is Harris recurrent. For a fuller discussion of this argument, please see the discussion around Theorems 9.0.1 and 9.1.5 in Meyn and Tweedie (1993).

Finally, what we need to show is that the posterior expectations are finite. From (4) and (6), let $f_t(y; s, m, M)$ be the density function of a t distribution with s degrees of freedom, mode m, and scale factor $M^{1/2}$ evaluated at the value y, where $M = (1+\tau)S/s$. The function $f_t(y; s, m, M)$ is bounded by $B_{n,s,S}$ for all values of y, where $B_{n,s,S} = S^{-1/2}\Gamma[(1+s+n)]/\{\Gamma[(s+n)/2]\Gamma(1/2)\}$. This is because,

$$\begin{aligned}
f_t(y; s, m, M) &= (sM)^{-1/2} \frac{\Gamma[(1+s)/2]}{\Gamma(s/2)\Gamma(1/2)} \Big[1 + \frac{(y-m)^2}{sM}\Big]^{-(1+s)/2} \\
&\leq (sM)^{-1/2} \frac{\Gamma[(1+s)/2]}{\Gamma(s/2)\Gamma(1/2)} \\
&= \big[(1+\tau)S\big]^{-1/2} \frac{\Gamma[(1+s)/2]}{\Gamma(s/2)\Gamma(1/2)} \\
&\leq S^{-1/2} \frac{\Gamma[(1+s+n)/2]}{\Gamma[(s+n)/2]\Gamma(1/2)} \\
&= B_{n,s,S}.
\end{aligned}$$

From (4) and (6), let $f_N(y; s, m, M)$ be the density function of a normal distribution with mean μ_j^* and variance V_j^* evaluated at the value y. As a function of y, the posterior expectation of $f_N(y; \mu_j^*, V_j^*)$ is also bounded by $B_{n,s,S}$ for all values of y. To see this, first note that $f_N(y; \mu_j^*, V_j^*) \leq (2V_j^*)^{-1/2}/\Gamma(1/2)$. Now note that for a fixed configuration, the posterior distribution of $(V_j^*)^{-1}$ is a gamma distribution with shape parameter $(s+n_j)/2$ and scale parameter $(S+N_j)/2$, where $N_j = \sum_{l=1}^{n_j}(Y_{jl}^* - \bar{Y}_j^*)^2$, Y_{jl}^* is the l-th value of Y_i such that $c_i = j$, and $\bar{Y}_j^* = n_j^{-1}\sum_{l=1}^{n_j} Y_{jl}^*$. Therefore,

$$\begin{aligned}
P(f_N(y; \mu_j^*, V_j^*)|C_k(c), m, \tau, D_n) &\leq P((2V_j^*)^{-1/2}/\Gamma(1/2)|C_k(c), m, \tau, D_n) \\
&= (S+N_j)^{-1/2} \frac{\Gamma[(s+n_j+1)/2]}{\Gamma[(s+n_j)/2]\Gamma(1/2)} \\
&\leq S^{-1/2} \frac{\Gamma[(1+s+n)/2]}{\Gamma[(s+n)/2]\Gamma(1/2)} \\
&= B_{n,s,S}.
\end{aligned}$$

Let $f_p(y)$ be the density of the random variable $(Y_{n+1}|\pi)$ defined in equations (4) and (5). With the above bounds on f_N and f_t, we now show that the posterior estimation of $f_p(y)$ is bounded by the finite constant $B_{n,s,S}$. For all y,

$$
\begin{aligned}
&P(f_p(y)|D_n) \\
&\quad = \int\int f(y)dP(\pi|C_k(c),m,\tau,D_n)dP(C_k(c),m,\tau|D_n) \\
&\quad = \int\int \Big\{\alpha a_n f_t(y;s,m,M) + a_n \sum_{j=1}^{k} n_j f_N(y;\mu_j^*,V_j^*)\Big\}dP(\pi|C_k(c),m,\tau,D_n)dP(C_k(c),m,\tau|D_n) \\
&\quad \le \int \Big\{\alpha a_n B_{n,s,S} + a_n \sum_{j=1}^{k} n_j \int f_N(y;\mu_j^*,V_j^*)dP(\pi|C_k(c),m,\tau,D_n)\Big\}dP(C_k(c),m,\tau|D_n) \\
&\quad \le \int \Big\{B_{n,s,S}\Big\}dP(C_k(c),m,\tau|D_n) \\
&\quad \le B_{n,s,S}.
\end{aligned}
$$

Therefore, the posterior expectation has a finite bound and by Theorem 3 of Tierney (in press), the proof is complete. $\diamondsuit$

Proof of Theorem 4. The proof is similar to the proof of our Theorem 3. After proving Theorem 3, what remains to be shown is that the posterior expectation is bounded. From the comments above the statement to Algorithm II, $(\tau^{-1}|m,\pi) \sim G((w+k)/2,(W+K)/2))$ where $K = \sum_{j=1}^{k}(\mu_j^* - m)^2/V_j^*$, and where w and W are fixed positive parameters of the prior distribution for τ. Let the density function of τ at the value τ_0 be $p(\tau_0|m,\pi)$. The density is maximized when $\tau_0 = (W+K)/(w+k+2)$. Now, by evaluating $p(\tau_0|m,\pi)$ at its mode to get the first inequality below and by using Stirling's formula for the gamma function (Johnson and Kotz, 1969, p. 6) to get the second inequality, we have for all τ_0

$$
\begin{aligned}
p(\tau_0|m,\pi) &\le \Big[\frac{w+k+2}{2}\Big]^{(w+k+2)/2} \ \frac{2\exp[-(k+w)/2]}{\Gamma[(w+k)/2]} \\
&< \frac{e\,(w+k)\sqrt{w+k+2}}{2\Gamma(1/2)} \\
&\le \frac{e\,(w+n)\sqrt{w+n+2}}{2\Gamma(1/2)},
\end{aligned}
$$

where $e = \exp(1)$. Since $p(\tau_0|m,\pi)$ is bounded by a constant for all values of m and π, then the posterior expectation, $P(p(\tau_0|m,\pi)|D_n)$, is also bounded and the theorem is proven. $\diamondsuit$

Proof of Theorem 5. Again, the proof is similar to the proof of Theorems 3 and 4, and what remains to be shown is that the posterior expectation is bounded. From the comments above the statement of Algorithm II, given τ and π, the random variable m is normally distributed with variance, $V(m|\tau,\pi) = \tau A\bar{V}/(A+\tau\bar{V})$, where $\bar{V}^{-1} = \sum_{j=1}^{k}(V_j^*)^{-1}$. Let the density function

of m at the value m_0 be $p(m_0|\tau,\pi)$. For all values of m_0, the density is bounded by $((A+\tau\bar{V})/(2\tau A\bar{V}))^{1/2}/\Gamma(1/2)$. Also, for all values of m, $P((\tau)^{-1/2}|m,\pi,C_k(c)) \leq \sqrt{2}\ \Gamma((w+k+1)/2)/(\Gamma((w+k)/2)\sqrt{W})$. Now define $(U_j)^{-1} = ((S+N_j)/S)(V_j^*)^{-1}$, and $(\bar{U}_j)^{-1} = \sum_{j=1}^{k}(U_j)^{-1}$. Then, $(\bar{V})^{-1} \leq (\bar{U})^{-1}$, and given $C_k(c)$ and D_n, $(\bar{U})^{-1} \sim G(k(s+1)/2, S/2)$. Finally, define the constant B^* as:

$$B^* = \Big(\frac{2\ \Gamma((w+n+1)/2)\Gamma((ns+n+1)/2)}{\Gamma((w+n)/2)\Gamma(n(s+1)/2)\sqrt{WS}} + A^{-1/2}\Big)\frac{1}{\sqrt{2}\ \Gamma(1/2)}.$$

Therefore, for all m_0:

$$\begin{aligned}
&P(p(m_0|\tau,\pi)|D_n) \\
&\quad = \int\int\int p(m_0|\tau,\pi)\, dP(\tau|\pi,C_k(c),D_n)\, dP(\pi|C_k(c),D_n)\, dP(C_k(c)|D_n) \\
&\quad \leq \int\int\int \Big(\frac{A+\tau\bar{V}}{2\tau A\bar{V}}\Big)^{1/2} \frac{1}{\Gamma(1/2)}\, dP(\tau|\pi,C_k(c),D_n)\, dP(\pi|C_k(c),D_n)\, dP(C_k(c)|D_n) \\
&\quad \leq \int\int\int \Big((\tau\bar{V})^{-1/2}+(A)^{-1/2}\Big)\frac{1}{\sqrt{2}\ \Gamma(1/2)}\, dP(\tau|\pi,C_k(c),D_n)\, dP(\pi|C_k(c),D_n)\, dP(C_k(c)|D_n) \\
&\quad \leq \int\int \Big(\frac{\sqrt{2}\ \Gamma((w+k+1)/2)}{\Gamma((w+k)/2)W^{1/2}(\bar{V})^{1/2}} + A^{-1/2}\Big)\frac{1}{\sqrt{2}\ \Gamma(1/2)}\, dP(\pi|C_k(c),D_n)\, dP(C_k(c)|D_n) \\
&\quad \leq \int\int \Big(\frac{\sqrt{2}\ \Gamma((w+k+1)/2)}{\Gamma((w+k)/2)W^{1/2}(\bar{U})^{1/2}} + A^{-1/2}\Big)\frac{1}{\sqrt{2}\ \Gamma(1/2)}\, dP(\pi|C_k(c),D_n)\, dP(C_k(c)|D_n) \\
&\quad \leq \int \Big(\frac{2\ \Gamma((w+k+1)/2)\Gamma((ks+k+1)/2)}{\Gamma((w+k)/2)\Gamma(k(s+1)/2)\sqrt{WS}} + A^{-1/2}\Big)\frac{1}{\sqrt{2}\ \Gamma(1/2)}\, dP(C_k(c)|D_n) \\
&\quad \leq B^*.
\end{aligned}$$

Therefore, the posterior expectation is bounded and the theorem is proven. $\diamondsuit$

REFERENCES

Antoniak, C.E. (1974), "Mixtures of Dirichlet Processes With Applications to Nonparametric Problems," *The Annals of Statistics*, 2, 1152-1174.

Besag, J. and Green, P.J. (1993) "Spatial Statistics and Bayesian Computation," *Journal of the Royal Statistical Society,* Ser. B, 55, 25-37.

Escobar, M.D. (1988) "Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means," unpublished Ph.D. dissertation, Yale University, Dept. of Statistics.

— (1992) Comment on "Bayesian Analysis of Mixtures: Some Results on Exact Estimability and Identification," by Florens, J.-P., Mouchart, M., and Rolin, J.-M., in *Bayesian Statistics 4*, eds: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford, U.K.: Oxford University Press, pp. 142-144.

— (1994) "Estimating Normal Means With a Dirichlet Process Prior," *Journal of the American Statistical Association,* 89, 268-277.

— (in press), "Nonparametric Bayesian Methods in Hierarchical Models," *The Journal of Statistical Inference and Planning.*

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications,* (Vol. 2, 2nd ed.), New York: John Wiley and Sons.

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

— (1983), "Bayesian Density Estimation by Mixtures of Normal Distributions," in *Recent Advances in Statistics*, eds. H. Rizvi and J. Rustagi, New York: Academic Press, pp. 287-302.

Gelfand, A.E., and Smith, A.F.M. (1990), "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association,* 85, 398-409.

Hartigan, J.A., and Hartigan, P.M. (1985), "The Dip Test of Unimodality," *The Annals of Statistics,* 13, 17-84.

Hasting, W.K. (1970)"Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika,* 57, 97-109.

Kuo, L. (1986), "Computations of Mixtures of Dirichlet Processes," *SIAM Journal of Scientific and Statistical Computing,* 7, 60-71.

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

Lo, A.Y. (1984), "On a Class of Bayesian Nonparametric Estimates: 1. Density Estimates," *The Annals of Statistics*, 12, 351-357.

MacEachern, S.N. (in press), "Estimating Normal Means With a Conjugate Style Dirichlet Process Prior," *Communication in Statistics.*

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., and Teller, A.H., and Teller, E. (1953) "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics,* 21, 1087-1091.

Meyn, S.P., and Tweedie, R.L. (1993) *Markov Chains and Stochastic Stability,* New York: Springer-Verlag.

Nummelin, E., (1984) *General Irreducible Markov Chains and Non-negative Operators,* Cambridge: Cambridge University Press.

Postman, M., Huchra, J.P., and Geller, M.J. (1986), "Probes of Large-Scale Structure in the Corona Borealis Region," *The Astronomical Journal*, 92, 1238-1247.

Press, W.H., Teukolsky, S.A. Vettering, W.H. and Flannery, B.P. (1992) *Numerical Recipes in Fortran,* (2nd Edition), Cambridge: Cambridge University Press.

Ritter, C. and Tanner, M.A. (1991) "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy Gibbs Sampler," *Journal of the American Statistical Association,* 87, 861-868.

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

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

— (1986), *Density Estimation for Statistics and Data Analysis*, New York: Chapman and Hall.

Smith, A.F.M., and Roberts, G.O. (1993) "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society,* Ser. B, 55, 3-23.

Tanner, M.A. and Wong, W.H. (1987) "The Calculation of Posterior Distributions by Data Augmentation (With Discussion)," *Journal of the American Statistical Association,* 82, 528-550.

Tierney, L. (1991) "Exploring Posterior Distributions using Markov Chains," in *Computer Science and Statistics: 23rd Symposium on the Interface,* ed: E.M. Keramidas, pp 563-570.

Tierney, L. (in press) "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics.*

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Chichester: John Wiley and Sons.

West, M. (1990), "Bayesian Kernel Density Estimation," Discussion Paper #90-A02, Institute of Statistics and Decision Sciences, Duke University.

— (1992a) "Modelling With Mixtures" (With Discussion), in *Bayesian Statistics 4*, eds: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford: Oxford University Press, pp. 503-524.

— (1992b) "Hyperparameter estimation in Dirichlet process mixture models," *ISDS Discussion Paper #92-A03*, Duke University.

—, and Cao, G. (1993) "Assessing Mechanisms of Neural Synaptic Activity," in *Bayesian Statistics in Science and Technology: Case Studies,* eds: C. Gatsonis, J. Hodges, R. Kass, N. Singpurwalla, New York: Springer-Verlag, pp. 416-428.

—, Müller, P., Escobar, M.D., (1994) "Hierarchical Priors and Mixture Models, With Applications in Regression and Density Estimation," in *Aspects of Uncertainty: A Tribute to D.V. Lindley*, eds: A.F.M. Smith and P. Freeman.