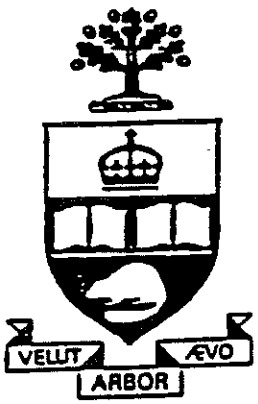

A note on convergence rates of Gibbs sampling for nonparametric mixtures*

by

Sonia Petrone
Dipartimento di Economia Politica
e Metodi Quantitativi
Università di Pavia

and

Gareth O. Roberts
Statistical Laboratory
University of Cambridge

and

Jeffrey S. Rosenthal
Department of Statistics
University of Toronto

Technical Report No. 9809, May 8, (1998)

A note on convergence rates of Gibbs sampling for nonparametric mixtures*

Sonia Petrone† Gareth O. Roberts‡ and Jeffrey S. Rosenthal§
Università di Pavia, University of Cambridge and University of Toronto

May 8, 1998

1 Introduction.

In many statistical problems it seems appropriate to specify the distribution of the data as a mixture of parametric densities, i.e. to assume that the data X_i are independent and identically distributed, conditionally to a distribution function G, with density $f(x \mid G) = \int f(x \mid \theta)\, dG(\theta)$. The mixing distribution G is unknown and, in a Bayesian nonparametric analysis, it is considered as a random distribution function which is usually given a Dirichlet process prior. Problems of mixtures arise in a great variety of applications, such as empirical Bayes problems (e.g. Berry and Christensen, 1978; Liu, 1996) or nonparametric hierarchical models (e.g. Ferguson, 1983; Lo, 1984; Escobar and West, 1995; Petrone, 1996).

Analytical computations in these models are difficult, and approximations have been suggested (Kuo, 1986, Newton and Zhang, 1996, Liu, 1996). In this paper we consider two Gibbs sampling algorithms. These have been proposed by Escobar (1994) and MacEachern (1994) for mixtures of normals and for ANOVA models. We first outline (section 2) the basic structure of the Markov chains resulting from the Gibbs samplers, for a general mixture model with a Dirichlet process mixing distribution. The Markov chains show interesting properties and are difficult to analyze. However, some knowledge about rates of convergence (cf. Rosenthal, 1995b; Roberts and Rosenthal, 1997) of Gibbs samplers in various contexts are of benefit to understanding these algorithms more deeply.

Key words and phrases: Dirichlet process, Mixture models, Markov chain Monte Carlo. *AMS Subject Classification*: primary 62G99, secondary 60J05. We thank Michael Escobar for a helpful discussion, and thank Deborah Tate and Marco Valvassori for their kind hospitality.

†Dipartimento di Economia Politica e Metodi Quantitativi, Università degli Studi di Pavia, Via San Felice, 5, I-27100 Pavia, Italia. Internet: spetrone@eco.unipv.it. Supported in part by TMR of the E.U.

‡Statistical Laboratory, University of Cambridge, Cambridge CB2 1SB, U.K. Internet: G.O.Roberts@statslab.cam.ac.uk.

§Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Internet: jeff@utstat.toronto.edu. Supported in part by NSERC of Canada.

In this paper, we prove (Section 3) that if the kernel $f(x \mid \theta)$ is bounded, then the Markov chains are uniformly ergodic, and we provide an explicit rate bound. The bound is not sharp, and is too big for providing useful indications about the number of iterations required in practical use of the Gibbs sampling. Improving sensibly the quantitative bound seems however quite difficult, except in a special case (Theorem 2). We also discuss some examples (Section 4).

2 The Markov chains.

Consider the following model. The data $X_1, \ldots, X_n$ are conditionally independent, given $\theta_1, \ldots, \theta_n$, with $X_i \mid \theta_1, \ldots, \theta_n \sim f(x_i \mid \theta_i)$. Conditionally on G, the θ_i are independent and identically distributed with distribution function G, where $G \sim \mathcal{D}(MG_0)$, i.e. G is a Dirichlet process with parameter MG_0, where M is a fixed constant (called the *scale parameter*) and G_0 is a probability distribution function (d.f.) on the parameter space Θ. For simplicity, we assume that G_0 is absolutely continuous with respect to (one-dimensional) Lebesgue measure, with density g_0.

The distribution of interest is the posterior distribution $\pi(d\theta_1, \ldots, d\theta_n \mid x_1, \ldots, x_n)$. This distribution gives positive probability to ties in $(\theta_1, \ldots, \theta_n)$. Let

$$\mathcal{C}(D, n_1, \ldots, n_D) = \{(\theta_1, \ldots, \theta_n) \in \Theta^n : \text{ there are } D \text{ distinct values; one of them}$$

$$\text{is repeated } n_1 \text{ times, one is repeated } n_2 \text{ times, ..., one is repeated } n_D \text{ times}\}$$

where $(n_1, \ldots, n_D)$ is a sequence of integers such that $n_i > 0$, $n_1 \leq n_2 \leq \cdots \leq n_D$ and $\sum_{i=1}^{D} n_i = n$. ¿From results in Antoniak (1974) it can be shown (see Petrone and Raftery, 1997) that

$$\pi(d\theta_1, \ldots, d\theta_n \mid x_1, \ldots, x_n) \propto$$

$$\sum_{D=1}^{n} \sum_{(n_1,\ldots,n_D)} \frac{M^D}{M^{(n)}} \prod_{i=1}^{D} (n_i - 1)! \prod_{i=1}^{D} g_0(\theta_i') \prod_{i=1}^{n} f(x_i \mid \theta_i) \; I_{\mathcal{C}(D,n_1,\ldots,n_D)}(\theta_1, \ldots, \theta_n) \lambda_D(d\theta_1', \ldots, d\theta_D'),$$

where $M^{(n)} = M(M+1)\cdots(M+n-1)$; $M^{(0)} = 1$; $(\theta_1', \ldots, \theta_D')$ is the vector of distinct values among $(\theta_1, \ldots, \theta_n)$, with θ_i' repeated n_i times, $i = 1, \ldots, D$, λ_D is D-dimensional Lebesgue measure, and $I_A(\cdot)$ is the indicator function of the set A. The structure of π is complicated and computation of functionals such as $\int h(\theta_1, \ldots, \theta_n)\, \pi(d\theta_1, \ldots, d\theta_n \mid x_1, \ldots, x_n)$ requires summations over the set of the partitions of $\{1, \ldots, n\}$ whose number becomes intractable if n is even moderately large. Therefore, alternative algorithms are required to estimate such functionals. We consider two Gibbs sampling algorithms for obtaining (approximate) samples from π.

Algorithm A. This is just ordinary Gibbs sampling (cf. Smith and Roberts, 1993), where $(\theta_1, \ldots, \theta_n)$ are updated one at the time. It was proposed by Escobar (1994) for mixtures of normals. The general structure of the resulting Markov chain is as follows. The state space is Θ^n. Let $f_0(x_i) = \int f(x \mid \theta) dG_0(\theta)$ and $g_0(\theta \mid x) = \frac{f(x|\theta)\, g_0(\theta)}{\int f(x|\theta)\, dG_0(\theta)}$. The algorithm proceeds

by updating one coordinate θ_i at a time. The i^{th} coordinate is updated by setting it equal to the current value of θ_j $(j = 1, 2, \ldots, n,\ j \neq i)$ with probability $q_{i,j} = \frac{f(x_i|\theta_j)}{Mf_0(x_i)+\sum_{l\neq i} f(x_i|\theta_l)}$; with the remaining probability $q_{i,0} = \frac{Mf_0(x_i)}{Mf_0(x_i)+\sum_{l\neq i} f(x_i|\theta_l)}$ we replace θ_i by a new value chosen independently from the distribution $g_0(\cdot \mid x_i)$. In summary, the value of θ_i is replaced either by a previous value θ_j, or by a fresh value chosen from $g_0(\cdot \mid x_i)$, with appropriate probabilities. It is straightforward to verify that these probabilities are chosen so that they preserve the stationary distribution $\pi(d\theta_1, \ldots, d\theta_n \mid x_1, \ldots, x_n)$.

The overall Markov chain on Θ^n will be one of two types. For the *systematic scan Gibbs sampler*, one iteration of the Markov chain consists of updating, in turn, the current value of θ_1, then θ_2, ..., then θ_n. For the *random scan Gibbs sampler*, one iteration of the Markov chain consists of choosing uniformly at random $I \in \{1, 2, \ldots, n\}$, and then updating θ_I. In either case, the coordinate updatings are done according to the above coordinate updating rules.

Algorithm B. It is clear that, if $Mf_0(x_i)$ is small (e.g. if M is very small), then the probability of generating fresh values from $g_0(\cdot \mid x_i)$ in the above Markov chain can be small so that the chain moves slowly. Therefore, a different algorithm has been proposed by MacEachern (1994) and Bush and MacEachern (1996). This involves updating all equal θ_i values simultaneously. This modification may allow the chain to forget its initial values much faster, especially for large n and small M, moving faster among the *configurations* of $(\theta_1, \ldots, \theta_n)$. A configuration can be described by the number D of distinct values among $(\theta_1, \ldots, \theta_n)$, and by a vector of r.v.'s $(S_1, \ldots, S_n)$ which, given the distinct values $(\theta'_1, \ldots, \theta'_D)$, are such that $S_i = j$ if $\theta_i = \theta'_j$, $1 \leq i \leq n$, $1 \leq j \leq D$. The Gibbs sampler consists of updating the vector $(S_1, \ldots, S_n)$ and the vector of distinct values in turn, from the appropriate full conditionals.

Specifically, to *update the configuration*, we generate a new vector $(S_1, \ldots, S_n)$ by updating one S_i value at a time, as follows. Let $\underline{S}_{[i]}$ be the vector $(S_1, \ldots, S_n)$ without the i-th element. With probability $q_{i,0} = \frac{Mf_0(x_i)}{Mf_0(x_i)+\sum_{l\neq i} f(x_i|\theta'_l)\, n_{l,[i]}}$, where $n_{j,[i]}$ is the number of elements in $\underline{S}_{[i]}$ which are equal to j, we let $S_i = 0$; otherwise, with probability $q_{i,j} = \frac{f(x_i|\theta'_j)\, n_{j,[i]}}{Mf_0(x_i)+\sum_{l\neq i} f(x_i|\theta'_l)\, n_{l,[i]}}$, we let $S_i = j$. Updating $(S_1, \ldots, S_n)$ also provides the updated number of distinct values, say D.

To *update the distinct values*, we generate $(\theta'_1, \ldots, \theta'_D)$ conditionally independently, choosing θ'_j $(1 \leq j \leq D)$ from a distribution proportional to

$$\prod_{\ell \in I_j} f(x_\ell \mid \theta'_j)\, g_0(\theta'_j)$$

where $I_j = \{i : S_i = j\}$ is the group of observations which share the same value θ'_j.

3 Results.

Our main method of proof below shall be the coupling method (see e.g. Lindvall, 1992). Recall that for a Markov chain $P(\theta, \cdot)$ on a state space Θ, with initial distribution μ_0 and stationary distribution π, the coupling method works as follows. If we can jointly construct chains $\{\theta_k\}_{k=0}^{\infty}$ and $\{Z_k\}_{k=0}^{\infty}$, such that $\theta_0 \sim \mu_0$, $Z_0 \sim \pi$, $\mathcal{L}(\theta_k \,|\, \theta_0, \ldots, \theta_{k-1}) = P(\theta_{k-1}, \cdot)$, and $\mathcal{L}(Z_k \,|\, Z_0, \ldots, Z_{k-1}) = P(Z_{k-1}, \cdot)$, then we have

$$\|\mu_0 P^n - \pi\| \equiv \sup_{A \subseteq \Theta} |\mu_0 P^n(A) - \pi(A)| \leq \mathbf{P}(T > n)\,,$$

where $\| \ldots \|$ is total variation distance, and where

$$T = \inf\{n \in \mathbf{N}\,;\, \theta_k = Z_k \text{ for all } k \geq n\}$$

is the *coupling time.*

Our first result shows that if the kernel $f(x \mid \theta)$ is uniformly bounded, then the Markov chain is uniformly ergodic with explicit rate bound (independent of starting distribution and of the g_0 and $f(x \mid \theta)$ distributions).

Theorem 1 *Let $m = \min_{i=1,\ldots,n} f_0(x_i)$. Suppose that $f(x_i \mid \theta) \leq J$ for all θ and i. Then, for the systematic scan Gibbs sampler as in algorithm A above, for any initial distribution μ_0, we have*

$$\|\mu_0 P^k - \pi\| \leq (1 - q^n)^k\,,$$

where $q = \frac{mM}{Mm + J(n-1)}$, with M the scale parameter of the Dirichlet process.

Proof. Observe that

$$q_{i,0} = \frac{M f_0(x_i)}{M f_0(x_i) + \sum_{\ell \neq i} f(x_\ell \mid \theta_j)} \geq \frac{M f_0(x_i)}{M f_0(x_i) + (n-1)J} \geq \frac{Mm}{Mm + (n-1)J} = q\ ,$$

the last inequality following by noting that $\frac{x}{x+a}$ in a nondecreasing function of x for $a > 0$.

Now, by inspection, we can define the chains $\{Z_k\}_{k=0}^{\infty}$ and $\{\theta_k\}_{k=0}^{\infty}$ so that for each coordinate i, with probability at least q, they *both* update from $g_0(\cdot \mid x_i)$, and furthermore in that case they choose the *same* value from $g_0(\cdot \mid x_i)$. But then for each iteration of the Markov chain, there is independent probability at least q^n that the two Markov chains will become equal in all coordinates. Thus, $\mathbf{P}(T > k) \leq (1 - q^n)^k$. The result follows. $\diamond$

Remark. Clearly, for the random-scan Gibbs sampler, we can obtain a similar (but even larger) bound.

The bounds of this proposition are impractically large (since q^n will usually be very small). To improve this, it is necessary to study the coupling more carefully. We are able to obtain substantial improvements in a special case, namely when certain function values are identically constant.

Theorem 2 *For the random-scan Gibbs sampler, suppose that*

$$f_0(x_i) = m\,, \quad \text{and} \quad f(x_i \mid \theta) = J \tag{1}$$

are independent of i and θ. Then setting $q = \frac{Mm}{Mm+(n-1)J}$, we have that for any initial distribution μ_0,

$$\|\mu_0 P^k - \pi\| \leq (1 - \frac{q}{n})^k n\,.$$

Proof. Similar to the previous proof, we define the chains $\{Z_k\}_{k=0}^{\infty}$ and $\{\theta_k\}_{k=0}^{\infty}$ so that for each coordinate i, with probability at least q, they *both* update from $g_0(\cdot \mid x_i)$, and furthermore in that case they choose the *same* value from $g_0(\cdot \mid x_i)$.

Let D_k be the number of coordinates at which the two chains differ at time k. Then by considering separately the cases where we pick a coordinate where the two chains do or do not differ, and where we update it either to a fresh value, to a value where the two chains do not differ, or to a value where the two chains do differ, we have that

$$\mathbf{E}(D_{k+1} - D_k \,|\, Z_k, \theta_k) \;=$$

$$\left(1 - \frac{D_k}{n}\right)\left(\frac{D_k}{n-1}\right)(1-q) \;-\; \left(\frac{D_k}{n}\right)\left[q + (1-q)\left(1 - \frac{D_k - 1}{n-1}\right)\right]$$

$$= \; -\frac{q}{n} D_k\,,$$

so that

$$\mathbf{E}(D_{k+1} \,|\, Z_k, \theta_k) \;\leq\; (1 - \frac{q}{n}) D_k\,.$$

Therefore,

$$\mathbf{E}(D_k) \;\leq\; (1 - \frac{q}{n})^k \mathbf{E}(D_0) \;\leq\; (1 - \frac{q}{n})^k n\,.$$

On the other hand, since D_k is non-negative integer values, clearly $\mathbf{E}(D_k) \geq \mathbf{P}(D_k > 0)$. Hence,

$$\mathbf{P}(T > k) \;=\; \mathbf{P}(D_k > 0) \;\leq\; \mathbf{E}(D_k) \;\leq\; (1 - \frac{q}{n})^k n\,.$$

The result follows. $\diamond$.

The bounds in this theorem are not as impractically large as those of the previous theorem. However, the hypotheses are much stronger. We return to this point below.

One might hope to achieve stronger convergence results for algorithm B, since it appears to get less stuck at particular parameter values. In fact, this is not easy, as we discuss below. However, can prove a general uniform ergodicity result for algorith, B, similar to that of Theorem 1.

Theorem 3 *Let $m = \min_{i=1,\ldots,n} f_0(x_i)$ and suppose that $f(x_i \mid \theta) \leq J$ for any i and θ. Then for the Gibbs sampling described as algorithm B, for any starting distribution μ_0 we have*

$$\|\mu_0 P^k - \pi\| \leq (1 - q^n)^k \,,$$

where $q = \frac{mM}{Mm+J(n-1)}$.

Proof. The proof is similar to that of Theorem 1. Indeed, we can define two chains $\{Z_k\}_{k=0}^n$ and $\{\theta_k\}_{k=0}^n$ such that, conditional on being in the same configuration $(S_1, \ldots, S_n)$, they both update the distinct values from the same distribution with the same values. Now, the probability of having the same configuration for both chains in one iteration is bigger than the probability of having the configuration $(S_1, \ldots, S_n) = (0, 0, \ldots, 0)$ in both chains in one iteration, and this is bigger than q^n.

Therefore, $\mathbf{P}(T > k) \leq (1 - q^n)^k$ and the result follows. $\diamond$

The above bound, like that of Theorem 1, is too large for practical use (because of the factors of $1 - q^n$ instead of $1 - \frac{q}{n}$). Indeed, for either algorithm A or algorithm B, we are only able to achieve reasonable (i.e., non-exponential as a function of n) bounds under the special case of "equal function values" as in equation (1).

Interestingly, there are several other natural approaches to proving bounds on the convergence rates of these Markov chains, but each such approach seems to work well (i.e., to give non-exponential bounds) precisely when the same condition (1) – or something similar – is satisfied! For example, for algorithm B, one can observe that after one iteration, with high probability there will only be about $O(\log n)$ distinct values θ_i'. If condition (1) is satisfied, then there is non-exponentially-small probability that two chains will jump to the same cluster structure on the following iteration, thus leading to good coupling bounds. However, if (1) is not satisfied, this does not seem to be the case.

Similarly, Jensen's inequality can be used in the coupling proofs to replace certain transition probabilities by their expected values. If (1) is satisfied, this appears to again lead to useful coupling bounds; but if not, then this approach does not seem helpful.

Finally, uniform ergodicity can sometimes be helpful for using perfect sampling algorithms, as in Propp and Wilson (1996). Such algorithms are more feasible if the chain is stochastically monotone with respect to some ordering. Now, there is a natural partial ordering on the space of configurations; specifically, two configurations $\mathcal{C}_1$ and $\mathcal{C}_2$ are ordered with $\mathcal{C}_1 \leq \mathcal{C}_2$ if each element of $\mathcal{C}_1$ is contained in some element of $\mathcal{C}_2$, i.e. if $\mathcal{C}_1$ is a finer partition than $\mathcal{C}_2$. If our Markov chain were stochastically monotone with respect to this ordering, then to run a perfect sampling algorithm it would only be necessary to keep track of the positions when starting from the two extreme partitions (i.e. $\mathcal{C}_1 = (\{1, 2, \ldots, n\})$ and $\mathcal{C}_2 = (\{1\}, \{2\}, \ldots, \{n\})$). Unfortunately, the Markov chain does not seem to be monotonic with respect to this partial ordering in general; however, it is monotonic in the case that (1) holds.

4 Examples.

(a) Empirical Bayes. Consider r.v.'s $X_1, \ldots, X_n$ which, conditionally on $(\theta_1, \ldots, \theta_n)$, are independently distributed, with $X_i \mid \theta_i$ having a binomial distribution of parameters l_i (known) and θ_i. Also, $\theta_1, \ldots, \theta_n$ are a sample from a d.f. G, where G is a Dirichlet process $\mathcal{D}(MG_0)$. This setting is treated in Berry and Christensen (1979) and Liu (1996). The Gibbs sampling algorithms A and B can be used in this problem, and the binomial kernel is uniformly bounded, so that from Theorems 1 and 3 the Markov chains resulting from the Gibbs sampling are uniformly ergodic.

Liu (1996) proposed a sequential imputation method for computation in this problem. It is not easy to make a general comparison between Gibbs sampling strategy and sequential imputation method. It can be noted (see Liu, 1996) that, since sequential imputation is advantageous in updating posterior distributions when new data arrive, Gibbs sampling and sequential imputation can be complementary to each other.

(b) Gaussian mixture models. Mixtures of normals are proposed in many applications, e.g. in Bayesian nonparametric density estimation (cf. Escobar and West, 1995; West, Müller, and Escobar, 1994). Let $\theta_i = (\mu_i, \sigma_i^2)$ and suppose that $X_i \mid \theta_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots$, $\theta_1, \theta_2, \ldots \mid G$ are i.i.d. according to G, and $G \sim \mathcal{D}(MG_0)$ where G and G_0 are bivariate distributions on $\mathbf{R} \times \mathbf{R}^+$. If the variances σ_i^2 are bounded away from zero, i.e. $\sigma_i^2 \geq \sigma^* > 0$, then the normal kernel is uniformly bounded so that Theorems 1 and 3 hold. (If instead the mixing distribution $G(\cdot)$ is parametric, then more precise convergence results are known; see for example Rosenthal, 1995a.)

(c) Finite mixture models. Consider a mixture model of the form $f(x \mid \{w_j\}) = \sum_{j=1}^k w_j f_j(x)$, where f_j are known probability density functions, and where the w_j are weights summing to 1. An example of $f(x \mid \{w_j\})$ are Bernstein densities (Petrone, 1996). Then $f(x \mid \{w_j\})$ can be written in the form $\int f(x \mid \theta)\, dG(\theta)$ if we let $f(x \mid \theta) = \sum_{j=1}^k f_j(x) I_{\Theta_j}(\theta)$, where $(\Theta_1, \ldots, \Theta_k)$ is a partition of the parameter space Θ such that $w_j = \int_{\Theta_j} dG(\theta)$. Thus, the results of the present paper can be applied to finite mixture models in this manner. The formulation of a finite mixture model in this form is useful for generalising to the case of a random number k of components (as done e.g. in Petrone, 1996).

5 Final remarks.

We have studied the Markov chains resulting from two Gibbs sampling algorithms that are useful in mixture models with a Dirichlet process mixing distribution. We showed (Theorems 1 and 3) that, if the mixture kernel is bounded, the Markov chains are uniformly ergodic, with explicit rate bounds. Improving the bound seems difficult in general, due to the complicated structure of the stationary distribution π, though we are able to give a much-improved result (Theorem 2) in a special case.

In fact, there might be examples where the Markov chain does actually converge slowly. For algortithm B, which in practice seems more efficient, this might perhaps happen if the prior and the likelihood suggest very different configurations, so that the posterior on the

space of configurations is multimodal. In this case, there might be a drift in the chain towards staying in one configuration mode.

Our last remark is about the possibility of allowing uncertainty about M and G_0 in the model. The importance of determining the weight M and the shape of G_0 is discussed in detail in Escobar and West (1995). We do not treat this generalisation here, but the results we show for the case of fixed M and G_0 may be of benefict for studying convergence rates of Gibbs sampling in the more general setting with random M and G_0.

References

[1] Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152-1174.

[2] Berry, D and Christensen, R. (1979). Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *Ann. Statist.* **7** 558-568.

[3] Bush, C.A. and MacEachern, S.N. (1996). A semiparametric Bayesian model for randomized block designs. *Biometrika*, **83**, 275-285.

[4] Escobar, M.D. (1994). Estimating normal means with a Dirichlet process prior. J. Amer. Stat. Assoc. **89**, 268–277.

[5] Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. J. Amer. Stat. Assoc. **90**, 577–587.

[6] Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. *Recent Advances in Statistics*, H. Rizvi and J. Rustagi eds. New York: Academic Press, 287-302.

[7] Kuo, L. (1986) Computation of mixtures of Dirichlet processes. *SIAM J. Sci. Stat. Comput.*, **7**, 1, 60-71.

[8] Lindvall, T. (1992). *Lectures on the Coupling Method.* Wiley & Sons, New York.

[9] Liu, J.S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *Ann. Statist.* **24**, 911-930.

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

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

[12] Newton, M.A. and Zhang, Y.-L. (1996). A partial predictive recursion. *Techn. Rep. 965*, University of Wisconsin-Madison.

[13] Petrone, S. (1996). Bayesian density estimation using Bernsetin polynomials. *Canadian J. Statist.*, to appear.

[14] Petrone, S. and Raftery, A.E. (1997). A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. *Statist. Prob. Letters*, **36**, 69-83.

[15] Propp, J. and Wilson, D. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanincs. *Random Structures and Algorithms*, **9**, 223-252.

[16] Roberts, G.O and Rosenthal, J.S. (1997). Markov chain Monte Carlo: Some practical implications of theoretical results. *Canadian J. Statist.*, to appear.

[17] Rosenthal, J.S. (1995a). Rates of convergence for Gibbs sampling for variance components models. *Ann. Statist.*, **23**, 740-761.

[18] Rosenthal, J.S. (1995b). Minorization conditions and convergence rates for Markov chain Monte Carlo. J. Amer. Stat. Assoc. **90**, 558-566.

[19] Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). J. Roy. Stat. Soc. Ser. B **55**, 3-24.

[20] West, M., Müller, P. and Escobar, M.D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of Uncertainty: A Tribute to D.V. Lindley*, P.R. Freeman and A.F.M. Smith eds. John Wiley & Sons Ltd, New York, 363-386.