Estimating Mixture of Dirichlet Process Models

Steven N. MacEachern and Peter Müller

Steven MacEachern is Associate Professor, Department of Statistics, Ohio State University, Columbus, OH 43210, and Peter Müller is Assistant Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251.

Abstract

Current Gibbs sampling schemes in mixture of Dirichlet process (MDP) models are restricted to using "conjugate" base measures which allow analytic evaluation of the transition probabilities when resampling configurations, or alternatively need to rely on approximate numeric evaluations of some transition probabilities. Implementation of Gibbs sampling in more general MDP models is an open and important problem since most applications call for the use of non-conjugate base measures.

In this paper we propose a conceptual framework for computational strategies. This framework provides a perspective on current methods, facilitates comparisons between them, and leads to several new methods that expand the scope of MDP models to non-conjugate situations. We discuss one in detail. The basic strategy is based on expanding the parameter vector, and is applicable for MDP models with arbitrary base measure and likelihood. Strategies are also presented for the important class of normal-normal MDP models and for problems with fixed or few hyperparameters. The proposed algorithms are easily implemented and illustrated with an application.

KEY WORDS: Gibbs sampling, hierarchical models, Markov chain Monte Carlo, simulation.

1 Introduction

1.1 The MDP model

This paper proposes a novel solution strategy to an open problem in implementing Gibbs sampling for mixture of Dirichlet process (MDP) models with non-conjugate base-measure and likelihood. MDP models have become increasingly popular for modeling when conventional parametric models would impose unreasonably stiff constraints on the distributional assumptions. Examples include empirical Bayes problems (Escobar, 1994), nonparametric regression (Müller, Erkanli and West, 1996), density estimation (Escobar and West, 1995; Gasparini, 1996), hierarchical modeling (MacEachern, 1994; West, Müller and Escobar, 1994; Bush and MacEachern, 1996), censored data settings (Doss, 1994; Kuo and Smith, 1992; Gelfand and Kuo, 1991), and estimating possibly non-standard link functions (Newton, Czado and Chappell 1996; Erkanli, Stangl and Müller 1993).

Despite the large variety of applications, the core of the MDP model can basically be thought of as a simple Bayes model given by the likelihood $y_i \sim p_{\theta_i}(y_i)$ and prior $\theta_i \sim G(\theta_i)$, with added uncertainty about the prior distribution G:

$$y_i \sim p_{\theta_i}(y_i),\ i = 1, ..., n, \quad \theta_i \sim G, \quad G \sim DP(G_0, \alpha), \tag{1}$$

where $G \sim DP(G_0, \alpha)$ refers to G being a random distribution generated by a Dirichlet process with base measure αG_0 and total mass parameter α. An important instance of the general model is the normal-normal MDP model, given by $p_{\mu,\Sigma}(y_i) = N(y_i; \mu, \Sigma)$ and $G_0(\mu, \Sigma) = N(\mu; m, B) \cdot W(\Sigma^{-1}; r, R)$. Here $F(x; \eta)$ means that the random variable x has distribution F with parameter η, N denotes the normal distribution, and W denotes the Wishart distribution. Models with more general applications typically require another portion to the hierarchy that allows the introduction of observation specific covariates, say x_i, and hyperparameters ν. The more complex models also introduce distributions on the hyperparameters ν, G_0, and α. But conditional on these additional parameters, the portion of the model involving the MDP has the form given above. See Antoniak (1974) and Ferguson (1973) for discussion of the Dirichlet process.

1.2 Estimating the MDP model

A key feature of the Dirichlet process in this MDP model is that the θ_i are marginally samples from G_0, and with positive probability some of the θ_i are identical. This is due to the discreteness of

the random measure G. This discreteness of G is the main impediment to an efficient estimation. Posterior integration, and thereby most inference, is made difficult by a combinatorial explosion in the number of terms in the posterior distribution, due to the need to account for all possible configurations of how the θ_i's are identical and distinct. See, for example, Antoniak (1974). However, implementation of Gibbs sampling is almost straightforward when one marginalizes over G and works directly with the θ_i. Except for a difficulty which arises when resampling θ_i conditional on all other parameters. The new value of θ_i can either be one of the θ_h's, $h \neq i$, or θ_i could be a new draw from G_0. Computing the probability of the latter alternative involves an integral of $p_{\theta_i}(y_i)$ with respect to $G_0(\theta_i)$: $P(\theta_i \neq \theta_h, h \neq i | \theta_h, h \neq i) \propto q_0 = \int p_{\theta_i}(y_i) dG_0(\theta_i)$. Using δ_x to indicate a point mass at x, the conditional posterior for θ_i can be written as:

$$p(\theta_i | \theta_{-i}, y) \propto \sum_{h \neq i} q_h \delta_{\theta_h} + q_0 G_i(\theta_i),$$

with $q_h = p_{\theta_h}(y_i)$ and $G_i(\theta_i) \propto G_0(\theta_i) p_{\theta_i}(y_i)$. We will discuss details in Section 2.2. Following a pragmatic definition we call a prior/likelihood pair $G_0(\theta_i)$ and $p_{\theta_i}(y_i)$ "conjugate" if the posterior $p(\theta_i | y_i)$ takes the form of some well known distribution, allowing efficient random variate generation and integration. Evaluation of the integral expression for q_0 is non-trivial unless G_0 and p_{θ_i} are a conjugate pair. Current implementations therefore either use a conjugate model or rely on approximate computations. Overcoming this computational hurdle is important because of the wide range of current and potential applications of MDP models, and the need in most applications to leave the conjugate framework. In this paper we propose a Gibbs sampling scheme which achieves this.

Recent work by Walker and Damien (1996) describes a promising alternative computational strategy for fitting MDP models.

1.3 Examples

In the description of current and proposed algorithms we will refer to the following models as examples (the term "conjugate" is used in the sense defined above):

Conjugate normal-normal MDP model:

$$p_{\mu,\Sigma}(y_i) = N(y_i; \mu, \Sigma),\ G_0(\mu, \Sigma) = N(\mu; m, \tau\Sigma) \cdot W(\Sigma^{-1}; r, R),$$

Conditionally conjugate normal-normal MDP model:

$p_{\mu,\Sigma}(y_i) = N(y_i; \mu, \Sigma)$, $G_0(\mu, \Sigma) = N(\mu; m, B) \cdot W(\Sigma^{-1}; r, R)$. The pair $p_{\mu,\Sigma}$, $G_0(\mu, \Sigma)$ is conjugate in μ if Σ is fixed, and conjugate in Σ when μ is fixed, but the posterior $p(\mu, \Sigma|y)$ does not allow efficient random variate generation and analytic integration (to compute the probability q_0 of a new draw from G_0 when resampling θ_i as described in Section 2.2(i)).

Non-conjugate uniform-normal MDP model: $p_{\mu,\Sigma}(y_i) = N(y_i; \mu, \Sigma)$, $G_0(\mu, \Sigma) = Unif(\mu; \mu_L, \mu_H) \cdot W(\Sigma^{-1}; r, R)$, where $Unif(\mu; \mu_L, \mu_H)$ denotes a uniform distribution on a rectangle with lower bounds μ_L and upper bounds μ_H.

1.4 Novel Gibbs sampling schemes for the MDP model

In this paper we propose Gibbs sampling schemes which allow estimation of MDP models without restriction to conjugate models. In Section 2 we review the sampling scheme which is currently most often applied. In Section 3 we develop a model augmentation which allows efficient implementation of Gibbs sampling in general, possibly non-conjugate MDP models. Section 4 illustrates the proposed scheme with an application to the non-conjugate uniform-normal MDP model. Section 5 concludes with a final discussion, including an outline of an alternative model augmentation that is sometimes preferable to the augmentation developed here.

2 Gibbs Sampling in MDP models

In this section, we briefly review Markov chain Monte Carlo schemes currently applied to estimate MDP models. For a more detailed discussion of the schemes, we refer the reader to Bush and MacEachern (1996) and West, Muller, and Escobar (1994).

2.1 Notation

We start out by fixing notation. In the general MDP model with continuous base measure G_0, let $\phi = \{\phi_1, ...\phi_k\}$ denote the set of distinct θ_i's, where $k \leq n$ is the number of distinct elements in the vector $\theta = (\theta_1, \ldots, \theta_n)$. Let $s = (s_1, \ldots, s_n)$ denote the vector of configuration indicators defined by $s_i = j$ iff $\theta_i = \phi_j$, $i = 1, \ldots, n$, and let n_j be the number of $s_i = j$. We will use the term "cluster" to refer to the set of all observations y_i, or just the indices i, or the corresponding θ_i's, with identical configuration indicators s_i. The n_j defined above is the size of the j-th cluster: $n_j = |\{i : s_i = j\}|$.

Since we allow arbitrary permutations of the ϕ indices $j = 1, \ldots, k$, any given θ vector corresponds to $k!$ pairs (ϕ, s). For later reference we note that if we assign equal probabilities to each of the $k!$ permutations then

$$P(s) = \frac{\alpha^k(\alpha-1)!\prod(n_j-1)!}{(\alpha+n-1)!k!} \tag{2}$$

This expression for $P(s)$ is easily obtained by multiplying the conditional prior distributions $P(s_i|s_1, \ldots, s_{i-1})$, $i = 1, \ldots, n$, that arise from the Polya urn scheme representation of the Dirichlet process. Under the convention that clusters are numbered consecutively as they arise, i.e., the first cluster is assigned the number 1, etc., we have $P(s_i = j|s_1, \ldots, s_{i-1}) = n_{i,j}/(\alpha+i-1)$, $j = 1, \ldots, k_i$, and $P(s_i = k_i + 1|s_1, \ldots, s_{i-1}) = \alpha/(\alpha + i - 1)$. Here k_i denotes the number of clusters among $\theta_1, \ldots, \theta_{i-1}$ and $n_{i,j} = |\{s_h = j; h < i\}|$ is the size of cluster j restricted to the first $i-1$ observations. The additional $1/k!$ term accounts for the permutations of the ϕ indices. This description of the configuration relies on the assumption of continuity for the underlying base measure of the Dirichlet process. When the base measure contains a discrete component, some of the θ_i's may be equal not because they belong to the same cluster, but because the draws from G_0 happen to be equal. The notion of a configuration may be extended in a straightforward fashion to these settings. We avoid the extension here to retain a clearer notation.

Gibbs sampling is implemented by iterative sampling from the full conditionals described in the next subsection. In the formulas below, the subscript "$-i$" means without the i^{th} element of the vector. The superscript "$-$" refers to a summary with the appropriate observation and/or parameter removed. For example, with θ_i removed, k^- refers to the number of clusters formed by θ_{-i}, and n_j^- represents the number of elements in cluster j when observation i is removed. The conditioning on the data, $y = (y_1, \ldots, y_n)$, does not appear in the notation, but should be understood.

2.2 The full conditionals

(i) Resampling (θ_i, s_i) given all other parameters: The new value of θ_i is equal to $\theta_h, h \neq i$ with probability proportional to $q_h = p_{\theta_h}(y_i)$, or with probability proportional to $q_0 = \alpha \int p_{\theta_i}(y_i) dG_0(\theta_i)$ is a draw from $G_i(\theta_i) \propto G_0(\theta_i)p_{\theta_i}(y_i)$. The distribution G_i is the posterior in a simple Bayes model given by likelihood $y_i \sim p_{\theta_i}(y_i)$ and prior $\theta_i \sim G_0(\theta_i)$. Using δ_x to indicate a point mass

at x, combining identical θ_h's, and redefining $q_j = p_{\phi_j}(y_i)$, this can be written as:

$$p(\theta_i|\theta_{-i}, s_{-i}, y) \propto \sum_{j=1}^{k} n_j^- q_j \delta_{\phi_j} + q_0 G_i(\theta_i). \tag{3}$$

Note that sampling θ_i implicitly samples a new configuration s_i. If the base distribution $G_0(\theta_i)$ and the likelihood $p_{\theta_i}(y_i)$ are chosen as a conjugate pair, then the integral q_0 can be analytically evaluated. If, however, G_0 is not conjugate with p_{θ_i} then resampling the configuration becomes difficult, as the integral q_0 may be computationally intensive.

Note that n_j^- could be zero for some j. This happens if the previous configuration put s_i into a cluster of size one, i.e., $n_j = 1$ for $j = s_i$. After resampling each s_i it is necessary to redefine the list ϕ of unique cluster locations if either a new cluster is created by sampling $\theta_i \sim G_i$, or an old cluster is left "empty", i.e., with $n_j = 0$, by reallocating the only element of a previous cluster to another cluster. In either case change ϕ accordingly by adding an element to ϕ or deleting ϕ_j and relabeling the remaining elements of ϕ if necessary. Redefine s and k to agree with the current clustering of θ.

(ii) Resampling ϕ_j conditional on the configuration s and all other parameters is straightforward. For a fixed j, it amounts to sampling from the posterior in the simple Bayes model given by $y_i \sim p_{\phi_j}$ and $\phi_j \sim G_0$, for $i \in \{i : s_i = j\}$. In the conjugate normal-normal MDP, for example, the conditional posterior for ϕ_j will simply be the appropriate inverse Wishart/normal posterior. In the conditionally conjugate normal-normal MDP model resampling $\phi_j = (\mu_j^\phi, \Sigma_j^\phi)$ would be broken into two parts: (iia) Resampling μ_j^ϕ conditional on Σ_j^ϕ (and all other parameters), and (iib) resampling Σ_j^ϕ conditional on μ_j^ϕ.

(iii) Resampling α and other hyperparameters: While not explicitly included in model (1), typical MDP applications would include a hyperprior on the total mass parameter α and other hyperparameters. For example, an unknown normal mean and covariance matrix or Wishart parameters would appear in the specification of G_0. Sampling of α is described in Escobar and West (1995), based on West (1992). See also Liu (1996) for an alternative approach based on sequential imputation. Sampling of other hyperparameters is typically straightforward, since conditioning on the configuration s reduces the problem to a conventional hierarchical model.

2.3 Current sampling schemes

MCMC implementations to estimate MDP models discussed in recent literature fit into the framework presented here. All may be represented in terms of steps (i) through (iii) with minor variations.

Escobar and West use a scheme similar to the one described above, but don't include the second step of moving the cluster locations. In terms of the latent variable notation, they use: (i) sample $[\theta_i|\theta_{-i}]$ for $i = 1, \ldots, n$. Drop (ii).

Bush and MacEachern use the above scheme. (i) sample $[\theta_i|\theta_{-i}]$ for $i = 1, \ldots, n$. (ii) sample$[\phi_1, \ldots, \phi_k|s]$.

MacEachern uses a scheme that dispenses with the cluster locations ϕ_j entirely. He uses: (i) sample $[s_i|s_{-i}]$ for $i = 1, \ldots, n$.

In each of these cases, the conditional distributions in (i) require an integration that is costly if p_{θ_i} and G_0 are not conjugate.

West, Müller, and Escobar (1994) present the first algorithm designed specifically for use with non-conjugate models. In this algorithm they approximate the draw in step (i) of the algorithm above by approximating q_0. Specifically, they take a random draw from G_0, say θ', and replace $\int p_{\theta_i}(y_i)dG_0(\theta_i)$ with $p_{\theta'}(y_i)$. The resulting rescaled probabilities typically lead to a Markov chain with a stationary distribution, but a stationary distribution which differs from the posterior. While this method does provide an approximation to the posterior, the accuracy of the approximation is difficult to evaluate because the approximation occurs within the transition probabilities. In some circumstances the approximation can be quite poor. Straightforward re-weighting of the output of the approximating chain to provide a weighted sample from the posterior is also prevented, since there appears to be no simple representation of the stationary distribution.

The new sampling plan entirely avoids the difficult integration and can replace 2.2.(i) when evaluation of q_0 is problematic.

3 Estimating non-conjugate MDP models

3.1 The novel algorithm

The problem of evaluating q_0 in (3) arises because we have to integrate over a new value of θ_i if the new indicator s_i opens a new cluster, i.e., $s_i \neq s_h$, $h \neq i$. We propose an alternative parametrization

by augmenting ϕ to

$$\{\underbrace{\phi_1, \ldots, \phi_k}_{\phi_F}, \underbrace{\phi_{k+1}, \ldots, \phi_n}_{\phi_E}\}.$$

with the same independent prior, $\phi_j \sim G_0$, on the ϕ_j, the same definition of configuration indicators s_i, and the prior $p(s)$ given in (2). The augmentation relies upon the constraint that there be *no gaps* in the values of the s_i, i.e., $n_j > 0$ for $j = 1, \ldots, k$, and $n_j = 0$, for $j = k+1, \ldots, n$. This corresponds to an interpretation of $\phi_E = \{\phi_{k+1}, \ldots, \phi_n\}$ as "potential", but not yet used cluster locations. We will refer to ϕ_E as "empty" clusters, and ϕ_F as "full" clusters.

In the augmented model the Gibbs sampler is simplified: Evaluation of integrals of the type q_0 is replaced by simple likelihood evaluations.

No Gaps Algorithm: Repeat (ia) and (ib) for $i = 1, \ldots, n$. Then perform step (ii).

(ia) Sample $(s, \phi_F)|\theta$. This step reduces to choosing a permutation of the cluster indices $1, ..., k$, with each permutation having probability $1/k!$.

(ib) Sample $s_i|(s_{-i}, \phi)$. The posterior conditional distributions are given by

$$P(s_i = j|s_{-i}, \phi, y) \propto P(s_i = j|s_{-i}, \phi) p_{\phi_j}(y_i). \tag{4}$$

The conditional prior distribution for $p(s_i|s_{-i}, \phi)$ is best described in two cases. First, if s_{-i} is a state where $n_j^- = 0$ for some $j \leq k^-$, then the *no gaps* constraint implies that the distribution of $s_i|(s_{-i}, \phi)$ is degenerate: $P(s_i = j|s_{-i}, \phi) = 1$. Second, when no gap would be created with the removal of s_i, we have

$$\begin{aligned} P(s_i = j|s_{-i}, \phi) &\propto n_j^- \text{ for } j = 1, \ldots, k^-, \\ P(s_i = k^- + 1|s_{-i}, \phi) &\propto \alpha/(k^- + 1). \end{aligned} \tag{5}$$

(ii) Sample $\phi|s$. The conditional distribution for ϕ consists of a product of n independent distributions. For $j = k+1, \ldots, n$ this simply amounts to draws, $\phi_j \sim G_0$, from the base measure (see the comment below about actually recording ϕ_j, $j = k+1, \ldots, n$). For $j = 1, \ldots, k$, the conditional posterior remains as in Section 2.2, step (ii).

The implementation of this algorithm may be simplified and speeded by discarding unnecessary draws that do not alter the chain itself. Thus, we recommend that the first step, where the indices

are permuted, be used only when $n_{s_i} = 1$. In this case, the permutation results in $s_i = k$ with probability $1/k$ and $s_i < k$ with probability $(k-1)/k$. The former case leads to a non-degenerate posterior conditional for s_i with distribution given above, while the latter leads to a degenerate posterior for s_i. Thus (ia) and (ib) can be combined to:

(i') If $n_{s_i} > 1$ then resample s_i with probabilities (5). Note that $k^- = k$.

If $n_{s_i} = 1$ then with probability $(k-1)/k$ leave s_i unchanged. Otherwise relabel clusters such that $s_i = k$ and then resample s_i with probabilities (5). Note that now $k^- = k - 1$. Also, if s_i happens to be resampled as $s_i = k^- + 1 = k$, then θ_i remains in effect unchanged because the preceding relabeling kept the previous value of θ_i as ϕ_k.

The values of $\phi = (\phi_1, \ldots, \phi_n)$ are never changed during execution of step (i') (except relabeling of the indices as described above). This implies in particular that locations ϕ_j of clusters which are left empty by resampling some s_i are not discarded. Only indices are changed if needed.

In a typical cycle of the algorithm, most of the ϕ_j's will not be used. Since most of the ϕ_j will be drawn from the prior distribution G_0 (in step (ii) of the previous cycle), we do not sample them until they are needed.

3.2 Extension to hyperparameters

The algorithm described above may be extended to models that incorporate hyperparameters with the inclusion of a draw of these parameters, conditional on the observed vector (s, ϕ). The conditional distributions are exactly those of the parametric hierarchical model that replaces the Dirichlet process with G_0 (see Escobar and West, 1995, or MacEachern, 1994). The conditional distribution depends on all of ϕ, not just ϕ_F. We recommend that the draws be made for the parameters $(\phi_E, \nu)|\phi_F$. In this conditioning, since the distribution of $\phi_E|\nu$ is independent of $\phi_F|\nu$, the updating of the hyperparameters reduces to the usual updating for the hierarchical model without the Dirichlet process, conditioning only on ϕ_F.

3.3 Predictive distributions

The posterior feature of greatest interest is often a predictive distribution. In the case of density estimation, the predictive distribution for a future observation is of direct interest. It is

also the Bayes estimate under a weighted quadratic loss function. In the basic MDP model, the posterior predictive distribution is most easily found by returning from the *no gaps* model to the parameterization in terms of θ. Then the predictive distribution is given by $p(y_{n+1}|y) = \int\int p(y_{n+1}|\theta_{n+1})dp(\theta_{n+1}|\theta, y)dp(\theta, y)$. The inner integral reduces to an integral of $p(y_{n+1}|\theta_{n+1})$ against $(\sum_{j=1}^{k} n_j\delta_{\phi_j} + \alpha G_0)/(\alpha + n)$. The term involving G_0 may be evaluated as $\alpha\delta_{\tilde{\theta}}$ where $\tilde{\theta}$ represents a new draw from G_0.

To obtain an estimate of the predictive distribution as the algorithm proceeds, we use an average over iterates of the resulting Markov chain. Let θ^t denote the imputed parameter vector θ after t iterations. After each complete cycle of the algorithm, just after stage (ii), one has the estimate $1/T\sum_{t=1}^{T} p(y_{n+1}|\tilde{\theta^t}, \theta^t)$ when evaluation of the conditional distributions are feasible. When this evaluation is not feasible, after each iteration a value y_{n+1} can be generated, with the resulting estimator based on the sample of T such values.

In more complex models that involve hyperparameters or observation specific covariates, predictive distributions are obtained in a similar fashion. Typically, one will condition on the values of these other parameters during the evaluations (see the forthcoming example in Section 4). In other circumstances, one is interested in distributions where the future values of an observation specific covariate are unknown. In these cases, either an integration over the distribution of the unknown covariate or a generation of its value is required.

3.4 Convergence of the new algorithm

In this section, we discuss convergence issues for the algorithm. The strictest interpretation of the Gibbs sampler is one in which only full conditional distributions are used and in which the order of generation of the parameters is fixed. Even in this seemingly simple setting, there are typically many sets of conditional distributions that may be used as the basis of the Gibbs sampler. Many of these conditional distributions will violate the conditions required for convergence of the sampler to the posterior distribution, even though the posterior distribution is naturally thought of as having conditionals that produce an irreducible, aperiodic chain.

As a simple example consider a nondegenerate, bivariate normal posterior with known covariance matrix. Define the conditional distribution of $X_1|X_2$ to be the appropriate normal distribution for all $X_2 \neq 0$ and to be degenerate at 0 when $X_2 = 0$. Define the conditional distribution of $X_2|X_1$ in

a similar fashion, with the distribution of $X_2|X_1$ degenerate at 0 when $X_1 = 0$. If started at $(0, 0)$, the chain will not converge to the bivariate normal distribution.

These technical details prevent overall statements of convergence such as "implementation of the Gibbs sampler in the general MDP model through the *no gaps* algorithm will provide convergence to the posterior distribution, from every initial condition." Instead, we provide a statement that there is a representation based on a wise choice of conditional distributions for which the algorithm will converge, and we provide guidance on the choice of the conditional distributions to use. The following argument illustrates the reasoning needed to ensure convergence, and it is readily applied to MDP models in the most common settings. The argument relies heavily on results in Tierney (1994). In particular, we present an easy method that allows us to check the absolute continuity condition in his Corollary 3.1. This corollary is reproduced below. Let $P(x, A)$ denote the transition probability in a Markov chain Monte Carlo scheme.

Corollary 3.1 (Tierney, 1994). Suppose P is π-irreducible and $\pi P = \pi$. If $P(x, .)$ is absolutely continuous with respect to π for all x, then P is Harris recurrent.

Tierney's work also shows that when the posterior distribution is proper, the conditions of Corollary 3.1 ensure convergence of the chain (in total variation norm) based on P to the unique invariant distribution, π, for all initial conditions x. In our applications of the MDP model, we begin with a proper prior distribution.

The following argument verifies Tierney's conditions to show convergence of the Markov chains based on the *no gaps* algorithm. The notion behind the argument is that we may easily demonstrate both π-irreducibility and absolute continuity with respect to π of $P(x, .)$ with an examination of the innards of a single Gibbs cycle. To this end, we focus on some set, A, for which $\pi(A) > 0$, with the intent of showing that $P(x, A) > 0$ for each x. Any set A may be represented as a partition $A = \cup_s A_s$, where the elements of the partition are indexed by the configuration vector. Also, the distribution π has a unique representation as $\pi = \sum_s \pi_s$, so that $\pi(A) = \sum_s \pi_s(A_s)$. We have $\pi(A) \geq \pi(A_s) = \pi_s(A_s) > 0$ for some s.

The first stage of the Gibbs cycle in the algorithm involves the generation of a new configuration, s, through a sequence of smaller generations. We assume that the conditional distributions used

are such that there is positive probability of a transition to each vector s which receives positive prior probability. At the second stage, we focus on some s for which $\pi(A_s) > 0$. Conditional on this s, the algorithm relies on a generation from a distribution which we take to be mutually absolutely continuous with respect to π_s, and so the conditional transition to A_s has positive probability. Thus, the overall transition kernel gives positive probability to the transition from x to A.

In the bivariate normal example, it is sufficient to choose families of conditional distributions that are mutually absolutely continuous with respect to Lebesgue measure on R^1 for each of the distributions $X_1|X_2$ and $X_2|X_1$. In the model (1), where the marginal posterior for each θ_i is mutually absolutely continuous with respect to Lebesgue measure on the same subset of R^d, it is enough to choose a family of conditional distributions for each ϕ_i of the *no gaps* model that is mutually absolutely continuous with respect to these marginal posteriors.

Convergence of the algorithm when hyperparameters are included may often be established with an argument similar to the one presented above. A third step is added to the algorithm, with the hyperparameters generated at this step.

When this additional level of hyperparameters is considered, an extra concern arises. At first glance, the generation of $(\phi_E, \nu)|\phi_F$ appears to be a random generation, depending on the results of previous steps in the same cycle of the sampler. Again, a simple example where (X, Y) has a bivariate normal distribution with covariance matrix I shows the dangers of such schemes. Using the "standard" conditional distributions, we may define an algorithm as (i) generate $X|Y$, (ii) generate $Y|X$, and (iii) generate $Y|X$ if $Y < 0$. This algorithm, though based on a sequence of conditional generations, does not yield a chain which converges to the joint distribution of (X, Y). The reason that the standard Gibbs sampling theory does not apply is that the sequence of conditional generations is neither a fixed scan Gibbs sampler nor a random scan Gibbs sampler with a randomization independent of the state of the chain.

To see that the algorithm we propose do not succumb to this difficulty, we provide a perspective for the generation of the hyper-parameters: Conditioning on (s, ϕ_F) partitions the state space. The generation of $(\phi_E, \nu)|(s, \phi_F)$ is just a generation from a conditional distribution defined on this partition. From this viewpoint, we simply return to the basic Gibbs sampler, with a fixed sequence of conditional generations. To ensure convergence for any initial condition, we repeat the argument above, invoking the requisite conditions on the posterior and conditional distributions as needed.

Note that for the example discussed in this paper, these arguments can be formalized. The "standard" choices for conditional distributions lead to algorithms that satisfy the conditions of Tierney's Corollary 3.1.

4 Example: The uniform-normal MDP

We illustrate application of the *no gaps* parameterization in the uniform-normal MDP model:

$$\begin{aligned} y_i &\sim N(\mu_i, \Sigma_i), \quad i = 1, \ldots, n, \\ (\mu_i, \Sigma_i) &\sim G, \quad G \sim DP(\alpha G_0), \\ G_0(\mu, \Sigma) &= U(\mu; \mu_L, \mu_H) \cdot W(\Sigma^{-1}; r, R), \end{aligned}$$

The model is completed by conjugate hyperpriors on the parameters α and R: $R \sim W(q, Q)$ and $\alpha \sim Gamma(a_0, b_0)$.

We estimate the model for a data set from Lubischew (1962). The data records five measurements of physical characteristics for male insects of the species chactocnema concina, chactocnema heikertinger, and chactocnema heptapotamica. We will only use two measurements in this illustration: y_{i1} and y_{i2}, the width of the first and second joint on the i-th beetle. We will use $y_i = (y_{i1}, y_{i2})$ to denote the observation on beetle i, and $y = (y_1, \ldots, y_n)$ to denote the whole data set. There are $n = 74$ observations. Although the classification into the three species was known, this was not used in the estimation of the model. The data are plotted in Figure 1.

Figure 2 shows the posterior predictive $p(y_{n+1}|y)$ for a future observation. This can be thought of as a density estimate for the unknown sampling distribution of beetle joint widths for the given species. The predictive $p(y_{n+1}|y)$ is estimated as an average over conditional predictives: $p(y_{n+1}|y) = \int p(y_{n+1}|\theta, \alpha) dp(\theta, \alpha|y) \approx 1/T \sum_{t=1}^{T} p(y_{n+1}|\theta^t, \alpha^t)$, where $\theta = (\theta_1, \ldots, \theta_n)$ and (θ^t, α^t) are the imputed values after t iterations of the Gibbs sampling scheme. Compare with the comments in Section 3.3 for a discussion of the parameterization used for computing the predictive distribution. For reasons of computational efficiency, the first 200 iterations are discarded, and thereafter only every 10-th iteration is used. Figures 3 through 5 show some more aspects of the posterior distribution on the MDP parameters and the Gibbs sampling scheme.

5 Discussion

We have discussed an augmented parameter model to allow implementation of efficient Gibbs sampling schemes for estimating MDP models. The heart of the augmentation is the explicit representation of θ in terms of (ϕ, s). Placing a distribution on (ϕ, s) induces a distribution on θ. The *no gaps* model ensures that the induced distribution on θ is identical to the distribution specified by the MDP model, and so whichever representation is computationally more convenient may be used to fit the model.

There are many distributions on (ϕ, s) other than the *no gaps* model that induce the MDP's distribution on θ. When the distribution on ϕ consists of independent draws from G_0, we need only create a distribution on s that matches the MDP's distribution over configurations. An easy way to do this is to begin with the simple distribution on s arising from the Polya urn scheme that leads to (2), and then to extend this to a more complex distribution by allowing permutations of the indices for ϕ and by allowing gaps in the sequence of indices, so that some of the k clusters may have indices larger than k.

Although it might at first seem detrimental to expand the distribution on s through introduction of permutations, or by allowing gaps in the sequence of cluster indices, these expansions are actually helpful. In small examples, the deliberate introduction of non-identifiability, as with the permutations for the *no gaps* model, can be demonstrated to speed convergence of the Markov chain to its limiting distribution. The reason for the improvement in convergence is that the individual updates in the Gibbs sampler are allowed to range over a larger set of potentially generated values. Viewed in this fashion, it is essentially this same reasoning that leads to recommendations for marginalizing unneeded parameters from the Gibbs sampler and for generating blocks of parameters all at once. In large problems, the same technique of expanding the distribution on s through natural symmetries in the labeling of the clusters seems to empirically improve the rate of convergence of the Markov chain.

One natural distribution on (ϕ, s), called the *complete* model, is described in the longer technical report version of this work, available from MacEachern and Müller (1994). This model allows one to fit current estimation schemes into the framework developed in this paper. Current estimation schemes based on special cases and approximations are shown to be specific choices of Gibbs scan-

ning schemes, skipping, approximating and/or integrating certain full conditionals of the general *complete* model scheme.

The important contribution of this paper is to provide a formal framework which encompasses all of these Markov chain Monte Carlo algorithms. While the formulation of the algorithm presented here is designed to fit a wide class of models, in many popular models simplifications are both possible and recommended. For example, in the conditionally conjugate normal-normal model we recommend an integration over ϕ when evaluating the multinomial resampling probabilities (4). Another example of efficient specifications for particular models occurs in problems with fixed hyperparameters (or a discrete hyperprior with few possible levels). The probabilities q_0 can then be evaluated before starting the simulation and stored on file.

References

Antoniak, C.E. (1974), "Mixtures of Dirichlet processes with applications to non-parametric problems," *Annals of Statistics*, 2, 1152-1174.

Bush, C.A. and MacEachern, S.N. (1996), "A semi-parametric Bayesian model for randomised block designs," *Biometrika*, 83, 275-286.

Doss, H. (1994), "Bayesian nonparametric estimation for incomplete data via successive substitution sampling," *Annals of Statistics*, 22, 1763 - 1786.

Erkanli, A., Stangl, D.K., and Müller, P. (1993), "A Bayesian analysis of ordinal data," ISDS Discussion Paper 93-A01, Duke University.

Escobar, M.D. (1994), "Estimating normal means with a Dirichlet process prior," *Journal of the American Statistical Association*, 89, 268-277.

Escobar, M.D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577 - 588.

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

Gasparini, M. (1996), "Bayesian density estimation via mixtures of Dirichlet processes", *Journal of Nonparametric Statistics*, 6, 355–366.

Gelfand, A.E. and Kuo, L. (1991), "Nonparametric Bayesian bioassay including ordered polytomous response," *Biometrika*, 78, 657-666.

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

Kuo, L. and Smith, A.F.M. (1992), "Bayesian Computations in survival models via the Gibbs sampler", in *Survival analysis: State of the Art*, ed. Klein, J.P. and Goel, P.K., Dodrecht: Kluwer Academics, pp. 11-24.

Liu, J. (1996), "Nonparametric hierarchical Bayes via sequential imputations," *Annals of Statistics*, 24, 911 - 930.

Lubischew, A. (1962), "On the use of discriminant functions in taxonomy," *Biometrics,* 18, 455-477.

MacEachern, S.N. (1994), "Estimating normal means with a conjugate style Dirichlet process prior," *Communications in Statistics B*, 23, 727-741.

MacEachern, S.N. and Müller, P. (1994), "Estimating Mixture of Dirichlet Process Models", Discussion paper 94-11, ISDS, Duke University.

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian curve fitting using multivariate normal mixtures," *Biometrika*, 83, 67 - 80.

Newton, M., Czado, C, and Chappell, R. (1996), "Semiparametric Bayesian inference for binary regression," *Journal of the American Statistical Association*, 91, 142-153.

Tierney, L. (1994), "Markov chains for exploring posterior distributions (with discussion)," *The Annals of Statistics*, 4, 1701-1762.

Walker, S. and Damien, P. (1996), "Sampling a Dirichlet Process Mixture Model", Technical Report, University of Michigan, Business School.

West, M. (1992), "Hyperparameter estimation in Dirichlet process mixture models," Technical Report92-A03, Duke University, ISDS.

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*, ed. A.F.M. Smith and P. Freeman, New York: Wiley, pp. 363-386

FIGURES

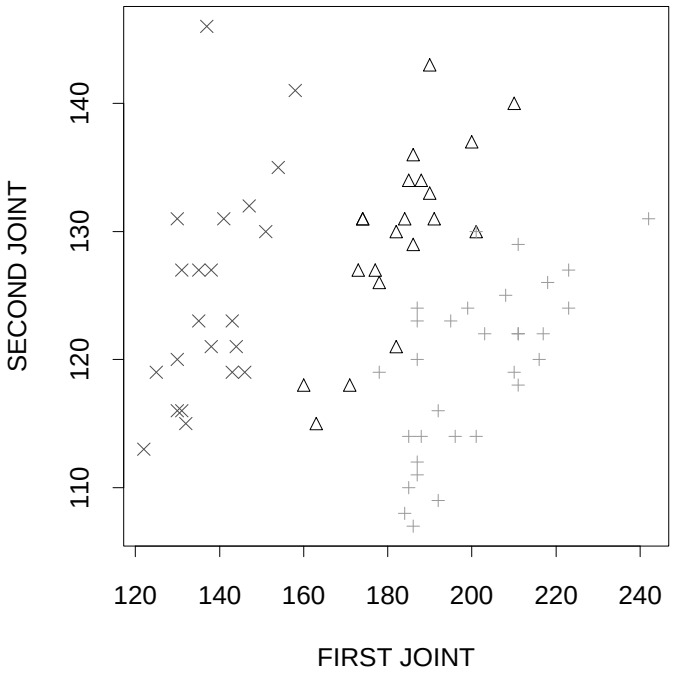

Figure 1: The data. The scatterplot shows a scatterplot of widths for the first (y_{i1}) and second joint (y_{i2}) for 74 beetles. The different plot symbols mark the three different species.

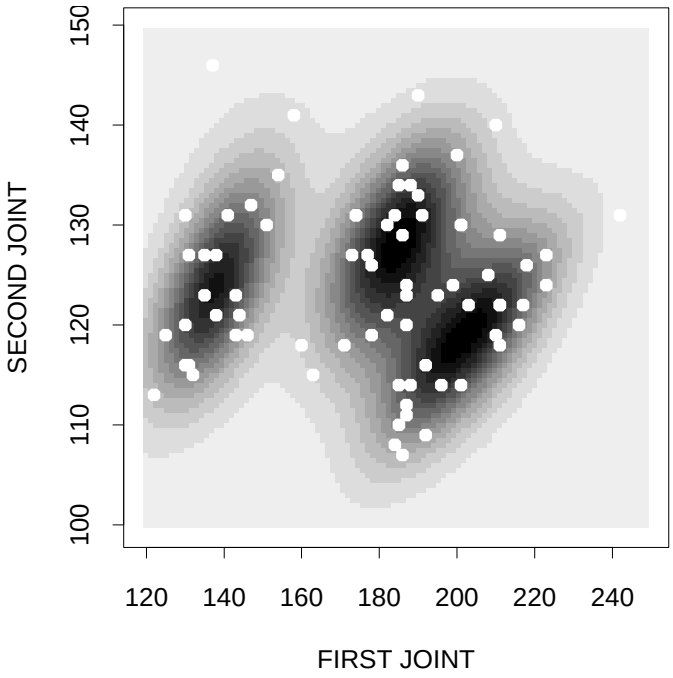

Figure 2: Predictive $p(y_{n+1}|y)$. The white dots show the observations y_i. The posterior predictive $p(y_{n+1}|y)$ can be thought of as a density estimate for the unknown sampling distribution of beetle joint widths for the given species. The format of the density estimate is similar to a conventional kernel density estimate. It is a mixture of normal kernels. However, the density estimate is model based, allows distinct correlation matrices for each normal term, and mixes over hyperparameters like the number of normal terms k, the prior parameters for cluster location (m and B) and the hyperparameters for cluster covariance matrices (Q and R).

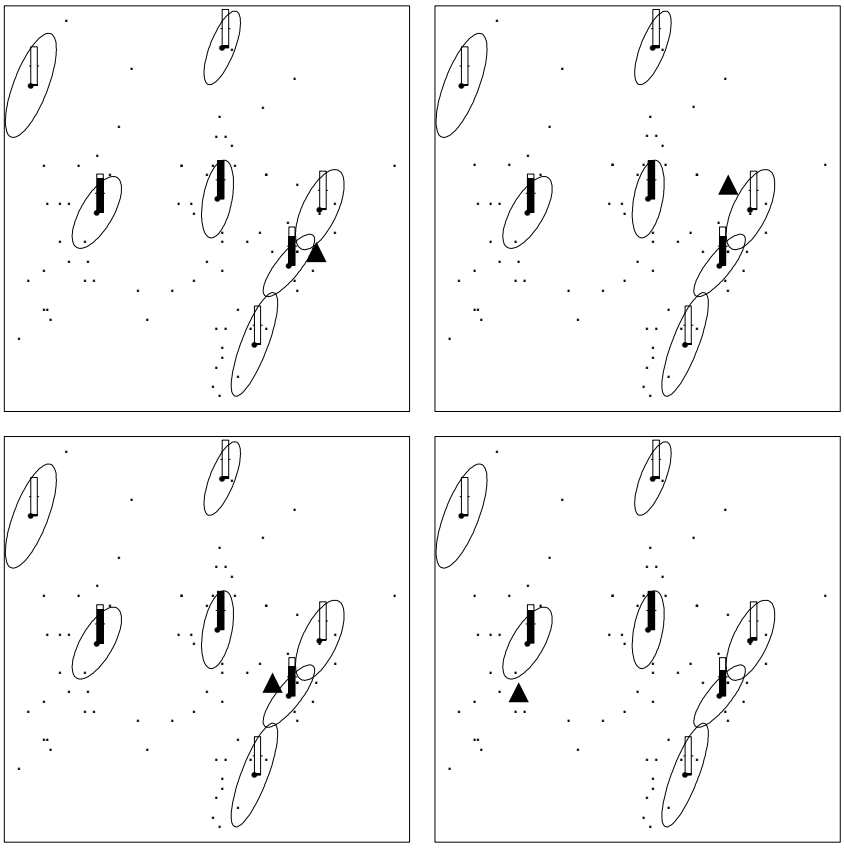

Figure 3: Cluster and relative weights at iteration 500. The four panels show cluster locations μ_j^ϕ (solid dots), covariance matrices Σ_j^ϕ (lines of constant Mahalanobis distance equal 0.5 from μ_j^ϕ) and cluster sizes n_j (thermometers) for the clusters as they are when resampling s_i for points $i = 27$, 31, 70 and 33 (clockwise from top). The solid triangles indicate points y_{27}, y_{31}, y_{70} and y_{33} respectively. The thin dots plot all other data points. Notice that in all four figures there are three clusters which take almost all weight, i.e. n_j is negligible for the remaining clusters compared to these three clusters. The three dominant clusters correspond roughly to the three beetle species in the data.

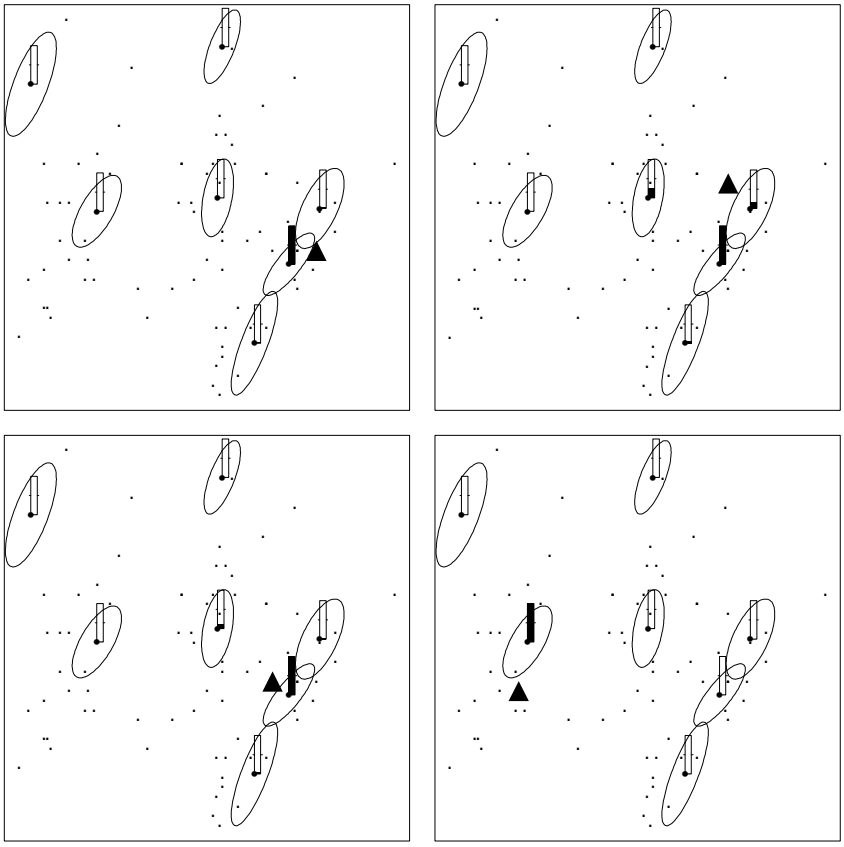

Figure 4: Probabilities for resampling the configuration indicators s_i. The figure shows for the same plots as the the previous figure, except that instead of the cluster sizes, the probabilities $\pi_j = Pr(s_i = j|\ldots)$ are plotted in the "thermometers". Notice, for example, in the first panel, that point y_{27} could be attributed to each of the three neighboring clusters with reasonably large probabilities.

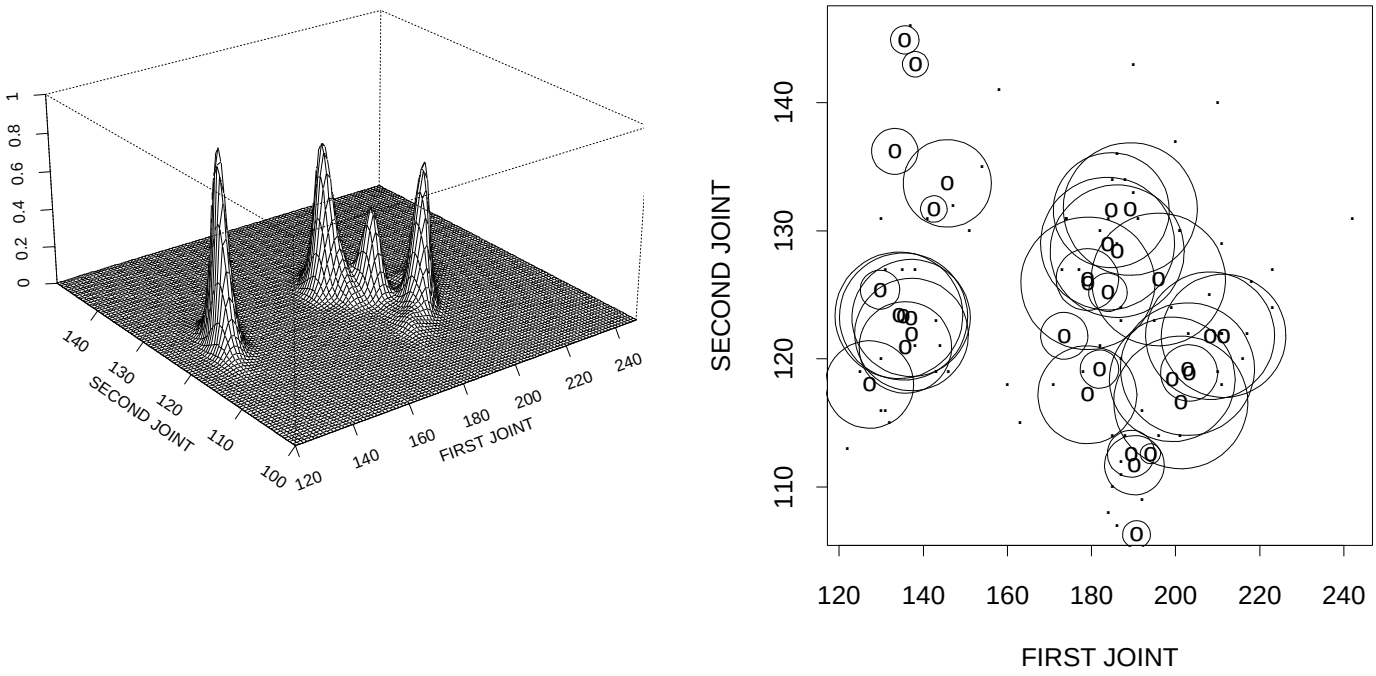

Figure 5: Some elements of the posterior distribution for the cluster locations μ_j. Panel (a) shows the posterior predictive distribution for a new latent variable μ_{n+1}. The posterior predictive for y_{n+1} shown in Figure 2 was the convolution of $p(\mu_{n+1}|y)$ with $p(y_{n+1}|\mu_{n+1}, y)$. The predictive distribution of μ_{n+1} shows the location of the three dominant clusters even more clearly than $p(y_{n+1}|y)$. Notice the fourth peak in between the two other modes on the right half of the plot. It is probably due to a combination of the two neighboring clusters corresponding to the two higher peaks. Panel (b) shows the sample of cluster locations μ_j^{ϕ} sampled at iterations 300, 1000, 2000, 3000, 4000 and 5000. Each circle corresponds to one cluster. The center indicates μ_j^{ϕ}. The area of the circle is proportional to the weight n_j of the cluster.