Computational methods for mixture of Dirichlet process models

Steven N. MacEachern

ABSTRACT This chapter lays out the basic computational strategies for models based on mixtures of Dirichlet processes. I describe the basic algorithm and give advice on how to improve this algorithm through a collapse of the state space of the Markov chain and through blocking of variates for generation. The computational methods are illustrated with a beta-binomial example and with the bioassay problem. Some advice is given for dealing with models that have little or no conjugacy present.

Key words and phrases: Dirichlet process, Gibbs sampler, MCMC, Rao-Blackwellization.

1 Introduction

The mixture of Dirichlet processes (MDP) model is one of extraordinary usefulness and generality. In its most common use, the model allows one to mimic a Bayesian, parametric hierarchical model a la Lindley and Smith (1972), but to relieve the strictures imposed by the parametric assumptions. A basic parametric model is

$$\begin{aligned} X_i|\theta_i, z_i &\sim F_{\theta_i, z_i}(.) \\ \theta_i|\nu &\sim G_{0,\nu} \\ \nu &\sim H_\nu(.) \end{aligned}$$

where $i = 1, \ldots, n$. The usual convention in writing a hierarchical model is followed, so that observations at a stage in the hierarchy are independent, and dependence upon all higher stages of the hierarchy is explicitly stated. Thus, for example, the $\theta_i|\nu$ are independent and identically distributed, from $G_{0,\nu}$.

As a specific example, the reader is encouraged to consider the simple normal theory hierarchical model. In this model, the data (the X_i), given the parameters and covariates, are normally distributed. The mid-level

parameters (the θ_i), given the hyperparameter, are normally distributed. The hyperparameter (ν) is often assigned a conjugate prior distribution. To formally match the notation, there are no covariates z_i, $\nu = (\mu, \tau^2)$, $F_{\theta_i, z_i} = N(\theta_i, \sigma^2)$, $G_{0,\nu} = N(\mu, \tau^2)$, and $H_\nu = N(\mu_0, \tau_0^2) \cdot IG(\alpha_0, \beta_0)$. In this description, $N(a, b)$ represents the normal distribution with mean a and variance b, and $IG(a, b)$ represents the inverse gamma distribution with shape a and scale b (hence mean $(b(a-1))^{-1}$ and variance $(b^2(a-1)^2(a-2))^{-1}$, provided these quantities exist). A model with unknown σ^2 would necessitate placing a distribution on σ^2, with a conjugate choice being an inverse gamma distribution.

In this simple hierarchical model, the θ_i are tied together through the portion of the model involving ν. Each of the θ_i contains information about ν which, in turn, provides information about $G_{0,\nu}$, and hence θ_j, $j \neq i$. Since θ_i contains information about θ_j, it will have an impact on inferences about θ_j. The model is appropriate when the θ_i are parameters for similar phenomena. See the example in section 3.1 for a case where this modelling strategy, up to the form of the likelihood, is reasonably appropriate.

The presence of the z_i greatly broadens the scope of the hierarchical model. It allows one to include case-specific covariates, that, in more complex models allow for such features as X_i which arise from varying sample sizes, a regression structure which is separate from the portion of the model described above (and which would include a prior distribution over additional parameters), an indication of censoring, or qualitatively varying forms for F_{θ_i, z_i}. This last allows for the explicit combination of data that are observed at different resolutions or which arise from different sampling plans.

The MDP analog of the parametric model is

$$\begin{aligned} X_i|\theta_i, z_i &\sim F_{\theta_i, z_i}(.) \\ \theta_i|G &\sim G \\ G|M, \nu &\sim Dir(MG_{0,\nu}) \\ \nu &\sim H_\nu(.) \\ M &\sim H_M(.) \end{aligned}$$

where $i = 1, \ldots, n$. Most of the components of the MDP model match its parametric analog. One novel component is $G|(M, \nu) \sim Dir(MG_{0,\nu})$. The distribution on G is a Dirichlet process with base measure α. As is commonly done, the mass of the base measure, a positive quantity, is denoted by M. $G_{0,\nu}$ is a distribution function that is proportional to the measure α. A second novel component, $H_M(.)$, is the distribution over the mass of the base measure. The prior distribution has traditionally taken M and ν to be independent, as is assumed in this chapter, although this need not be the case.

The difference between the two models lies in the middle stage of the hierarchy, where, to create the MDP model, $(\theta_i|\nu) \sim G_{0,\nu}$ has been ripped out

to be replaced by the nonparametric $(\theta_i|G) \sim G$; $(G|M,\nu) \sim Dir(MG_{0,\nu})$; $M \sim H_M(.)$. It is this alteration that results in the additional flexibility of the MDP model, enabling the model to adapt to a wide variety of data.

The Dirichlet process provides a means of placing a distribution on the space of distribution functions. It satisfies the two goals outlined in Ferguson's (1973) seminal paper of having large support and being easy to work with. The support of the Dirichlet process is large, and, with appropriate choice of $G_{0,\nu}$ can be the set of all distributions on, say, the real line. With such large support, the distribution from which the θ_i are drawn is no longer restricted to lie in the set of distributions $G_{0,\nu}$, indexed by ν. This set, in a parametric hierarchical model, may be very small–say the set of normal distributions. The switch to the MDP model allows the data to adapt to a G that is skewed, has "shoulders", is multimodal, or departs from the parametric form $G_{0,\nu}$ in less evident ways.

The Dirichlet process is easy to work with when the hyperparameters, ν and M, are fixed and the θ_i are observed directly. In this instance, the distribution on G, after observing $\theta_1, \ldots, \theta_{n-1}$, would also be a Dirichlet process, with measure $MG_{0,\nu} + \sum_{i=1}^{n-1} \delta_{\theta_i}$, with δ_{θ_i} representing a point mass, of measure 1, at θ_i. The distribution for $\theta_n|(M,\nu,\theta_1,\ldots,\theta_{n-1})$ is just a rescaled version of this updated base measure: The distribution for $\theta_n|(M,\nu,\theta_1,\ldots,\theta_{n-1})$ can be written as a mixture, with $\theta_n \sim G_{0,\nu}$ with probability $M/(M+n-1)$ and θ_n set equal to θ_i, $i = 1,\ldots,n-1$ with probability $1/(M+n-1)$. Since the $\theta_i|G$ are independent and identically distributed, they are exchangeable, and so one can rewrite the conditional distribution for any $\theta_i|(M,\nu,\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_n)$ as the mixture $\theta_i \sim G_{0,\nu}$ with probability $M/(M+n-1)$ and θ_i set equal to θ_j, for $j = 1,\ldots,i-1,i+1,\ldots,n$ with probability $1/(M+n-1)$. It is this representation of the conditional distribution of θ_n, which is most easily seen from the Polya urn scheme representation of the Dirichlet process (Blackwell and MacQueen, 1973), that lies both at the heart of the MDP modelling strategy and at the heart of the modern computational methods for fitting the MDP model.

To see the parallel between the MDP model and the parametric model that gives rise to it, consider the marginal distribution of a θ_i, say θ_1. As seen from the previous description, the distribution of $\theta_1|(M,\nu)$ is just $G_{0,\nu}$. With $n = 1$ the middle stage of the hierarchy collapses, and the two models are identical. The impact of the MDP model is seen when $n > 1$. In this instance, the θ_i divide into $k \leq n$ clusters, with values of θ_i within a cluster identical and with common cluster locations independent draws from $G_{0,\nu}$. It is this description of the clustering of the vector θ that leads to the view of the MDP model as a mixture model.

While the conditional distributions are tractable, the entirety of the MDP model is analytically intractable for all but the simplest cases or very small data sets, due to a combinatorial explosion in the number of possible cluster structures. As a result, one must turn to approximation in order to fit the model. With a modern computational environment, this most often

means that a simulation based approximation is chosen. The early work on simulation methods was done by Kuo (1986) who developed an importance sampler for Dirichlet process based models, and Escobar (1994) who developed the first Markov chain Monte Carlo methods for fitting Dirichlet process based models. Further work has improved Escobar's initial algorithm and extended the methods to deal with the full MDP model (Escobar and West, 1995; MacEachern, 1994). Modifications have produced Markov chains that have been empirically shown to move through the parameter space more quickly and that naturally suggest more accurate estimators. These current algorithms are the focus of this chapter.

2 The Basic Algorithm

A basic computational algorithm that can be applied to a wide variety of problems extends Escobar's (1994) algorithm. The third step in this algorithm is due to Bush and MacEachern (1996). A nice description of this algorithm appears in West, Müller and Escobar (1994). This algorithm, still a Gibbs sampler (Gelfand and Smith, 1990) in the loose sense of the term, where parameters are generated from a sequence of conditional distributions, is easily programmed for the MDP model. It involves little effort beyond writing the Gibbs sampler for the corresponding parametric model.

In order to describe the algorithm, some additional notation is introduced. The vector θ^* consists of those elements of θ that lie in distinct clusters. It is of length $k \leq n$. The vector $s = (s_1, \ldots, s_n)$ indicates in which cluster each of the θ_i lies. If $s_i = j$, then θ_i lies in cluster j, and $\theta_i = \theta^*_j$. The number of s_i such that $s_i = j$ is denoted by n_j. In the algorithms below, a value will be generated for θ_i, conditional on θ_{-i}, the other parameters, and the data. When performing such a conditioning, the superscript $-$ denotes the fact that θ_i is left out of the conditioning. Thus n_j^- represents the number of θ's in cluster j, not counting θ_i, and k^- represents the number of clusters when θ_i is removed. If removal of θ_i shrinks the number of clusters, so that $k^- < k$, then relabel the k^{th} cluster as cluster s_i: Set $s_j = s_i$ for all j with $s_j = k$, and set $\theta^*_{s_i} = \theta^*_k$. Since there are now only $k-1$ clusters, remove θ^*_k.

The generations from the conditional distributions are

- Generate $\nu|\theta^*$. For this conditional distribution, ν has density proportional to $h(\nu)\prod_{i=1}^{k} g_{0,\nu}(\theta^*_i)$.
- Generate $(s_i, \theta_i)|(\nu, M, s_{-i}, \theta_{-i}, X)$. This distribution is derived by taking each θ_i, in turn, to be the last parameter arising from a Polya urn scheme. The parameter θ_i comes from cluster j with probability

$$q_j \propto (\frac{n_j^-}{M+n-1})(f_{\theta^*_j, z_i}(x_i)),$$

for $j = 1, \ldots, k^-$. It comes from cluster $k^- + 1$ with probability

$$q_0 \propto (\frac{M}{M+n-1})(\int f_{\tilde{\theta},z_i}(x_i) g_{0,\nu}(\tilde{\theta}) d\tilde{\theta}).$$

In the sequel, the two terms in these probabilities are referred to as **1** and **2**, respectively. Term **1** comes from the prior probability of θ_i joining an established cluster or beginning a new cluster under the Polya urn scheme. Term **2** accounts for the appropriate marginal density of $X_i|\theta_j^*$.

In the event that $s_i = k^- + 1$, a value is needed for the newly created θ_k^*. This value is drawn from the distribution with density proportional to $g_{0,\nu}(\theta_k^*) f_{\theta_k^*,z_i}(x_i)$.

This step is implemented, for $i = 1, \ldots, n$. The values of parameters are updated, immediately upon their generation.

- Generate $\theta_i^*|(\nu, X)$. For this conditional distribution, θ_i^* has density proportional to $g_{0,\nu}(\theta_i^*) \prod_{j|s_j=i} f_{\theta_i^*,z_j}(x_j)$.

  This step is implemented for $i = 1, \ldots, k$.

If a distribution is placed on M, its value may be generated as an additional stage. Alternatively, the distribution of M may be removed through integration, as indicated in the next section on the collapse of the state space.

3 More Efficient Algorithms

There are a number of techniques that are known to improve the performance of Gibbs samplers in many problems. When regularity conditions are satisfied, these techniques increase the rate of convergence of the Markov chain that underlies the Gibbs sampler to its limiting distribution. The improvement in rate of convergence may be qualitative, with the rate moving from geometric to uniform, for example. Such improvement can be more useful than merely obtaining better constants for the same qualitative rate.

Two of the most popular (and easiest to implement) techniques are that of collapsing the state space and that of blocking several parameters together. The general heuristic that motivates both of these techniques arises from consideration of a single generation in the Gibbs sampler. To illustrate the heuristic, its application to a model with three parameters is examined. Call the parameters θ_1, θ_2, and θ_3. There are three Markov chains to compare. One of the chains is described by the generations $\theta_1|(\theta_2, \theta_3)$, $\theta_2|(\theta_1, \theta_3)$ and $\theta_3|(\theta_1, \theta_2)$, the second is described by the generations $(\theta_1, \theta_2)|\theta_3$, $(\theta_1, \theta_2)|\theta_3$ and $\theta_3|(\theta_1, \theta_2)$, and the third is described by the generations $(\theta_1, \theta_2)|\theta_3$, $(\theta_1, \theta_2)|\theta_3$ and $(\theta_1, \theta_3)|\theta_2$.

Compare the first two chains, generation by generation. For the first generation, chain 2 is preferable to chain 1. The values of both θ_1 and θ_2 are updated instead of merely updating θ_1, and this enables us to move closer to the joint posterior distribution of $(\theta_1, \theta_2, \theta_3)$. For the second generation, chain 2 again is preferable, since θ_1 and θ_2 are both updated instead of merely updating θ_2. The third generations are equivalent. Since, for each generation, chain 2 is at least as good as chain 1 and sometimes better, there is strong heuristic evidence that chain 2 is better.

Compare the second and third chains, generation by generation. The first two generations are identical. For the third generation, chain 3 appears preferable since both θ_1 and θ_3 are updated instead of merely updating θ_3. Following the earlier reasoning, chain 3 is preferable to chain 2.

The heuristic can be formalized by taking a fixed initial distribution for the state of the Markov chain and measuring the distance between the distribution on the state after a single generation and the limiting distribution of the chain. A chain appears preferable to another if it results in a smaller total variation distance after a single generation. Additional work yields theorems about which chain, as a complete collection of generations, is preferable, although the conditions are non-trivial.

The three chains described above are a standard Gibbs sampler, a Gibbs sampler that uses blocking, and a Gibbs sampler that collapses the state space. In practice, the implementation of chains 2 and 3 would be simplified: The redundant generation of $(\theta_1, \theta_2)|\theta_3$ would be eliminated from both chains. Additionally, θ_1 wouuld be entirely omitted from chain 3, leaving only the two generations $\theta_2|\theta_3$ and $\theta_3|\theta_2$. See Liu (1994) and Liu, Wong and Kong (1994) for theory on both collapse of the state space and blocking, and MacEachern (1994) for theory on the collapse of the state space. Subsections 3.1 and 3.2 illustrate the use of collapsing and blocking with the MDP model.

3.1 Collapsing

The collapse of the state space for a Gibbs sampler, illustrated with the chains above, exploits the idea that, since one is only simulating to avoid a difficult integration, one should perform as much integration as possible before simulating. As in the small example's chain 3, one can view any Gibbs sampler for which a parameter is completely marginalized (integrated out) as a heuristically improved Gibbs sampler where the marginalized parameter is generated, in a block, with each other parameter. This subsection illustrates how the collapse applies, in two different fashions, to the MDP model. A formula which is helpful in the following development is $P(\theta) = M^k \prod_{j=1}^{k} [(n_j - 1)! g_{0,\nu}(\theta_j^*)] / [\prod_{i=1}^{n} (M + i - 1)]$. The formula implicitly connects values of θ on the left side of the equation with values of θ^* on the right side of the equation. It may be derived from the Polya urn

scheme representation of the Dirichlet process. When using the formula, one must, at times, be careful to work with the appropriate dominating measure.

Preintegration, or how to marginalize the mass parameter, M

The first part of the collapse is illustrated with a pre-integration over the distribution of the mass parameter, M. I first show how the integration works in a model with no data, and then apply the result to the model with data. Considering generation of $(M, \theta_i)|(\nu, s_{-i}, \theta_{-i})$, the Polya urn scheme leads to,

$$h_{M|(\nu,s_{-i},\theta_{-i})}(M) \quad = \quad \frac{h_M(M)M^{k^-}/\prod_{j=1}^{n-1}(M+j-1)}{\int h_M(M)M^{k^-}/\prod_{j=1}^{n-1}(M+j-1)dM}$$

and for $\theta_i|(M, \nu, s_{-i}, \theta_{-i})$ the mixture described in the introduction. Taking the product of these two distributions, one has

$$P(s_i = j|\nu, s_{-i}, \theta_{-i}) = n_j^- \int h_{M|(\nu,s_{-i},\theta_{-i})}(M)\frac{1}{M+n-1}dM \quad \mathbf{1A}$$

for $j = 1, \ldots, k^-$, and

$$P(s_i = k^- + 1|\nu, s_{-i}, \theta_{-i}) = \int h_{M|(\nu,s_{-i},\theta_{-i})}(M)\frac{M}{M+n-1}dM. \quad \mathbf{1B}$$

As before, if $s_i = k^- + 1$, the value of θ_i is drawn from $G_{0,\nu}$.

Evaluation of **1A** and **1B** can can be done quickly, to any degree of accuracy desired, before the simulation is begun, justifying the terminology pre-integration. When working with data, simply substitute **1A** for n_j^- and **1B** for M in step 2 of the basic algorithm.

How to marginalize the θ_i

The second technique relies on a conjugate structure for a portion of the parametric model. If the form $\int f_{\theta,z_i}g_{0,\nu}(\theta)d\theta$ is easily integrable, as is often the case when one chooses the basic algorithm, the mixing of the Markov chain can be improved with another collapse. This collapse, however, is dynamic, proceeding as the algorithm moves along. To accomplish this collapse, one integrates over the cluster locations. To do this, replace the marginal probabilities, $f_{\theta_j^*,z_i}(x_i)$, in step 2 of the basic algorithm with **2A**

$$\int f_{\theta_j^*,z_i}(x_i)g_{0,\nu,s_{-i},x}(\theta_j^*)d\theta_j^*, \quad \mathbf{2A}$$

where

$$g_{0,\nu,s_{-i},x}(\theta_j^*) = \frac{g_{0,\nu}(\theta_j^*)\prod_{l\neq i|s_l=j} f_{\theta_j^*,z_l}(x_l)}{\int g_{0,\nu}(\theta_j^*)\prod_{l\neq i|s_l=j} f_{\theta_j^*,z_l}(x_l)d\theta_j^*}$$

is the posterior density of θ_j^*, calculated as if it had prior density $g_{0,\nu}$ and updated with the data attached to cluster j. Make this replacement for calculation of $q_1, \ldots, q_{k-}$. In the evaluation of q_0, **2A** matches the term $\int f_{\tilde{\theta},z_i}(x_i) g_{0,\nu}(\tilde{\theta}) d\tilde{\theta}$, and so no modification is needed. With conjugate forms, the integral **2A** reduces to calculation of a normalizing constant for a conjugate pair.

This collapse can be used in multivariate as well as univariate settings. The formulas for the collapse remain the same, with integrals taken over more than one dimension. A particularly common multivariate setting is the multivariate normal, where results on quadratic forms produce the normalizing constant. The normalizing constants can be programmed directly, with little effort.

Occasionally, even more integration can be performed, and the hyperparameter ν can be marginalized. Such an integration requires a great deal of conjugacy. An example is the simple hierarchical MDP model with normal likelihood, normal base measure, and normal distribution for ν–and with known variances. A second example is the conjugate normal hierarchical model where variances are unknown but the variances at different stages are tied together.

The baseball data

As a light-hearted example, I analyze the baseball batting average data of Efron and Morris (1975). The data set consists of the number of hits for players in the major leagues having exactly 45 at bats, as recorded in the April 29, 1970 edition of the *New York Times*. The goal of Efron and Morris' analysis is to predict the batting average, for each player, for the remainder of the season. They employ an empirical Bayesian analysis to produce predictions that are found to be much superior to predictions based on maximum likelihood. Here, an analysis based on the MDP model is used to produce predictions.

$$
\begin{aligned}
X_i|(\theta_i, z_i) &\sim Binomial(z_i, \theta_i) \\
\theta_i|G &\sim G \\
G|(M, \nu) &\sim Dir(MG_{0,\nu}), \text{ with } G_{0,\nu} \text{ the } beta(\alpha, \beta) \text{ distribution} \\
\alpha/(\alpha+\beta) &\sim beta(\nu_1, \nu_2) \\
\alpha+\beta &= 216.6 \\
M &\sim lognormal(2.81, 1.186^2)
\end{aligned}
$$

The model follows the simple hierarchical MDP model with F_{θ_i,z_i} the binomial(z_i, θ_i) distribution. In this instance, $z_i = 45$ for all i. The base measure, G_0, has a beta(α, β) shape. The parameters α and β will have a prior distribution chosen after a peek at the data. Given α and β, the

prior mean for θ_i is $\alpha/(\alpha+\beta)$. The prior mean for X_i/z_i, integrating over θ_i, is also $\alpha/(\alpha+\beta)$. Hence, I set the prior mean for $\alpha/(\alpha+\beta)$ equal to the overall batting average of the 18 players: $\sum_{i=1}^{18} X_i / \sum_{i=1}^{18} z_i = .265$. For convenience, I chose a beta (ν_1, ν_2) prior distribution for $\alpha/(\alpha+\beta)$. To find the two parameters of this prior distribution, I set $\nu_1/(\nu_1+\nu_2) = .265$, and selected $\nu_1+\nu_2$ to produce a sample standard deviation that matches the value $.265 * .735/810$ (where $810 = \sum_{i=1}^{18} z_i$). This results in $\nu_1+\nu_2 = 811$. Focusing on the marginal distribution for θ_i, I set the standard deviation to .030, creating a mean ± 2 standard deviation interval of width .120. To obtain such an interval, with $\alpha/(\alpha+\beta)$ fixed at .265, requires $\alpha+\beta = c = 216.6$. As a consequence of the prior distribution on the hyperparameter, the actual marginal prior distribution of θ_i will show some additional variability. The prior distribution on M was taken to be lognormal (μ, σ^2). The values of μ and σ were chosen to produce a mean of approximately 12 clusters, with a .1 chance of 6 or fewer clusters. The specific values chosen were $\mu = 2.81$ and $\sigma = 1.186$.

With the prior distribution specified, the analysis proceeds through simulation. Conditional generations for the Gibbs sampler are

- Generate $s_i|(s_{-i}, \alpha, X)$ from the distribution with $q_j \propto \tilde{n}_j^- \frac{B(\alpha_j+X_i-1, \beta_j+z_i-X_i-1)}{B(\alpha_j,\beta_j)}$, $j = 1, \ldots, k^-$, and with $q_0 \propto \tilde{M} \frac{B(\alpha+X_i-1, \beta+z_i-X_i-1)}{B(\alpha,\beta)}$, where $\alpha_j = \alpha + \sum_{l\neq i|s_l=j} X_l$ and $\beta_j = \beta + \sum_{l\neq i|s_l=j}(z_l - X_l)$ for $j = 1, \ldots, k^-$, and where $\tilde{n}_j^-$ and $\tilde{M}$ are derived from the preintegration in **1A** and **1B**. Note that neither M nor θ_j^* appears in the formula for q_j. This generation relies on both collapses described earlier in this section.

  Perform this generation for $i = 1, \ldots, n$, updating the vector s immediately upon generation of each s_i.

- Generate $\theta_j^*|(s, \alpha, X)$ from the beta $(\alpha + \sum_{l|s_l=j} X_l, \beta + \sum_{l|s_l=j}(z_l - X_l))$ distribution, for $j = 1, \ldots, k^-$.

- Generate $\alpha|(\theta, s)$ from its posterior distribution. The density is proportional to $B(\alpha, c-\alpha)^{-k} \theta_i^{*\alpha-1}(1-\theta_i^*)^{c-\alpha-1}\alpha^{\nu_1-1}(c-\alpha)^{\nu_2-1}$.

The Gibbs sampler described above was initialized with $s = (1, 2, \ldots, 18)$, $\theta_i^* = X_i/n_i$, and $\alpha = .265c$. The sampler was run for 10,000 iterations, with the first 1,000 iterations discarded as burn-in. Table 1 contains a summary of the data and estimates, rounded to the traditional three decimal places. Conservative values for the standard errors of the estimates, calculated with the batch means method (batches of size 500) range from .00023 to .00048. Consequently, the estimates are expected to be within .001 of the actual estimate under the model. Note that the strong prior used for the MDP analysis leads to considerably more shrinkage than does the empirical Bayes analysis. For the final criterion of predicting the batting average for the rest of the season, the MDP analysis results in a sum of squared error of

prediction of 0.021388 while the empirical Bayes analysis results in a sum of squared error of 0.021611. The difference, on the order of one percent of the sum of squared error, is small, but favors the MDP analysis.

TABLE 1.1. The baseball data, the Stein estimate and an MDP estimate.

Player	X/z	Rest of Season	Stein	MDP
1	.400	.346	.290	.286
2	.378	.298	.286	.283
3	.356	.276	.281	.279
4	.333	.222	.277	.276
5	.311	.273	.273	.273
6	.311	.270	.273	.273
7	.289	.263	.268	.269
8	.267	.210	.264	.266
9	.244	.269	.259	.262
10	.244	.230	.259	.262
11	.222	.264	.254	.259
12	.222	.256	.254	.259
13	.222	.303	.254	.259
14	.222	.264	.254	.259
15	.222	.226	.254	.259
16	.200	.285	.249	.255
17	.178	.316	.244	.252
18	.156	.200	.239	.248

3.2 Blocking

The use of blocking in MDP models is more specialized than the use of the collapse. Its main application is with models that involve censored data, and it is this type of model that is used to illustrate the computational technique. To set up the technique, consider the case where all observations are right censored. Let $z_1 \geq \ldots \geq z_n$ denote the censoring times, ordered from greatest to least. In this instance the i^{th} survival time, say θ_i, is known to exceed z_i, but its exact value is not known. Take the model to be $G \sim Dir(MG_0)$, with $\theta_i|G \sim G$ for $i = 1, \ldots, n$. The data part of the model, since all observations are right censored, is just a collection of indicators that guarantee $\theta_i > z_i$ for each i. The following algorithm generates $(\theta_1, \ldots, \theta_n)$ directly from its posterior distribution. See MacEachern (1992) for a brief description.

- Let $G_{0;z}$ denote the conditional distribution function defined as $G_{0;z}(t) = 1 - G_0(t)/G_0(z)$ for $t \geq z$, and defined as $G_{0;z}(t) = 0$ for $t < z$. Generate the θ_i sequentially. For $i = 1, \ldots, n$, draw $\theta_i \sim G_{0;z_i}$ with

probability proportional to $M(1-G_0(z))$, and set $\theta_i = \theta_j, j < i$ with probability proportional to 1.

The algorithm can be extended to include uncensored observations as well. In this case, exact survival times are first absorbed into the base measure to create an effective base measure. If, for example, there are an additional m observations with exact survival times $\theta_{n+1}, \ldots, \theta_{n+m}$, the effective base measure is $MG_0 + \sum_{i=n+1}^{n+m} \delta(\theta_i)$, where $\delta(\theta_i)$ is a point mass of size one at the point θ_i. Generation from the posterior is accomplished using the algorithm described above, with the effective base measure substituted for the base measure.

The accuracy of the algorithm can be verified by writing the joint distribution of $(\theta_1, \ldots, \theta_n)$ as the mixture distribution induced by the Polya urn scheme, multiplied by the indicators due to the censoring times. An alternate, more intuitive derivation relies on a redistribution of mass argument.

As with the collapse of the state space, the fragment of an algorithm described above is worked into a larger problem. The problem described below, the bioassay problem, is one of the earliest proposed uses for the hierarchical MDP model, and one that I studied intensively while writing my dissertation (MacEachern, 1988). The problem may be viewed as one involving doubly censored data, where each observation is either left censored or right censored, or it may be considered the simplest example of a logistic regression problem, where there is a single covariate.

The bioassay problem

The goal of the bioassay problem is to learn about the dose-response relationship in a population. As an example, a drug to reduce hypertension could be administered in various doses. In order to determine an effective dose, one would need to collect data on the relationship between reduction in hypertension and dose of the drug. Each individual in a study would be assigned some dose of the drug. A sufficient reduction in blood pressure for an individual would be considered a success, and that person would be said to have responded to treatment. Individuals without a sufficient reduction in blood pressure would fail to respond.

The impact of the drug on members of the population is summarized by a cumulative distribution function, say $G(z)$. $G(z)$ gives the proportion of the population that would respond to dose z. This distribution, often called the dose-response curve, is used to find doses to which varying fractions of the population respond: At least half the population will respond to the median of G. The dose producing this response fraction is variously called the effective dose 50 (the ED50) or the lethal dose 50 (the LD50), depending on the nature of the response.

To fit the bioassay problem into the formal MDP model, one more quantity is needed: The smallest dose which would induce an individual to respond. This dose is called the tolerance of the individual. The tolerance is

a latent variable and is not directly observed for any individual. However, some information is collected about the tolerance, in that after being assigned a dose, an individual will either be known to have a tolerance that does not exceed the dose (i.e., the individual responds, and the tolerance is known to be left censored) or a tolerance that exceeds the dose (i.e., the individual does not respond, and the tolerance is known to be right censored). Thus, the data are binary responses. The tolerances are introduced as latent variables, as an aid to simulation.

In the example to come, the unknown distribution function is modelled with an MDP model that is patterned after the logistic regression model. Formally, z_i represents the dose assigned to individual i. The individual's tolerance is θ_i. A right censored observation is denoted by $X_i = 0$ and a left censored observation is denoted by $X_i = 1$. The shape of the base measure is the logistic distribution, with $G_{0;a,b}(x) = (1+exp(-a(x-b)))^{-1}$, where b is the location parameter and $a > 0$ is the reciprocal of a scale parameter. The above discussion, coupled with a specific choice for the prior distribution for the coefficients of the logistic distribution, results in

$$\begin{aligned} X_i|(\theta_i, z_i) &= \begin{cases} 1 & \text{if } \theta_i \le z_i \\ 0 & \text{if } \theta_i > z_i \end{cases} \\ \theta_i|G &\sim G \\ G|(M, a, b) &\sim Dir(MG_{0,a,b}) \\ b &\sim N(\mu_0, \tau^2) \\ a &\sim H(a) \end{aligned}$$

where $i = 1, \ldots, n$. $H(a)$ is a distribution chosen to introduce a uniform(a_0, b_0) distribution on the standard deviation of a variate with distribution G_0. As a consequence, a has a density proportional to a^{-2} over its range. To make the algorithm easier to describe, the data are reordered so that the first m observations have $X_i = 0$ and the last $n-m$ observations have $X_i = 1$. As a further convention, the first m observations have doses listed in decreasing order, $z_1 \ge \ldots \ge z_m$, and the last $n - m$ observations have doses listed in increasing order, $z_{m+1} \le \ldots \le z_n$. Since the θ_i are exchangeable, the only impact of this reordering is a relabelling of the cases. The essence of the posterior is not affected.

The Gibbs sampler uses blocking in two stages:

- Generate $(\theta_1, \ldots, \theta_m)|(\theta_{m+1}, \ldots, \theta_n, a, b, X)$. Absorb the left censored observations into the base measure, so that an effective base measure is defined for this step. The effective base measure is $\alpha + \sum_{i=m+1}^{n} \delta(\theta_i)$ and hence the effective M is $M + n - m$. The effective G_0 is proportional to the effective base measure. With this set-up, the algorithm fragment described above is used to generate the block of parameters $(\theta_1, \ldots, \theta_m)$.

- Generate $(\theta_{m+1}, \ldots, \theta_n)|(\theta_1, \ldots, \theta_m, a, b, X)$. Absorb the right censored observations into the base measure, so that an effective base measure is defined for this step. The effective base measure is $\alpha + \sum_{i=1}^{m} \delta(\theta_i)$ and hence the effective M is $M + m$. The effective G_0 is proportional to the effective base measure. With this set-up, the algorithm fragment described above is used, with the dose scale reversed, to generate the block of parameters $(\theta_{m+1}, \ldots, \theta_n)$.

- Generate (a, b) with a Metropolis step. To do so, a proposal value for the pair (a, b) is generated, and then either accepted or declined. The actual proposals used in the simulation below are a uniform proposal for a over its entire range and a standard normal proposal for b, centered at the current value of b. The proposal is accepted with the usual Metropolis acceptance probability.

- Generate $\theta^*|(a, b, s, X)$. The conditional distribution of θ^* is a truncated logistic distribution. For θ_i^*, examine cluster i, the set of observations for which $s_j = i$. Find the largest right censored observation in the cluster, call it r, and the smallest left censored observation in the cluster, call it l. If there are no right censored observations in the cluster, set $r = -\infty$. If there are no left censored observations in the cluster, set $l = \infty$. Since the value of θ_i^* must lie in the interval $(r, l]$, the conditional distribution of θ_i^* has $G_{\theta_i^*;a,b,s,z,X}(\theta^*) = 0$ for $\theta^* < r$ and $G_{\theta_i^*;a,b,s,z,X}(\theta^*) = 1$ for $\theta^* \geq l$. In between, the distribution function interpolates as a logistic, with $G_{\theta_i^*;a,b,s,z,X}(\theta^*) = (G_{0;a,b}(\theta^*) - G_{0;a,b}(r))/(G_{0;a,b}(l) - G_{0;a,b}(r))$

The first two generations described above make use of the blocking described in the previous subsection. For the right censored observations, the left censored observations are absorbed into the base measure of the Dirichlet process, and then blocking is used, exactly as described above. For the left censored observations, the dose axis is reversed, the right censored observations are absorbed into the base measure, and the algorithm as described above is applied. If the base-measure had a discrete component, and a left censoring time exactly matched an atom of mass from the discrete component, a slight modification to the algorithm would be needed, since the left censored observation could have a tolerance exactly matching its censoring time. The modification merely requires a slight change in the definition of the conditional distribution in the earlier fragment of an algorithm. The Metropolis step for generation of (a, b) can be replaced with a pair of Gibbs steps, generating a and then b from their full conditional distributions. These conditional distributions do not fall in any of the usual families of distributions.

The Ryanodine data

Dixon (1965) presents the results of a study which investigated the effect of Ryanodine on male mice. Ryanodine was administered intravenously, with dose adjusted for weight, in saline, under 12 experimental conditions.

The experimental conditions arise from a two-way table, with the margins of the table based on body weight and a time cut-off. The mice were grouped into three weight classes, so that the appropriateness of the dosage (defined as dose adjusted for body weight) to body weight scale could be investigated. The response of direct interest was time of last visible movement. Since it can be difficult to determine time of last visible movement, the data were collected on a coarser scale. Four time cut-offs, 64, 96, 144 and 216 seconds, were selected. For a particular time point, it was determined whether there was visible movement after that time point or not.

For each experimental condition, an up-and-down sequential design was conducted. Thus, data was collected on several mice, with the mice assigned varying doses and a binary response/non-response recorded for each mouse. The data for one experimental condition appear in Table 2. The sample sizes for the 12 experiments range from 5 to 8. The reader is referred to Dixon (1965) for the complete data set.

TABLE 1.2. The Ryanodine data for the weight class 18 - 20 and the time cut-off of 64 seconds. The data are presented in the order of experimentation.

z_i:	0	1	2	3	2	1	2
X_i:	0	0	0	1	1	0	1

The data are analyzed on the dosage scale, as twelve separate data sets, with a prior distribution for each chosen to reflect the basic MDP modelling strategy. The prior distribution relies on the information used to implement the up-and-down method. With the up-and-down method, the investigator attempts to guess the median of the tolerance distribution (also called the $LD50$). The first mouse is assigned this guess as his dosage. As the experiment progresses, steps of a constant size are taken, either up or down, depending on the result for the most recent mouse. For this experiment, the step size was targeted as the standard deviation of the distribution of tolerances. We take G_0 to be a logistic distribution with parameters a and b. The prior distribution on b is normal. The 12 prior means for the individual b's were chosen by fitting the two-way additive model to the doses assigned to the first mouse in each weight class by time cut-off pair. The two-way model was fit to smooth out the prior means. The standard deviation of the prior distribution is taken to be 1.61, a value which is consistent with Dixon's assessment of the investigator's prior beliefs. The prior distribution on a was chosen to produce a standard deviation of the tolerances near the dosage step size of 1. It induces a uniform$[0.1, 1.9]$ distribution on the standard deviation. a and b are a priori independent.

To complete the prior distribution, a value (or distribution) is needed for the mass of the base measure. This value was set to $M = 10$, a value that is reasonably large when compared to the sample sizes. Choice of an M that is either large in an absolute sense or large relative to the sample size is necessary when working with bioassay data in order to avoid the excessive clustering first described by Antoniak (1974).

A simulation was performed for each of the 12 data sets. For each data set, the algorithm described above was run, with a burn-in period of 1000 iterates and an estimation period of 5000 iterates. The results of the simulation were used to estimate the predictive distribution of a future tolerance, say $\tilde{\theta}$, via $\hat{G}(\tilde{\theta}) = R^{-1}\sum_{i=1}^{R}\hat{G}_i(\tilde{\theta})$ with the generated values at the end of iterate i yielding $\hat{G}_i(\tilde{\theta}) = (M+n)^{-1}\sum_{j=1}^{n} I(\theta_{s_j} \leq x) + MG_{0;a,b}(\tilde{\theta})/(M+n)$. This estimate, computed on a grid of values for $\tilde{\theta}$, was used to produce an estimate of the $LD50$. Table 3 briefly summarizes the results of the model fitting procedure. The fit also produces estimates of the tolerance distribution for each weight-time combination. Comparison of these estimates facilitates a visual assessment of the similarities and differences between weight groups and across the time cut-off points.

TABLE 1.3. Estimates of the LD50 for each of the 12 Ryanodine data sets. The top number in each cell is the prior predictive LD50; the middle number is the estimate from the MDP simulation described here; the bottom number is the estimate from Dixon's paper.

Wt.	Time	64	96	144	216
18-20		-0.55	-3.66	-6.76	-8.10
		1.71	-0.48	-3.29	-6.74
		1.86	-0.25	-2.84	-6.35
21-23		0.95	-2.16	-5.26	-6.60
		3.78	-1.51	-5.38	-6.87
		4.18	-0.55	-5.06	-6.95
24-26		0.95	-2.16	-5.26	-6.60
		1.66	-2.07	-4.87	-6.26
		1.86	-2.25	-4.52	-6.13

Estimates of the LD50 from the posterior are fairly close to Dixon's estimates. The great advantage of the MDP analysis is that it allows one to investigate many features of the posterior, and it allows one to phrase decision problems that are appropriate for a nonparametric analysis. Estimates of other LD's are calculated in a fashion similar to that for the LD50. Once an estimate of an LD is found, traditional analyses, motivated by parametric models, attach a standard error to the estimate. The full posterior distribution described by the MDP model allows one to investigate the more relevant question of just what the impact of a proposed dose is. Such questions are about the distribution of $G(z)$ for a particular

z rather than about how close z is to some particular LD.

3.3 Estimation

Estimation on the basis of MCMC output can be greatly improved by a variety of techniques. One of the most straightforward improvements, aimed at increasing the accuracy of estimates of marginal features of the posterior distribution, is the Rao-Blackwellization suggested by Gelfand and Smith (1990). With this technique, values generated as part of the simulation are replaced by conditional expectations, most often immediately before a parameter is generated. For example, in the easy Gibbs sampler described at the beginning of section 3, the tabulation based estimate of $E[h(\theta_1)]$, $R^{-1}\sum_{i=1}^{R} h(\theta_1^{(i)})$, can be replaced with $R^{-1}\sum_{i=1}^{R} E[h(\theta_1^{(i)})|\theta_2^{(i-1)}, \theta_3^{(i-1)}]$. Such replacements are, of course, not limited to estimation within the Gibbs sampler or to use of conditional distributions used in the Markov chain. Were one to post-process the output of the simulation, one would find other estimators, including one which is a direct parallel to the usual Rao-Blackwellization, namely $R^{-1}\sum_{i=1}^{R} E[h(\theta_1^{(i)})|\theta_2^{(i)}, \theta_3^{(i-1)}]$.

The technique of Rao-Blackwellization can be used for the MDP model in a variety of ways. One natural use, developed by Bush and MacEachern (1996), is to create estimates at the stage where the θ_i^* are generated. To estimate $E[h(\theta_1)]$, one might use $R^{-1}\sum_{i=1}^{R} E[h(\theta_{s_1}^{(i)})|s^{(i)}, \nu^{(i)}, X]$, which often has a closed form expression. In the baseball example, the distribution of $\theta_{s_i}^*|(s, \nu, x)$ is a beta distribution, and so with polynomial h, the expectations are easily evaluated.

The bioassay model admits a wide range of Rao-Blackwellizations. To estimate the predictive density of the tolerance of an individual not included in the study, a particularly nice Rao-Blackwellization comes from Kuo's (1983) work. The distribution of $\theta_i^*|(s, \nu, X)$, used as part of the Gibbs sampler, is described in the last step of the algorithm of section 3.2. It is a truncated logistic distribution. To estimate the predictive distribution of a future tolerance, say $\tilde{\theta}$, compute $\hat{G}(\tilde{\theta}) = R^{-1}\sum_{i=1}^{R} \hat{G}_i(\tilde{\theta})$ with $\hat{G}_i(\tilde{\theta}) = (M+n)^{-1}[\sum_{j=1}^{k^{(i)}} n_j^{(i)} G_{\theta_j^*;s^{(i)},\nu^{(i)},X}(\tilde{\theta}) + MG_{0;a,b}(\tilde{\theta})]$. The predictive density is the derivative of $\hat{G}$ which exists for all $\tilde{\theta}$ other than assigned doses. Since the expression for $\hat{G}$ is linear in the individual $G_{\theta_j^*}$ and G_0, the density can be computed directly, avoiding the need to differentiate $\hat{G}$.

This Rao-Blackwellization for the bioassay problem produces an estimate with a nice property. It matches two known features of the estimand. First, although the model assigns probability one to the event that the tolerance distribution is discrete, the predictive distribution of tolerances is known to be continuous. The estimator above, an average of continuous distributions, is continuous. There has been no need to artificially smooth the estimates in order to produce this continuity. Second, the estimand may be represented

as a mixture over cluster structures and values of a and b. The pieces of this mixture are of the form $G_{\theta;s,\nu,X}$. The estimate has this same structure. The mixture estimate also allows one to pick up and display more subtle features of the estimand such as the discontinuity of the predictive density at the assigned dosages.

The variety of potential Rao-Blackwellizations in MDP models is enormous. My personal preference is to identify what seem to be key qualitative features of the posterior and to choose a Rao-Blackwellization that has the same features. There will typically be many such Rao-Blackwellizations. The final decision of which Rao-Blackwellization to choose involves a trade-off between the additional effort required to program the estimator, the time needed to compute the estimator, and the perceived benefits of use of the estimator.

4 Non-Conjugate Models

In many instances, the base measure, conditioned on the hyperparameters, will not have a conjugate form with the likelihood. In this instance, the integral to obtain q_0 may be difficult to compute, or even to approximate. In such a situation, one must turn to another method to perform the simulation. MacEachern and Müller (1998) describe an algorithm which can be routinely implemented. This algorithm relies on the creation of a model that underlies the Polya urn scheme. The model remains hidden throughout the simulation, and entails only small changes to the basic algorithm. It nevertheless enables one to avoid the integral needed to find q_0. The cost of avoiding the integration is an algorithm which appears to suffer to some extent from poorer convergence and mixing properties.

There is one situation, however, in which the basic algorithm can be rescued with a slight alteration. When the density $g_{0,\nu}(\theta_k^*)f_{\theta_k^*,z_i}(X_i)$ appearing in term **2** of step 2 of the basic algorithm is log-concave, one can make use of a variant on Gilks and Wild's (1992) adaptive rejection method to generate $\theta_i|(\nu, M, \theta_{-i}, X)$. The variant arises from the fact that the rejection envelope which those authors adaptively build can also be built for a piecewise log-concave (or log-convex!) density. The density at each of the cluster locations, θ_i^*, is trivially log-concave. So, to demonstrate piecewise log-concavity, we need only verify that the aforementioned density is log-concave. Evans and Swartz (1998) develop a more general version of this technique that allows for transformations more general than the log transform. Further work on generation from non-conjugate models can be found in Walker and Damien (1998).

5 Discussion

The development of the MDP model, as the development of the Dirichlet process itself, reflects the tension between devising a model that captures important features of the data and finding a model that can actually be fit to the data. Escobar's (1994) development of the first MCMC techniques for the Dirichlet process provided major impetus to the field, allowing the MDP modelling strategy to be implemented for an array of problems that had previously been beyond then current computational procedures.

Computational developments have included a focus on the cluster structure of θ, relying either on an integration (MacEachern, 1994) or an additional stage appended to the Gibbs sampler (Bush and MacEachern, 1996) to improve the mixing of the Markov chain used to fit the model. The material in Section 3 furthers these developments. It should also be mentioned that a number of alternative computational methods have also been developed for the MDP model. Notable among them are an MCMC algorithm developed by Doss (1994) which takes a different tack, essentially generating the random G rather than marginalizing it, and the sequential imputation method whose application to the MDP problem is due to Liu (1994). Some of the same modifications developed for the MCMC methods also pay dividends for these alternative strategies (MacEachern, Clyde and Liu, 1994; Doss and Huffer, 1998). Finally, techniques developed for finite mixture models, such as splitting and combining clusters (see Richardson and Green, 1997, and references therein), can also be used to improve MCMC methods for MDP models.

The models have also been refined, moving beyond a model that essentially describes one-sample problems but which can be used for problems that appear to require a more complex model (see, for a sterling example, Müller, West and Erkanli, 1996). The more refined models develop particular roles for the parametric and nonparametric portions of the model. Two directions are now well established: One direction, exemplified by the early paper of Newton, Czado and Chappell (1996) in the context of logistic regression, is to model the impact of covariates in a parametric fashion, but to replace the logistic distribution with a nonparametric model (in their case a variant on the Dirichlet process). The second direction is to draw a distinction between fixed effects–those effects for which we can model the individual effect–and random effects–those effects for which we most naturally model the distribution from which the effects arise. This distinction, first drawn for the MDP model in Bush and MacEachern's (1996) analysis of the randomized block design, and further developed in Bush (1994), suggests use of a parametric model for the fixed effects and a nonparametric model for the random effects. Further modelling developments are currently an active area of research. This author is pursuing research that will more sharply define the dividing line between the parametric and nonparametric portions of the model, and also the development of models that

automatically display more robustness than do the current MDP models.

References

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

Blackwell, D. and MacQueen, J.B. (1973). Ferguson distributions via Polya urn schemes. *Ann. Statist.* **1** 353-355.

Bush, C.A. (1994). Semiparametric Bayesian Linear Models. Ph.D. dissertation, Ohio State University.

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

Dixon, W.J. (1965). The up-and-down method for small samples *J. Amer. Statist. Assoc.* **60** 967-978.

Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Ann. Statist.* **22** 1763-1786.

Doss, H. and Huffer, F. (1998). Monte Carlo methods for Bayesian analysis of survival data using mixtures of Dirichlet priors. Technical Report, Department of Statistics, Ohio State University.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311-319.

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

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

Evans, M. and Swartz, T. (1998). Random variable generation using concavity properties of transformed densities. To appear in *J. Comp. Graph. Statist.*

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

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

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398-409.

Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *J. Applied Statist.* **41** 337-348.

Kuo, L. (1983). Bayesian bioassay design. *Ann. Statist* **11** 886-895.

Kuo, L. (1986). Computations of mixtures of Dirichlet processes. *SIAM J. Sci. Statist. Comput.* **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. J.P. Klein and P.K. Goel, 11-22.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear

model (with discussion). *J. R. Statist. Soc. B* **34** 1-42.

Liu, J.S. (1994). The collapsed Gibbs sampler in Bayesian computations with application to a gene regulation problem. *J. Amer. Statist. Assoc.* **89**, 958-966.

Liu, J.S., Wong, W.H. and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27-40.

MacEachern, S.N. (1988). Sequential Bayesian bioassay design. Unpublished Ph.D. Dissertation. University of Minnesota.

MacEachern, S.N. (1992). Discussion of "Bayesian computations in survival models via the Gibbs sampler" by Kuo and Smith. In Survival Analysis: State of the Art, ed. J.P. Klein and P.K. Goel, 22-23.

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

MacEachern, S.N., Clyde, M. and Liu, J. (1994). Sequential importance sampling for nonparametric Bayesian models: The next generation. To appear, *Can. J. Statist.*

MacEachern, S.N. and Müller, P. (1998). Estimating mixtures of Dirichlet process models. To appear, *J. Comp. Graph. Statist.*

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

Newton, M.A., Czado, C. and Chappell, R. (1996). Bayesian inference for semiparametric binary regression. *J. Amer. Statist. Assoc.* **91** 142-153.

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

Walker, S. and Damien, P. (1998). Sampling methods for Bayesian nonparametric inference involving stochastic processes.

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, 363-368.