15

Efficient MCMC Schemes for Robust Model Extensions using Encompassing Dirichlet Process Mixture Models

Steven MacEachern and Peter Müller

ABSTRACT We propose that one consider sensitivity analysis by embedding standard parametric models in model extensions defined by replacing a parametric probability model with a nonparametric extension. The nonparametric model could replace the entire probability model, or some level of a hierarchical model. Specifically, we define nonparametric extensions of a parametric probability model using Dirichlet process (DP) priors. Similar approaches have been used in the literature to implement formal model fit diagnostics (Carota, Parmigiani and Polson, 1996).
In this paper we discuss at an operational level how such extensions can be implemented. Assuming that inference in the original parametric model is implemented by Markov chain Monte Carlo (MCMC) simulation, we show how minimal additional code can turn the same program into an implementation of MCMC in the larger encompassing model, allowing formal sensitivity analysis with respect to prior and likelihood assumptions. If the base measure of the DP is assumed conjugate to the appropriate component of the original probablity model, then implementation is straightforward. The main focus of this paper is to discuss general strategies allowing implementation of models without this conjugacy.

1 Introduction

We propose that one consider sensitivity analysis by embedding standard parametric models in nonparametric extensions. We use random measures with DP priors to define these encompassing nonparametric extensions. We present a framework which makes the implementation of posterior inference in such extensions always possible with minimum additional effort, essentially requiring only one additional multinomial sampling step in a Markov chain Monte Carlo (MCMC) posterior simulation. This is straightforward for models which are conjugate (conjugate in a sense which we shall make formal). In models without such conjugate structure, however, computational problems render posterior simulation difficult, and hinder the routine application of such nonparametric model augmentations. In this paper we present a scheme which overcomes this hurdle and allows the implementation of robust nonparametric model extensions with equal ease in nonconjugate models.

In this chapter we shall use models based on Dirichlet process prior distributions (Ferguson, 1973; Antoniak, 1974). Many alternative approaches are possible for the encompassing nonparametric model. Among the many models proposed for nonparametric Bayesian modelling in the recent literature are Polya trees (Lavine 1992, 1994), Gaussian processes (O'Hagan, 1992; Angers and Delampady, 1992), beta processes (Hjort, 1990), beta-Stacy processes (Walker and Muliere, 1997), extended gamma processes (Dykstra and Laud, 1981), random Bernstein polynomials (Petrone, 1999a,b). See Walker et al. (1999) for a recent review of these alternative forms of nonparametric Bayesian modelling.

Consider a generic Bayes model for a collection of n nominally identical problems with likelihood

$$y_i \stackrel{iid}{\sim} p_{\theta,\nu}(y_i), \quad i = 1, \ldots, n, \tag{1}$$

and prior $\theta \sim G_0(\theta|\nu)$ and $\nu \sim H(\nu)$. In anticipation of the later generalization the parameter vector is partitioned into (θ, ν), where θ is the subvector of parameters with respect to which the model extension will be defined below. Model (1) could, for example, be a normal distribution with unknown location θ and variance ν. Inference from such a model is extremely restrictive in that a single parameter θ indexes the conditional distribution for each and every y_i. Estimation of an observation specific parameter – say θ_i, representing the mean of the conditional distribution for y_i in our simple example – is identical for every i since there is only a single θ. At the far extreme from model (1), we may write

$$y_i \stackrel{iid}{\sim} p_{\theta_i,\nu_i}(y_i), \quad i = 1, \ldots, n, \tag{2}$$

and prior $\theta_i \sim G_0(\theta_i|\nu_i)$ and $\nu_i \sim H(\nu_i)$, creating n separate problems. Since the joint distribution on the n collections of parameters, θ_i, ν_i, y_i, form a set of n independent distributions, inference is made independently in the n cases. This model does not permit any pooling of information across the n problems, leading to potentially poor inference.

We consider generalizations of (1) to

$$y_i \stackrel{iid}{\sim} \int p_{\theta,\nu}(y_i)\, dG(\theta), \quad G \sim DP\left(M\, G_0(\cdot|\nu)\right). \tag{3}$$

The original sampling model $p_{\theta,\nu}$ is replaced by a mixture over such models, with a mixing measure G. For example, we might replace a simple normal sampling model by a location mixture of normals. As a probability model for the random mixing measure we assume a Dirichlet process (DP) with base measure MG_0, where G_0 is a probability measure. See, for example, Antoniak (1974) or Ferguson (1973) for a definition and discussion of DP's. The model contains the original model (1) as a special case when G is a point mass. The DP prior puts non-zero prior probability on G being arbitrarily close to such a single point mass, and implies that the point mass be a sample from G_0. The base measure of the DP need not be the same as the prior in the original parametric model, but this is a natural choice since it implies the same marginal distribution $p(y_i)$ as under (1). The model provides a nice alternative to (2), allowing us to pool information obtained from the entire collection of problems to make better inference for each individual problem.

The perspective of providing a flexible, nonparametric version of the parametric Bayes model motivated much early work in the area (see, for example, Susarla and van Ryzin, 1976; Kuo, 1983; MacEachern, 1988; Escobar, 1988). The flexibility of the nonparametric analysis both allows one to conduct a formal sensitivity analysis by comparing the fit of the parametric model and its elaboration and also provides a fresh look at the data with what can alternatively be considered a larger model.

For the sake of presentation it is convenient to consider the case of a parametric hierarchical model which is to be elaborated separately from the case of non-hierarchical models. To wit,

$$\begin{aligned} y_i &\overset{iid}{\sim} p_{\theta_i,\nu}(y_i), \\ \theta_i &\overset{iid}{\sim} G_0(\theta_i|\nu), \end{aligned} \tag{4}$$

with prior $\nu \sim H(\nu)$. The model is generalized by replacing the prior G_0 with a random distribution G:

$$\begin{gathered} y_i \overset{iid}{\sim} p_{\theta_i,\nu}(y_i) \\ \theta_i \overset{iid}{\sim} G(\theta_i), \quad G \sim DP\left(M\, G_0(\cdot|\nu)\right). \end{gathered} \tag{5}$$

As can easily be seen by marginalizing over θ_i in (5) model (5) is identical to (3). Following traditional terminology we refer to (5) as the mixture of Dirichlet process model (MDP). Given a MDP model it is often a matter of perspective whether it is seen as a generalization of a basic model (1) or a hierarachical model (4), although we believe the latter is the more common view in the literature. See Escobar and West (1998) for a recent summary of this perspective. Below, in examples (i) through (xii), we give examples of both.

In the rest of this chapter we will argue that Markov chain Monte Carlo (MCMC) posterior simulation in model (5), and thus in (3), can be easily implemented by adding just one additional (multinomial) sampling step to an MCMC scheme for the original models (4) or (1). Posterior inference under the augmented model (3) or (5) provides a basis for investigating model sensitivity and robustness.

2 Survey of MDP models

A number of models in the recent literature fit into the framework of (5). Recent versions of these models, and new developments include those that follow. When likelihoods do not depend on certain parameters, the corresponding subscripts have been omitted. Most of these applications include priors on ν which have been omitted. Using the notation of (1) and (4), for each application we point out the corresponding G_0 and parameter θ or θ_i, respectively. Depending on what we think is the more natural perspective, we write $p_{\theta,\nu}$ as in (1), or $p_{\theta_i,\nu}$ as in (4). We use $N(x; m, S)$ to indicate that the random variable x follows a normal distribution with mean and variance (m, S). Also, we use $Bin(x;\ n, \theta)$, $W(x;\ \nu, A)$, $Ga(x;\ a, b)$, $Exp(x;\ \lambda)$, $U(x; a, b)$, $Dir(x; \lambda)$ and $Be(x;\ a, b)$ to denote a binomial, Wishart, gamma, exponential, uniform, Dirichlet and beta distribution, respectively. Our notation ignores distinctions between random variables and their realizations.

(i) Nonparametric regression: Müller, Erkanli and West (1996) use

$$\theta_i = (\mu_i, \Sigma_i) \text{ and } p_{\mu_i,\Sigma_i}(y_i) = N(y_i; \mu_i, \Sigma_i)$$

where $G_0(\mu, \Sigma) = N(\mu; a, B)\ W(\Sigma^{-1}; s, S)$;

(ii) Density estimation: West, Müller, and Escobar (1994) have

$$\theta_i = (\mu_i, \Sigma_i),\ p_{\mu_i,\Sigma_i}(y_i) = N(y_i; \mu_i, \Sigma_i)$$

and $G_0(\mu, \Sigma) = N(\mu; a, B)\ \ W(\Sigma^{-1}; s, S)$;

Gasparini's (1993) model can be reformulated as an MDP model with

$$p_{\theta_i,\nu}(y_i) = \mathrm{U}(y; \theta_i - \nu, \theta_i + \nu)$$

and $G_0(\theta)$ a discrete measure on $\{a, a + 2\nu, a + 4\nu, \dots\}$;

(iii) Estimation of a monotone density. Brunner (1995) has

$$p_{\theta_i}(y_i) = \mathrm{U}(y; 0, \theta_i),$$

where $G_0(\theta)$ is an arbitrary distribution on the positive half-line. Brunner and Lo (1989) use a similar model for estimation of a symmetric, unimodal density.

(iv) Hierarchical modelling: Escobar and West (1995) have

$$\theta_i = (\mu_i, \sigma_i),\ p_{\mu_i,\sigma_i}(y_i) = N(y_i; \mu_i, \sigma_i)$$

and $G_0(\mu, \sigma) = Ga(\sigma^{-2}; s/2, S/2)\ \ N(\mu; m, \tau\sigma^2)$.

MacEachern (1994) uses $p_{\theta_i}(y_i) = N(y_i; \theta_i, \sigma^2)$.

Liu (1996) proceeds from (1), the non-hierarchical model, and uses $p_\theta(y_i) = Bin(n_i, \theta)$, where $G_0(\theta) = Be(a, b)$.

(v) Fixed and random effects modelling: Bush and MacEachern (1996) have

$$p_{\theta_i,\lambda}(y_i) = N(y_i; \lambda' x_i + \theta_i, \sigma_i^2)$$

where $G_0(\theta) = N(m, \sigma^2)$ and x_i is a vector of covariates for observation i. Malec and Müller (1999) use a similar random effects model in the context of small area estimation.

(vi) Contingency tables: Quintana (1998) has

$$\theta_i = (p_{i1}, \ldots, p_{il}) \text{ and } p_{\theta_i}(n_i) = Multin(n_i, p_i),$$

with $G_0(p) = Dir(p; \lambda)$;

(vii) Longitudinal data models: Müller and Rosner (1997) and Kleinman and Ibrahim (1998) use for patient-specific random effects z_i:

$$p_{\mu,\Sigma}(z_i) = N(z_i;\ \mu, \Sigma),\ G_0(\mu) = N(m, S).$$

(viii) Estimating possibly non-standard link functions: Erkanli, Stangl, and Müller (1993).

$$y_i = \begin{cases} 0 \text{ if } z_i < 0, \\ 1 \text{ if } z_i \geq 0, \end{cases}$$

$p_\theta(z_i) = N(z_i; \mu, 1)$, and $G_0(\mu) = N(\mu; m, \tau^2)$.

(ix) Censored data: Doss' (1991) model for survival data and one of the proposed models in Gelfand and Kuo (1991) for dose-response data can be rewritten as:

$$p_{\theta_i}(y_i) = \begin{cases} 0 \text{ with prob. 1 if } \theta_i > x_i, \\ 1 \text{ with prob. 1 if } \theta_i \leq x_i, \end{cases}$$

for those data values that are right censored. Left and interval censored data values have similarly defined likelihoods. Uncensored observations are absorbed into the base measure. Doss uses $G_0(\theta) = Exp\,(\mu)$. while Gelfand and Kuo take $G_0(\theta) = N(\theta; \mu, \sigma^2)$.

(x) Survival analysis with covariates: Kuo and Mallick (1997) use the accelerated failure time model where

$$p_{\mu_i,\sigma_i,\beta}(\log T_i) = N(\log T_i;\ \mu_i - \beta' x_i,\ \sigma_i)$$

for failure times T_i, and – in one example – $G_0(\mu, \sigma) = Exp(1)\ \delta_{0.1}(\sigma)$, where $\delta_x(\cdot)$ is a point mass at x. Alternatively they consider a similar DP mixture on $v_i = T_i\ \exp(x_i \beta)$ instead of $w_i = \log T_i + \beta' x_i$.

(xi) Generalized linear models: Mukhopadhyay and Gelfand (1997) use

$$p_{\theta,\beta}(y_i) = f(y_i|\eta = \theta + x_i'\beta)$$

where $f(y|\eta)$ is a generalized linear model with linear predictor η.

(xii) Errors in variables models: Müller and Roeder (1997) use for the joint distribution of the missing covariate x_i and observed proxy w_i:

$$p_{\mu,\Sigma}(w_i, x_i) = N(w_i, x_i;\ \mu, \Sigma),\ G_0(\mu) = N(m, S).$$

Other related models are used in Lavine and Mockus (1995), Kuo and Smith (1992) and Newton, Czado and Chappell (1996) Numerous other authors are currently working with models that fit into this MDP framework.

3 Gibbs Sampling in Conjugate MDP Models

We briefly review Markov chain Monte Carlo schemes currently applied to estimate MDP models. Estimation of the MDP model (5) can be efficiently implemented by a Gibbs sampling scheme if $p_{\theta,\nu}$ and G_0 are conjugate (cf. Escobar and West, 1995, MacEachern, 1994, West, Müller and Escobar, 1994, Bush and MacEachern, 1996). In MacEachern and Müller (1998), we define a model augmentation and outline a Markov chain Monte Carlo implementation which allows the use of nonconjugate pairs $p_{\theta,\nu}$ and G_0. The focus is on discussing the conceptual framework.

The next two sections summarize the discussion in MacEachern and Müller (1998) which is relevant for a practical implementation of a Gibbs sampling algorithm. It has the added benefit of providing an explicit description of the "complete model" algorithm which was trimmed from the published version of that manuscript. Building on this general discussion, we give specific Gibbs sampler algorithms suitable for a practical implementation.

A key feature of the DP is the almost sure discreteness of the random measure G which gives positive probability to some of the θ_i's being equal. When the base measure G_0 is continuous with probability 1, the θ_i's are equal only due to the discreteness inherent in the Dirichlet process, and not to discreteness of G_0. In this case, write $\{\theta_1^*, \ldots, \theta_k^*\}$ for the set of $k \leq n$ distinct elements in $\{\theta_1, \ldots, \theta_n\}$. Thus θ is partitioned into k sets. Call this partitioning a configuration, and let $s_i = j$ iff $\theta_i = \theta_j^*$ denote configuration indicators. Also let n_j be the number of s_i equal to j, i.e., the size of the j-th element of the partition (also called the jth cluster).

A Gibbs sampling scheme to estimate MDP models is described by the following conditional distributions.

(i) Resampling s_i given all other parameters:

We marginalize over θ_i and sample s_i from

$$Pr(s_i = j|\theta_{-i}, s_{-i}, \nu, y) \propto \begin{cases} n_j^- \; p_{\theta_j^*,\nu}(y_i) & j = 1, \dots, k^- \\ M \; \int p_{\theta,\nu}(y_i) \; dG_0(\theta) & j = k^- + 1 \end{cases} \quad (6)$$

Here, θ_{-i} denotes the vector $(\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n)$, n_j^- denotes the size of the j-th cluster with θ_i removed from consideration (i.e., $n_{s_i}^- = n_{s_i} - 1$ while for other j, $n_j^- = n_j$), and k^- denotes the number of clusters with θ_i removed from consideration. If $n_{s_i}^- = 0$, we relabel the remaining clusters $j = 1, \dots, k^- = k - 1$.

After sampling s_i, redefine k accordingly, i.e., set $k = k^-$ if $s_i \leq k^-$, and $k = k^- + 1$ if $s_i = k^- + 1$.

(ii) Resampling θ_j^* is straightforward. The posterior $p(\theta_j^*|s, \nu, y)$ is the same as in the simple Bayes model (4) with i going over all indices with $s_i = j$:

$$y_i \sim p_{\theta_j^*,\nu}(y_i), \; \theta_j^* \sim G_0(\theta_j^*|\nu)$$

for all i such that $s_i = j$.

(iii) Resampling M: If one wishes to express uncertainty about the total mass parameter, it can be included in the parameter vector and resampled in the MCMC simulation. West (1992) shows that if M is given a $Ga(a, b)$ hyperprior, it can be resampled by introducing an additional latent variable x with $p(x|k, M) = \text{Be}(M + 1, n)$ and

$$p(M|x, k) = \\ \pi \; Ga\left(a + k, b - log(x)\right) + (1 - \pi) \; Ga\left(a + k - 1, b - log(x)\right),$$

where $\pi/(1 - \pi) = (a + k - 1)/n(b - log(x))$. Alternatively, uncertainty about the mass parameter can be expressed and then eliminated from the sampling scheme through a preintegration, as described in MacEachern (1998).

(iv) Resampling ν given all other parameters: The portion of the model involving ν is a conventional parametric model. Hence, conditioning on all other parameters leaves a standard Bayes model. Often, this will be of conjugate form and a standard generation will suffice.

Only step (i) and, if included, step (iii) go beyond the MCMC for the original parametric model. Step (i) is a multinomial draw and Step (iii) is a gamma and a beta random variate generation. Steps (ii) and (iv) might require complicated posterior simulation, depending on the application. These steps remain almost unchanged.

4 Novel Algorithms for Non-conjugate MDP Models

The Gibbs sampler described in Section 3 is practicable only if $p_{\theta,\nu}$ and $G_0(\theta|\nu)$ are conjugate in θ allowing analytic evaluation of $q_0 = Pr(s_i = k^- + 1|\ldots)$ in equation (6). In many applications, however, a nonconjugate setup is required. West, Müller, and Escobar (1994) present an algorithm for nonconjugate MDP models using an approximate evaluation of q_0. In MacEachern and Müller (1998), we propose a general framework which allows nonconjugate pairs G_0 and $p_{\theta,\nu}$. The scheme is based on a model augmentation introducing latent variables, $\{\theta^*_{k+1}, \ldots, \theta^*_n\}$, for up to n possible cluster locations. At any time, $n-k$ of the clusters are empty, i.e. $n_j = 0$ for these j. For a detailed definition and discussion we refer to MacEachern and Müller (1998). Here we build on the conceptual framework described there to formulate a practical implementation of a Gibbs sampling scheme for continous base measures G_0.

Alternative approaches for MCMC in nonconjugate models are described in Neal (1998) and Green and Richardson (1998) and in Walker and Damien (1998) and MacEachern (1998) for one-dimensional distributions. Neal (1998) proposes alternative algorithms using Metropolis-Hastings type moves to propose new configuration indicators s_i. Similar to Neal (1998), Green and Richardson (1998) exploit the relationship of the DP mixture model with a dirichlet/multinomial allocation model and propose an algorithm based on split/merge moves. Walker and Damien (1998) use the auxiliary variable technique introduced in Damien, Wakefield and Walker (1998), essentially avoiding evaluation of the integral in (6) by introducing a uniform latent variable u with $p(u_i|\theta_i) \sim U[0, p_{\theta_i,\nu}(y_i)]$. MacEachern (1998) suggests the use of adaptive rejection techniques for the special case of log-concavity in the complete conditional posterior distribution for θ_i.

We define two alternative algorithms, based on the "no gaps" model and the "complete model" defined in MacEachern and Müller (1998). Choice of the algorithm depends on the particular application. As a guideline, if k is typically much smaller than n, and n is large, then we recommend the "no gaps" algorithm. In other nonconjugate situations, we recommend the complete model.

4.1 No Gaps Algorithm

Application of the "no gaps" model results in the following changes of the Gibbs sampler steps (i) through (v) described in Section 3:

(i') If $n_{s_i} > 1$, then resample s_i from

$$Pr(s_i = j|\theta^*, s_{-i}, \nu, y) \propto \begin{cases} n_j^- \; p_{\theta^*_j,\nu}(y_i) & j = 1, \ldots, k^- \\ \frac{M}{k^-+1} \; p_{\theta^*_{k^-+1},\nu}(y_i) & j = k^- + 1. \end{cases} \quad (7)$$

Here, k^- and n_j^- denote the number of distinct values in $\{\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n\}$ and the number of observations in cluster j after removing observation i, respectively.

If $n_{s_i} = 1$, then with probability $(k-1)/k$ leave s_i unchanged. With probability $1/k$, resample s_i from (8).

If a cluster j with $n_j = 1$ is removed by resampling s_i then switch the labels of clusters j and k and decrement k by 1. But keep the old value of θ_j^* recorded as θ_{k+1}^*.

After repeating step (i') for $i = 1, \ldots, n$ to resample all indicators s_i, marginalize over $\theta_{k+1}^*, \ldots, \theta_n^*$ by simply dropping them from the simulation. Steps (ii), (iii) and (iv) are executed conditioning on $\theta_1^*, \ldots, \theta_k^*$ only. Before returning to step (i') in the next iteration, augment the θ^* vector again by sampling $\theta_{k+1}^*, \ldots, \theta_n^*$ from $G_0(\theta_j^*|\nu)$. Of course, this could be done when and as θ_j^* is needed in step (i') only, i.e., θ_j^*, $j = k+1, \ldots, n$ need not be actually generated and kept in memory until they are needed in step (i').

4.2 The Complete Model Algorithm

(i") The complete conditional posterior for resampling s_i is given by

$$Pr(s_i = j|\theta^*, s_{-i}, \nu, y) \propto \begin{cases} n_j^- \; p_{\theta_j^*,\nu}(y_i) & j = 1, \ldots, k^- \\ \frac{M}{n-k^-} \; p_{\theta_j^*,\nu}(y_i) & j = k^- + 1, \ldots, n \end{cases} \tag{8}$$

where k^- and n_j^- are defined as before.

Again, after step (i") update k and relabel the clusters such that all "empty" clusters (with $n_j = 0$) have higher indices than the "non-empty" clusters (with $n_j > 0$).

After completing step (i") for $i = 1, \ldots, n$, marginalize over $\theta_{k+1}^*, \ldots, \theta_n^*$ by dropping them from the simulation and execute steps (ii) through (iv) conditioning on $\theta_1^*, \ldots, \theta_k^*$ only. Before returning to step (i") augment the θ^* vector again by generating $\theta_j^* \sim G_0(\theta_j^*|\nu)$, $j = k+1, \ldots, n$.

5 Non-identifiability of the algorithms

The primitive notion of identifiability is easily described. An identifiable model has the property that, for every point in the parameter space, a different distribution is implied for the data that are to be collected in an experiment. Unfortunately, this notion becomes a bit fuzzy in the hierarchical Bayesian model, as just what is considered a parameter is open to several interpretations. Since several different parameterizations of a model, some of which may even be nested in others, can be considered identifiable, it is difficult to write a concise discussion of identifiability and its impact on fitting nonparametric Bayesian models. Nevertheless, the issue is important enough for fitting these models that we provide a brief discussion of the issue here.

We stress that identifiability is often important for interpretation of an analysis, and that the question can appear in subtle forms in nonparametric Bayesian analysis. Newton et al. (1996) create a model so that a set of parameters that are useful for interpretation are identifiable and also explicitly appear in their model. Green and Richardson (1997) provide a focused treatment of identifiability in the context of finite mixture models of varying dimension. In general, the interpretation of a model is often tied to a particular parameterization at an intuitive level if not in mathematical terms. As an example, the two distinct routes to creation of the MDP model lead to different identifiable parameterizations of the model. While the eventual use of the model may follow from either generalization, for computational purposes, we need only ensure that the strategy used to fit the models allows us to make inference under either parameterization.

First, an example, to illustrate the variety of ways in which one can term models as identifiable or not in the context of the MDP model. The parameter for model (3) or (5) can be considered either G, or $(\theta_1, \ldots, \theta_p)$, or $(G, \theta_1, \ldots, \theta_p)$. Under relatively weak conditions on the likelihood, the first parameterization leads to a model that is identifiable in the sense that the joint distribution on y differs for each differing G. This parameterization is most naturally thought of as the generalization of model (3). The second model is identifiable in that the distribution for y differs for each differing vector θ. This parameterization is most naturally thought of as the generalization of model (5). The third model only becomes identifiable when one steps outside of the current experiment, as when performing a predictive analysis. The joint distribution of the data collected from the current experiment and a future observation, say y_{p+1} which depends on θ_{p+1} with $\theta_{p+1} \sim G$, depends both on G and on $\theta_1, \ldots, \theta_p$.

In terms of computation, the algorithms which we have just described facilitate inference under any of the three parameterizations given above (see the next section for the treatment of predictive inference relevant to the third parameterization). At any stage in the Gibbs sampler, the stored parameters – the vector s and the vector θ^* – enable us to reconstruct the entire vector θ. Consequently, inference about the individual θ_i can be performed on the basis of the standard MCMC formulas. This handles inference for the second parameterization given above.

Inference for the first parameterization follows from the distribution of G given θ. Recalling the early result from Ferguson (1973), we know that the distribution of $G|\theta$ follows a Dirichlet process with parameter $\alpha + \sum_{i=1}^{p} \delta_{\theta_i}$. Subsequent inference about functionals against G can be made with any of the many varied tricks that have been described in the literature. The two main approaches for functionals that do not have tidy, closed form expressions are to either pin down the random G at a collection of sites (the joint distribution of $G(x_1), \ldots, G(x_n)$ for $x_1 < \ldots, x_n$ follows from a Dirichlet distribution on the increments between successive values of the distribution function) or to approximate the countably infinite discrete distribution by a finite discrete distribution, perhaps of arbitrary size. The distribution on the finite approximation follows from Sethuraman's (1994) representation of the Dirichlet process and the rule for determining the (possibly random) number of components in the finite mixture. See, in particular, Tardella and Muliere (1998) for the ϵ-Dirichlet process and Gelfand and Kottas (1999) for the first sort of approximation. Guglielmi (1998) provides a means for calculating what are effectively exact values for functionals of Dirichlet processes.

The computational algorithms described in the preceeding section rely on an additional non-identifiability in terms of the model that is written out for simulation. Instead of describing the value of the parameters $\theta_1, \ldots, \theta_p$ at any stage of the algorithm directly in terms of the θ's, a latent structure is introduced. The vector (s, θ^*) contains all information needed to reconstruct the vector θ through the relation $\theta_i = \theta^*_{s_i}$. There will be many vectors (s, θ^*) that result in the same value for the vector θ. We term these models non-identifiable because for any inference made from the posterior distribution for G and θ, the inference will not depend on the particular (s, θ^*) which produced this θ, G pair. The point that we wish to emphasize is that the model devised for computational purposes provides a finer scale of latent structure than does a model (3) or (5) elaboration.

The no-gaps algorithm and the complete model algorithm result from particular models for the latent structure. Each grouping of the θ_i into clusters corresponds to several vectors s. We have deliberately created non-identifiable models in order to improve the mixing/convergence of the Markov chain used to fit the models. In particular, the models are created by first writing an identifiable version of the model that leads to a 1-1 relationship between s and the grouping of the θ_i. This identifiable version of the model is then symmetrized by creating many labellings that correspond to that particular grouping of the θ_i and by apportioning the probability assigned to that grouping to each of the possible labellings.

For the no-gaps model, symmetrization proceeds by first labelling the, say k, groups $1, \ldots, k$. Next, all permutations of the labels $1, \ldots, k$ are considered for the group names. Each such permutation receives $1/k!$ of the probability assigned to that particular grouping of the θ_i.

For the complete model, symmetrization proceeds by first labelling the k groups $1, \ldots, k$, in an identifiable fashion. Next, all subsets of k distinct labels chosen from the integers $1, \ldots, n$ are considered as labellings of the groups. Each of the $n!/(n-k)!$ labellings receives an equal share of the probability assigned to that particular grouping of the θ_i.

The motivation behind the introduction of non-identifiable models for simulation as well as the idea behind symmetrization can be found in MacEachern (1996, 1998). West (1997) and Huerta and West (1999) use the technique to improve simulation in related models. For theory on the improvement that non-identifiable models can bring to MCMC simulation see Meng and van Dyk (1999).

6 The 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. In the basic MDP model, the posterior predictive distribution is most easily found by returning from the *no gaps* or *complete* 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 $(\sum n_j\delta_{\theta_j^*} + MG_0)/(M+n)$. The term involving G_0 may be evaluated as $M\tilde{\theta}$ 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. 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. Here θ^t refers to the imputed parameter vector θ after t iterations. When this evaluation is not feasible, after each iteration a value y_{n+1} can be generated, with the resulting estimator $1/T\sum_{t=1}^{T} y_{n+1}$.

7 Discussion

Inference in the encompassing nonparametric model provides an alternative to traditional sensitivity analysis. In this paper we have introduced practically feasible approaches to implement such model extensions with – at least in principle – minimal additional effort. Considering nonparametric extensions can also be used for formal robustness measures in the form of model diagnostics. Such approaches are considered in Carota, Parmigiani and Polson (1996) and Florens, Richard and Rolin (1996) using DP priors, in Berger and Guglielmi (1999) using Polya tree priors, and in Verdinelli and Wasserman (1998) using Gaussian processes.

The beauty of the nonparametric Bayesian sensitivity analysis/model elaboration is that it plays this dual role. As a sensitivity analysis technique, one can monitor the range of posterior inferences as the prior distribution is varied over a class of nonparametric priors; as a more general class of models, one can monitor summaries of the fit of the model as the prior varies over a class. In any realistic setting, the sensitivity analysis produces a range of inferences; often, the general class of models exhibits a substantially better fit than does the parametric model. Qin (1998) provides a prime example of this improvement in fit where, in a fairly complex model, the posterior standard deviations for several parameters are *smaller* under the nonparametric elaboration than under the parametric model. Such examples illustrate the benefit of a technique that plays both exploratory and confirmatory roles.

References

Angers, J.-F. and Delampady, M. (1992). Hierarchical Bayesian curve fittin and smoothing. *Canadian J. Statist.*, 20, 35-49.

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

Berger, J. O. and Guglielmi, A. (1999). "Bayesian testing of a parametric model versus nonparametric alternatives," Technical report 99-04, ISDS, Duke University.

Brunner, L.J. (1995). Using the Gibbs sampler to simulate from the Bayes estimate of a decreasing density. *Comm. Statist. A – Theory Methods*, 24, 215-226.

Brunner, L.J. and Lo, A. (1989). Bayes methods for a symmetric unimodal density and its mode. *Ann. Statist.*, 17, 1550-1566.

Bush, C. A. and MacEachern, S. N., (1996), "A semiparametric Bayesian model for randomised block designs," *Biometrika*, 83, 275–285.

Carota, C., Parmigiani, G. and Polson, N. G. (1996), "Diagnostic measures for model criticism," *Journal of the American Statistical Association*, 91, 753 – 762.

Damien, P., Wakefield, J.C. and Walker, S.G. (1998). "Gibbs sampling for Bayesian non-conjugate and hierarchical models using auxiliary variables." *Journal of the Royal Statistical Society, Series B*, to appear.

Doss, H. (1991). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. Florida State University Technical Report No. 850.

Dykstra, R.L. and Laud, P. (1981). A Bayesian nonparametric approach to reliability. *Ann. Statist.*, 9, 356-367.

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

Escobar, M. (1988). Estimating the means of several normal populations by estimating the distribution of the means, Ph.D. thesis, Yale University, New Haven, CT.

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

Escobar, M.D. and West, M. (1998), "Computing nonparametric hierarchical models," in *Practical Nonparametric and Semiparametric Bayesian Statistics* (eds: Dey, D. and Müller, P. and Sinha, D.), 1–22, New York: Springer-Verlag.

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

Florens, Richard, and Rolin (1996), "Bayesian encompassing specifications tests of a parametric model against a nonparametric alternative", T.R., Universite' Catholique de Louvain.

Gasparini, M. (1993). Bayesian density estimation via mixtures of Dirichlet processes. Technical Report No. 93-23, Department of Statistics, Purdue University.

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

Gelfand, A.E. and Kottas, A. (1999) Full Bayesian Inference for the Nonparametric Analysis of Single and Multiple Sample Problems, Technical Report, University of Conneticut.

Green, P. and Richardson, S. (1997). On Bayesian analyisi of mixtures with an unknown number of components (Disc: 758-792). *J. Royal Statist. Soc. Ser. B*, 59, 731-758.

Green, P. and Richardson, S. (1998). "Modelling heterogeneity with and without the Dirichlet process," Technical report, University of Bristol.

Guglielmi, A. (1998). Results with distribution functions of means of a Dirichlet process. *Bollettino della unione matematica Italiana*, 1A (supp. S), 125–128.

Hjort, N.L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.*, 18, 1259-1294.

Huerta, G. and West, M. (1999). Priors and component structures in autoregressive time series models. *J. Royal Statist. Soc. B*, 61, 881–899.

Kleinman K.P., Ibrahim, J.G. (1998). A Semi-parametric Bayesian Approach to the Random Effects Model. *Biometrics*, to appear.

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

Kuo, L. and Mallick, B. (1997). Bayesian Semiparametric Inference for the Accelerated Failure Time Model, *Canadian Journal of Statistics* 25, 457-472.

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.), Kluwer Academic Publishers: Dordrecht, The Netherlands.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *Ann. Statist.*, 20, 1203-1221.

Lavine, M. (1994). More aspects of Polya tree distributions for statistical modelling. *Ann. Statist.*, 22, 1161-1176.

Lavine, M. and Mockus, A. (1995). A nonparametric Bayes method for isotonic regression. *Journal for Statistical Planning and Inference*, 46, 235-248.

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

MacEachern, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. To appear in *Communications in Statistics: Simulation and Computation*, 23, 727–741.

MacEachern, S.N. (1996). Identifiability and Markov chain Monte Carlo methods. Unpublished manuscript.

MacEachern, S.N. (1998), "Computations for MDP models," in *Practical Nonparametric and Semiparametric Bayesian Statistics* (eds: Dey, D. and Müller, P. and Sinha, D.), 23–43, New York: Springer-Verlag.

MacEachern, S.N. and Muller, P. (1998). Estimating mixture of Dirichlet process models, *J. Comp. Graph. Statist.*, 7, 223-238.

Malec, D. and Müller, P. (1999). A Bayesian Semi-Parametric Model for Small Area Estimation, Discussion Paper, Duke University, Durham, NC.

Meng, X.L. and Van Dyk, D.A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, 86, 301–320.

Mukhopadhyay, S. and Gelfand, A.E. (1997). "Dirichlet Process Mixed Generalized Linear Models," *Journal of the American Statistical Association*, 92, 633–639.

Müller, P., Erkanli, A., and West, M. (1996). "Bayesian curve fitting using multivariate normal mixtures," *Biometrika*, 83, 67-79.

Müller, P. and Roeder, K. (1997). "A Bayesian Semiparametric Model for Case-Control Studies With Errors in Variables," *Biometrika*, 84, 523-537.

Müller, P. and Rosner, G. (1997). "A Bayesian population model with hierarchical mixture priors applied to blood count data," *Journal of the American Statistical Association*, 92, 1279-1292.

Neal, R. M., (1998), "Markov chain sampling methods for Dirichlet process mixture models," Technical report, University of Toronto.

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

O'Hagan, A. (1992). "Some Bayesian numerical analysis," in *Bayesian Statistics 4 Proceedings of the Fourth Valencia International Meeting,* eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith, 355-363, Clarendon Press, Oxford.

Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials. *Can. J. Statist.*, 27, 105-126.

Petrone, S. (1999b). Random Bernstein polynomials. *Scand. J. Statist.*, 26, 373-393.

Qin, L. (1998). Nonparametric Bayesian models for item response data. Unpublished Ph.D. dissertation, The Ohio State University.

Quintana, F. (1998). Nonparametric Bayesian analysis for assessing homogeneity in kxl contingency talbes with fixed right margin totals, *J. Amer. Statist. Assoc.*, 93, 1140-1149.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica*, 4, 639-650.

Susarla, J. and van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.*, 71, 897-902.

Tardella, L. and Muliere, P. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Can. J. Statist.*, 26, 283–297.

Verdinelli,I. and Wasserman, L. (1998). Bayesian Goodness of Fit Testing Using Infinite Dimensional Exponential Families, *Annals of Stats*, 20, 1203-1221.

Walker, S. and Damien, P. (1998). Sampling methods for Bayesian nonparametric inference involving stochastic processes, in *Practical Nonparametric and Semiparametric Bayesian Statistics* (eds: Dey, D. and Müller, P. and Sinha, D.), 243–254, New York: Springer-Verlag.

Walker, S. and Muliere, P. (1997). Beta-Stacy processes and a generalization of the Polya urn scheme. *Ann. Statist.*, 25, 1762-1780.

Walker, S.G., Damien, P., Laud, P.W. and Smith, A.F.M. (1999). Bayesian nonparametric inference for distributions and related functions. *J. Royal Statist. Soc. B*, 61, 485-527.

West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. ISDS Discussion Paper # 92-A03, Duke University.

West, M. (1997). Hierarchical mixture models in neurological transmission anslysis. *J. Amer. Statist. Assoc.*, 92, 587-606.

West, M., Müller, P., and Escobar, M. (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.