

# Sequential Importance Sampling for Nonparametric Bayes Models: The Next Generation

Steven N. MacEachern, Merlise Clyde, and Jun S. Liu

There are two generations of Gibbs sampling methods for semi-parametric models involving the Dirichlet process. The first generation suffered from a severe drawback, namely that the locations of the clusters, or groups of parameters, could essentially become fixed, moving only rarely. Two strategies that have been proposed to create the second generation of Gibbs samplers are integration and appending a second stage to the Gibbs sampler wherein the cluster locations are moved. We show that these same strategies are easily implemented for the sequential importance sampler, and that the first strategy dramatically improves results. As in the case of Gibbs sampling, these strategies are applicable to a much wider class of models. They are shown to provide more uniform importance sampling weights and lead to additional Rao-Blackwellization of estimators.

*Key words and phrases:* Beta-binomial, Dirichlet process, Gibbs sampler, importance sampling, MCMC, posterior distribution, Rao-Blackwellization, Sequential Imputation.

*AMS 1991 subject classifications:* 62C10, 62G07.

Steve MacEachern is Associate Professor, Department of Statistics, Ohio State University, Merlise Clyde is Assistant Professor, Institute of Statistics and Decision Sciences, Duke University, and Jun Liu is Assistant Professor, Department of Statistics, Stanford University. The work of the first author was supported in part by the National Science Foundation grant DMS-9305699, and that of the last author by the National Science Foundation grants DMS-9406044, DMS-9501570, and the Terman Fellowship.

Address for Correspondence: Jun S. Liu, Department of Statistics, Stanford University, Stanford, CA 94305, USA

# Sequential Importance Sampling for Nonparametric Bayes Models: The Next Generation

There are two generations of Gibbs sampling methods for semi-parametric models involving the Dirichlet process. The first generation suffered from a severe drawback, namely that the locations of the clusters, or groups of parameters, could essentially become fixed, moving only rarely. Two strategies that have been proposed to create the second generation of Gibbs samplers are integration and appending a second stage to the Gibbs sampler wherein the cluster locations are moved. We show that these same strategies are easily implemented for the sequential importance sampler, and that the first strategy dramatically improves results. As in the case of Gibbs sampling, these strategies are applicable to a much wider class of models. They are shown to provide more uniform importance sampling weights and lead to additional Rao-Blackwellization of estimators.

*Key words and phrases:* Beta-binomial, Dirichlet process, Gibbs sampler, importance sampling, MCMC, posterior distribution, Rao-Blackwellization, Sequential Imputation.

*AMS 1991 subject classifications:* 62C10, 62G07.

# 1 Introduction

Nonparametric and semi-parametric hierarchical Bayes methods have enjoyed a resurgence of interest with the development of modern Monte Carlo techniques. One popular class of models is based on the Dirichlet process (Ferguson, 1973), and has become known as mixture of Dirichlet process (MDP) models (Antoniak, 1974). The advantage of a MDP model over a standard parametric hierarchical model as in Lindley and Smith (1972) is that it allows for more flexibility through the incorporation of a nonparametric hierarchical distribution. In such a model, the observable random variable  $X_i$  is generated according to the following process:

$$\begin{aligned} X_i | \theta_i &\sim g_{\theta_i}(\cdot) \\ \theta_i | F &\sim F \\ F &\sim Dir(\alpha_\nu) \\ \nu &\sim h(\cdot). \end{aligned}$$

For a sample of size  $n$  from this process, the  $\theta$  are iid with given  $F$  and the  $X_i$  are mutually independent conditional on the parameters  $\theta_i$ .  $Dir(\alpha_\nu)$  represents a Dirichlet process characterized by measure  $\alpha_\nu$ . Early work related to theory and computation of such models can be found in Antoniak (1974), Berry and Christensen (1979), Kuo (1986), just to start a list. Some more recent work in connection with Monte Carlo methods was done by Escobar (1994), Escobar and West (1994), Doss (1994), Kong, Liu, and Wong (1994), MacEachern (1994), etc. The models have been further developed and applied to a wide variety of settings, with much research currently under way.

Escobar (1994) describes the use of Gibbs sampling for a class of normal MDP models. Others have extended this work, creating more realistic models and devising more efficient Gibbs samplers. The more efficient samplers rely on one of two improvements: either a simple integration, if feasible, is used to collapse the parameter space upon which the Gibbs sampler runs, or the addition of an extra stage for the Gibbs sampler, which results in quicker convergence and mixing over the parameter space. These modifications also suggest a stronger form of Rao-Blackwellization that has been empirically demonstrated to improve estimation. Together, the quicker mixing and better estimation result in dramatically improved computations.

Kong et al. (1994) and Liu (1996) propose the use of sequential importance sampling (SIS) for nonparametric Bayes models. SIS provides a method whereby the data automatically help to construct an importance sampling density. It relies on much the same calculations as the Gibbs sampler, and so provides a rival Monte Carlo technique for large, complex problems. One potential advantage of the SIS is that it does not rely on an underlying Markov chain as does Gibbs sampling. Instead, many independent and identically distributed replicates are

run to create an importance sample. This may result in more efficient estimators, and greatly simplifies assessment of the accuracy of the estimators.

In spite of the great success in many problems, the development of modern Monte Carlo techniques for nonparametric Bayes models has not completely solved the computational question. In particular, there is concern about the methods' effectiveness for large problems. In these cases, squeezing every bit of efficiency out of the computational routines is necessary. In this paper, we exploit the similarities between the Gibbs sampler and the SIS, bringing over the improvements for Gibbs sampling algorithms to the SIS setting for nonparametric Bayes problems. These improvements result in an improved sampler and help satisfy questions of Diaconis (1995) pertaining to convergence. Such an effort can see wide applications in many other problems related to dynamic systems where the SIS is useful (Berzuini et al. 1996; Liu and Chen 1996).

Section 2 describes the specific model that we consider. For illustration we focus discussion on the beta-binomial model, although the methods are applicable to other conjugate families. In Section 3, we describe the first generation of the SIS and Gibbs sampler in this context, and present the necessary conditional distributions upon which the techniques rely. Section 4 describes the alterations that create the second generation techniques, and provides specific algorithms for the model we consider. Section 5 presents a comparison of the techniques on a large set of data. Section 6 provides theory that ensures the proposed methods work and that is generally applicable to many other problems using importance sampling approaches. The final section presents discussion.

## 2 The Model

The model that we consider is representative of the general class of mixture of Dirichlet process models. The motivation for the model is hierarchical Bayes (or empirical Bayes) problems, where we wish to pool information across a number of similar experiments. We use the model

$$F \sim Dir(\alpha) \tag{1}$$

$$\theta_1, \dots, \theta_n | F \sim F \tag{2}$$

$$X_i | \theta_i \sim Binom(t_i, \theta_i), \tag{3}$$

where  $\theta_1, \dots, \theta_n | F$  are independent and identically distributed and the  $X_i | \theta_i$  are independent binomial random variable from a sample of size  $t_i$ , for  $i = 1, \dots, n$ .

### 2.1 Simple properties

This model places a mixture of binomial distributions on the observable  $X_i$ , where  $F$  is the mixing distribution. The Dirichlet process places a distribution over this mixing distribution,

and if  $\alpha$  assigns mass to each open interval on  $[0, 1]$ , then the support of the distribution on  $F$  is all distributions on  $[0, 1]$ . The model allows for pooling of information across the samples, in that  $X_i$  will have an effect on  $\theta_j|x$  for  $j \neq i$  through the hierarchical stage involving the Dirichlet process.

With the model (1) - (3), a variety of features of the posterior are of interest: Berry and Christensen (1979) use the model for the quality of welding material submitted to a naval shipyard, implying an interest in  $\theta_i|x$ . Liu (1996) uses the model for the results of flicks of thumbtacks, and focuses on the distribution of  $\theta_{n+1}|x$ . Gopalan (1994) uses this model for Bayesian multiple comparisons, and so focuses on  $Pr(\theta_i = \theta_j|x)$ . Other natural estimands are the predictive distribution of a new observable,  $X_{n+1}|x$ , the results of a future binomial sample from a distribution with one of the current  $\theta_i$ , and the posterior on the mixing distribution,  $F|x$ . We may obtain estimates of all of these quantities on the basis of the simulation techniques that we present.

The Polya urn scheme representation of the Dirichlet process (Blackwell and MacQueen, 1973) provides the basis for most modern computational strategies: for exceptions, see Doss (1994), Gelfand and Kuo (1991), and Kuo and Smith (1992). Let the measure  $\alpha$  be written as:  $\alpha((-\infty, x]) = MF_0(x)$ , where  $M$  is a normalizing constant, and  $F_0$  a probability cdf. Then  $(\theta_1, \dots, \theta_n)$  has a distribution identical to the one produced by the following Polya Urn scheme:  $\theta_1 \sim F_0$ , and conditional on  $(\theta_1, \dots, \theta_{i-1})$ ,

$$\theta_i \sim \begin{cases} F_0 & \text{with probability } M/(M+i-1) \\ \delta_{\theta_j} & \text{with probability } 1/(M+i-1) \quad \text{for } j = 1, \dots, i-1, \end{cases} \quad (4)$$

where  $\delta_{\theta_j}$  represents the distribution that is degenerate at  $\theta_j$ . This representation of the joint distribution of  $(\theta_1, \dots, \theta_n)$  implicitly marginalizes over the mixing distribution  $F$ .

A realization of the Polya urn scheme partitions the vector  $\theta$  into a batch of clusters, where the  $\theta_i$  belonging to the same cluster assume the same value and  $\theta_i$  belonging to different clusters may assume different values. In the case of a continuous  $F_0$ , the  $\theta_i$  belonging to different clusters assume different values with probability one. We note that this view of the Dirichlet process also leads to the representation of  $\theta$  as  $(s, \theta^*)$  where the vector  $s$  captures the clustering of the  $\theta_i$  and  $\theta^*$  captures the locations of the clusters. The relationship between  $\theta$  and  $(s, \theta^*)$  is given by  $\theta_i = \theta_{s_i}^*$ ,  $i = 1, \dots, n$ . Equivalent to  $(s, \theta^*)$ , we often write  $(s, \theta)$ .

## 2.2 Notation

We introduce the following notation to make the subsequent algorithms clear. For the remainder of the paper,  $\pi$  will provide generic notation for the prior or posterior distribution or density. We assume that  $F_0$  is a beta distribution with parameters  $\alpha_0$  and  $\beta_0$ . During the algorithms' progress, some of the  $\theta_i$  will be grouped together in a cluster. We will let  $k$  represent the number of clusters at a given point in time, and will let  $n_j$  represent the number of the  $\theta_i$  in cluster

$j$ . We let  $\alpha_j = \alpha_0 + \sum_{i'|s_{i'}=j} x_{i'}$  and  $\beta_j = \beta_0 + \sum_{i'|s_{i'}=j} (t_{i'} - x_{i'})$ . The algorithms allow  $\theta_i$  to begin a new cluster. When considering this potential new cluster, we label it  $k + 1$ , and define  $\alpha_{k+1} = \alpha_0$ ,  $\beta_{k+1} = \beta_0$ , and  $n_{k+1} = M$ . The  $n_j$ ,  $\alpha_j$ , and  $\beta_j$  are implicitly assumed to include only those observations which  $i' < i$  for the SIS. The subscript  $< i$  will denote these conditions.

In the sequel, we discuss importance sampling. In general discussion, we use  $f$  for the target density,  $g$  for the importance sampling density, and  $h$  for the function of interest. We use  $z$  for the argument of these densities, and let  $w$  denote the importance sampling weight.

### 3 Simulation Techniques

The model (1) - (3), though simple to describe, results in a posterior that is difficult to evaluate. The clustering of the  $\theta_i$  described by the Polya urn scheme results in a posterior that may be represented as an enormous mixture. Each component of the mixture corresponds to a partition of  $\theta$  into clusters, and the number of components in the mixture grows exponentially. Because of the analytic intractability of this type of model, early investigators developed approximations (see Berry and Christensen, 1979) or simulation methods (see Kuo, 1986).

For both the SIS and the Gibbs sampler, the key calculations stem from the Polya urn representation of the Dirichlet process:

$$\begin{aligned} Pr(s_i = j \mid s_{<i}, \theta_{<i}^*, x_{<i}, x_i) &\propto Pr(X_i = x_i, s_i = j \mid s_{<i}, \theta_{<i}^*, x_{<i}) \\ &= Pr(s_i = j \mid s_{<i}, \theta_{<i}^*, x_{<i}) Pr(X_i = x_i \mid s_{<i}, \theta_{<i}^*, x_{<i}, s_i) \\ &= \begin{cases} \frac{n_j}{M+i-1} \binom{t_i}{x_i} \theta_{s_i}^{*x_i} (1 - \theta_{s_i}^*)^{t_i - x_i}, & \text{for } j = 1, \dots, k \\ \frac{M}{M+i-1} \binom{t_i}{x_i} \frac{B(\alpha_0 + x_i, \beta_0 + t_i - x_i)}{B(\alpha_0, \beta_0)}, & \text{for } j = k + 1 \end{cases} \end{aligned}$$

where  $B(\alpha, \beta)$  is the beta function. Simplifying, we have

$$Pr(s_i = j \mid s_{<i}, \theta_{<i}^*, x_{<i}, x_i) \propto \begin{cases} n_j \theta_{s_i}^{*x_i} (1 - \theta_{s_i}^*)^{t_i - x_i}, & \text{for } j = 1, \dots, k \\ MB(\alpha_0 + x_i, \beta_0 + t_i - x_i) / B(\alpha_0, \beta_0), & \text{for } j = k + 1. \end{cases} \quad (5)$$

The SIS relies on a trick to construct an importance sampling density. Once constructed, this importance sampler behaves like any other: assume a target density  $f(z)$ , an importance sampling density,  $g(z)$ , and a function of interest  $h(z)$ . A large number,  $R$ , of independent draws are made from  $g(\cdot)$ , resulting in  $z^{(1)}, \dots, z^{(R)}$ . The importance sampling weights are calculated as  $w_r = f(z^{(r)})/g(z^{(r)})$ , and  $E^f[h(z)]$  is approximated by  $\hat{h} = \sum_{r=1}^R w_r h(z^{(r)})/R$ . Under relatively mild conditions ( $f$  and  $g$  are mutually absolutely continuous and  $\text{var}(w_1 h_1) < \infty$  are sufficient),  $\hat{h}$  provides an unbiased estimator of  $E^f[h(z)]$  and, as  $R \rightarrow \infty$ ,  $\hat{h} \rightarrow E^f[h(z)]$ . In practice, when  $f$  is evaluated only up to a constant of proportionality, and also for the purpose of reducing Monte Carlo variation, the weights are renormalized to have mean 1:

$$w_r^* = R w_r / \sum_{j=1}^R w_j.$$

Importance samplers are used for two main reasons: either to reduce the variance in a simulation, or because it is difficult to obtain a sample from the density of interest. The latter use motivates our choice of the SIS. With this use in mind, we judge the quality of our importance sampler by how closely we approximate the target distribution. A measure of this is the variance of the importance sampling weights, or equivalently the effective sample size  $ESS = R/(1 + \text{var}(w))$ . In the sequel, we estimate  $ESS$  by replacing  $\text{var}(w)$  with  $\hat{\text{var}}(w^*)$ . The following scheme was proposed in Liu (1996):

*Sequential Importance Sampler S1 (Liu, 1996):*

Repeat steps (A) and (B) for  $i = 1, \dots, n$ .

A Generate  $(s_i, \theta_i) | (s_{<i}, \theta_{<i}^*, x_{<i}, x_i)$  by first generating  $s_i$  from the distribution in (5). If  $s_i \leq k$ , then set  $\theta_i = \theta_{s_i}^*$ . If  $s_i = k+1$ , then generate a new  $\theta_{k+1}^*$  from a  $Beta(\alpha_0 + x_i, \beta_0 + t_i - x_i)$  distribution.

B Calculate  $Pr(X_i = x_i | s_{<i}, \theta_{<i}, x_{<i}), i = 1, \dots, n$ .

After values  $(s_i, \theta_i), i = 1, \dots, n$ , have been generated, calculate

$$w_r = \prod_{i=1}^n Pr(X_i = x_i | s_{<i}, \theta_{<i}, x_{<i})$$

Repeat this procedure to obtain  $R$  replicates. The weights are then normalized so that  $w_r^* = R w_r / (\sum_{j=1}^R w_j)$ . We note that the probabilities needed to calculate the weights must be evaluated in order to generate values of  $(s_i, \theta_i)$  in step (A). Finding the weights requires almost no additional computational effort.

*Gibbs sampler (Escobar, 1994):*

Each  $\theta_i$  is, in turn, viewed as the last observation (of  $n$ ) from a Polya urn scheme. The remaining  $n - 1$  observations form  $k$  clusters with locations  $\theta_1^*, \dots, \theta_k^*$ . A cluster membership and location are generated for  $\theta_i$ .

A second generation of algorithms has been developed to speed the Gibbs sampler. These algorithms work by improving the mixing of the sampler, and are motivated by the difficulty encountered by the first generation of Gibbs samplers. The locations of the clusters may become stuck, and only rarely move. When this problem is encountered, the Gibbs sampler will mix very slowly over the parameter space, and will, as a practical matter, provide poor estimates.

## 4 More Efficient Algorithms

The two principal fixes for Escobar's Gibbs sampler both aim at alleviating the difficulty with the sticky cluster locations. The first fix, described in MacEachern (1994), accomplishes this by

removing the locations  $\theta_i$  entirely via integration. The state space of the Markov chain on which the Gibbs sampler is defined is then collapsed to the space of the clustering vector  $s$ . The second fix, introduced by Bush and MacEachern (1996), is appropriate when integration is difficult or time consuming. With this fix, the cluster locations are moved at appropriate intervals, say in an extra stage appended to the end of each complete cycle of the Gibbs sampler.

#### 4.1 Integration

For a new  $\theta_i$ , the prior probability of its joining cluster  $j$ ,  $j = 1, \dots, k$ , is proportional to  $n_j$ , and that of its beginning a new cluster,  $j = k + 1$ , is proportional to  $M$ . The effective prior distribution of the location for the current cluster  $j$  is a  $Beta(\alpha_j, \beta_j)$  distribution, where  $\alpha_j$  and  $\beta_j$  are as defined in Section 2.1. The conditional probability for the data, given that  $\theta_i$  is in cluster  $j$  and integrating over the cluster location, is then

$$Pr(X_i = x_i | s_{<i}, s_i = j, x_{<i}) \propto \int B(\alpha_j, \beta_j)^{-1} \theta^{\alpha_j + x_i - 1} (1 - \theta)^{\beta_j + t_i - x_i - 1} d\theta.$$

Putting the pieces together, we have the following conditional probabilities:

$$Pr(s_i = j | s_{<i}, x_{<i}, x_i) \propto n_j B(\alpha_j + x_i, \beta_j + t_i - x_i) / B(\alpha_j, \beta_j) \quad \text{for } j = 1, \dots, k + 1,$$

where we set  $n_{k+1} = M$ ,  $\alpha_{k+1} = \alpha_0$ , and  $\beta_{k+1} = \beta_0$ . Using this we can specify a version of the sequential importance sampler based on integrating out the locations.

#### *Sequential Importance Sampler S2*

For  $i = 1, \dots, n$ , repeat steps (A) and (B).

A Generate  $s_i$  from the multinomial distribution with

$$Pr(s_i = j | s_{<i}, x_{<i}, x_i) \propto n_j B(\alpha_j + x_i, \beta_j + t_i - x_i) / B(\alpha_j, \beta_j) \quad \text{for } j = 1, \dots, k + 1.$$

where we set  $n_{k+1} = M$ ,  $\alpha_{k+1} = \alpha_0$ , and  $\beta_{k+1} = \beta_0$ .

B Calculate

$$Pr(X_i = x_i | s_{<i}, x_{<i}) \propto \sum_{j=1}^{k+1} \frac{n_j B(\alpha_j + x_i, \beta_j + t_i - x_i)}{(M + i - 1) B(\alpha_j, \beta_j)},$$

for  $i = 1, \dots, n$ .

After values  $s_i$ ,  $i = 1, \dots, n$ , have been generated, calculate the importance sampling weights

$$w_r = \prod_{i=1}^n Pr(X_i = x_i | s_{<i}, x_{<i}).$$

Repeat to obtain  $R$  replicates. Then normalize the weights so that  $w_r^* = R w_r / (\sum_{j=1}^R w_j)$ .

The underlying parameter space for  $s$  stems from the set of all partitions of  $\theta$  and the indexing generated by the sequential nature of the sequential importance sampler. Both the importance sampling density and the posterior assign positive probability to each element of this space. The weights follow from Kong et al. (1994). In Section 6 we show that S2 is a Rao-Blackwellization of S1 and is therefore always more efficient.

## 4.2 Gibbs Iteration within SIS

In designing a Gibbs sampler, it is often useful for improving its mixing rate to introduce special moves that help the sampler escape from a local mode. This fix for the Dirichlet process related problems is described in Bush and MacEachern (1996) and West, Müller and Escobar (1994). For the SIS, we introduce two similar approaches. For S1, we move the cluster locations once in a while as we proceed through a single replicate. Fix a set of times to move the cluster locations, say  $T = (t_1, \dots, t_l)$ , we have the following scheme:

### *Sequential Importance Sampler S3*

For  $i = 1, \dots, n$ , repeat steps (A) and (B).

A Stage 1. Generate  $(s_i, \theta_i)$  as in the sampler S1.

Stage 2. If  $i \in T$ , then for  $j = 1, \dots, k$ , generate  $\theta_j^*$  from a  $Beta(\alpha_j, \beta_j)$  distribution, where  $k$  is the number of clusters at time  $i$ .

B Calculate  $Pr(X_i = x_i | s_{<i}, \theta_{<i}, x_{<i}), i = 1, \dots, n$  as the generations proceed. These probabilities are calculated immediately before the value of  $s_i$  is generated. They are not recalculated when the new values of  $\theta$  are generated. After values of  $(s_i, \theta_i), i = 1, \dots, n$  have been generated, compute  $w_r = \prod_{i=1}^n Pr(X_i = x_i | s_{<i}, \theta_{<i}, x_{<i})$ .

Repeat to obtain  $R$  replicates. Then normalize the weights so that  $w_r^* = R w_r / (\sum_{j=1}^R w_j)$ .

### *Sequential Importance Sampler S4*

For  $i = 1, \dots, n$ , repeat steps (A) and (B).

A Stage 1. Generate  $s_i$  as in the sampler S2.

Stage 2. If  $i \in T$ , then iterate through  $s_1, \dots, s_{i-1}$  using the Gibbs sampler. That is, each  $s_t$  is substituted by a draw from  $Pr(s_t | s_{<i[-t]}, x_{<i}),$  for  $t = 1, \dots, i - 1$ , where  $s_{<i[-t]}$  is the collection of all  $s_j$  with  $j \neq t$  and  $j < i$ .

B Calculate  $Pr(X_i = x_i | s_{<i}, x_{<i}), i = 1, \dots, n$  the same way as in S2. These probabilities are calculated immediately before the value of  $s_i$  is generated. They are not recalculated when the new values of the  $s_t$  are generated.

Obtain weights as in the previous procedures.

### 4.3 Estimation

In Section 2, we noted that there are many features of interest of the posterior. This section is devoted to a discussion of estimation for these features. The over-riding principle that guides our choice of an estimator, both for Gibbs sampling and for sequential importance sampling, is Rao-Blackwellization. There are strong parallels in estimation for the two Monte Carlo techniques, with the main difference being the inclusion of weights for the sequential importance sampler. See MacEachern (1994) or MacEachern and Müller (1994) for a discussion of estimation for the Gibbs sampler, and Kong et al.(1994) or Liu (1996) for discussion for the sequential importance sampler. We turn to the specific estimators for the sequential importance sampler.

The density for  $\theta_{n+1}$  for the next observation given the current data  $x$  may be estimated by

$$\hat{\pi}(\theta_{n+1}|x) = \sum_{r=1}^R \frac{w_r^*}{M+n} \left( \sum_{j=1}^{k^{(r)}+1} n_j^{(r)} \text{Beta}(\alpha_j^{(r)}, \beta_j^{(r)}) \right), \quad (6)$$

where the superscript  $(r)$  denotes the replicate and  $\text{Beta}(\cdot, \cdot)$  represents the Beta density. Likewise, the predictive density  $X_{n+1}|x$  for a new observation may be estimated by

$$\hat{Pr}(X_{n+1}|X_{<(n+1)}) = \sum_{r=1}^R \frac{w_r^*}{M+n} \sum_{j=1}^{k^{(r)}+1} n_j^{(r)} \text{BB}(x_{n+1}; \alpha_j^{(r)}, \beta_j^{(r)}, t_{n+1}). \quad (7)$$

Here, we define the beta-binomial probabilities

$$\text{BB}(x'; \alpha, \beta, t) = \binom{t}{x'} \int_0^1 B(\alpha, \beta)^{-1} u^{\alpha-1} (1-u)^{\beta-1} u^{x'} (1-u)^{t-x'} du \quad (8)$$

$$= \binom{t}{x'} B(\alpha + x', \beta + t - x') / B(\alpha, \beta). \quad (9)$$

The density of  $\theta_i|x, i \leq n$ , may be simply estimated by averaging over the groups in which  $\theta_i$  falls, obtaining

$$\hat{\pi}(\theta_i|x) = \sum_{r=1}^R w_r^* \text{Beta}(\alpha_{s_i}, \beta_{s_i}).$$

However, we may also use Rao-Blackwellization to reduce the estimation variance, which is done through temporarily removing  $(s_i, \theta_i)$  from the vector  $(s, \theta)$  and performing calculations similar to those in a Gibbs sampling step. This gives us

$$\hat{\pi}(\theta_i|x) = \frac{1}{C} \sum_{r=1}^R w_r^* \sum_{j=1}^{k^{(r)}+1} n_j^{(r)} \frac{B(\alpha_j + x_i, \beta_j + t_i - x_i)}{B(\alpha_j, \beta_j)} \text{Beta}(\alpha_j + x_i, \beta_j + t_i - x_i) \quad (10)$$

where  $C = \sum_{j=1}^{k^{(r)}+1} n_j^{(r)} B(\alpha_j + x_i, \beta_j + t_i - x_i) / B(\alpha_j, \beta_j)$ . Other functions of interest may be evaluated by means of integrating these functions against the beta densities inside these summations, or by taking draws of  $\theta_i$  from these distributions and evaluating the function at these draws. The former method is preferable when the integrals are not too time consuming.

A further technique is occasionally available for improving the estimator  $\hat{\pi}(\theta_i|x)$ . When some of the binomial data are identical, say  $t_i = t_{i'}$  and  $x_i = x_{i'}$ , then we know that  $\hat{\pi}(\theta_i|x) \equiv \hat{\pi}(\theta_{i'}|x)$ . Averaging these two estimators leads to some improvement. It also brings a feature of the estimated posterior into agreement with a known feature of the posterior, something we find comforting.

Both  $\hat{\pi}(\theta_i|x)$  and  $\tilde{\pi}(\theta_i|x)$  are mixtures of Beta distributions, hence are preferable to the early estimators used in Liu (1996), say,

$$\tilde{\pi}(\theta_{n+1}|x) = \sum_{r=1}^R \frac{w_r^*}{M+n} (MF_0 + \sum_{j=1}^n \delta_{\theta_j^{(r)}}). \quad (11)$$

There is no need for binning or kernel smoothing to produce a picture of a density that is known to be continuous. But the use of smoothers may still be helpful when  $\theta$  is partitioned into a small number of clusters, where each cluster has many observations. In this instance, the Beta distributions in the mixture are very peaked, and the estimator  $\hat{\pi}$  may still be overly bumpy.

Estimators may also be developed for models (1) - (3) when the base measure  $\alpha$  of the Dirichlet process is not a beta distribution or when there is a distribution over the mass parameter  $M = \|\alpha\|$ . Perhaps the simplest way to construct these estimators is through further use of importance sampling techniques (Liu 1996). For the sampler S2 which is based only on  $s$ , reweighting may be accomplished by generating  $\theta|(s, x)$  and applying Liu's method. When feasible, this generation may be avoided with an integration to find the expected weight.

Another feature of interest is the mixing distribution  $F$ . Previous work in the area has focused on the point-wise posterior mean of this distribution,  $F(\theta)|x$ , for values of  $\theta$ . When looking at all values of  $\theta$ , this mean gives the predictive density for  $\theta_{n+1}|x$  that was discussed earlier, and (6) provides a means to estimate it. We may also sample  $F$  from the posterior, or more properly sample almost all of the mass of  $F$  from an approximation to the posterior, with the upcoming algorithm. The justification for this algorithm is a consequence of elementary properties of the Dirichlet process, and is standard.

*Algorithm for Drawing  $F|x$ :*

1. Choose a small value  $\epsilon$ .
2. Choose a replicate from the importance sample. Select replicate  $r$  with probability  $w_r^*$  for  $r = 1, \dots, R$ .

| X        | 0    | 1 | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  |
|----------|------|---|----|----|----|----|----|----|----|----|
| counts   | 0    | 3 | 13 | 18 | 48 | 47 | 67 | 54 | 51 | 19 |
| mean     | 5.84 |   |    |    |    |    |    |    |    |    |
| variance | 3.46 |   |    |    |    |    |    |    |    |    |

Table 1: Summary of tack data.

3. Draw the amounts of mass for the discrete points in the distribution. Continue until  $\prod_{j=1}^i (1 - m_j) < \epsilon$ .

For  $i = 1, \dots, k$ , draw  $m_i \sim \text{Beta}(n_i, M + n - \sum_{j=1}^i n_j)$

For  $i > k$  draw  $m_i \sim \text{Beta}(1, M)$ .

Assume that there are  $I$  lumps of mass when this stage of the algorithm ends. Then, the mass assigned to lump  $i$  is given by  $m_i^* = m_i \prod_{j=1}^{i-1} (1 - m_j)$ ,  $i = 1, \dots, I$ .

4. Draw the locations of the lumps of mass.

For  $i = 1, \dots, k$ , draw  $\xi_i \sim \text{Beta}(\alpha_i, \beta_i)$ .

For  $i > k$ , draw  $\xi_i \sim \text{Beta}(\alpha_0, \beta_0)$ .

5. Approximate  $F$  as  $\sum_{i=1}^I m_i^* \delta_{\xi_i}$ .

## 5 An Example

We analyze the data from Beckett and Diaconis (1994), which consists of the results from flicking thumbtacks. Liu (1996) used the importance sampler S1 to analyze this data, focusing on the distribution of  $\theta_{n+1}|x$ . For the data set, 320 tacks were flicked, 9 times each, and the number of times that they landed point up was recorded. A summary of the data is recorded in Table 1. We consider several analyses with the baseline model given by  $M = .1, 1, 5$ , and 10. In all cases, we take  $F_0$  to be a uniform distribution. 10,000 replicates of the sequential importance samplers were used for the comparative analysis.

In our experience we found that the sampler S3 did not result in an improvement in ESS, so the remainder of the discussion will focus on the comparison between the original sequential importance sampler S1 and the sequential importance sampler S2 that is obtained by integrating out the locations. Table 2 contains the variances of the normalized sequential importance sampling weights and the effective sample sizes, ESS, under S1 and S2 for the four values of  $M$ . The integration appears to greatly reduce the variability of the importance sampling weights, leading to a larger effective sample size. In particular, the integration method (S2) has an ESS that

| M        | .1    | 1     | 5    | 10   |
|----------|-------|-------|------|------|
| S1       |       |       |      |      |
| variance | 378   | 43    | 34   | 33   |
| ESS      | 26    | 227   | 285  | 294  |
| S2       |       |       |      |      |
| variance | 94.94 | 11.29 | 3.08 | 1.67 |
| ESS      | 104   | 814   | 2452 | 3751 |
| Gain     | 4.00  | 3.58  | 8.60 | 12.7 |

Table 2: Comparison of the original sequential importance sampler (S1) and the sequential importance sampler with integration (S2) for different values of the mass parameter  $M$ .

is 12.7 times the ESS from the original sequential importance sampler (S1) for  $M = 10$ . The gains appear to be greater for larger values of  $M$ .

Estimates of the predictive density (7) for  $X_{n+1}$  for the next experiment with  $t_{n+1} = 9$  for the four values of  $M$  are shown in Figure 1. The plots are virtually identical. Using 10 batches of 1000 iterations to estimate the densities results in very similar plots. The most noticeable difference occurs for the small values of  $X_{n+1}$ . This difference is attributable to the mass of the Dirichlet process: The estimate (7) is a convex combination of estimates from the prior in cluster  $k + 1$ , and the data for clusters  $1, \dots, k$ . The weights for the convex combination are  $M/(M + n)$  and  $n/(M + n)$ . As  $M$  ranges from 0.1 to 10,  $Pr(X_{n+1} = 0|x)$  increases along with the weight given to the prior.

Figure 2 shows the effect of changing  $M$  on the distribution of  $\theta_{n+1}|x$ . Each density was estimated using 10,000 iterations of the S2 sequential importance sampler. Unlike the densities for  $X_{n+1}$  given  $x$ , the value of  $M$  in the prior specification has a large impact on the distribution for  $\theta_{n+1}|x$ . Each run of 10,000 was also divided into ten batches of 1,000 and the density was estimated from each batch (Figure 3) to address the issue of sampling variation in the density estimation. While for  $M = 1$ , it is clear that there are two peaks, there is a lot of uncertainty about the exact location and height in the density. This is even more pronounced with  $M = .1$  and may be related to the larger variation in the importance sampling weights for smaller  $M$ . We note from an examination of these plots that as  $M$  increases, the peaks gradually merge together. The qualitative differences for various values of  $M$  do not disappear as the Monte Carlo size increases.

## 6 Theory

The previous sections dealt with the construction of algorithms and estimates based on the improved SIS algorithms. In this section, we verify that these algorithms are legitimate and can be applicable to a wider class of problems. They are presented at a moderate level, omitting measure theoretic details.

### 6.1 Mixing MCMC with Importance Sampling

Let the variate  $z$  assume values in a subset of  $R^n$ , and let  $f$  and  $g$  have the same support in  $R^n$ . We define the weight function as  $w(z) = f(z)/g(z)$ . Consequently,  $w(z)g(z) = f(z)$  for all  $z$ . Furthermore, we let  $P(z, z')$  be the transition kernel for a Markov chain on  $R^n$  with  $f$  as its invariant distribution. The following theorem, and indeed algorithm S3, is motivated by the idea that the weighted importance sample is in practice equivalent to a random sample from  $f$ .

**Theorem 6.1** *Assume that  $w, f, g,$  and  $P$  are as defined above. Then  $\int w(z)g(z)P(z, z')dz = f(z')$  for all  $z'$ .*

*Proof:*

$$\begin{aligned} \int w(z)g(z)P(z, z')dz &= \int f(z)P(z, z') \\ &= f(z'). \end{aligned}$$

The last equality follows from the invariance of  $f$  under  $P$ .  $\diamond$

The sample that we obtain after generation from  $g$  and the transition according to  $P$  may be thought of in two parts: a generation of  $z$  according to a distribution that is difficult to specify and its accompanying weight  $w$ . But we retain the key property that the weighted average of our  $z$ 's can be used to estimate features of the target distribution  $f$ . The proposed Algorithms S3 and S4 iterate the two steps. A portion of the parameter vector is initially generated, then a transition is made according to some kernel  $P$ , then generation of more of the parameter vector, another transition according to another kernel, and so on. Intuitively, one step of transition  $P$  brings the trial distribution  $g$  closer to  $f$  which benefits the latter generations. and makes estimation less variable.

Theorem 6.2 extends the earlier result to this iterated situation by considering each of the two types of steps. We define  $z_1$  and  $z_2$  to be portions of the parameter vector  $z$ . We let  $g_1(z_1, w_1)$  represent the joint density for  $(z_1, w_1)$  based on a legitimate importance sampler for the target distribution  $f_1(z_1)$ . Note that  $w_1$  is not necessarily  $f_1(z_1)/g_1(z_1)$ , but only that  $E[w_1|z_1] = f_1(z_1)/g_1(z_1)$ . Let  $g_{2|1}(z_2|z_1)$  be a conditional distribution, and define  $w_2 = f_{2|1}(z_2)/g_{2|1}(z_2)$ . Let  $g_{12}(z, w_1)$  denote the joint distribution of these quantities and let  $w = w_1w_2$ . We have the following theorem.

**Theorem 6.2** *Assume that  $f$  and  $P$  are as defined above. Suppose  $E(w_1 | z_1) = f_1(z_1)$  for all  $z_1$ . Then*

(A)  $\int \int w_1 g_1(z_1, w_1) P(z_1, z'_1) dw_1 dz_1 = f_1(z'_1)$  for all  $z'_1$ .

(B)  $\int w g_{12}(z, w_1) dw_1 = f(z)$  for all  $z$ .

*Proof:* For (A) we have

$$\int \int w_1 g_1(z_1, w_1) P(z_1, z'_1) dw_1 dz_1 = \int f_1(z_1) P(z_1, z'_1) dz_1 = f_1(z'_1).$$

Then (B) follows from

$$\begin{aligned} \int w g_{12}(z, w_1) dw_1 &= \int w_1 \int w_2 g_{2|1}(z_2 | z_1) g_1(z_1, w_1) dw_1 \\ &= f_{2|1}(z_2 | z_1) \int w_1 g(z_1, w_1) dw_1 = f(z) \quad \diamond \end{aligned}$$

Taken together, the above results suggest that we may freely mix Gibbs sampling or other steps based on a conditional generation into the interior of a sequential importance sampling algorithm. We may also incorporate other steps such as Metropolis-Hastings steps that retain the posterior as an invariant distribution. We do need to take care to specify how we will mix in these steps, by choosing beforehand when they will be implemented. If choice of a step is governed by the state  $z$ , or the weight  $w_1$ , we may arrive at an illegitimate importance sampler.

## 6.2 Rao-Blackwellizing the Importance Sampling

The next theorem provides a theoretical reason for the improvement that we see with the SIS algorithm S2 in which the state space is collapsed.

**Theorem 6.3** *Let  $f(z_1, z_2)$  and  $g(z_1, z_2)$  be two probability densities, and the support of  $f$  is a subset of that of  $g$ . Then*

$$\text{var}_g \left\{ \frac{f(Z_1, Z_2)}{g(Z_1, Z_2)} \right\} \geq \text{var}_g \left\{ \frac{f_1(Z_1)}{g_1(Z_1)} \right\},$$

where  $f_1(z_1) = \int f(z_1, z_2) dz_2$  and  $g_1(z_1) = \int g(z_1, z_2) dz_2$  are marginal densities. The variances are taken with respect to  $g$ .

*Proof:* It is easy to see that

$$\frac{f_1(z_1)}{g_1(z_1)} = \int \frac{f(z_1, z_2)}{g_1(z_1) g_{2|1}(z_2 | z_1)} g_{2|1}(z_2 | z_1) dz_2 = E_g \left\{ \frac{f(Z_1, Z_2)}{g(Z_1, Z_2)} \mid Z_1 = z_1 \right\}.$$

Hence

$$\text{var}_g \left\{ \frac{f(Z_1, Z_2)}{g(Z_1, Z_2)} \right\} \geq \text{var}_g \left\{ E_g \left[ \frac{f(Z_1, Z_2)}{g(Z_1, Z_2)} \mid Z_1 \right] \right\} = \text{var}_g \left\{ \frac{f_1(Z_1)}{g_1(Z_1)} \right\}.$$

We can also obtain an explicit expression of the variance reduction:

$$\text{var}_g\left\{\frac{f(Z_1, Z_2)}{g(Z_1, Z_2)}\right\} - \text{var}_g\left\{\frac{f_1(Z_1)}{g_1(Z_1)}\right\} = E_g\left\{\text{var}_g\left[\frac{f(Z_1, Z_2)}{g(Z_1, Z_2)} \mid Z_1\right]\right\},$$

which, in ANOVA terminology, is the average “within group” variation with group indexed by  $Z_1$ .  $\diamond$

The estimation method of Section 4, and also scheme S2, can be more generally regarded as the Rao-Blackwellization of an importance sampler. More precisely, if we have an importance sampling estimate  $\hat{\theta}$  of quantity  $\theta = E_f\{h(Z_1, Z_2)\}$ , then

$$\begin{aligned} E_g\{w(Z_1, Z_2)h(Z_1, Z_2) \mid Z_2 = z_2\} &= \int h(z_1, z_2)\frac{f(z_1, z_2)}{g(z_1, z_2)}g_{1|2}(z_1 \mid z_2)dz_1 \\ &= w(z_2)E_f\{h(Z_1, Z_2) \mid Z_2 = z_2\}; \end{aligned}$$

Hence, an importance sampling estimate using  $w(Z_1, Z_2)$  and  $h(Z_1, Z_2)$  is always less efficient than the one using  $w(Z_2)$  and  $E_f\{h(Z_1, Z_2) \mid Z_2\}$ . In the special case when  $h$  is a function of  $Z_2$  alone, the conditional expectation is reduced to  $h(Z_2)$ . In a complicated problem when the marginal weight  $w(Z_2)$  is difficult to come by, a partial Rao-Blackwellization scheme can be used, as implemented in Section 4. That is, the joint weight  $w(Z_1, Z_2)$  is used together with the conditional expectation  $E_f\{h(Z_1, Z_2) \mid Z_2\}$ . Although simulations show great improvement by using partial Rao-Blackwellization, the mathematical properties of such an approach are less clear.

## 7 Discussion

Modern Monte Carlo methods have opened the door to much more complex and realistic models. While the early hope was that the methods would provide a panacea, letting simple programs do the work, the current consensus is that care needs to be taken with these methods, and that it can often be worthwhile to develop general strategies for improving their performance. The practitioner can then examine a specific problem, select the improvements that are feasible for this problem, and with reasonable effort, produce a satisfactory Monte Carlo technique.

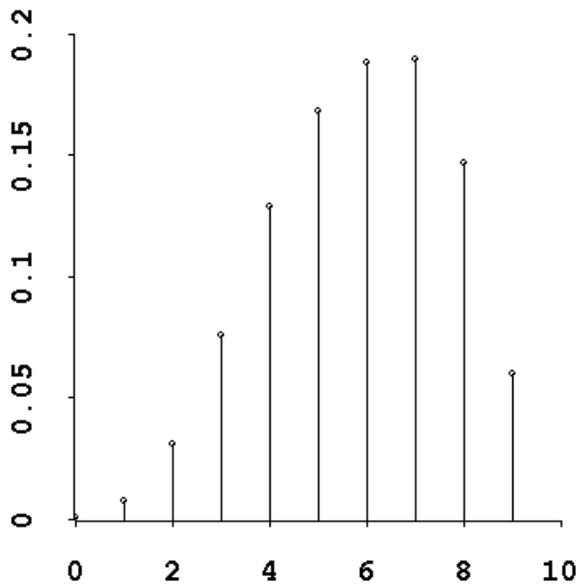
The close parallels between the conditional distributions needed to run the Gibbs sampler and the SIS suggest that the same techniques that improve one are likely to improve the other as well. With the large literature on Gibbs sampling, much is known about improving the samplers, either through collapsing the state space of the Markov chain on which the sampler runs, or by designing special moves to enable the chain mix more quickly. We have shown how the two principal fixes of this sort for models involving the Dirichlet process can be applied to the SIS and result in dramatic improvement. This creates a hierarchy of algorithms for the SIS parallel to that for the Gibbs sampler.

A substantial statistical issue that remains to be tackled is the great discrepancies between pictures of the distribution of  $\theta_{n+1}|x$  as the value of  $M$  changes. Extensive Monte Carlo results suggest that there is a large, real difference, beyond the Monte Carlo variation. With given  $F$  the distribution of  $X$  depends only on the first nine moments of  $F$ . This leads to a two-stage view of the posterior of  $F$ . First, there is a distribution on these nine moments. Second, there is a distribution on  $F$  given these nine moments. This suggests that we obtain consistency for  $X_{n+1}|(x_1, \dots, x_n)$  as  $n \rightarrow \infty$ . The story differs for  $\theta_{n+1}|(x_1, \dots, x_n)$ , however. We believe that the magnitude of the difference in the distributions is due to two features: that the map from  $F$  to its first nine moments is many to one, and that the various values of  $M$  assign very different distributions to  $F$ , conditional on its first nine moments. This view makes it clear that if  $F$  is allowed to be an arbitrary distribution, any estimator of  $F$  will suffer from inconsistency.

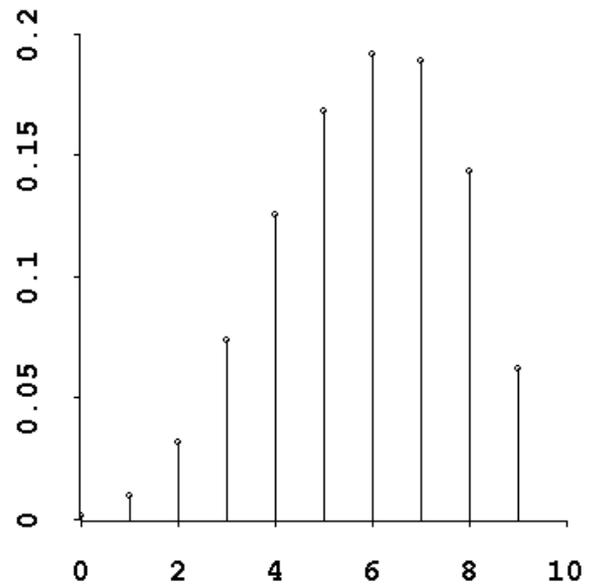
## References

- Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152-1174.
- Beckett, L. and Diaconis, P. (1994). Spectral analysis for discrete longitudinal data. *Adv. in Math.* **103** 107-128.
- Berry, D.A. and Christensen, R. (1979). Empirical Bayes estimation of a binomial parameter via mixture of Dirichlet processes. *Ann. Statist.* **7** 558-568.
- Berzuini, C., Best, N.G., Gilks, W.R., and Larizza, C. (1996). Dynamic graphical models and Markov chain Monte Carlo methods. *J. Amer. Statist. Assoc.* **91**, forthcoming.
- Blackwell, D. and MacQueen, J.B. (1973). Ferguson distributions via Polya urn schemes. *Ann. Statist.* **1** 353-355.
- Bush, C.A. and MacEachern, S.N. (1996). A semi-parametric Bayesian model for randomized block designs. *Biometrika* **83** 275-285.
- Diaconis, P. (1995). Personal communication.
- Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Ann. Statist.* **22** 1763-1786.
- 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.
- 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.

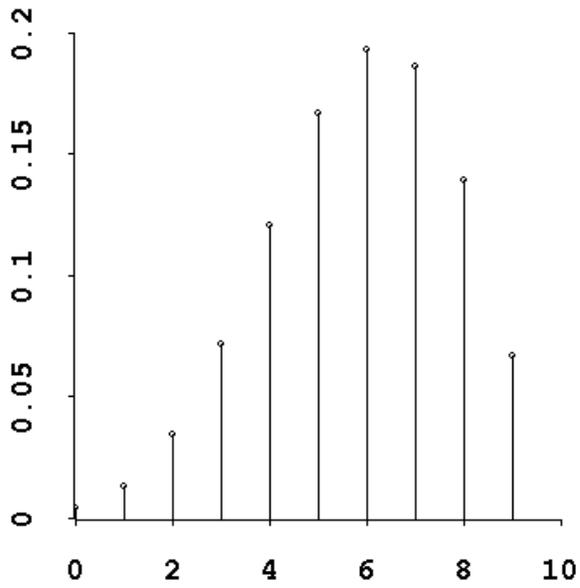
- Gopalan, R. (1994). Unpublished Ph.D. dissertation, Institute of Statistics and Decision Sciences, Duke University.
- Kong, A., Liu, J.S. and Wong, W.H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.* **89** 278-288
- Kuo, L. (1986). Computations of mixtures 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. (1996). Nonparametric hierarchical Bayes via sequential imputations. *Ann. Statist.*, **24**, 910-930.
- Liu, J.S. and Chen, R. (1996). A note on Monte Carlo methods for dynamic systems. *Technical Report*, Department of Statistics, Stanford University.
- 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. and Müller, P. (1994). Estimating mixtures of Dirichlet process models. *ISDS Discussion Paper*, Duke University.
- 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.



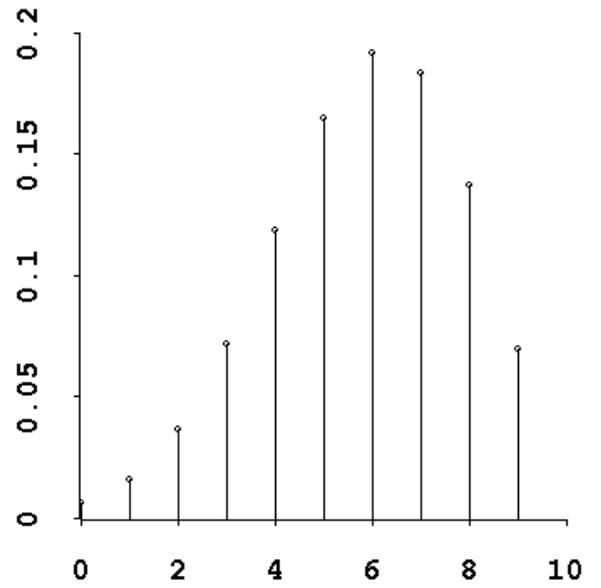
$M = .1$



$M = 1$

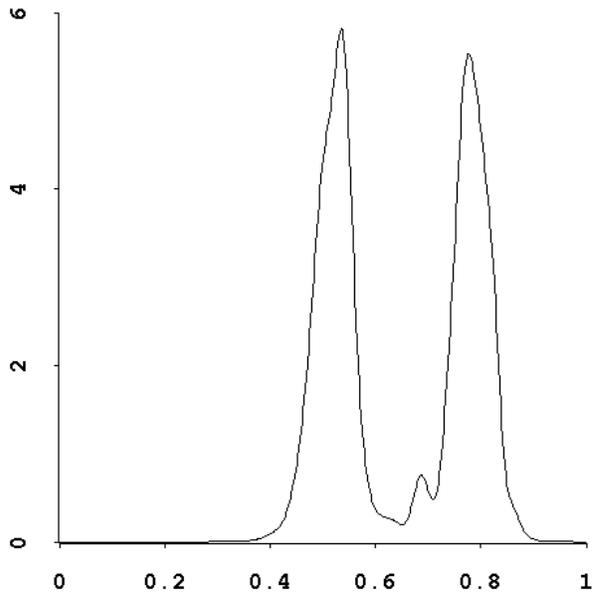


$M = 5$

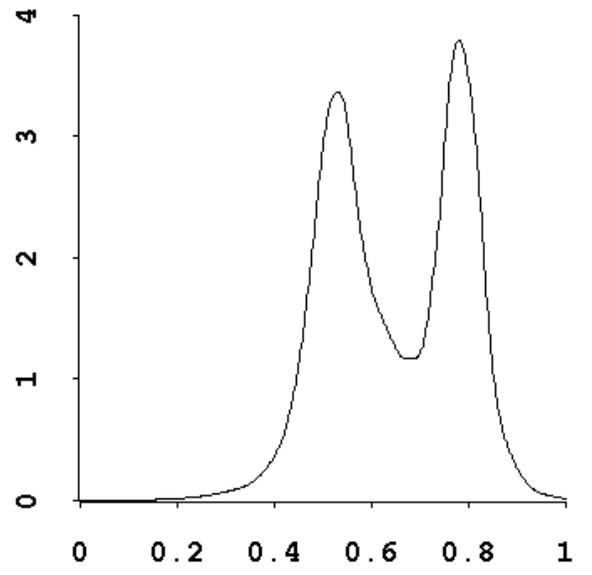


$M = 10$

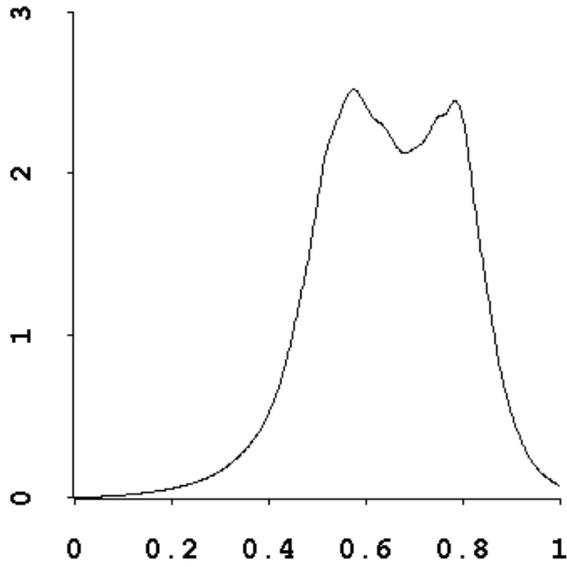
Figure 1: Estimates of the densities for  $X_{n+1}$  given  $x$  for the four values of  $M$ . Each density is based on 10,000 iterations of the sequential importance sampler S2.



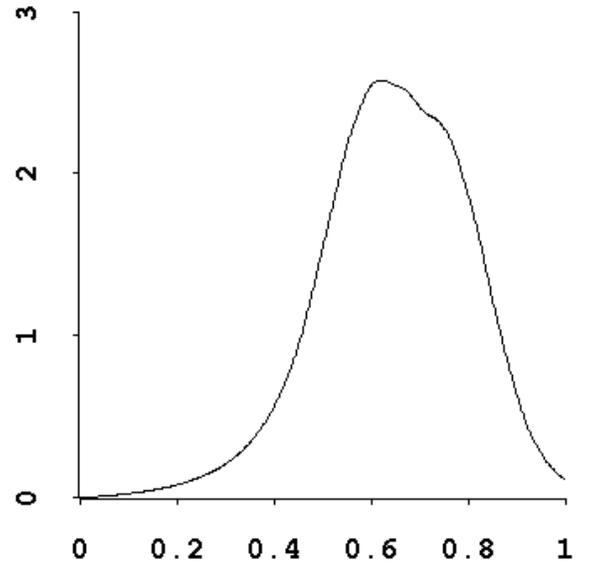
$M = .1$



$M = 1$



$M = 5$



$M = 10$

Figure 2: Estimates of the densities for  $\theta_{n+1}$  given  $x$  for the four values of  $M$ . Each density is based on 10000 iterations of the sequential importance sampler S2.

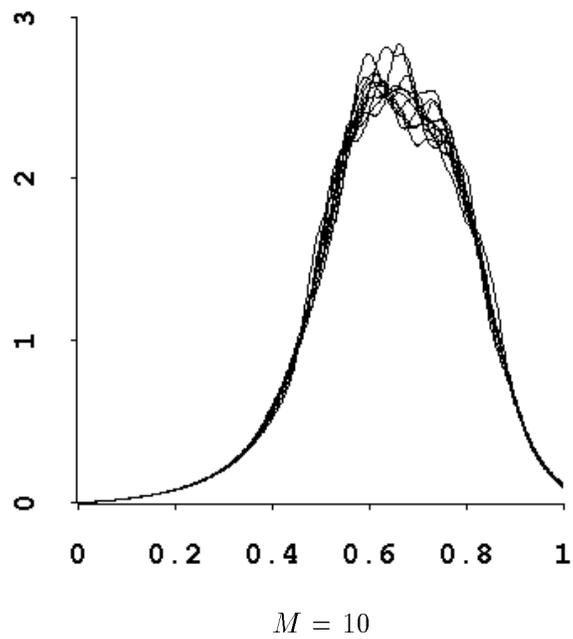
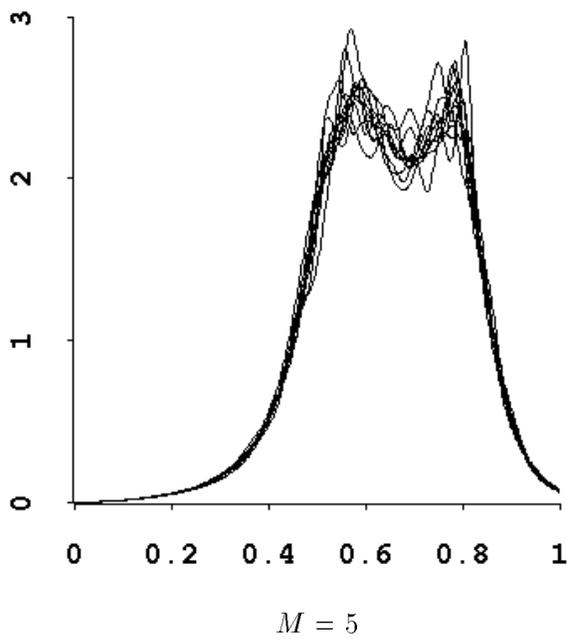
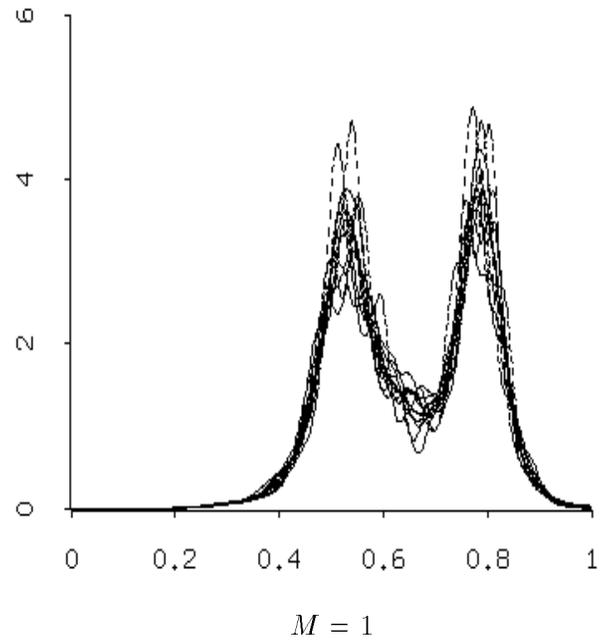
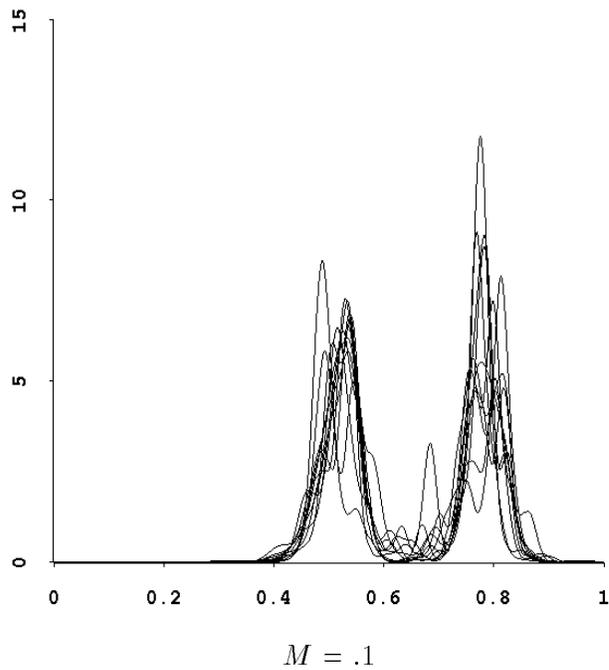


Figure 3: Simulation variation in the estimates of the densities for  $\theta_{n+1}$  given  $x$  for the four values of  $M$ . Ten batches based on 1000 iterations of the sequential importance sampler S2 were used to estimate the densities.