Statistical Methodology 28 (2016) 1–17

Contents lists available at ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

# Bayesian predictive inference under a Dirichlet process with sensitivity to the normal baseline

Balgobin Nandram *, Jiani Yin

*Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, United States*

ARTICLE INFO

*Article history:*
Received 6 June 2014
Received in revised form
12 February 2015
Accepted 12 July 2015
Available online 26 July 2015

*MSC:*
62D05
62G35

*Keywords:*
Dirichlet process mixture
Grid method
Leave-one-out kernel density
Polya posterior
Sensitivity analysis
Skewed distribution
Simulation

ABSTRACT

It is well known that the Dirichlet process (DP) model and Dirichlet process mixture (DPM) model are sensitive to the specifications of the baseline distribution. Given a sample from a finite population, we perform Bayesian predictive inference about a finite population quantity (e.g., mean) using a DP model. Generally, in many applications a normal distribution is used for the baseline distribution. Therefore, our main objective is empirical and we show the extent of the sensitivity of inference about the finite population mean with respect to six distributions (normal, lognormal, gamma, inverse Gaussian, a two-component normal mixture and a skewed normal). We have compared the DP model using these baselines with the Polya posterior (fully nonparametric) and the Bayesian bootstrap (sampling with a Haldane prior). We used two examples, one on income data and the other on body mass index data, to compare the performance of these three procedures. These examples show some differences among the six baseline distributions, the Polya posterior and the Bayesian bootstrap, indicating that the normal baseline model cannot be used automatically. Therefore, we consider a simulation study to assess this issue further, and we show how to solve this problem using a leave-one-out kernel baseline. Because the leave-one-out kernel baseline cannot be easily applied to the DPM, we show theoretically how one can solve the sensitivity problem for the DPM as well.



* Corresponding author.
*E-mail addresses:* balnan@wpi.edu (B. Nandram), jianiyin@wpi.edu (J. Yin).

http://dx.doi.org/10.1016/j.stamet.2015.07.003
1572-3127/

## 1. Introduction

We consider the problem of robustness to the misspecification of the baseline model of a Dirichlet process model [11]. We restrict attention to finite population sampling and we assume that a simple random sample is drawn from a finite population and the population values follow a Dirichlet process (DP) model. The sampled values are observed and the nonsampled values are to be predicted using the DP model. Because the data could be skewed, heavy tailed, multimodal or a combination of these features, we use several baseline models. We also compare our method with the Bayesian bootstrap [27], a special case of multinomial sampling first introduced by Ericson [9] for finite population sampling, and the Polya posterior [13], a special case of the DP model. We show how to use a leave-one-out kernel density for the baseline model. In addition, we show how to obtain a finite mixture of distributions as a baseline for the Dirichlet process mixture model which is used routinely in applications.

Nonparametric Bayesian statistics is currently a very active area (e.g., see [16]) and it can be used to analyze data from complex surveys (e.g., [4,24,7]). However, in these works no attention was paid to the sensitivity of inference to the baseline model which has to be chosen very carefully because posterior inference is sensitive to such a choice (e.g., [8,22,10,18]).

Within the Bayesian nonparametric paradigm, there are two standard choices which are the Dirichlet process (DP) model and the Dirichlet process mixture (DPM) model. The DP model is

$$y_1, \ldots, y_n \mid G \overset{\text{i.i.d.}}{\sim} G \quad \text{and} \quad G \mid \alpha, H \sim DP(\alpha, H),$$

where $\alpha$ is the concentration parameter and $H$, the baseline distribution which is generally assumed to be absolutely continuous. The DPM model is

$$y_i \mid \mu_i \overset{ind}{\sim} f(y_i \mid \mu_i, \tau), \quad i = 1, \ldots, n,$$

$$\mu_1, \ldots, \mu_n \mid G \overset{\text{i.i.d.}}{\sim} G \quad \text{and} \quad G \mid H \sim DP(\alpha, H).$$

It is worth noting that in the DPM model the parametric distributions, $f(y_i \mid \mu_i, \tau)$, have to be specified. Besides, in practice, inference is likely to be sensitive to the specification of $f(y \mid \mu_i, \tau)$ and model diagnostics will be needed. Nevertheless, the whole idea is that the discreteness [11] of $G$ in the DP is removed by using the DPM model [12,20]. The issue is then how different will inference be with baselines other than the normal.

For an unknown $H$ there are two approaches, parametric or nonparametric. If one specifies a parametric distribution, one is essentially using an empirical Bayes procedure and it creates the sensitivity issue. Nevertheless, most applied statisticians inadvertently or carelessly use this approach. Referring to the choice of base measure and other hyperparameters, Hannah, Blei and Powell [14] wrote, "The base measure is chosen in line with data size, distribution type, distribution features (such as heterogeneity and others) and computational constraints". In the nonparametric approach, a kernel density is used (e.g., [22]). The problem here is that the data have to be used to construct the kernel density estimator. Although this empirical Bayes procedure is incoherent from a Bayesian perspective, it is apparent that this is the only sensible choice. For the DP model $E(G(y)) = H(y)$, $-\infty < y < \infty$. Thus, the empirical cdf $\widehat{H(y)} = \sum_{i=1}^{n} \Delta_{y_i}(y)$ where $\Delta_{y_i}(y)$ is the cdf of a point mass at $y_i$ and $\widehat{h(y)} = \sum_{i=1}^{n} \delta_{y_i}(y)$ is a "density" estimator with $\delta_{y_i}(y)$ being the weak derivative of $\Delta_{y_i}(y)$. The problem is that $\widehat{H(y)}$ is not smooth and therefore a kernel density estimator has to be used. In either case the data are used twice. Another solution is to put a Dirichlet process on $H$ [30] but, as correctly pointed out by McAuliffe, Blei and Jordan [22] the same problem arises in specifying the hyperparameters.

For the DPM model McAuliffe, Blei and Jordan [22] used an iterative procedure, iterating between an estimation and inference phase with both $H$ and $\alpha$ being updated simultaneously. Starting with a sample of size $B$ of $\theta_1, \ldots, \theta_\ell$, with distinct values denoted by $\theta^*_{1b}, \ldots, \theta^*_{k_b b}$, $b = 1, \ldots, B$, a kernel density estimator of $\theta$ is constructed as

$$\widehat{H(\theta)} = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{k_b} \sum_{i=1}^{k_b} \frac{1}{h_b} \phi \left\{ \frac{\theta - \theta^*_{ib}}{h_b} \right\},$$

where $\phi(\cdot)$ is the standard normal density, $h_b$ is the optimal window width and $k_b$ is the number of distinct values among the $\ell$ values of $\theta$. Then, fit the model with $\widehat{H(\theta)}$ as the baseline, update $\widehat{H(\theta)}$

and iterate this process. There are at least five additional problems besides being empirical Bayes. First, it is inconvenient to monitor the convergence of the Gibbs sampler each time it is run. Second, in a situation when one run of the Gibbs sampler takes a long time, repeating it would take a lot more time. Third, it is difficult to assess the overall convergence property of this procedure. Fourth, a relatively large number of iterations is needed for convergence; roughly 200 iterations as reported by McAuliffe, Blei and Jordan [22]. Fifth, if $k_b$ is small, the kernel density cannot be efficiently constructed and this is a general difficulty with the Dirichlet process. Moreover, this procedure cannot be applied to the DP model which is distinct from the DPM model. In this paper, although we focus on the DP model, not the DPM model, in Section 4 we show a procedure, simpler than the one of McAuliffe, Blei and Jordan [22], for the DPM model, and our purpose is to give an applied statistician a method to proceed, not a detailed discussion.

For survey sampling Ericson [9] assumed a multinomial-Dirichlet model for the $k$, $1 \leq k \leq n$, distinct values $y_1^*, \ldots, y_k^*$ and $y_i^*$, with the $i$th value occurring $n_i \geq 1$ times, in the observed data and $\sum_{i=1}^{k} n_i = n$. The model assumes that the only values that can occur in the population are the $y_i^*$, an obvious weakness, which the DP model [4] can potentially get around. The Dirichlet prior, with all parameters set to 0, is the Haldane prior which models the proportions of the $y_i^*$ values in the population. This was an original idea of Ericson [9] although he did not use the Haldane prior which is improper. Instead he used a small positive value for the parameters of the Dirichlet distribution to accommodate a slightly higher degree of smoothness. But with $n_i \geq 1,\ i = 1, \ldots, k$, the posterior density of the proportions of values in the population is proper. Posterior inference is available for the number of nonsampled values, $N_i - n_i,\ i = 1, \ldots, k$, in the population, where $N_i$ is the number (assumed known) of $y_i^*$ values in the population. The Bayesian bootstrap [27] draws samples from the posterior distribution under a Haldane prior. Basically, $\underset{\sim}{n} \mid k, \underset{\sim}{\pi} \sim \text{Multinomial}(n, \underset{\sim}{\pi}), \underset{\sim}{\pi} \sim \text{Dirichlet}(0, \ldots, 0)$. Then, $\underset{\sim}{\pi} \sim \text{Dirichlet}(n_1, \ldots, n_k)$ from which $\underset{\sim}{\pi}$ is drawn and then bootstrapping is done by drawing samples from Multinomial$(1, \underset{\sim}{\pi})$. To bootstrap a finite population of size $N$, the nonsampled values are drawn in the same manner.

Another nonparametric procedure is the Polya posterior [13] and it is, in fact, a special case of the DP model with $\alpha$ approaching zero; see [21]. An urn contains $n$ balls with $k$ different colors $n_1, \ldots, n_k,\ k \geq 2$. Draw a ball from this urn, note its color, and return this ball and a ball of the same color into the urn. Continue this procedure until there are now $N$ balls, finite population size, in this urn. We now have a copy of the entire finite population. We can now repeat the entire procedure many times to have a large sample (a sample from the Polya posterior); see [13] for detailed mathematical properties of the Polya posterior. Of course, the Bayesian bootstrap and the Polya posterior are very similar.

In this paper we use the DP model to make inference about a finite population mean when a simple random sample is obtained from a finite population. This requires generating the nonsampled finite population. Specifically, we have compared six baseline distributions (normal, lognormal, gamma, inverse Gaussian, a two-component normal and a skewed normal). We have made comparisons with the Bayesian bootstrap and the Polya posterior. This is how we study sensitivity to the normal baseline.

The plan of the rest of the paper is as follows. In Section 2 we describe the six baselines. Our objective is to show the extent of the sensitivity of inference about the finite population mean to the specification of the baseline. In agreement with McAuliffe, Blei and Jordan [22] we also believe that the best choice is a kernel density which, unfortunately, "double" uses the data. In Section 3 we consider two numerical examples. We also describe a simulation study to assess frequentist differences in the three procedures (the six baselines, Polya posterior and Bayesian bootstrap), and we suggest a solution to this problem using a leave-one-out kernel density estimator as a baseline. In Section 4 we provide a theoretical extension to the Dirichlet process mixture (DPM) model. In Section 5 we have concluding remarks. In the Appendix A we have a theoretical discussion on the consistency of the Bayes estimator of the finite population mean. Detailed discussions are given in the supplement (see Appendix B).

## 2. Bayesian methodology

We have a simple random sample of size $n$ from a population of size $N$. We assume that the sampled values are $y_1, \ldots, y_n$ and nonsampled values are $y_{n+1}, \ldots, y_N$. Inference is required for the finite

population mean, $\bar{Y} = \sum_{i=1}^{N} y_i/N$, and data $y_1, \ldots, y_n$ are available. Note that $\bar{Y} = \sum_{i=1}^{N} y_i/N = f\bar{y}_s + (1-f)\bar{y}_{ns}$, where $f = n/N$ is the sampling fraction, $\bar{y} = \sum_{i=1}^{n} y_i/n$, the sample mean, and $\bar{y}_{ns} = \sum_{i=n+1}^{N} y_i/(N-n)$, the nonsample mean which is to be predicted. Thus, we need random samples from the posterior density of $\bar{y}_{ns}$ given $\underset{\sim}{y}_s$.

### 2.1. General discussions

Henceforth, for the DP model we assume that

$$y_1, \ldots, y_N \mid G \overset{\text{i.i.d.}}{\sim} G \quad \text{and} \quad G \mid \{\alpha, H_{\underset{\sim}{\psi}}(y), \underset{\sim}{\psi}\} \sim DP\{\alpha, H_{\underset{\sim}{\psi}}(y)\}, \tag{1}$$

where $H_{\underset{\sim}{\psi}}(y)$ is the smooth baseline cdf and the pdf is $h_{\underset{\sim}{\psi}}(y)$. The parameters $\alpha$ and $\underset{\sim}{\psi}$ are unknown and a priori we assume that they are independent with prior distributions, $\pi(\underset{\sim}{\psi})$ and $\pi(\alpha)$. We want to study empirically the extent of the sensitivity of inference about the finite population mean for specifications of $h_{\underset{\sim}{\psi}}(y)$ which differ from the normal baseline. Then, integrating out $G$ [5], the joint posterior density under the full DP model is

$$\pi(\alpha, \underset{\sim}{\psi} \mid \underset{\sim}{y}) \propto h_{\underset{\sim}{\psi}}(y_1) \left[ \prod_{i=2}^{n} \left\{ \frac{i-1}{\alpha+i-1} \left\{ \frac{\sum_{j=1}^{i-1} \delta_{y_j}(y_i)}{i-1} \right\} + \frac{\alpha}{\alpha+i-1} h_{\underset{\sim}{\psi}}(y_i) \right\} \right] \pi(\underset{\sim}{\psi})\pi(\alpha). \tag{2}$$

Let $k$, $1 \le k \le n$, denote the number of distinct values among $y_1, \ldots, y_n$. That is, letting $y_1^*, \ldots, y_k^*$ denote the $k$ distinct sample values, the baseline model is

$$y_1^*, \ldots, y_k^* \mid k, \underset{\sim}{\psi} \overset{\text{i.i.d.}}{\sim} h_{\underset{\sim}{\psi}}(y)$$

with the prior in $\pi(\underset{\sim}{\psi})$ (same as in the full DP model). Note that $k$ is also a random variable and Antoniak [3] showed that $p(k \mid \alpha) = s_n(k)\alpha^k\Gamma(\alpha)/\Gamma(\alpha+n)$, $k = 1, \ldots, n, \alpha > 0$, where the $s_n(k)$ are the absolute values of the Stirling numbers of the first kind [1]. The joint posterior density under the baseline model is

$$\pi(\alpha, \underset{\sim}{\psi} \mid k, \underset{\sim}{y^*}) = \pi(\alpha \mid k)\pi(\underset{\sim}{\psi} \mid \underset{\sim}{y^*}), \tag{3}$$

where $\pi(\alpha \mid k) \propto \alpha^k\{\Gamma(\alpha)/\Gamma(\alpha+n)\}\pi(\alpha)$, $\alpha > 0$ and $\pi(\underset{\sim}{\psi} \mid \underset{\sim}{y^*}) \propto \{\prod_{i=1}^{k} h_{\underset{\sim}{\psi}}(y_i^*)\}\pi(\underset{\sim}{\psi})$. Note that a posteriori while $\alpha$ and $\underset{\sim}{\psi}$ are independent in (3), they are not independent in (2).

We will use a 'Cauchy' type prior for $\alpha$, sometimes called a shrinkage prior, of the form $\alpha$, $p(\alpha) = 1/(\alpha+1)^2$, $\alpha > 0$. It is slightly more convenient to use $p(\alpha) = 1/(\alpha+1)^2$, $\alpha > 0$ rather than the half Cauchy density $p(\alpha) = 2/\pi(\alpha^2+1)$, $\alpha > 0$ [26]. We will specify appropriate noninformative prior for $\underset{\sim}{\psi}$, denoted by $\pi(\underset{\sim}{\psi})$. Under the assumption of independence of $\alpha$ and $\underset{\sim}{\psi}$, we have

$$\pi(\alpha, \underset{\sim}{\psi}) \propto \frac{1}{(\alpha+1)^2}\pi(\underset{\sim}{\psi}), \quad \alpha > 0 \tag{4}$$

with appropriate support for $\psi$ depending on the baseline. The prior specification in (4) is the same in both the baseline model and the full DP model.

It is easy to show that under the baseline model or the full DP with prior in (4) the posterior density of $\alpha$ is proper. Transforming $\alpha$ according to $\rho = 1/(\alpha+1)$ (correlation in Dirichlet process) a priori and simplifying we get

$$\pi(\rho \mid k) \propto \frac{(1-\rho)^{k-1}\rho^{n-k}}{\prod_{j=1}^{n-1}\{1-\rho+\rho j\}}, \quad 0 \le \rho \le 1. \tag{5}$$

This function is bounded for all $\rho$. Thus, under the baseline model, the joint posterior is proper if and only if $\pi(\underset{\sim}{\psi} \mid \underset{\sim}{y^*})$ is proper. The following theorem may be known, but it is difficult to retrieve.

**Theorem.** *If the posterior density under the baseline model is proper, the posterior density under the full Dirichlet process is proper.*

**Proof.** Without loss of generality, we assume that the $k$ distinct values come first. Then using the form of the joint posterior density in (2) and noting that $(i-1)/(\alpha+i-1) \leq 1,\ i=1,\ldots,n$, and $\sum_{j=1}^{i-1}\delta_{y_j}(y_i)/(i-1) \leq 1,\ i=2,\ldots,n$, we have

$$\frac{\pi(\underset{\sim}{\psi})}{(\alpha+1)^2}\left[\prod_{i=1}^{k} h_{\underset{\sim}{\psi}}(y_i)\right]\left[\prod_{i=k+1}^{n}\left\{\frac{1}{\alpha+i-1}\sum_{j=1}^{i-1}\delta_{y_j}(y_i)+\frac{\alpha}{\alpha+i-1}h_{\underset{\sim}{\psi}}(y_i)\right\}\right]$$
$$\leq \frac{\pi(\underset{\sim}{\psi})}{(\alpha+1)^2}\left[\prod_{i=1}^{n} h_{\underset{\sim}{\psi}}(y_i)\right].$$

It is convenient to use $\prod_{i=1}^{n} h_{\underset{\sim}{\psi}}(y_i)$ in the inequality. Therefore, we only need to show that

$$\int\int_0^\infty \frac{\pi(\underset{\sim}{\psi})}{(\alpha+1)^2}\prod_{i=1}^{n} h_{\underset{\sim}{\psi}}(y_i)\,d\alpha\,d\underset{\sim}{\psi} < \infty. \tag{6}$$

Integrating out $\alpha$ (any proper prior will do), we now only need to show that

$$\int \pi(\underset{\sim}{\psi})\prod_{i=1}^{n} h_{\underset{\sim}{\psi}}(y_i)\,d\underset{\sim}{\psi} < \infty. \tag{7}$$

This is simply the condition needed for propriety of the posterior density under the baseline model.

Typically, it is straight forward to draw $\underset{\sim}{\psi}$. However, it is not really trivial to draw $\alpha$ without using a special kind of prior; see [24] for a discussion of the gamma prior which was discussed earlier in [10] and it is used by numerous statisticians as well. It is easy to draw $\alpha$ via $\rho$ in (5) using a grid method (i.e., convert the posterior density into a probability mass function by discretizing $[0, 1]$ into a large number of tiny intervals) with jittering (e.g., [23,31]). This procedure works exactly the same way for the DPM and no monitoring is needed.

### 2.2. Some baseline distributions

We specify various density functions for $\underset{\sim}{\psi}$. We also show how to draw samples from the posterior density of $\underset{\sim}{\psi}$. Specifically, we consider the normal, lognormal, gamma, inverse Gaussian, two-component mixture and skewed normal distributions. We will also state conditions for the posterior density to be proper under the baseline model. To avoid the asterisk notation, we will let $y_1,\ldots,y_k$ denote the distinct values.

Here, for convenience we present the details of the normal baseline distribution. Here $\underset{\sim}{\psi} = (\mu,\sigma^2)$ and the model in (1) is

$$y_1,\ldots,y_k \mid k,\mu,\sigma^2 \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu,\sigma^2), \qquad p(\mu,\sigma^2)\propto 1/\sigma^2, \quad -\infty<\mu<\infty,\ \sigma^2>0.$$

Then, letting $\bar{y}_k = \sum_{i=1}^{k} y_i/k$ and $s_k^2 = \sum_{i=1}^{k}(y_i-\bar{y}_k)^2/(k-1)$, we have the standard result that $\mu \mid \sigma^2, k, \bar{y}_k, s_k^2 \sim \text{Normal}(\bar{y}_k, \sigma^2/k)$ and $\sigma^{-2} \mid s_k^2, k \sim \text{Gamma}\{(k-1)/2, (k-1)s_k^2/2\}$. That is, $\sqrt{k}(\mu-\bar{y}_k)/s_k \mid \bar{y}_k, s_k^2, k \sim t_{k-1}, k>1$. Thus, the posterior distribution of $(\mu,\sigma^2)$ is proper and it is trivial to draw $\mu$ and $\sigma^2$.

Similar results for the lognormal, gamma, inverse Gaussian, two-component mixture and skewed normal are given in Table 1 (suggested by a referee). Details of model fitting and propriety are given in the supplement (see Appendix B).

**Table 1**
Summaries of different baseline distributions of the Dirichlet process model.

| | |
|---|---|
| Normal | |
| Model | $y_1, \ldots, y_k \mid k, \mu, \sigma^2 \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu, \sigma^2)$; $p(\mu, \sigma^2) \propto 1/\sigma^2,\ -\infty < \mu < \infty,\ \sigma^2 > 0$. |
| Posterior | $\mu \mid \sigma^2, k, \bar{y}_k, s_k^2 \sim \text{Normal}(\bar{y}_k, \sigma^2/k)$; $\sigma^{-2} \mid s_k^2, k \sim \text{Gamma}\{(k-1)/2, (k-1)s_k^2/2\}$. |
| Remarks | $\sqrt{k}(\mu - \bar{y}_k)/s_k \mid \bar{y}_k, s_k^2, k \sim t_{k-1}$, $k > 1$ for propriety. |
| Lognormal | |
| Model | $z_1, \ldots, z_k \mid k, \mu, \sigma^2 \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu, \sigma^2)$; $p(\mu, \sigma^2) \propto 1/\sigma^2,\ -\infty < \mu < \infty,\ \sigma^2 > 0$. (Define $z_i = \ln(y_i),\ y_i > 0,\ i = 1, \ldots, k$.) |
| Posterior | $\mu \mid \sigma^2, k, \bar{z}_k, s_k^2 \sim \text{Normal}(\bar{z}_k, \sigma^2/k)$; $\sigma^{-2} \mid s_k^2, k \sim \text{Gamma}\{(k-1)/2, (k-1)s_k^2/2\}$. |
| Remarks | The moments of the nonsampled $y_i$ may not exist. |
| Gamma | |
| Model | $y_1, \ldots, y_k \mid k, \mu, \eta \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\eta, \mu^{-1}\eta)$; $p(\mu, \eta) \propto \frac{1}{\mu(1+\eta)^2},\ \mu > 0,\ \eta > 0$. |
| Posterior | $\mu \mid \eta, \underset{\sim}{y}_k \sim \text{Inverse-Gamma}(k\eta, k\eta a)$; $\pi(\eta \mid \underset{\sim}{y}_k) \propto \frac{1}{\mu(1+\eta)^2} \left(\frac{\eta^\eta}{\mu\Gamma(\eta)}\right)^k g^{k(\eta-1)} \left(\frac{1}{k\eta a}\right)^{k\eta}$. |
| Remarks | By transforming $\eta$ to $\tau = 1/(1+\eta)$, $\pi(\tau \mid \underset{\sim}{y}_k)$ is proper if $0 < \tau < 1$. |
| Inverse Gaussian | |
| Model | $y_1, \ldots, y_k \mid \mu, \lambda \overset{\text{i.i.d.}}{\sim} \text{IGauss}(\mu, \lambda)$, where $f(y \mid \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\},\ y > 0$; $p(\mu, \eta) \propto \frac{1}{\mu(1+\eta)^2},\ \mu > 0,\ \eta > 0$. |
| Posterior | $\mu \mid \eta, \underset{\sim}{y}_k \sim \text{Inverse-Gamma}(k\eta, k\eta a)$; $\pi(\eta \mid \underset{\sim}{y}_k) \propto \frac{1}{\mu(1+\eta)^2} \left(\frac{\eta^\eta}{\mu\Gamma(\eta)}\right)^k g^{k(\eta-1)} \left(\frac{1}{k\eta a}\right)^{k\eta}$. |
| Remarks | Computation is similar to the gamma baseline. |
| Two-component mixture | |
| Model | $y_i \mid z_i = r \overset{ind}{\sim} \text{Normal}(\mu_r, \sigma^2)$; $z_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi),\ i = 1, \ldots, k$, where $1 \le \sum_{i=1}^k z_i \le k-1$; $\pi \sim \text{Uniform}(0, 1)$; $\pi(\mu_0, \mu_1) \propto 1,\ -\infty < \mu_0 < \mu_1 < \infty$; independently $\pi(\sigma^2) \propto 1/\sigma^2,\ \sigma^2 > 0$. |
| Posterior | $\pi(\underset{\sim}{z}, \pi, \mu_0, \mu_1, \sigma^2 \mid \underset{\sim}{y}_k) \propto \frac{1}{\sigma^2} \pi^{\sum_{i=1}^k z_i} (1-\pi)^{\sum_{i=1}^k (1-z_i)} \prod_{i=1}^k \left(\frac{1}{\sigma}\phi\{\frac{y_i-\mu_0}{\sigma}\}\right)^{1-z_i} \times \left(\frac{1}{\sigma}\phi\{\frac{y_i-\mu_1}{\sigma}\}\right)^{z_i}$, where $\phi(\cdot)$ is the standard normal density function. |
| Remarks | $\pi(\underset{\sim}{z}, \pi, \mu_0, \mu_1, \sigma^2 \mid \underset{\sim}{y}_k)$ is proper if $k \ge 3$. Use the Gibbs sampler to fit the model. |
| Skewed normal | |
| Model | $y_i \mid \mu, \sigma^2, \gamma \overset{\text{i.i.d.}}{\sim} \text{SN}(\mu, \sigma^2, \gamma),\ i = 1, \ldots, k,\ -\infty < y_i < \infty$, where $f(y \mid \mu, \sigma^2, \gamma) = \frac{2}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\Phi\{\frac{\gamma}{\sqrt{1-\gamma^2}}\left(\frac{y-\mu}{\sigma}\right)\}$, $\phi(\cdot)$ is pdf of $N(0, 1)$, $\Phi(\cdot)$ is the cdf of $N(0, 1)$; $\pi(\mu, \sigma^2, \gamma) \propto 1/\sigma^2,\ -\infty < \mu < \infty,\ \sigma^2 > 0,\ |\gamma| < 1$. |
| Posterior | $\pi(\gamma \mid \mu, \sigma^2, \underset{\sim}{y}_k) \propto \prod_{i=1}^k \Phi\{\frac{\gamma}{\sqrt{1-\gamma^2}}(\frac{y_i-\mu}{\sigma})\}$; $\pi(\mu, \sigma^2 \mid \underset{\sim}{y}_k) \propto A(\mu, \sigma)\frac{1}{\sigma^2}\prod_{i=1}^k \frac{2}{\sigma}\phi(\frac{y_i-\mu}{\sigma})$, where $A(\mu, \sigma) = \int_{-1}^1 \prod_{i=1}^k \Phi\{\frac{\gamma}{\sqrt{1-\gamma^2}}(\frac{y_i-\mu}{\sigma})\}d\gamma$. |
| Remarks | $\pi(\mu, \sigma^2, \gamma \mid \underset{\sim}{y}_k)$ is proper if $k > 1$. |

### *2.3. Posterior inference about the finite population mean*

Assuming that we already have 10,000 iterates, say, from the joint posterior density of $\alpha, \underset{\sim}{\psi}$ given $\underset{\sim}{y}_s$, samples of the finite population mean can now be drawn. We just need to be able to obtain samples from the joint posterior density of $\underset{\sim}{y}_{ns}, \alpha, \underset{\sim}{\psi}$ given $\underset{\sim}{y}_s$ (i.e., Polya urn scheme, $G$ is integrated out). We have

$$p(\underset{\sim}{y}_{ns}, \alpha, \underset{\sim}{\psi} \mid \underset{\sim}{y}_s) = p(\underset{\sim}{y}_{ns} \mid \alpha, \underset{\sim}{\psi}, \underset{\sim}{y}_s)\pi(\alpha, \underset{\sim}{\psi} \mid \underset{\sim}{y}_s).$$

Once samples are taken from $\pi(\alpha, \underset{\sim}{\psi} \mid \underset{\sim}{y}_s)$, using the composition rule, samples are obtained from $p(\underset{\sim}{y}_{ns} \mid \alpha, \underset{\sim}{\psi}, \underset{\sim}{y}_s)$. Therefore, samples can be taken from $p(\underset{\sim}{y}_{ns}, \alpha, \underset{\sim}{\psi} \mid \underset{\sim}{y}_s)$.

It is easy to draw $\bar{y}_{ns}$. To each of the 10,000 iterates, simply fill in the values $y_{n+1}, \ldots, y_N$ (data augmentation). Letting $F_{y_i}(y)$ denote the cdf of a point mass at $y_i$, it is easy to show, using the

generalized Polya urn scheme (e.g., [5]), that

$$y_{n+1} \mid \{\alpha, \mu, \sigma^2, \underset{\sim}{y}_s\} \sim \frac{n}{\alpha+n}\bar{F}_n(y) + \frac{\alpha}{\alpha+n}H(\underset{\sim}{\psi}),$$

$$y_{n+k+1} \mid \{\alpha, \mu, \sigma^2, \underset{\sim}{y}_s, y_{n+1}, \ldots, y_{n+k}\} \sim \frac{n+k}{\alpha+n+k}\bar{F}_{n+k}(y) + \frac{\alpha}{\alpha+n+k}H(\underset{\sim}{\psi}), \tag{8}$$

$k = 1, \ldots, N-n-1$, where $\bar{F}_{n+k}(y) = \sum_{i=1}^{n+k} F_{y_i}(y)/(n+k)$ and $\bar{F}_n(y)$ has $k = 0$. Thus, one can draw the nonsampled values one by one using (8). The speed of this process is increased by drawing from $\bar{F}_{n+k}(y)$ using the multinomial distribution because there are repeats among the values already drawn. Thus, we get 10,000 values of $\bar{Y}$. Order these values and pick the 95% prediction interval as $(\bar{Y}_{(250)}, \bar{Y}_{(9750)})$, where the values are arranged in increasing order.

In the Appendix A we describe some features of the Bayes estimator of $\bar{Y}$ for the normal baseline under a squared error loss function. For completeness, we discuss the Bayes estimator of $\bar{Y}$ in the Appendix A. We have shown that under the marginal distribution (integrating out $G$) of $y_1, \ldots, y_n$ in the DP model, the expected difference between the Bayes estimator and the finite population mean is 0 and the expected squared difference is zero for large $n$, $k$, $N$.

## 3. Examples, simulation studies, and leave-one-out kernel baseline

We discuss two examples and a simulation study to assess the extent of the sensitivity of inference about the finite population mean of the normal baseline. Specifically, we apply the DP model with the six baselines, which are normal (NO), lognormal (LN), gamma (GA), inverse Gaussian (IG), two-component mixture (MI) and skewed normal (SN) and two nonparametric baselines which are the Polya posterior (PP) and the Bayesian bootstrap (BB); occasionally, for convenience, we have called PP and BB baselines as well. As a solution, we provide a leave-one-out kernel density estimator for the baseline model.

### 3.1. Examples

The first example is on the third National Health and Nutrition Examination Survey (NHANES III). These are the data on body mass index for females older than forty-five years where we assume that an equivalent simple random sample is taken from a US state. The sample size is 45 with 20 distinct values and the population size is 190,472, making the prediction problem challenging in terms of time. The second data set is taken from [2] on income which he used to discuss finite population sampling. This is a much smaller population with a size of 648 and a sample size of 40 with 30 distinct values. In both cases, histograms (omitted) of the sampled values are right skewed.

We consider inference for the finite population mean in Table 2 for the income data and Table 3 for the body mass index data. Specifically, we have used posterior mean (PM), posterior standard deviation (PSD), numerical standard errors (NSE) and 95% credible intervals for PP, BB and the six baseline models. For both data sets there are small differences among the PMs but much larger differences among the PSDs. For the income data IG stands out with much larger PSD (PSD = 7.870) and for the body mass index data it is MI that stands out with a largest PSD (PSD = 1.436). These differences are reflected in the 95% credible intervals. The NSEs are small.

We have plotted the posterior densities of the finite population mean in Fig. 1 for the income data and Fig. 2 for the body mass index data. There are large differences among PP, BB and the six parametric baselines. These differences occur mainly around the modes. For the income data the inverse Gaussian (IG) gives a much wider distribution than the others, PP and BB are similar but are different from the normal, lognormal, gamma, two-component mixture and the skewed normal baselines which are also different among themselves. For these six baselines, the differences are much more pronounced for the body mass data. Also it is the skewed normal baseline, not the inverse Gaussian, which is much different from the others.

Thus, it is clear that inference about the finite population can be different from the normal baseline when other appropriate baselines are used. In particular, if a baseline, other than the normal is used,

**Table 2**
Posterior inference of the finite population mean for the income data using the Polya posterior, the Bayesian bootstrap and six baseline distributions.

| Baseline | PM | PSD | NSE | 95% CI |
|---|---|---|---|---|
| PP | 67.102 | 3.390 | 0.035 | (60.995, 74.289) |
| BB | 67.091 | 3.370 | 0.034 | (60.716, 73.736) |
| NO | 67.910 | 3.484 | 0.034 | (60.919, 74.701) |
| LN | 68.352 | 3.804 | 0.038 | (60.900, 75.688) |
| GA | 68.021 | 3.665 | 0.039 | (60.692, 75.132) |
| IG | 67.685 | 7.870 | 0.078 | (54.098, 82.978) |
| MI | 68.297 | 3.600 | 0.046 | (61.201, 75.301) |
| SN | 68.378 | 3.389 | 0.032 | (61.973, 75.217) |

Note: PM is the posterior mean; PSD is the posterior standard deviation; NSE is the numerical standard error; CI is the credible interval. Each procedure uses 10,000 draws from the posterior density. The Polya posterior (PP) takes $\alpha = 0$ in the simple Dirichlet process and the Bayesian bootstrap (BB) uses Haldane prior for multinomial sampling. The income data has $N = 648$ and $n = 40$.

**Table 3**
Posterior inference of the finite population mean for the body mass index data using the Polya posterior, the Bayesian bootstrap and six baseline distributions.

| Baseline | PM | PSD | NSE | 95% CI |
|---|---|---|---|---|
| PP | 28.473 | 1.126 | 0.041 | (26.365, 30.679) |
| BB | 28.381 | 1.092 | 0.034 | (26.505, 30.535) |
| NO | 28.740 | 1.257 | 0.037 | (26.575, 31.446) |
| LN | 28.748 | 1.210 | 0.034 | (26.485, 31.115) |
| GA | 28.812 | 1.244 | 0.043 | (26.680, 31.470) |
| IG | 28.318 | 1.314 | 0.030 | (26.065, 30.786) |
| MI | 29.823 | 1.436 | 0.063 | (27.311, 32.810) |
| SN | 28.806 | 1.169 | 0.041 | (26.756, 31.316) |

Note: PM is the posterior mean; PSD is the posterior standard deviation; NSE is the numerical standard error; CI is the credible interval. Each procedure uses 1000 draws from the posterior density. The Polya posterior (PP) takes $\alpha = 0$ in the simple Dirichlet process and the Bayesian bootstrap (BB) uses Haldane prior for multinomial sampling. The BMI data are positively skewed. The BMI data set has a single US state for females older than 45 years, $N = 190{,}472$ and $n = 45$.

inference about the finite population mean can change. Although not reported here, it is also true that inference about a population quantile (e.g., median) will vary with these baselines. This depends on the sample size and the population size as well.

### 3.2. Simulation study

We assess the frequentist properties of the posterior mean of the finite population mean. We have restricted our simulation study to population sizes similar to Aitkin [2].

We have drawn our sample using a Parzen–Rosenblatt kernel density estimate with a window width obtained using the income data. Let $y_1, \ldots, y_n$ denote the sample of size $n = 40$ observations from the income data. So that

$$\widehat{f(y)} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_o}\phi\left(\frac{y - y_i}{h_o}\right), \quad -\infty < y < \infty,$$

where $h_o$ is the optimal window width [29] and $\phi(.)$ is the standard normal density. We have drawn a random sample (population) of size $N$ from this kernel density and we have taken a simple random sample of size $n$ from this selected population. We have considered sample sizes and population sizes $(n, N) = [(25, 250), (50, 500), (100, 1000)]$. We have drawn 1000 samples at each of the three design points $(n, N)$. Thus, we know the finite population mean. For each run we have computed the posterior mean (PM), posterior standard deviation (PSD) and the 95% credible interval of the finite

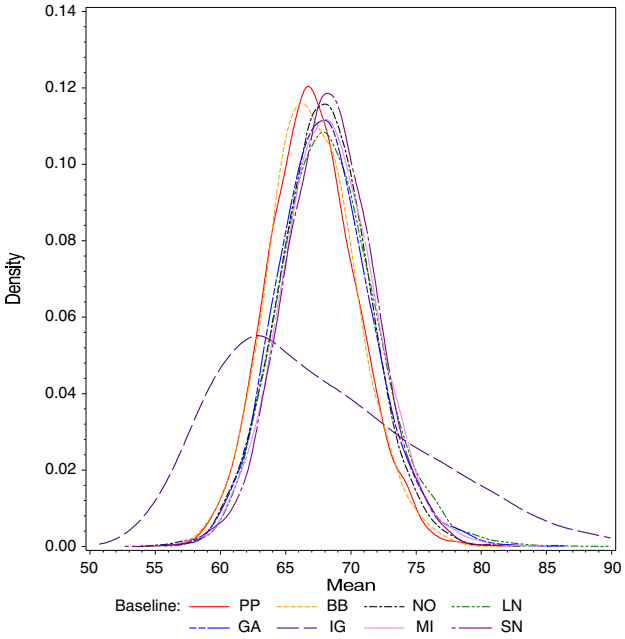


**Fig. 1.** Plots of the posterior density of the finite population mean by baseline model for income data.

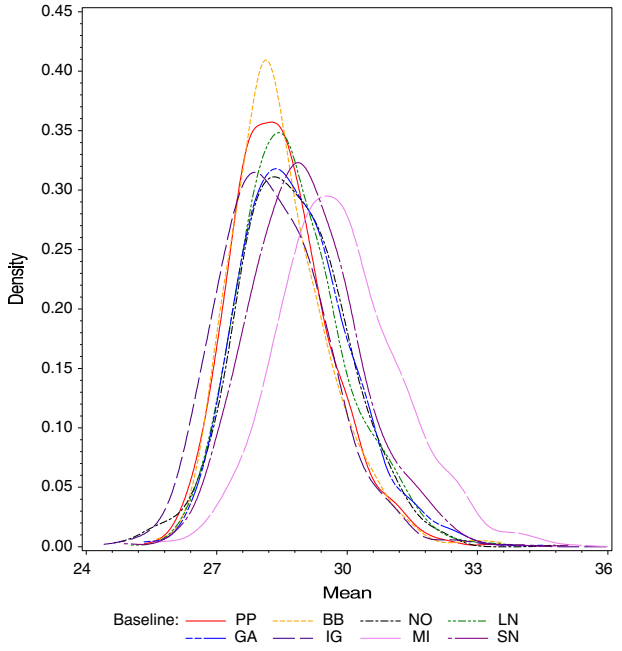


**Fig. 2.** Plots of the posterior density of the finite population mean by baseline model for body mass index data.

population mean in the same manner as for the two examples. We have done this for all six baseline models (NO, LN, GA, IG, MI, SN), the Polya posterior (PP) and the Bayesian bootstrap (BB).

**Table 4**
Simulation study: Comparison of coverage ($C$), relative bias (Rbias), posterior root mean squared error (PRMSE) and width (Wid) by sample size ($n$), population size ($N$) and baseline model.

| $n$ | $N$ | Baseline | Rbias | PRMSE | $C$ | Wid |
|---|---|---|---|---|---|---|
| 25 | 250 | PP | −0.002; 0.002 | 5.497; 0.055 | 0.923; 0.008 | 15.60; 0.073 |
| | | BB | −0.004; 0.002 | 5.636; 0.059 | 0.911; 0.009 | 15.55; 0.073 |
| | | NO | 0.002; 0.002 | 5.636; 0.051 | 0.941; 0.007 | 16.25; 0.067 |
| | | LN | 0.009; 0.002 | 5.982; 0.053 | 0.946; 0.007 | 17.29; 0.080 |
| | | GA | 0.009; 0.002 | 5.759; 0.048 | 0.957; 0.006 | 17.02; 0.070 |
| | | IG | 0.005; 0.001 | 7.467; 0.203 | 1.000; 0.000 | 24.74; 0.090 |
| | | MI | 0.013; 0.002 | 5.666; 0.052 | 0.951; 0.007 | 16.90; 0.065 |
| | | SN | 0.005; 0.002 | 5.571; 0.053 | 0.938; 0.008 | 16.07; 0.068 |
| 50 | 500 | PP | 0.001; 0.001 | 4.088; 0.042 | 0.918; 0.009 | 11.33; 0.038 |
| | | BB | 0.001; 0.001 | 4.025; 0.040 | 0.929; 0.008 | 11.34; 0.037 |
| | | NO | 0.002; 0.001 | 3.888; 0.036 | 0.952; 0.007 | 11.43; 0.033 |
| | | LN | 0.006; 0.001 | 4.019; 0.036 | 0.953; 0.007 | 11.89; 0.037 |
| | | GA | 0.007; 0.001 | 3.927; 0.034 | 0.959; 0.006 | 11.70; 0.032 |
| | | IG | 0.003; 0.001 | 4.806; 0.025 | 0.998; 0.001 | 16.13; 0.039 |
| | | MI | 0.008; 0.001 | 3.930; 0.035 | 0.957; 0.006 | 11.61; 0.031 |
| | | SN | 0.004; 0.001 | 3.889; 0.035 | 0.957; 0.006 | 11.35; 0.032 |
| 100 | 1000 | PP | −0.001; 0.001 | 2.846; 0.028 | 0.938; 0.008 | 8.093; 0.020 |
| | | BB | 0.001; 0.001 | 2.855; 0.027 | 0.940; 0.008 | 8.118; 0.020 |
| | | NO | 0.002; 0.001 | 2.739; 0.024 | 0.962; 0.006 | 8.073; 0.018 |
| | | LN | 0.004; 0.001 | 2.783; 0.024 | 0.965; 0.006 | 8.247; 0.018 |
| | | GA | 0.003; 0.001 | 2.796; 0.026 | 0.956; 0.006 | 8.165; 0.017 |
| | | IG | 0.001; 0.001 | 3.154; 0.023 | 0.985; 0.004 | 9.886; 0.017 |
| | | MI | 0.004; 0.001 | 2.809; 0.025 | 0.960; 0.006 | 8.143; 0.016 |
| | | SN | 0.004; 0.001 | 2.803; 0.027 | 0.944; 0.007 | 8.067; 0.017 |

Note: There are 1000 runs in the simulations and for each run the 95% credible intervals, Rbias, posterior root mean squared error and credible incidence (whether an interval contains the true value) and wid of the credible intervals are calculated. The first number in each entry is the average over the 1000 runs and of the second number is the standard error of the average. The Polya posterior (PP) takes $\alpha = 0$ in the simple Dirichlet process and the Bayesian bootstrap (BB) uses the Haldane prior for multinomial sampling.

In Table 4 we investigate frequentist properties such as relative bias, mean squared error, width of 95% credible intervals and the coverage of the intervals. The relative bias is Rbias $= (PM - \bar{Y})/\bar{Y}$, the posterior root mean squared error is PRMSE $= \sqrt{(PM - \bar{Y})^2 + PSD^2}$, the width (wid) is the difference between the upper end and the lower end of the 95% credible interval, and the credible incidence is 1 if the 95% credible interval contains the true value and 0 otherwise. We have taken the average of these quantities over the 1000 runs. The average of the credible incidences is the coverage ($C$).

Rbias is negligible for all baselines, PP and BB; the largest bias occurs for MI at the three design points. As expected, PRMSE must get smaller as ($n$, $N$) increase; and the PRMSE varies from different baseline models. The variation is not so large for (50, 500) and (100, 1000) but it is a bit larger for (25, 250) with MI standing out (PRMSE $= 7.467$) and PP and SN with the smallest PRMSE, not NO which has the smallest PRMSE at (50, 500) and (100, 1000). The coverage C is generally good for all baseline models conservative for IG (C $=$ 1.000, 0.998, 0.985) which has larger width (wid $=$ 24.74, 16.13, 9.886) at the three design points.

We have also looked at the sampling distributions of the Bayes estimator (posterior mean under square error loss) of the finite population mean at the three design points. Again, we have used a Parzen–Rosenblatt kernel density estimator for the 1000 simulated posterior means. In Figs. 3–5 we have shown the three plots which give much differences around the mode of the sampling distributions especially at (25, 250); at this design point there are also much differences in the tails of the sampling distributions.

Thus, it is clear that there are considerable differences from the normal baseline. In a practical application we cannot simply assume normality as the baseline model. Perhaps, a nonparametric continuous distribution (a kernel density) should be used, but it is not clear how to do a coherent Bayesian analysis with a nonparametric baseline.

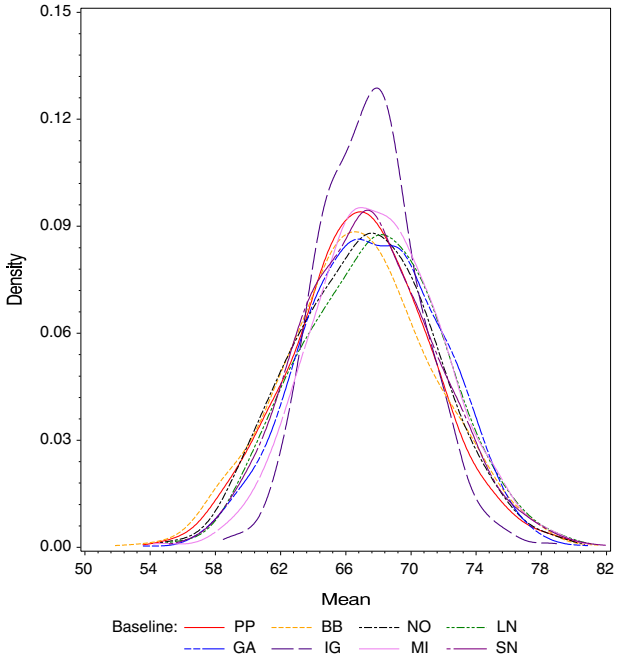


**Fig. 3.** Plots of the sampling distribution of the Bayes estimator of the finite population mean by baseline model for simulated data ($n = 25, N = 250$).

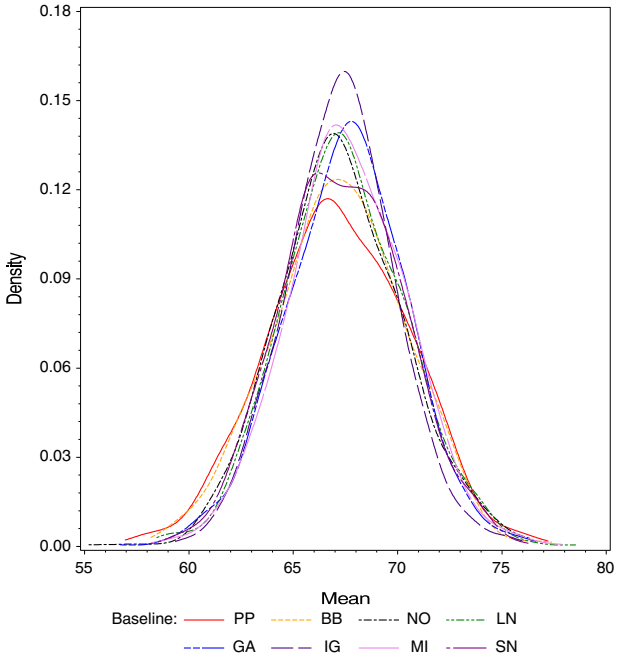


**Fig. 4.** Plots of the sampling distribution of the Bayes estimator of the finite population mean by baseline model for simulated data ($n = 50, N = 500$).

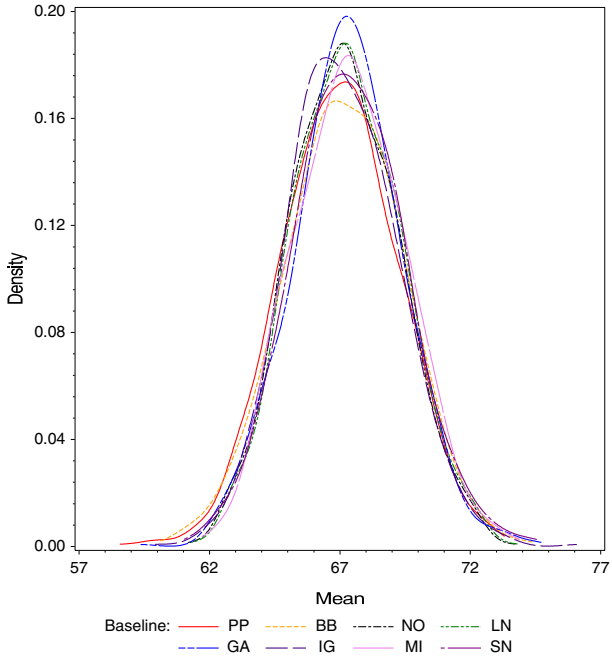


**Fig. 5.** Plots of the sampling distribution of the Bayes estimator of the finite population mean by baseline model for simulated data ($n = 100, N = 1000$).

### 3.3. Leave-one-out kernel baseline

We present a solution to the sensitivity problem faced by the DP model. Clearly, a solution has to be based on a nonparametric distribution. As we noted, the Monte Carlo method of McAuliffe, Blei and Jordan [22] cannot be used for the DP model. So we use the leave-one-out kernel density estimator.

Hardle [15] described the leave-one-out kernel density estimator; a Bayesian version (again not fully within the Bayesian paradigm) is available (e.g., [6,17]). With a single parameter $\psi$ for the window width, we assume that $y_1, \ldots, y_k \mid \psi$ are independent with

$$f(y_i \mid \psi) = \frac{1}{k-1} \sum_{j=1, j\neq i}^{k} \frac{1}{\psi} \phi\left(\frac{y_i - y_j}{\psi}\right), \quad -\infty < y_i < \infty,$$

where $\phi(\cdot)$ is the standard normal density function (e.g., [29]). We take

$$\pi(\psi) = \frac{1}{(1+\psi)^2}, \quad \psi \geq 0.$$

So, the posterior density of $\psi$ is

$$\pi(\psi \mid \underset{\sim}{y}_k) \propto \frac{1}{(1+\psi)^2} \prod_{i=1}^{k} \frac{1}{k-1} \sum_{j=1, j\neq i}^{k} \frac{1}{\psi} \phi\left(\frac{y_i - y_j}{\psi}\right), \quad \psi \geq 0.$$

Therefore, the data are used many times and again this procedure is a bit problematic for Bayesian inference because $\psi$ is not really a parameter of the DP model. Otherwise, there is not much that one can do.

In the spirit of our computations, it is easy to use a grid method to draw samples of $\psi$. For prediction of a future $y$ value we use

$$f(y \mid \psi) = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{\psi} \phi\left(\frac{y - y_i}{\psi}\right), \quad -\infty < y < \infty,$$

where a random value, say $t$, in $(1, \ldots, k)$, is drawn and $y \sim \text{Normal}(y_t, \psi^2)$; see Section 2.3 for prediction from the DP.

For the income data we obtained $PM = 67.846$, $PSD = 3.387$, $NSE = 0.037$ with a 95% credible interval of (61.428, 74.872). The PM is closest to that of the normal baseline and the PSD is closest to that of the Polya posterior or the Bayesian bootstrap in Table 2. For the body mass index data we got $PM = 28.706$, $PSD = 1.264$, $NSE = 0.044$ with a 95% credible interval of (26.537, 31.287). The PM is closest to the normal or the lognormal baseline and the PSD is closest to that of normal or gamma baseline in Table 3. Although these differences from the normal baseline are small, again one cannot naively use the normal baseline.

## 4. Extension to Dirichlet process mixture model

The Dirichlet process mixture (DPM) model is

$$y_{ij} \mid \mu_i \overset{ind}{\sim} f(y_{ij} \mid \mu_i, \tau), \quad i = 1, \ldots, \ell,\ j = 1, \ldots, n_i,$$
$$\mu_1, \ldots, \mu_\ell \mid G \overset{\text{i.i.d.}}{\sim} G \quad \text{and} \quad G \mid G_o(\underset{\sim}{\psi}) \sim DP\{\alpha, G_o(\underset{\sim}{\psi})\}$$

for $\ell$ groups (areas) of data. The procedure for $\alpha$ is exactly the same as in the DP model. However, the problem is not so simple for $G_o(\psi)$. We would set up a mixture model for the $\mu_i$ (i.e., $G_o(\psi)$). It is apparent that this cannot be done without "double" using the data (e.g., see [22]) (i.e., a Bayes empirical Bayes analysis has to be done). This is true because using only the Dirichlet process prior a density estimator cannot be constructed. Thus, one must use a data-based Dirichlet process prior and we discuss how to construct one in a sensible manner with little computational effort. Our main purpose is to show how an applied statistician might proceed in a sensitivity analysis.

Let us consider a specific DPM,

$$y_{ij} \mid \theta, \nu_i, \sigma^2 \overset{ind}{\sim} \text{Normal}(\theta + \nu_i, \sigma^2), \quad i = 1, \ldots, \ell,\ j = 1, \ldots, n_i,$$
$$\nu_1, \ldots, \nu_n \mid G \overset{\text{i.i.d.}}{\sim} G \quad \text{and} \quad G \mid G_o(\sigma^2, \rho) \sim DP\{\alpha, G_o(\sigma^2, \rho)\},$$
$$\pi(\theta, \sigma^2, \rho, \alpha) \propto \frac{1}{\sigma^2(1+\alpha)^2}, \quad -\infty < \theta < \infty,\ 0 < \rho < 1,\ \sigma^2, \alpha > 0,$$

where $G_o(\sigma^2, \rho)$ is the cdf of a normal random variable with mean 0 and variance $\frac{\rho}{1-\rho}\sigma^2$ (the baseline distribution). Taking the limit as $\alpha$ goes to infinity, the baseline model is

$$y_{ij} \mid \theta, \nu_i, \sigma^2 \overset{ind}{\sim} \text{Normal}(\theta + \nu_i, \sigma^2), \quad i = 1, \ldots, \ell,\ j = 1, \ldots, n_i,$$
$$\nu_i \mid \sigma^2, \rho \overset{ind}{\sim} \text{Normal}\left(0, \frac{\rho}{1-\rho}\sigma^2\right),$$
$$\pi(\theta, \sigma^2, \rho) \propto \frac{1}{\sigma^2}, \quad -\infty < \theta < \infty,\ \sigma^2 > 0,\ 0 < \rho < 1;$$

see [25]. We would use this baseline model to construct a finite mixture of $\ell$ components.

Sufficient statistics are $(\bar{y}_i, s_i^2),\ i = 1, \ldots, \ell$, where $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and $s_i^2 = \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(n_i - 1)$. Letting $\lambda_i = n_i\rho/\{(n_i - 1)\rho + 1\},\ i = 1, \ldots, \ell, \bar{y} = \sum_{i=1}^{\ell} \lambda_i \bar{y}_i / \sum_{i=1}^{\ell} \lambda_i$ and $n = \sum_{i=1}^{\ell} n_i$, it is well known that

$$\nu_i \mid \theta, \sigma^2, \rho, \underset{\sim}{y} \overset{ind}{\sim} \text{Normal}\left\{\lambda_i(\bar{y}_i - \theta), (1 - \lambda_i)\frac{\rho}{1-\rho}\sigma^2\right\}, \quad i = 1, \ldots, \ell, \tag{9}$$
$$\theta \mid \sigma^2, \rho, \underset{\sim}{y} \sim \text{Normal}\left\{\bar{y}, \frac{\rho\sigma^2}{(1-\rho)} \Big/ \sum_{i=1}^{\ell} \lambda_i\right\},$$

$$\sigma^{-2} \mid \rho, \underset{\sim}{y} \sim \text{Gamma}\left\{\frac{n-1}{2}, \frac{\sum_{i=1}^{\ell}(n_i-1)s_i^2 + (1-\rho)/\rho \sum_{i=1}^{\ell}\lambda_i(\bar{y}_i-\bar{y})^2}{2}\right\},$$

$$\pi(\rho \mid \underset{\sim}{y}) \propto \frac{1}{\sqrt{\sum_{i=1}^{\ell}\lambda_i}}\sqrt{\frac{\rho}{1-\rho}}\left\{\prod_{i=1}^{\ell}\sqrt{1-\lambda_i}\right\}\left\{1+\frac{1-\rho}{\rho}\frac{\sum_{i=1}^{\ell}\lambda_i(\bar{y}_i-\bar{y})^2}{\sum_{i=1}^{\ell}(n_i-1)s_i^2}\right\}^{-(n-1)/2},$$

$$0 \le \rho \le 1.$$

It is easy to show that $\pi(\rho \mid \underset{\sim}{y})$ is well defined for all $\rho$ in $[0, 1]$. Hence, the joint posterior density is proper provided $n_i \ge 2,\ i = 1, \ldots, \ell, \ell \ge 2$; see [25].

When $\theta$ is integrated out of (9), the $\nu_i$ will become correlated. Specifically,

$$\nu_i \mid \sigma^2, \rho, \underset{\sim}{y} \sim \text{Normal}\left[\lambda_i(\bar{y}_i-\bar{y}), \left\{(1-\lambda_i)+\lambda_i^2\Big/\sum_{i=1}^{\ell}\lambda_i\right\}\frac{\rho}{1-\rho}\sigma^2\right], \quad i = 1, \ldots, \ell,$$

and

$$\text{cov}(\nu_i, \nu_{i'} \mid \sigma^2, \rho, \underset{\sim}{y}) = \frac{\lambda_i\lambda_{i'}}{\sum_{i=1}^{\ell}\lambda_i}\frac{\rho}{1-\rho}\sigma^2, \quad i \ne i' = 1, \ldots, \ell.$$

It is convenient that when $\rho$ goes to zero, $\text{cov}(\nu_i, \nu_{i'} \mid \sigma^2, \rho, \underset{\sim}{y})$ goes to zero. In fact, for small area estimation this correlation is generally small.

Next, for $i = 1, \ldots, \ell$, let

$$\tilde{y}_i = \lambda_i(\bar{y}_i-\bar{y}),$$

$$\tilde{s}_i^2 = \left\{(1-\lambda_i)+\lambda_i^2\Big/\sum_{i=1}^{\ell}\lambda_i\right\}\left\{\frac{\sum_{i=1}^{\ell}\lambda_i(\bar{y}_i-\bar{y})^2 + \frac{\rho}{1-\rho}\sum_{i=1}^{\ell}(n_i-1)s_i^2}{n-1}\right\}.$$

Then, integrating $\sigma^2$ out, we get

$$\pi(\nu_i \mid \rho, \underset{\sim}{y}) = \frac{\Gamma(n/2)}{\Gamma\{(n-1)/2\}\tilde{s}_i\sqrt{\pi(n-1)}}\frac{1}{\left\{1+\frac{1}{n-1}\left\{\frac{\nu_i-\tilde{y}_i}{\tilde{s}_i}\right\}^2\right\}^{(n-1)/2}},$$

a Student's $t$ density with location $\tilde{y}_i$, scale $\tilde{s}_i^2$ and degrees of freedom $n-1$ which we denote by $t_{n-1}(\nu \mid \tilde{y}_i, \tilde{s}_i^2)$. These are the $\ell$ marginal densities of the full multivariate Student's $t$ density which is easy to write down, but it is not pertinent here.

However, $\rho$ is unknown and it must be estimated as well. Fortunately, in these problems $n$ is large enough making the Student's $t$ approximately normal. So we just need to know the mean and variance of normal density. Therefore, we integrate $\rho$ from $\tilde{y}_i$ and $\tilde{s}_i^2$ to get updated values as

$$\tilde{y}_i = E\{\lambda_i(\bar{y}_i-\bar{y}) \mid \underset{\sim}{y}\},$$

$$\tilde{s}_i^2 = E\left[\left\{(1-\lambda_i)+\lambda_i^2\Big/\sum_{i=1}^{\ell}\lambda_i\right\}\left\{\frac{\sum_{i=1}^{\ell}\lambda_i(\bar{y}_i-\bar{y})^2 + \frac{\rho}{1-\rho}\sum_{i=1}^{\ell}(n_i-1)s_i^2}{n-1}\right\} \mid \underset{\sim}{y}\right]$$
$$+ \text{Var}\{\lambda_i(\bar{y}_i-\bar{y}) \mid \underset{\sim}{y}\}, \quad i = 1, \ldots, \ell,$$

where expectation is taken over the posterior density of $\rho$.

We are now ready to construct $g_o(\nu \mid \underset{\sim}{\omega})$ where $\omega_{i'} > 0$, $\sum_{i'=1}^{\ell} \omega_{i'} = 1$. That is, the mixture distribution is

$$g_o(\nu \mid \underset{\sim}{\omega}) = \sum_{i'=1}^{\ell} \frac{\omega_{i'}}{\tilde{s}_{i'}} \phi\left(\frac{\nu - \tilde{y}_{i'}}{\tilde{s}_{i'}}\right),$$

where $\phi(\cdot)$ is the standard normal density. Let $G_o(\nu \mid \underset{\sim}{\omega})$ denote the corresponding cdf. Clearly, while $g_o(\nu \mid \underset{\sim}{\omega})$ depends on the sufficient statistics, and therefore, we have essentially constructed a finite mixture of normal random variables which we use as a basis to get an exchangeable sequence of random variables from a Dirichlet process prior.

Therefore, our full DPM model is

$$\begin{aligned}
&y_{ij} \mid \theta, \nu_i, \sigma^2 \overset{ind}{\sim} \text{Normal}(\theta + \nu_i, \sigma^2), \quad i = 1, \ldots, \ell,\ j = 1, \ldots, n_i,\\
&\nu_1, \ldots, \nu_\ell \mid G \overset{\text{i.i.d.}}{\sim} G, \qquad G \sim DP\{\alpha, G_o(\nu \mid \underset{\sim}{\omega})\}\\
&\pi(\theta, \sigma^2, \alpha) \propto \frac{1}{\sigma^2(1+\alpha)^2}, \quad -\infty < \theta < \infty,\ \sigma^2 > 0,\ \alpha > 0 \quad \text{and}\\
&\underset{\sim}{\omega} \sim \text{Dirichlet}(1, \ldots, 1).
\end{aligned}$$

This model can be fit efficiently using the algorithm of Kalli, Griffin and Walker [19] which is particularly attractive because it uses the stick-breaking construction [28] to obtain a properly mixing Markov chain. A similar DPM can be obtained when there are covariates. Details are given in the supplement (see Appendix B).

## 5. Concluding remarks

Our key contribution has been to investigate the extent of the sensitivity of inference of the specification of the baseline distribution in the Dirichlet process using simple random sampling from a finite population. We have compared six baselines which are the normal, lognormal, gamma, inverse Gaussian, a mixture of two normals and positively skewed normal. We have also compared the DP model with the Polya posterior and the Bayesian bootstrap. We have used two examples with positively skewed data and we performed a simulation study with simple random samples drawn from the Parzen–Rosenblatt kernel density estimator (a procedure not favorable to any of our baselines, Polya posterior or Bayesian bootstrap). We have provided a possible solution using a leave-one-out kernel density estimator.

Because our work is on Bayesian predictive inference in sample surveys, we have used this area to illustrate our findings. Two examples show that the posterior densities of the finite population mean are generally different among the baselines especially around their modes and the tails. This can change the posterior inference of the finite population mean or quantile. The simulation helps to confirm that the frequentist properties of the posterior mean can be different over baselines.

For the Dirichlet process mixture (DPM) model, while McAuliffe, Blei and Jordan [22] have provided a good solution to the sensitivity problem, we have proposed a solution which is based on the baseline model of the DPM. It is interesting that in the DPM our baseline model is a mixture of $\ell$ components (corresponding to the groups or areas). Our proposed model can be fit with less computational effort and we can potentially provide a coherent Bayesian analysis rather than an empirical Bayes analysis. In survey sampling there are similar past data or a census, typically similar to the current data, which can be used to construct our proposed baseline distribution.

## Appendix A. Features of the Bayes estimator

Let $\hat{\bar{Y}}$ denote the Bayes estimator of $\bar{Y}$ under squared error loss. We want to check how close $\hat{\bar{Y}}$ is to $\bar{Y}$.

Under the normal baseline, using the DP model and letting $\lambda = \frac{n(\alpha+N)}{N(\alpha+n)}$, we have $\mathrm{E}(\bar{Y} \mid \mu, \sigma^2, \alpha, \underset{\sim}{y}) = \lambda\bar{y} + (1-\lambda)\mu$. Also, under the normal baseline $\mu \mid \sigma^2, \underset{\sim}{y}_k \sim \mathrm{Normal}(\bar{y}_k, \sigma^2/k)$. Thus, the Bayes estimator of $\bar{Y}$, where $\hat{\bar{Y}} = \mathrm{E}(\bar{Y} \mid \sigma^2, \alpha, \underset{\sim}{y})$, is

$$\hat{\bar{Y}} = \lambda\bar{y} + (1-\lambda)\bar{y}_k.$$

We evaluate $\hat{\bar{Y}}$ under the marginal distribution of $y_1, \ldots, y_N$ after integrating out $G$ from the DP (i.e., conditional on $(\mu, \sigma^2, \alpha)$). For the rest of the appendix we drop the conditioning on $\mu$ and $\sigma^2$. Denote the marginal distribution by $\mathcal{M}$. Now, $y_1, \ldots, y_N$ form a generalized Polya urn scheme and are exchangeable [5], and $\mathrm{E}_{\mathcal{M}}(y_i) = \mu$, $\mathrm{Var}_{\mathcal{M}}(y_i) = \sigma^2$, $\mathrm{Cov}_{\mathcal{M}}(y_i, y_j) = \sigma^2/(\alpha+1)$, $i \neq j = 1, \ldots, N$. This is true for any baseline distribution with mean, $\mu$, and variance, $\sigma^2$. We will examine the difference, $\hat{\bar{Y}} - \bar{Y}$, under $\mathcal{M}$, and henceforth, we will drop the subscript $\mathcal{M}$.

First, $\mathrm{E}(\hat{\bar{Y}}) = \mu$ and $\mathrm{E}(\bar{Y}) = \mu$. Thus, $\mathrm{E}(\hat{\bar{Y}} - \bar{Y}) = 0$. That is, $\hat{\bar{Y}} - \bar{Y}$ is an unbiased estimator of 0. Second, we compute $\mathrm{Var}(\hat{\bar{Y}} - \bar{Y})$. It is easy to show that

$$\mathrm{Var}(\hat{\bar{Y}} - \bar{Y}) = \mathrm{Var}(\hat{\bar{Y}}) + \mathrm{Var}(\bar{Y}) - 2\,\mathrm{Cov}(\hat{\bar{Y}}, \bar{Y}),$$

where

$$\mathrm{Var}(\bar{Y}) = \frac{\sigma^2(\alpha+N)}{N(\alpha+1)},$$

$$\mathrm{Var}(\hat{\bar{Y}}) = \lambda^2\frac{\sigma^2(\alpha+n)}{n(\alpha+1)} + (1-\lambda)^2\frac{\sigma^2(\alpha+k)}{k(\alpha+1)} + 2\lambda(1-\lambda)\frac{\sigma^2(\alpha+n)}{n(\alpha+1)},$$

$$\mathrm{Cov}(\hat{\bar{Y}}, \bar{Y}) = \frac{\sigma^2(\alpha+N)}{N(\alpha+1)}.$$

Consider the standard survey sampling asymptotics as $n, k, N-n$ go to infinity such that $nN^{-1}$ goes to $a$ and $kN^{-1}$ goes to $b$, where $a \geq b$ are constants. Then, $\lambda$ goes to 1 and $\mathrm{Var}(\hat{\bar{Y}} - \bar{Y})$ goes to zero. Thus, $\hat{\bar{Y}} - \bar{Y}$ is an unbiased and a consistent estimator of 0.

## Appendix B. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.stamet.2015.07.003.

## References

[1] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables, Dover Publications, New York, 1965.
[2] M. Aitkin, Statistical Inference: An Integrated Bayesian/Likelihood Approach, Chapman & Hall, New York, 2010.
[3] C.E. Antoniak, Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, Ann. Statist. 2 (1974) 1152–1174.
[4] D.A. Binder, Non-parametric Bayesian models for samples from finite populations, J. R. Stat. Soc. Ser. B 44 (1982) 388–393.
[5] D. Blackwell, J.B. MacQueen, Ferguson distributions via Polya urn schemes, Ann. Statist. 1 (2) (1973) 353–355.
[6] M.J. Brewer, A Bayesian model for local smoothing in kernel density estimation, Stat. Comput. 10 (2000) 299–309.
[7] S. Chaudhuri, M. Ghosh, Empirical likelihood for small area estimation, Biometrika 87 (2011) 633–649.
[8] D.B. Dunson, Nonparametric Bayes applications to biostatistics, in: N.L. Hjort, C. Holmes, P. Muller, S.G. Walker (Eds.), Bayesian Nonparametrics, Chapman and Hall, New York, 2010, pp. 223–273 (Chapter 7).
[9] W.A. Ericson, Subjective Bayesian models in sampling finite populations (with discussions), J. R. Stat. Soc. Ser. B 31 (1969) 195–233.
[10] M.D. Escobar, M. West, Bayesian density estimation and inference using mixtures, J. Amer. Statist. Assoc. 90 (1995) 577–588.
[11] T.S. Ferguson, A Bayesian analysis of some nonparametric problems, Ann. Statist. 1 (1973) 209–230.
[12] T.S. Ferguson, Bayesian density estimation by mixtures of normal distributions, in: Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday, Academic Press, New York, 1983, pp. 287–302.
[13] M. Ghosh, G. Meeden, Bayesian Methods for Finite Population Sampling, Cambridge University Press, Cambridge, 1997.

[14] L.H. Hannah, D.M. Blei, W.B. Powel, Dirichlet process mixtures of generalized linear models, J. Mach. Learn. Res. (2011) 1–33.
[15] W. Hardle, Smoothing Techniques: With Implementation in S, Springer, New York, 1991.
[16] N.L. Hjort, C. Holmes, P. Muller, S.G. Walker, Bayesian Nonparametrics, Chapman and Hall, New York, 2010.
[17] S. Hu, D.S. Poskitt, X. Zhang, Bayesian adaptive bandwidth kernel density estimation of irregular multivariate distributions, Comput. Statist. Data Anal. 56 (2012) 732–740.
[18] H. Ishwaran, L.F. James, Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information, J. Comput. Graph. Statist. 11 (2002) 508–532.
[19] M. Kalli, J.E. Griffin, S.G. Walker, Slice sampling mixture models, Stat. Comput. 21 (2011) 93–105.
[20] A.Y. Lo, On a class of Bayesian nonparametric estimates: I. Density estimates, Ann. Statist. 12 (1984) 351–357.
[21] A.Y. Lo, A Bayesian bootstrap for a finite population, Ann. Statist. 26 (1988) 1684–1695.
[22] J.D. McAuliffe, D.M. Blei, M., I. Jordan, Nonparametric empirical Bayes for the Dirichlet process mixture model, Stat. Comput. 16 (2006) 5–14.
[23] I. Molina, B. Nandram, J.N.K. Rao, Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach, Ann. Appl. Stat. 8 (2) (2014) 852–885.
[24] B. Nandram, J.W. Choi, Nonparametric Bayesian analysis of a proportion for small area under nonignorable nonresponse, J. Nonparametr. Stat. 16 (2004) 821–839.
[25] B. Nandram, M.C.S. Toto, J.W. Choi, A Bayesian benchmarking of the Scott–Smith model for small areas, J. Stat. Comput. Simul. 81 (11) (2011) 1593–1608.
[26] N.G. Polson, J.G. Scott, On the half-Cauchy prior for a global scale parameter, Bayesian Anal. 7 (2012) 887–902.
[27] D.B. Rubin, The Bayesian bootstrap, Ann. Statist. 9 (1981) 130–134.
[28] J. Sethuraman, A constructive definition of Dirichlet priors, Statist. Sinica 4 (1994) 639–650.
[29] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, New York, 1986.
[30] Y.W. Teh, M.I. Jordan, M.J. Beal, D.M. Blei, Hierarchical Dirichlet processes, J. Amer. Statist. Assoc. 101 (2006) 1566–1580.
[31] M.C.S. Toto, B. Nandram, A Bayesian predictive inference for small area means incorporating covariates and sampling weights, J. Statist. Plann. Inference 140 (2010) 2963–2979.