Yin, Jiani; Nandram, Balgobin

**Article**

# A Bayesian Small Area Model with Dirichlet Processes on the Responses

Statistics in Transition New Series

**Provided in Cooperation with:**
Polish Statistical Association

*Suggested Citation:* Yin, Jiani; Nandram, Balgobin (2020) : A Bayesian Small Area Model with Dirichlet Processes on the Responses, Statistics in Transition New Series, ISSN 2450-0291, Exeley, New York, Vol. 21, Iss. 3, pp. 1-19,
https://doi.org/10.21307/stattrans-2020-041

*STATISTICS IN TRANSITION new series, September 2020*
*Vol. 21, No. 3, pp. 1–19, DOI 10.21307/stattrans-2020-041*
Received – 30.01.2020; accepted – 03.08.2020

# A Bayesian Small Area Model with Dirichlet Processes on the Responses

**Jiani Yin**[1], **Balgobin Nandram**[2]

## ABSTRACT

Typically survey data have responses with gaps, outliers and ties, and the distributions of the responses might be skewed. Usually, in small area estimation, predictive inference is done using a two-stage Bayesian model with normality at both levels (responses and area means). This is the Scott-Smith (S-S) model and it may not be robust against these features. Another model that can be used to provide a more robust structure is the two-stage Dirichlet process mixture (DPM) model, which has independent normal distributions on the responses and a single Dirichlet process on the area means. However, this model does not accommodate gaps, outliers and ties in the survey data directly. Because this DPM model has a normal distribution on the responses, it is unlikely to be realized in practice, and this is the problem we tackle in this paper. Therefore, we propose a two-stage non-parametric Bayesian model with several independent Dirichlet processes at the first stage that represents the data, thereby accommodating some of the difficulties with survey data and permitting a more robust predictive inference. This model has a Gaussian (normal) distribution on the area means, and so we call it the DPG model. Therefore, the DPM model and the DPG model are essentially the opposite of each other and they are both different from the S-S model. Among the three models, the DPG model gives us the best head-start to accommodate the features of the survey data. For Bayesian predictive inference, we need to integrate two data sets, one with the responses and other with area sizes. An application on body mass index, which is integrated with census data, and a simulation study are used to compare the three models (S-S, DPM, DPG); we show that the DPG model might be preferred.

**Key words:** Bayesian computation, bootstrap, predictive inference, robust modeling, computational and model diagnostics, survey data.

## 1. Introduction

There are many methods in the current statistical literature for making inferences based on samples selected from a finite population. The most widely used approach is design-based inference, which is nonparametric but requires large sample sizes. Model-based inference for survey sampling has been proposed as an alternative to the design-based theory, and this is particularly useful for small area estimation (Rao and Molina 2015) when there are sparse data from many areas. We consider the simplest version of a small area model, and we show how to robustify it to fit survey responses with gaps, outliers and ties.

[1]Takeda Pharmaceuticals. USA. E-mail: jianiyin@gmail.com. ORCID: https://orcid.org/0000-0002-5007-2833.
[2]Worcester Polytechnic Institute. USA. E-mail: balnan@wpi.edu. ORCID: https://orcid.org/0000-0002-3204-0301.


Figure 1: Dot plots of body mass index (BMI) for thirty-five areas (counties)

Generally, in a unit-level model, the responses from each area might have a distribution with a mean and a variance. The variance is usually taken constant over areas, but the mean varies over the areas. Sometimes each mean is written as global constant plus a random effect, different over areas. The area random effects or means share a common distribution allowing a borrowing of strength adaptively across areas (sample sizes are generally different). Complete pooling is generally a bad idea, because there is usually heterogeneity across areas. A degree of heterogeneity can be accommodated using covariates, but while useful covariates are particularly important in any analysis, this is not enough because there will

still be heterogeneity across areas. So there is a need to model area means, for example, to provide a small area model. Here, we consider continuous responses from a number of small areas using a unit-level model.

Our application is on body mass index (BMI), a continuous variable used to measure lifestyle. (Because of how survey data are collected, the BMI data can be discrete and there will be gaps, outliers and ties.) We use data from the 35 largest counties (areas) with at least 500,000 people from the third National Health and Nutrition Examination Survey (NHANES III), a survey conducted during the period October 1988 through September 1994. We do not have access to data from other smaller counties. In fact, these are the BMI data from the same 35 counties we analyzed in Nandram and Choi (2005, 2010) and in other places too numerous to mention. However, we have used data for adults who are older than 20 years because these data have very few nonresponse, rather than for children younger than 19 years, because our current study is not about nonresponse. We use BMI data where the BMI values are given up to the first decimal place. Dot plots of the data from 35 counties are shown in Figure 1. There are three things we observe in these data. First, there are ties because several adults have the same BMI values. This is clear because an adult BMI value is some value from about 18.0 $kgm^{-2}$ to about 40.0 $kgm^{-2}$ (one decimal place). Second, there are gaps (i.e. no BMI values between two adjacent values) and this is especially true in the right extreme areas of the dot plots. Third, there are outliers, which occur mostly in the right tails of the dot plots, thereby showing some right skewness with outliers. Therefore, it is clear that these BMI values do not follow normal distributions; a kernel density estimator will hide these features in the data. The data have natural gaps (e.g. there are no values in between 20.1 and 20.2) and ties (e.g. several values at say 20.1); these will exist in the population as well. This is why we model the gaps, outliers and ties in these data. We note that there are some demographic variables such as age, race and sex, which we do not study here, but we discuss in the concluding section how to incorporate covariates into our models.

Our goal is to predict the finite population mean, $85^{th}$ (overweight) and $95^{th}$ (obese) finite population percentiles of BMI for all eligible adults from each county. The sample from each area is at least about 100; we have a small area problem because these sample sizes are about a 0.01% of the population. Our problem is how to take care of the gaps, outliers and ties in the BMI data. To this end, we use two-stage Bayesian models with one model having a component that addresses directly these non-standard features in the responses. To do Bayesian predictive inference for the finite population quantities, we also need a data set with the population sizes of the areas (counties). To achieve this end, we integrate the NHANES BMI data with the population counts from the US 1990 Census.

Let $y_{ij}$ denote the value for the $j^{th}$ unit within the $i^{th}$ area, $i = 1,\ldots,\ell, j = 1,\ldots,N_i$. Throughout, we assume that $y_{ij}$, $i = 1,\ldots,\ell, j = 1,\ldots,n_i$, are the samples from $i^{th}$ area and are observed, and $y_{ij}, j = n_i+1,\ldots,N_i$ are not observed. Inference is required for the finite population mean or a finite population percentile. For example, the finite population mean of the $i^{th}$ area is $\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij}/N_i, i = 1,\ldots,\ell$. We use Bayesian predictive inference that requires specification of parametric distributions. Moreover, to help protect against posterior impropriety, we use non-informative (vague) independent priors, which are proper, for all hyper-parameters. Specifically, we have used Cauchy priors for location parameters

(e.g. Gelman, Jakulin, Pittau and Su 2008) and shrinkage priors for non-negative parameters (e.g. Nandram and Yin 2016 a, b and the references therein).

Scott and Smith (1969) introduced the basic two-stage model for cluster sampling, but the same model has been used for small areas. The difference is that in small area estimation, we are interested in inference about the population of each small area, but in cluster sampling, we are interested in all sub-populations combined into a single one. Nandram, Toto and Choi (2011) has given a Bayesian analysis of this model. However, the use of models raises the question of the robustness of the inference to possible model mis-specification. Again, in particular, survey data tend to have gaps, outliers and ties and we need to remedy this defect. A generalization to include covariates is the model of Battese, Harter and Fuller (1988) in non-Bayesian survey sampling; for a full Bayesian formulation, see Toto and Nandram (2010) and Molina, Nandram and Rao (2014); but this is not our key issue here.

The Bayesian Scott-Smith (S-S) model, as formulated by Nandram, Toto and Choi (2011), is

$$y_{ij}|\mu_i,\sigma^2 \overset{ind}{\sim} N\left(\mu_i,\sigma^2\right), \quad j=1,\ldots,N_i, \tag{1}$$

$$\mu_i \mid \theta,\sigma^2,\rho \overset{ind}{\sim} N\left(\theta,\frac{\rho}{1-\rho}\sigma^2\right), \quad i=1,\ldots,\ell, \tag{2}$$

$$\pi(\theta,\sigma^2,\rho) = \frac{1}{\pi(1+\theta^2)}\frac{1}{(1+\sigma^2)^2}, \tag{3}$$

where $-\infty<\theta<\infty$, $\sigma^2>0$, $0\le\rho\le 1$. Here, $\rho$ is the intra-cluster correlation. It is worth noting that we have taken $\rho\sim\text{Uniform}(0,1)$, $\theta$ to have a standard Cauchy distribution and $\sigma^2$ to have a shrinkage distribution (i.e. $f(2,2)$ distribution), all independent. Here, we have used vague proper priors on all parameters.

Suppose we have written $\mu_i \mid \theta,\delta^2 \overset{ind}{\sim} \text{Normal}(\theta,\delta^2)$ and define $\rho=\delta^2/(\delta^2+\sigma^2)$, then we will get $\delta^2=\frac{\rho}{1-\rho}\sigma^2$. Clearly, $0\le\rho\le 1$ and this makes one variance component bounded instead of two unbounded ones, $\sigma^2$ and $\delta^2$. This simplifies the computations by permitting a random sampler, which requires no monitoring, rather than a Gibbs sampler, which requires monitoring; see Appendix A.

Another standard model that relaxes some parametric assumptions is the Dirichlet process mixture (DPM) model,

$$y_{ij}|\mu_i,\sigma^2 \overset{ind}{\sim} \text{Normal}(\mu_i,\sigma^2), \quad j=1,\ldots,N_i, \tag{4}$$

$$\mu_i|G \sim G, \quad i=1,\ldots,\ell,$$

$$G \mid \theta,\sigma^2,\gamma,\rho \sim \text{DP}\left\{\gamma,\text{Normal}(\theta,\frac{\rho}{1-\rho}\sigma^2)\right\}, \tag{5}$$

$$\pi(\theta,\sigma^2,\gamma,\rho) = \frac{1}{\pi(1+\theta^2)}\frac{1}{(1+\sigma^2)^2}\frac{1}{(1+\gamma)^2}, \tag{6}$$

where $-\infty<\theta<\infty$, $\sigma^2>0,\gamma>0$, $0\le\rho\le 1$, and $\gamma$ is called the concentration parameter; see Ferguson (1973) for a definition of the Dirichlet process (DP) and Lo (1984), who extended the DP to DPM. Here, in this formulation the S-S model is a baseline model;

the DPM model is centred on the S-S model and $\gamma$ controls how close DPM model gets to the S-S model. Here, G is a random distribution function, discrete with probability one, and had distribution $DP(\cdot,\cdot)$. Escobar and West (1995) proposed a simple (not necessarily efficient) algorithm by integrating out the random distribution function in the model. Kalli, Griffin and Walker (2011) suggested slice-efficient samplers, an improved slice sampling scheme that we use in our work, and it is based on the stick-breaking construction of Sethuraman (1994); see Appendix B. Nandram and Choi (2004) and Polettini (2017) have applications on small area estimation, but they did not use the slice-efficient sampler of Kalli, Griffin and Walker (2011).

However, this DPM model does not address our main concern. It does not model the responses non-parametrically to take care of gaps, outliers and ties in the survey data in general, not just BMI data. It models ties among the $\mu_i$, thereby clustering the $\mu_i$. Indeed, this is the strength of the Dirichlet process prior. In reality, we want to do the opposite. That is, we want to have independent Dirichlet processes on the responses and possibly a normal distribution on the random effects. This is the key issue we address in this paper, and we will call this model the DPG model (G refers to the normal assumption on the $\mu_i$). However, the DPM model gives a good sense of how to proceed to meet our requirement.

The plan of the rest of the paper is as follows. In Section 2, we discuss the DPG model with independent Dirichlet processes on the responses. In Section 2.1, we discuss the methodology and inferences. In Section 2.2, we discuss the prediction for a finite population quantity using a data integration. In Section 3, we compare the three models (S-S, DPM, DPG). Specifically, in Section 3.1, we discuss an illustrative example on the body mass index (BMI) data and in Section 3.2 a small simulation study. In Section 4, we present our conclusion and two important extensions.

## 2. DPG Model, Computations and Prediction

In this section, we describe the DPG model that has independent Dirichlet processes on the responses and a normal distribution on the area means. This robustifies the S-S model in the opposite direction to the DPM, our novel contribution. In Section 2.1, we describe the DPG model, in Section 2.2, we describe how to draw samples from it, and in Section 2.3, we show how to do the prediction.

### 2.1. DPG Model

Using DPs in the first level and a parametric distribution as prior gives us,

$$\begin{aligned} y_{ij}|G_i &\overset{ind}{\sim} G_i, \quad j=1,\ldots,N_i, \\ G_i|\mu_i,\alpha_i,\sigma^2 &\overset{ind}{\sim} \mathrm{DP}\{\alpha_i,\mathrm{Normal}(\mu_i,\sigma^2)\}, \quad i=1,\ldots,\ell, \\ \mu_i|\rho i.e.\theta,\sigma^2,\rho &\overset{iid}{\sim} \mathrm{Normal}(\theta,\frac{\rho}{1-\rho}\sigma^2). \end{aligned} \tag{7}$$

A full Bayesian model can be obtained by adding prior distributions. We use proper non-informative priors,

$$\pi(\alpha_i) = \frac{1}{(\alpha_i+1)^2}, \qquad \alpha_i > 0, \quad i = 1,\ldots,\ell, \tag{8}$$

$$\pi(\theta,\sigma^2,\rho) = \frac{1}{\pi(1+\theta^2)}\frac{1}{(1+\sigma^2)^2}, \quad -\infty < \theta < \infty, 0 < \sigma^2 < \infty, 0 \le \rho \le 1, \tag{9}$$

with independence. Here, (7), (8) and (9) define the DPG model. Note that the concentration parameters $\alpha_i$ are not included in the S-S model or the DPM model. It is not sensible to assume that the $\alpha_i$ are identically distributed, because they can be very different.

We give a brief comparison of the three models and how they are related. The S-S model is a special case of the DPG model and the DPM model, and both are centred on the S-S model. This occurs when the $\alpha_i$ are large for the DPG model and when $\gamma$ is large for the DPM model. The DPM model is actually the opposite of the DPG model with normal distribution for the data in each area and a DP prior on the area means. In the DPG model, each area has a distinct DP (i.e. $\ell$ DPs with different $\mu_i$ and $\alpha_i$) and there is pooling across areas because the $\mu_i$ share an effect and $\sigma^2$ is common.

We look at the sampling process for the DPG model. When we integrate out the random probability measure (Blackwell and MacQueen, 1973), we get

$$f(\underset{\sim}{y}_i \mid \mu_i,\sigma^2,\alpha_i) = \frac{1}{\sigma}\phi(\frac{y_{i1}-\mu_i}{\sigma})$$

$$\times \prod_{k=2}^{n_i}\left\{\frac{k-1}{\alpha_i+k-1}\frac{\sum_{j=1}^{k-1}\delta_{y_{ij}}(y_{ik})}{k-1} + \frac{\alpha_i}{\alpha_i+k-1}\frac{1}{\sigma}\phi(\frac{y_{ik}-\mu_i}{\sigma})\right\}, \tag{10}$$

where $\delta_a(y)$ means that $y$ is a point mass at $a$ and $\phi(.)$ is the standard normal density. Therefore, in each area we are mixing the distributions in (10) using normal mixing distributions in the DPG model. The DPM is different being a Dirichlet process mixture of normals. The DPM model actually produces ties among the random effects (clustering), its major strength, but it does not model gaps, outliers, ties and possibly skewness among the responses. By putting DPs on the responses in different areas, we are actually taking a head-start on the data, because they accommodate the gaps, ties and outliers in the data; see Figure 1. It is important to note that $\delta_{y_{ij}}(y_{ik})$ is a statement that for each $i$, $y_{ik}$ is a point mass at $y_{ij}, j = 1,\ldots,k-1$. That is, for the $i^{th}$ area, $y_{ik}$ can be the same as $y_{ij}$ with nonzero probability and this is crucial in our new model. Therefore, equation (10) is the key to how we attempt to accommodate gaps, outliers and ties, particularly ties, in the data. The DPG model is attractive even if there are a few ties (or no ties at all) because the data may have heavy tails where the normal distribution is not appropriate (true for the BMI data).

### 2.2. Computations

Letting $\underset{\sim}{\psi} = \{\underset{\sim}{\mu}, \theta, \sigma^2, \rho\}$ and $\underset{\sim}{\alpha} = \{\alpha_1, \ldots, \alpha_\ell\}$, it is easy to get a sample from the joint posterior density of $(\underset{\sim}{\psi}, \underset{\sim}{\alpha})$, and therefore inference under the DPG model can be easily performed.

The posterior densities of the $\alpha_i$ are independent of the other parameters $\underset{\sim}{\psi}$ in the model, conditioning on only the distinct values. Let $k_i$ denote the number of distinct values for each area in the observed data, $\underset{\sim}{k} = \{k_i, i = 1, \ldots, \ell\}$ be the vector of $k_i$, $y^*_{i1}, \ldots, y^*_{ik_i}$ be the $k_i$ distinct sample values for each $i$ and $\underset{\sim}{y}^* = \{y^*_{i1}, \ldots, y^*_{ik_i}, i = 1, \ldots, \ell\}$ be the vector of $y^*_{ij}$. Thus, the joint posterior density is

$$\pi(\underset{\sim}{\alpha}, \underset{\sim}{\psi} \mid \underset{\sim}{k}, \underset{\sim}{y}^*) = \left[\prod_{i=1}^{\ell} \pi(\alpha_i \mid k_i)\right] \pi(\underset{\sim}{\psi} \mid \underset{\sim}{y}^*), \tag{11}$$

where $\pi(\alpha_i|k_i) \propto \pi(k_i \mid \alpha_i)\pi(\alpha_i)$. For the parameters $\underset{\sim}{\psi}$, we have

$$\begin{aligned} y^*_{ij}|\mu_i &\overset{ind}{\sim} N\left(\mu_i, \sigma^2\right), \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, k_i, \\ \mu_i &\overset{iid}{\sim} N\left(\theta, \frac{\rho}{1-\rho}\sigma^2\right), \\ \pi(\theta, \sigma^2, \rho) &= \frac{1}{\pi(1+\theta^2)} \frac{1}{(1+\sigma^2)^2}, -\infty < \theta < \infty, 0 < \sigma^2 < \infty, 0 \le \rho \le 1. \end{aligned} \tag{12}$$

Therefore, the algorithm for the DPG model is
Step 1 : For each $i$, $i = 1, \ldots, \ell$, draw $\alpha_i$ from $\pi(\alpha_i|k_i) \propto \alpha^{k_i} \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i+n_i)} \frac{1}{(\alpha_i+1)^2}$; see Antoniak (1974).
Step 2: Draw $\underset{\sim}{\psi}$ from the parametric model (12), which is easy to fit; see Appendix A for the S-S model.

Step 1 is easily realized using the grid method (Nandam and Yin 2016 a,b). Step 2 is accomplished using a random sampler together with the sampling importance resampling (SIR) algorithm. Therefore, samples can be drawn from the DPG model using a random sampler rather than a Gibbs sampler (as in the DPM, Markov chain samplers need monitoring).

### 2.3. Prediction for the Finite Population

We have a simple random sample of size $n_i$ from a finite population of size $N_i$, $i = 1, \ldots, \ell$. Let $y_{i1}, \ldots, y_{in_i}$ denote the sampled values. We want to predict $y_{in_i+1}, \ldots, y_{iN_i}$, the nonsampled values, and obtain the predictive distribution and the prediction interval for any finite population quantity (e.g. $\bar{Y}_i$ for the $i^{th}$ area). Prediction under the S-S model and the DPM model is straightforward.

For the DPG model, the sampling process is

$$\begin{aligned} y_{ij}|G_i &\overset{ind}{\sim} G_i, \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, N_i, \\ G_i|\mu_i &\overset{ind}{\sim} \mathrm{DP}\{\alpha_i, G_0(\mu_i)\}. \end{aligned}$$

Predictive inference for the DPG model simply uses the generalized Polya urn scheme (Blackwell and MacQueen 1973) for each $i$, since all areas are independent (see Nandram and Yin 2016 a,b). Once the nonsampled $y_{ij}, j = n_i + 1, \ldots, N_i, i = 1, \ldots, \ell$, are obtained, one can now calculate any finite population quantity of interest. Here, we are interested in the finite population mean, the $85^{th}$ percentile (overweight individuals) and the $95^{th}$ percentile (obese individuals). The $N_i$ are assumed known, and they can be obtained from a census.

Binder (1982), a very nice paper, frustrated with the bootstrap method (discussed later) that does not produce values different from the sample values, introduced the Dirichlet process into finite population sampling. We note that when prediction is done using any of the three models, including the DPG model, new values different from the samples will be generated. For the S-S model and the DPM model, this will happen with probability one, but for the DPG model with just a positive probability. For the DPG model, because the nonsample values are generated from the generalized Polya urn scheme, values already sampled can be repeated. However, for the DPG model, as the prediction proceeds in an order for a long run (population sizes are large here), the $\alpha_i$ will be dominated, thereby making the process draw more and more values that have already been drawn as in "the rich gets richer scheme".

Letting $f_i = \frac{n_i}{N_i}, i = 1, \ldots, \ell$, denote the sample fractions, the finite population mean is the composite, $\bar{Y}_i = f_i \bar{y}_{i,s} + (1 - f_i)\bar{Y}_{i,ns}$, where $\bar{y}_{i,s} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, the mean of the sample values, and $\bar{Y}_{i,ns} = \frac{1}{N_i - n_i}\sum_{j=n_i+1}^{N_i} y_{ij}$, the mean of the non-sample values. To obtain the percentiles, one simply sorts all the data (sample values and predicted non-sample values) in increasing order. Then, for the $85^{th}$ percentile, pick the value at $.85N_i$ (nearest integer) position, and for the $95^{th}$ percentile, pick the value at $.95N_i$ (nearest integer) position.

It is worth noting that it is easy to estimate the finite population mean; it is more difficult to estimate the two percentiles because they are in the right tail of the posterior distributions. It is interesting that in finite population mean, the sample mean, $\bar{y}_{i,s}$, is constant a posteriori but $\bar{y}_{i,ns}$ is dynamic (i.e. changes with the iterations). However, when the finite population percentiles are estimated, all the sample values and the predicted values are ordered at each iteration (i.e. the actual positions of the sample values in the ordering will change). Therefore, computation of the finite population percentiles at each iteration takes more time than the finite population mean.

## 3. Empirical Studies

In this section, we compare the three models (S-S model, DPM model and DPG model). Specifically, in Section 3.1, we describe an application on body mass index (BMI) data, and in Section 3.2, we present a small simulation study.

### 3.1. Application to Body Mass Index Data

As described in the introduction, we use the example on BMI data for illustration. Since the predictive inference for the overweight and obese population is very important, the heavy tail of the distribution cannot be ignored. Thus, we cannot automatically use the S-S model

nor the DPM model to accommodate the gaps, outliers and ties in the BMI data; see Figure 1. A more robust assumption on the responses, such as the DPG model, needs to be considered.

For the DPM, we ran $10,000$ MCMC iterations, used $5,000$ as a "burn in" and thinned every $5^{th}$ to obtain $1,000$ converged posterior samples. We have monitored the parameters, $\sigma^2$, $\theta$, $\delta^2$ and $\gamma$, for the DPM model. The Geweke test of stationarity gives p-values of .483, .414, .459, 0.620 respectively; therefore the iterates pass the test of stationarity. The effective sample sizes are 1000, 1000, 698, 1084 respectively, thereby showing that the iterates form an efficient sample. Numerical summaries such as trace plots and auto-correlation plots (not shown) indicate that the MCMC chains converge and mix well, and a 'random sample' is obtained from the joint posterior density. To get samples from the S-S model and the DPG model, we do not need a Gibbs sampler; a random sampler suffices and monitoring of a Gibbs sampler is not needed.

As a comparison, we also use the Bayesian bootstrap to do prediction in each county individually without borrowing across counties. This will allow us to see how much improvement we can have over direct estimation. Note that for each area (county), all sample sizes are over 100. Here, we describe the Bayesian bootstrap (see Rubin 1981 for more details). Momentarily we consider a single subscript (drop subscript $i$), so that we have $y_1,\ldots,y_n$ (sample values) from an area and we need to predict $y_{n+1},\ldots,y_N$ (nonsample values), where $N$ is the population size of this area. First, we find the distinct values among $y_1,\ldots,y_n$ and we assume that there are $d$ distinct values, denoted by $y_1^*,\ldots,y_d^*$. Let $n_j$ denote the number of times the $j^{th}$ value occurs in the sample. In the bootstrap it is assumed that only $y_1^*,\ldots,y_d^*$ can occur, and let $N_j$ denote the number of times the $j^{th}$ distinct value occurs in the population; the $N_j$ are unknown. The Bayesian bootstrap has the following model,

$$\underset{\sim}{n} \mid \underset{\sim}{p} \sim \text{Multinomial}(n, \underset{\sim}{p}), \underset{\sim}{p} \sim \text{Dirichlet}(\underset{\sim}{0}),$$

where the improper Haldane's prior is used. Then, the posterior density of $\underset{\sim}{p}$ is

$$\underset{\sim}{p} \mid \underset{\sim}{n} \sim \text{Dirichlet}(\underset{\sim}{n}),$$

which is proper. The Bayesian bootstrap has the following steps,

1. Sample $\underset{\sim}{p} \mid \underset{\sim}{n} \sim \text{Dirichlet}(\underset{\sim}{n})$;
2. Sample $(N_1 - n_1,\ldots,N_d - n_d) \mid \underset{\sim}{p}, \underset{\sim}{n} \sim \text{Multinomial}(N-n, \underset{\sim}{p})$;
3. Repeat (1) and (2) a large number of times.

We have repeated the bootstrap procedure 1000 times. At each repetition, for the nonsamples, $y_j^*$ occurs $N_j - n_j$ times, $j = 1,\ldots,d$; so we have got the entire population with $y_j^*$ occurring $N_j$ times with 1000 repetitions. It is worth noting that the Bayesian bootstrap is different from the DPG model (one-level DP model) when it is applied to an individual area because while the Bayesian bootstrap cannot produce new values, the DPG model can do so.

The $85^{th}$ and $95^{th}$ percentiles are also important and the methodology is essentially the same. We perform the predictive inference of the population mean, $85^{th}$ and $95^{th}$ percentiles

for each area using the three models (S-S, DPM, DPG). We have compared the DPG model to the S-S model, the DPM model and Bayesian bootstrap. We have computed summary statistics, posterior mean (PM), posterior standard deviation (PSD), and coefficient of variation (CV $= 100\times$ PSD/PM), as a measure of reliability.

We have looked at the five-number summaries (Min, Q1, Med, Q3, Max) of the shrinkage coefficients over the areas; for example, see Appendix A. For the S-S model and the DPM model, these are virtually the same $(.48,\ .55,\ .58,\ .60,\ .84)$ but there is only a small difference from the DPG model, which has $(.52,\ .59,\ .62,\ .64,\ .87)$. These numbers indicate that there is comparable and moderate pooling for all three models.

In Table 1, as a further measure of shrinkage, we have presented five-number summaries over the areas of $PB = (PM - B)/B$, where $B$ is the posterior mean from the Bayesian bootstrap method; recall that the bootstrap is a method to obtain the finite population quantities for each area separately (no pooling). We observe that for the finite population mean, the five-number summaries are virtually the same with 50% negative PBs and 50% positive PBs. The three models are almost the same for the finite population $85^{th}$ percentile; Min is negative for all three models but they are different. They differ for the finite population $95^{th}$ percentile; virtually all the PBs are positive under the DPG model, but 25% are positive under the S-S and DPM models. Therefore, there is some evidence that the DPG model is more responsive to the gaps, outliers and ties in the BMI data. The assumption of independent normal responses in the S-S and DPM models is overly restrictive, especially when we get out into the tails of the BMI data.

Table 1: Five-number summaries of $PB = (PM - B)/B$ of the finite population mean, $85^{th}$ percentile and $95^{th}$ percentile for BMI data by three models (S-S, DPM, DPG)

| | Mean | | | | | $85^{th}$ Percentile | | | | | $95^{th}$ Percentile | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Min | Q1 | Med | Q3 | Max | Min | Q1 | Med | Q3 | Max | Min | Q1 | Med | Q3 | Max |
| S-S | -0.02 | -0.01 | 0.00 | 0.01 | 0.02 | -0.05 | 0.00 | 0.01 | 0.02 | 0.05 | -0.11 | -0.04 | -0.01 | 0.00 | 0.04 |
| DPM | -0.02 | -0.01 | 0.00 | 0.01 | 0.02 | -0.05 | 0.00 | 0.01 | 0.03 | 0.05 | -0.11 | -0.04 | -0.01 | 0.00 | 0.04 |
| DPG | -0.02 | -0.01 | 0.00 | 0.01 | 0.02 | -0.02 | 0.01 | 0.02 | 0.03 | 0.05 | -0.05 | 0.00 | 0.01 | 0.02 | 0.04 |

NOTE: Min=Minimum; Q1= 1st quartile; Med=median; Q3=3rd quartile; Max= Maximum.

In Table 2, as a measure of reliability, we present the five-number summaries of the coefficient of variation (CV $= 100\times$PSD/PM) over the areas. Overall these are very good for all finite population quantities and models (including the bootstrap) although under the bootstrap these CVs should be a bit bigger because the bootstrap generally underestimates variability. For the finite population mean, the CVs from the three models are mostly similar and those under the S-S, DPM and DPG models are mostly smaller than the bootstrap. For the finite population $85^{th}$ percentile and the finite population $95^{th}$ percentile, the S-S model and DPM model are similar, but their CVs are mostly to the left of those of the bootstrap. However, the five-number summaries of DPG model for estimating the finite population $95^{th}$ percentile are to the right of those of the S-S and DPM models, but still to the left of the bootstrap. Nevertheless, all three models appear to show good reliability.

Table 2: Five-number summaries of coefficient of variation (CV = 100×PSD/PM), of the finite population mean, 85th percentile and 95th percentile for BMI data by three models (S-S, DPM , DPG) and Bayesian bootstrap (Boot)

| | Mean | | | | | 85th Percentile | | | | | 95th Percentile | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Min | Q1 | Med | Q3 | Max | Min | Q1 | Med | Q3 | Max | Min | Q1 | Med | Q3 | Max |
| S-S | 0.62 | 1.20 | 1.26 | 1.37 | 1.57 | 0.59 | 1.11 | 1.19 | 1.25 | 1.48 | 0.60 | 1.12 | 1.18 | 1.27 | 1.46 |
| DPM | 0.66 | 1.28 | 1.36 | 1.45 | 1.60 | 0.60 | 1.20 | 1.29 | 1.32 | 1.52 | 0.63 | 1.21 | 1.25 | 1.30 | 1.53 |
| DPG | 0.61 | 1.18 | 1.23 | 1.30 | 1.55 | 1.13 | 1.60 | 1.84 | 2.24 | 2.65 | 1.85 | 2.19 | 2.44 | 2.54 | 3.97 |
| Boot | 0.61 | 1.34 | 1.49 | 1.62 | 1.97 | 1.19 | 2.04 | 2.67 | 3.04 | 4.49 | 2.17 | 2.99 | 3.58 | 4.10 | 7.39 |

NOTE: Min=Minimum; Q1= 1st quartile; Med=median; Q3=3rd quartile; Max= Maximum, Boot=Bootsrap.

We have looked at plots (not shown) of the posterior densities of the finite population mean, $85^{th}$ and $95^{th}$ percentiles for the three models (S-S, DPM and DPG) and Bayesian bootstrap for the 35 areas of BMI data. For the population mean, most parts of the density under the S-S, DPM and DPG models are similar, the DPG model has slightly smaller variation. Plots of the estimated densities of the population $85^{th}$ and $95^{th}$ percentiles under the DPG model are not smooth and the estimated densities of the population $85^{th}$ and $95^{th}$ percentiles under the S-S and DPM models are similar. Because the BMI data have some gaps, ties and outliers in the right tails, the estimations given by parametric models may be incorrect. Thus, based on a belief that the parametric model is too restrictive, we prefer the analysis based on the nonparametric DPG model.

Finally, we compare predictive inference of the finite population mean, $85^{th}$ and $95^{th}$ percentile for each area by the three models (S-S, DPM and DPG). We use three plots (not shown), which contain posterior means with credible bands versus direct estimates for BMI data. The posterior means are very similar under the S-S, DPM and DPG models and the predictive inferences of the population percentile are similar under the S-S and DPM models. For the finite population mean, the points (plot not shown) are all roughly on a straight line crossing the 45-degree straight line with slightly smaller slope, as it is should be. For the $85^{th}$ percentile, the points (plot not shown) are a little bit more spread out. However, as expected, the DPG model tends to have higher predictions (closer to the 45-degree straight line) of the population percentiles with similar credible bands when it is compared to the other two models. We suspect that S-S and DPM model might underestimate the $85^{th}$ and $95^{th}$ population percentiles when the data are right skewed. Without the restrictive parametric assumptions, the DPG model tends to provide less biased estimation with similar variation comparing to the other candidate models, thereby showing a distinct advantage of the DPG model; see Figure 2 for the finite population $95^{th}$ percentile. We investigate this issue in a small simulation study.

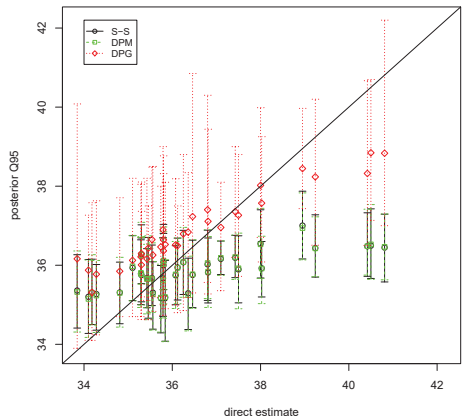


Figure 2: Comparison for body mass index (BMI) data (posterior means with credible bands versus direct estimates): the predictive inference of the finite population $95^{th}$ percentile for each county under the three models (S-S, DPM, DPG)

### 3.2. Simulation Study

We conduct a small simulation study. We choose $\ell = 50$ and the sample sizes, $n_i$, for 50 areas. The sample sizes are 35 for each of the first 10 areas, 50 for each of the second 10 areas, 100 for each of the third 10 areas, 200 for each of the fourth 10 areas and 500 for each of the last 10 areas. Then, the population sizes are selected as $N_i = 100n_i, i = 1, \ldots, \ell$. These are comparable to the BMI data. For convenience, we have taken $\theta = 0.0$, $\sigma^2 = 0.01$, $\delta^2 = 0.04$, thereby making $\rho = 0.8$. For the concentration parameters of the Dirichlet processes, we have selected $\gamma = 0.5$, and $\alpha_i \overset{ind}{\sim} 0.5 + \text{Beta}(5,5), i = 1, \ldots, \ell$. These choices allow us to have data similar to the BMI data with some flexibility to get gaps, outliers and ties when data are simulated from the DPG model.

We have simulated the entire finite population separately under the three models. This is done the same way under each model separately. For example, under the S-S model, because we have set $\theta$, $\sigma^2$ and $\rho$, we have generated $\mu_1, \ldots, \mu_\ell$ from (2) and for the $i^{th}$ area, we have generated $y_{ij}, j = 1, \ldots, N_i$ from (1). Therefore, we have all three finite population quantities. Given the parameters, because the observations are independent and identically distributed within each area, we simply take the first $n_i$ values as the sample. In the case of the DPG model, the population values are exchangeable and so we still take the first $n_i$ values as our sample.

When data are generated from the S-S model and the DPM model, there could be gaps and outliers in different areas. We note, in particular, there will be no ties because two data values cannot be the same (this happens with probability zero). Of course, the data from distinct areas will show some differences. However, as we have explained in this paper, when data are generated from the DPG model, there will be gaps, outliers and ties because

the data values are generated via the Polya urn scheme. By the nature of a DP, two values can be the same with nonzero probability; so there can be ties.

We fit the three models to the simulated data in exactly the same manner as for the NHANES-III BMI data. When the DPM model was fit, the Geweke tests show stationarity and the effective sample sizes are comparable to 1000. We have looked at plots (not shown) of the posterior means with 95% credible bands and true population means for the simulated S-S, DPM and DPG data under three different models (S-S, DPM and DPG models). The values are close to the true population mean. We also use absolute bias (AB) and posterior root mean squared error (PRMSE) to compare the models. We know the true value of the finite population quantities, denoted by $T$. Then, $AB = \mid PM - T \mid$ and $PRMSE = \sqrt{(PM-T)^2 + PSD^2}$. We compute these quantities for each of the fifty counties for the finite population mean and the $85^{th}$ and $95^{th}$ finite population percentiles, and respectively we average them. We present AB and PRMSE in Table 3; note that the entries in the table must be divided by $10,000$.

First, consider the finite population mean in Table 3 (a). We observe that AB is always too large when the DPM model is fit to any of the three simulated data sets. The S-S model and DPG model show comparable AB, much smaller than those for DPM. The PRMSEs under the DPG model are larger than those from S-S model and the DPM model (first two rows) by about 7% (marginal) but they are not larger than that of the DPM model when data are generated from the S-S model (0.01272 vs. 0.01008). The DPG is almost always better when data are generated from it; there is only a minor difference for AB under the S-S model and the DPG model (0.0001409 vs. 0.0001484).

Second, consider the finite population $85^{th}$ percentile in Table 3 (b). When data are generated from the S-S model, the S-S model and the DPG model are comparable and better than the DPM model. When data are generated from the DPM model, the three models are comparable with the PRMSE under the DPG model slightly higher than the other two models. When data are generated from the DPG model, the DPG model is a clear winner by far.

Third, consider the finite population $95^{th}$ percentile in Table 3 (c). When data are generated from the S-S model, the S-S model and the DPG model are comparable and better than the DPM model. When data are generated from the DPM model, the three models are more comparable. When data are generated from the DPG model, the DPG model is enormously better than the S-S model and DPM model.

When data are generated from the DPG model, in terms of AB and PRMSE, it performs much better than the S-S model and the DPM model for all three finite population quantities. This is strong evidence that when there are gaps, outliers and ties, the DPG model is the best. It is risky to use the S-S model or the DPM model for such data. The DPG model does not have to do better for data that are generated from the S-S model or the DPM model. By drawing a dot plot, one can see clearly which model is appropriate; data generated from the DPG model will have gaps, outliers and ties. Therefore, it is safe to conclude that the DPG model will perform better for data like the BMI data; see Figure 1.

Table 3: Comparison of absolute bias (AB) and posterior root mean squared error (PRMSE) of the finite population mean, $85^{th}$ percentile and $95^{th}$ percentile for each simulated data by three models (S-S, DPM and DPG) averaged over areas

| | (a) Mean | | | | | |
|---|---|---|---|---|---|---|
| | S-S | | DPM | | DPG | |
| Data | AB | PRMSE | AB | PRMSE | AB | PRMSE |
| S-S | 6.172 | 94.21 | 87.32 | 127.2 | 6.536 | 100.8 |
| DPM | 6.169 | 93.82 | 43.27 | 85.44 | 6.32 | 100.70 |
| DPG | 1.409 | 42.88 | 28.95 | 64.55 | 1.484 | 27.64 |
| | (b) $85^{th}$ Percentile | | | | | |
| | S-S | | DPM | | DPG | |
| Data | AB | PRMSE | AB | PRMSE | AB | PRMSE |
| S-S | 70.21 | 130.7 | 111.2 | 155.1 | 77.39 | 137.4 |
| DPM | 69.93 | 133.9 | 75.49 | 123.4 | 78.38 | 141.3 |
| DPG | 379.0 | 385.9 | 384.1 | 394.6 | 18.05 | 40.0 |
| | (c) $95^{th}$ Percentile | | | | | |
| | S-S | | DPM | | DPG | |
| Data | AB | PRMSE | AB | PRMSE | AB | PRMSE |
| S-S | 120.6 | 182.1 | 150.3 | 203.5 | 133.8 | 188.4 |
| DPM | 104.0 | 168.9 | 114.9 | 166.0 | 118.9 | 176.1 |
| DPG | 550.2 | 556.3 | 555.3 | 563.7 | 35.5 | 101.0 |

NOTE: Each row gives a model that generates the data and each column gives a model that is fit to the simulated data. The same three data sets are used in (a), (b) and (c). [The numbers in the table must be multiplied by $10^{-4}$.]

## 4. Concluding Remarks and Future Work

If the parametric distribution assumption does not hold, the model is mis-specified and the inference may be invalid. The Bayesian nonparametric methods are motivated by the desire to avoid overly restrictive assumptions. We believe that our DPG model, which has independent Dirichlet processes on the responses and a normal distribution on the area means, can accommodate survey responses with gaps, outliers and ties reasonably well.

Our illustration using the BMI data in our novel DPG model is a step forward. Our simulation shows the advantage of the DPG model when the finite population mean and the $85^{th}$ and $95^{th}$ finite population percentiles are being estimated. In the illustrative example on BMI data, it is interesting that Bayesian predictive inference can be performed using a data integration because the area sizes are available from the 1990 census. In future, we can adjust the DPG model to include a DP prior on the area means, rather than a normal distribution (Nandram and Yin 2019).

For future work, we may also include covariates in the DPG model in a manner in which Battese, Harter and Fuller (1988) actually extended the model of Scott and Smith (1969) to include covariates. The two-stage nonparametric alternative of the DPG model

with $p$ covariates and an intercept, $\mathbf{x_{ij}} = (\mathbf{1}, \mathbf{x_{ij}}'^{(\mathbf{0})})'$, is

$$\begin{aligned}
y_{ij} - \mathbf{x_{ij}}'^{(\mathbf{0})}\beta^{(\mathbf{0})}|\mathbf{G_i} &\overset{ind}{\sim} G_i, \quad i = 1,\ldots,\ell, \quad j = 1,\ldots,N_i, \\
G_i|\beta_{0i} &\overset{ind}{\sim} \mathrm{DP}\left\{\alpha_i, \mathrm{Normal}(\beta_{0i}, \sigma^2)\right\}, \\
\beta_{0i} \quad | \ \theta, \sigma^2, \rho &\overset{ind}{\sim} \mathrm{Normal}(\theta, \frac{\rho}{1-\rho}\sigma^2), \\
\pi(\alpha_i) &= \frac{1}{(\alpha_i+1)^2}, \qquad \alpha_i > 0, \quad i = 1,\ldots,\ell, \\
\pi(\beta^{(\mathbf{0})}, \theta, \sigma^{\mathbf{2}}, \rho) &\propto \frac{1}{1+\theta^2}\frac{1}{(1+\sigma^2)^2},
\end{aligned}$$

$-\infty < \theta, \beta_s^{(0)} < \infty, s = 1,\ldots,p, 0 < \sigma^2 < \infty, 0 \leq \rho \leq 1$, where $\rho$ is the intra-cluster correlation, $\mathbf{x_{ij}}^{(\mathbf{0})}$ and $\beta^{(\mathbf{0})}$ denote $\mathbf{x_{ij}}$ and $\beta$ without the intercepts. Note also that a priori the $\alpha_i$ are independent and there is a flat prior on $\beta^{(\mathbf{0})}$. This is how we can incorporate demographic variables (age, race and sex) for the BMI data from NHANES III.

In many complex surveys, there are also survey weights; this is also true for NHANES III. We may include the survey weights in the model using a normalized composite likelihood. However, if the survey weights for the nonsampled values are unknown, it is not obvious how to perform predictive inference under the model. One solution may be to use surrogate sampling (Nandram 2007).

## Acknowledgements

This research was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram). The authors thank Professor Wlodzimierz Okrasa for his invitation and encouragement.

## REFERENCES

ANTONIAK, C. E., (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2 (6), pp. 1152–1174.

BATTESE, G. E., HARTER, R. M. and FULLER, W. A., (1988). An Error-components Model for Prediction of County Crop Areas using Survey and Satellite Data. *Journal of the American Statistical Association*, 83 (401), pp. 28–36.

BINDER, D. A., (1982). Non-parametric Bayesian Models for Samples from Finite Populations. *Journal of the Royal Statistical Society*, Series B, 44 (3), pp. 388–393.

BLACKWELL, D., MACQUEEN, J. B., (1973). Ferguson Distributions via Polya Urn Schemes. *The Annals of Statistics*, 1 (2), pp. 353–355.

ESCOBAR, M. D., WEST, M., (1995). Bayesian Density Estimation and Inference using Mixtures. *Journal of the American Statistical Association*, 90 (430), pp. 577–588.

FERGUSON, T. S., (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1 (2), pp. 209–230.

GELMAN, A., JAKULIN, A., PITTAU, M. P. and SU, Y-S., (2008). A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. *Annals of Applied Statistics*, 2 (4), pp. 1360–1383.

KALLI, M., GRIFFIN, J. E. and WALKER, S. G., (2011). Slice Sampling Mixture Models. *Statistics and Computing*, 21 (1), pp. 83–105.

LO, A. Y., (1984). On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics*, 12 (1), pp. 351–357.

MOLINA, I., NANDRAM, B. and RAO, J. N. K., (2014). Small Area Estimation of General Parameters with Application to Poverty Indicators: A Hierarchical Bayes Approach. *The Annals of Applied Statistics*, 8 (2), pp. 852–885.

NANDRAM, B., CHOI, J. W., (2010). Bayesian Analysis of Body Mass Index Data from Small Domains Under Nonignorable Nonresponse and Selection. *Journal of the American Statistical Association*, 105 (489), pp. 120–135.

NANDRAM, B., CHOI, J. W., (2005). Hierarchical Bayesian Nonignorable Nonresponse Regression Models for Small Areas: An Application to the NHANES Data. *Survey Methodology*, 31 (1), pp. 73–84.

NANDRAM, B., CHOI, J. W., (2004). Nonparametric Bayesian Analysis of a Proportion for a Small Area under Nonignorable Nonresponse. *Journal of Nonparametric Statistics*, 16 (6), pp. 821–839.

NANDRAM, B., (2007). Bayesian Predictive Inference under Informative Sampling via Surrogate Samples. *In Bayesian Statistics and Its Applications*, edited by S.K. Upadhyay, U.Singh and D.K. Dey, Anamaya New Delhi, pp. 356–374.

NANDRAM, B., TOTO, M. C. S. and CHOI, J. W., (2011). A Bayesian Benchmarking of the Scott-Smith Model for Small Areas. *Journal of Statistical Computation and Simulation*, 81 (11), pp. 1593–1608.

NANDRAM, B., YIN, J., (2016a). Bayesian Predictive Inference under a Dirichlet Process with Sensitivity to the Normal Baseline. *Statistical Methodology*, 28, pp. 1–17.

NANDRAM, B., YIN, J., (2016b). A Nonparametric Bayesian Prediction Interval for a Finite Population Mean. *Journal of Statistical Computation and Simulation*, 86 (16), pp. 3141–3157.

NANDRAM, B., YIN, J., (2019). Hierarchical Bayesian Models for Small Areas With Dirichlet Processes. *JSM Proceedings, pp. 2594–2613, Survey Research Methods Section. Alexandria, VA: American Statistical Association*.

POLETTINI, S., (2017). A Generalized Semiparametric Bayesian Fay-Herriot Model for Small Area Estimation Shrinking Both Means and Variances. *Bayesian Analysis*, 12 (3), pp. 729–752.

RAO, J. N. K., MOLINA, I., (2015). *Small Area Estimation*, John Wiley & Sons, NY.

RUBIN, D. B., (1981). The Bayesian Bootstrap. *The Annals of Statistics*, 9 (1), pp. 130–134.

SETHURAMAN, J., (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4, pp. 639–650.

SCOTT, A., SMITH, T. M. F., (1969). Estimation in Multi-Stage Surveys. *Journal of the American Statistical Association*, 64 (327), pp. 830–840.

TOTO, M. C. S., NANDRAM, B., (2010). A Bayesian Predictive Inference for Small Area Means Incorporating Covariates and Sampling Weights. *Journal of Statistical Planning and Inference*, 140, pp. 2963–2979.

# APPENDICES

## A: Fitting the S-S Model

Let $\underset{\sim}{y} = (\underset{\sim}{y_s}, \underset{\sim}{y_{ns}})$, where $\underset{\sim}{y_s} = \{y_{ij}, i = 1, \ldots, \ell, j = 1, \ldots, n_i\}$ is the vector of observed values and $\underset{\sim}{y_{ns}} = \{y_{ij}, i = 1, \ldots, \ell, j = n_i + 1, \ldots, N_i\}$ vector of unobserved values. First, define the sample means and sample variances, $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ and $s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2, n_i > 1, i = 1, \ldots, \ell$. Second, let $\lambda_i = \frac{n_i}{n_i + (1-\rho)/\rho}, i = 1, \ldots, \ell$, $\tilde{y} = \sum_{i=1}^{\ell} \lambda_i \bar{y}_i / \sum_{i=1}^{\ell} \lambda_i$, and $A_1 = \frac{1-\rho}{\rho}\sum_{i=1}^{\ell} \lambda_i (\tilde{y} - \bar{y}_i)^2 + \sum_{i=1}^{\ell}(n_i - 1)s_i^2$. Here, the $\lambda_i$ are *shrinkage* coefficients.
Then, using Bayes' theorem, the joint posterior density of $\underset{\sim}{\mu}, \theta, \sigma^2, \rho$ is

$$\begin{aligned}
\pi(\underset{\sim}{\mu}, \theta, \sigma^2, \rho | \underset{\sim}{y_s}) \quad \propto \quad & \left(\frac{1}{\sigma^2}\right)^{(n+\ell)/2} \left(\frac{1-\rho}{\rho}\right)^{\ell/2} \exp\Bigg\{ -\frac{1}{2\sigma^2}\Bigg\{ \sum_{i=1}^{\ell} \Bigg\{ (n_i - 1)s_i^2 \\
+ \quad & \left(n_i + \frac{1-\rho}{\rho}\right)(\mu_i - [\lambda_i \bar{y}_i + (1-\lambda_i)\theta])^2 \\
+ \quad & \lambda_i \left(\frac{1-\rho}{\rho}\right)(\bar{y}_i - \theta)^2 \Bigg\}\Bigg\}\Bigg\} \times \frac{1}{(1+\sigma^2)^2} \times \frac{1}{\pi(1+\theta^2)}. \qquad \text{(A.1)}
\end{aligned}$$

We use a simple method called the sampling importance resampling (SIR) algorithm to draw from the posterior distribution $\pi(\underset{\sim}{\mu}, \theta, \sigma^2, \rho | \underset{\sim}{y_s})$ in (A.1). That is, we take a sample of draws from a proposal density $\pi_a(\underset{\sim}{\mu}, \theta, \sigma^2, \rho | \underset{\sim}{y_s})$, then use these draws to produce a sample from $\pi(\underset{\sim}{\mu}, \theta, \sigma^2, \rho | \underset{\sim}{y_s})$. As a well-known result, one would need $\pi(\underset{\sim}{\mu}, \theta, \sigma^2, \rho | \underset{\sim}{y_s}) / \pi_a(\underset{\sim}{\mu}, \theta, \sigma^2, \rho | \underset{\sim}{y_s})$ to be uniformly bounded in its parameters. A reasonable approximation to the joint posterior density (A.1) and one from which it is easy to draw samples will suffice. We use the same likelihoods (1) and (2) in the two-level normal model together with an improper prior $\pi(\theta, \sigma^2, \rho) \propto \frac{1}{\sigma^2}, -\infty < \theta < \infty, 0 < \sigma^2 < \infty, 0 \le \rho \le 1$ as a Bayesian model from which we use the posterior density as a proposal density,

$$\begin{aligned}
\pi_a(\underset{\sim}{\mu}, \theta, \sigma^2, \rho | \underset{\sim}{y_s}) \quad \propto \quad & \pi_a(\underset{\sim}{\mu} | \theta, \sigma^2, \rho, \underset{\sim}{y_s}) \pi_a(\theta | \sigma^2, \rho, \underset{\sim}{y_s}) \pi_a(\sigma^2 | \rho, \underset{\sim}{y_s}) \pi_a(\rho | \underset{\sim}{y_s}) \quad \text{(A.2)} \\
\propto \quad & \prod_{i=1}^{\ell} N\left[\mu_i; \lambda_i \bar{y}_i + (1-\lambda_i)\theta, (1-\lambda_i)\frac{\rho}{1-\rho}\sigma^2\right] \\
\times \quad & N\left(\theta; \tilde{y}, \frac{\sigma^2 \rho}{\sum_{i=1}^{\ell} \lambda_i (1-\rho)}\right) \times \mathrm{IG}\left[\sigma^2; (n-1)/2, A_1/2\right] \\
\times \quad & \frac{\Gamma[(n-1)/2]}{(A_1/2)^{(n-1)/2}} \prod_{i=1}^{\ell} (1-\lambda_i)^{1/2} \left[\frac{\rho}{\sum_{i=1}^{\ell} \lambda_i (1-\rho)}\right]^{1/2}.
\end{aligned}$$

Note that $\pi(\underset{\sim}{\mu}, \theta, \sigma^2, \rho | \underset{\sim}{y_s}) / \pi_a(\underset{\sim}{\mu}, \theta, \sigma^2, \rho | \underset{\sim}{y_s}) = \frac{1}{\pi(1+\theta^2)} \frac{\sigma^2}{(1+\sigma^2)^2} \le \frac{1}{\pi}$ (uniformly bounded as required). We draw a sample from the approximate joint posterior density (A.2) by first drawing a sample from $\pi_a(\rho | \underset{\sim}{y_s})$ using the grid method and continue using the multiplication rule of probability. The algorithm works fine because the sub-sampling weights are nearly uniform.

**B: Fitting the DPM Model**

Kalli, Griffin and Walker (2011) suggested slice-efficient samplers, and it is based on the stick-breaking algorithm (Sethuraman 1994). Letting $G = \sum_{s=1}^{\infty} \pi_s \delta_{\mu_s^*}$, where

$$\pi_1 = \beta_1, \quad \pi_s = \beta_s \prod_{j=1}^{s-1}(1-\beta_j), \quad \beta_s \overset{iid}{\sim} \text{Beta}(1,\gamma), \quad \mu_s^* \overset{iid}{\sim} G_0,$$

and $G_0$ is a baseline distribution. Here, for convenience, we will use a short-hand notation for the formulas below, $h(y_{ij};\mu_i) = \text{Normal}_{y_{ij}}(\mu_i,\sigma^2), j = 1,\ldots,n_i, i = 1,\ldots,\ell$ and so $h(\underset{\sim}{y_i};\mu_i) = \prod_{j=1}^{n_i}\{\text{Normal}_{y_{ij}}(\mu_i,\sigma^2)\}, i = 1,\ldots,\ell$. Also, we use $g(\mu_i) = \text{Normal}_{\mu_i}(\theta, \frac{\rho}{1-\rho}\sigma^2)$, $i = 1,\ldots,\ell$.

The idea is to introduce latent variables $\{u_1,u_2,\ldots,u_\ell\}$, which allows us to sample a finite number of variables at each iteration. One can introduce further latent variables, $\{d_1,d_2,\ldots,d_\ell\}$ that indicate the components of the mixture from which observations are to be taken to give a general class of slice samplers,

$$f(\underset{\sim}{y_i},u_i,d_i|\pi,\mu^*) = \mathbf{1}(u_i < \xi_{d_i})\pi_{d_i}/\xi_{d_i}h(\underset{\sim}{y_i};\mu_{d_i}^*),$$

where $\xi_1,\xi_2,\ldots$ is any positive sequence. Typically, the sequence will be deterministic decreasing sequence. In our computation, we use $\xi_s = (1-\kappa)\kappa^{s-1}$ where the tuning constant $\kappa$ is between 0 and 1; other choices are possible. Let $K = \max_{i=1}^{\ell}(K_i)$, where $K_i$ is the largest integer $t$ such that $\xi_t > u_i$.

Specifically, for our DPM model, the joint posterior distribution is proportional to

$$\pi(\theta,\sigma^2,\rho,\gamma)\prod_{s=1}^{K}\text{Beta}(\beta_s;1,\gamma)g_0(\mu_s^*)\prod_{i=1}^{\ell}\mathbf{1}(u_i < \xi_{d_i})\pi_{d_i}/\xi_{d_i}h(\underset{\sim}{y_i};\mu_{d_i}^*).$$

The variables $\{(\mu_s^*,\beta_s), s = 1,2,\ldots,K; (d_i,u_i), i = 1,\ldots,\ell\}$ need to be sampled at each iteration. The Gibbs sampler is obtained by drawing samples, each in turn, from the conditional posterior distributions, (a) $\pi(u_i|\ldots) \propto \mathbf{1}(0 < u_i < \xi_{d_i})$; (b) $\pi(\mu_s^*|\ldots) \propto g_0(\mu_s^*)\prod_{\{i|d_i=s\}} h(\underset{\sim}{y_i};\mu_s^*)$; (c) $\pi(\beta_s|\ldots) \propto \text{Beta}(a_s,b_s)$, where $a_s = 1+\sum_{i=1}^{\ell}\mathbf{1}(d_i = s)$ and $b_s = \gamma+\sum_{i=1}^{\ell}\mathbf{1}(d_i > s)$; (d) $P(d_i = r|\ldots) \propto \mathbf{1}(r : \xi_r > u_i)\pi_r/\xi_r h(\underset{\sim}{y_i};\mu_r^*)$, $r = 1,\ldots,K$. The other parameters are included in the Gibbs sampler, and the grid method is used to draw some of them (e.g. $\gamma$).