Hierarchical Bayesian Models for Small Areas With Dirichlet Processes

Balgobin Nandram* Jiani Yin†

Abstract

We consider a two-stage unit-level model for small areas with continuous survey responses. Typically survey data have responses with outliers, gaps and ties, and the distributions of the responses might be skewed. Therefore, predictive inference using a hierarchical Bayesian model with normality at both levels (responses and random effects) might not be robust against these features. So we provide a two-level non-parametric Bayesian model with a Dirichlet process at each stage, thereby permitting a more robust predictive inference. We show how to fit the four versions of this model (e.g., one version has Dirichlet processes at both levels) using Markov chain Monte Carlo methods. An application on body mass index and a simulation study are discussed to compare the four models. While it is difficult to tell which model is preferred, one might want to use the model with Dirichlet processes at both levels (it robustifies both levels against non-normality).

Key Words: Bayesian computation, Booststrap, Diagnostics, Predictive inference, Robust model, Survey data.

1. Introduction

There are many methods in 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 population has been proposed as an alternative to the design-based theory. Survey data, structured hierarchically, are quite common. For example, students are in classes, classes are in schools, schools are in counties and counties are in states. Hierarchical models are often applicable to modeling data from complex surveys such as multistage cluster sampling, because usually such sample designs are used when the population has a hierarchical structure.

In many surveys, we want to estimate quantities not only for the population as a whole, but also for subpopulations (e.g., to estimate the average income for every county in the United States in order to allocate funds for needed areas). Once a hierarchical model is specified, inference can be drawn from available data for the population quantities at any level. From a Bayesian perspective, these estimators which can be regarded as posterior means often have better properties than area-specific direct estimators. This makes hierarchical Bayesian models useful in the problem of small area estimation (SAE) (e.g., Rao and Molina 2015). That is, the sample size for a given area or domain may be too small to provide reliable estimates and it may be needed to borrow information from neighboring areas, or from areas with similar characteristics.

Hierarchical Bayesian methods, studied in the literature, have been mostly parametric, based on specified parametric likelihoods with conjugate or non-conjugate parametric priors. The normal likelihood is the most popular choice; see Scott and Smith (1969), Malec and Sedransk (1985), Battese, Harter and Fuller (1988) and Nandram, Toto and Choi (2011). The use of models raises the question about the robustness of the inference to possible model misspecification. Particularly, survey data tend to have gaps, ties and outliers.

*Worcester Polytechnic Institute, 100 Institute Road,Worcester, MA 01609

†Alkermes, Inc., 852 Winter Street, Waltham, MA 02451

There are extensive researches to relax the assumption of normality. One way to do this is to use heavy-tailed distributions (e.g., t distribution rather than a normal distribution to account for outliers (e.g., Lange, Little and Talyor 1989), and skew-normal distribution for heavy-skewed data (e.g., Azzalini 2013). Alternatively, the use of a mixture of normal distributions takes into account the presence of subgroups or multimodal data (e.g., Verbeke and Lesaffre 1996).

However, we often know very little about the specific parametric forms of the distributions, and it is also difficult to validate the parametric assumptions. The parametric Bayesian models based on distributional assumptions may be problematic because inferences are sensitive to such assumptions. It may be more appealing to use a nonparametric Bayesian approach.

In this paper, we discuss statistical modeling associated with the analysis of two-level survey data. Our intention is to propose nonparametric Bayesian alternatives using the Dirichlet process (DP) to robustify the inference by embedding parametric models in nonparametric models, thereby reducing critical dependence on parametric assumptions and to allow for heterogeneity, outliers, skewness, etc. These are expansion models, and a baseline parametric model is expanded using Dirichlet processes. The existence of the DP was established by Ferguson (1973). It is a distribution over distributions; each draw from a DP itself is a distribution (i.e., functional spaces). The DP has gained a lot of attention recently. It has nice properties such as clustering and borrowing information with reduced shrinkage, which is attractive to SAE.

In Section 2, we briefly review the Dirichlet process (DP), the Dirichlet process mixture (DPM) model and the hierarchical Dirichlet process mixture (HDPM) model. In Section 3, we discuss the two-stage Dirichlet process models as expansion models of a baseline model. We also discuss posterior propriety and prediction for finite population quantities. In Section 4, we discuss an illustrative example on body mass index (BMI) data and a small simulation study. Section 5 has concluding remarks and we provide some extensions to accommodate more stages, covariates and survey weights.

2. A Review of Dirichlet Process and Dirichlet Process Mixture Models

In Section 2.1, we present a review of the Dirichlet process (DP) prior, in Section 2.2, we present a review the Dirichlet process mixture (DPM) model, and in Section 2.3, we briefly review the hierarchical Dirichlet process mixture (HDPM) model and discuss why it might not be adequate for survey sampling.

2.1 Dirichlet Process Model

Let $(\Theta, \mathcal{B})$ be a measurable space, with G_0 a baseline measure (nonrandom) on the space, and let α be a positive real number. A Dirichlet process, DP(α, G_0), is defined as the distribution of a random probability measure G over $(\Theta, \mathcal{B})$ such that, for any finite measurable partition of the measurable space Θ, $\{A_i\}_{i=1}^n$,

$$\{G(A_1), \ldots, G(A_n)\} \sim \text{Dirichlet}\left\{\alpha G_0(A_1), \ldots, \alpha G_0(A_n)\right\}.$$

We write $G \sim \text{DP}(\alpha, G_0)$, if G is a random probability measure with a distribution given by the DP, where α is the concentration parameter. For an measurable set, A, we have $\text{E}[G(A)] = G_0(A)$, that is the mean of the DP is the baseline distribution G_0 and $\text{Var}[G(A)] = G_0(A)[1 - G_0(A)]/(\alpha + 1)$. The larger α is, the smaller the variance (i.e., the DP concentrates more of its mass around the baseline distribution). Here G_0 and α

are both parameters and they play intuitive roles in the definition of the DP. Here G is constrained to be around G_o and this is regulated by α.

Let $G \sim \text{DP}(\alpha, G_0)$ and $\theta_1, \ldots, \theta_n$ be a sequence of independent draws from G. The posterior distribution, $G|\theta_1, \ldots, \theta_n$, is

$$\text{DP}\left(\alpha + n, \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i}\right),$$

where δ_{θ_i} is the cdf of a point mass at θ_i. This conjugate property of the DP was motivated by Ferguson (1973), desirable for easy algebra and computations.

For a one-sample problem, one might take

$$X_1, \ldots, X_n \mid G \overset{ind}{\sim} G, G \sim DP(\alpha, G_o),$$

where G_o is the baseline measure and α the concentration parameter. Assuming that there are k distinct values among $X_1, \ldots, X_n$, the baseline model is $X_1, \ldots, X_k \mid k \sim G_o$. Note that k is a random variable. The baseline measure G_o is assumed continuous. Binder (1982) was the first to introduce this model to survey sampling; more recently, see Nandram and Yin (2016 a,b). Although G_o can be discrete, it appears that this latter case was not discussed by Antoniak (1994).

Now considering the predictive distribution for θ_{n+1} conditioned on $\theta_1, \ldots, \theta_n$ with G integrated out, we have

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i}.$$

The sequence of predictive distributions for $\theta_1, \theta_2, \ldots$ is called the Polya urn scheme (Blackwell and MacQueen 1973). Here, it is interesting that the probability measure G is discrete with probability one, but $\theta_1, \ldots, \theta_n \overset{iid}{\sim} G_o$, a continuous measure (i.e., the θ_i are continuous, yet $\theta_i = \theta_j, i \neq j$ with positive probability).

The discreteness property of draws from a DP also implies a clustering property. Since the values of draws are repeated, let $\theta_1^*, \ldots, \theta_k^*$ be the k distinct values among $\theta_1, \ldots, \theta_n$ and n_s be the number of θ_s^*, $s = 1, \ldots, k$. The predictive distribution can be equivalently written as:

$$\theta_{n+1}|k, \theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{s=1}^{k} n_s \delta_{\theta_s^*}.$$

Notice that the value θ_s^* will be repeated by θ_{n+1} with probability proportional to n_s. The larger n_s is the higher the probability that it will grow.

Antoniak (1974) wrote down the distribution of k given α and he proved that k is a sufficient statistic for α. This is true when G_o is continuous. It is easy to write down the posterior density with an appropriate prior. Nandram and Yin (2016 a, b) used a grid method to sample α from the posterior density of $\rho = 1/(1 + \alpha)$; they have used the prior $\pi(\alpha) = 1/(1 + \alpha)^2$, a proper, but nearly noninformative prior for α, different from the proper (informative) prior suggested by Escobar and West (1995). Antonelli, Trippa and Haneuse (2016) reviewed several methods and suggested a more complex method. The problem of sampling the posterior density of α is a difficult one, and we believe that the best method is the one of Nandram and Yin (2016 a, b) that we are still trying to improve.

However, if G_o is discrete, k is no longer a sufficient statistic; this result appears to be not so well known. Therefore, if the result is used, this is a violation of the sufficiency principle; further work is needed and this has not been presented any where. While this issue is not the subject of the current paper, we are trying to resolve it.

Sethuraman (1994) provided an elegant equivalent constructive definition of the DP called the stick-breaking construction, which is $G = \sum_{s=1}^{\infty} \pi_s \delta_{\theta_s^*}$, where

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

The construction of $\underset{\sim}{\pi} = \{\pi_1, \pi_2, \pi_3, \dots\}$ can be understood as follows. Starting with a stick of length 1, we break it at β_1 assigning π_1 to be the length of the stick we just broke off. Now continually break the remaining part of the stick to obtain π_2, π_3 and so forth. Despite the continuity of the baseline distribution, samples from DP are discrete distribution with probability one. For computational purposes we use this form of the DP repeatedly.

2.2 Dirichlet Process Mixture Model

In many applications, the discreteness of the DP measure may be inappropriate. As we noted, the most popular application of the DP is in clustering data using mixture models. We model a set of observations $\{y_1, \dots, y_n\}$ using a set of latent parameters $\{\theta_1, \dots, \theta_n\}$ as,

$$\begin{aligned} y_i|\theta_i &\overset{ind}{\sim} h(y_i;\theta_i), \quad i = 1, \dots, n, \\ \theta_i|G &\sim G, \\ G &\sim \text{DP}(\alpha, G_0). \end{aligned} \tag{1}$$

This model is referred to as a Dirichlet process mixture (DPM) model; see Lo (1984) where the DPM was introduced. Nandram and Choi (2004) and Polettini (2017) have applications on SAE. Each θ_i is a latent parameter modeling y_i, while G is the unknown distribution over parameters modeled using a DP. It can be seen as a Dirichlet process mixture of $h(y_i;\theta_i)$, where y_i's with the same value of θ_i belong to the same cluster. The DPM model removes the constraint from discrete measures. The corresponding parametric baseline model with G_0 replacing the random probability measure G is,

$$\begin{aligned} y_i|\theta_i &\overset{ind}{\sim} h(y_i;\theta_i), \quad i = 1, \dots, n, \\ \theta_i &\sim G_0. \end{aligned}$$

Thus, the DPM model is an expansion model of the baseline model, and therefore it should be more robust against non-normality. However, this model does not do much to accommodate outliers, gaps and ties in the data.

There are many Markov chain Monte Carlo (MCMC) methods that can be used to fit the DPM model. Escobar and West (1995) proposed a simple (not necessarily efficient) algorithm by integrating out the random distribution function in the model. Neal (2000) constructed efficient algorithms to fit nonconjugate DPM models. Another idea, which has been used, is to leave the infinite dimensional distribution in the model and find ways of sampling a sufficient but finite number of variables at each iteration. There are two classes of such methods: retrospective samplers (Papaspiliopoulos and Roberts 2008) and slice samplers (Ishwaran and James 2001, Kalli, Griffin and Walker 2011). The slice-efficient sampler is easier to use, as opposed to the complexity of the set up of the retrospective sampling steps, while both samplers are approximately the same in terms of efficiency and performance.

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

(Sethuraman 1994) without truncation error. Again, the stick-breaking algorithm is $G = \sum_{s=1}^{\infty} \pi_s \delta_{\theta_s^*}$ where

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

The key contribution of Kalli, Griffin and Walker (2011) is a general class of slice samplers that can be defined as

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

where $u_i, i = 1, \ldots, n$, are a set of uniform random variables in $(0, 1)$, $d_1, \ldots, d_n$ are a set of integer random variables that map the areas to clusters, and $\xi_1, \xi_2, \ldots$ are a positive sequence. Typically, this latter sequence will be a deterministic decreasing sequence. In our computation, we use $\xi_s = (1-\kappa)\kappa^{s-1}$ where the tuning constant κ is between 0 and 1. Let $K = \max_{i=1}^{n}(K_i)$, where K_i is the largest integer t such that $\xi_t > u_i$. The joint posterior distribution is proportional to

$$\prod_{s=1}^{K} \text{Beta}(\beta_s; 1, \alpha) g_0(\theta_s^*) \prod_{i=1}^{n} \mathbf{1}(u_i < \xi_{d_i})\pi_{d_i}/\xi_{d_i} h(y_i; \theta_{d_i}^*).$$

The variables $\{(\theta_s^*, \beta_s), s = 1, 2, \ldots, K; (d_i, u_i), i = 1, \ldots, n\}$ need to be sampled at each iteration. The Gibbs sampler is as follows.

1. $\pi(u_i|\ldots) \propto \mathbf{1}(0 < u_i < \xi_{d_i})$;
2. $\pi(\theta_s^*|\ldots) \propto g_0(\theta_s^*) \prod_{\{i|d_i=s\}} h(y_i; \theta_s^*)$;
3. $\pi(\beta_s|\ldots) \propto \text{Beta}(a_s, b_s)$, where $a_s = 1 + \sum_{i=1}^{n} \mathbf{1}(d_i = s)$ and $b_s = \alpha + \sum_{i=1}^{n} \mathbf{1}(d_i > s)$;
4. $P(d_i = r|\ldots) \propto \mathbf{1}(r : \xi_r > u_i)\pi_r/\xi_r h(y_i; \theta_r^*), r = 1, \ldots, K.$

Escobar and West (1995) showed how the other parameters are sampled from the baseline model conditional on the ones that are already obtained when G_o is specified (e.g., Normal(θ, δ^2)).

2.3 Hierarchical Dirichlet Process Mixture Model

In a celebrated paper, Teh, Jordan, Beal and Blei (2006) introduced and described hierarchical Dirichlet process mixture (HDPM) model. We give a brief review here and we discuss its relevance to survey sampling.

The HDPM model is

$$y_{ij} \mid \theta_{ij} \overset{ind}{\sim} F(y_{ij} \mid \theta_{ij}), \tag{2}$$

$$\theta_{ij} \mid G_i \overset{ind}{\sim} G_i, \tag{3}$$

$$G_i \mid \alpha, G_o \overset{ind}{\sim} DP(\alpha_o, G_o), \tag{4}$$

$$G_o \mid \gamma, H \sim DP(\gamma, H). \tag{5}$$

Vague gamma priors are assumed for α_o and γ. The specifications, (4) and (5), form the hierarchical Dirichlet process prior and (2), (3), (4) and (5) form the HDPM model. The base measure H is assumed known and conjugate to $F(\cdot \mid \cdot)$, a standard assumption in hierarchical Bayesian models to simplify algebraic manipulation and numerical computation.

It is well-known that inference is sensitive to the specification of baseline measure (e.g., McAuliffe, Blei and Jordan 2006 and Nandram and Yin 2016 a). Therefore, it is good to

have an unspecified baseline distribution like G_0. However, the discreteness of G_0 means that the same value can come from either G_0 or from the balls already drawn in the Polya urn scheme. But it is mandatory to have G_o discrete in this model; note that the G_i must be discrete also. In a lengthy discussion about the discreteness of G_0, Teh, Jordan, Beal and Blei (2006) wrote that if G_o were to be continuous, it "is ruinous for our problem of sharing clusters ... our restriction to discrete G_o has important implications for the design of efficient Markov chain Monte Carlo (MCMC) inference algorithms." They claimed that the HDPM model is appropriate for applications in genetics, information retrieval and text modeling.

Teh, Jordan, Beal and Blei (2006) stated, "We are interested in problems in which the observations are organized into groups and assumed to be exchangeable both within each group and across groups." The HDPM model is used when the y_{ij} are exchangeable within groups and the vectors of the observations are exchangeable across groups; as they clearly pointed out, this is their key interest. Actually, because data come with heterogeneity, this model is generally useful, and the data can be organized within groups after they are collected.

In survey sampling, this model may be fine for simple random sampling. But it is short for complex survey designs with cluster sampling and stratification (i.e., weighting). The θ_{ij} cannot be all exchangeable. In fact, the stick-breaking algorithm (Sethuraman 1994) gives

$$G_o(\theta) = \sum_{s=1}^{\infty} \lambda_s \delta_{\phi_s}(\theta), \;\; \lambda_1 = \tau_1, \lambda_s = \tau_s \prod_{r=1}^{s-1}(1-\tau_r),$$

$$\tau_s \mid \gamma \overset{ind}{\sim} \text{Beta}(1,\gamma), \phi_s \overset{ind}{\sim} H.$$

Therefore, the θ_{ij} can take only the values ϕ_s (a limited set of values in practice). Moreover, for our application on BMI data in which there are outliers, gaps and ties, this model will be inadequate; a more robust model is needed for the observations as well, not just the other levels. This is the reason why we develop two new hierarchical Dirichlet process models in which the observations follow a Dirichlet process.

3. Two-level Dirichlet Process Model

We assume that data are obtained from a two-stage sample survey, for example, a two-stage cluster sampling, stratified or post-stratified sampling that is often seen in SAE problems. The sampled values are observed and the nonsampled values are to be predicted using the two-level models. To gain robustness, these models start with a simple idea that uses a random distribution drawn from the DP in the model instead of some parametric distributions. Especially for the area means, it is hard to know the correct parametric distribution. Assuming a specific parametric form is typically motivated by technical convenience rather than by genuine prior beliefs. One drawback of the parametric model is the over-shrinkage; the posterior mean of certain areas may be shrunk too much towards the overall mean. Using the DP for the area means allows borrowing information moderately within some of the areas instead of all. Moreover since there are gaps and ties in the survey data, it is reasonable to introduce a correlation among area means. Thus, it is important to use a non-parametric procedure. Although presented in a survey sampling framework, the proposed approach can be adapted to general random and mixed effect models.

Let y_{ij} denote the value for the jth unit within the ith area, $i = 1, \ldots, \ell, j = 1, \ldots, N_i$. We assume that y_{ij}, $i = 1, \ldots, \ell, j = 1, \ldots, n_i$, are observed, and inference is required for $\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij}/N_i, i = 1, \ldots, \ell$, the finite population mean of the ith area, and finite population quantiles.

3.1 Two-level Baseline Model

We assume that there are ℓ areas, and within the ith area there are N_i (known) individuals. A sample of size n_i is available from the ith area, and the remaining $N_i - n_i$ values are unknown. Inference is required for the finite population mean and quantile of each area.

For continuous data y_{ij}, $i = 1, \ldots, \ell, j = 1, \ldots, N_i$, one can assume that

$$\begin{aligned} y_{ij}|\nu_i &\overset{ind}{\sim} N\left(\theta + \nu_i, \sigma^2\right), \\ \nu_i &\overset{ind}{\sim} N\left(0, \delta^2\right), \end{aligned} \tag{6}$$

where priors are chosen for θ, δ^2 and σ^2 to form a full Bayesian model. This is the simplest hierarchical Bayesian model (Scott and Smith 1969) without covariates, called the Scott-Smith model, where θ is an overall mean and $\underset{\sim}{\nu} = \{\nu_i, i = 1, \ldots, \ell\}$ are area effects. Letting $\underset{\sim}{\mu} = \{\mu_i, i = 1, \ldots, \ell\}$, where $\mu_i = \theta + \nu_i$, we can write the Scott-Smith model equivalently to a two-level normal model,

$$\begin{aligned} y_{ij}|\mu_i &\overset{ind}{\sim} N\left(\mu_i, \sigma^2\right), \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, N_i, \\ \mu_i &\overset{ind}{\sim} N\left(\theta, \delta^2\right). \end{aligned} \tag{7}$$

Letting $\delta^2 = \frac{\rho}{1-\rho}\sigma^2$, our two-level normal model (baseline parametric model) is then

$$y_{ij}|\mu_i \overset{ind}{\sim} N\left(\mu_i, \sigma^2\right), \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, N_i, \tag{8}$$

$$\mu_i \overset{ind}{\sim} N\left(\theta, \frac{\rho}{1-\rho}\sigma^2\right), \tag{9}$$

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

Here we consider a reparameterization of the Scott-Smith model (7) together with proper non-informative priors that allow computation of marginal likelihood and Bayes factors. We replace δ^2 by $\frac{\rho}{(1-\rho)}\sigma^2$ to gain some analytical and computational simplicity. Note that $\rho = \delta^2/(\delta^2 + \sigma^2)$ is a common intra-class correlation; see Nandram, Toto and Choi (2011) and Molina, Nandram and Rao (2014). We have used the Cauchy prior centered at 0 for θ; one can use a location-scale Cauchy prior distribution (e.g., Gelman, Jakulin, Pittau and Su, 2008), but one would need to specify the location and scale parameters using the data (double using the data is forbidden in Bayesian statistics). The prior on σ^2 is a standard shrinkage prior (almost noninformative).

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. 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 $A_1 = \frac{1-\rho}{\rho}\sum_{i=1}^{\ell} \lambda_i (\tilde{y} - \bar{y}_i)^2 + \sum_{i=1}^{\ell}(n_i - 1)s_i^2$.

Using Bayes' theorem, the joint posterior density of $\underset{\sim}{\mu}, \theta, \sigma^2, \rho$ is

$$\begin{aligned} \pi(\underset{\sim}{\mu}, \theta, \sigma^2, \rho|\underset{\sim}{y}_s) \propto & \left(\frac{1}{\sigma^2}\right)^{(n+\ell)/2} \left(\frac{1-\rho}{\rho}\right)^{\ell/2} \exp\Bigg\{ -\frac{1}{2\sigma^2}\Bigg\{ \sum_{i=1}^{\ell} \Big\{ (n_i - 1)s_i^2 \\ & + \left(n_i + \frac{1-\rho}{\rho}\right)\left(\mu_i - [\lambda_i \bar{y}_i + (1-\lambda_i)\theta]\right)^2 \\ & + \lambda_i \left(\frac{1-\rho}{\rho}\right)(\bar{y}_i - \theta)^2 \Big\} \Bigg\} \Bigg\} \times \frac{1}{(1+\sigma^2)^2} \times \frac{1}{\pi(1+\theta^2)}. \end{aligned} \tag{10}$$

We use a simple method called the sample importance resampling (SIR) algorithm to draw from the posterior distribution $\pi(\underset{\sim}{\mu}, \theta, \sigma^2, \rho|\underset{\sim}{y}_s)$ in (10). 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 $\pi(\underset{\sim}{\mu}, \theta, \sigma^2, \rho|\underset{\sim}{y}_s)$. 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; clearly this is true here. A rough approximation to the joint posterior density (10) and one form which it is easy to draw samples will suffice. We use the same likelihoods (8) and (9) in the two-level normal model together with an improper prior $\pi(\theta, \sigma^2, \rho) \propto \frac{1}{\sigma^2}, -\infty < \theta < \infty, 0 \le \sigma^2 < \infty, 0 \le \rho \le 1$ as the proposal model, that is,

$$\begin{aligned}
\pi_a(\underset{\sim}{\mu}, \theta, \sigma^2, \rho|\underset{\sim}{y}_s) &\propto \pi_a(\underset{\sim}{\mu}|\theta, \sigma^2, \rho, \underset{\sim}{y}_s)\pi_a(\theta|\sigma^2, \rho, \underset{\sim}{y}_s)\pi_a(\sigma^2|\rho, \underset{\sim}{y}_s)\pi_a(\rho|\underset{\sim}{y}_s) \qquad (11)\\
&\propto \prod_{i=1}^{\ell} N\left[\mu_i; \lambda_i \bar{y}_i + (1-\lambda_i)\theta, (1-\lambda_i)\frac{\rho}{1-\rho}\sigma^2\right]\\
&\times N\left(\theta; \tilde{y}, \frac{\sigma^2\rho}{\sum_{i=1}^{\ell}\lambda_i(1-\rho)}\right) \times \text{IG}\left[\sigma^2; (n-1)/2, A_1/2\right]\\
&\times \frac{\Gamma[(n-1)/2]}{(A_1/2)^{(n-1)/2}} \prod_{i=1}^{\ell}(1-\lambda_i)^{1/2}\left[\frac{\rho}{\sum_{i=1}^{\ell}\lambda_i(1-\rho)}\right]^{1/2}.
\end{aligned}$$

We draw a sample from the approximate joint posterior density (11) by first drawing a sample from $\pi_a(\rho|\underset{\sim}{y}_s)$ using the grid method; the remaining parameters being obtained using the multiplication rule of probability.

3.2 Two-level Expansion Models

We describe two models, one with a Dirichlet process at both levels and the other with a Dirichlet process for the sampling process only.

Let us consider a nonparametric hierarchical Bayesian extension of the parametric baseline model,

$$\begin{aligned}
y_{ij}|G_i &\overset{ind}{\sim} G_i, \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, N_i, \qquad (12)\\
G_i|\mu_i &\overset{ind}{\sim} \text{DP}\{\alpha_i, G_0(\mu_i)\},\\
\mu_i|H &\overset{iid}{\sim} H,\\
H &\sim \text{DP}\{\gamma, H_0(\cdot)\},
\end{aligned}$$

where $G_0(\mu_i)$ and $H_0(\cdot)$ can be any parametric distributions. In particular, we consider $G_0 = N(\mu_i, \sigma^2)$ and $H_0(\cdot) = N(\theta, \delta^2)$, where $\delta^2 = \frac{\rho}{1-\rho}\sigma^2$ in (12) to be consistent with the two-level normal model. A full Bayesian model can be obtained by adding prior distributions. For example, we can use proper non-informative priors,

$$\begin{aligned}
\pi(\alpha_i) &= \frac{1}{(\alpha_i+1)^2}, \qquad \alpha_i > 0, \quad i = 1, \ldots, \ell, \qquad (13)\\
\pi(\gamma) &= \frac{1}{(\gamma+1)^2}, \qquad \gamma > 0, \qquad (14)\\
\pi(\theta, \sigma^2, \rho) &= \frac{1}{\pi(1+\theta^2)}\frac{1}{(1+\sigma^2)^2},\\
&\quad -\infty < \theta < \infty, 0 \le \sigma^2 < \infty, 0 \le \rho \le 1, \qquad (15)
\end{aligned}$$

with independence. We call (12), (13), (14) and (15) together with the two-level Dirichlet process a (DPDP) model. Note that the concentration parameters α_i and γ are not included

in the two-level normal model. One might not want to assume that the α_i are identically distributed because they can be very different.

The inference of the DPDP model can be easily performed. We denote $(\underset{\sim}{\mu}, \gamma, \theta, \sigma^2, \rho)$ as $\underset{\sim}{\psi}$ and $\underset{\sim}{\alpha} = \{\alpha_1, \ldots, \alpha_\ell\}$. The posterior density of α_i are independent with 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

$$\pi(\underset{\sim}{\alpha}, \underset{\sim}{\psi} \mid \underset{\sim}{k}, \underset{\sim}{y}^*) = \left[\prod_{i=1}^{\ell} \pi(\alpha_i \mid k_i)\right] \pi(\underset{\sim}{\psi} \mid \underset{\sim}{y}^*), \tag{16}$$

where $\pi(\alpha_i|k_i) \propto \pi(k_i \mid \alpha_i)\pi(\alpha_i)$. For each i, we can draw posterior samples of α_i in the manner similar to the one-level DP model. For the other parameters $\underset{\sim}{\psi}$, we have

$$\begin{aligned} y^*_{i1}, \ldots, y^*_{ik_i} \mid k_i, \mu_i, \sigma^2 &\overset{ind}{\sim} \text{Normal}(\mu_i, \sigma^2), \quad i = 1, \ldots, \ell, \\ \mu_i|H &\overset{iid}{\sim} H, \\ H &\sim \text{DP}\{\gamma, N(\theta, \delta^2)\}, \end{aligned}$$

with the prior in $\pi(\gamma, \theta, \sigma^2, \rho)$. We know that H can be expressed as $H = \sum_{s=1}^{\infty} p_s \delta_{\mu^*_s}$ where

$$p_1 = v_1, \quad p_s = v_s \prod_{j=1}^{s-1}(1 - v_j), \quad v_s \overset{iid}{\sim} \text{Beta}(1, \gamma), \quad \mu^*_s \overset{iid}{\sim} N(\theta, \delta^2).$$

Note that this is a DPM model. So the slice sampler (Kalli, Griffin and Walker 2011) can be used easily to obtain posterior samples of μ and γ. We need to add a few steps in the Gibbs sampler to draw the hyper-parameters, θ, σ^2, ρ. For this specific prior, an accept-reject algorithm is used for the $\pi(\sigma^2, \theta, \rho| \ldots)$ within the Gibbs sampler update.

The algorithm has two steps.

Step 1: For each i ($i = 1, \ldots, \ell$), draw α_i from $\pi(\alpha_i|k_i) \propto \alpha^{k_i} \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i+n_i)} \frac{1}{(\alpha_i+1)^2}$;

Step 2: Draw $\underset{\sim}{\psi}$. Let $K = \max_{i=1}^{n}(K_i)$, where K_i is the largest integer t such that $\xi_t > u_i$.

The Gibbs sampler is as follows.

1. $\pi(u_i| \ldots) \propto \mathbf{1}(0 < u_i < \xi_{d_i})$;
2. $\pi(\mu^*_s| \ldots) \propto N(\mu^*_s; \theta, \delta^2) \prod_{\{i|d_i=s\}} \prod_{j=1}^{k_i} N(y^*_{ij}; \mu^*_s, \sigma^2)$;
3. $\pi(v_s| \ldots) \propto \text{Beta}(a_s, b_s)$, $a_s = 1 + \sum_{i=1}^{\ell} \mathbf{1}(d_i = s)$ and $b_s = \gamma + \sum_{i=1}^{\ell} \mathbf{1}(d_i > s)$;
4. $\pi(\gamma| \ldots) \propto \gamma^{k_0} \frac{\Gamma(\gamma)}{\Gamma(\gamma+\ell)} \frac{1}{(\gamma+1)^2}$, k_0 is the number of distinct $d_1, \ldots, d_\ell$;
5. $P(d_i = t| \ldots) \propto \mathbf{1}(t : \xi_t > u_i) p_t / \xi_t \prod_{j=1}^{k_i} N(y^*_{ij}, \mu^*_t, \sigma^2)$, $t = 1, \ldots, K$;
6. $\pi(\sigma^2, \theta, \rho| \ldots) \propto \prod_{i=1}^{\ell} \prod_{j=1}^{k_i} N(y^*_{ij}; \mu^*_{d_i}, \sigma^2) \times \prod_{s=1}^{K} N(\mu^*_s; \theta, \delta^2) \times \frac{1}{\pi(1+\theta^2)} \frac{1}{(1+\sigma^2)^2}$.

When we have strong beliefs that our sampling population or the area means are from normal distributions, we may choose to use the normal likelihood instead of a random distribution drawn from the DP. Thus, we can have three additional models which are easy to fit. Using normal distributions in both levels gives us the normal model. Using the normal distribution in the first level and the DP as prior,

$$\begin{aligned} y_{ij}|\mu_i &\overset{ind}{\sim} N\left(\mu_i, \sigma^2\right), \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, N_i, \\ \mu_i|H &\overset{iid}{\sim} H, \\ H &\sim \text{DP}\{\gamma, N(\theta, \delta^2)\}, \end{aligned} \tag{17}$$

together with (14) and (15) gives us the DPM model which is easy to fit.

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

$$\begin{aligned} y_{ij}|G_i &\overset{ind}{\sim} G_i, \quad i=1,\ldots,\ell, \quad j=1,\ldots,N_i, \\ G_i|\mu_i &\overset{ind}{\sim} \text{DP}\{\alpha_i, N\left(\mu_i,\sigma^2\right)\}, \\ \mu_i &\overset{iid}{\sim} N(\theta,\delta^2). \end{aligned} \tag{18}$$

We call (18), (13) and (15) the DP normal (DPnormal) model. The algorithm for the DP-normal model is

Step 1 : For each i ($i=1,\ldots,\ell$), draw α_i from $\pi(\alpha_i|k_i) \propto \alpha^{k_i}\frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i+n_i)}\frac{1}{(\alpha_i+1)^2}$;

Step 2: Draw $\underset{\sim}{\psi}$ from the following parametric model which is easy to fit,

$$\begin{aligned} y^*_{ij}|\mu_i &\overset{ind}{\sim} N\left(\mu_i,\sigma^2\right), \quad i=1,\ldots,\ell, \quad j=1,\ldots,k_i, \\ \mu_i &\overset{iid}{\sim} N\left(\theta,\frac{\rho}{1-\rho}\sigma^2\right), \\ \pi(\theta,\sigma^2,\rho) &= \frac{1}{\pi(1+\theta^2)}\frac{1}{(1+\sigma^2)^2}, -\infty<\theta<\infty, 0\le\sigma^2<\infty, 0\le\rho\le 1. \end{aligned} \tag{19}$$

Finally, we look at the sampling process for the DPnormal and the DPDP models. When we integrate out the random probability measure (Blackwell and MacQueen (1973), we get

$$f(\underset{\sim}{y_i} \mid \mu_i,\sigma^2,\alpha_i) = \frac{1}{\sigma}\phi(\frac{y_{i1}-\mu_i}{\sigma})$$
$$\times\prod_{k=2}^{n_i}\left\{\frac{k-1}{\alpha_i+k-1}\frac{\sum_{j=1}^{k-1}\delta_{y_{ij}}(y_{ik})}{k-1}+\frac{\alpha_i}{\alpha_i+k-1}\frac{1}{\sigma}\phi(\frac{y_{ik}-\mu_i}{\sigma})\right\}. \tag{20}$$

Therefore, in each area we are mixing the distributions in (20) using normal mixing distributions and Dirichlet process mixing distributions in the DPnormal and DPDP models respectively. This is how we attempt to accommodate ties, gaps and outliers in the data.

3.3 Posterior Propriety and Prediction

In this section, we demonstrate the propriety of the four joint posterior densities; thereby adding a degree of credence to our Bayesian models. Although we use proper priors throughout to accommodate the use of Bayes factors, our proofs remain correct with an improper prior on θ. We also discuss how to do the prediction.

Theorem 1: The joint posterior density $\pi(\underset{\sim}{\mu},\theta,\sigma^2,\rho|\underset{\sim}{y_s})$ (10) under the two-level normal model is proper if $\ell \ge 2$.

Proof: Since $\frac{\sigma^2}{(1+\sigma^2)^2}\times\frac{1}{\pi(1+\theta^2)} < 1$, we have $\pi(\underset{\sim}{\mu},\theta,\sigma^2,\rho|\underset{\sim}{y_s}) < \pi_a(\underset{\sim}{\mu},\theta,\sigma^2,\rho|\underset{\sim}{y_s})$ which is shown proper in Nandram, Toto and Choi (2011).

Theorem 2: If the posterior density under the normal baseline model is proper, the posterior densities under the DPM, DPnormal and DPDP models are proper.

Proof: For the DPM model, use Lemma 2 of Lo (1984). For the DPnormal and DPDP models, use theorems in Nandram and Yin (2016 a, b) for the simple DP model and Lemma 2 of Lo (1984).

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 prediction intervals for

the finite population quantities for each county. The prediction of the nonsampled values is straight forward under the baseline model and the DPM model.

For the DPnormal and the DPDP models, the sampling process is

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

So conditional on all parameters and for each area, these can be obtained using the Polya urn scheme (Blackwell and MacQueen 1973); see also Nandram and Yin (2016 a,b). So that we can obtain $y_{in_i+1}, \ldots, y_{iN_i}, i = 1, \ldots, \ell$.

The finite population mean is calculated as $\frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$. For the finite population percentiles, we simply order $y_{ij}, j = 1, \ldots, N_i$ from the smallest to the largest, and pick the value in the $.85N_i^{th}$ position and the value in the $.95N_i^{th}$ position for the 85^{th} and 95^{th} percentiles.

4. Empirical Studies

In this section, we compare the four models. In Section 4.1, we describe an application on body mass index (BMI) data, and in Section 4.2, we describe a small simulation study.

4.1 Application to Body Mass Index Data

As an illustrative example, we use a limited set of data from the third National Health and Nutrition Examination Survey (NHANES III), where we study only the 35 largest counties with a population at least 500,000. One of the variables in this survey is body mass index (BMI), the demographic variables are age, race and sex, and there are survey weights. We study BMI data for adults who are older than 20 years since the observed nonresponse rates (see Nandram and Choi 2010) for children and adolescents are high. Our goal is to predict the mean, 85th and 95th percentiles of BMI for the finite population of adults for each county without using the covariates or survey weights; we discuss how to do so in the concluding section.

As we mentioned in previous sections, survey data tend to have ties, outliers and gaps. The BMI data set is an example because in practice, BMI is rounded to one decimal place which creates many ties; BMI data are artificial when presented with more than one decimal place. We plotted dot plots for all thirty-five areas (not shown). The observations are more concentrated and having ties within the range around 25. It is also clear that the data are clustered with gaps. Especially outside the normal weight range, the data become sparse and present bigger gaps. The box plots (not shown) suggest that the distributions are right skewed with outliers in the right tails. Since the predictive inference for the overweight and obese population is very important, the heavy tail of the distribution can not be ignored. Thus we can not automatically use the standard normal assumptions. More robust hierarchical models are desired.

Note that for the county level, all sample sizes are over 100, but these are relatively small compared with the population sizes of the counties. We have compared the DPDP model to the normal model, the DPM model, the DPnormal model and Bayesian bootstrap.

For the DPM and DPDP model, we ran $10,000$ MCMC iterations, used $5,000$ as a "burn in" and used every 5^{th} to obtain $1,000$ converged posterior samples. Table 1 gives the p-values of the Geweke test and the effective sample sizes for the parameters σ^2, θ, δ^2 and γ for the DPM and DPDP model. The p-values are all large so we do not reject the null hypothesis that the Markov chain is stationary. The effective sample sizes are not too

far away from $1,000$. In addition, numerical summaries, trace plots, and auto-correlation plots (not shown) indicate that the MCMC chains converge.

Tables 2, 3 and 4 give the summary statistics, posterior mean (PM) and posterior standard deviation (PSD), of the finite population mean, 85^{th} and 95^{th} percentiles for each county of BMI data under the two-level DP models (normal, DPM, DPnormal and DPDP models) and Bayesian bootstrap respectively. To show the effect of pooling, we also use a Bayesian bootstrap to do prediction in each county individually (no borrowing across counties).

These tables show that roughly similar results are obtained from the two-level DP models. As expected, in terms of efficiency, all four models beat the Bayesian bootstrap. For the finite population mean, Table 2 shows that roughly half of the counties with smaller PSD are under the DPDP model than the normal model; otherwise they are comparable. The PMs under the DPDP model are closer to the PMs under the Bayesian bootstrap, which does not allow for pooling. By comparison, the PMs under the normal model are shrunk towards the overall mean; it is well known that when the area mean is far way from the overall mean, the normal model has the risk of over-shrinkage. It is good that the shrinkages are less under the DPnormal and DPDP models, and there are larger PSDs reflecting the ties, outliers and skewness in the BMI data. In fact, the PSDs are much larger under the DPnormal and DPDP models when the 85^{th} and 95^{th} finite population percentiles are being inferred (Tables 3 and 4).

4.2 Simulation Study

We conduct a small simulation study. We simulated data from the baseline model, the DPM model with $\gamma = 0.5$, the DPnormal model with $\alpha = 0.3$, and the DPDP model with $\alpha = 0.3$ and $\gamma = 0.5$. We also took $\sigma^2 = .01$; $\theta = 0$ and $\rho = .80$. As for the sample sizes, we took $N_i = 5 \times n_i, i = 1, \ldots, \ell = 50$ small areas. The first 10 n_i are set at 35, the second 10 n_i at 50, the third 10 n_i at 100, the fourth 10 n_i at 200 and the fifth 10 n_i at 500. We generated a single data set from each model, and fit all four models; we can do more runs but the computation is time consuming.

We 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 5. For the finite population mean, when data are generated from the baseline and the DPM models, they perform better than the DPnormal and DPDP models. This situation is reversed when data are generated from the DPnormal and DPDP models. Similar results hold when posterior inference is made about the 85^{th} and 95^{th} finite population percentiles, but the DPnormal and DPDP models are much better than the baseline and DPM model when data are generated from them.

The simulation examples show some evidence that the nonparametric models perform better when predictive inference is done for data generated from the DPnormal and DPDP models. This appears to meet our objective of looking at data that have outliers, gaps and ties. More extensive simulation is planned for the future.

5. Concluding Remarks and Future Work

Bayesian nonparametric models are motivated by the desire to avoid overly restrictive assumptions. We have proposed two new nonparametric models for two-stage survey data

using DPs, and we have shown predictive inference for extreme percentiles, where a robust model might be preferred. Using an example on NHANES III and a simulated study, we have demonstrated that the two models with DP sampling processes, DPnormal and DPDP models, might be preferred over the baseline model with normality at both levels and the DPM model with normality in the sampling process. Our key goal was to overcome the gaps, ties and outliers, and possible skewness, that are inherent in the BMI data. The results in Table 5 are particularly informative.

However, it is difficult to provide convincing evidence as to which model is the best. The DPnormal and DPDP models are more complex than the baseline model and the DPM model. Measures to assess which of these infinite dimensional model is preferred are not available, and current literature shows that standard measures such Bayes factor, conditional predictive ordinate, Bayesian predictive p-values and deviance information criterion tend to choose the finite dimensional models, although these models may not be better. See, for example, Carota and Parmigiani (1996) and Petrone and Raftery (1997). Really one is uncertain about what these measures actually do in problems with infinite number of parameters like the Dirichlet process. Yet, one can proceed by predicting a number of data points (not just one) that have been deleted, but in small area inference there are just a handful of observations in some areas.

We consider three extensions. We extend the two-level DP models to three-level DP models; thereby showing a natural procedure for multi-stage (more than three stages) sampling. We also show also show how to include covariates and survey weights.

The first extension to a three-level Dirichlet process model (DPDPDP) is given by

$$\begin{aligned}
y_{ijk}|G_{ij} &\overset{ind}{\sim} G_{ij}, \quad i=1,\ldots,\ell, \quad j=1,\ldots,N_i, \quad k=1,\ldots,M_{ij},\\
G_{ij}|\mu_{ij} &\overset{ind}{\sim} \text{DP}\{\alpha_{ij}, G_0(\mu_{ij})\},\\
\mu_{ij}|H_i &\overset{ind}{\sim} H_i,\\
H_i|\theta_i &\overset{ind}{\sim} \text{DP}\{\gamma_i, H_0(\theta_i)\},\\
\theta_i|F &\overset{iid}{\sim} F,\\
F &\sim \text{DP}\{\gamma_0, F_0(\cdot)\}.
\end{aligned}$$

Here $G_0(\cdot)$, $H_0(\cdot)$ and $F_0(\cdot)$ are parametric distributions. In particular, we consider $G_0 = N(\mu_{ij}, \sigma^2)$, $H_0 = N(\theta_i, \delta_1^2)$ and $F_0 = N(\theta_0, \delta_2^2)$, where $\delta_1^2 = \frac{\rho_1}{1-\rho_1}\sigma^2$ and $\delta_2^2 = \frac{\rho_2}{1-\rho_2}\sigma^2$. A full Bayesian model can be obtained by adding prior distributions. Similar to two-level models, we can use proper non-informative priors,

$$\begin{aligned}
\pi(\alpha_{ij}) &= \frac{1}{(\alpha_{ij}+1)^2}, \quad \alpha_{ij} > 0, \quad i=1,\ldots,\ell, \quad j=1,\ldots,N_i,\\
\pi(\gamma_i) &= \frac{1}{(\gamma_i+1)^2}, \qquad \gamma_i > 0,\\
\pi(\gamma_0) &= \frac{1}{(\gamma_0+1)^2}, \qquad \gamma_0 > 0,\\
\pi(\theta_0, \sigma^2, \rho_1, \rho_2) &= \frac{1}{\pi(1+\theta_0^2)}\frac{1}{(1+\sigma^2)^2},\\
&\quad -\infty < \theta_0 < \infty, 0 \le \sigma^2 < \infty, 0 \le \rho_1 \le 1, 0 \le \rho_2 \le 1,
\end{aligned}$$

with independence.

The second extension is to include covariates in the model. Battese, Harter and Fuller

(1988) extended the Scott-Smith model (6) to include covariates, assuming

$$\begin{aligned} y_{ij}|\nu_i &\overset{ind}{\sim} N\left(\mathbf{x}'_{\mathbf{ij}}\boldsymbol{\beta}+\nu_i,\sigma^2\right), \quad i=1,\ldots,\ell, \quad j=1,\ldots,N_i, \\ \nu_i &\overset{iid}{\sim} N\left(0,\delta^2\right), \end{aligned}$$

where $\boldsymbol{\beta}$ is a p-vector of fixed effects and the ν_i are the random effects. The DPM model with covariates can be easily written as

$$\begin{aligned} y_{ij} \mid \boldsymbol{\beta},\nu_i,\sigma^2 &\overset{ind}{\sim} N(\mathbf{x}'_{\mathbf{ij}}\boldsymbol{\beta}+\nu_i,\sigma^2), \quad i=1,\ldots,\ell, \quad j=1,\ldots,N_i, \\ \nu_i|\mathrm{G} &\overset{iid}{\sim} \mathrm{G} \\ \mathrm{G} &\sim DP\left[\alpha, N\left(0,\frac{\rho}{1-\rho}\sigma^2\right)\right] \\ \pi(\boldsymbol{\beta},\sigma^2,\rho) &\propto 1/\sigma^2, \quad \boldsymbol{\beta}\in R^p, \quad \sigma^2>0, \quad 0<\rho<1, \end{aligned}$$

where ρ is the intracluster correlation. The two-level nonparametric alternative with covariates can be

$$\begin{aligned} y_{ij}-\mathbf{x}'^{(\mathbf{0})}_{\mathbf{ij}}\boldsymbol{\beta}^{(\mathbf{0})}|\mathbf{G_i} &\overset{ind}{\sim} G_i, \quad i=1,\ldots,\ell, \quad j=1,\ldots,N_i, \\ G_i|\boldsymbol{\beta}_{0i} &\overset{ind}{\sim} \mathrm{DP}\left\{\alpha_i, N(\beta_{0i};\sigma^2)\right\}, \\ \boldsymbol{\beta}_{0i}|H &\overset{iid}{\sim} H, \\ H &\sim \mathrm{DP}\left\{\gamma, N(\theta,\frac{\rho}{1-\rho}\sigma^2)\right\}, \end{aligned}$$

where $\mathbf{x}'^{(\mathbf{0})}_{\mathbf{ij}}$ and $\boldsymbol{\beta}^{(0)}$ denote $\mathbf{x}'_{\mathbf{ij}}$ and $\boldsymbol{\beta}$ with the intercepts excluded respectively.

Finally, in many complex surveys, there are also survey weights, but these survey weights are available only for the sample units. An initial step is to standardize the survey weights. These standardized survey weights can be added to any model in one of two possible ways. First, we can add them as a covariate. In this way, after fitting the model, we set the regression parameter associated with the survey weights to zero, and given the other parameters, predict all values (samples and nonsamples). Second, a normalized composite likelihood can be constructed. After this model is fit, set the weights equal unity, and conditional on the parameters, again all the values are predicted. These two ways fall under the general scheme of surrogate sampling (Nandram 2007), but the first approach is much easier.

REFERENCES

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

Antonelli, J., Trippa, L. and Haneuse, S. (2016), "Mitigating bias in generalized linear mixed models: The case for Bayesian nonparametrics,"*Statistical Science*, 31 (1), 80-95.

Azzalini, A. (2013), *The Skew-normal and Related Families*, Cambridge University Press.

Binder, D. A. (1982), "Non-parametric Bayesian models for samples from finite populations," *Journal of the Royal Statistical Society*, Series B, 44 (3), 388-393.

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), 28-36.

Blackwell, D. and MacQueen, J. B. (1973), "Ferguson distributions via Polya urn schemes," *The Annals of Statistics*, 1 (2), 353-355.

Carota, C. and Parmigiani, G., (1996), "On Bayes factor for nonparametric alternatives," *Bayesian Statistics*, 5, 507-511.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90 (430), 577-588.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1 (2), 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), 1360-1383.

Ishwaran, H. and James, L. F. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96 (453), 161-173.

Kalli, M. and Griffin, J. E. and Walker, S. G. (2011), "Slice sampling mixture models," *Statistics and Computing*, 21 (1), 83-105.

Lange, K. L. and Little, R. J. A. and Taylor, J. M. G. (1989), "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, 84 (408), 881-896.

Lo, A. Y. (1984), "On a class of Bayesian nonparametric estimates: I. Density estimates," *The Annals of Statistics*, 12 (1), 351-357.

McAuliffe, J. D., Blei, D. M. and Jordan, M. I. (2006), "Nonparametric empirical Bayes for the Dirichlet process mixture model," *Statistics and Computing*, 16 (1), 5-14.

Malec, D. and Sedransk, J. (1985), "Bayesian inference for finite population parameters in multistage cluster sampling," *Journal of the American Statistical Association*, 80 (392), 897-902.

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), 852-885.

Nandram, B. and Choi, J. W. (2004), 'Nonparametric Bayesian analysis of a proportion for a small area under nonignorable nonresponse," *Journal of Nonparametric Statistics*, 16 (6), 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, 356-374.

Nandram, B. and 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), 1593-1608.

Nandram, B. and Yin, J. (2016a), "Bayesian predictive inference under a Dirichlet process with sensitivity to the normal baseline," *Statistical Methodology*, 28, 1-17

Nandram, B. and Yin, J. (2016b), "A nonparametric Bayesian prediction interval for a finite population mean," *Journal of Statistical Computation and Simulation*, 86 (16), 3141-3157.

Neal, R. M. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, 9 (2), 249-265.

Papaspiliopoulos, O. and Roberts, G. O. (2008), "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models," *Biometrika*, 95 (1), 169-186.

Petrone, S. and Raftery, A. E. (1997), "A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability," *Statistics & Probability Letters*, 36 (1), 69-83.

Polettini, S. (2017), "A generalized semiparametric Bayesian Fay-Herriot model for small area estimation shrinking both means and variances," *Bayesian Analysis*, 12 (3), 729-752.

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

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

Scott, A. and Smith, T. M. F. (1969), "Estimation in multi-stage surveys," *Journal of the American Statistical Association*, 64 (327),830-840.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006), "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, 101 (476), 1566-1581.

Verbeke, G. and Lesaffre, E. (1996), "A linear mixed-effects model with heterogeneity in the random-effects population," *Journal of the American Statistical Association*, 91 (433), 217-221.

Table 1: Summary of Markov chain Monte Carlo (MCMC) diagnostics: the p-values of the Geweke test and the effective sample sizes for the parameters σ^2, θ, δ^2 and γ for the DPM and DPDP model

	P-values for the Geweke test			
Model	σ^2	θ	δ^2	γ
DPM	.483	.414	.459	.620
DPDP	.522	.676	.752	.110
	Effective sample sizes			
Model	σ^2	θ	δ^2	γ
DPM	1000	1000	698	1085
DPDP	1000	938	627	732

NOTE: Random draws are used for the baseline and the DPnormal models.

Table 2: Comparison of posterior mean (PM) and posterior standard deviation (PSD) of the finite population mean for each county of body mass index (BMI) data by four models (normal, DPM, DPnormal and DPDP models) and Bayesian bootstrap

	Bootstrap		Normal		DPM		DPDP		DPnormal	
	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD
1	26.93	0.36	26.93	0.32	26.93	0.36	26.92	0.32	26.92	0.33
2	27.48	0.54	27.24	0.37	27.25	0.36	27.38	0.42	27.24	0.42
3	26.28	0.44	26.51	0.35	26.47	0.38	26.35	0.36	26.55	0.36
4	26.00	0.37	26.34	0.36	26.30	0.36	26.14	0.33	26.35	0.32
5	25.67	0.41	26.18	0.41	26.17	0.40	25.87	0.37	26.16	0.36
6	28.40	0.43	27.85	0.40	27.78	0.40	28.13	0.36	27.84	0.35
7	27.08	0.34	27.03	0.31	27.03	0.35	27.04	0.32	27.02	0.32
8	26.88	0.47	26.88	0.33	26.90	0.39	26.88	0.38	26.93	0.35
9	27.83	0.39	27.46	0.36	27.46	0.36	27.68	0.34	27.49	0.34
10	27.65	0.45	27.39	0.36	27.39	0.34	27.53	0.35	27.33	0.33
11	27.26	0.26	27.18	0.23	27.20	0.24	27.24	0.23	27.19	0.24
12	25.72	0.34	26.15	0.37	26.14	0.34	25.87	0.32	26.11	0.32
13	26.67	0.39	26.75	0.32	26.74	0.39	26.71	0.34	26.80	0.33
14	27.28	0.17	27.23	0.17	27.25	0.18	27.28	0.17	27.25	0.17
15	27.33	0.50	27.15	0.39	27.17	0.39	27.23	0.39	27.10	0.35
16	27.31	0.40	27.17	0.33	27.17	0.34	27.22	0.33	27.15	0.32
17	26.08	0.38	26.39	0.34	26.36	0.37	26.20	0.35	26.41	0.33
18	26.71	0.37	26.79	0.32	26.77	0.41	26.75	0.36	26.81	0.33
19	26.19	0.41	26.46	0.34	26.44	0.37	26.30	0.34	26.51	0.32
20	26.81	0.44	26.86	0.34	26.88	0.38	26.86	0.38	26.89	0.35
21	26.90	0.43	26.90	0.34	26.92	0.39	26.91	0.35	26.91	0.34
22	27.28	0.36	27.12	0.33	27.15	0.33	27.23	0.32	27.15	0.32
23	25.87	0.41	26.27	0.37	26.23	0.37	26.03	0.35	26.31	0.35
24	27.12	0.42	27.04	0.34	27.07	0.37	27.09	0.36	27.05	0.35
25	26.75	0.44	26.80	0.34	26.82	0.37	26.79	0.38	26.83	0.37
26	26.58	0.47	26.74	0.37	26.71	0.42	26.65	0.42	26.77	0.35
27	26.77	0.36	26.82	0.29	26.83	0.35	26.78	0.32	26.83	0.30
28	27.52	0.49	27.28	0.34	27.30	0.35	27.42	0.36	27.25	0.37
29	26.59	0.43	26.75	0.38	26.76	0.43	26.68	0.40	26.79	0.39
30	25.91	0.40	26.32	0.37	26.27	0.38	26.10	0.34	26.35	0.34
31	27.82	0.33	27.52	0.34	27.48	0.34	27.71	0.30	27.52	0.30
32	27.64	0.41	27.37	0.32	27.37	0.33	27.52	0.33	27.38	0.34
33	26.35	0.32	26.53	0.32	26.53	0.37	26.44	0.32	26.58	0.31
34	27.39	0.30	27.22	0.28	27.26	0.29	27.35	0.27	27.27	0.27
35	26.80	0.38	26.84	0.30	26.85	0.36	26.83	0.33	26.87	0.31

Table 3: Comparison of posterior mean (PM) and posterior standard deviation (PSD) of the finite population 85^{th} percentile for each county of body mass index (BMI) data by four models (normal, DPM, DPnormal and DPDP models) and Bayesian bootstrap

	Bootstrap		Normal		DPM		DPDP		DPnormal	
	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD
1	32.14	0.50	32.48	0.35	32.50	0.39	32.27	0.46	32.46	0.47
2	34.76	1.24	32.93	0.45	32.95	0.43	33.77	0.83	34.08	0.82
3	30.76	0.78	32.05	0.39	32.00	0.44	31.34	0.62	31.94	0.63
4	31.57	1.07	31.97	0.43	31.93	0.42	31.84	0.72	32.48	0.61
5	30.51	0.90	31.75	0.47	31.75	0.45	31.11	0.70	31.87	0.72
6	33.82	1.22	33.42	0.44	33.35	0.44	33.51	0.64	33.55	0.67
7	31.59	0.85	32.58	0.36	32.58	0.39	32.07	0.69	32.45	0.72
8	32.25	0.67	32.46	0.36	32.48	0.42	32.32	0.48	32.70	0.53
9	32.81	1.18	33.03	0.41	33.01	0.42	32.99	0.74	33.15	0.75
10	34.01	0.74	33.07	0.39	33.08	0.36	33.53	0.47	33.73	0.48
11	32.75	0.54	32.78	0.26	32.79	0.27	32.76	0.45	32.90	0.49
12	30.26	0.80	31.67	0.42	31.67	0.38	30.92	0.53	31.45	0.53
13	31.91	0.88	32.34	0.36	32.32	0.43	32.15	0.56	32.64	0.57
14	32.37	0.38	32.80	0.19	32.82	0.20	32.46	0.37	32.50	0.37
15	33.39	0.50	32.84	0.40	32.85	0.41	33.10	0.47	33.39	0.42
16	32.21	0.75	32.72	0.37	32.71	0.40	32.41	0.56	32.73	0.62
17	30.88	0.83	31.95	0.40	31.91	0.42	31.41	0.65	32.07	0.72
18	31.18	0.80	32.29	0.39	32.28	0.49	31.68	0.70	32.21	0.85
19	32.03	0.97	32.09	0.38	32.08	0.42	32.05	0.64	32.77	0.56
20	32.71	0.96	32.50	0.39	32.52	0.42	32.63	0.66	33.08	0.61
21	33.08	0.98	32.57	0.40	32.58	0.44	32.87	0.62	33.28	0.56
22	32.06	0.72	32.65	0.36	32.68	0.37	32.34	0.54	32.57	0.57
23	31.18	0.77	31.85	0.42	31.81	0.42	31.47	0.56	32.19	0.70
24	32.66	0.66	32.64	0.37	32.68	0.40	32.67	0.52	32.96	0.52
25	31.63	0.98	32.37	0.39	32.39	0.42	32.05	0.74	32.47	0.73
26	32.02	0.96	32.34	0.40	32.30	0.45	32.22	0.61	32.77	0.57
27	31.56	0.44	32.34	0.31	32.36	0.39	31.85	0.44	32.16	0.50
28	33.51	1.51	32.87	0.39	32.89	0.40	33.00	0.70	33.33	0.80
29	31.53	0.97	32.30	0.45	32.31	0.49	31.99	0.79	32.57	0.80
30	30.62	0.94	31.89	0.43	31.83	0.45	31.37	0.67	32.13	0.71
31	32.36	0.57	33.02	0.38	32.99	0.38	32.62	0.49	32.72	0.49
32	33.24	0.89	32.96	0.37	32.96	0.37	33.05	0.57	33.31	0.62
33	30.54	0.51	32.03	0.37	32.01	0.42	31.20	0.53	31.61	0.57
34	32.48	0.49	32.78	0.31	32.82	0.31	32.59	0.44	32.71	0.45
35	31.78	1.04	32.40	0.35	32.41	0.42	32.09	0.65	32.54	0.75

Table 4: Comparison of posterior mean (PM) and posterior standard deviation (PSD) of the finite population 95^{th} percentile for each county of body mass index (BMI) data by four models (normal, DPM, DPnormal and DPDP models) and Bayesian bootstrap

	Bootstrap		Normal		DPM		DPDP		DPnormal	
	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD
1	35.52	1.27	35.79	0.42	35.81	0.45	35.63	0.83	36.21	0.88
2	40.88	2.32	36.45	0.46	36.47	0.45	38.38	1.48	38.83	1.54
3	34.90	2.58	35.36	0.47	35.32	0.51	34.83	1.18	36.16	1.43
4	35.59	1.12	35.31	0.45	35.27	0.45	35.47	0.73	36.26	0.85
5	35.82	1.61	35.19	0.51	35.19	0.50	35.57	1.03	36.53	0.92
6	39.32	1.58	37.00	0.44	36.94	0.44	38.25	0.73	38.45	0.74
7	35.93	1.12	35.95	0.40	35.94	0.44	35.93	0.81	36.50	0.69
8	37.32	1.49	35.90	0.43	35.92	0.48	36.57	0.92	37.26	0.86
9	38.76	1.54	36.55	0.45	36.53	0.46	37.72	0.83	38.02	0.84
10	39.82	1.64	36.48	0.41	36.48	0.41	37.83	1.13	38.32	1.14
11	37.49	0.94	36.19	0.28	36.21	0.29	37.15	0.72	37.36	0.71
12	35.84	1.50	35.17	0.47	35.18	0.44	35.64	0.86	36.46	0.89
13	36.13	1.20	35.68	0.40	35.66	0.45	35.88	0.81	36.65	0.93
14	36.90	0.80	36.16	0.22	36.19	0.23	36.85	0.70	36.96	0.69
15	36.04	1.47	36.00	0.48	36.03	0.49	35.98	0.71	36.64	0.89
16	36.44	1.40	36.08	0.41	36.08	0.44	36.20	0.84	36.79	0.93
17	34.70	0.99	35.27	0.44	35.23	0.45	34.95	0.77	35.77	0.83
18	35.57	0.81	35.68	0.38	35.65	0.46	35.58	0.58	36.16	0.78
19	34.88	0.88	35.31	0.40	35.30	0.44	35.04	0.62	35.85	0.78
20	37.08	1.89	35.82	0.42	35.84	0.46	36.34	1.04	37.11	1.14
21	35.75	1.03	35.75	0.44	35.77	0.47	35.69	0.66	36.30	0.84
22	35.56	1.08	35.94	0.43	35.98	0.42	35.65	0.81	36.12	0.89
23	36.46	1.46	35.29	0.45	35.24	0.46	35.83	0.98	36.84	0.92
24	37.80	2.17	36.02	0.44	36.06	0.45	36.65	1.17	37.40	1.33
25	37.29	2.60	35.76	0.43	35.77	0.46	36.37	1.39	37.23	1.47
26	36.18	1.92	35.67	0.52	35.62	0.55	35.80	1.11	36.90	1.10
27	36.09	1.30	35.75	0.38	35.77	0.44	35.92	0.82	36.51	0.78
28	40.33	1.37	36.50	0.44	36.53	0.46	38.46	1.03	38.84	0.96
29	35.71	1.10	35.66	0.52	35.67	0.52	35.76	0.88	36.43	0.78
30	34.57	1.11	35.20	0.48	35.15	0.49	34.88	0.80	35.87	0.83
31	35.43	1.06	36.28	0.39	36.26	0.39	35.77	0.62	36.01	0.68
32	39.12	1.40	36.43	0.41	36.43	0.40	37.75	1.03	38.24	1.00
33	34.10	0.83	35.31	0.42	35.30	0.46	34.63	0.63	35.32	0.88
34	35.98	1.02	36.09	0.36	36.12	0.36	35.98	0.79	36.36	0.85
35	37.83	1.13	35.92	0.38	35.92	0.44	37.03	0.89	37.57	0.92

Table 5: Comparison of absolute bias (AB) and posterior root mean squared error (PRMSE) of the finite population mean, 85th percentile and 95th percentile for each simulated data by four models (normal, DPM, DPnormal and DPDP) averaged over areas

	(a) Finite population mean							
	Normal Model		DPM Model		DPnormal Model		DPDP Model	
	AB	PRMSE	AB	PRMSE	AB	PRMSE	AB	PRMSE
Normal Data	6.172	94.21	87.32	127.2	6.536	100.8	63.57	109.8
DPM Data	6.169	93.82	43.27	85.44	6.32	100.70	43.26	85.28
DPnormal Data	1.409	42.88	28.95	64.55	1.484	27.64	1.022	26.35
DPDP Data	1.81	37.98	19.67	57.38	1.74	25.84	1.614	24.17
	(b) Finite population 85th percentile							
	Normal Model		DPM Model		DPnormal Model		DPDP Model	
	AB	PRMSE	AB	PRMSE	AB	PRMSE	AB	PRMSE
Normal Data	70.21	130.7	111.2	155.1	77.39	137.4	80.96	140.3
DPM Data	69.93	133.9	75.49	123.4	78.38	141.3	75.1	122.8
DPnormal Data	379.0	385.9	384.1	394.6	18.05	40.0	16.42	37.09
DPDP Data	307.2	313.4	305.1	315.4	0.8099	7.095	0.4063	4.824
	(c) Finite population 95th percentile							
	Normal Model		DPM Model		DPnormal Model		DPDP Model	
	AB	PRMSE	AB	PRMSE	AB	PRMSE	AB	PRMSE
Normal Data	120.6	182.1	150.3	203.5	133.8	188.4	113.1	196.0
DPM Data	104.0	168.9	114.9	166.0	118.9	176.1	115.2	165.6
DPnormal Data	550.2	556.3	555.3	563.7	35.5	101.0	36.81	103.8
DPDP Data	481.6	486.3	475.8	483.5	25.71	84.85	22.99	74.5

NOTE: The numbers in the table must be multiplied by 10^{-4}.