Bayesian Nonparametric Models for Multi-stage Sample Surveys

by

Jiani Yin

A PhD Dissertation

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Doctor of Philosophy

in

Mathematical Sciences

by

April 4, 2016

APPROVED:

Professor Balgobin Nandram, Advisor
Department of Mathematical Sciences
Worcester Polytechnic Institute

Professor Lynn Kuo
Department of Statistics
University of Connecticut

Dr. Jai Won Choi
Statistical Consultant, Meho Inc.
9504 Mary Knoll Dr., Rockville MD
20850

Professor Marcus Sarkis
Department of Mathematical Sciences
Worcester Polytechnic Institute

Assistant Professor Jian Zou
Department of Mathematical Sciences
Worcester Polytechnic Institute

Abstract

It is a standard practice in small area estimation (SAE) to use a model-based approach to borrow information from neighboring areas or from areas with similar characteristics. However, survey data tend to have gaps, ties and outliers, and parametric models may be problematic because statistical inference is sensitive to parametric assumptions. We propose nonparametric hierarchical Bayesian models for multi-stage finite population sampling to robustify the inference and allow for heterogeneity, outliers, skewness, etc. Bayesian predictive inference for SAE is studied by embedding a parametric model in a nonparametric model. The Dirichlet process (DP) has attractive properties such as clustering that permits borrowing information. We exemplify by considering in detail two-stage and three-stage hierarchical Bayesian models with DPs at various stages. The computational difficulties of the predictive inference when the population size is much larger than the sample size can be overcome by the stick-breaking algorithm and approximate methods. Moreover, the model comparison is conducted by computing log pseudo marginal likelihood and Bayes factors. We illustrate the methodology using body mass index (BMI) data from the National Health and Nutrition Examination Survey and simulated data. We conclude that a nonparametric model should be used unless there is a strong belief in the specific parametric form of a model.

Acknowledgements

I would like to express my deep appreciation and gratitude to my advisor, Professor Balgobin Nandram, for the patient guidance and mentorship he provided to me, all the way from when I was first considering applying to the PhD program in the Department of Mathematical Sciences, through to completion of this degree. Your advice on both research as well as on my career have been priceless. I would also like to thank my committee members, Dr. Jai Won Choi, Professor Lynn Kuo, Professor Marcus Sarkis and Assistant Professor Jian Zou for their time and guidance.

I would especially like to thank the department for providing the support, all the professors I have taken classes with and worked with as a TA and the staff for their help and kindness.

A special thanks to my family. Words cannot express how grateful I am to my mother-in law, father-in-law, my mother, and father for all of the sacrifices that you have made on my behalf. I would also like to thank all of my friends who supported me. At the end I would like express appreciation to my beloved husband Lihan. He has always been there cheering me up and stood by me through the good times and bad times.

Contents

List of Figures

List of Tables

Chapter 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. The survey data structured hierarchically is 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 cluster or multistage 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, inferences 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). That is, the sample size for a given area or domain maybe 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 of the robustness of the inference to possible model misspecification. Particularly, survey data tend to have gaps, ties and outliers. There are extensive research 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 dissertation, we discuss the statistical modeling associated with the analysis of multilevel 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, to avoid critical dependence on

parametric assumptions and to allow for heterogeneity, outliers, skewness, etc. The DP has gained a lot of attention recently. It has nice properties such as clustering and borrowing information which is attractive to SAE. In practice, a model for SAE generally includes covariates to further borrow information. However, it is a reasonable start to explore robust extensions of hierarchical Bayesian models without covariates.

In Section 1.1 we briefly review the DP, the Dirichlet process mixture (DPM) model and other applications of the DP. In Section 1.2 we discuss the methodology for model comparisons. In Section 1.3 we discuss body mass index (BMI) data that we use for illustration.

1.1 Literature Review

1.1.1 Dirichlet Process

In this section we provide a brief overview of the DP starting with a discussion of the basic definition and then some properties. The existence of the DP was established by Ferguson (1973). It is a distribution over distributions, that is, each draw from a DP itself is a distribution (i.e., we are working on functional spaces).

Let $(\Theta, \mathcal{B})$ be a measurable space, with G_0 a baseline measure on the space. Let α be a positive real number. A Dirichlet process, $\mathrm{DP}(\alpha, G_0)$, 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, $\{A_i\}_{i=1}^n$,

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

We write $G \sim \mathrm{DP}(\alpha, G_0)$, if G is a random probability measure with a dis-

tribution given by the DP and α is called the concentration parameter. We have $\mathrm{E}[G(A)] = G_0(A)$, that is the mean of the DP is the baseline distribution G_0 and $\mathrm{Var}[G(A)] = G_0(A)[1 - G_0(A)]/(\alpha + 1)$. The larger α is, the smaller the variance, that is 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.

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

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

where δ_{θ_i} is the cdf of a point mass at θ_i.

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).

A standard interpretation of this scheme is as follows. Each value in Θ is a unique color and draws from G are balls with the drawn value being the color of the ball. We have an urn containing previous seen balls. We start with no balls in the urn. We randomly draw a color from G_0, paint a ball with that color and drop it into the urn. For the $(n+1)^{st}$ ball, we will either randomly draw a new color with probability $\frac{\alpha}{\alpha+n}$, paint a ball with that color and drop it into the urn, or with probability $\frac{n}{\alpha+n}$, randomly draw a ball from the urn, paint a new ball with the same color and drop both balls back to the urn. We observe that with positive probability

draws from G can take the same value regardless of the smoothness of G_0. That is, G is a discrete distribution with probability one.

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

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

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.

Due to a different metaphor, the Polya urn scheme is closely related to a distribution on partitions known as the Chinese restaurant process (Aldous 1985). We have a Chinese restaurant with an infinite number of tables, each of which can seat an infinite number of customers. The first customer enters the restaurant and sits at the first table. The second one enters and decides to either sit with the first customer or at a new table. In general, the $(n+1)^{st}$ customer either joins an already occupied table s with probability proportional to the number n_s of customers already sitting there, or sits a new table with probability proportional to α. We can imagine the tables as clusters and customers sitting at the same table belong to the same cluster. Notice that α controls the number of clusters, with larger α implying a larger number of clusters.

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 stick we just broke off. Now continually break the remaining part of the stick to obtain π_2, π_3 and so forth. In Figure 1.1, we show one possible draw from G, where $G \sim DP[10, N(0,1)]$. Despite of the continuousness of the baseline distribution, samples from DP are discrete distribution with probability one. For computational purposes we use this form of the DP repeatedly.

1.1.2 Dirichlet Process Mixture

In many applications, the almost sure 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,

$$\begin{aligned} y_i|\theta_i &\overset{ind}{\sim} h(\theta_i), \quad i = 1, \dots, n, \\ \theta_i|G &\sim G, \\ G &\sim \mathrm{DP}(\alpha, G_0). \end{aligned} \tag{1.1}$$

This model is referred to as a Dirichlet process mixture (DPM) model. 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,

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

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 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, Walker 2007). 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 which we use in our work, and it is based on the stick-breaking construction without truncation error. We briefly describe the basis of the algorithm here. We know that $G = \sum_{s=1}^{\infty} \pi_s \delta_{\theta_s^*}$ where

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

Given the form of G, we can write

$$\begin{aligned} f(y_i|G) &= \int h(y_i;\theta_i)dG(\theta_i) \\ &= \int h(y_i;\theta_i)\left[\sum_{s=1}^{\infty}\pi_s\delta_{\theta_s^*}(\theta_i)\right]d\theta_i \\ &= \sum_{s=1}^{\infty}\pi_s\int h(y_i;\theta_i)\delta_{\theta_s^*}(\theta_i)d\theta_i \\ &= \sum_{s=1}^{\infty}\pi_s h(y_i;\theta_s^*). \\ &\stackrel{def}{=} f(y_i|\pi,\theta^*) \end{aligned}$$

The idea is to introduce latent variables $\{u_1, u_2, \ldots, u_n\}$ allowing us to sample finite number of variables at each iteration. So the joint density of (y_i, u_i) conditional on π, θ^* is given by

$$f(y_i, u_i|\pi,\theta^*) = \sum_{s=1}^{\infty}\mathbf{1}(u_i < \pi_s)h(y_i;\theta_s^*).$$

One can introduce further latent variables $\{d_1, d_2, \ldots, d_n\}$ which indicate the components of the mixture from which observations are to be taken to give the joint density

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

Updating $\{u_1, u_2, \ldots, u_n\}$ can lead to the simulation of more π's. This problem can be addressed by a more general approach to slice sampling.

A general class of slice sampler can be defined by writing

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

where $\xi_1, \xi_2, \ldots$ 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 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. $P(d_i = r| \ldots) \propto \mathbf{1}(r : \xi_r > u_i) \pi_r / \xi_r h(y_i; \theta_r^*)$, $r = 1, \ldots, K$.

In the next section, we discuss applications of the DP for survey data.

1.1.3 Applications of the Dirichlet Process for Survey Data

The DP and Bayesian nonparametric statistics in general is an active area of research. The DP can be applied to different types of problems that involve clustering and borrowing information which is very attractive to SAE.

An early work using the DP for survey data can be traced back to the work of Binder (1982), an extension of Ericson (1969). Ericson (1969) introduced the Bayesian approach via an exchangeable prior using superpopulation models in survey sampling. Using a multinomial distribution, Ericson (1969) assumed that the superpopulation distribution is discrete and he used a Dirichlet prior distribution as a convenient approximation. The multinomial-Dirichlet model, which we also

call it Bayesian bootstrap, assumes that among the n observed values, $y_1, \ldots, y_n$, there are $1 \leq k \leq n$ distinct values $y_1^*, \ldots, y_k^*$ and y_i^* occurs $m_i \geq 1$ times, in the observed data and $\sum_{i=1}^{k} m_i = n$. The model assumes that the only values that can occur in the population are the y_i^*, an obvious weakness. The Dirichlet prior, with all parameters set to 0, is the Haldane prior which models the proportions of the y_i^* values in the population. This was an original idea of Ericson (1969) although he did not use the Haldane prior which is improper. Instead he used a small positive value for the parameters of the Dirichlet distribution to accommodate a slightly higher degree of smoothness. But with $m_i \geq 1, i = 1, \ldots, k$, the posterior density of the proportions of values in the population is proper. Posterior inference is available for the number of nonsampled values, $M_i - m_i, i = 1, \ldots, k$, in the population, where M_i is the number (assumed known) of y_i^* values in the population.

One drawback of this approach is that the discrete values of the superpopulation distribution are assumed to be a subset of some known countable set. Motivated by Ericson (1969), Binder (1982) extended the method of Ericson (1969) to allow the discrete values to take any real value by using the one-level Dirichlet process (DP) model which is

$$y_1, \ldots, y_N \mid G \overset{iid}{\sim} G \text{ and } G \mid \{\alpha, H_{\underset{\sim}{\psi}}(y), \underset{\sim}{\psi}\} \sim \mathrm{DP}\{\alpha, H_{\underset{\sim}{\psi}}(y)\}, \tag{1.2}$$

where α is the concentration parameter and $H_{\underset{\sim}{\psi}}(y)$, the baseline distribution which is generally assumed to be absolutely continuous. Besides simple random sampling, Binder (1982) also studied stratified random sampling and obtained asymptotic (large sample) results corresponding to standard design-based procedures.

The main reason for using the one-level DP model is to accommodate the gaps and ties in the data. It is easy to show that $\mathrm{Cor}(y_i, y_j) = 1/(1+\alpha), i \neq j$. This

correlation is useful because when there are gaps or ties in the data, it is reasonable to believe that the data are correlated. The exchangeability and the gaps and ties in the responses appear contradictory. However, this is not true because we are not restricted to independent and identically distributed responses from a common parametric distribution but rather from a random distribution which follows a DP. When the random distribution is integrated out, the responses become equicorrelated. Moreover, under the DP the responses are discrete with probability one, thereby making the DP a natural clustering algorithm. Even when a simple random sample is taken from a population, there may be hidden structures in the data that the DP can accommodate because it is essentially nonparametric.

Nandram and Yin (2016a, b) reported some results for simple random sampling when the one-level DP model is used. We discuss them briefly here, more details are given in Chapter 2.

Nandram and Yin (2016a) discussed the sensitivity of inference about the finite population mean with respect to different baseline distributions (other than the normal) and a possible solution using a leave-one-out kernel for the DP and a mixture distribution for the DPM.

Nandram and Yin (2016b) gave a closed-form nonparametric Bayesian prediction interval estimate for the finite population mean of a large population, since under the one-level DP model, when the population size is much larger than the sample size, the computational task becomes expensive. An approximate Bayesian procedure which is very close to the exact intervals is provided by using the exchangeability property of the DP together with normality.

Within the Bayesian nonparametric paradigm, there is another choice. The

attractive DPM model is

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

$$\mu_i \mid G \overset{iid}{\sim} G \text{ and } G \mid H \sim \text{DP}(\alpha, H),$$

where the parameters of H are assumed known. It is worth noting that in the DPM model the parametric distribution, $f(y \mid \mu_i, \tau)$, has to be specified. Besides, in practice, inference is likely to be sensitive to the specification of $f(y \mid \mu_i, \tau)$ and model diagnostics will be needed. Nevertheless, the whole idea is that the discreteness (Ferguson 1973) of G in the DP is removed by using the DPM model (Ferguson 1983, Lo 1984). This is advantageous for many applications (e.g., estimation of a density function), but with a simple random sample, the DPM model appears to be inappropriate that it may need to specify H at least partially. For simple random sample the data may have gaps and ties, and it seems more appropriate to use the one-level DP model, which is more nonparametric than the DPM model. Generally, a normal baseline is used, but clearly other distributions can be used. A disadvantage of the DP prior is that if $G \sim \text{DP}(\alpha, H)$, then with probability one, G is a discrete distribution. However, for the finite populations this is not a serious restriction since all survey data are inherently discrete due to limitations of measuring instruments, etc.

In SAE, having only a small sample in a given area, we borrow strength from related areas or domains to produce estimates with adequate precision and increase the effective sample size. The DP definitely provides a promising solution to this type of problem. However, owing to DP's complexity it has received very little attention in the survey methodology community.

Currently, most of the existing models using the DP in the survey sampling

are simple applications of the DPM model. Nandram and Choi (2004) proposed a nonparametric Bayesian analysis of a proportion for a small area under nonignorable nonresponse, an application of the DPM model. The use of the DP prior gains more flexibility and robustness to departures from the assumption of a parametric distribution. Malec and Müller (2008) provided an application of the DPM model in the context of logistic regression, a semi-parametric model to describe the geographic diversity of the U.S. population, where Dirichlet process mixtures of multivariate normals for county-specific random effects are assumed. Chaudhuri and Ghosh (2011) considered the use of empirical likelihood method, together with a DP prior, in SAE instead of full parametric likelihood as another way of robustifying the inference. When combined with appropriate proper priors, it defines a semi-parametric Bayesian approach, which can handle continuous and discrete outcomes in area and unit level models, without specifying the distribution of the outcomes as in the classical Bayesian approach.

Next, we discuss models that provide the hierarchical structure using the DP and borrow strength in some ways with applications in other fields of statistics.

1.1.4 Other Applications of the Dirichlet Process

One example is, Müller, Quintana and Rosner (2004) who considered an extension of the DPM model to produce a combined inference over related nonparametric Bayes models. The hierarchical extension formalizes borrowing of strength across different but related studies, e.g. combining inference from related pharmacological studies. Their model allows linking the submodels at an intermediate level.

Another example is an application in the machine learning. Teh, Jordan, Beal and Blei (2006) proposed a hierarchical model, specifically one in which the base measure for the child DP is itself distributed according to a DP which allows sharing

mixture components between different groups with the application in document modeling, a nonparametric extension of the latent Dirichlet allocation model (Blei et al. 2003). This hierarchical model is also borrowing information, but the groups or clusters are latent and have overlaps, which is different from what is normally observed in survey sampling.

Dunson (2009) focused on the problem of choosing a prior for an unknown random effects distribution within a Bayesian hierarchical model. He obtains a sparse representation by allowing a combination of global and local borrowing of information which can be applied in the analysis of longitudinal and functional data.

Although various hierarchical Bayesian models using the DP are proposed, they are not specially designed for the problems we typically considered in the survey sampling, e.g. multi-stage sample surveys or SAE, and they cannot be applied directly. We propose nonparametric and semiparametric models to provide different degrees of robustness and to accommodate the hierarchical structure in multi-stage sampling. In the next section, we discuss the general methodology we used to conduct model comparison.

1.2 Model Comparison

Here we review some model comparison approaches including Bayes factor (ratio of marginal likelihoods), log pseudo marginal likelihood (LPML), delete-one cross validation (CV) and deviance information criterion (DIC).

For the Bayes factors, we have to use proper priors. Basu and Chib (2003) presented a method for comparing semiparametric Bayesian models, constructed under the DPM framework which is based on the basic marginal likelihood identities (Chib 1995). But this is a very complicated method. Nandram and Kim (2002)

proposed an easier approach to compute the marginal likelihood. We use their calculation to evaluate our models. We give the general approach here and discuss the details for each model in Chapter 3 and 4. Let $\underset{\sim}{y}$ denote the observations and Ω denote the parameters. We can write the marginal likelihood as,

$$\begin{aligned} M(\underset{\sim}{y}) &= \int f(\underset{\sim}{y}|\Omega)\pi(\Omega)d\Omega \\ &= \frac{\int \left\{f(\underset{\sim}{y}|\Omega)\pi(\Omega)/\pi_a(\Omega|\underset{\sim}{y})\right\}\pi_a(\Omega|\underset{\sim}{y})d\Omega}{\int \left\{\pi(\Omega)/\pi_a(\Omega|\underset{\sim}{y})\right\}\pi_a(\Omega|\underset{\sim}{y})d\Omega}, \end{aligned}$$

where $f(\underset{\sim}{y}|\Omega)$ is the likelihood function, $\pi(\Omega)$ is the prior distribution and $\pi_a(\Omega|\underset{\sim}{y})$ is a reasonable approximation to the posterior distribution $\pi(\Omega|\underset{\sim}{y})$. It should be easy to draw samples from $\pi_a(\Omega|\underset{\sim}{y})$. Assuming samples $\{\Omega^{(h)}, h = 1,\ldots,M\}$ drawn from $\pi_a(\Omega|\underset{\sim}{y})$, then $\widehat{M(\underset{\sim}{y})} = \sum_{h=1}^{M} W^{(h)} f(\underset{\sim}{y}|\Omega^{(h)})$, where $W^{(h)} = \left\{\pi(\Omega^{(h)})/\pi_a(\Omega^{(h)}|\underset{\sim}{y})\right\} \Big/ \left\{\sum_{h=1}^{M}[\pi(\Omega^{(h)})/\pi_a(\Omega^{(h)}|\underset{\sim}{y})]\right\}$. Note that integrating out parameters in the model as much as possible leads to improved estimations. See Lo (1984) and Kuo (1986). Possible accurate Monte Carlo methods, e.g. thermodynamic integration (Lartilliot and Philippe 2006), can be used but they are much more complicated and need relatively much longer computing time. For the normal baseline model, since this is a parametric model it is easy to write down $f(\underset{\sim}{y}|\Omega)$, $\pi(\Omega)$ and $\pi_a(\Omega|\underset{\sim}{y})$. For the models having DPs, we need to use the Polya urn scheme by integrating out DPs to obtain the specific forms of $f(\underset{\sim}{y}|\Omega)$, $\pi(\Omega)$ and $\pi_a(\Omega|\underset{\sim}{y})$.

The conditional predictive ordinate (CPO) proposed by Geisser (1980) is a statistic that can detect observations that were fitted poorly by a given parametric model. If having observations $\underset{\sim}{y}$, it is defined as the predictive density of observation i given

all the other observations, that is

$$\begin{aligned} \text{CPO}_i &= f(y_i|y_{(i)}) = \frac{f(\underset{\sim}{y})}{f(y_{(i)})} \\ &= \left(\int \frac{1}{f(y_i|y_{(i)}, \Omega)} f(\Omega|\underset{\sim}{y}) d\Omega \right)^{-1}, \end{aligned}$$

where $y_{(i)}$ is the data $\underset{\sim}{y}$ without ith observation. A Monte Carlo approximation of the CPO_i is given by $\widehat{\text{CPO}_i} = \left\{ \frac{1}{M} \sum_{h=1}^{M} \frac{1}{f(y_i|y_{(i)}, \Omega^{(h)})} \right\}^{-1}$, where $\{\Omega^{(h)}, h = 1, \ldots, M\}$ are samples from the posterior distribution $f(\Omega|y)$. If observations are conditionally independent, $\widehat{\text{CPO}_i} = \left\{ \frac{1}{M} \sum_{h=1}^{M} \frac{1}{f(y_i|\Omega^{(h)})} \right\}^{-1}$. The CPO statistics can be used to detect outliers. Ntzoufras (2009) established that inverse CPO values larger than 40 can be considered as possible outliers and higher than 70 as extreme values. A summary statistic of the CPO_i is the log pseudo marginal likelihood (LPML) which is given by

$$\text{LPML} = \sum_i \log(\widehat{\text{CPO}_i}).$$

Larger values of LMPL indicate better fit. Note that the value of LPML is very similar to the log of marginal likelihood under the same model.

The deviance information criterion (DIC) (Spiegelhalter et al. 2002) is another Bayesian measure of goodness-of-fit,

$$\text{DIC} = 2 \left\{ \frac{1}{M} \sum_{h=1}^{M} D(\underset{\sim}{y}, \Omega^{(h)}) \right\} - D(\underset{\sim}{y}, \hat{\Omega}),$$

where $\hat{\Omega}$ is a point estimate for Ω such as the mean of the posterior simulations, $\Omega^{(h)}$ are posterior simulations and $D(y, \Omega) = -2\log f(y|\Omega)$. DIC has been suggested as a criterion of model fit when the goal is to pick a model with best out-of-sample predictive power. A smaller value of DIC indicates a better fit.

The delete-one cross validation (CV) divergence measure, defined in Wang et al.

(2012), is

$$\begin{aligned} \text{CV} &= \frac{1}{\#\{y_i\}} \sum_i |y_i - \text{E}(y_i|y_{(i)})|, \\ \text{E}(y_i|y_{(i)}) &= \text{E}_{\Omega|y_{(i)}} \text{E}(y_i|y_{(i)}, \Omega) \\ &= \int \text{E}(y_i|y_{(i)}, \Omega) f(\Omega|y_{(i)}) d\Omega \\ &\approx \sum_{h=1}^{M} \text{E}(y_i|y_{(i)}, \Omega^{(h)}) V_i^{(h)}, \end{aligned}$$

where $V_i^{(h)} \propto 1/f(y_i|y_{(i)}, \Omega^{(h)})$. DIC and CV provide reasonable assessments of model fit while considering the model complexity.

The Bayes factor provides some evidence about the fit of the embedded parametric model. Unfortunately, it might suffer some flaws when comparing a parametric model with an infinite dimensional nonparametric model using DPs. Carota (2006) considered the inconsistency problems arising when using the Bayes factor under certain conditions. The difficulties arise when the parametric model is nested in the nonparametric alternative if there are no ties in the data. In this case, the Bayes factor depends on the data only through the sample size (Carota and Parmigiani 1996) and as the number of distinct observations get larger, the Bayes factor increasingly favors the parametric model in the presence of very extreme departures even if the parametric model is incorrect. Other methods have similar problems because they are based on the likelihood function directly or indirectly. We may need a cross-validation method to compare the models or to perform simulation study. However, they are all computational intensive and time-consuming procedures.

1.3 Applications

1.3.1 Body Mass Index (BMI) Data

For illustrative purpose, we discuss the third National Health and Nutrition Examination Survey (NHANES III), a survey conducted during the period October 1988 through September 1994. Due to confidentiality reasons, the final data set for this study uses only the 35 largest counties with a population at least 500,000.

One of the variables in this survey is body mass index (BMI) and the demographic variables are age, race and sex. We study BMI data for adults who are older than 20 years since the observed nonresponse rate for children and adolescents are high. So nonresponse is not an important issue and we do not address it here. Our goal is to predict the mean, 85th and 95th percentiles of BMI for the finite population of adults, post-stratified by county for each sub-domain formed by age, race and sex. Many sub-domains by county are very small or some have no sample.

Higher than what is considered as a healthy weight for a given height is described as overweight or obese. Obesity-related conditions include heart disease, stroke, type 2 diabetes and certain types of cancer, some of the leading causes of preventable death. BMI is a person's weight in kilograms divided by the square of height in meters and used as a screening tool for overweight or obesity. A high BMI can be an indicator of high body fatness. If your BMI is less than 18.5 kgm^{-2}, it falls within the underweight range. If your BMI is 18.5 kgm^{-2} to 24.9 kgm^{-2}, it falls within the normal or healthy weight range. If your BMI is 25.0 kgm^{-2} to 29.9 kgm^{-2}, it falls within the overweight range. If your BMI is 30.0 kgm^{-2} or higher, it falls within the obese range. A child's weight status is determined using an age- and sex-specific percentile for BMI rather than BMI categories used for adults. Overweight is defined as a BMI at or above the 85th percentile and below the 95th percentile for children

and teens of the same age and sex. Obesity is defined as a BMI at or above the 95th percentile for children and teens of the same age and sex.

As we mentioned in previous sections, survey data tend to have ties and gaps. BMI data set is an example because in practice, BMI is rounded to one decimal place which creates many ties. We present the dot plots for all thirty-five areas (see Figure 1.2). The observations are more concentrated and having ties within the range around 25. It is also clear that the data are clustered and present gaps. Especially outside the normal weight range, the data become sparse and present bigger gaps. The box plots (see Figure 1.3) suggest that the distributions are right skewed with outliers in the right tail. 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.

1.4 Plan of the Dissertation

This dissertation has four additional chapters. In Chapter 2, we discuss the one-level DP model for simple random sampling which is used to avoid assumptions regarding the shape of the finite population distribution. Posterior propriety, predictive inference and sensitivity are discussed. In Chapter 3, we propose two-level Bayesian models using the DP in each level for more complex designs. Model comparison and predictive inference are conducted. The results for BMI data and simulated data are presented. In Chapter 4, we extend two-level models to three-level models. Finally, in Chapter 5 we summarize our results, present concluding remarks, and discuss future research work.

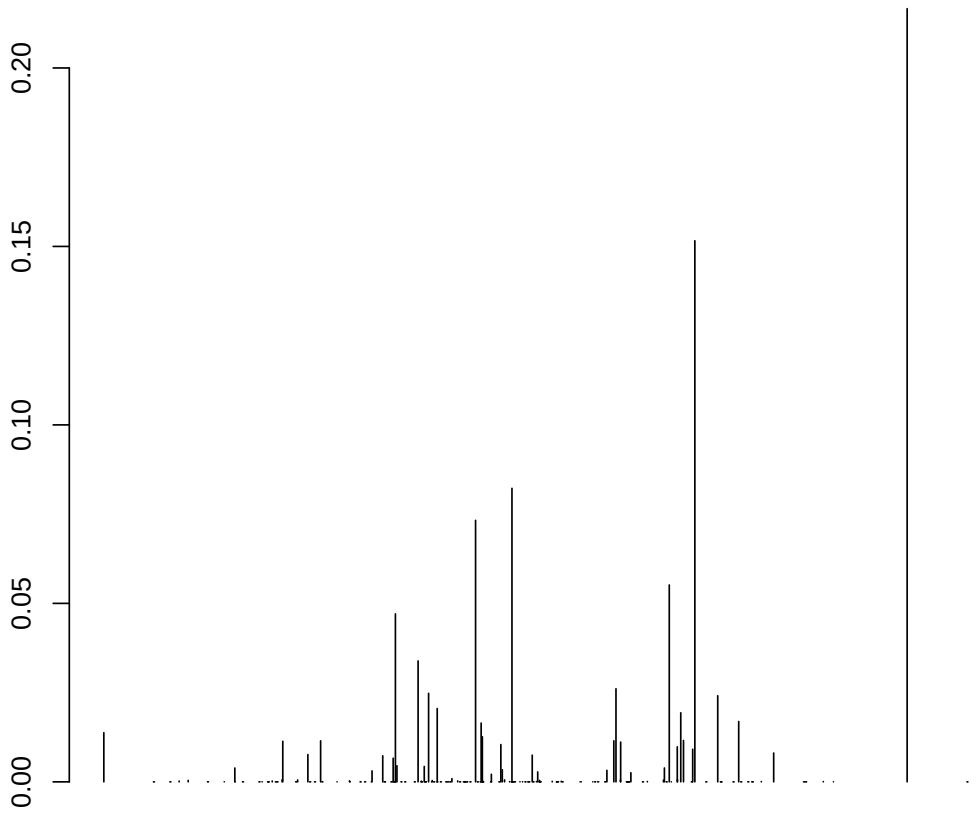

Figure 1.1: Plot of one possible draw from G, where $G \sim DP[10, N(0, 1)]$

Figure 1.2: Dot plots of body mass index (BMI) for thirty-five counties

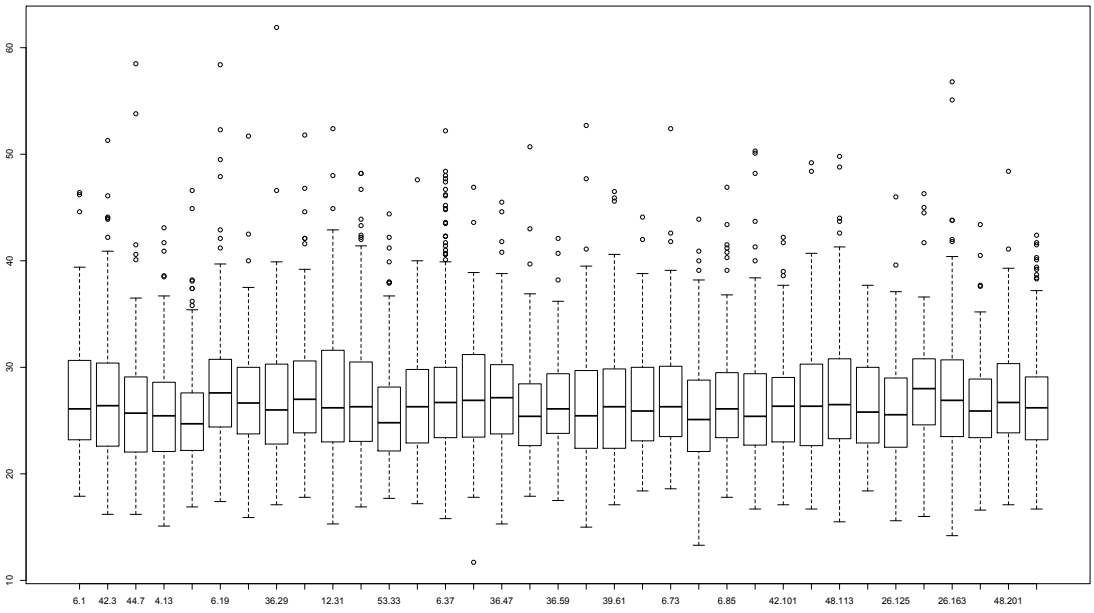

Figure 1.3: Box plots of body mass index (BMI) for thirty-five counties

Chapter 2

One-level Dirichlet Process Models

In Chapter 2, we summarize the methodology and results discussed in Nandram and Yin (2016a, b). We assume that a simple random sample is drawn from a finite population and the population values follow a random distribution which is drawn from the DP. The sampled values are observed and the nonsampled values are to be predicted using the one-level Dirichlet process (DP) model. In Section 2.1, we discuss the inference of the one-level DP model. In Section 2.2, we prove that the posterior distribution is proper under the one-level DP model. In Section 2.3, we discuss the prediction for the finite population when the DP is used for the sampling process including exact and approximate methods. In Section 2.4, we explore the sensitivity to the normal baseline and provide a possible solution.

We have a simple random sample of size n from a population of size N. We assume that the sampled values are $y_1, \ldots, y_n$ and nonsampled values are $y_{n+1}, \ldots, y_N$. Let $\underset{\sim}{y} = (\underset{\sim}{y}_s, \underset{\sim}{y}_{ns})$, where $\underset{\sim}{y}_s = \{y_i, i = 1, \ldots, n\}$ is the vector of observed values and $\underset{\sim}{y}_{ns} = \{y_i, i = n+1, \ldots, N\}$ vector of unobserved values. Inference is re-

quired for the finite population mean, $\bar{Y} = \sum_{i=1}^{N} y_i/N$, and data $y_1, \ldots, y_n$ are available. Note that $\bar{Y} = \sum_{i=1}^{N} y_i/N = f\bar{y}_s + (1-f)\bar{y}_{ns}$, where $f = n/N$ is the sampling fraction, $\bar{y}_s = \sum_{i=1}^{n} y_i/n$, the sample mean, and $\bar{y}_{ns} = \sum_{i=n+1}^{N} y_i/(N-n)$, the nonsample mean which is to be predicted. The sample variance denotes as $s^2 = \sum_{i=1}^{n}(y_i - \bar{y}_s)^2/(n-1)$. Thus, we need random samples from the posterior density of $\underset{\sim}{y}_{ns}$ given $\underset{\sim}{y}_s$.

2.1 Basic Methodology

We use the one-level DP model for the population values to construct a 95% nonparametric Bayesian prediction interval for a finite population mean. For the one-level DP model we assume that

$$y_1, \ldots, y_N \mid G \overset{iid}{\sim} G \text{ and } G \mid \{\alpha, H_{\underset{\sim}{\psi}}(y), \underset{\sim}{\psi}\} \sim \text{DP}\{\alpha, H_{\underset{\sim}{\psi}}(y)\}, \tag{2.1}$$

where $H_{\underset{\sim}{\psi}}(y)$ is the smooth baseline cdf and the pdf is $h_{\underset{\sim}{\psi}}(y)$. The parameters α and $\underset{\sim}{\psi}$ are unknown and a priori we assume that they are independent with prior distributions, $\pi(\underset{\sim}{\psi})$ and $\pi(\alpha)$. We will use a 'Cauchy' type prior for α, sometimes called a shrinkage prior, of the form α, $p(\alpha) = 1/(\alpha+1)^2, \alpha > 0$ (a f density with two degrees of freedom in both the numerator and denominator). It is slightly more convenient to use $p(\alpha) = 1/(\alpha+1)^2, \alpha > 0$ rather than the half Cauchy density $p(\alpha) = 2/\pi(\alpha^2+1), \alpha > 0$ (Polson and Scott 2012). We will specify appropriate noninformative prior for $\underset{\sim}{\psi}$, denoted by $\pi(\underset{\sim}{\psi})$. Under the assumption of independence of α and $\underset{\sim}{\psi}$, we have

$$\pi(\alpha, \underset{\sim}{\psi}) \propto \frac{1}{(\alpha+1)^2}\pi(\underset{\sim}{\psi}), \alpha > 0 \tag{2.2}$$

with appropriate support for $\underset{\sim}{\psi}$ depending on the baseline. We call (2.1) and (2.2) the one-level Dirichlet process (DP) model. And the corresponding nested parametric baseline model is

$$y_1, \ldots, y_N \mid \underset{\sim}{\psi} \overset{iid}{\sim} H_{\underset{\sim}{\psi}}(y), \tag{2.3}$$

with prior $\pi(\underset{\sim}{\psi})$.

Integrating out G (Blackwell and MacQueen 1973), we have $y_1 \sim h_{\underset{\sim}{\psi}}(y_1)$ and for $i = 2, \ldots, n$,

$$y_i | y_1, \ldots, y_{i-1} \sim \frac{i-1}{\alpha + i - 1} \left\{ \frac{\sum_{j=1}^{i-1} \delta_{y_j}(y_i)}{i-1} \right\} + \frac{\alpha}{\alpha + i - 1} h_{\underset{\sim}{\psi}}(y_i).$$

So the joint posterior density $\pi(\alpha, \underset{\sim}{\psi} \mid \underset{\sim}{y_s})$ under the one-level DP model is proportional to

$$h_{\underset{\sim}{\psi}}(y_1) \left[\prod_{i=2}^{n} \left\{ \frac{i-1}{\alpha + i - 1} \left\{ \frac{\sum_{j=1}^{i-1} \delta_{y_j}(y_i)}{i-1} \right\} + \frac{\alpha}{\alpha + i - 1} h_{\underset{\sim}{\psi}}(y_i) \right\} \right] \pi(\underset{\sim}{\psi}) \pi(\alpha). \tag{2.4}$$

Let k, $1 \leq k \leq n$, denote the number of distinct values among $y_1, \ldots, y_n$. Antoniak (1974) showed that $p(k \mid \alpha) = s_n(k) \alpha^k \Gamma(\alpha) / \Gamma(\alpha + n), \alpha > 0$, where $s_n(k)$, the absolute values of the Stirling numbers of the first kind (Abramowitz and Stegun 1965), are independent of $\underset{\sim}{\psi}$. Then, the joint posterior density of $\underset{\sim}{\psi}$ comes from the baseline model conditional on only the distinct values. Letting $y_1^*, \ldots, y_k^*$ denote the k distinct sample values ($k \geq 2$) and $\underset{\sim}{y^*} = \{y_1^*, \ldots, y_k^*\}$, we have

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

with the prior in $\pi(\underset{\sim}{\psi})$. Thus the joint posterior density is

$$\pi(\alpha, \underset{\sim}{\psi} \mid k, \underset{\sim}{y}^*) = \pi(\alpha \mid k)\pi(\underset{\sim}{\psi} \mid \underset{\sim}{y}^*), \tag{2.5}$$

where $\pi(\alpha \mid k) \propto p(k \mid \alpha) \times \pi(\alpha), \alpha > 0$, and $\pi(\underset{\sim}{\psi} \mid \underset{\sim}{y}^*) \propto \{\prod_{i=1}^{k} h_{\underset{\sim}{\psi}}(y_i^*)\}\pi(\underset{\sim}{\psi})$.

Typically, it is straight forward to draw $\underset{\sim}{\psi}$. However, it is not really trivial to draw α without using a special kind of prior; see Nandram and Choi (2004) for a discussion of the gamma prior which was introduced earlier by Escobar and West (1995). We present two improved methods to draw α from its posterior density,

$$\pi(\alpha \mid k) \propto \frac{\alpha^k \Gamma(\alpha)}{\Gamma(\alpha + n)(\alpha + 1)^2}, \alpha > 0. \tag{2.6}$$

The first method would be transforming α according to $\rho = 1/(\alpha+1)$ (correlation in the DP) and simplifying (2.6) we get

$$\pi(\rho \mid k) \propto \frac{(1-\rho)^{k-1}\rho^{n-k}}{\prod_{j=1}^{n-1}\{1 - \rho + \rho j\}}, 0 \leq \rho \leq 1. \tag{2.7}$$

Note that $\lim_{\rho \to 0} \pi(\rho \mid k) = 0 = \lim_{\rho \to 1} \pi(\rho \mid k)$, and $\pi(\rho \mid k)$ is well defined and differentiable everywhere in the closed interval $[0, 1]$. Because the posterior density of ρ is not in a simple form, we use a one-dimensional grid method to draw samples from it, thereby avoiding MCMC methods (e.g., Metropolis sampler). The unit interval is simply divided into 100 sub-intervals of equal width, and the joint posterior density is approximated by a discrete distribution with probabilities proportional to the heights of the continuous distribution at the mid-points of these sub-intervals. Now, it is easy to draw a sample from this univariate discrete distribution of $\pi(\rho \mid k)$. It is efficient to remove sub-intervals with small probabilities (smaller than 10^{-6}); we call the others probable sub-intervals. To draw a single deviate, we first draw one of the

probable sub-intervals. After we have obtained this sub-interval, a uniform random variable is drawn within this sub-interval. This is a standard jittering procedure which provides different deviates with probability one.

However, this method tends to give larger value of α. We use another transformation of α. Letting $\alpha = e^{\psi}$, the posterior density for ψ is

$$\pi(\psi \mid k) \propto \frac{e^{k\psi}}{(1+e^{\psi})^2 \prod_{j=1}^{n-1}(j+e^{\psi})}, -\infty < \psi < \infty.$$

We note that $\pi(\psi \mid k)$ is logconcave (i.e., strongly unimodal with a unique mode), and describe an iterative procedure for finding the posterior mode of ψ and α. Taking the first derivative of $\pi(\psi \mid k)$ and setting it equal zero, we get the fixed point solution

$$\psi = \ln\left\{\frac{k}{\sum_{j=1}^{n-1}(j+e^{\psi})^{-1} + 2(1+e^{\psi})^{-1}}\right\}.$$

Thus, starting at $\psi = 0$ after a few iterations we get the posterior mode $\hat{\psi}$ and therefore the posterior mode $\hat{\alpha} = e^{\hat{\psi}}$. This is similar to a procedure described in Liu (1996) which we have discovered independently. Then taking the second derivative of $\pi(\psi \mid k)$ to approximate the variance of ψ, that is

$$\widehat{\mathrm{Var}}(\psi) \approx -\frac{1}{\pi(\psi \mid k)''} = \frac{1}{e^{\psi}\left[\sum_{j=1}^{n-1} \frac{j}{(e^{\psi}+j)^2} + \frac{2}{(e^{\psi}+1)^2}\right]}.$$

Use the grid method on the range of $\hat{\psi} \pm 10\,\widehat{\mathrm{Var}}(\psi)$ to obtain posterior samples. Since $\pi(\psi \mid k)$ is logconcave, probabilities outside this range can be ignored.

Here, for convenience we present the details of the normal baseline distribution. We take $H_{\underset{\sim}{\psi}}(y)$ to be

$$H_{\underset{\sim}{\psi}}(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t-\mu)^2} dt, -\infty < y < \infty,$$

the cdf of the normal random variable with mean μ and variance σ^2 (i.e., $\underset{\sim}{\psi} = (\mu, \sigma^2)$) and $\pi(\mu, \sigma^2) \propto 1/\sigma^2, -\infty < \mu < \infty, \sigma^2 > 0$. It is easy to draw $\underset{\sim}{\psi}$ from the posterior density. Letting $\bar{y}_* = \sum_{i=1}^k y_i^*/k$ and $s_*^2 = \sum_{i=1}^k (y_i^* - \bar{y}_*)^2/(k-1)$, we have $\mu \mid \sigma^2, k, \bar{y}_*, s_*^2 \sim \text{Normal}(\bar{y}_*, \sigma^2/k)$ and $\sigma^{-2} \mid s_*^2, k \sim \text{ Gamma}\{(k-1)/2, (k-1)s_*^2/2\}$. That is, $\sqrt{k}(\mu - \bar{y}_*)/s_* \mid \bar{y}_*, s_*^2, k \sim t_{k-1}$. Thus, the posterior distribution of (μ, σ^2) is proper and it is trivial to draw μ and σ^2.

2.2 Propriety of the Posterior Distributions

Theorem 2.2.1 is a statement about propriety of the joint posterior density under the one-level DP model. This is useful because if the posterior density is improper, inference about the finite population mean will be defective (i.e., the coverage of the prediction interval will be unknown). Thus, Theorem 2.2.1 adds credibility to the Bayesian procedure. Theorem 2.2.1 may be known, but it is difficult to retrieve.

Theorem 2.2.1 *If the posterior density under the baseline model is proper, the posterior density under the one-level Dirichlet process model is proper.*

Proof: Without loss of generality, we assume that the k distinct values come first. Then using the form of the joint posterior density in (2.4) and noting that $(i-1)/(\alpha+i-1) \leq 1, i = 1, \ldots, n$, and $\sum_{j=1}^{i-1} \delta_{y_j}(y_i)/(i-1) \leq 1, i = 2, \ldots, n$, we have

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

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

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

$$\int \int_0^\infty \frac{\pi(\underset{\sim}{\psi})}{(\alpha+1)^2} \prod_{i=1}^{n} h_{\underset{\sim}{\psi}}(y_i) d\alpha d\underset{\sim}{\psi} < \infty. \tag{2.8}$$

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

$$\int \pi(\underset{\sim}{\psi}) \prod_{i=1}^{n} h_{\underset{\sim}{\psi}}(y_i) d\underset{\sim}{\psi} < \infty. \tag{2.9}$$

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

2.3 Prediction for the Finite Population

Nandram and Yin (2016b) discussed the predictive inference under the one-level DP model. Given a sample from a finite population, we provided a nonparametric Bayesian prediction interval for a finite population mean when a standard normal assumption may be tenuous. The predictions for the two-level and multi-level DP model discussed in Chapter 3 and 4 follow in a similar manner. We showed how to compute the exact prediction interval and useful approximations to the prediction interval. We compared the exact interval and the approximate interval with three standard intervals, design-based interval under simple random sampling, an empirical Bayes interval and a moment-based interval which uses the mean and variance under the DP. However, these latter three intervals do not fully utilize the posterior distribution of the finite population mean under the DP. Using several numerical examples and a simulation study we showed that the approximate Bayesian interval is a good competitor to the exact Bayesian interval for different combinations of sample sizes and population sizes.

First, we review a well-known prediction interval. Under simple random sampling a 95% prediction interval for $\bar{Y}$ is

$$\bar{y}_s \pm z_{2.5}\sqrt{\frac{1-f}{n}}s, \tag{2.10}$$

where $z_{2.5}$ is the 2.5^{th} percentile point of the standard normal density. We call this interval the design based interval (DBI) and the design based method (DBM), and it is pertinent to start with it.

Note that if we assume the Bayesian model

$$y_1, \ldots, y_N \mid \mu, \sigma^2 \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2), \pi(\mu, \sigma^2) \propto 1/\sigma^2, -\infty < \mu < \infty, \sigma^2 > 0,$$

the Bayesian prediction interval is

$$\bar{y}_s \pm t_{n-1,2.5}\sqrt{\frac{1-f}{n}}s,$$

where $t_{n-1,2.5}$ is the 2.5^{th} percentile of the Student's t density on $n-1$ degrees of freedom. This is true because the prior predictive distribution of $\bar{Y}$ is normal with mean $f\bar{y}_s+(1-f)\mu$ and variance $(1-f)\frac{\sigma^2}{N}$, $\mu \mid \sigma^2, \bar{y}_s \sim N(\bar{y}_s, \sigma^2/n)$ and $(n-1)s^2/\sigma^2 \mid s^2 \sim \chi^2_{n-1}$. For large n the prediction interval in (2.10) is an approximate (normality) 95% Bayesian prediction interval. However, it is well-known that this latter interval is not robust to non-normality especially when the sample size is small.

In order to obtain the exact prediction interval, we show how to obtain samples from the joint posterior density of $\underset{\sim}{y}_{ns}, \alpha, \underset{\sim}{\psi}$ given $\underset{\sim}{y}_s$. We have

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

Once samples are taken from $\pi(\alpha, \underset{\sim}{\psi} \mid \underset{\sim}{y_s})$, using the composition rule, samples are obtained from $p(\underset{\sim}{y_{ns}} \mid \alpha, \underset{\sim}{\psi}, \underset{\sim}{y_s})$.

Thus, we get 10,000 values of $\bar{Y}$; order these values and pick the 95% prediction interval to be $(\bar{y}_{(250)}, \bar{y}_{(9750)})$, where the values are arranged in increasing order. We call this interval the full (exact) Bayesian interval (FBI) and the method the full Bayesian method (FBM). Clearly, this procedure can be used for inference about quantiles. For each draw of the entire population compute the required quantile (e.g., median, Q) and then a 95% credible interval is $(Q_{(250)}, Q_{(9750)})$.

In theory it is easy to draw $\underset{\sim}{y_{ns}}$. To each of the 10,000 iterates, simply fill in the values $y_{n+1}, \ldots, y_N$ (data augmentation). Using the generalized Polya urn scheme, for $j = 1, \ldots, N - n$, we have

$$y_{n+j} \mid \{\alpha, \underset{\sim}{\psi}, y_1, \ldots, y_{(n+j-1)}\} \sim \frac{\alpha}{\alpha + n + j - 1} H + \frac{\sum_{s=1}^{n+j-1} \delta_{y_s}(y_{n+j})}{\alpha + n + j - 1}. \qquad (2.11)$$

It is now easy to draw the nonsampled values one by one using (2.11).

However, when the population size is much larger than the sample size, the computation becomes prohibitive. Thus, we obtain an approximate interval which is virtually the same as the FBI for large populations. This is obtained using the central limit theorem for exchangeable random variables. As competitors we also consider other approximations such as those based on the posterior mean and variance of the finite population mean together with the assumption of normality. We develop several approximate calculations to overcome this difficulty.

Approximate Bayesian Prediction Interval

Let

$$\lambda = n(\alpha + N)/N(\alpha + n) \text{ and } \phi = 1/(\alpha + n + 1),$$

where $0 \leq \lambda \leq 1$ is a shrinkage parameter and ϕ is the posterior correlation. Momentarily, let $\mathrm{E}'(\bar{Y}) = \mathrm{E}(\bar{Y} \mid \mu, \sigma^2, \alpha, \underset{\sim}{y}_s)$ and $\mathrm{Var}'(\bar{Y}) = \mathrm{Var}(\bar{Y} \mid \mu, \sigma^2, \alpha, \underset{\sim}{y}_s)$.

Theorem 2.3.1 *Assuming that the one-level Dirichlet process model holds,*

$$E'(\bar{Y}) = \lambda \bar{y}_s + (1-\lambda)\mu,$$

$$Var'(\bar{Y}) = \lambda \left[(n-1)\phi(1-f)\frac{s^2}{n} + (1-\lambda)\left\{ \phi(\bar{y}_s - \mu)^2 + (1-\phi)\frac{\sigma^2}{n} \right\} \right].$$

Therefore, it is easy to describe the approximate Bayesian interval (ABI). As $y_{n+1}, \ldots, y_N \mid \underset{\sim}{y}_s$ are exchangeable, using Theorem 2.3.1,

$$\bar{Y} \mid \mu, \sigma^2, \alpha, \underset{\sim}{y}_s \sim \mathrm{Normal}\{\mathrm{E}(\bar{Y} \mid \mu, \sigma^2, \alpha, \underset{\sim}{y}_s), \mathrm{Var}(\bar{Y} \mid \mu, \sigma^2, \alpha, \underset{\sim}{y}_s)\}, \tag{2.12}$$

asymptotically (as n and N go to infinity with $n < N$). In our case this is a very reasonable approximation for finite population sampling because N is generally large enough. With this normal approximation, we can proceed in the same manner as we did for the FBM; the difference is that we do not have to draw the nonsampled values. We call this method the approximate Bayesian method (ABM).

This is an enormous saving over the FBI because as we will see this approximation is very good for large population sizes where there are large computational savings. However, if quantiles are needed, the ABI must be abandoned and the exact method must be used.

Empirical Bayes and Exact Moment Prediction Intervals

Like the design based prediction intervals, we construct two additional approximate prediction intervals which are based on the DP. The first method is empirical Bayes and the second obtains the exact mean and variance via numerical integration

(not a sampling based method).

First, we describe the empirical Bayes method. We will substitute posterior modes of μ, σ^2 and α into (2.12). The posterior modes of μ and σ^2 are in closed forms and they are respectively $\hat{\mu} = \bar{y}_*$ and $\hat{\sigma}^2 = (k-1)s_*^2/(k+1), k > 1$. However, the posterior mode of α is a bit more involved. We study two procedures for finding the posterior mode of α, one is the iterative procedure as we discussed in Section 2.1 by taking transformation of $\alpha = e^{\psi}$ and the other uses stochastic optimization.

The stochastic optimization to get the posterior mode is easy to perform. We have already shown how to get 10,000 iterates from the posterior density of ρ in (2.7). Note that $\pi(\rho \mid k)$ is unimodal but not logconcave because it is the density of $\log(\rho)$ which is logconcave. Simply compute the $\pi(\rho \mid k)$ at each of the iterates. Then, take the value $\hat{\rho}$ where $\pi(\rho \mid k)$ takes the largest value. So $\hat{\alpha} = \hat{\rho}/(1-\hat{\rho})$ is the posterior mode. Both the iterative procedure and the stochastic optimization give essentially the same posterior mode. We will call this interval the empirical Bayes interval (EBI) and the method to construct it the empirical Bayes method (EBM).

Second, we describe the integration to obtain the exact moments (mean and variance). In Theorem 2.3.2 we obtain almost the complete forms of the moments.

Theorem 2.3.2 *Assuming that the one-level Dirichlet process model holds and* $k \geq 4$*,*

$$E(\bar{Y} \mid \underset{\sim}{y}_s, k) = E(\lambda \mid k)\bar{y}_s + \{1 - E(\lambda \mid k)\}\bar{y}_* \quad \textit{and} \quad Var(\bar{Y} \mid \underset{\sim}{y}_s, k) = V_1 + V_2,$$

$$V_1 = (n-1)(1-f)\frac{s^2}{n}E(\lambda\phi \mid k)$$

$$+\{(\bar{y}_s - \bar{y}_*)^2 + \frac{(k-2)s_*^2}{k(k-3)}\}E\{\lambda(1-\lambda)\phi \mid k\} + \frac{(k-1)s_*^2}{(k-3)n}E\{\lambda(1-\lambda)(1-\phi) \mid k\},$$

$$V_2 = (\bar{y}_s - \bar{y}_*)^2\, Var(\lambda \mid k) + \frac{(k-2)s_*^2}{k(k-3)}E\{(1-\lambda)^2 \mid k\},$$

where expectations are taken over the posterior density of α.

Proof: We integrate $\Omega = (\mu, \sigma^2, \alpha)$ out of the moments, stated in Theorem 2.3.1, using the conditional mean and variance formulas. That is,

$$\mathrm{E}(\bar{Y} \mid \underset{\sim}{y}_s) = \mathrm{E}\{\mathrm{E}(\bar{Y} \mid \underset{\sim}{y}_s, \Omega)\}, \tag{2.13}$$

$$\mathrm{Var}(\bar{Y} \mid \underset{\sim}{y}_s) = V_1 + V_2, V_1 = \mathrm{E}\{\mathrm{Var}(\bar{Y} \mid \underset{\sim}{y}_s, \Omega)\}, V_2 = \mathrm{Var}\{\mathrm{E}(\bar{Y} \mid \underset{\sim}{y}_s, \Omega)\}, \tag{2.14}$$

where $\mathrm{E}(\bar{Y} \mid \underset{\sim}{y}_s, \Omega)$ and $\mathrm{Var}(\bar{Y} \mid \underset{\sim}{y}_s, \Omega)$ are given by Theorem 2.3.1. We need to determine V_1 and V_2. It is worth noting that α and (μ, σ^2) are independent a posteriori with $(k-1)s_*^2/\sigma^2 \mid \underset{\sim}{y}_*, k \sim \chi^2_{k-1}$ and $\sqrt{k}(\mu - \bar{y}_*)/s_*^2 \mid \underset{\sim}{y}_*, k \sim t_{k-1}$, a Student's t density. Then, $\mathrm{Var}(\mu \mid \underset{\sim}{y}_s, k) = \frac{(k-2)s_*^2}{k(k-3)}$ and $\mathrm{E}(\sigma^2 \mid \underset{\sim}{y}_s, k) = \frac{(k-1)s_*^2}{k-3}, k \geq 4$.

For (2.13), using the independence of μ and α,

$$\mathrm{E}(\bar{Y} \mid \underset{\sim}{y}_s) = \mathrm{E}\{\lambda \bar{y}_s + (1-\lambda)\mu \mid \underset{\sim}{y}_s\} = \mathrm{E}(\lambda \mid k)\bar{y}_s + \{1 - \mathrm{E}(\lambda \mid k)\}\bar{y}_*, \tag{2.15}$$

where $\lambda = n(\alpha + N)/N(\alpha + n)$ as in Theorem 2.3.1. Next, we find V_1 and V_2 in (2.14).

First, we find V_1 in (2.14). It is easy to show that

$$V_1 = (n-1)(1-f)\frac{s^2}{n}\mathrm{E}(\lambda\phi \mid \underset{\sim}{y}_*, k) + \mathrm{E}\{\lambda(1-\lambda)\phi(\mu - \bar{y}_s)^2 + \lambda(1-\lambda)(1-\phi)\frac{\sigma^2}{n} \mid \underset{\sim}{y}_s, k\},$$

where $\phi = 1/(\alpha + n + 1)$ as in Theorem 2.3.1. Now because α and μ are independent,

$$\mathrm{E}\{\lambda(1-\lambda)\phi(\mu - \bar{y}_s)^2 \mid \underset{\sim}{y}_s, k\} = \{(\bar{y}_s - \bar{y}_*)^2 + \mathrm{Var}(\mu \mid \underset{\sim}{y}_s, k)\}\mathrm{E}\{\lambda(1-\lambda)\phi \mid \underset{\sim}{y}_s, k\}.$$

Because $\mathrm{E}(\sigma^2 \mid \underset{\sim}{y}_*) = \frac{(k-1)s_*^2}{k-3}$ and $\mathrm{Var}(\mu \mid \underset{\sim}{y}_s, k) = \frac{(k-2)s_*^2}{k(k-3)}, k \geq 4$, we have

$$V_1 = (n-1)(1-f)\frac{s^2}{n}\mathrm{E}(\lambda\phi \mid k)$$

$$+\{(\bar{y}_s - \bar{y}_*)^2 + \frac{(k-2)s_*^2}{k(k-3)}\}\mathrm{E}\{\lambda(1-\lambda)\phi \mid k\} + \frac{(k-1)s_*^2}{(k-3)n}\mathrm{E}\{\lambda(1-\lambda)(1-\phi) \mid k\}. \quad (2.16)$$

Second, we find V_2 in (2.14). We use the standard formula for variance,

$$V_2 = \mathrm{E}[\{\mathrm{E}(\bar{Y} \mid \underset{\sim}{y}_s, \Omega) - \mathrm{E}(\bar{Y} \mid \underset{\sim}{y}_s)\}^2]$$

where $\mathrm{E}(\bar{Y} \mid \underset{\sim}{y}_s)$ is given by (2.15). It is easy to show that

$$\mathrm{E}(\bar{Y} \mid \underset{\sim}{y}_s, \Omega) - \mathrm{E}(\bar{Y} \mid \underset{\sim}{y}_s) = (\bar{y} - \bar{y}_*)\{\lambda - \mathrm{E}(\lambda \mid \underset{\sim}{y}_*, k)\} + (\mu - \bar{y}_*)(1-\lambda).$$

Then, completing the squares and using the independence of μ and α again, we have

$$V_2 = (\bar{y}_s - \bar{y}_*)^2\mathrm{Var}(\lambda \mid k) + \frac{(k-2)s_*^2}{k(k-3)}\mathrm{E}\{(1-\lambda)^2 \mid k\}. \quad (2.17)$$

Finally, the integration over α can be obtained either by numerical or Monte Carlo techniques. We use the latter with the 10,000 draws we already made from the posterior density of α, described in (2.6), which we write fully as

$$\pi(\alpha \mid k) = \frac{\alpha^{k-1}\{\prod_{j=1}^{n-1}(j+\alpha)\}^{-1}(1+\alpha)^{-2}}{\int_0^\infty \alpha^{k-1}\{\prod_{j=1}^{n-1}(j+\alpha)\}^{-1}(1+\alpha)^{-2}d\alpha}, \alpha > 0.$$

Letting $g(\alpha)$ be any integrable function of α,

$$\mathrm{E}\{g(\alpha) \mid k\} = \frac{\int_0^\infty g(\alpha)\alpha^{k-1}\{\prod_{j=1}^{n-1}(j+\alpha)\}^{-1}(1+\alpha)^{-2}d\alpha}{\int_0^\infty \alpha^{k-1}\{\prod_{j=1}^{n-1}(j+\alpha)\}^{-1}(1+\alpha)^{-2}d\alpha}.$$

Then, a good Monte Carlo estimate of $\mathrm{E}\{g(\alpha) \mid k\}$ is

$$\widehat{\mathrm{E}}\{g(\alpha) \mid k\} = \sum_{h=1}^{10000} w_h g(\alpha_h),$$

where $w_h \propto \alpha_h^{k-1}\{\prod_{j=1}^{n-1}(j+\alpha_h)\}^{-1}(1+\alpha_h)^{-2}, h = 1, \ldots, 10000$, and $\alpha_h \overset{iid}{\sim} \pi(\alpha \mid k)$. We apply this method to each of the required integrals. The computation of the expectations took only a few seconds. We will call this interval the exact moment interval (EMI) and the method to construct it the exact moment method (EMM).

Example and Simulation Studies

To compare our five intervals/methods, we discuss fourteen examples and a simulation study. We are particularly interested in the comparison between the approximate Bayesian method (ABI/ABM) and the full (exact) Bayesian method (FBI/FBM) but we also make comparisons with the other intervals/methods: design based (DBI/DBM), empirical Bayes (EBI/EBM) and exact moment (EMI/EMM).

In the fourteen examples the population sizes vary considerably. The first thirteen examples are on the third National Health and Nutrition Examination Survey (NHANES III). These are the data on BMI where we assume that equivalent simple random samples are taken from thirteen states. This example data is about females older than 45 years which is different from the examples in Chapter 3 and 4. The population sizes for the obesity study are around one million and the sample sizes are considerably smaller making the prediction problem challenging in terms

of time. The fourteenth dataset is taken from Aitkin (2010) on income which he used to discuss finite population sampling. This is a much smaller population which creates little difficulty in terms of time for the FBM.

We show three tables here. In Table 2.1 we present a comparison of four methods (DBM, EBM, EMM and ABM) by examples. We have used the posterior mean (PM) and posterior standard deviation (PSD) of the finite population mean. As expected, PSD is directly related to the sample size n, i.e. smaller sample size larger PSD. However, our main purpose here is to compare PM and PSD across different methods. There are some differences among the four methods. Sometimes the differences are large. In Table 2.2 we have first assessed normality of the posterior distribution of $\bar{Y}$ using the Kolmogorov-Smirnov test (KST). In Table 2.3 we have compared the time (hours) it takes to do the computation on our Linux Computational Node with 2.70GHz and 8 CPU Cores. For further discussion of the results, see Nandram and Yin (2016b).

For population sizes of $1,000$ the time to run EBM is not significant. However, the time to run population sizes of $1,000,000$ is intolerable, and therefore, an approximation such as the one we have developed is useful. More importantly the posterior distributions of the finite population mean under ABM and FBM are approximate normal distributions and posterior inferences are similar. So it is reasonable to use ABI for large populations and the FBI for small populations.

We have used several numerical examples and a simulation study with simple random samples drawn from the Parzen-Rosenblatt kernel density estimator. We have one recommendation. The FBM should be used when prediction is to be done for small to moderate populations (size less than 500) and the ABM should be used for much larger populations.

Prediction inference based on the stick-breaking algorithm

The exact method must be used if quantile estimation is needed, but the computational time can be prohibitive for large populations. We develop an approximate calculation using the stick-breaking structure to overcome this difficulty. Instead of integrating out G, the posterior distribution of G is still a DP with a different concentration parameter and baseline distribution. Let us denote $G|y_1, \ldots, y_n$ as G^*. That is

$$y_{n+1}, \ldots, y_N \mid G^* \overset{iid}{\sim} G^* \text{ and } G^* \sim \text{DP}\left\{\alpha + n, \frac{\alpha}{\alpha+n} H_{\underset{\sim}{\psi}}(y) + \frac{\sum_{j=1}^{n} \delta_{y_j}}{\alpha+n}\right\}. \quad (2.18)$$

We can draw the non-sampled values $\underset{\sim}{y}_{ns}$ from G^* which has the following stick-breaking structure (Sethuraman 1994),

$$\begin{aligned} G^* &= \sum_{s=1}^{\infty} \omega_s \delta_{\phi_s}, \quad \omega_1 = \upsilon_1, \quad \omega_s = \upsilon_s \prod_{t=1}^{s-1}(1-\upsilon_t), \\ \upsilon_s &\overset{iid}{\sim} \text{Beta}(1, \alpha+n), \quad \phi_s \overset{iid}{\sim} \frac{\alpha}{\alpha+n} H_{\underset{\sim}{\psi}}(y) + \frac{\sum_{j=1}^{n} \delta_{y_j}}{\alpha+n}. \end{aligned}$$

Then set $E(\omega_s) < \epsilon$, where ϵ is a very small number, and draw non-sampled values $\underset{\sim}{y}_{ns}$ from G^* that is $G|y_1, \ldots, y_n$.

We have many alternative methods to perform the prediction when the population size is too large (Yin and Nandram, working paper).

2.4 Sensitivity to the Normal Baseline

It is well known that the one-level DP model and DPM model are sensitive to the specifications of the baseline distribution. Generally, in many applications a normal distribution is used for the baseline distribution. Therefore, Nandram and Yin

(2016a) showed the extent of the sensitivity of inference about the finite population mean with respect to six distributions (normal, lognormal, gamma, inverse Gaussian, a two-component normal mixture and a skewed normal).

We specify various density functions for $\underset{\sim}{\psi}$. We also show how to draw samples from the posterior density of $\underset{\sim}{\psi}$. Specifically, we consider the normal, lognormal, gamma, inverse Gaussian, two-component mixture and skewed normal distributions. We state conditions for the posterior density to be proper under the baseline model. Following Theorem 2.2.1, the posterior density under the corresponding one-level DP model is proper under the same conditions. To avoid the asterisk notation, we will let $y_1, \ldots, y_k$ denote the distinct values. Results for the normal, lognormal, gamma, inverse Gaussian, two-component mixture and skewed normal are given in Table 2.4.

Example and Simulation Studies

We have compared the one-level DP model using these baselines with the Polya posterior (fully nonparametric) and the Bayesian bootstrap (sampling with a Haldane prior). We used two examples, one on income data and the other on BMI data, to compare the performance of these three procedures. These examples show some differences among the six baseline distributions, the Polya posterior and the Bayesian bootstrap, indicating that the normal baseline model cannot be used automatically. In addition, we consider a simulation study to assess this issue further. Here we present the example on BMI data which is on the NHANES III. These are the data on BMI for females older than forty-five years where we assume that an equivalent simple random sample is taken from a US state. The sample size is 45 with 20 distinct values and the population size is $190,472$, making the prediction problem challenging in terms of time. In both cases, histograms (omitted) of the

sampled values are right skewed.

We consider inference for the finite population mean in Table 2.5 for BMI data. We have plotted the posterior densities of the finite population mean in Figure 2.1 for BMI data.

It is clear that inference about the finite population can be different from the normal baseline when other appropriate baselines are used. In particular, if a baseline, other than the normal is used, inference about the finite population mean can change. Although not reported here, it is also true that inference about a population quantile (e.g., median) will vary with these baselines. This depends on the sample size and the population size as well.

Leave-one-out Kernel Baseline for the one-level DP model

Nandram and Yin (2016a) presented a solution to the sensitivity problem faced by the one-level DP model. Clearly, a solution has to be based on a nonparametric distribution. As we noted, the Monte Carlo method of McAuliffe, Blei and Jordan (2006) is difficult to used for the one-level DP model. So we use the leave-one-out kernel density estimator.

Hardle (1991) described the leave-one-out kernel density estimator; a Bayesian version (again not fully within the Bayesian paradigm) is available (e.g., Brewer 2000; Hu, Poskitt and Zhang 2012). To avoid the asterisk notation, we will let $y_1, \ldots, y_k$ denote the distinct values. With a single parameter ψ for the window width, we assume that $y_1, \ldots, y_k \mid \psi$ are independent with

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

where $\phi(\cdot)$ is the standard normal density function (e.g., Silverman 1986). We take

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

So, the posterior density of ψ is

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

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

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

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

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

In Chapter 2, we have proposed a nonparametric Bayesian model for the simple random sampling. We generalize the one-level DP model to complex surveys where a hierarchical model is needed in Chapters 3 and 4. We do not consider the sensitivity issue further although this is still important. Our main goal is to develop hierarchical DP models for multi-stage sample surveys.

Table 2.1: Comparison of posterior mean (PM) and posterior standard deviation (PSD) of the finite population mean for fourteen examples by methods

	DBM		EBM		EMM		ABM	
$n; N$	PM	PSD	PM	PSD	PM	PSD	PM	PSD
25; 608491	25.880	1.205	26.254	1.046	25.879	1.158	25.807	1.307
556; 4453263	28.045	0.272	28.185	0.276	28.046	0.271	28.131	0.275
162; 2704478	28.086	0.490	28.160	0.495	28.088	0.487	28.250	0.508
86; 1985501	28.860	0.676	29.113	0.657	28.862	0.668	29.038	0.701
47; 1086648	26.213	0.844	26.493	0.799	26.216	0.826	26.522	0.904
80; 1562869	28.150	0.642	28.297	0.632	28.152	0.634	28.339	0.676
59; 947239	27.458	0.669	27.558	0.667	27.460	0.659	27.628	0.725
322; 3310865	28.009	0.339	28.086	0.342	28.010	0.338	28.079	0.343
83; 1949322	27.229	0.687	27.451	0.663	27.230	0.678	27.382	0.708
129; 2358615	26.690	0.534	26.803	0.534	26.692	0.530	26.871	0.552
45; 190472	28.444	1.131	30.324	1.089	28.447	1.106	28.703	1.259
240; 2524603	28.521	0.361	28.574	0.364	28.522	0.360	28.602	0.369
64; 776246	27.031	0.683	27.619	0.663	27.035	0.672	27.247	0.711
48; 648	67.075	3.471	70.775	2.458	67.076	3.385	67.845	3.518

NOTE: PM is the posterior mean; PSD is the posterior standard deviation. The first thirteen examples are from NHANES III and the fourteenth one is a data set on income (Aitkin 2010). DBM is the design-based method, EBM is the empirical Bayes method, EMM is the exact moment method and ABM is the approximate Bayesian method.

Table 2.2: Comparison of the approximate Bayesian method (ABM) and the full (exact) Bayesian method (FBM) for posterior inference of the finite population mean for fourteen examples

	ABM				FBM			
$n; N^{\dagger}$	PM	PSD	95% CI	Pval	PM	PSD	95% CI	Pval
25; 0.6	25.807	1.307	(23.317, 28.484)	.158	25.794	1.319	(23.233, 28.226)	.881
556; 4.5	28.131	0.275	(27.588, 28.658)	.982	28.155	0.279	(27.587, 28.665)	.996
162; 2.7	28.250	0.508	(27.296, 29.296)	.604	28.259	0.522	(27.239, 29.207)	.429
86; 2.0	29.038	0.701	(27.670, 30.392)	.731	29.056	0.708	(27.817, 30.472)	.450
47; 1.1	26.522	0.904	(24.792, 28.349)	.725	26.578	0.928	(24.842, 28.427)	.280
80; 1.6	28.339	0.676	(27.020, 29.676)	.512	28.358	0.713	(26.987, 29.714)	.491
59; 0.9	27.628	0.725	(26.302, 29.125)	.984	27.640	0.742	(26.231, 29.098)	.689
322; 3.3	28.079	0.343	(27.416, 28.753)	.930	28.070	0.361	(27.401, 28.820)	.818
83; 1.9	27.382	0.708	(25.967, 28.748)	.985	27.377	0.688	(26.014, 28.657)	.632
129; 2.4	26.871	0.552	(25.799, 27.962)	.956	26.873	0.561	(25.877, 27.967)	.903
45; 0.2	28.703	1.259	(26.198, 31.136)	.316	28.656	1.214	(26.087, 30.985)	.720
240; 2.5	28.602	0.369	(27.875, 29.323)	.984	28.606	0.372	(27.866, 29.294)	.660
64; 0.8	27.247	0.711	(25.827, 28.634)	.543	27.227	0.688	(25.930, 28.535)	.464
40; 644	67.845	3.518	(61.051, 74.903)	.410	67.868	3.558	(61.173, 75.169)	.763

NOTE: PM is the posterior mean; PSD is the posterior standard deviation; CI is the credible interval; Pval refers to the Kolmogorov test for normality. † Except for the last example N must be multiplied by 10^6; see the note to Table 2.1 for the exact population sizes. The procedure uses 10,000 draws from the approximate posterior density. The BMI data set has a single US state for females older than 45 years from NHANES III and the last example is on the income data (Aitkin 2010).

Table 2.3: Comparison of the times (hours) for the approximate Bayesian method (ABM) and the full (exact) Bayesian method (FBM) to perform the computations for the finite population mean by example

$n; N$	FBM
25; 608491	6.055
556; 4453263	44.311
162; 2704478	26.910
86; 1985501	19.756
47; 1086648	10.812
80; 1562869	15.551
59; 947239	9.425
322; 3310865	32.944
83; 1949322	19.396
129; 2358615	23.469
45; 190472	1.895
240; 2524603	25.120
64; 776246	7.724
48; 648	0.006

NOTE: The total time it took to compute all 14 examples just 8.8 seconds using the approximate Bayesian method (ABM). The computations to obtain the samples from the joint posterior density of μ, σ^2, α is common to both methods. The first thirteen examples are from NHANES III and the fourteenth one is a data set on income (Aitkin 2010).

Table 2.4: Summaries of different baseline distributions of the one-level Dirichlet process model

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

Table 2.5: Posterior inference of the finite population mean for body mass index (BMI) data using the Polya posterior, the Bayesian bootstrap and six baseline distributions

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

NOTE: PM is the posterior mean; PSD is the posterior standard deviation; NSE is the numerical standard error; CI is the credible interval. Each procedure uses 1,000 draws from the posterior density. The Polya posterior (PP) takes $\alpha = 0$ in the simple Dirichlet process and the Bayesian bootstrap (BB) uses Haldane prior for multinomial sampling. The BMI data are positively skewed. The BMI data set has a single US state for females older than 45 years, $N = 190,472$ and $n = 45$.

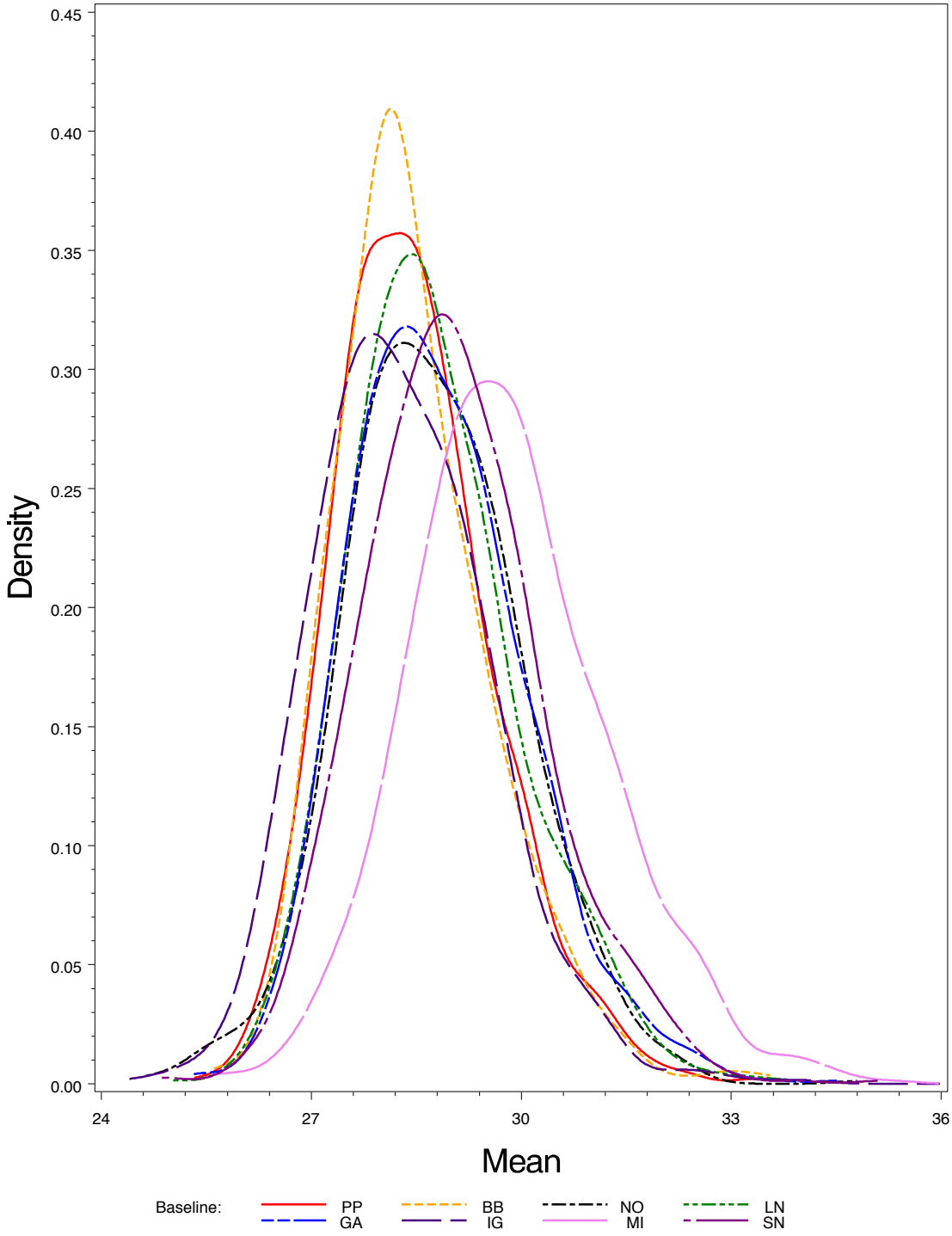

Figure 2.1: Plots of the posterior density of the finite population mean by baseline model for body mass index (BMI) data

Chapter 3

Two-level Dirichlet Process Models

In Chapter 3, we assume that data are obtained from a two-stage sample survey, for example, a two-stage cluster sampling, stratified or post-stratified sampling which is often seen in SAE problems. The sampled values are observed and the non-sampled 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 Scott-Smith model is the over-shrinkage, the mean of certain area maybe pooled too much toward the overall mean. Using the DP for the area mean 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 nonparametric procedure. Although presented in a survey sampling frame-

work, the proposed approach can be adapted to general random and mixed effect models.

In Section 3.1, we discuss the methodology and inferences of two-level DP models. In Section 3.2, we discuss the propriety of the posterior distributions. In Section 3.3, we discuss the prediction for the finite population when the DP is used for the sampling process. In Section 3.4, for model comparison, we discuss the computation of Bayes Factors. In Section 3.5, we discuss the results of the application to BMI data and simulated data.

3.1 Two-level Dirichlet Process Models

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.

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, also the finite population quantile. Let $n = \sum_{i=1}^{\ell} n_i$ be the total sample size and $N = \sum_{i=1}^{\ell} N_i$ be the total population size. Note that under simple random sampling, a design-based (direct) estimator of $\bar{Y}_i$ is $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i, i = 1, \ldots, \ell$; and we let $s_i^2 = \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(n_i - 1), i = 1, \ldots, \ell$. The estimation of the standard deviation of the design-based (direct) estimator is $\sqrt{(1 - f_i)s_i^2/n_i}$, where $f_i = n_i/N_i$ is the sampling fraction for each area.

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

$$\begin{aligned} y_{ij}|\nu_i &\overset{ind}{\sim} N\left(\theta + \nu_i, \sigma^2\right), \\ \nu_i &\overset{ind}{\sim} N\left(0, \delta^2\right), \end{aligned} \tag{3.1}$$

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 the $\underset{\sim}{\nu} = \{\nu_i, i = 1, \ldots, \ell\}$ 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,

$$\begin{aligned} y_{ij}|\mu_i &\overset{ind}{\sim} N\left(\mu_i, \sigma^2\right), \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, N_i, \\ \mu_i &\overset{ind}{\sim} N\left(\theta, \delta^2\right). \end{aligned} \tag{3.2}$$

Our two-level normal model (baseline parametric model) is then

$$y_{ij}|\mu_i \overset{ind}{\sim} N\left(\mu_i, \sigma^2\right), \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, N_i, \tag{3.3}$$

$$\mu_i \overset{ind}{\sim} N\left(\theta, \frac{\rho}{1-\rho}\sigma^2\right), \tag{3.4}$$

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

Here we consider a reparameterization of the Scott-Smith model (3.2) 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), Molina, Nandram and Rao (2014).

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

$$\begin{aligned}
\pi(\underset{\sim}{\mu}, \theta, \sigma^2, \rho | \underset{\sim}{y}_s) \quad \propto \quad & \left(\frac{1}{\sigma^2}\right)^{(n+\ell)/2} \left(\frac{1-\rho}{\rho}\right)^{\ell/2} \exp\Bigg\{ -\frac{1}{2\sigma^2} \Bigg\{ \sum_{i=1}^{\ell} \Big\{ (n_i - 1) s_i^2 \\
+ \quad & \left(n_i + \frac{1-\rho}{\rho}\right) (\mu_i - [\lambda_i \bar{y}_i + (1 - \lambda_i)\theta])^2 \\
+ \quad & \lambda_i \left(\frac{1-\rho}{\rho}\right) (\bar{y}_i - \theta)^2 \Big\} \Bigg\} \Bigg\} \times \frac{1}{(1+\sigma^2)^2} \times \frac{1}{\pi(1+\theta^2)}. \qquad (3.5)
\end{aligned}$$

We use a simple method called the simple important resampling (SIR) algorithm to draw from the posterior distribution $\pi(\underset{\sim}{\mu}, \theta, \sigma^2, \rho | \underset{\sim}{y}_s)$ (3.5). That is to take a simulated 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 $\pi(\underset{\sim}{\mu}, \theta, \sigma^2, \rho | \underset{\sim}{y}_s)$. The proposal model needs to be a rough approximation to the joint posterior density (3.5) and easy to draw samples from. We use the same likelihoods (3.3) and (3.4) in the two-level normal model together with an improper prior $\pi(\theta, \sigma^2, \rho) \propto \frac{1}{\sigma^2}, -\infty < \theta < \infty, 0 \leq \sigma^2 < \infty, 0 \leq \rho \leq 1$ as the proposal model, that is,

$$\begin{aligned}
\pi_a(\underset{\sim}{\mu}, \theta, \sigma^2, \rho | \underset{\sim}{y}_s) \quad \propto \quad & \pi_a(\underset{\sim}{\mu} | \theta, \sigma^2, \rho, \underset{\sim}{y}_s) \pi_a(\theta | \sigma^2, \rho, \underset{\sim}{y}_s) \pi_a(\sigma^2 | \rho, \underset{\sim}{y}_s) \pi_a(\rho | \underset{\sim}{y}_s) \qquad (3.6) \\
\propto \quad & \prod_{i=1}^{\ell} N\left[\mu_i; \lambda_i \bar{y}_i + (1 - \lambda_i)\theta, (1 - \lambda_i) \frac{\rho}{1-\rho} \sigma^2\right] \\
\times \quad & N\left(\theta; \tilde{y}, \frac{\sigma^2 \rho}{\sum_{i=1}^{\ell} \lambda_i (1-\rho)}\right) \times \mathrm{IG}\left[\sigma^2; (n-1)/2, A_1/2\right] \\
\times \quad & \frac{\Gamma[(n-1)/2]}{(A_1/2)^{(n-1)/2}} \prod_{i=1}^{\ell} (1 - \lambda_i)^{1/2} \left[\frac{\rho}{\sum_{i=1}^{\ell} \lambda_i (1-\rho)}\right]^{1/2}.
\end{aligned}$$

We draw a sample from the approximate joint posterior density (3.6) by first drawing

a sample from $\pi_a(\rho|\underset{\sim}{y}_s)$ using the grid method.

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

$$\begin{aligned} y_{ij}|G_i &\overset{ind}{\sim} G_i, \quad i=1,\ldots,\ell, \quad j=1,\ldots,N_i, \\ G_i|\mu_i &\overset{ind}{\sim} \mathrm{DP}\{\alpha_i, G_0(\mu_i)\}, \\ \mu_i|H &\overset{iid}{\sim} H, \\ H &\sim \mathrm{DP}\{\gamma, H_0(\cdot)\}, \end{aligned} \tag{3.7}$$

where $G_0(\cdot)$ and $H_0(\cdot)$ can be any parametric distributions. In particular, we consider $G_0 = N(\mu_i, \sigma^2)$ and $H_0 = N(\theta, \delta^2)$, where $\delta^2 = \frac{\rho}{1-\rho}\sigma^2$ in (3.7) 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,

$$\pi(\alpha_i) = \frac{1}{(\alpha_i+1)^2}, \qquad \alpha_i > 0, \quad i=1,\ldots,\ell, \tag{3.8}$$

$$\pi(\gamma) = \frac{1}{(\gamma+1)^2}, \qquad \gamma > 0, \tag{3.9}$$

$$\begin{aligned} \pi(\theta,\sigma^2,\rho) &= \frac{1}{\pi(1+\theta^2)}\frac{1}{(1+\sigma^2)^2}, \\ &\quad -\infty < \theta < \infty, 0 \le \sigma^2 < \infty, 0 \le \rho \le 1, \end{aligned} \tag{3.10}$$

with independence. We call (3.7), (3.8), (3.9) and (3.10) together the two-level Dirichlet process (DPDP) model. Note that the concentration parameters α_i and γ are not included in the two-level normal model.

The inference of the DPDP model can be easily performed. The idea is similar to the one-level DP model in Chapter 2. 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 which 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

$$\pi(\underset{\sim}{\alpha}, \underset{\sim}{\psi} \mid \underset{\sim}{k}, \underset{\sim}{y}^*) = \left[\prod_{i=1}^{\ell} \pi(\alpha_i \mid k_i)\right] \pi(\underset{\sim}{\psi} \mid \underset{\sim}{y}^*), \tag{3.11}$$

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) algorithm can be used easily to obtain posterior samples of $\underset{\sim}{\mu}$ and γ. We need to add a few steps in the Gibbs sampler to draw 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 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 $\underset{\sim}{\psi}$. 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. $\pi(u_i|\dots) \propto \mathbf{1}(0 < u_i < \xi_{d_i})$.

2. $\pi(\mu_s^*|\dots) \propto N(\mu_s^*; \theta, \delta^2) \prod_{\{i|d_i=s\}} \prod_{j=1}^{k_i} N(y_{ij}^*; \mu_s^*, \sigma^2)$.

3. $\pi(v_s|\dots) \propto \text{Beta}(a_s, b_s)$, where

$a_s = 1 + \sum_{i=1}^{\ell} \mathbf{1}(d_i = s)$ and $b_s = \gamma + \sum_{i=1}^{\ell} \mathbf{1}(d_i > s)$.

4. $\pi(\gamma|\dots) \propto \gamma^{k_0} \frac{\Gamma(\gamma)}{\Gamma(\gamma+\ell)} \frac{1}{(\gamma+1)^2}$, where k_0 is the number of distinct $d_1, \dots, d_\ell$.

5. $P(d_i = t|\dots) \propto \mathbf{1}(t : \xi_t > u_i) p_t/\xi_t \prod_{j=1}^{k_i} N(y_{ij}^*; \mu_t^*, \sigma^2)$, $t = 1, \dots, K$.

6. $\pi(\sigma^2, \theta, \rho|\dots) \propto \prod_{i=1}^{\ell} \prod_{j=1}^{k_i} N(y_{ij}^*; \mu_{d_i}^*, \sigma^2) \times \prod_{s=1}^{K} N(\mu_s^*; \theta, \delta^2) \times \frac{1}{\pi(1+\theta^2)} \frac{1}{(1+\sigma^2)^2}$.

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,

$$\begin{aligned} y_{ij}|\mu_i &\overset{ind}{\sim} N\left(\mu_i, \sigma^2\right), \quad i = 1, \dots, \ell, \quad j = 1, \dots, N_i, \\ \mu_i|H &\overset{iid}{\sim} H, \\ H &\sim \text{DP}\{\gamma, N(\theta, \delta^2)\}, \end{aligned} \tag{3.12}$$

together with (3.9) and (3.10) gives us the DPM model which is easy to fit.

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

$$\begin{aligned} y_{ij}|G_i &\overset{ind}{\sim} G_i, \quad i = 1, \dots, \ell, \quad j = 1, \dots, N_i, \\ G_i|\mu_i &\overset{ind}{\sim} \text{DP}\{\alpha_i, N\left(\mu_i, \sigma^2\right)\}, \\ \mu_i &\overset{iid}{\sim} N(\theta, \delta^2). \end{aligned} \tag{3.13}$$

We call (3.13), (3.8) and (3.10) the DP normal (DPnormal) model. The algorithm for the DPnormal 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,

$$\begin{aligned} y^*_{ij}|\mu_i &\overset{ind}{\sim} N\left(\mu_i, \sigma^2\right), \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, k_i, \\ \mu_i &\overset{iid}{\sim} N\left(\theta, \frac{\rho}{1-\rho}\sigma^2\right), \\ \pi(\theta, \sigma^2, \rho) &= \frac{1}{\pi(1+\theta^2)} \frac{1}{(1+\sigma^2)^2}, -\infty < \theta < \infty, 0 \leq \sigma^2 < \infty, 0 \leq \rho \leq 1. \end{aligned} \tag{3.14}$$

3.2 Propriety of the Posterior Distributions

Lemma 3.2.1 *The joint posterior density $\pi(\underset{\sim}{\mu}, \theta, \sigma^2, \rho|\underset{\sim}{y}_s)$ (3.5) under the two-level normal model is proper if $\ell \geq 2$.*

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).

We restate the Lemma 2 in Lo (1984) in our notation in order to prove the propriety of the posterior distributions. Let m be a positive integer and $g_i, i = 1, \ldots, \ell$ positive functions. Let $\boldsymbol{P}$ be a partition of $1, \ldots, \ell$ and $N(\boldsymbol{P})$ be the number of cells in the partition. Thus, $\boldsymbol{P} = \{C_i, i = 1, \ldots, N(\boldsymbol{P}\}$, where C_i is the ith cell of the partition. Let e_i be the number of elements in C_i. Note that C_i and $e_i, i = 1, \ldots, N(\boldsymbol{P})$ depend on $\boldsymbol{P}$.

Lemma 3.2.2

$$\int_{\mathbb{R}^\ell} \prod_{i=1}^{\ell} g_i(\mu_i) dG_0(\mu_1) \prod_{i=2}^{\ell} d\left\{\frac{\alpha G_0(\mu_i) + \sum_{j=1}^{i-1} \delta_{\mu_j}(\mu_i)}{\alpha + i - 1}\right\} = \left[\prod_{i=1}^{\ell} \frac{1}{\alpha + i - 1}\right] \sum_{\boldsymbol{P}} \phi(\boldsymbol{P})$$

where

$$\phi(\boldsymbol{P}) = \prod_{i=1}^{N(\boldsymbol{P})} \left\{ (e_i - 1)! \int_{\mathbb{R}} \left[\prod_{c \in C_i} g_c(\mu) \right] \alpha g_0(\mu) d\mu \right\}.$$

Theorem 3.2.3 *If the posterior density under the normal baseline model is proper, the posterior density under the DPM model is proper.*

Proof: Let $\prod_{j=1}^{n_i} N(y_{ij}|\mu_i, \sigma^2) = g_i(\mu_i; \sigma^2)$ for $i = 1, \ldots, \ell$ and $e = \max_{i=1}^{\ell}\{(e_i - 1)!\}$. We have

$$\int \prod_{i=1}^{\ell} g_i(\mu_i; \sigma^2) N(\mu_1; \theta, \delta^2) d\mu_1 \times \prod_{i=2}^{\ell} \left(\frac{\gamma}{\gamma + i - 1} N(\mu_i; \theta, \delta^2) \right.$$

$$\left. + \frac{1}{\gamma + i - 1} \sum_{s=1}^{i-1} \delta_{\mu_s}(\mu_i) \right) d\mu_i \times \frac{1}{(\gamma+1)^2} \times \pi(\theta, \sigma, \rho) d\theta d\sigma d\rho d\gamma$$

$$\begin{aligned}
&= \int \left[\prod_{i=1}^{\ell} \frac{1}{\gamma + i - 1} \right] \sum_{\boldsymbol{P}} \prod_{i=1}^{N(\boldsymbol{P})} \left\{ (e_i - 1)! \int_{\mathbb{R}} \prod_{c \in C_i} [\gamma g_c(\mu; \sigma^2)] N(\mu; \theta, \sigma^2) d\mu \right\} \\
&\times \frac{1}{(\gamma+1)^2} \times \pi(\theta, \sigma, \rho) d\theta d\sigma d\rho d\gamma \\
&< e \sum_{\boldsymbol{P}} \int \prod_{i=1}^{N(\boldsymbol{P})} \left\{ \int_{\mathbb{R}} \left[\prod_{c \in C_i} g_c(\mu; \sigma^2) \right] N(\mu; \theta, \sigma^2) d\mu \right\} \pi(\theta, \sigma, \rho) d\theta d\sigma d\rho < \infty.
\end{aligned}$$

The integral within the summation is finite because it is the marginal distribution of regrouped data for one particular partition under the baseline model which is proper.

Theorem 3.2.4 *If the posterior density under the normal baseline model is proper, the posterior density under the DPDP model is proper.*

Proof: We need to show that

$$\begin{aligned}\int \quad & \prod_{i=1}^{\ell}\left\{N(y_{i1};\mu_1,\sigma^2)\prod_{j=2}^{n_i}\left[\frac{\alpha_i}{\alpha_i+j-1}N(y_{ij};\mu_i,\sigma^2)+\frac{1}{\alpha_i+j-1}\sum_{s=1}^{j-1}\delta_{y_{is}}(y_{ij})\right]\right\} \\ \times \quad & N(\mu_1;\theta,\delta^2)d\mu_1\prod_{i=2}^{\ell}\left[\frac{\gamma}{\gamma+i-1}N(\mu_i;\theta,\delta^2)+\frac{1}{\gamma+i-1}\sum_{s=1}^{i-1}\delta_{\mu_s}(\mu_i)\right]d\mu_i \\ \times \quad & \left[\prod_{i=1}^{\ell}\frac{1}{(\alpha_i+1)^2}\right]\frac{1}{(\gamma+1)^2}\times\pi(\theta,\sigma,\rho)d\underset{\sim}{\alpha}d\theta d\sigma d\rho d\gamma<\infty.\end{aligned}$$

Following the same arguments in Theorem 2.2.1, we now only need to show that

$$\begin{aligned}\int \quad & \prod_{i=1}^{\ell}\prod_{j=1}^{n_i}N(y_{ij}|\mu_i,\sigma^2)\times N(\mu_1;\theta,\delta^2)d\mu_1 \\ \times \quad & \prod_{i=2}^{\ell}\left(\frac{\gamma}{\gamma+i-1}N(\mu_i;\theta,\delta^2)+\frac{1}{\gamma+i-1}\sum_{s=1}^{i-1}\delta_{\mu_s}(\mu_i)\right)d\mu_i \\ \times \quad & \left[\prod_{i=1}^{\ell}\frac{1}{(\alpha_i+1)^2}\right]\frac{1}{(\gamma+1)^2}\times\pi(\theta,\sigma,\rho)d\underset{\sim}{\alpha}d\theta d\sigma d\rho d\gamma<\infty,\end{aligned}$$

which is shown in Theorem 3.2.3.

Theorem 3.2.5 *If the posterior density under the normal baseline model is proper, the posterior density under the DPnormal model is proper.*

Proof: Following similar arguments it is clear that DPnormal model is also proper.

3.3 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 prediction intervals for the finite population mean $\bar{Y}_i$ for each area. The sampling

process is as,

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

As we discuss in Chapter 2, the predictive inference for the two-level DP model follows the same way in the one-level DP model for each i, since all areas are independent. Also it is essentially the same methodology for the three-level models discussed in Chapter 4.

3.4 Bayes Factor

Let $\Omega = (\underset{\sim}{\mu}, \theta, \sigma^2, \rho, \gamma, \underset{\sim}{\alpha})$ and $\Omega' = (\theta, \sigma^2, \rho, \gamma, \underset{\sim}{\alpha})$. As in the methodology discussed for Bayes factors in Chapter 1, we can write the marginal likelihood as,

$$M(\underset{\sim}{y}_s) = \int f(\underset{\sim}{y}_s|\Omega)\pi(\Omega)d\Omega$$

$$= \int \left\{ \frac{f(\underset{\sim}{y}_s|\Omega)\pi(\Omega)}{\pi_a(\Omega|\underset{\sim}{y}_s)} \right\} \pi_a(\Omega|\underset{\sim}{y}_s)d\Omega \Big/ \int \left\{ \frac{\pi(\Omega)}{\pi_a(\Omega|\underset{\sim}{y}_s)} \right\} \pi_a(\Omega|\underset{\sim}{y}_s)d\Omega.$$

In this section, the formula we need for each model is given. It is easy to write down the likelihood function of a parametric distribution. For the DP, we need to use the Polya urn scheme by integrating out this random measure to obtain a closed-form for $f(\underset{\sim}{y}_s|\Omega), \pi(\Omega)$ and $\pi_a(\Omega|\underset{\sim}{y}_s)$. Thus, we discuss in detail for the DPDP model and summarize the equations for other models in Table 3.1.

For each i, integrating out G, it is easy to write the likelihood function

$f^{\mathrm{DPDP}}(\underset{\sim}{y}_s|\Omega)$ and prior $\pi^{\mathrm{DPDP}}(\Omega)$, where

$$\begin{aligned} f^{\mathrm{DPDP}}(\underset{\sim}{y}_s|\Omega) &= \prod_{i=1}^{\ell}\left\{N(y_{i1};\mu_1,\sigma^2)\prod_{j=2}^{n_i}\left(\frac{\alpha_i}{\alpha_i+j-1}N(y_{ij};\mu_i,\sigma^2)\right.\right. \\ &+ \left.\left.\frac{1}{\alpha_i+j-1}\sum_{s=1}^{j-1}\delta_{y_{is}}(y_{ij})\right)\right\}, \end{aligned}$$

and

$$\begin{aligned} \pi^{\mathrm{DPDP}}(\Omega) &= \pi(\underset{\sim}{\mu}|\theta,\rho,\sigma^2,\gamma)\pi(\theta,\rho,\sigma^2)\left[\prod_{i=1}^{\ell}\pi(\alpha_i)\right]\pi(\gamma) \\ &= N(\mu_1;\theta,\delta^2)\prod_{i=2}^{\ell}\left[\frac{\gamma}{\gamma+i-1}N(\mu_i;\theta,\delta^2)+\frac{1}{\gamma+i-1}\sum_{s=1}^{i-1}\delta_{\mu_s}(\mu_i)\right] \\ &\times \left[\prod_{i=1}^{\ell}\frac{1}{(\alpha_i+1)^2}\right]\frac{1}{(\gamma+1)^2}\frac{1}{\pi(1+\theta^2)}\frac{1}{(1+\sigma^2)^2}. \end{aligned}$$

It is a little bit involved for $\pi_a^{\mathrm{DPDP}}(\Omega|\underset{\sim}{y}_s)$, since $\mu_1,\ldots,\mu_\ell$ are correlated. We use $\pi(\mu_i|\mu_{i-1},\ldots,\mu_1)$ as an approximate of $\pi(\mu_i)$. Thus the approximate posterior $\pi_a^{\mathrm{DPDP}}(\Omega|\underset{\sim}{y}_s)$ is equal to

$$\left[\prod_{i=2}^{\ell}\pi(\mu_i|\mu_{i-1},\ldots,\mu_1,\Omega',\underset{\sim}{y}_s)\right]\pi(\mu_1|\Omega',\underset{\sim}{y}_s)\pi_a(\theta,\sigma^2,\rho|\underset{\sim}{y}_s)\pi_a(\gamma|k)\prod_{i=1}^{\ell}\pi_a(\alpha_i|k_i),$$

where

$$\pi(\mu_1|\Omega',\underset{\sim}{y}_s)=N\left(\mu_1;\lambda_1\bar{y}_1+(1-\lambda_1)\theta,(1-\lambda_1)\frac{\rho\sigma^2}{(1-\rho)}\right),$$

and for $i = 2, \ldots, \ell$, $\pi_a(\mu_i|\mu_{i-1}, \ldots, \mu_1, \Omega', \underset{\sim}{y}_s)$

$$
\begin{aligned}
&= \frac{1}{i-1+\gamma}\sum_{s=1}^{i-1}\left\{\left(\frac{1}{2\pi\sigma^2}\right)^{n_i/2}\exp\left\{-\frac{1}{2\sigma^2}\Big[n_i(\bar{y}_i-\mu_s)^2\right.\right.\\
&+ \left.\left.(n_i-1)s_i^2\Big]\right\}\delta_{\mu_s}(\mu_i)\right\} + \frac{\gamma}{i-1+\gamma}\left(\frac{1}{2\pi\sigma^2}\right)^{n_i/2}\\
&\times (1-\lambda_i)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}\left[(1-\lambda_i)n_i(\bar{y}_i-\theta)^2+(n_i-1)s_i^2\right]\right\}\\
&\times N\left(\mu_i; \lambda_i\bar{y}_i+(1-\lambda_i)\theta, (1-\lambda_i)\frac{\rho\sigma^2}{(1-\rho)}\right).
\end{aligned}
$$

For the concentration parameter γ, taking the transformation $\rho_\gamma = 1/(\gamma+1)$, we can compute a_γ and b_γ, the MLEs of parameters in the Beta distribution by using posterior samples of γ. We have $\pi_a(\gamma|k) = \gamma^{(b_\gamma-1)}/[(\gamma+1)^{(a_\gamma+b_\gamma)}B(a_\gamma, b_\gamma)]$, where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. For α_i, we proceed in the similar way by transformation $\rho_{\alpha_i} = 1/(\alpha_i+1)$ for each i. And $\pi_a(\theta, \sigma^2, \rho|\underset{\sim}{y}_s)$ is the same as in the baseline model.

Table 3.1 gives the equations for the computation of Bayes factors for normal model, DPM model and DPnormal model.

3.5 Empirical Studies

3.5.1 Application to Body Mass Index (BMI) Data

We first fit the two-level models by collapsing the sub-domains formed by age, race and sex to obtain the population mean for each county. 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 two-level DP models. We also use a Bayesian bootstrap, which is discussed in Chapter 1, without borrowing across counties as a comparison.

Note that for the county level, all sample sizes are over 100. But we have a SAE problem when it comes to the sub-domains. 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 run 10000 MCMC iterations, burn in 5000 and thin every 5th to obtain 1000 converged posterior samples. Table 3.2 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 test which is that the Markov chain is in the stationary distribution. And effective sample sizes are not too far away from 1000. These numerical summaries, trace plots, and autocorrelation plots indicate that the MCMC chains converge.

Tables 3.3, 3.4 and 3.5 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. These tables show that roughly similar results 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 3.3 shows that roughly half of the counties with smaller PSD under the DPDP model than the normal model. And PMs under the DPDP model are closer to the PMs under the Bayesian bootstrap which is considered as an unbiased estimator. Meanwhile, PMs under the normal model are pooled toward to 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. We examine several plots to further compare results of BMI data.

The predictive inference of the finite population mean, 85^{th} and 95^{th} percentile for each county by four different models (normal, DPM, DPnormal and DPDP mod-

els) are compared respectively. Figures 3.1, 3.2 and 3.3 plot posterior means with credible bands versus direct estimates for BMI data. In Figure 3.1, the posterior means are very similar under the normal, DPM and DPnormal models. They are pooled toward the overall means. The posterior means under the DPDP model are closer to the direct estimators (less pooling) meanwhile with similar credible bands comparing to other models. Some evidence of the advantage of the nonparametric alternative, DPDP model, when predicting population means are presented. Without the restrictive parametric assumptions, the DPDP model tends to provide less biased estimation with similar variation comparing to the other candidate models. The predictive inferences of the population percentile are similar under the normal and DPM model. However, the predictive inference of the population percentile is not so good under the DPDP and DPnormal model. We suspect that it may be due to the discreteness of the DP when it is used as sampling process.

Figures 3.4, 3.5, and 3.6 are plots of the posterior density of the finite population mean, 85^{th} and 95^{th} percentiles for the four models (normal, DPM, DPnormal, DPDP models) and Bayesian bootstrap for the first eight counties of BMI data. We show these density plots as examples to further confirm our observations. In Figure 3.4, for the population mean, most parts of the density under the normal, DPM and DPnormal models are similar and the DPnormal model have slightly smaller variation. The results from the DPDP model are close to the unbiased estimation under the Bayesian bootstrap with smaller variation. However, the DPDP model does not always have the smallest variation, since in general one expects a more flexible model will have larger variability. Figures 3.5 and 3.6 show that the estimated density of the population 85^{th} and 95^{th} percentiles under the DPnormal and DPDP model are not smooth and the estimated density of the population 85^{th} and 95^{th} percentiles under the normal and DPM model are similar. Other counties have

similar phenomenon which is not shown here.

Several comparison measurements are also computed. Table 3.6 gives the log of the marginal likelihood (LML) with Monte Carlo errors, log pseudo marginal likelihood (LPML), delete-one cross validation (CV) divergence measure, deviance information criterion (DIC) and percentages of conditional predictive ordinate (CPO) less than .025 ($P_{CPO} < .025$) and .014 ($P_{CPO} < .014$) of each two-level model for BMI data. CV of four models are comparable. And the differences among the percentages of CPO less than .025 and .014 in these models are very small. These comparison measurements suggest choosing parametric baseline model. However, as we discussed in Chapter 1, when the parametric model is nested in the nonparametric alternative, the Bayes factor may be misleading. Intuitively any likelihood-based diagnostic will be misleading because we are comparing infinite dimensional distributions.

Since BMI data suffers right skewness with outliers in the right tails, ties and gaps, 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 DPDP model.

3.5.2 Simulation

We conduct a simple simulation study. We have simulated three data sets to fit the normal model (that is, the Scott-Smith model), the DPM model, the DP normal (DPnormal) model and the two-level DP (DPDP) model respectively. We simulated data from the normal model, the DPM model with $\gamma = 0.5$ and the DPDP model with $\alpha = 0.3$ and $\gamma = 0.5$.

Figures 3.7, 3.8 and 3.9 show the comparison of posterior means with credible bands and true population means for the simulated normal, DPM and DPDP data

under four different models (normal, DPM, DPnormal and DPDP models). We can see that the results are similar, all close to the true population mean. Table 3.7 gives Log of the marginal likelihood with Monte Carlo errors, Log pseudo marginal likelihood (LPML) and delete-one cross validation (CV) divergence measure of each model for each simulated data set.

The simulation examples show some evidence that the nonparametric method performs well for the predictive inference of the population mean. We may want to conduct more extensive simulation study on repeated simulated data. However, this process is time consuming because parallel computing in R is needed and is not well developed.

Table 3.1: The equations for the computation of Bayes factors for normal model, DPM model and DPnormal model

	Normal Model
$f(\underset{\sim}{y}_s\|\Omega)$	$\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\prod_{i=1}^{\ell}(1-\lambda_i)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{\ell}\left(\lambda_i\left(\frac{1-\rho}{\rho}\right)(\bar{y}_i-\theta)^2+(n_i-1)s_i^2\right)\right]\right\}$.
$\pi(\Omega)$	$\frac{1}{\pi(1+\theta^2)}\frac{1}{(1+\sigma^2)^2}$.
$\pi_a(\Omega\|\underset{\sim}{y}_s)$	$\mathrm{N}(\theta;\tilde{y},\frac{\rho\sigma^2}{(1-\rho)\sum_{i=1}^{\ell}\lambda_i})\mathrm{IG}\left\{\sigma^2;(n-1)/2,\left[\sum_{i=1}^{\ell}\left(\lambda_i\left(\frac{1-\rho}{\rho}\right)(\bar{y}_i-\tilde{y})^2+(n_i-1)s_i^2\right)\right]/2\right\}$ $\times\mathrm{Beta}(\rho;a,b)$.
Remarks:	We can integrate out $\underset{\sim}{\mu}$. $\tilde{y}=\sum_{i=1}^{\ell}\lambda_i\bar{y}_i/\sum_{i=1}^{\ell}\lambda_i$ and parameters a and b are the MLEs by using posterior samples of ρ to fit a beta distribution.
	DPM Model
$f(\underset{\sim}{y}_s\|\Omega)$	$\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}\left\{n_i(\bar{y}_i-\mu_i)^2+(n_i-1)s_i^2\right\}\right\}$.
$\pi(\Omega)$	$N(\mu_1;\theta,\delta^2)\prod_{i=2}^{\ell}\left(\frac{\gamma}{\gamma+i-1}N(\mu_i;\theta,\delta^2)+\frac{1}{\gamma+i-1}\sum_{s=1}^{i-1}\delta_{\mu_s}(\mu_i)\right)\frac{1}{(\gamma+1)^2}\frac{1}{\pi(1+\theta^2)}\frac{1}{(1+\sigma^2)^2}$.
$\pi_a(\Omega\|\underset{\sim}{y}_s)$	$\left[\prod_{i=2}^{\ell}\pi(\mu_i\|\mu_{i-1},\ldots,\mu_1,\Omega',\underset{\sim}{y}_s)\right]\pi(\mu_1\|\Omega',\underset{\sim}{y}_s)\pi_a(\theta,\sigma^2,\rho\|\underset{\sim}{y}_s)\pi_a(\gamma\|k)$.
Remarks:	The computation of $\pi_a(\Omega\|\underset{\sim}{y}_s)$ proceeds in the same manner as in the DPDP model excluding $\underset{\sim}{\alpha}$.
	DPnormal Model
$f(\underset{\sim}{y}_s\|\Omega)$	$f^{\mathrm{DPDP}}(\underset{\sim}{y}_s\|\Omega)$.
$\pi(\Omega)$	$\prod_{i=1}^{\ell}N(\mu_i;\theta,\delta^2)\prod_{i=1}^{\ell}\frac{1}{(\alpha_i+1)^2}\frac{1}{\pi(1+\theta^2)}\frac{1}{(1+\sigma^2)^2}$.
$\pi_a(\Omega\|\underset{\sim}{y}_s)$	$\pi(\underset{\sim}{\mu}\|\theta,\rho,\sigma^2,\underset{\sim}{y}_s)\pi_a(\theta\|\sigma^2,\rho,\underset{\sim}{y}_s)\pi_a(\sigma^2\|\rho,\underset{\sim}{y}_s)\pi_a(\rho\|\underset{\sim}{y}_s)\prod_{i=1}^{\ell}\pi_a(\alpha_i\|k_i)$, where $\pi(\underset{\sim}{\mu}\|\theta,\rho,\sigma^2,\underset{\sim}{y}_s)=\prod_{i=1}^{\ell}N[\mu_i;\lambda_i\bar{y}_i+(1-\lambda_i)\theta,(1-\lambda_i)\rho\sigma^2/(1-\rho)]$.
Remarks:	$\pi_a(\theta\|\sigma^2,\rho,\underset{\sim}{y}_s)$, $\pi_a(\sigma^2\|\rho,\underset{\sim}{y}_s)$ and $\pi_a(\rho\|\underset{\sim}{y}_s)$ are same as normal model with $\underset{\sim}{y}^*$ replacing $\underset{\sim}{y}_s$ and $\pi_a(\alpha_i\|k_i)$ same as DPDP model.

Table 3.2: 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	0.4831612	0.4140493	0.4592166	0.6196973
DPDP	0.5221358	0.6755549	0.7519071	0.1104736
	effective sample sizes			
Model	σ^2	θ	δ^2	γ
DPM	1000	1000	697.8087	1084.5006
DPDP	1000	938.2965	626.9378	732.3416

Table 3.3: 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 3.4: 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 3.5: 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 3.6: Log of the marginal likelihood (LML) with Monte Carlo errors , Log pseudo marginal likelihood (LPML), delete-one cross validation (CV) divergence measure, deviance information criterion (DIC) and percentages of conditional predictive ordinate (CPO) less than .025 ($P_{CPO<.025}$) and .014 ($P_{CPO<.014}$) of each two-level model for body mass index (BMI) data

	BMI data (two-level models)					
	LML	LPML	$P_{CPO<.025}$	$P_{CPO<.014}$	CV	DIC
Normal	$-9288.92_{0.01}$	-9292.26	0.0265	0.0197	0.7652	18583.2
DPM	$-9292.76_{0.08}$	-9445.15	0.0273	0.0200	0.7655	18588.3
DPnormal	$-32177.71_{3.40}$	-16110.97	0.0601	0.0288	0.7719	26686.5
DPDP	$-32348.29_{4.01}$	-16136.36	0.0743	0.0397	0.7721	26686.0

Table 3.7: Log of the marginal likelihood with Monte Carlo errors, Log pseudo marginal likelihood (LPML) and delete-one cross validation (CV) divergence measure of each model for each simulated data set. (DPM data: $\gamma = 0.5$; DPDP data: $\alpha = 0.3, \gamma = 0.5$)

	(a) Log of the marginal likelihood			
	Normal model	DPM model	DPnormal model	DPDP model
Normal data	-7136.083	-7135.931	-7141.158	-7180.218
	(8.973×10^{-7})	(0.1800)	(0.0010)	(40.3708)
DPM data	-7161.715	-7151.729	-7162.483	-7246.303
	(2.376×10^{-5})	(0.3162)	(0.0008)	(73.5941)
DPDP data	-3805.430	-3811.510	-2840.229	-2838.449
	(0.0280)	(0.0358)	(0.0008)	(0.2113)
	(b) LPML			
	Normal model	DPM model	DPnormal model	DPDP model
Normal data	-7146.061	-7176.803	-7149.160	-7179.017
DPM data	-7171.468	-7155.752	-7174.504	-7157.872
DPDP data	-3821.925	-3886.685	-2683.769	-2683.673
	(c) CV			
	Normal model	DPM model	DPnormal model	DPDP model
Normal data	0.4334	0.4350	0.4335	0.4351
DPM data	0.4339	0.4332	0.4340	0.4332
DPDP data	0.1703	0.1767	0.1703	0.1703

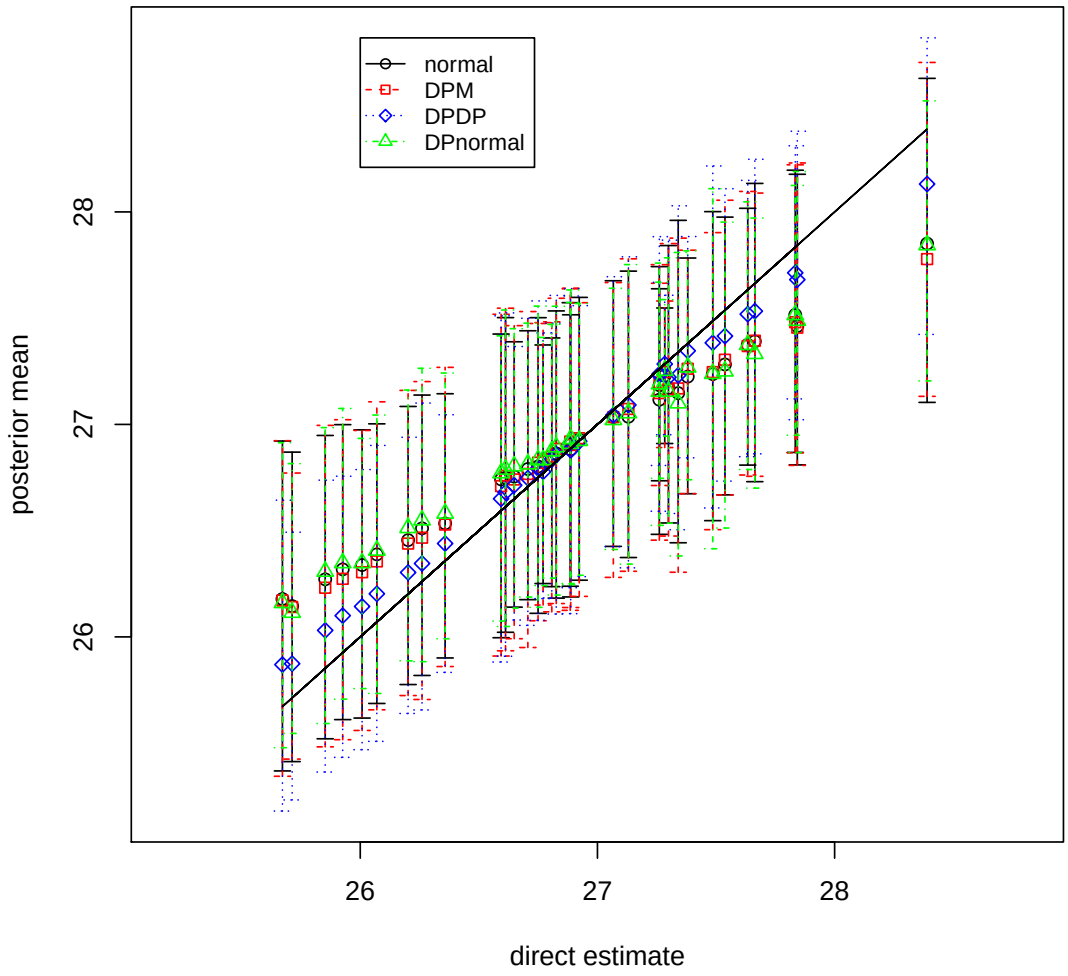

Figure 3.1: Comparison for body mass index (BMI) data (posterior means with credible bands versus direct estimates): the predictive inference of the finite population mean for each county under four different models (normal, DPM, DPnormal and DPDP models)

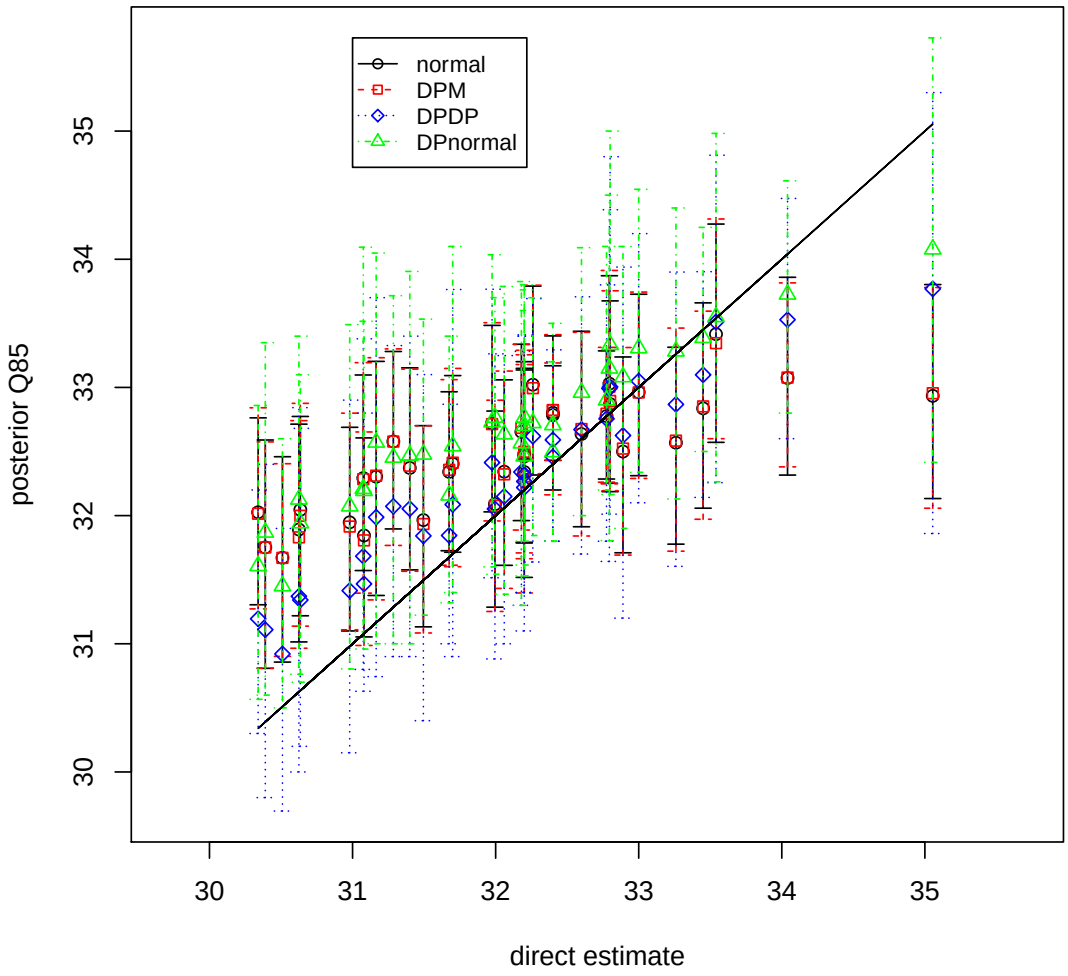

Figure 3.2: Comparison for body mass index (BMI) data (posterior means with credible bands versus direct estimates): the predictive inference of the finite population 85^{th} percentile for each county under four different models (normal, DPM, DPnormal and DPDP models)

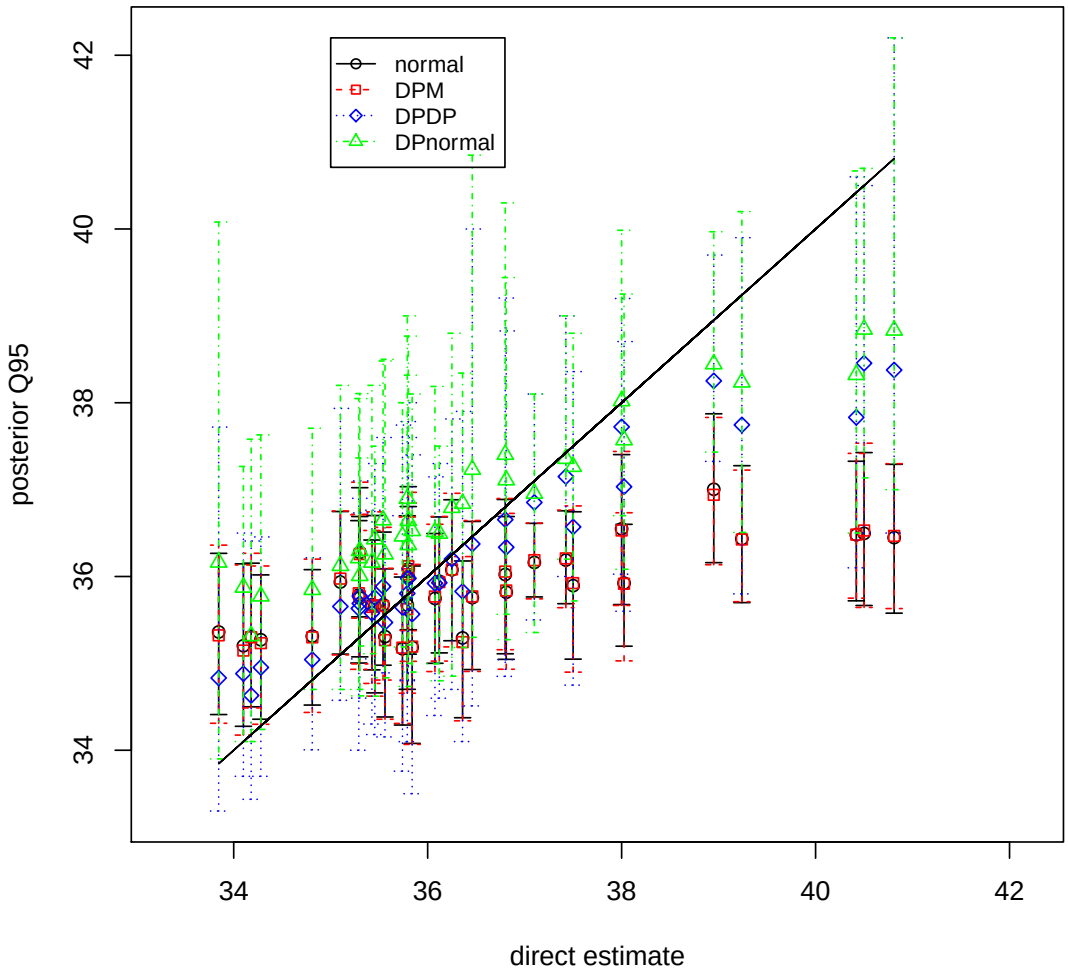

Figure 3.3: Comparison for body mass index (BMI) data (posterior means with credible bands versus direct estimates): the predictive inference of the finite population 95^{th} percentile for each county under four different models (normal, DPM, DPnormal and DPDP models)

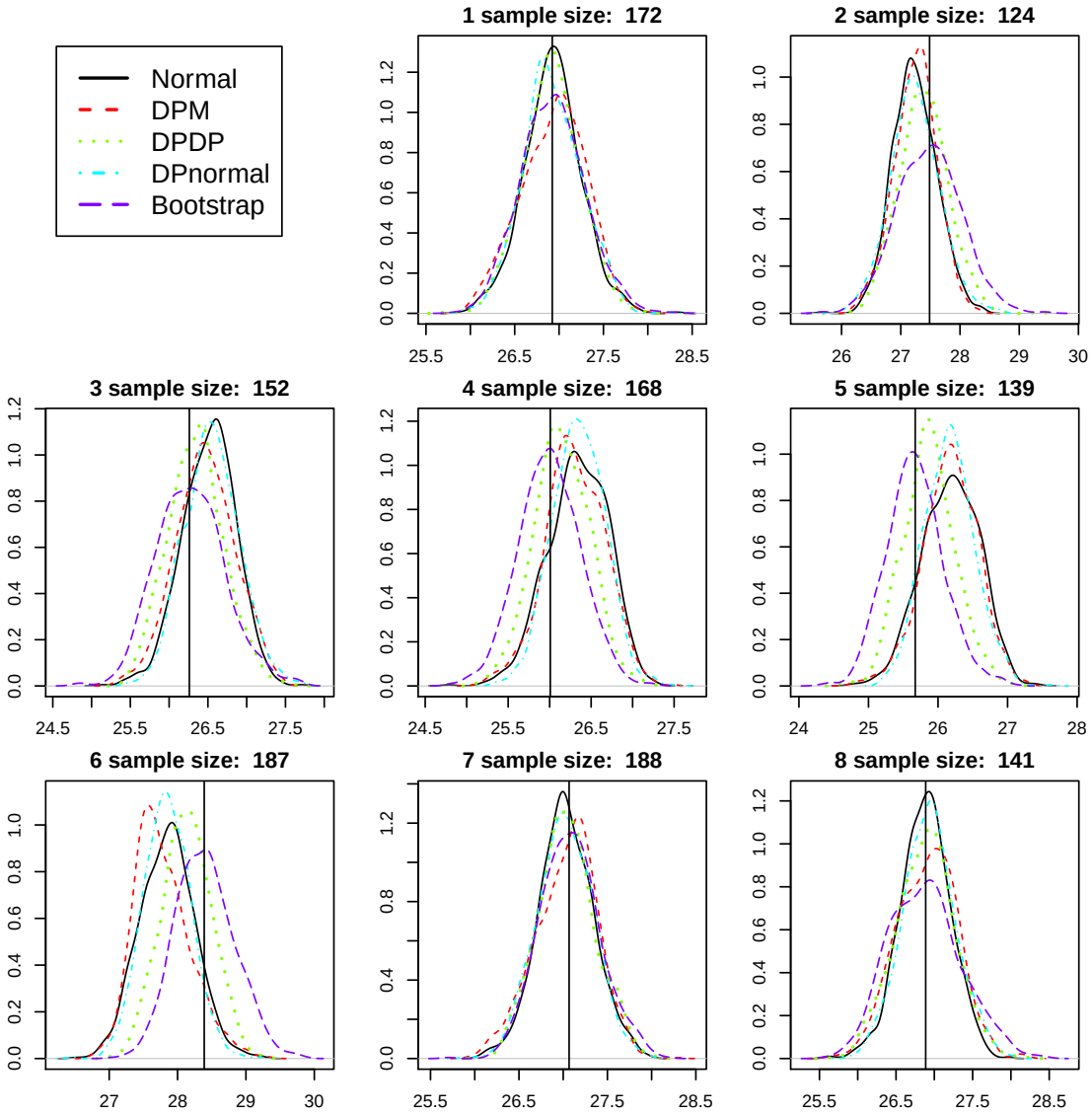

Figure 3.4: Plots of the posterior density of the finite population mean by four models (normal, DPM, DPnormal, DPDP models) and Bayesian bootstrap for the first eight counties of body mass index (BMI) data

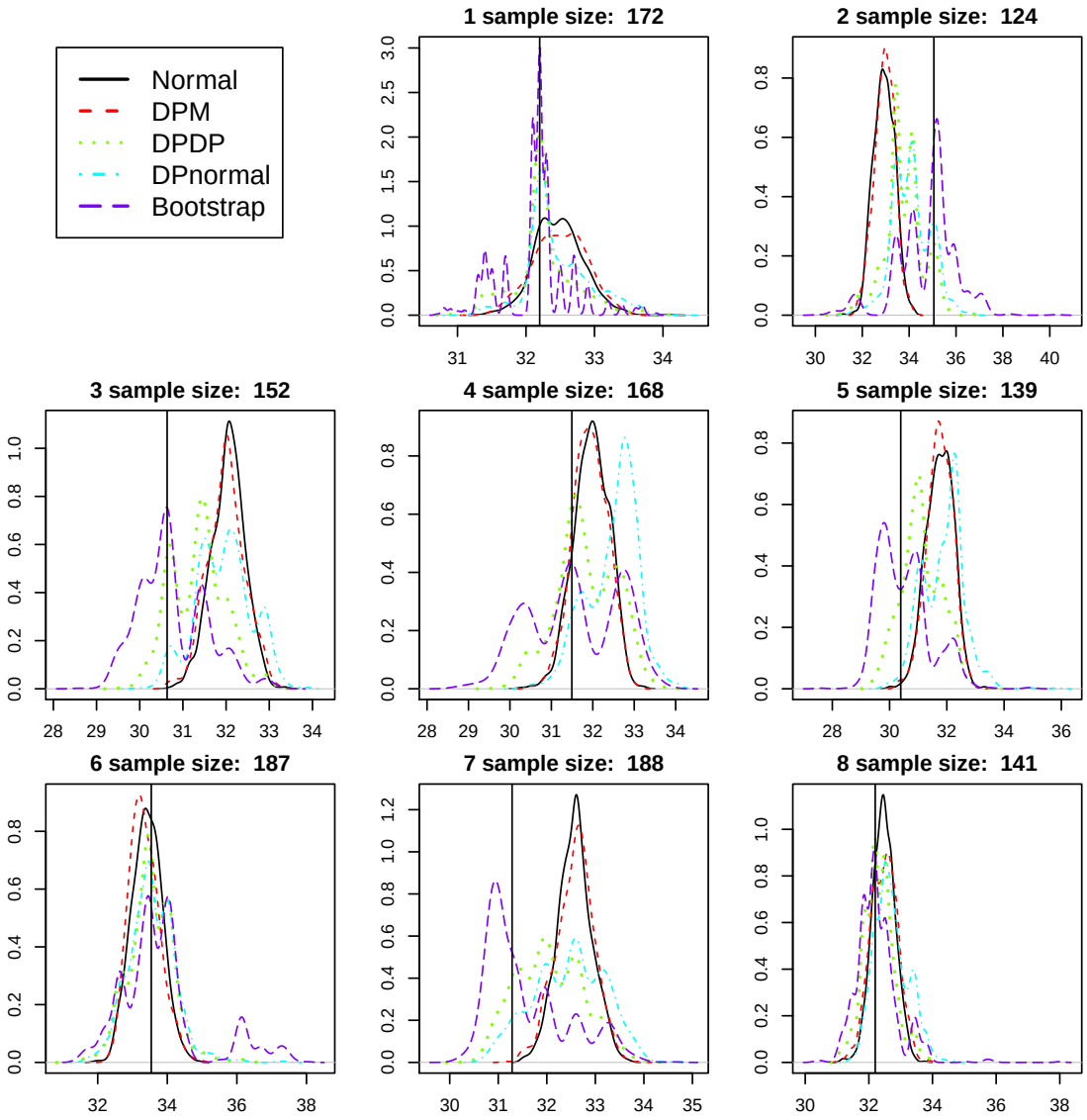

Figure 3.5: Plots of the posterior density of the finite population 85^{th} percentile by four models (normal, DPM, DPnormal, DPDP models) and Bayesian bootstrap for the first eight counties of body mass index (BMI) data

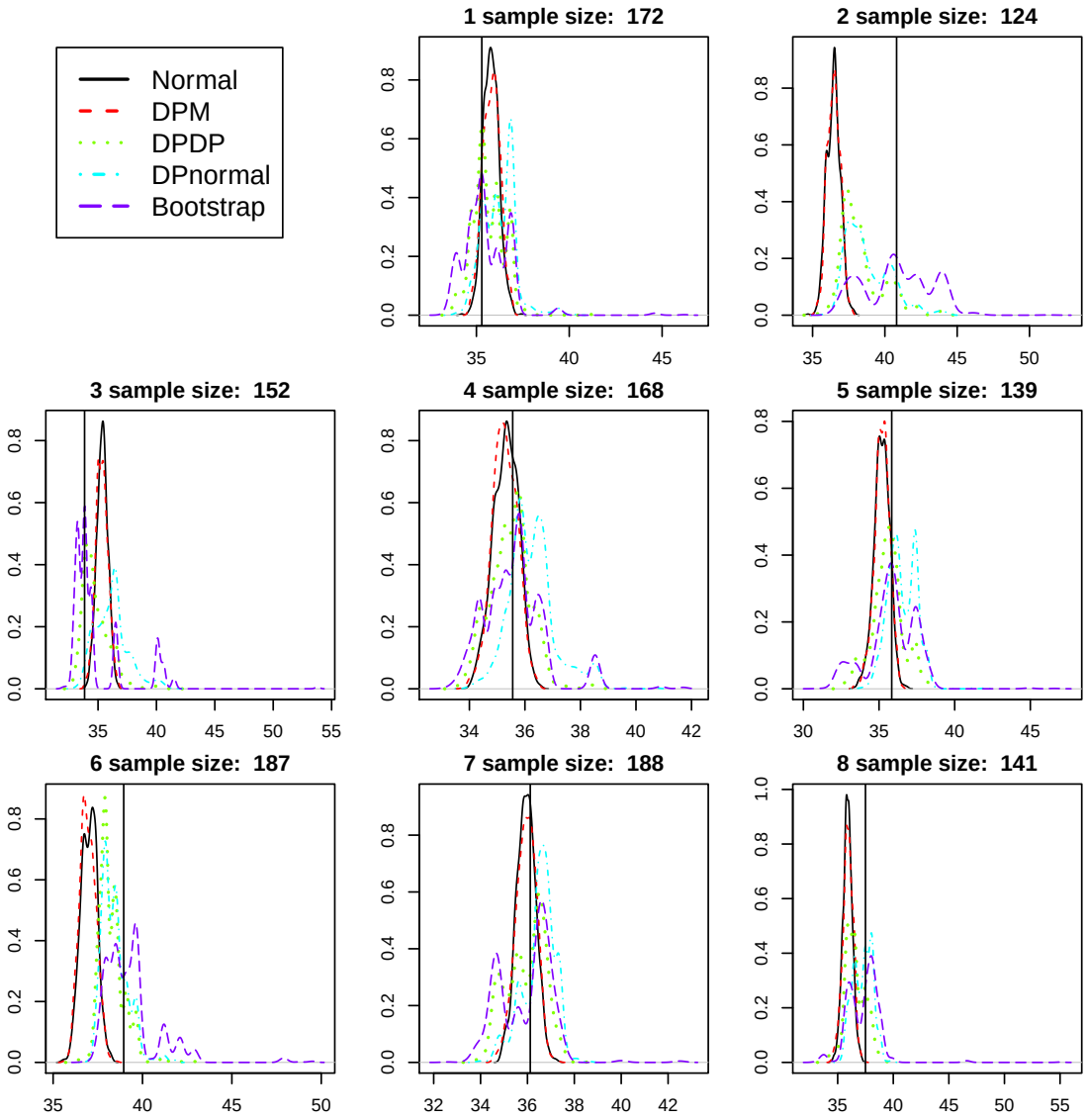

Figure 3.6: Plots of the posterior density of the finite population 95^{th} percentile by four models (normal, DPM, DPnormal, DPDP models) and Bayesian bootstrap for the first eight counties of body mass index (BMI) data

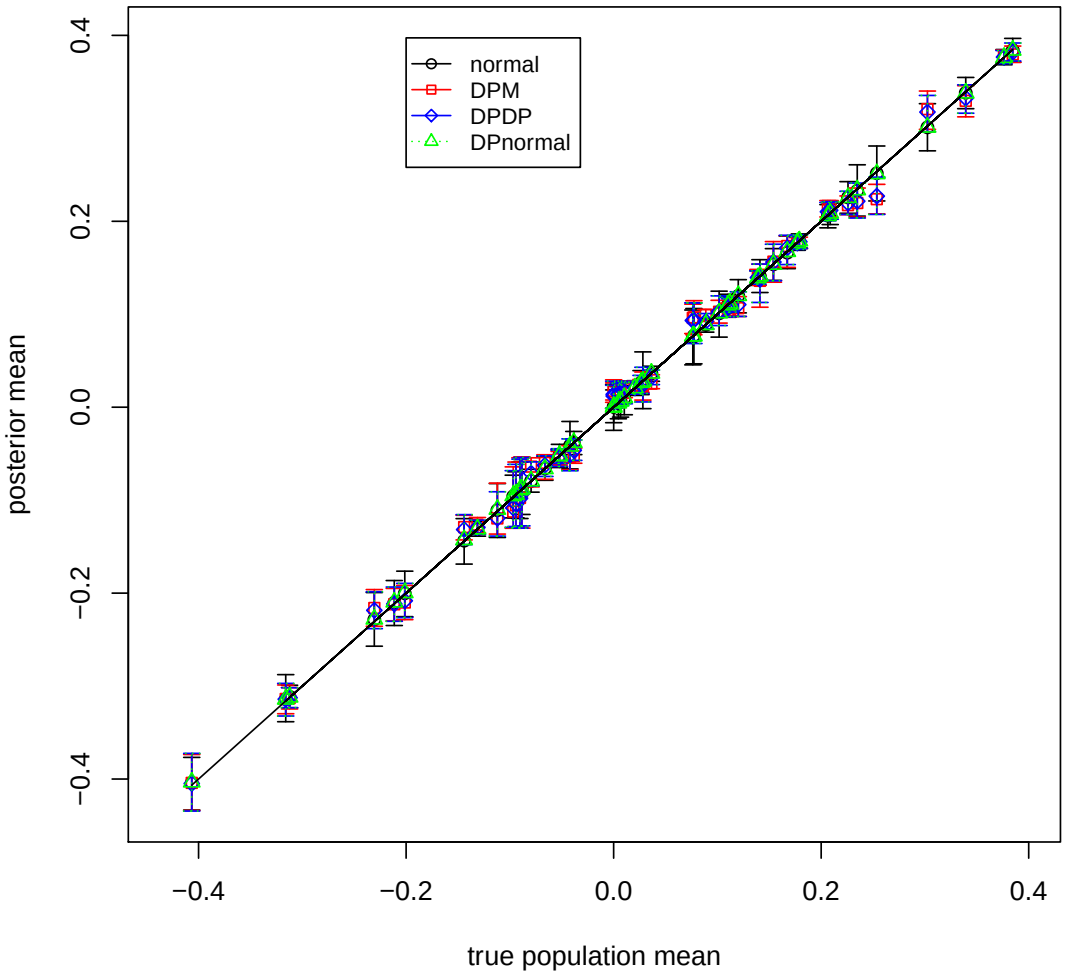

Figure 3.7: Comparison for the simulated normal data (posterior means with credible bands versus true population means): the predictive inference of the finite population mean for each county under four different models (normal, DPM, DPnormal and DPDP models).

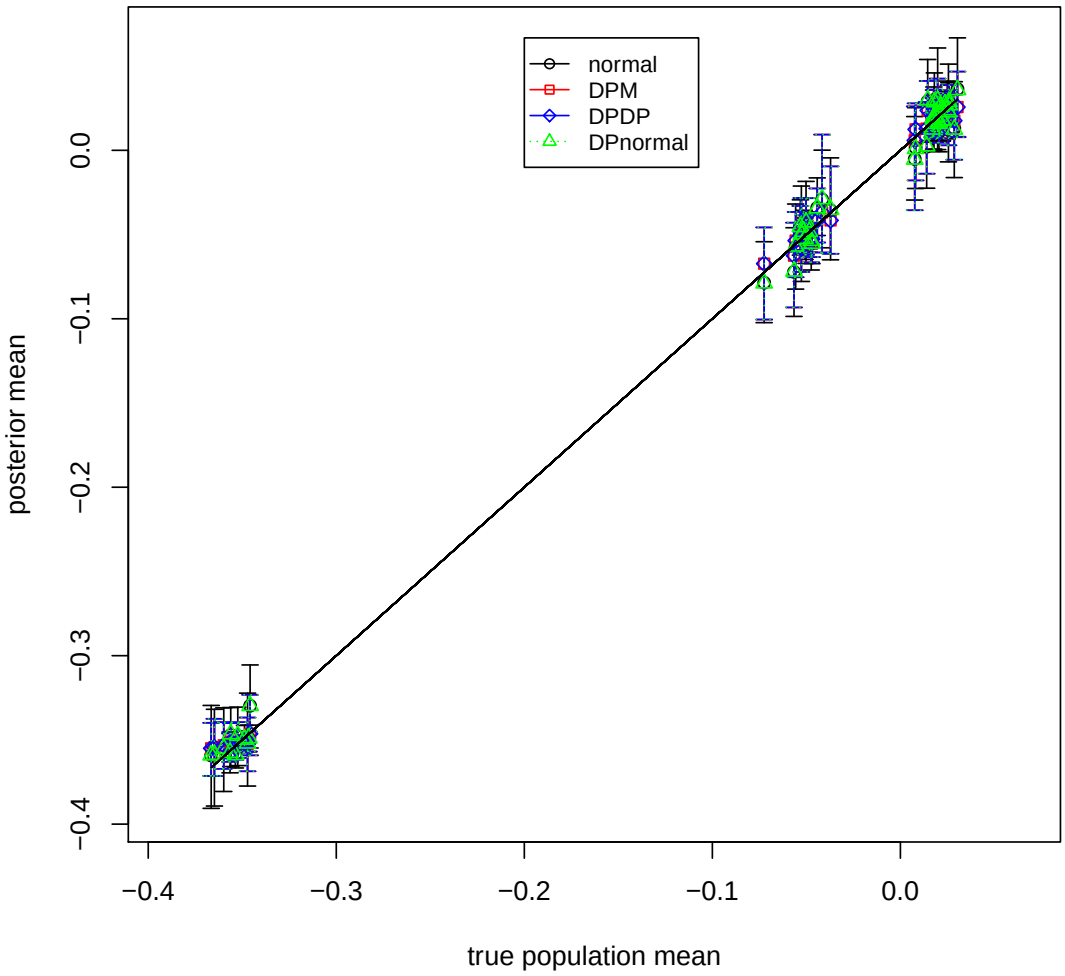

Figure 3.8: Comparison for the simulated DPM data (posterior means with credible bands versus true population means): the predictive inference of the finite population mean for each county under four different models (normal, DPM, DPnormal and DPDP models).

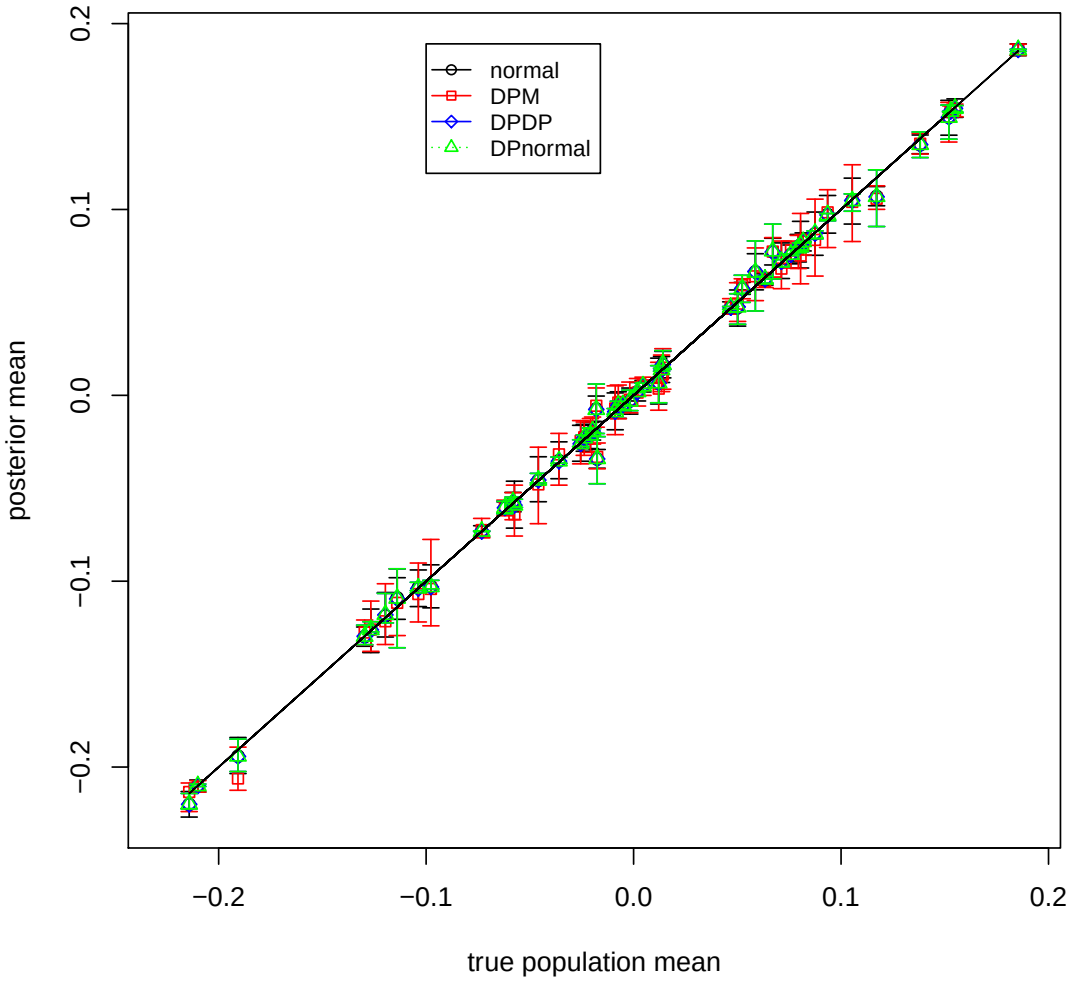

Figure 3.9: Comparison for the simulated DPDP data (posterior means with credible bands versus true population means): the predictive inference of the finite population mean for each county under four different models (normal, DPM, DPnormal and DPDP models).

Chapter 4

Three-level Dirichlet Process Models

In this chapter, we generalize the two-level Dirichlet process models to three levels, e.g. state-county-individual in a multi-stage finite population sampling. We assume that there are ℓ areas, within the ith area there are N_i sub-domains, and within the jth sub-domain there are M_{ij} (known) individuals. For sampling, n_i second-stage units are selected from the N_i units available, and m_{ij} third-stage units (elements) are sampled from the M_{ij} elements available. Inference is required for the finite population quantities of each area.

Let y_{ijk} denote the value for the kth unit within the jth sub-domain and ith area, $i = 1, \ldots, \ell, j = 1, \ldots, N_i, k = 1, \ldots, M_{ij}$. We assume that y_{ijk}, $i = 1, \ldots, \ell, j = 1, \ldots, n_i, k = 1, \ldots, m_{ij}$ are observed. Let $\underset{\sim}{y} = (\underset{\sim}{y}_s, \underset{\sim}{y}_{ns})$, where $\underset{\sim}{y}_s = \{y_{ijk}, i = 1, \ldots, \ell, j = 1, \ldots, n_i, k = 1, \ldots, m_{ij}\}$ is the vector of observed values and $\underset{\sim}{y}_{ns} = \{y_{ijk}, i = 1, \ldots, \ell, j = n_i + 1, \ldots, N_i, k = m_{ij} + 1, \ldots, M_{ij}\}$ vector of unobserved values. Inferences are required for $\bar{Y}_i = \sum_{j=1}^{N_i} \sum_{k=1}^{M_{ij}} y_{ijk} / \sum_{j=1}^{N_i} M_{ij}, i = 1, \ldots, \ell$, the finite population mean of the ith area and the 85th and 95th population quantiles

for each area. For $i = 1, \ldots, \ell, j = 1, \ldots, n_i$, we let $\bar{y}_{ij} = \sum_{k=1}^{m_{ij}} y_{ijk}/m_{ij}$, $s_{ij}^2 = \sum_{k=1}^{m_{ij}} (y_{ijk} - \bar{y}_{ij})^2/(m_{ij} - 1)$ and $m_0 = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} m_{ij}$.

The three-level Dirichlet process model (DPDPDP) is given by

$$\begin{aligned}
y_{ijk}|G_{ij} &\overset{ind}{\sim} G_{ij}, \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, N_i, \quad k = 1, \ldots, M_{ij}, \qquad (4.1)\\
G_{ij}|\mu_{ij} &\overset{ind}{\sim} \mathrm{DP}\{\alpha_{ij}, G_0(\mu_{ij})\},\\
\mu_{ij}|H_i &\overset{ind}{\sim} H_i,\\
H_i|\theta_i &\overset{ind}{\sim} \mathrm{DP}\{\gamma_i, H_0(\theta_i)\},\\
\theta_i|F &\overset{iid}{\sim} F,\\
F &\sim \mathrm{DP}\{\gamma_0, F_0(\cdot)\}.
\end{aligned}$$

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,

$$\begin{aligned}
\pi(\alpha_{ij}) &= \frac{1}{(\alpha_{ij}+1)^2}, \quad \alpha_{ij} > 0, \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, N_i, \qquad (4.2)\\
\pi(\gamma_i) &= \frac{1}{(\gamma_i+1)^2}, \quad \gamma_i > 0, \qquad (4.3)\\
\pi(\gamma_0) &= \frac{1}{(\gamma_0+1)^2}, \quad \gamma_0 > 0, \qquad (4.4)\\
\pi(\theta_0, \sigma^2, \rho_1, \rho_2) &= \frac{1}{\pi(1+\theta_0^2)} \frac{1}{(1+\sigma^2)^2}, \qquad (4.5)\\
&\quad -\infty < \theta_0 < \infty, 0 \le \sigma^2 < \infty, 0 \le \rho_1 \le 1, 0 \le \rho_2 \le 1,
\end{aligned}$$

with independence.

The corresponding embedded three-level parametric baseline model is,

$$
\begin{aligned}
y_{ijk}|\mu_{ij} &\overset{ind}{\sim} N(\mu_{ij}, \sigma^2), \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, N_i, \quad k = 1, \ldots, M_{ij}, \\
\mu_{ij}|\theta_i &\overset{ind}{\sim} N(\theta_i, \delta_1^2), \\
\theta_i &\overset{iid}{\sim} N(\theta_0, \delta_2^2).
\end{aligned} \tag{4.6}
$$

We note that Malec and Sedransk (1985) first proposed this type of model, an extension of the Scott-Smith model (3.1), initially used to model continuous data from three-stage cluster sampling. We call this parametric baseline (4.6) together with (4.5) the NNN model.

Similar to the two-level DP models, if we choose not to use DPs in all levels, we have seven additional models, which are NNN, NNDP, NDPN, NDPDP, DPNN, DPNDP, DPDPN. Here, we use letter N if the normal baseline distribution is chosen in that level, and DP for the random distribution drawn from DP.

4.1 Inference

Here, we start with the parametric baseline model. Letting $\underset{\sim}{\mu} = \{\mu_{ij}, i = 1, \ldots, \ell, j = 1, \ldots, n_i\}$ and $\underset{\sim}{\theta} = \{\theta_i, i = 1, \ldots, \ell\}$, the joint posterior density is

$$
\begin{aligned}
\pi \quad & (\underset{\sim}{\mu}, \underset{\sim}{\theta}, \theta_0, \sigma^2, \rho_1, \rho_2 | \underset{\sim}{y_s}) \\
\propto \quad & \left(\frac{1}{\sigma^2}\right)^{m_0/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \left[m_{ij}(\bar{y}_{ij} - \mu_{ij})^2 + (m_{ij} - 1) s_{ij}^2 \right] \right\} \\
\times \quad & \left(\frac{1}{\delta_1^2}\right)^{\sum_{i=1}^{\ell} n_i/2} \exp\left\{ -\frac{1}{2\delta_1^2} \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} (\mu_{ij} - \theta_i)^2 \right\} \\
\times \quad & \left(\frac{1}{\delta_2^2}\right)^{\ell/2} \exp\left\{ -\frac{1}{2\delta_2^2} \sum_{i=1}^{\ell} (\theta_i - \theta_0)^2 \right\} \times \frac{1}{\pi(1 + \theta_0^2)} \frac{1}{(1 + \sigma^2)^2}.
\end{aligned} \tag{4.7}
$$

Like the two-level hierarchical models, we use the SIR algorithm to draw from the posterior distribution (4.7). The proposal model is (4.6) together with an improper prior $\pi(\theta_0, \sigma^2, \rho_1, \rho_2) \propto \frac{1}{\sigma^2}, -\infty < \theta_0 < \infty, 0 \leq \sigma^2 < \infty, 0 \leq \rho_1 \leq 1, 0 \leq \rho_2 \leq 1$. Using the multiplication rule, we have

$$\begin{aligned}
\pi_a(\underset{\sim}{\mu}, \underset{\sim}{\theta}, \theta_0, \sigma^2, \rho_1, \rho_2 | \underset{\sim}{y}_s) &\propto \pi_a(\underset{\sim}{\mu} | \underset{\sim}{\theta}, \theta_0, \sigma^2, \rho_1, \rho_2, \underset{\sim}{y}_s) \pi_a(\underset{\sim}{\theta} |, \theta_0, \sigma^2, \rho_1, \rho_2, \underset{\sim}{y}_s) \quad (4.8) \\
&\times \pi_a(\theta_0 | \sigma^2, \rho_1, \rho_2, \underset{\sim}{y}_s) \pi_a(\sigma^2 | \rho_1, \rho_2, \underset{\sim}{y}_s) \pi_a(\rho_1, \rho_2 | \underset{\sim}{y}_s) \\
&\propto N\left(\mu_{ij}; \lambda_{ij}\bar{y}_{ij} + (1-\lambda_{ij})\theta_i, (1-\lambda_{ij})\frac{\rho_1}{1-\rho_1}\sigma^2\right) \\
&\times N\left(\theta_i; \lambda_i \tilde{y}_i + (1-\lambda_i)\theta_0, (1-\lambda_i)\frac{\rho_2}{1-\rho_2}\sigma^2\right) \\
&\times N\left(\theta_0; \tilde{y}, \frac{\sigma^2 \rho_2}{\sum_i \lambda_i (1-\rho_2)}\right) \mathrm{IG}\left[\sigma^2; \frac{m_0 - 1}{2}, A_2/2\right] \\
&\times \frac{\Gamma[(m_0-1)/2]}{(A_2/2)^{(m_0-1)/2}} \prod_{i=1}^{\ell} \prod_{j=1}^{n_i} (1-\lambda_{ij})^{1/2} \\
&\times \prod_{i=1}^{\ell} (1-\lambda_i)^{1/2} \left[\frac{\rho_2}{\sum_{i=1}^{\ell} \lambda_i (1-\rho_2)}\right]^{1/2},
\end{aligned}$$

where $\lambda_{ij} = m_{ij}/(m_{ij} + \frac{1-\rho_1}{\rho_1})$, $\tilde{y}_i = \sum_{j=1}^{n_i} \lambda_{ij}\bar{y}_{ij} / \sum_{j=1}^{n_i} \lambda_{ij}$, $\tilde{y} = \sum_{i=1}^{\ell} \lambda_i \tilde{y}_i / \sum_{i=1}^{\ell} \lambda_i$ and $A_2 = \frac{1-\rho_2}{\rho_2} \sum_{i=1}^{\ell} \lambda_i (\tilde{y} - \tilde{y}_i)^2 + \frac{1-\rho_1}{\rho_1} \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \lambda_{ij} (\tilde{y}_i - \bar{y}_{ij})^2 + \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} (m_{ij} - 1) s_{ij}^2$. We draw samples from the approximate joint posterior density (4.8) by first drawing samples from $\pi_a(\rho_1, \rho_2 | \underset{\sim}{y}_s)$.

Let us consider the NDPDP model,

$$\begin{aligned}
y_{ijk}|\mu_{ij} &\overset{ind}{\sim} N(\mu_{ij},\sigma^2), \quad i=1,\ldots,\ell, \quad j=1,\ldots,N_i, \quad k=1,\ldots,M_{ij}, \quad (4.9)\\
\mu_{ij}|H_i &\overset{ind}{\sim} H_i,\\
H_i|\theta_i &\overset{ind}{\sim} \mathrm{DP}\{\gamma_i, N(\theta_i,\delta_1^2)\},\\
\theta_i|F &\overset{iid}{\sim} F,\\
F &\sim \mathrm{DP}\{\gamma_0, N(\theta_0,\delta_2^2)\},
\end{aligned}$$

together with priors (4.3) (4.4) and (4.5). We develop an algorithm that is an extension of the slice sampler (Kalli, Griffin and Walker 2011) to obtain samples from the joint posterior density. The idea here is linking parameters of different levels. We use the slice sampler repeatedly to obtain samples from the conditional posterior distributions in the Gibbs sampling. We know that

$$H_i=\sum_{s=1}^{\infty}\omega_{is}\delta_{\mu_{is}^*}, \quad \omega_{i1}=\upsilon_{i1}, \quad \omega_{is}=\upsilon_{is}\prod_{m=1}^{s-1}(1-\upsilon_{im}),$$

$$\upsilon_{is}\overset{iid}{\sim}\mathrm{Beta}(1,\gamma_i), \quad \mu_{is}^*\overset{iid}{\sim}N(\theta_i,\delta_1^2),$$

and

$$F=\sum_{t=1}^{\infty}\omega_{0t}\delta_{\theta_t^*}, \quad \omega_{01}=\upsilon_{01}, \quad \omega_{0t}=\upsilon_{0t}\prod_{m=1}^{t-1}(1-\upsilon_{0m}),$$

$$\upsilon_{0t}\overset{iid}{\sim}\mathrm{Beta}(1,\gamma_0), \quad \theta_t^*\overset{iid}{\sim}N(\theta_0,\delta_2^2).$$

The Gibbs sampler proceeds as follows.

1. For each i, update $\{\mu^*_{is}, s = 1, \dots\}$ and γ_i as if the model is

$$\begin{aligned} y_{ijk}|\mu_{ij} &\overset{ind}{\sim} N(\mu_{ij}, \sigma^2), \quad i = 1, \dots, \ell, \quad j = 1, \dots, n_i, \quad k = 1, \dots, m_{ij}, \\ \mu_{ij}|H_i &\overset{ind}{\sim} H_i, \\ H_i|\theta_i &\overset{ind}{\sim} \mathrm{DP}\{\gamma_i, N(\theta_i, \delta_1^2)\}, \end{aligned}$$

which can be fit as a DPM model.

2. Update $\{\theta^*_t, t = 1, \dots\}$ and γ_0 as if the model is

$$\begin{aligned} \mu^*_{is}|\theta_i &\overset{ind}{\sim} N(\theta_i, \delta_1^2), \quad i = 1, \dots, \ell, \quad s = 1, \dots \\ \theta_i|F &\overset{iid}{\sim} F, \\ F &\sim \mathrm{DP}\{\gamma_0, N(\theta_0, \delta_2^2)\}. \end{aligned}$$

Again this can be considered as a DPM model.

3. Update other hyper-parameters $\{\theta_0, \sigma^2, \rho_1, \rho_2\}$. We have

$$\begin{aligned} \pi(\theta_0, \sigma^2, \rho_1, \rho_2|\dots) &\propto \prod_i \prod_j \prod_k N(y_{ijk}; \mu_{ij}, \sigma^2) \times \prod_i \prod_s N(\mu^*_{is}; \theta_i, \delta_1^2) \\ &\times \prod_t N(\theta^*_t; \theta_0, \delta_2^2) \times \pi(\sigma^2, \theta_0, \rho_1, \rho_2). \end{aligned}$$

Next, let us consider the NNDP model,

$$y_{ijk}|\mu_{ij} \overset{ind}{\sim} N(\mu_{ij}, \sigma^2), \quad i = 1, \dots, \ell, j = 1, \dots, N_i, k = 1, \dots, M_{ij}, \tag{4.10}$$

$$\mu_{ij}|\theta_i \overset{ind}{\sim} N(\theta_i, \delta_1^2), \tag{4.11}$$

$$\theta_i|F \overset{iid}{\sim} F, \tag{4.12}$$

$$F \sim \mathrm{DP}\{\gamma_0, N(\theta_0, \delta_2^2)\}, \tag{4.13}$$

together with priors (4.4) and (4.5).

By integrating out μ_{ij}, we have

$$\begin{aligned} f(\underset{\sim}{y}|\underset{\sim}{\theta}) &\propto \left(\frac{1}{\sigma^2}\right)^{(m_0+\ell)/2} \left(\frac{1-\rho_2}{\rho_2}\right)^{\ell/2} \prod_{i=1}^{\ell}\prod_{j=1}^{n_i}(1-\lambda_{ij})^{1/2} \\ &\times \exp\left\{-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{\ell}\sum_{j=1}^{n_i}\left[(m_{ij}-1)s_{ij}^2 + \lambda_{ij}\frac{1-\rho_1}{\rho_1}(\theta_i-\bar{y}_{ij})^2\right]\right\}\right\}. \end{aligned} \tag{4.14}$$

Here, (4.14), (4.12) and (4.13) now form a DPM model, that is (4.14) as a likelihood function with parameter $\underset{\sim}{\theta}$ which has a DP prior. Moreover, for $i = 1, \ldots, \ell$, $j = 1, \ldots, n_i$,

$$\mu_{ij}|\theta_i, \theta_0, \sigma^2, \rho_1, \underset{\sim}{y}_s \sim N\left[\lambda_{ij}\bar{y}_{ij} + (1-\lambda_{ij})\theta_i, (1-\lambda_{ij})\frac{1-\rho_1}{\rho_1}\sigma^2\right].$$

Next, we consider the NDPN model,

$$\begin{aligned} y_{ijk}|\mu_{ij} &\overset{ind}{\sim} N(\mu_{ij}, \sigma^2), \quad i = 1, \ldots, \ell, j = 1, \ldots, N_i, k = 1, \ldots, M_{ij}, \\ \mu_{ij}|H_i &\overset{ind}{\sim} H_i, \\ H_i|\theta_i &\overset{ind}{\sim} \mathrm{DP}\{\gamma_i, N(\theta_i, \delta_1^2)\}, \\ \theta_i &\overset{iid}{\sim} N(\theta_0, \delta_2^2), \end{aligned} \tag{4.15}$$

together with priors (4.3) and (4.5). The Gibbs sampler proceeds as follows.

1. For each i, update $\{\mu_{is}^*, s = 1, \ldots\}$ and γ_i as if the model is

$$\begin{aligned} y_{ijk}|\mu_{ij} &\overset{ind}{\sim} N(\mu_{ij}, \sigma^2), \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, n_i, \quad k = 1, \ldots, m_{ij}, \\ \mu_{ij}|H_i &\overset{ind}{\sim} H_i, \\ H_i|\theta_i &\overset{ind}{\sim} \mathrm{DP}\{\gamma_i, N(\theta_i, \delta_1^2)\}, \end{aligned}$$

which can be fit as a DPM model.

2. Update $\{\theta_i, i = 1, \ldots, \ell\}$ and other hyper-parameters $\{\theta_0, \sigma^2, \rho_1, \rho_2\}$ as if the model is

$$\begin{aligned} \mu^*_{is}|\theta_i &\overset{ind}{\sim} N(\theta_i, \delta_1^2), \quad i = 1, \ldots, \ell, \quad s = 1, \ldots \\ \theta_i &\overset{iid}{\sim} N(\theta_0, \delta_2^2), \\ \pi(\theta_0, \sigma^2, \rho_1, \rho_2) &= \frac{1}{\pi(1+\theta_0^2)} \frac{1}{(1+\sigma^2)^2}, \end{aligned}$$

which is easy to fit as a two-level normal model.

At last, when the DP is used for the sampling process, the idea is similar to the two-level DP models. Inference of DPDPDP, DPNDP, DPDPN and DPNN model can be obtained easily. For example, the DPDPDP model can be reduced to NDPDP model with additional sampling of α_{ij}. The algorithm is as follows.

Step 1 : For each i, j, draw α_{ij} from the posterior distribution $\pi(\alpha_{ij}|k_{ij})$, where k_{ij} denotes the number of distinct values among observations for fixed i and j.

Step 2: Draw other parameters from the NDPDP model with distinct values as data.

4.2 Propriety of the Posterior Distributions

Lemma 4.2.1 *The joint posterior density* $\pi(\underset{\sim}{\mu}, \underset{\sim}{\theta}, \theta_0, \sigma^2, \rho_1, \rho_2|\underset{\sim}{y}_s)$ *(4.7) under the NNN model is proper if* $\ell \geq 2$ *and* $n_i \geq 2$ *at least for one* i.

Proof: Similar to the two-level normal model, we only need to show that the joint posterior density $\pi_a(\underset{\sim}{\mu}, \underset{\sim}{\theta}, \theta_0, \sigma^2, \rho_1, \rho_2|\underset{\sim}{y}_s)$ (4.8) under the proposal model is proper.

That is to show

$$\begin{aligned}
& \int\int \pi_a(\rho_1,\rho_2|\underset{\sim}{y}_s)d\rho_1 d\rho_2 \\
= & \int\int \frac{\Gamma[(m_0-1)/2]}{(A_2/2)^{(m_0-1)/2}} \prod_{i=1}^{\ell}\prod_{j=1}^{n_i}(1-\lambda_{ij})^{1/2} \\
\times & \prod_{i=1}^{\ell}(1-\lambda_i)^{1/2}\left[\frac{\rho_2}{\sum_{i=1}^{\ell}\lambda_i(1-\rho_2)}\right]^{1/2} d\rho_1 d\rho_2 < \infty.
\end{aligned}$$

It is clear that $\pi_a(\rho_1,\rho_2|\underset{\sim}{y}_s)$ is well defined because of $A_2 \neq 0$ if $\ell \geq 2$ and $n_i \geq 2$ at least for one i. Thus, $\int\int \pi_a(\rho_1,\rho_2|\underset{\sim}{y}_s)d\rho_1 d\rho_2 < \infty$ since $0 \leq \rho_1 \leq 1, 0 \leq \rho_2 \leq 1$.

Theorem 4.2.2 *If the posterior density under the NNN baseline model is proper, the posterior density under all other three-level DP models are all proper.*

Proof: We prove for the NDPDP model, others are similar. Letting $\underset{\sim}{\mu}_i = \{\mu_{ij}, j = 1,\ldots,n_i\}$ for $i = 1,\ldots,\ell$, $\Omega' = \{\theta_0, \sigma^2, \rho_1, \rho_2, \gamma_0\}$ and $\underset{\sim}{\gamma} = \{\gamma_1,\ldots,\gamma_\ell\}$, we have

$$\begin{aligned}
f(\underset{\sim}{y}_s) &= \int f(\underset{\sim}{y}_s|\Omega)\pi(\Omega)d\Omega \\
&= \int \left[\prod_{i=1}^{\ell}\prod_{j=1}^{n_i}\prod_{k=1}^{m_{ij}} N(y_{ijk}|\mu_{ij},\sigma^2)\right]\left[\prod_{i=1}^{\ell}\pi(\underset{\sim}{\mu}_i)d\underset{\sim}{\mu}_i\right]\pi(\underset{\sim}{\theta})d\underset{\sim}{\theta}\pi(\underset{\sim}{\gamma})d\underset{\sim}{\gamma}\pi(\Omega')d\Omega',
\end{aligned}$$

where

$$\begin{aligned}
\pi(\underset{\sim}{\mu}_i) &= N(\mu_{i1};\theta_i,\delta_1^2)\prod_{j=2}^{n_i}\left[\frac{\gamma_i}{\gamma_i+j-1}N(\mu_{ij};\theta_i,\delta_1^2) + \frac{1}{\gamma_i+j-1}\sum_{s=1}^{j-1}\delta_{\mu_{is}}(\mu_{ij})\right], \\
\pi(\underset{\sim}{\theta}) &= N(\theta_1;\theta_0,\delta_2^2)\prod_{i=2}^{\ell}\left[\frac{\gamma_0}{\gamma_0+i-1}N(\theta_i;\theta_0,\delta_2^2) + \frac{1}{\gamma_0+i-1}\sum_{s=1}^{i-1}\delta_{\theta_s}(\theta_i)\right], \\
\pi(\Omega') &= \frac{1}{(\gamma_0+1)^2}\frac{1}{\pi(1+\theta_0^2)}\frac{1}{(1+\sigma^2)^2}, \\
\pi(\underset{\sim}{\gamma}) &= \left[\prod_{i=1}^{\ell}\frac{1}{(\gamma_i+1)^2}\right].
\end{aligned}$$

It is convenient to write

$$\begin{aligned} f(\underset{\sim}{y}_s) &\leq \int \prod_{i=1}^{\ell} \left\{ \prod_{j=1}^{n_i} \left[N(\mu_{ij}; \theta_i, \delta_1^2) \prod_{k=1}^{m_{ij}} N(y_{ijk}|\mu_{ij}, \sigma^2) \right] \right\} d\underset{\sim}{\mu}_i \\ &\times N(\theta_1; \theta_0, \delta_2^2) \prod_{i=2}^{\ell} \left[\frac{\gamma_0}{\gamma_0 + i - 1} N(\theta_i; \theta_0, \delta_2^2) + \frac{1}{\gamma_0 + i - 1} \sum_{s=1}^{i-1} \delta_{\theta_s}(\theta_i) \right] d\underset{\sim}{\theta} \\ &\times \frac{1}{(\gamma_0 + 1)^2} \frac{1}{\pi(1 + \theta_0^2)} \frac{1}{(1 + \sigma^2)^2} d\Omega'. \end{aligned}$$

Now this is the marginal distribution under the NNDP model. It is easy to show the NNDP model is proper since it is a DPM model with the $\underset{\sim}{\mu}$ integrated out.

4.3 Bayes Factor

Here we give the key formula needed for the computation of Bayes factors under the NNN, NDPDP and NNDP model. Others are similar.

For the NNN model, it is easy to integrate out $\underset{\sim}{\mu}$ and $\underset{\sim}{\theta}$. We have the marginal likelihood function for the NNN model as

$$\begin{aligned} M(\underset{\sim}{y}_s) &= \int f(\underset{\sim}{y}_s|\theta_0, \sigma^2, \rho_1, \rho_2) \times \pi(\theta_0, \sigma^2, \rho_1, \rho_2) d\theta_0 d\sigma^2 d\rho_1 d\rho_2 \qquad (4.16) \\ &= \int \left(\frac{1}{2\pi\sigma^2} \right)^{m_0/2} \left(\frac{1}{2\pi\delta_1^2} \right)^{\sum_{i=1}^{\ell} n_i/2} \prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \left[2\pi(1 - \lambda_{ij}) \frac{\rho_1}{1 - \rho_1} \sigma^2 \right]^{1/2} \\ &\times \exp\left\{ -\frac{1}{2\sigma^2} \left[\frac{1 - \rho_2}{\rho_2} \sum_{i=1}^{\ell} \lambda_i (\tilde{y}_i - \theta_0)^2 + \frac{1 - \rho_1}{\rho_1} \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \lambda_{ij} (\tilde{y}_i - \bar{y}_{ij})^2 \right. \right. \\ &+ \left. \left. \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} (m_{ij} - 1) s_{ij}^2 \right] \right\} \left(\frac{1}{2\pi\delta_2} \right)^{\ell/2} \prod_{i=1}^{\ell} \left[2\pi(1 - \lambda_i) \frac{\rho_2}{1 - \rho_2} \sigma^2 \right]^{1/2} \\ &\times \frac{1}{\pi(1 + \theta_0^2)} \frac{1}{(1 + \sigma^2)^2} d\theta_0 d\sigma^2 d\rho_1 d\rho_2. \end{aligned}$$

The rest of the computation follows.

For the NDPDP model, the most tricky part is to obtain the approximate posterior density for θ_i. Other parts are very similar to the two-level DP models. Letting $\sigma_{ij}^2 = \delta_1^2 + \sigma^2/m_{ij}$, the idea is to use the following model,

$$\begin{aligned} \bar{y}_{ij} &\overset{ind}{\sim} N\left(\theta_i, \sigma_{ij}^2\right), \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, n_i, \\ \theta_i | H_0 &\overset{iid}{\sim} H_0, \\ H_0 &\sim \mathrm{DP}\{\gamma_0, N(\theta_0, \delta_2^2)\}, \end{aligned} \tag{4.17}$$

as an approximation of $\pi(\underset{\sim}{\theta}|\Omega', \underset{\sim}{y}_s) = \pi(\theta_1|\Omega', \underset{\sim}{y}_s) \prod_{i=2}^{\ell} \pi(\theta_i|\theta_{i-1}, \ldots, \theta_1, \Omega', \underset{\sim}{y}_s)$. Applying the idea to the DPM model, we have

$$\begin{aligned} \pi_a(\theta_i|\Omega', \underset{\sim}{y}_s) &= \frac{1}{i-1+\gamma_0} \sum_{s=1}^{i-1} \left[\prod_{j=1}^{n_i} \left(\frac{1}{2\pi\sigma_{ij}^2} \right)^{1/2} \exp\left\{ -\frac{(\bar{y}_{ij} - \theta_s)^2}{2\sigma_{ij}^2} \right\} \right] \delta_{\theta_s}(\theta_i) \\ &+ \frac{\gamma_0}{i-1+\gamma_0} \left(\frac{1}{2\pi\sigma_{ij}^2} \right)^{n_i/2} \left[(1-\tilde{\lambda}_i)\rho_2\sigma^2/(1-\rho_2) \right]^{1/2} \\ &\times \exp\left\{ -\frac{1}{2} \left[\sum_{j=1}^{n_i} \tilde{\lambda}_{ij}(\tilde{\bar{y}}_i - \bar{y}_{ij})^2 + \frac{\tilde{\lambda}_{ij}(1-\rho_2)}{\rho_2\sigma^2} (\tilde{\bar{y}}_i - \theta_0)^2 \right] \right\} \\ &\times N\left[\theta_i; \tilde{\lambda}_i \tilde{\bar{y}}_i + (1-\tilde{\lambda}_i)\theta_0, (1-\tilde{\lambda}_i) \frac{\rho_2}{(1-\rho_2)} \sigma^2 \right], \quad i = 2, \ldots, \ell, \end{aligned}$$

and

$$\pi_a(\theta_1|\Omega', \underset{\sim}{y}_s) = N\left[\theta_1; \tilde{\lambda}_1 \tilde{\bar{y}}_1 + (1-\tilde{\lambda}_1)\theta_0, (1-\tilde{\lambda}_1) \frac{\rho_2}{(1-\rho_2)} \sigma^2 \right],$$

where $\tilde{\lambda}_{ij} = 1/\sigma_{ij}^2$, $\tilde{\bar{y}}_i = \sum_{j=1}^{n_i} \tilde{\lambda}_{ij}\bar{y}_{ij} / \sum_{j=1}^{n_i} \tilde{\lambda}_{ij}$, $\tilde{\lambda}_i = \sum_{j=1}^{n_i} \tilde{\lambda}_{ij} / (\sum_{j=1}^{n_i} \tilde{\lambda}_{ij} + 1/\delta_2^2)$.

For the NNDP model, integrating out μ_{ij}, the likelihood function is

$$\begin{aligned} f(\underset{\sim}{y}_s|\underset{\sim}{\theta}, \Omega') &= \left(\frac{1}{2\pi\sigma^2} \right)^{m_0/2} \prod_{i=1}^{\ell} \prod_{j=1}^{n_i} (1-\lambda_{ij})^{1/2} \\ &\times \exp\left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \left[(m_{ij}-1)s_{ij}^2 + \lambda_{ij} \frac{1-\rho_1}{\rho_1} (\theta_i - \bar{y}_{ij})^2 \right] \right] \right\}. \end{aligned}$$

The prior is

$$\begin{aligned}\pi(\Omega') &= N(\theta_1;\theta_0,\delta_2^2)\prod_{i=2}^{\ell}\left(\frac{\gamma_0}{\gamma_0+i-1}N(\theta_i;\theta_0,\delta_2^2)+\frac{1}{\gamma_0+i-1}\sum_{s=1}^{i-1}\delta_{\theta_s}(\theta_i)\right)\\ &\times \frac{1}{(\gamma_0+1)^2}\frac{1}{\pi(1+\theta_0^2)}\frac{1}{(1+\sigma^2)^2}.\end{aligned}$$

Here $\pi_a(\theta_i|\Omega', \underset{\sim}{y}_s)$ is same as the approximate posterior distribution in the NDPDP model.

4.4 Empirical Studies

The three-level models are desirable since BMI data are post-stratified to three-level. The sub-domains are formed by age, race and sex. We fit the three-level DP models (NDPDP, NNDP, NDPN, NNN, DPDPDP, DPNDP, DPDPN, DPNN model) to obtain the finite population mean, 85^{th} and 95^{th} percentile for each county of BMI data. We have conducted model comparisons under the three-level DP models.

For the three-level models, it is harder to converge than the two-level models, so longer runs are needed. For the NNDP, NDPN model, we run 35000 MCMC iterations, burn in 25,000 and thin every 10th to obtain 1000 converged posterior samples. For the NDPDP model, we run 75000 iterations, burn in 70000 and thin every 5th to obtain 1000 posterior samples. For the DPNDP model, we run 55000 iterations, burn in 45000 and thin every 10th to obtain 1000 posterior samples. For the DPDPN model, we run 45000 iterations, burn in 35000 and thin every 10th to obtain 1000 posterior samples. For the DPDPDP model, we run 90000 iterations, burn in 80000 and thin every 10th to obtain 1000 posterior samples. Table 4.1 gives the p-values of the Geweke test and the effective sample sizes for the parameters σ^2,

θ_0, δ_1^2, δ_2^2 and γ_0 under each model. The p-values are not significant and effective sample sizes are not too far from 1000. These numerical summaries, trace plots, and autocorrelation plots indicate that the MCMC chains converge.

Tables 4.2, 4.3 and 4.4 give the summary statistics, posterior mean (PM) and posterior standard deviation (PSD), of the finite population mean, 85^{th} and 95^{th} percentile for each county of BMI data under the three-level DP models (NNN, NNDP, NDPN, NDPDP, DPNN, DPNDP, DPDPN, DPDPDP models) and Bayesian bootstrap respectively. These tables show that roughly similar results are obtained from the eight models. We examine several plots to further compare the results of BMI data.

The predictive inference of the finite population mean, 85^{th} and 95^{th} percentile for each county by eight different models (NNN, NNDP, NDPN, NDPDP, DPNN, DPNDP, DPDPN, DPDPDP models) are compared respectively. Figures 4.1, 4.2 and 4.3 plot posterior means with credible bands versus direct estimates for BMI data. In Figure 4.1, we compare the difference between the predictive inference of the finite population means under models and the direct estimates. The posterior means under the NNN and DPNN models are shrank toward to the overall mean. The posterior means under the other models are closer to the direct estimates with less pooling. Similar to the two-level DP models, the predictive inference of the population percentile is not so good under the DPNN, DPNDP, DPDPN and DPDP model (see Figures 4.2 and 4.3).

We present the density estimations of the population mean, 85^{th} and 95^{th} percentile for the first eight counties as an example (see Figures 4.4, 4.5 and 4.6). Since the existence of the third stage, the NNN has reduced the bias comparing to the two-level normal model. The estimated densities under the eight three-level models are similar. The density under the DPNN model is very close to the NNN model

with slightly smaller variation. Consistent with the observations from Figure 4.1, results from the nonparametric alternative tend to have bigger variation however less bias.

The log of the marginal likelihood (LML) with Monte Carlo errors, log pseudo marginal likelihood (LPML) and percentages of conditional predictive ordinate (CPO) less than .025 ($P_{CPO} < .025$) and .014 ($P_{CPO} < .014$) for BMI data under the NNN, NNDP, NDPN, NDPDP model are given in Table 4.5. These measurements may be inconsistent when the three-level parametric models embedded in the nonparametric models.

In conclusion, it may be not obvious to say which model is better. For quantile estimation, it does not seem reasonable to use a DP for the sampling process, but this may be fine for the finite population mean. BMI data are certainly not normally distributed. Typically a log transformation is used, but this is also uncertain of the form of distribution after transformation. In addition, another problem of the log transformation is that when transforming back to the original scale, the expectation dose not exist. Of course, there will be some loss in efficiency under an nonparametric model. But the nonparametric alternatives seem to be the right direction.

Table 4.1: 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_0, δ_1^2, δ_2^2 and γ_0 for the NNDP, NDPDP, DPNDP, DPDPN, and DPDPDP model

	p-values for the Geweke test				
Model	σ^2	θ_0	δ_1^2	δ_2^2	γ_0
NNDP	0.9496993	0.3050090	0.3878581	0.5864042	0.8140230
NDPDP	0.8337016	0.3316585	0.4926789	0.0824082	0.8636205
DPNDP	0.9799888	0.7478633	0.6661014	0.7090474	0.2504232
DPDPN	0.2989892	0.2899847	0.2523066	0.8983445	NA
DPDPDP	0.8799581	0.3183782	0.9755728	0.3202928	0.3073552
	effective sample sizes				
Model	σ^2	θ_0	δ_1^2	δ_2^2	γ_0
NNDP	1000	1000	1000	870.4086	680.3195
NDPDP	1000	757.5346	700.8814	1000	658.7319
DPNDP	1000	907.0290	1000	1000	818.0584
DPDPN	1000	1000	808.5789	879.5892	NA
DPDPDP	1000	1000	1000	1051.920	1009.381

Table 4.2: 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 eight three-level DP models and Bayesian bootstrap

	Bootstrap		NNN		NNDP		NDPN		NDPDP		DPNN		DPNDP		DPDPN		DPDPDP	
	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD
1	26.93	0.36	26.93	0.33	26.91	0.35	26.92	0.47	26.94	0.44	26.91	0.32	26.90	0.34	26.87	0.43	26.90	0.43
2	27.48	0.54	27.29	0.39	27.40	0.45	27.47	0.59	27.47	0.51	27.38	0.39	27.50	0.45	27.58	0.53	27.56	0.50
3	26.28	0.44	26.37	0.35	26.27	0.38	26.28	0.51	26.24	0.50	26.42	0.36	26.27	0.38	26.30	0.49	26.27	0.52
4	26.00	0.37	26.16	0.35	26.01	0.37	26.00	0.41	26.04	0.43	26.29	0.34	26.18	0.36	26.17	0.43	26.19	0.41
5	25.67	0.41	26.09	0.38	25.81	0.43	25.69	0.53	25.70	0.52	25.96	0.40	25.76	0.40	25.64	0.46	25.68	0.48
6	28.40	0.43	28.00	0.33	28.28	0.40	28.41	0.47	28.38	0.47	28.13	0.35	28.39	0.38	28.46	0.40	28.43	0.45
7	27.08	0.34	26.96	0.33	27.02	0.34	27.06	0.50	27.05	0.43	26.96	0.32	27.03	0.34	27.07	0.45	27.04	0.44
8	26.88	0.47	26.91	0.35	26.88	0.39	26.90	0.52	26.91	0.50	26.99	0.37	26.98	0.40	27.02	0.47	27.02	0.46
9	27.83	0.39	27.62	0.36	27.76	0.39	27.88	0.48	27.85	0.46	27.77	0.34	27.90	0.39	27.97	0.49	27.93	0.44
10	27.65	0.45	27.39	0.35	27.57	0.38	27.65	0.47	27.59	0.46	27.46	0.34	27.62	0.37	27.71	0.45	27.68	0.45
11	27.26	0.26	27.20	0.23	27.23	0.25	27.29	0.30	27.28	0.30	27.29	0.24	27.34	0.24	27.38	0.27	27.38	0.26
12	25.72	0.34	26.03	0.35	25.80	0.38	25.73	0.43	25.73	0.41	26.00	0.33	25.78	0.34	25.73	0.40	25.72	0.38
13	26.67	0.39	26.78	0.35	26.69	0.37	26.65	0.49	26.66	0.47	26.73	0.35	26.64	0.36	26.61	0.44	26.62	0.42
14	27.28	0.17	27.26	0.17	27.26	0.17	27.29	0.20	27.29	0.19	27.44	0.16	27.45	0.17	27.47	0.18	27.47	0.18
15	27.33	0.50	27.22	0.38	27.30	0.46	27.27	0.63	27.32	0.65	27.25	0.40	27.32	0.44	27.33	0.56	27.32	0.57
16	27.31	0.40	27.17	0.35	27.24	0.39	27.33	0.50	27.30	0.47	27.33	0.35	27.42	0.38	27.46	0.47	27.46	0.46
17	26.08	0.38	26.30	0.35	26.14	0.38	26.05	0.47	26.08	0.44	26.31	0.34	26.17	0.37	26.12	0.46	26.13	0.43
18	26.71	0.37	26.76	0.36	26.68	0.42	26.71	0.52	26.69	0.52	26.78	0.36	26.69	0.39	26.74	0.51	26.69	0.45
19	26.19	0.41	26.54	0.34	26.31	0.37	26.22	0.45	26.27	0.43	26.45	0.35	26.24	0.36	26.20	0.44	26.19	0.45
20	26.81	0.44	26.87	0.35	26.84	0.38	26.88	0.52	26.87	0.51	26.87	0.35	26.83	0.38	26.83	0.46	26.84	0.47
21	26.90	0.43	26.81	0.35	26.82	0.41	26.90	0.53	26.81	0.52	26.97	0.37	26.94	0.41	26.98	0.48	27.01	0.49
22	27.28	0.36	27.12	0.34	27.17	0.37	27.30	0.48	27.23	0.48	27.23	0.33	27.27	0.36	27.37	0.44	27.33	0.42
23	25.87	0.41	26.26	0.34	25.95	0.41	25.92	0.52	25.95	0.53	26.11	0.35	25.89	0.36	25.81	0.49	25.85	0.51
24	27.12	0.42	27.07	0.35	27.10	0.39	27.11	0.51	27.11	0.51	27.12	0.37	27.14	0.39	27.12	0.50	27.16	0.46
25	26.75	0.44	26.78	0.34	26.74	0.40	26.75	0.50	26.75	0.46	26.87	0.37	26.84	0.38	26.89	0.47	26.85	0.45
26	26.58	0.47	26.78	0.37	26.61	0.44	26.62	0.55	26.63	0.55	26.75	0.41	26.68	0.45	26.64	0.55	26.65	0.55
27	26.77	0.36	26.78	0.32	26.74	0.34	26.77	0.42	26.79	0.42	26.79	0.31	26.76	0.36	26.80	0.48	26.77	0.39
28	27.52	0.49	27.33	0.35	27.44	0.40	27.55	0.52	27.51	0.53	27.36	0.36	27.43	0.39	27.54	0.47	27.52	0.49
29	26.59	0.43	26.68	0.39	26.63	0.47	26.65	0.63	26.64	0.61	26.70	0.41	26.64	0.43	26.64	0.55	26.65	0.61
30	25.91	0.40	26.22	0.35	25.98	0.41	25.94	0.53	26.00	0.53	26.22	0.36	25.98	0.39	25.94	0.45	25.98	0.45
31	27.82	0.33	27.62	0.33	27.76	0.36	27.81	0.43	27.82	0.40	27.61	0.32	27.75	0.35	27.77	0.40	27.76	0.38
32	27.64	0.41	27.38	0.31	27.55	0.37	27.63	0.42	27.62	0.42	27.47	0.34	27.61	0.36	27.66	0.43	27.66	0.39
33	26.35	0.32	26.53	0.34	26.40	0.36	26.37	0.44	26.40	0.45	26.51	0.32	26.40	0.36	26.37	0.44	26.38	0.43
34	27.39	0.30	27.25	0.28	27.32	0.31	27.39	0.38	27.42	0.37	27.30	0.27	27.38	0.30	27.42	0.37	27.40	0.38
35	26.80	0.38	26.86	0.32	26.82	0.34	26.81	0.44	26.79	0.45	26.90	0.32	26.85	0.34	26.84	0.41	26.83	0.41

Table 4.3: Comparison of posterior mean (PM) and posterior standard deviation (PSD) of the finite population 85^{th} percentile for each county of the BMI data by eight three-level DP models and Bayesian bootstrap

	Bootstrap		NNN		NNDP		NDPN		NDPDP		DPNN		DPNDP		DPDPN		DPDPDP	
	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD
1	32.14	0.50	32.48	0.36	32.57	0.40	32.60	0.56	32.56	0.48	32.43	0.38	32.47	0.42	32.46	0.55	32.47	0.50
2	34.76	1.24	32.99	0.46	33.18	0.52	33.17	0.71	33.14	0.58	33.47	0.59	33.67	0.67	33.71	0.81	33.64	0.70
3	30.76	0.78	31.90	0.41	31.95	0.44	31.96	0.66	31.89	0.54	31.85	0.48	31.77	0.56	31.81	0.84	31.76	0.70
4	31.57	1.07	31.71	0.39	31.48	0.43	31.55	0.52	31.54	0.48	32.13	0.58	31.95	0.60	31.99	0.69	31.98	0.66
5	30.51	0.90	31.66	0.44	31.44	0.48	31.33	0.64	31.30	0.57	31.53	0.57	31.32	0.56	31.23	0.66	31.24	0.65
6	33.82	1.22	33.56	0.39	33.93	0.46	34.08	0.59	33.99	0.51	33.72	0.49	34.10	0.56	34.19	0.66	34.13	0.59
7	31.59	0.85	32.53	0.38	32.70	0.41	32.75	0.60	32.70	0.49	32.52	0.48	32.66	0.50	32.73	0.69	32.67	0.56
8	32.25	0.67	32.48	0.38	32.52	0.44	32.50	0.56	32.48	0.54	32.63	0.48	32.70	0.51	32.72	0.68	32.67	0.54
9	32.81	1.18	33.15	0.42	33.27	0.45	33.41	0.63	33.35	0.53	33.42	0.53	33.55	0.57	33.66	0.76	33.59	0.60
10	34.01	0.74	33.15	0.38	33.60	0.48	33.75	0.58	33.64	0.51	33.40	0.43	33.81	0.50	33.93	0.67	33.88	0.54
11	32.75	0.54	32.78	0.26	32.79	0.28	32.89	0.38	32.84	0.34	32.90	0.34	32.94	0.36	33.02	0.50	33.01	0.38
12	30.26	0.80	31.51	0.39	31.25	0.41	31.21	0.52	31.18	0.46	31.39	0.45	31.12	0.44	31.12	0.54	31.06	0.49
13	31.91	0.88	32.35	0.40	32.31	0.43	32.22	0.58	32.20	0.50	32.33	0.48	32.26	0.49	32.21	0.68	32.20	0.56
14	32.37	0.38	32.81	0.20	32.75	0.20	32.87	0.31	32.83	0.22	32.94	0.26	32.92	0.28	32.99	0.35	32.97	0.28
15	33.39	0.50	32.90	0.40	33.14	0.50	33.04	0.63	33.06	0.66	33.12	0.46	33.28	0.48	33.26	0.66	33.21	0.57
16	32.21	0.75	32.72	0.41	32.89	0.47	32.97	0.59	32.90	0.52	32.90	0.53	33.09	0.60	33.17	0.77	33.12	0.64
17	30.88	0.83	31.83	0.41	31.64	0.43	31.58	0.50	31.59	0.50	31.85	0.50	31.65	0.51	31.66	0.65	31.63	0.59
18	31.18	0.80	32.28	0.44	32.37	0.52	32.38	0.64	32.32	0.59	32.05	0.65	32.05	0.67	32.10	0.90	32.02	0.69
19	32.03	0.97	32.20	0.40	32.09	0.43	32.01	0.53	32.04	0.50	32.36	0.51	32.20	0.53	32.16	0.76	32.12	0.60
20	32.71	0.96	32.53	0.41	32.63	0.46	32.64	0.66	32.61	0.55	32.76	0.51	32.81	0.55	32.81	0.73	32.79	0.61
21	33.08	0.98	32.45	0.41	32.53	0.47	32.60	0.66	32.48	0.58	32.91	0.56	32.92	0.59	32.98	0.72	32.96	0.66
22	32.06	0.72	32.63	0.37	32.75	0.42	32.84	0.58	32.75	0.51	32.73	0.45	32.81	0.49	32.90	0.77	32.81	0.54
23	31.18	0.77	31.86	0.40	31.77	0.45	31.71	0.66	31.73	0.59	31.82	0.49	31.71	0.49	31.61	0.63	31.63	0.62
24	32.66	0.66	32.65	0.39	32.71	0.44	32.70	0.58	32.65	0.54	32.78	0.45	32.83	0.48	32.79	0.66	32.82	0.55
25	31.63	0.98	32.33	0.40	32.29	0.46	32.28	0.61	32.23	0.53	32.46	0.53	32.45	0.54	32.49	0.79	32.43	0.61
26	32.02	0.96	32.37	0.41	32.28	0.49	32.22	0.64	32.19	0.60	32.53	0.55	32.53	0.60	32.45	0.75	32.43	0.67
27	31.56	0.44	32.26	0.36	32.21	0.37	32.27	0.49	32.25	0.44	32.26	0.41	32.19	0.44	32.31	0.76	32.22	0.51
28	33.51	1.51	32.90	0.41	33.11	0.49	33.21	0.65	33.16	0.58	33.08	0.55	33.25	0.62	33.41	0.86	33.34	0.73
29	31.53	0.97	32.30	0.47	32.68	0.61	32.67	0.73	32.64	0.70	32.28	0.67	32.56	0.69	32.60	0.85	32.56	0.81
30	30.62	0.94	31.81	0.41	31.72	0.48	31.69	0.65	31.71	0.58	31.77	0.55	31.64	0.61	31.61	0.75	31.62	0.67
31	32.36	0.57	33.12	0.37	33.25	0.40	33.34	0.52	33.31	0.44	32.99	0.43	33.14	0.46	33.22	0.55	33.15	0.50
32	33.24	0.89	32.98	0.37	33.18	0.42	33.31	0.55	33.26	0.46	33.11	0.44	33.29	0.47	33.40	0.66	33.34	0.49
33	30.54	0.51	32.00	0.39	31.91	0.42	31.85	0.53	31.85	0.52	31.74	0.49	31.64	0.52	31.60	0.65	31.56	0.56
34	32.48	0.49	32.81	0.31	32.91	0.34	32.98	0.45	32.99	0.40	32.83	0.35	32.90	0.37	32.97	0.56	32.93	0.42
35	31.78	1.04	32.46	0.38	32.54	0.42	32.54	0.55	32.49	0.50	32.55	0.51	32.61	0.54	32.62	0.75	32.56	0.60

Table 4.4: 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 eight three-level DP models and Bayesian bootstrap

	Bootstrap		NNN		NNDP		NDPN		NDPDP		DPNN		DPNDP		DPDPN		DPDPDP	
	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD	PM	PSD
1	35.52	1.27	35.80	0.43	35.94	0.48	35.95	0.70	35.90	0.56	35.87	0.60	35.97	0.65	35.94	0.96	35.94	0.72
2	40.88	2.32	36.51	0.48	36.69	0.55	36.66	0.79	36.60	0.58	37.33	0.79	37.58	0.91	37.60	1.13	37.51	0.89
3	34.90	2.58	35.23	0.48	35.74	0.66	35.76	0.83	35.67	0.75	35.39	0.83	35.71	1.00	35.78	1.35	35.70	1.09
4	35.59	1.12	35.02	0.43	34.76	0.46	34.87	0.63	34.83	0.52	35.56	0.69	35.41	0.70	35.47	0.92	35.43	0.75
5	35.82	1.61	35.09	0.50	34.91	0.55	34.82	0.78	34.75	0.65	35.33	0.70	35.13	0.78	35.04	0.99	35.01	0.82
6	39.32	1.58	37.14	0.40	37.51	0.48	37.65	0.67	37.55	0.52	37.91	0.52	38.19	0.57	38.28	0.83	38.19	0.58
7	35.93	1.12	35.92	0.43	36.12	0.47	36.16	0.72	36.10	0.55	36.10	0.55	36.28	0.57	36.35	0.90	36.25	0.64
8	37.32	1.49	35.93	0.44	36.01	0.50	35.94	0.66	35.91	0.59	36.50	0.68	36.63	0.72	36.62	1.05	36.56	0.74
9	38.76	1.54	36.62	0.46	36.79	0.47	36.93	0.74	36.84	0.56	37.43	0.57	37.48	0.57	37.61	0.94	37.56	0.61
10	39.82	1.64	36.62	0.44	37.40	0.67	37.59	0.76	37.45	0.68	37.07	0.66	37.85	0.83	38.05	1.09	37.93	0.82
11	37.49	0.94	36.18	0.28	36.17	0.31	36.30	0.50	36.23	0.38	36.71	0.43	36.73	0.43	36.85	0.79	36.79	0.46
12	35.84	1.50	34.97	0.45	34.69	0.50	34.65	0.67	34.59	0.54	35.51	0.64	35.22	0.70	35.26	0.93	35.14	0.76
13	36.13	1.20	35.67	0.44	35.66	0.48	35.57	0.66	35.52	0.52	35.84	0.61	35.78	0.60	35.79	0.97	35.72	0.66
14	36.90	0.80	36.15	0.23	36.05	0.23	36.23	0.47	36.17	0.28	36.89	0.44	36.84	0.46	36.97	0.70	36.93	0.45
15	36.04	1.47	36.07	0.48	36.41	0.63	36.28	0.75	36.28	0.73	36.15	0.63	36.39	0.76	36.39	1.11	36.29	0.79
16	36.44	1.40	36.09	0.45	36.32	0.53	36.40	0.71	36.30	0.57	36.53	0.65	36.79	0.76	36.86	1.05	36.79	0.74
17	34.70	0.99	35.13	0.45	34.95	0.49	34.90	0.59	34.88	0.54	35.26	0.56	35.08	0.59	35.08	0.89	35.04	0.69
18	35.57	0.81	35.65	0.44	35.89	0.57	35.92	0.73	35.83	0.63	35.70	0.52	35.83	0.59	35.89	1.05	35.79	0.62
19	34.88	0.88	35.42	0.43	35.37	0.46	35.32	0.62	35.33	0.54	35.45	0.54	35.34	0.58	35.34	1.09	35.24	0.62
20	37.08	1.89	35.86	0.47	36.01	0.53	36.03	0.78	35.98	0.60	36.27	0.72	36.45	0.76	36.44	1.17	36.38	0.84
21	35.75	1.03	35.62	0.43	35.84	0.55	35.87	0.77	35.73	0.60	35.86	0.60	35.95	0.67	35.97	0.96	35.95	0.70
22	35.56	1.08	35.90	0.43	36.12	0.52	36.15	0.76	36.04	0.58	36.07	0.63	36.23	0.69	36.32	1.16	36.22	0.74
23	36.46	1.46	35.33	0.44	35.33	0.51	35.26	0.79	35.27	0.65	35.67	0.61	35.60	0.67	35.56	0.90	35.55	0.77
24	37.80	2.17	36.01	0.45	36.09	0.51	36.06	0.71	35.99	0.61	36.57	0.77	36.55	0.75	36.57	1.10	36.56	0.83
25	37.29	2.60	35.70	0.44	35.67	0.51	35.66	0.72	35.57	0.56	36.15	0.73	36.13	0.75	36.23	1.23	36.12	0.90
26	36.18	1.92	35.69	0.52	35.61	0.60	35.53	0.79	35.47	0.69	35.97	0.85	35.97	0.87	35.88	1.14	35.85	0.95
27	36.09	1.30	35.63	0.42	35.62	0.44	35.66	0.62	35.63	0.50	36.02	0.53	35.96	0.60	36.08	1.09	35.96	0.64
28	40.33	1.37	36.53	0.46	36.84	0.59	36.90	0.76	36.85	0.65	37.62	0.79	37.80	0.78	37.98	1.12	37.87	0.88
29	35.71	1.10	35.69	0.51	36.27	0.71	36.26	0.84	36.21	0.78	35.94	0.72	36.37	0.74	36.41	1.01	36.32	0.84
30	34.57	1.11	35.14	0.46	35.11	0.53	35.08	0.77	35.08	0.62	35.26	0.67	35.19	0.70	35.19	1.01	35.15	0.75
31	35.43	1.06	36.36	0.40	36.49	0.44	36.60	0.66	36.56	0.49	36.29	0.47	36.41	0.54	36.53	0.82	36.42	0.54
32	39.12	1.40	36.45	0.41	36.66	0.48	36.80	0.68	36.72	0.51	36.96	0.65	37.13	0.68	37.27	0.97	37.19	0.66
33	34.10	0.83	35.26	0.44	35.19	0.47	35.13	0.66	35.12	0.55	35.16	0.60	35.07	0.67	35.06	0.95	35.02	0.67
34	35.98	1.02	36.10	0.36	36.24	0.40	36.31	0.57	36.32	0.46	36.25	0.48	36.37	0.53	36.46	0.90	36.39	0.60
35	37.83	1.13	36.01	0.41	36.16	0.46	36.19	0.63	36.12	0.53	36.68	0.59	36.84	0.64	36.90	1.02	36.83	0.70

	BMI data (three-level models)			
	LML	LPML	$P_{CPO<.025}$	$P_{CPO<.014}$
NNN	−8964.996	−9230.771	0.0257	0.0192
NNDP	−10964.55	-9307.957	0.0274	0.0196
NDPN	−9189.928	-9325.564	0.0266	0.0193
NDPDP	−9248.959	-9284.62	0.0271	0.0193

Table 4.5: Log of the marginal likelihood (LML) with Monte Carlo errors, Log pseudo marginal likelihood (LPML) and percentages of conditional predictive ordinate (CPO) less than .025 ($P_{CPO<.025}$) and .014 ($P_{CPO<.014}$) for body mass index (BMI) data under the NNN, NNDP, NDPN, NDPDP model

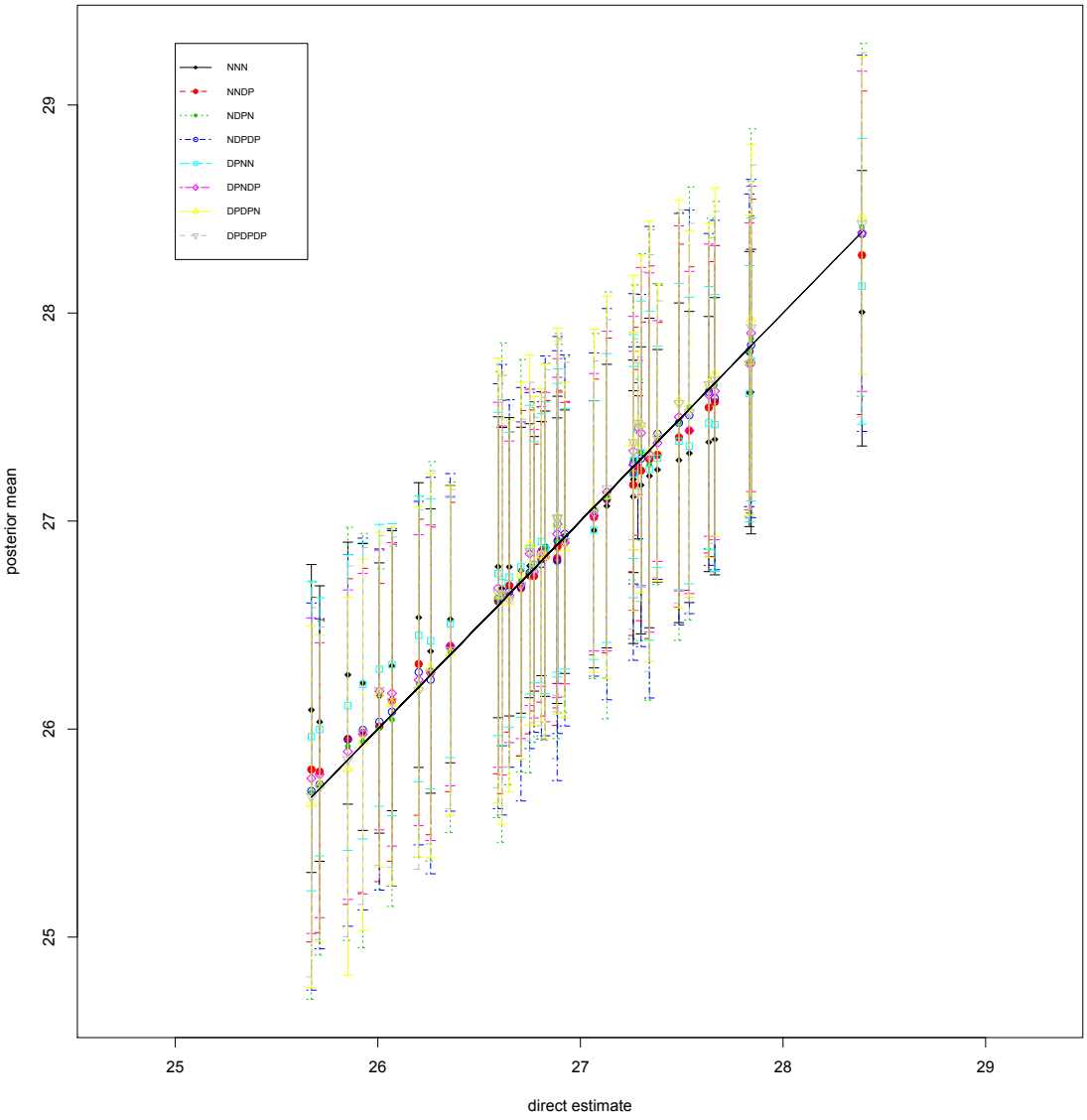

Figure 4.1: Comparison for body mass index (BMI) data (posterior means with credible bands versus direct estimates): the predictive inference of the finite population mean for each county under eight three-level DP models

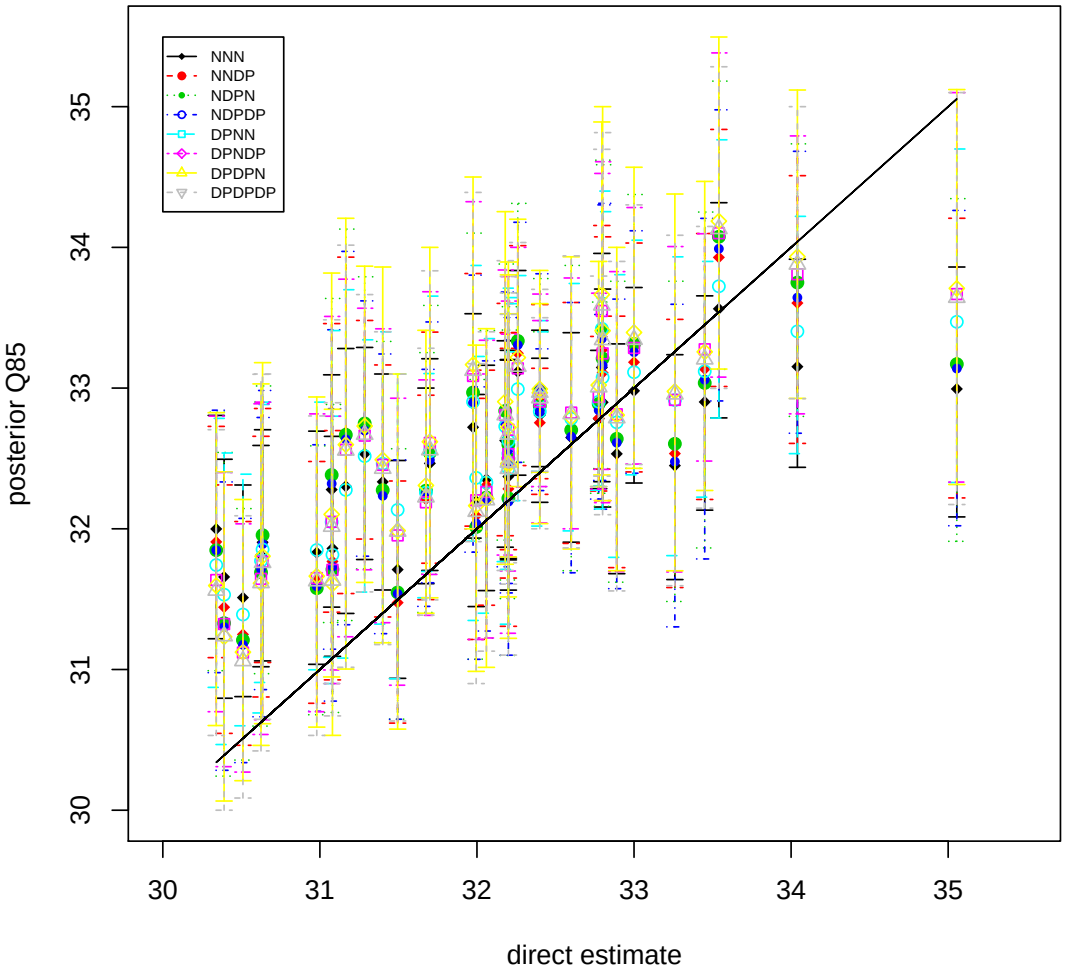

Figure 4.2: Comparison for body mass index (BMI) data (posterior mean with credible bands versus direct estimates): the predictive inference of the finite population 85^{th} percentile for each county under eight three-level DP models

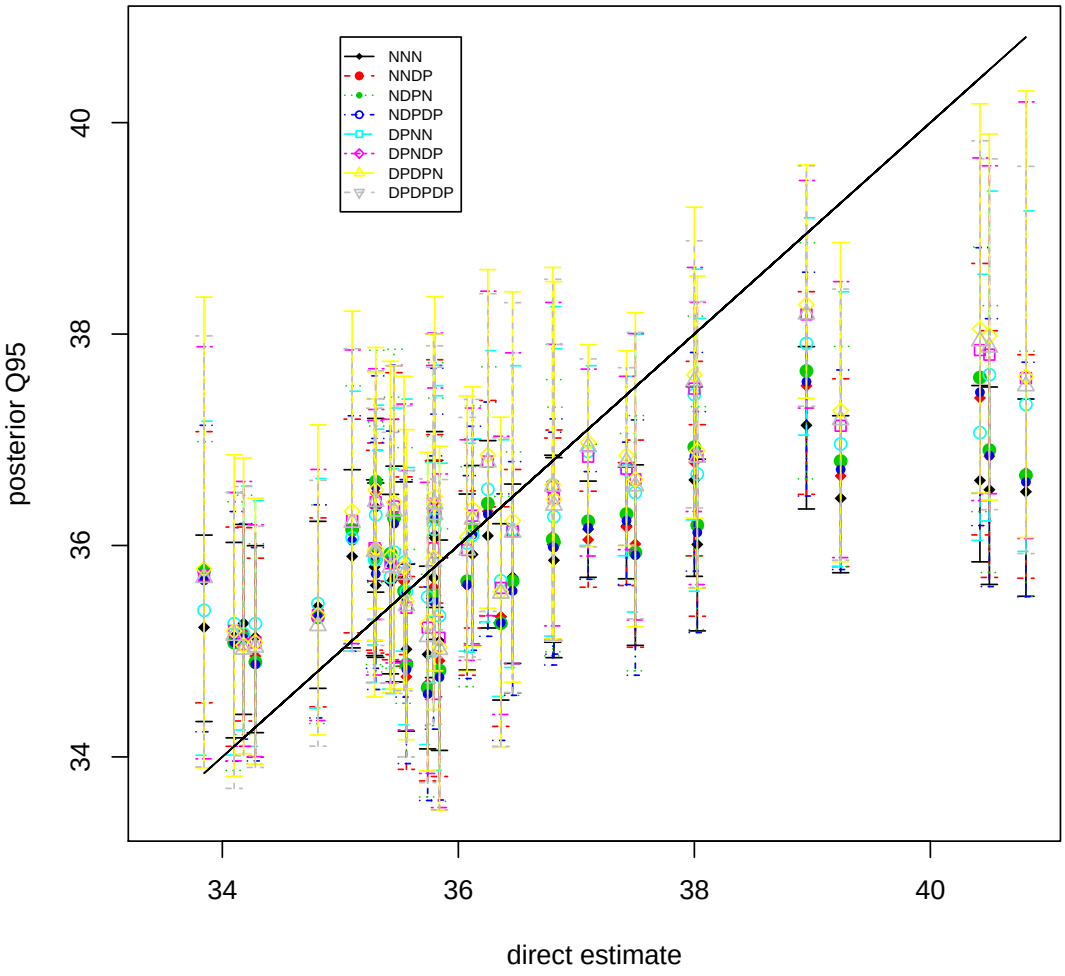

Figure 4.3: Comparison for body mass index (BMI) data (posterior mean with credible bands versus direct estimates): the predictive inference of the finite population 95^{th} percentile for each county under eight three-level DP models

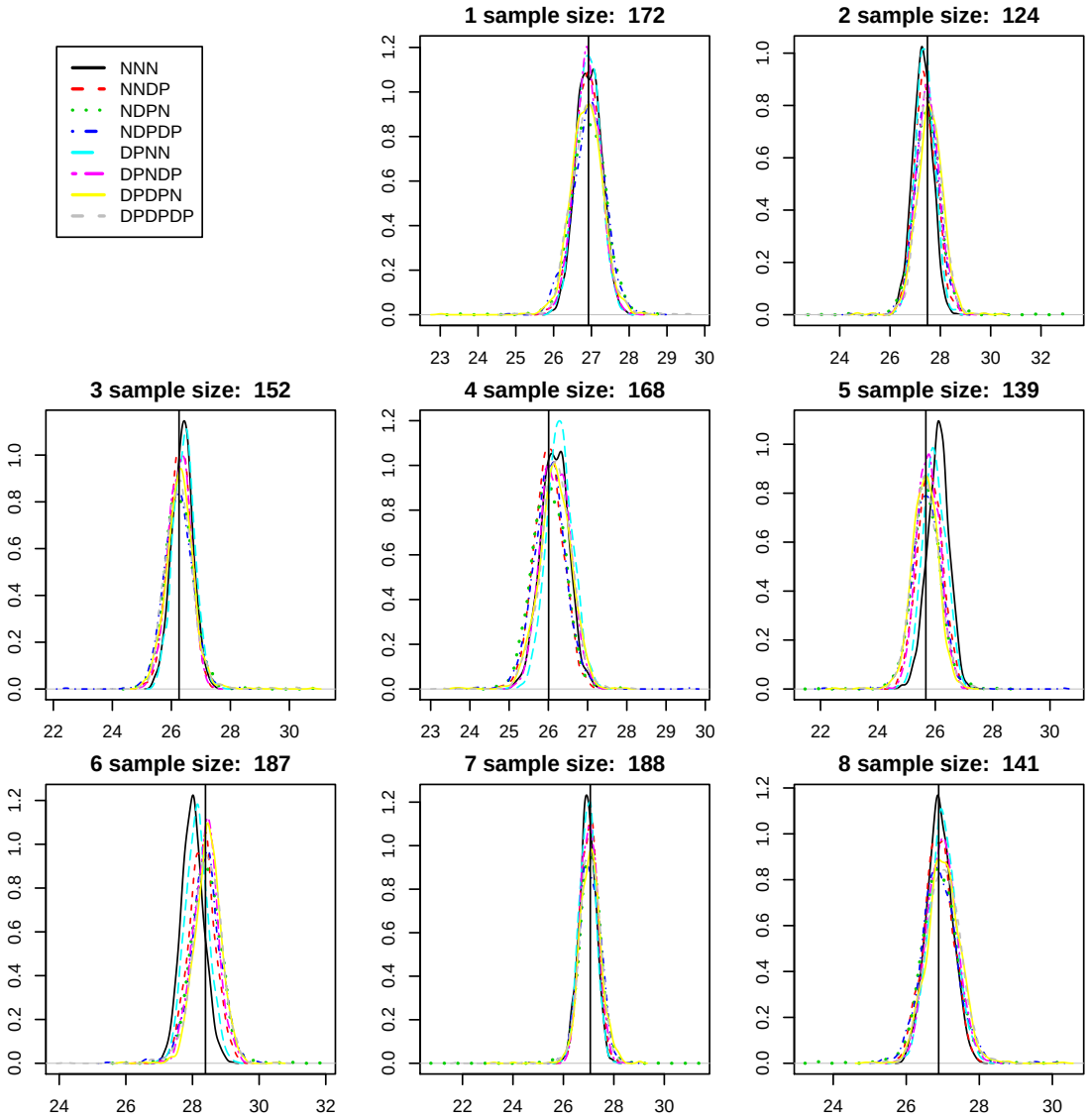

Figure 4.4: Plots of the posterior density of the finite population mean by eight three-level DP models for the first eight counties of body mass index (BMI) data

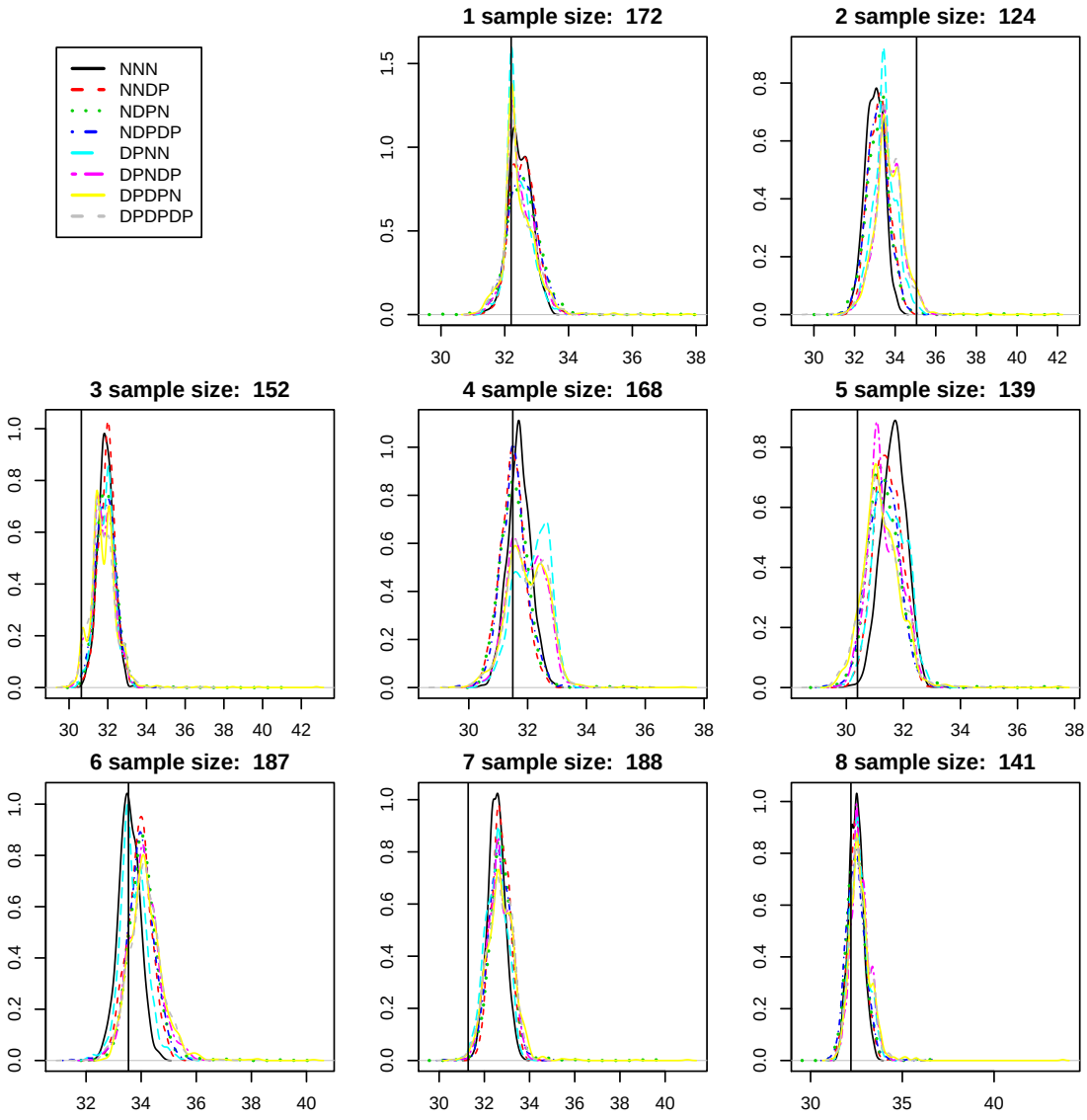

Figure 4.5: Plots of the posterior density of the finite population 85^{th} percentile by eight three-level DP models for the first eight counties of body mass index (BMI) data

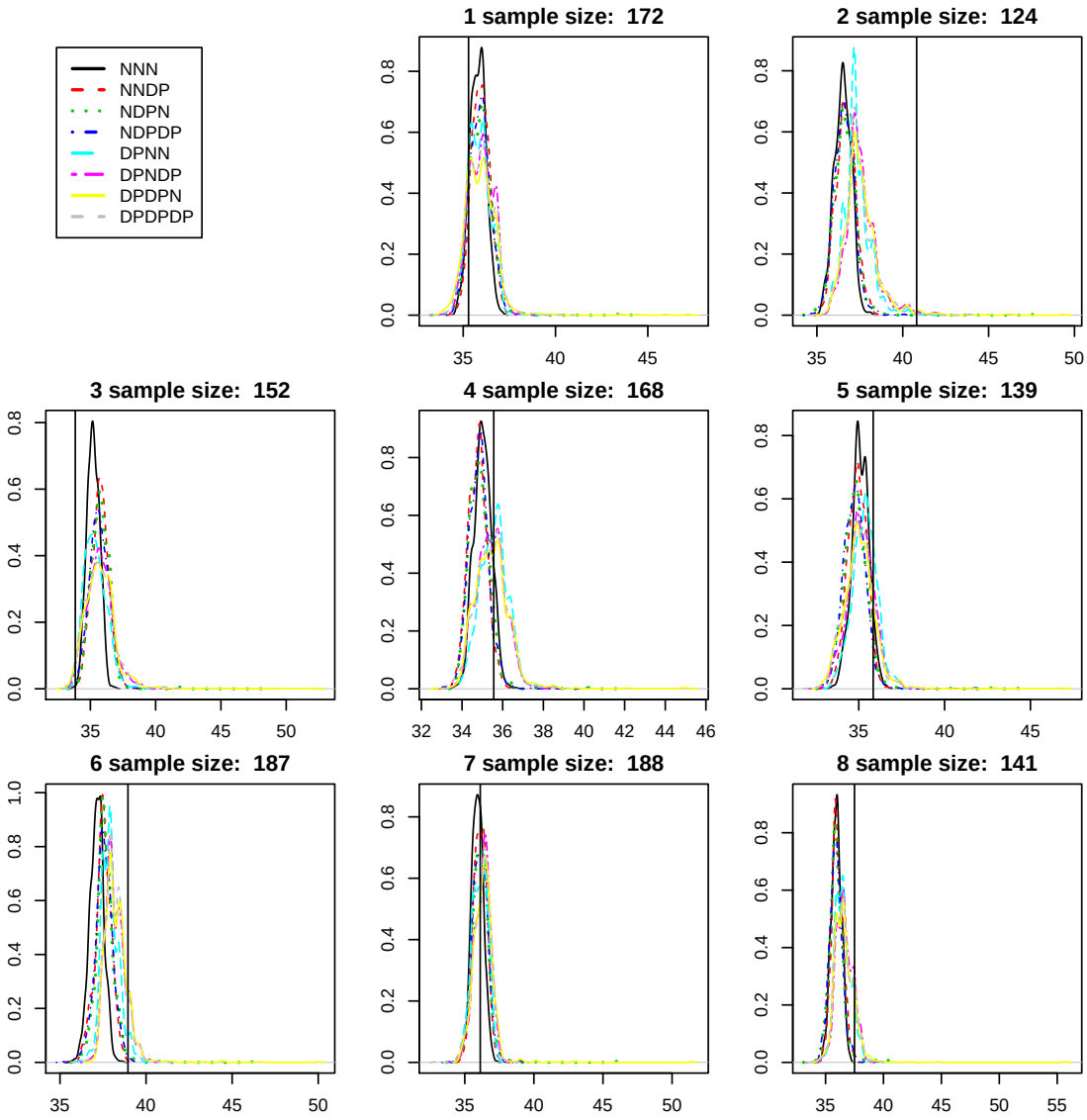

Figure 4.6: Plots of the posterior density of the finite population 95^{th} percentile by eight three-level DP models for the first eight counties of body mass index (BMI) data

Chapter 5

Concluding Remarks and Future Work

If the parametric distribution assumption does not hold, the model is misspecified and the inference may be invalid. The Bayesian nonparametric methods are motivated by the desire to avoid overly restrictive assumptions. We have proposed several nonparametric models for multi-stage survey data using DPs. We extend the two-level DP models to three-level DP models and also can naturally extend to multi-stage (more than three stages) sampling. The predictive inference and comparison are conducted. The results of an illustrated example and a small stimulation study are given. In Chapter 5, we compare the results of BMI data under two- and three-level models, summarize our findings and discuss some future problems.

5.1 Comparison of Two- and Three-level Models

It is possible that the fitted model has two-stage hierarchical structure while the data may come from a model with three-stage structure. We compare the two-

and three-level models for BMI data. We select the best candidates in models using DPs, the DPDP and DPNDP models, then compare them to the parametric baseline models, the normal and NNN models. We plot the results under these four models along with results under Bayesian bootstrap.

In Figure 5.1, the predictions of the population means under the normal model are mostly biased. The posterior means under the NNN model are slightly closer to the direct estimates due to the introduction of the additional hierarchical structure. However, the parametric model assumptions may be incorrect resulting in misleading conclusions. The two-level nonparametric alternative, the DPDP model, results in large reductions of bias together with similar or even smaller variation for some areas comparing to the baseline models. The best three-level nonparametric candidate results in further reduction of bias, however with increasing of variation.

Figure 5.2 gives the plots of the estimated posterior density of the finite population means by the normal, DPDP, NNN, DPNDP models and Bayesian bootstrap for the first eight counties as examples. The same observations as in Figure 5.1 can be obtained that the DPNDP model gives almost unbiased estimations however with the sacrifice of variations. The DPDP model has the smallest variation for most of the areas with small bias. Maybe the three-level structure is redundant for this data set, and the two-level model using DPs is sufficient.

In general, we need diagnostic techniques when the fitted model includes some hierarchical structure, but the data are from a model with additional, unknown hierarchical structure (Yan and Sedransk 2007; Yan and Sedransk 2010). It is important to detect unknown hierarchical structure and check model assumptions under parametric models. It seems promising that the use of DPs in the models can reduce the bias with manageable penalty in terms of variation. Antonelli, Trippa and Haneuse (2016) pointed out similar findings when the DP prior is used in modeling the

random effect distribution in a logistic generalized linear mixed model for repeated measures binary data. Thus, the robust nonparametric models are recommended especially where there is little knowledge of the distribution or hierarchical structure of the data.

5.2 Future Work

We describe nonparametric alternatives with the normal baseline parametric model assumed. Other parametric baseline distributions instead of normal distribution are possible. For example, for size data, a gamma distribution as the baseline distribution may be desired. For the two-level DP model, one may write

$$\begin{aligned} y_{ij}|G_i &\overset{ind}{\sim} G_i, \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, N_i, \\ G_i|\mu_i &\overset{ind}{\sim} \mathrm{DP}\{\alpha_i, \mathrm{Gamma}(a, a/\mu_i)\}, \\ \mu_i|H &\overset{iid}{\sim} H, \\ H &\sim \mathrm{DP}\{\gamma, G_0\}, \end{aligned} \tag{5.1}$$

where μ_i is the mean of the gamma random variable.

It is important to study sensitivity to posterior inference, not only in the prior specifications, but also in the baseline model. As pointed out by Nandram and Yin (2016a) and others, posterior inference in the DP is sensitive to the specification of the baseline model. A more robust specification is needed; it is obvious that using a DP for the baseline distribution is not sensible. There is sensitivity to the prior specifications as well. Also recently Bayesian models have been called "brittle" especially for problems with infinite number of parameters (Owhadi et al. 2015).

As we mentioned in previous chapters, we have some difficulties in inference of

the population quantiles and computation of Bayes factors, when a sample from a DP. One possible explanation of this fact is that the DP generates discrete distributions with probability one. This phenomenon can arise, more generally, in different contexts, e.g. using the DP in goodness of fit testing (Carota and Parmiginani 1996). Petrone and Raftery (1997) pointed out that the discreteness of the DP can have a large effect on inferences (posterior distributions and Bayes factors), when the data are partially exchangeable with an unknown partition. One possible solution is by introducing the nugget effect (Gelfand et al. 2005). Another alternative is to use Polya trees (e.g., Lavine 1992), a generalization of the DP. This needs further investigations.

For future work, we may also include the covariates in the model. Battese, Harter and Fuller (1988) extended the Scott-Smith model (3.1) to include covariates, assuming

$$\begin{aligned} y_{ij}|\nu_i &\overset{ind}{\sim} N\left(\boldsymbol{x'_{ij}}\boldsymbol{\beta}+\nu_i,\sigma^2\right), \quad i=1,\ldots,\ell, \quad j=1,\ldots,N_i, \\ \nu_i &\overset{iid}{\sim} N\left(0,\delta^2\right), \end{aligned} \tag{5.2}$$

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

$$y_{ij} \mid \boldsymbol{\beta},\nu_i,\sigma^2 \overset{ind}{\sim} N(\boldsymbol{x'_{ij}}\boldsymbol{\beta}+\nu_i,\sigma^2), \quad i=1,\ldots,\ell, \quad j=1,\ldots,N_i, \tag{5.3}$$

$$\nu_i|\mathrm{G} \overset{iid}{\sim} \mathrm{G} \tag{5.4}$$

$$\mathrm{G} \sim DP\left[\alpha, N\left(0,\frac{\rho}{1-\rho}\sigma^2\right)\right] \tag{5.5}$$

$$\pi(\underset{\sim}{\beta},\sigma^2,\rho) \propto 1/\sigma^2, \quad \boldsymbol{\beta}\in R^p, \quad \sigma^2>0, \quad 0<\rho<1, \tag{5.6}$$

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

covariates can be

$$
\begin{aligned}
y_{ij} - \boldsymbol{x'_{ij}}^{(0)}\boldsymbol{\beta}^{(0)}|G_i &\overset{ind}{\sim} G_i, \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, N_i, \\
G_i|\beta_{0i} &\overset{ind}{\sim} \mathrm{DP}\left\{\alpha_i, N(\beta_{0i}; \sigma^2)\right\}, \\
\beta_{0i}|H &\overset{iid}{\sim} H, \\
H &\sim \mathrm{DP}\left\{\gamma, N(\theta, \delta^2)\right\},
\end{aligned} \tag{5.7}
$$

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

In many complex surveys, there are also survey weights. We may include them as covariates in the model, 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 the surrogate sampling (Nandram 2007).

There are also other possible datasets to explore. For example, Behavioral Risk Factor Surveillance System (BRFSS) is the world's largest, on-going telephone health survey system, tracking health conditions and risk behaviors among adults in all 50 states and selected territories. In the Trends in International Mathematics and Science Study (TIMSS), one can consider mathematics or science test scores along with other covariates. We have worked on the public-used TIMSS data, however, it is really masked data drawn from normal distributions. The results under the nonparametric model are very similar to the results under the normal models. One may proceed to the restricted data for further investigations.

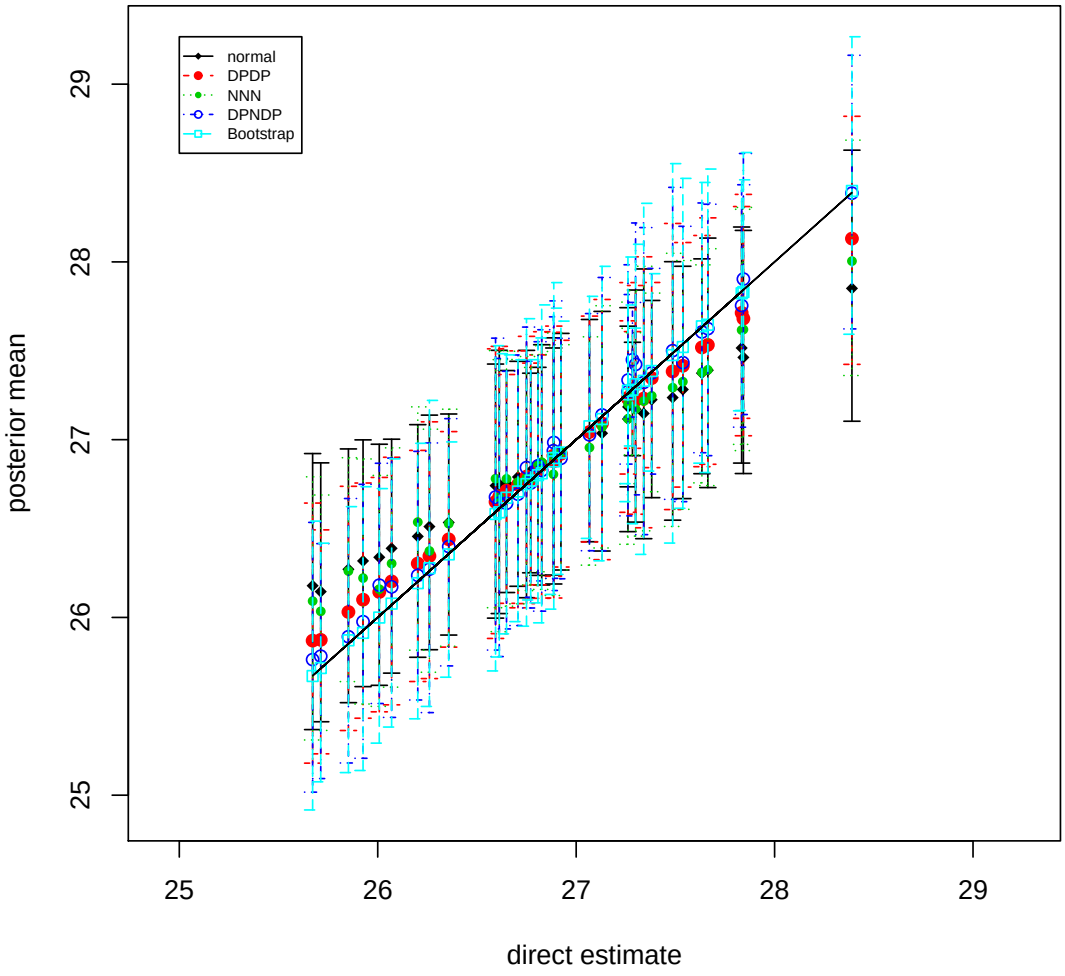

Figure 5.1: Comparison for body mass index (BMI) data (posterior means with credible bands versus direct estimates): the predictive inference of the finite population mean for each county under the normal, DPDP, NNN, DPNDP models and Bayesian bootstrap

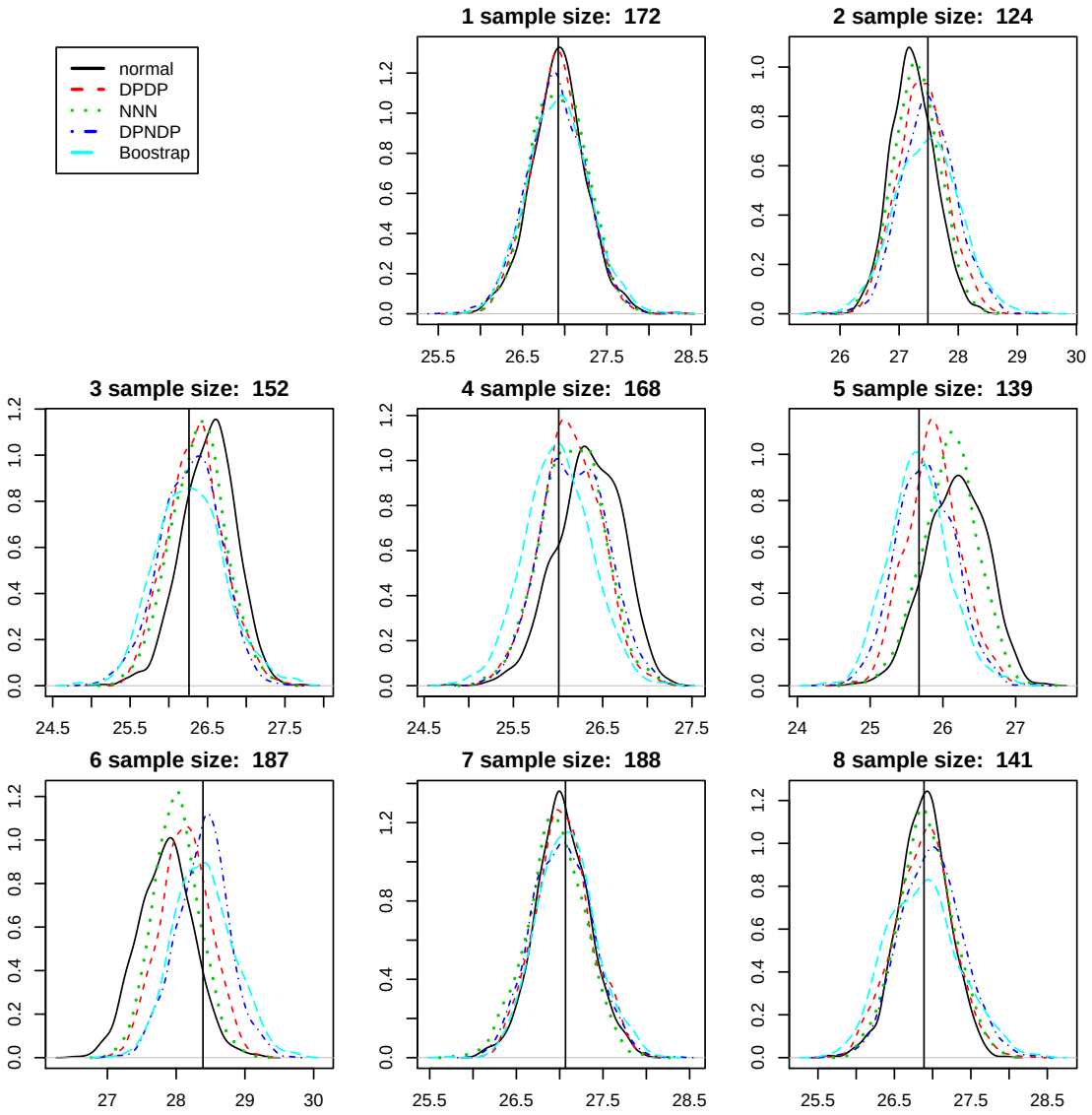

Figure 5.2: Plots of the posterior density of the finite population mean by the normal, DPDP, NNN, DPNDP models and Bayesian bootstrap for the first eight counties of body mass index (BMI) data

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover Publications, New York, 1965.

[2] M. Aitkin. *Statistical Inference: An Integrated Byesian/Likelihood Approach*. CRC Press, 2010.

[3] D. J. Aldous. *Exchangeability and Related Topics*. Springer, 1985.

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

[5] C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[6] A. Azzalini. *The Skew-normal and Related Families*, volume 3. Cambridge University Press, 2013.

[7] S. Basu and S. Chib. Marginal likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association*, 98(461):224–235, 2003.

[8] G. E. Battese, R. M. Harter, and W. A. Fuller. 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, 1988.

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

[10] D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *the Journal of Machine Learning Research*, 3:993–1022, 2003.

[12] M. J. Brewer. A Bayesian model for local smoothing in kernel density estimation. *Statistics and Computing*, 10(4):299–309, 2000.

[13] C. Carota. Some faults of the Bayes factor in nonparametric model selection. *Statistical Methods and Applications*, 15(1):37–42, 2006.

[14] C. Carota and G. Parmigiani. On Bayes factor for nonparametric alternatives. *Bayesian Statistics*, 5:507–511, 1996.

[15] S. Chaudhuri and M. Ghosh. Empirical likelihood for small area estimation. *Biometrika*, 98(2):473–480, 2011.

[16] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

[17] D. B. Dunson. Nonparametric Bayes local partition models for random effects. *Biometrika*, 96(2):249–262, 2009.

[18] W. A. Ericson. Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 195–233, 1969.

[19] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

[20] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[21] T. S. Ferguson. Bayesian density estimation by mixtures of normal distributions. *Recent Advances in Statistics*, 24(1983):287–302, 1983.

[22] S. Geisser. Discussion on sampling and Bayes' inference in scientific modelling and robustness (by G.E.P. Box). *Journal of the Royal Statistical Society. Series A (General)*, pages 383–430, 1980.

[23] A. E. Gelfand, A. Kottas, and S. N. MacEachern. Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035, 2005.

[24] W. Hardle. *Smoothing Techniques: With Implementation in S*. Springer, New York, 1991.

[25] S. Hu, D. Poskitt, and X. Zhang. Bayesian adaptive bandwidth kernel density estimation of irregular multivariate distributions. *Computational Statistics & Data Analysis*, 56(3):732–740, 2012.

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

[27] M. Kalli, J. E. Griffin, and S. G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011.

[28] L. Kuo. Computations of mixtures of Dirichlet processes. *SIAM Journal on Scientific and Statistical Computing*, 7(1):60–71, 1986.

[29] K. L. Lange, R. J. Little, and J. M. Taylor. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989.

[30] N. Lartillot and H. Philippe. Computing Bayes factors using thermodynamic integration. *Systematic Biology*, 55(2):195–207, 2006.

[31] M. Lavine. Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, pages 1222–1235, 1992.

[32] J. S. Liu. Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics*, pages 911–930, 1996.

[33] A. Y. Lo. On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, 12(1):351–357, 1984.

[34] D. Malec and P. Müller. A Bayesian semi-parametric model for small area estimation. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, volume 3, pages 223–236. Institute of Mathematical Statistics, 2008.

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

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

[37] I. Molina, B. Nandram, and J. Rao. Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *The Annals of Applied Statistics*, 8(2):852–885, 2014.

[38] P. Müller, F. Quintana, and G. Rosner. A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):735–749, 2004.

[39] B. Nandram. 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*, pages 356–374, 2007.

[40] B. Nandram and J. W. Choi. Nonparametric Bayesian analysis of a proportion for a small area under nonignorable nonresponse. *Journal of Nonparametric Statistics*, 16(6):821–839, 2004.

[41] B. Nandram and H. Kim. Marginal likelihood for a class of Bayesian generalized linear models. *Journal of Statistical Computation and Simulation*, 72(4):319–340, 2002.

[42] B. Nandram, M. C. S. Toto, and J. W. Choi. A Bayesian benchmarking of the Scott–Smith model for small areas. *Journal of Statistical Computation and Simulation*, 81(11):1593–1608, 2011.

[43] B. Nandram and J. Yin. Bayesian predictive inference under a Dirichlet process with sensitivity to the normal baseline. *Statistical Methodology*, 28:1–17, 2016a.

[44] B. Nandram and J. Yin. A nonparametric Bayesian prediction interval for a finite population mean. *Journal of Statistical Computation and Simulation*, pages 1–17, 2016b.

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

[46] I. Ntzoufras. *Bayesian Modeling using WinBUGS*. Wiley, Hoboken, NJ, 2009.

[47] H. Owhadi, C. Scovel, and T. Sullivan. On the brittleness of Bayesian inference. *SIAM Review*, 57(4):566–582, 2015.

[48] O. Papaspiliopoulos and G. O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

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

[50] N. G. Polson and J. G. Scott. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012.

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

[52] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[53] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26. CRC press, 1986.

[54] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

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

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

[57] S. G. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics. Simulation and Computation*, 36(1-3):45–54, 2007.

[58] J. C. Wang, S. H. Holan, B. Nandram, W. Barboza, C. Toto, and E. Anderson. A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, 17(1):84–106, 2012.

[59] G. Yan and J. Sedransk. A note on Bayesian residuals as a hierarchical model diagnostic technique. *Statistical Papers*, 51(1):1–10, 2010.

[60] G. Yan, J. Sedransk, et al. Bayesian diagnostic techniques for detecting hierarchical structure. *Bayesian Analysis*, 2(4):735–760, 2007.

[61] J. Yin and B. Nandram. Rapid prediction methods under the one-level Dirichlet process model. (Working paper).