STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 3, pp. 21–36, DOI 10.2478/stattrans-2022-0040
Received – 26.01.2021; accepted – 14.08.2022

Sampling methods for the concentration parameter and discrete baseline of the Dirichlet Process

Yang Liu[1], **Balgobin Nandram**[2]

ABSTRACT

There are many models in the current statistical literature for making inferences based on samples selected from a finite population. Parametric models may be problematic because statistical inference is sensitive to parametric assumptions. The Dirichlet process (DP) prior is very flexible and determines the complexity of the model. It is indexed by two hyperparameters: the baseline distribution and concentration parameter. We address two distinct problems in the article. Firstly, we review the current sampling methods for the concentration parameter, which use the continuous baseline distribution. We compare three different methods: the adaptive rejection method, the mixture of Gammas method and the grid method. We also propose a new method based on the ratio of uniforms. Secondly, in practice, some survey responses are known to be discrete. If a continuous distribution is adopted as the baseline distribution, the model is misspecified and standard inference may be invalid. We propose a discrete baseline approach to the DP prior and sample the unobserved responses from the finite population both using a Polya urn scheme and a Multinomial distribution. We applied our discrete baseline approach to a Phytophthora data set.

Key words: concentration parameter, discrete baseline, empirical study, grid method, nonparametric Bayesian statistics.

1. Introduction

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. The existence of the Dirichlet Process (DP) was established by Ferguson (1973) and further developed by Blackwell and MacQueen (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). In this paper we provide an improved method to sample the concentration parameter and show that it is affected by a discrete baseline.

Another representation of the DP is the generalized Polya urn scheme (Blackwell and MacQueen, 1973). We consider two urns. Urn I is empty and Urn II contains an infinite number of balls, each with a different colour. Pick a ball from Urn II and put it in Urn I. For the next ball, we draw Urn I with probability $\frac{1}{\alpha+1}$. If Urn I is selected, we replace

[1]Worcester Polytechnic Institute, USA. E-mail: yliu22@wpi.edu. ORCID: https://orcid.org/0000-0003-1414-0349.
[2]Worcester Polytechnic Institute, USA. E-mail: balnan@wpi.edu. ORCID: https://orcid.org/0000-0002-3204-0301.

the selected ball with two balls of the same colour, and if Urn II is selected, we take a ball and place it into Urn I. This procedure is repeated until n balls are in Urn I; this is the sample. We observe that with positive probability draws from G (distribution of Urn I) can take the same value regardless of the smoothness of G_0 (distribution of Urn II). That is, G is a discrete distribution with probability one. Current literature have been using smooth functions such as Gaussian distribution as G_0; however this is not always reasonable. This paper will explore a different choice of G_0.

To sample the concentration parameter α of the DP is still an open topic. One can use Gilks' (1992) adaptive rejection sampling method, which relies on the logconcavity of the distribution of the logarithmic transformation of α. Nandram and Yin (2016 a, b) used a grid method to sample α from the posterior density of $\rho = 1/(1+\alpha)$; they have used a noninformative prior for α, different from the proper (informative) prior suggested by Escobar and West (1995). Antonelli, Trippa and Haneuse (2016) reviewed several methods and suggested a more complex method. The problem of sampling the posterior density of α is a difficult one. One of the reasons why it is difficult to estimate α is because it is based on a 'single' observation, k. There are no repeated sampling. So there will be computational instability. There is some research in which the authors set $\alpha = 1$ (e.g. Chaudhuri and Ghosh 2011) to overcome the difficulty in estimating α, thereby leading to an underestimation in variability. In this paper we will propose a new method based on the ratio of uniforms in random sampling.

Another concern that will be addressed is regarding the discreteness of the baseline distribution G_0. It is well known that inference is sensitive to the specification of baseline measure (e.g. McAuliffe, Blei and Jordan 2006 and Nandram and Yin 2016 a). So it is more robust if we have an unspecified distribution G_0. Camerlenghi et al. (2019) discussed ties across samples at the observed or latent level. In the discrete case we mention here, an observation can look like a tie, but it may not be. We are not actually talking about ties, although it is a part of what we are doing. The discreteness of G_0 means that the same value can come from either G_0 or from the balls already drawn in the Polya urn scheme. But it is mandatory to have G_0 discrete in this model if we have a strong belief that the observations are from a discrete family. In such a case, the number of distinct values in the sample, k, is no longer a sufficient statistic for α. This paper will correct this.

We demonstrate our discrete baseline approach to Phytophthora epidemics in bell pepper. The pathogen Phytophthora Capsici Leonian is a severe infectious disease and could rapidly cause death of the plant (Gumpertz 1997). Disease presence or absence was recorded for each cell in a 20×20 quadrats study field. We group the quadrats and count the number of diseased plants in each group so we know the response is guaranteed to be discrete. Now our goal is to obtain an estimator of the finite population mean provided by a nonparametric approach. It is apparent that this approach is more robust than the parametric models such as those based on normality. On the other hand, current nonparametric methods are all based on continuous baseline distribution (i.e. normal baseline). Our approach, with relaxation to the baseline distribution, gives a more realistic estimator when we know the response is discrete.

This paper is an extension of Nandram and Yin (2016 a,b), who studied the sensitivity of the baseline distribution to the finite population mean. They proposed the DP approach to

predict the nonsampled observations by using the Polya urn scheme. We choose a discrete baseline to the DP when the response is known to be from a discrete family. When G_0 is discrete, the number of distinct values in the sample is no longer a sufficient statistic of the concentration parameter α. We proposed a way to correct this by adding a latent variable to indicate which urn the observation is from.

When faced with a discrete baseline, researchers might resort to a DP-based mixture model (DPM) involving a continuous kernel density, however, this is not what we are trying to discuss here. Note that DPMs are often miscalled as "mixture of Dirichlet process model" (Neal 2000). There have been many computational methods to run the model over the past two decades (e.g. Escobar and West 1995, Neal 2000 and Kalli, Griffin and Walker 2011). The DPM is not appropriate in some applications like the example we discuss in this paper because we do not have well defined groups of data. For the DPM, we need different groups of data with different parameters and then a DP is assigned to these parameters. Of course, in applications the DPM is the workhorse of nonparametric Bayesian statistics, yet we need to solve the problem associated with discrete baseline distributions as they may be included as a step in a hierarchical Bayesian model.

The plan of this paper is as follows. In Section 2, we briefly review the Dirichlet process (DP), and different sampling algorithms for α, the concentration parameter. We also introduce our approach, the ratio of uniforms algorithm, and a simulation study to compare the different methods. In Section 3, we discuss one limitation that current literature has regarding the baseline distribution of the DP and how we resolve it. We also discuss the implementation of our method to the finite population mean. In Section 4, we discuss an illustrative example on Phytophthora data. We conclude this paper in Section 5. An appendix has technical details.

2. Dirichlet Process and Sampling the Concentration Parameter

In Section 2.1, we give a brief review of the Dirichlet process, and in Section 2.2, we review current methods to sample the posterior density of α. In Section 2.3 we present our new method based on the ratio of uniforms. In Section 2.4, we provide a small simulation study to compare our new method with few selected ones that we review.

2.1. Review of the Dirichlet Process

Let $(\Theta, \mathscr{B})$ be a measurable space, with G_0 the baseline measure (nonrandom) on the space, and let α be the concentration parameter. A Dirichlet process, $\mathrm{DP}(\alpha, G_0)$, is defined as the distribution of a random probability measure G over $(\Theta, \mathscr{B})$ such that, for any finite measurable partition of the measurable space, $(\Theta, \{A_i\}_{i=1}^n)$, with $A_i \bigcap A_j = \phi$, $\bigcup_{i=1}^n A_i = \Theta$,

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

We write $G \sim \mathrm{DP}(\alpha, G_0)$, if G is a random probability measure with a distribution given by the DP. For any measurable set, A, 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 (i.e. the DP concentrates more of its mass around the baseline distribution). Here, G_0 and α are both parameters and they play intuitive roles in the definition of the DP. Here, G is constrained to be around G_0 and this is regulated by α.

Let $G \sim \text{DP}(\alpha, G_0)$ and $y_1, \cdots, y_n$ be a sequence of independent draws from G. The posterior distribution, $G|y_1, \cdots, y_n$ is

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

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

For a one-sample problem, one might take

$$Y_1, \cdots, Y_n|G \sim G,$$

$$G \sim DP(\alpha, G_0),$$

where G_0 is the baseline measure and α the concentration parameter. Assuming that there are k distinct values among $Y_1, \cdots, Y_n$, the baseline model is $Y_1^*, \cdots, Y_k^*|k \sim G_0$. Note that k is a random variable. The baseline measure G_0 is assumed continuous. Binder (1982) was the first to introduce this model to survey sampling; more recently, see Nandram and Yin (2016 a,b). Although G_0 can be discrete, it appears that this latter case was not discussed by Antoniak (1974).

Antoniak (1974) wrote down the distribution of the number of distinct values k given α and he proved that k is a sufficient statistic for α where G_0 is continuous. It is easy to write down the posterior density with an appropriate prior. The sampling methods being discussed in this section are all based on continuous baseline. We write here that

$$p(k|\alpha) = C \cdot \frac{\Gamma(\alpha)\alpha^k}{\Gamma(\alpha+n)}, \;\; k = 1, \cdots, n,$$

where C is a constant.

However, if G_0 is discrete, k is no longer a sufficient statistic; this result appears to be not so well known. Therefore, if the result is used, this is a violation of the sufficiency principle; we will discuss this issue in Section 3.

2.2. Current Sampling Methods

In this section, we will discuss three current sampling methods for the concentration parameter: the adaptive reject sampling method (ARS), the mixture of Gamma method and the grid method.

We first review the ARS method (Gilks 1992).

Theorem. Let $\phi = \ln(\alpha)$, where α is the concentration parameter. With a logconcave prior $\pi(\phi)$, the posterior density $\pi(\phi|k)$ is logconcave, (i.e. strongly unimodal with a unique mode).

Rasmussen (2000) first demonstrated the logconcavity of $\pi(\phi|k)$ but here we provide

our own proof in the appendix. More generally, we show that if prior $\pi(\phi)$ is logconcave, i.e. $\frac{d^2\ln(\pi(\phi))}{d\phi^2} < 0$, then the posterior density on the transformed scale is logconcave. We mention two useful priors in the appendix when $\phi = \ln(\alpha)$. The shrinkage prior, $f(2,2)$ distribution, is

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

Another example, the half-Cauchy prior, is

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

Knowing that $\pi(\phi|k)$ is logconcave, we can use the adaptive reject sampling method (Gilks 1992) to draw ϕ. This sampling procedure was performed with the R package `ars`. Then we can compute α in the form $\alpha = e^{\phi}$. There is limitation to the ARS method due to tail problem, i.e. the sampling distribution for the two tails of the distribution is not accurate and this can be seen in the simulation section.

Nandram and Choi (2004) discussed the use of the gamma prior, which was introduced earlier by Escobar and West (1995). One concern is that the mix of Gamma method gives bimodal sampling distribution whereas a unimodal density of α is preferred. Another problem is that it requires informative Gamma prior and this remains to be validated.

Nandram and Yin (2016) transformed α according to $\rho = \frac{1}{1+\alpha}$, this is actually the correlation, $\text{Cor}(y_i, y_j)$, $i \neq j$, in the DP. The posterior density of ρ is

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

Note that $\pi(\rho|k)$ is well defined on $[0,1]$. However, we see that it is not in a simple form and a one-dimensional grid method was used to draw samples from it, thereby avoiding Markov chain Monte Carlo methods (e.g. Metropolis - Hastings 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 $\pi(\rho|k)$; the discreteness is removed by jittering. Nonetheless, there is a drawback of this method, because it may not perform well when ρ has substantial probability near 0 or 1.

2.3. Ratio of Uniforms Method

Liu and Nandram (2020) proposed to use the ratio of uniforms method to obtain posterior samples of α. Originally introduced by Kingderman and Monahan (1977), a point is generated uniformly over a certain region in the plane.

To achieve this, independent uniform random variables are simulated,

$$U, V \sim \text{Uniform}(0,1)$$

and those that fall outside some set are discarded. The ratio V/U is then calculated for those points inside the set. The ratio values obtained are used as observations from the required distribution.

There are other priors that can be used but here for illustration purpose, we use the posterior distribution of α with a noninformative prior, $\pi(\alpha) = \frac{1}{(1+\alpha)^2}$,

$$h(\alpha) = \pi(\alpha|k) \propto \frac{\alpha^k \Gamma(\alpha)}{\Gamma(\alpha+n)(\alpha+1)^2}, \ \ \alpha > 0.$$

A half Cauchy prior can be used for the prior of α, but there is very little difference between the two when transformed to $[0,1]$ a posteriori. This method can proceed using the following algorithm.

1. Generate u and v independently from $U(0,b)$ and $U(c,d)$.
2. Set $\alpha = v/u$ if $u^2 \leq h(v/u)$ and return to (i) otherwise.

Here, b, c and d are given by

$$b = \sup_{\alpha} \sqrt{h(\alpha)}, \quad c = -\sup_{\alpha} \alpha\sqrt{h(\alpha)}, \quad d = \sup_{\alpha} \alpha\sqrt{h(\alpha)}.$$

Because α is positive, we set $c = 0$. This algorithm is very easy to implement and it is very efficient to get samples.

2.4. Simulation Study

It is convenient to compare different sampling methods using simulations because we can obtain the true distribution of α and compare the theoretical values with the sampled values. Firstly we find the theoretical percentiles of α using fine grids of width 0.0025. Then we perform the four sampling methods to get 10,000 sample points. We can find the sample percentiles by ordering the sample values and find corresponding quantiles as the theoretical values. Lastly, we compare the theoretical value versus the sampled value using a quantile-quantile plot.

In order to compare the four sampling methods, we take the sample size $n = 12$, 25, 100 and the number of distinct values k to be roughly equal to $\ln n$, with $k = 2$, 3 and 5 respectively. We choose a common prior, the shrinkage prior,

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

to be used for all four sampling methods.

Results are shown in Figures 1, 2 and 3. All four methods provide reasonable sampling distributions for α. However, the ratio of uniforms method is most accurate. In all these figures, the points of ratio of uniforms method fall on almost a 45 degree straight line through the origin and there is some problem with the other plots at various places (e.g. not fitting exactly on the 45 degree straight line through the origin). As we mentioned in Section 2, the ARS and grid method have tail problems and mixture of gamma uses an informative prior

which remains to be validated. Our method does not require informative gamma prior and it is easy to implement. So we recommend using ratio of uniforms to get random samples of α.

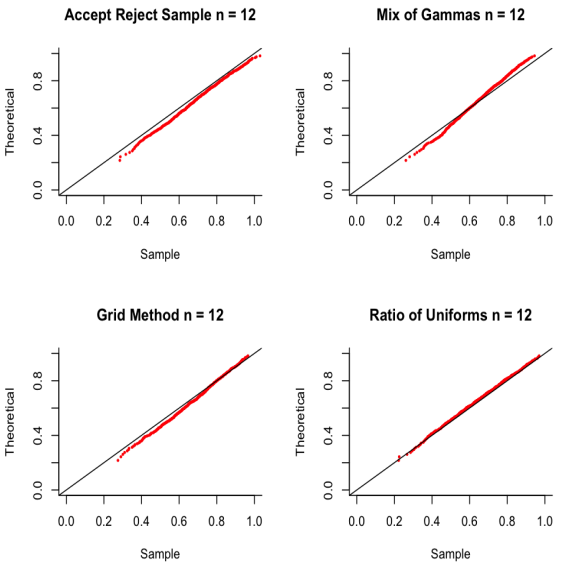

Figure 1: Comparison for the posterior distributions of the concentration parameter using the four sampling methods (n= 12, k = 2)

3. Discrete Baseline

Current literature on DP has been using continuous baseline distributions, see Antonelli, Trippa and Haneuse (2016). Teh, Jordan, Beal and Blei (2006) developed a hierarchical Dirichlet process model with a discrete baseline distribution. Apparently, they were not aware of the problem with the discrete baseline distribution when sampling the concentration parameter and they inadvertently attempted to "sweep the problem under the rug."

Here, we explore a possibility of using a discrete baseline. One problem is that the distinct values in the sample are no longer the true distinct ones because of discrete baseline. We allow observing a "new" value from the baseline distribution that is the same as one that is already in the sample. To solve this problem, we introduce latent variables,

$$Z_i = \begin{cases} 1, & \text{if a draw is made from the baseline,} \\ 0, & \text{if a draw is from the value that is already observed} \end{cases}$$

with $Z_i \overset{ind}{\sim} \text{Ber}(\frac{\alpha}{\alpha+i-1})$, $i = 1, \cdots, n$. Therefore, the true number of distinct values is $k = \sum_{i=1}^{n} z_i$.

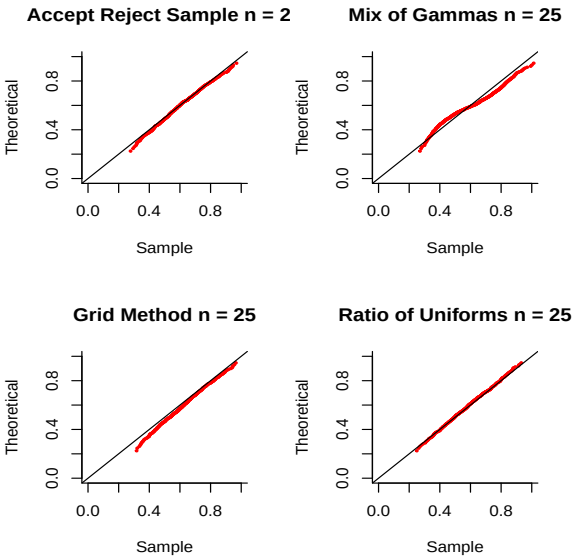

Figure 2: Comparison for the posterior distributions of the concentration parameter using the four sampling methods (n= 25, k = 3)

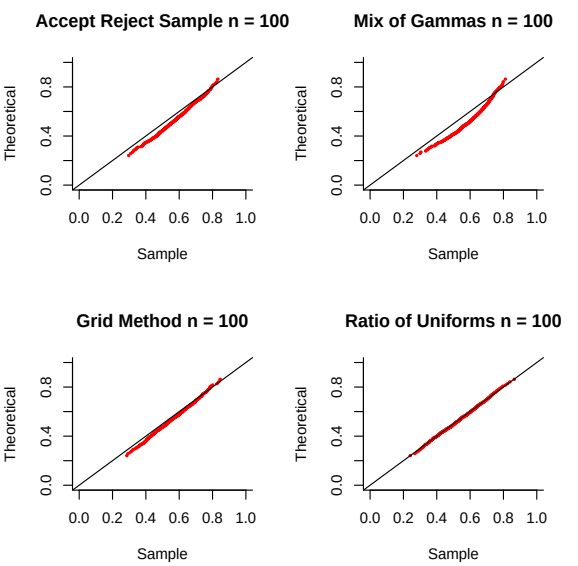

Figure 3: Comparison for the posterior distributions of the concentration parameter using the four sampling methods (n= 100, k = 5)

Our goal is to predict the finite population proportion for a given area based on a random sample from it. This could be applied to many areas of study, for example we want to predict the infectious rate (like a proportion where the denominator is fixed) of a given farmland for some disease and it is not feasible to observe all the plants on the farm, however, we could take a random sample and estimate the posterior mean using this sample. We have observed n of them and want to make predictions to the $N-n$ individuals. We consider three cases.

Case 1. We use the one-level DP model for the population values to make inference for a finite population mean. For this case, the baseline distribution is chosen to be normal. We assume that

$$\begin{aligned} y_1 \cdots, y_N | G &\sim G, \\ G &\sim DP(\alpha, G_0), \\ G_0 &\sim N(\mu, \sigma^2), \\ \pi(\mu, \sigma^2) &\propto \frac{1}{\sigma^2}, \\ \pi(\alpha) &= \frac{1}{(1+\alpha)^2}. \end{aligned}$$

Here, we observe the number of distinct values k and then sample α as discussed in Section 2. For each sampled α value, we predict the unobserved $Y_{n+1}, \cdots, Y_N$ using the Polya urn scheme,

$$Y_{n+i+1} | y_1, \cdots, y_n, y_{n+1}, \cdots, y_{n+i} \sim \frac{\alpha}{\alpha+n+i} G_0 + \frac{n+i}{\alpha+n+i} \sum_{j=1}^{n+i} \delta_{y_j},$$

for $i = 1, \cdots, N-n-1$, (Nandram and Yin 2016 a, b). So it is easy to draw the nonsampled values one by one using the Polya urn scheme.

Case 2a. We correct the true number of observations from the baseline distribution $\sum_{i=1}^{n} z_i$, where $z_i = 1$ when obervation i is a distinct value from G_0. The discrete model is

$$\begin{aligned} y_i | G &\overset{ind}{\sim} G, \; i = 1, \cdots, n, \\ G | p, \alpha &\sim \mathrm{DP}(\alpha, \mathrm{Bin}(m, p)), \\ z_i | \alpha, p &\overset{ind}{\sim} \mathrm{Ber}(\frac{\alpha}{\alpha+i-1}), \\ \pi(\alpha) &= \frac{1}{(1+\alpha)^2}, \pi(p) = 1. \end{aligned}$$

Where n is the sample size and m is the predefined number of the total trials in the Binomial distribution. The joint posterior density can be written as

$$\begin{aligned} \pi(z, p, \alpha | y) \propto & \frac{1}{(1+\alpha)^2} \times \\ & \prod_{i=1}^{n} [p^{y_i}(1-p)^{m-y_i}]^{z_i} [p^{y_i}(1-p)^{m-y_i}]^{1-z_i} \cdot \left[\frac{\alpha}{\alpha+i-1}\right]^{z_i} \left[\frac{1}{\alpha+i-1}\right]^{1-z_i}. \end{aligned}$$

We obtain the conditional distribution of the Gibbs sampler

$$z_i|\alpha,p,y \overset{ind}{\sim} \text{Ber}(Q_i),$$

where

$$Q_i = \frac{\frac{\alpha}{\alpha+i-1}p^{y_i}(1-p)^{m-y_i}}{\frac{\alpha}{\alpha+i-1}p^{y_i}(1-p)^{m-y_i}+\frac{i-1}{\alpha+i-1}p^{y_i}(1-p)^{m-y_i}} = \frac{\alpha}{\alpha+i-1},\ \ i=1,\cdots,n,$$

$$p|z,\alpha,y \sim \text{Beta}(\sum_{\{z_i=1\}} y_i+1,\ \sum_{\{z_i=1\}}(m-y_i)+1),$$

$$\pi(\alpha|z,p) \propto \frac{\alpha^{\sum_{i=1}^n z_i}\Gamma(\alpha)}{\Gamma(\alpha+n)(\alpha+1)^2}.$$

Here, α is drawn from its conditional posterior distribution using our preferred ratio of uniforms method. For each sampled α, we predict one set of unobserved values and compute the finite population mean using the Polya urn scheme. We used a burn in of 1000 and thinning of 10 to get a sample of 10,000. We diagnosed the Gibbs sampler after the chain is run. For the data example we use in this paper, the diagnostic result shows that the effective sample sizes are 4537, 5042 and 4978 for α, p and $\sum_{i=1}^n z_i$ respectively. P-values from the Geweke's tests are 0.384, 0.533 and 0.628 respectively. So at this setting the Gibbs sampler is mixing well. It took about 22 seconds to obtain 10,000 sample values on our computer (see Section 2).

Case 2b. It would be interesting to see the difference of the prediction between a Polya urn scheme and a stick breaking procedure with the idea borrowed from Sethuraman (1994), Ishwaran and James (2001), Kalli, Griffin and Walker (2011). Using the model in Case 2a, but suppose we have already observed $y_1^*,\cdots,y_d^*$, d distinct values ($1 \le d \le n$), with $n_1 \ge n_2,\cdots,\ge n_d$ being their corresponding counts. Here, we allow some values to be unobserved. Now we want to predict $N_1-n_1,\cdots,N_d-n_d$, for convenience, we write $N_1^*,\cdots,N_d^*$. Let $N^* = N-n$, so that we know $N^* = \sum_{i=1}^d N_i^*$. Now

$$N_1^*,\cdots,N_d^* \sim \text{Multinomial}\Big\{N^*,(w_1,\cdots,w_d)\Big\},$$

where $w_1,\cdots,w_d$ are the weights in stick-breaking algorithm with $\sum_{s=1}^\infty w_s = 1$, $w_1 = \nu_1$, $w_2 = \nu_2(1-\nu_1),\cdots,w_{d-1} = \nu_{d-1}\prod_{i=1}^{d-2}(1-\nu_i)$, $w_d = \prod_{i=1}^{d-1}(1-\nu_i)$, and

$$\nu_i \overset{iid}{\sim} \text{Beta}(1,\alpha).$$

Given α from the Gibbs sampler, we can draw ν_i and thus draw the predicted values from a Multinomial distribution.

With the posterior samples of α from the Gibbs sampler in Case 2a, the conditional

posterior distribution of ν is

$$\begin{aligned}\pi(\nu|\alpha,d) \propto\ & \nu_1^{n_1}[\nu_2(1-\nu_1)]^{n_2}\cdots[\nu_{d-1}(1-\nu_1)\cdots(1-\nu_{d-2})]^{n_{d-1}} \\ & [(1-\nu_1)\cdots(1-\nu_{d-2})(1-\nu_{d-1})]^{n_d}\times\prod_{i=1}^{d}(1-\nu_i)^{\alpha-1} \\ \propto\ & \nu_1^{n_1}(1-\nu_1)^{n_2+\cdots+n_d+\alpha-1}\nu_2^{n_2}(1-\nu_2)^{n_3+\cdots+n_d+\alpha-1}\cdots\nu_{d-1}^{n_{d-1}}(1-\nu_{d-1})^{n_d+\alpha-1}.\end{aligned}$$

Therefore,

$$\pi(\nu_i|\alpha,d) \overset{ind}{\sim} \text{Beta}(n_i+1, \sum_{j=i+1}^{d} n_j+\alpha).$$

Once samples of ν_i are obtained, we can predict the unobserved response values from a Multinomial distribution discussed above.

We will implement the three cases we discussed in the following example.

4. Real Data Analysis

The data we present here are about Phytophthora Epidemic in Bell Pepper from Gumpertz (1997). The pathogen Phytophthora Capsici Leonian causes lesions on the crown, stem, and leaves of bell pepper, and rapidly causes the plant to die. For their analyses, they took one field which was a square lattice of 20×20 quadrats with 2 to 3 bell pepper plants per quadrat as an example. The response variable within each quadrat was presence or absence of disease in a quadrat. If any plant was wilted, dead, or had lesions on stem, crown, or leaves, disease was considered to be present in the quadrat. Disease presence or absence was recorded for each quadrat on nine dates throughout the growing season, from 6/16/92 to 8/5/92. Figure 2 shows the disease incidence on 6/25/92.

We want to make this data set usable to mimic our discrete response scenario so we perform the following sampling procedure. We divide each row of the field by every fifth quadrats and then we take one random sample within each row of the field. We assume that the sampled value follows a binomial distribution with total number of trials being 5. Now, our goal is to predict the unobserved quadrats and estimate the infectious rate, which is really a finite population proportion (mean) in this application. We performed the estimation using both discrete baseline and continuous baseline approaches, as discussed in Section 3.2.

We report the posterior means (PM), posterior standard deviations (PSD) and the credible intervals (CI) in Table 1. Given the true infectious rate of 0.1525, we found that the continuous (Normal) baseline distribution provides an unbiased estimation to the infectious rate. However, the lower end of 95% credible interval is negative. This is because the posterior sample is taken over the whole real line as a nature of the Gaussian distribution. We know in reality, the infectious rate is a probability and should always be positive. Here, we naively use normal baseline because this is often chosen to be the G_0 in many practices. But for obvious reasons we now want to avoid it. PMs are roughly the same for Case 1 and Case 2a but Case 2b has some bias, with a larger estimation, however, the PSD is the smallest

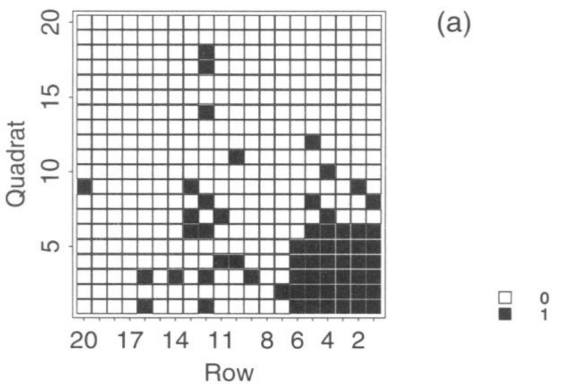

Figure 4: Map of Disease Incidence.

for Case 2b. Binomial baseline (Case 2a) has slightly more variability because it considers the uncertainty of which urn a new observation is drawn. Last but not least, the prediction using a Multinomial approach (Case 2b) can significantly reduce the standard deviation and it provides a realistic result because it is based on a discrete distribution. Figure 5 shows the posterior distribution of the three estimates of the finite population mean. The plots are similar for Cases 2a and 2b but Case 2a is more spread out. Note that the plots are in the same range (easier for us to visualize and compare). Similar to the table, the Multinomial approach gives most concentrated plot. The Normal approach exceeds zero to the negative side.

Table 1. Estimation of the Infectious Rate (True Rate*: 0.1525)

Baseline Distribution	PM	PSD	95% CI
Case 1. Normal (μ, σ^2)	0.1551	0.1728	(-0.2221,0.4735)
Case 2a. Binomial (n^{**}, p)	0.1588	0.219052	(0,0.6800)
Case 2b. Multinomial (N^{**}, w)	0.1698	0.0661	(0.0667,0.3267)

PM = Posterior Mean; PSD = Posterior Standard Deviation; CI = Credible Interval. * We can compute the true rate with data from the whole 20×20 study site. ** In our case $n = 5$ and $N = 80 - 20 = 60$.

5. Concluding Remarks

We have proposed a new sampling method for the standard concentration parameter of the Dirichlet Process and compared it with three methods. The Ratio of Uniforms is more accurate and it is faster considering the computational time. In the meantime, we pointed out a problem that current researchers have ignored regarding the baseline distribution of the DP. We have corrected the true number of distinct values in the sample by introducing a latent variable which indicated which urn a new observation is from. By using this approach,

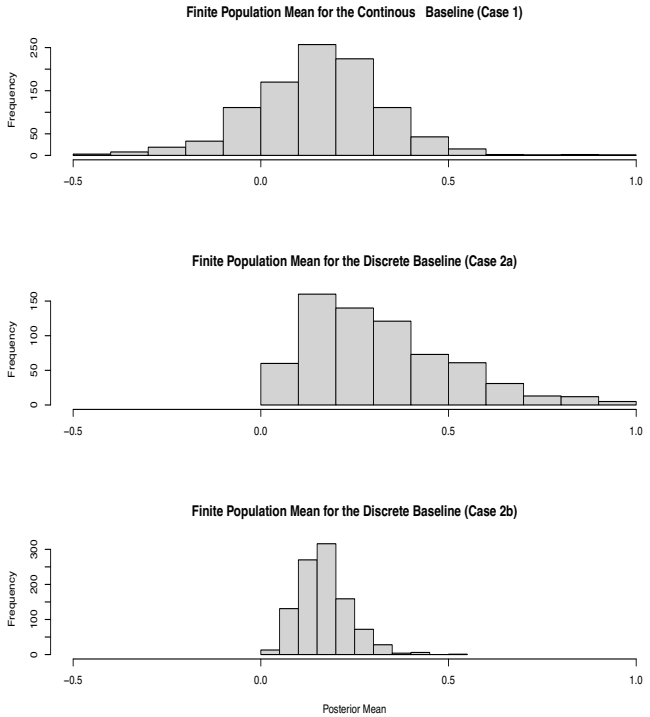

Figure 5: Posterior Distributions of the Finite Population Mean (Proportion) for the Three Cases

we are able to give a more accurate estimation of the finite population mean when the observations are discrete. We used a Phytophthora example to illustrate our approach. We concluded the discrete baseline method is more reasonable.

There are two directions we could proceed to extend our current work. First, we might consider a spatial model for the example provided in this paper. However, it is not the purpose of our paper to provide a complete analysis of these data.

Second, we could extend the one-level DP model to a two-level DP model, where there are groups naturally occur in the data. The two-level model is the Dirichlet process mixture (DPM) model with a DP on the second level. Recently, Yin and Nandram (2020 a,b) placed the DP on the first level but not on the second. They claimed that their approach is good for data with gaps, outliers and ties.

Third, the work we have done in this paper also inspired us to study sensitivity to the baseline. We may give a very weak assumption to the baseline, i.e. either logconcave or unimodal. These can be discretized nicely. For a logconcave density the slopes of the tangent lines decrease all the way or the chords joining any two points will have non-increasing slopes all the way from left to right on the real line. Also, a unimodal density has heights increasing to the mode and then decreasing (i.e. the cumulative distribution function is first

convex up to the mode and concave after the mode). So, essentially, we can use a discrete baseline distribution in the DPM.

Therefore, our work on discrete baseline distribution is an important start. However, although we have a good algorithm for the concentration parameter of the Dirichlet process, based on the ratio of uniforms, some improvement may be possible.

References

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

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

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

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

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prunster, I. and Rodrigue,A., (2019). Latent nested nonparametric priors. *Bayesian Analysis*, 14, pp. 1303–1356.

Chaudhuri, S., Gosh, M., (2011). Empirical likelihood for small area estimation. *Biometrika*, 98, 2, pp. 473–480.

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

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

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

Gumpertz, M. L., Graham, J. M. and Ristaino, J. B., (1997). Autologistic Model of spatial pattern of Phytophthora Epidemic in bell pepper: Effects of Soil Variables on Disease Presence. *Journal of Agricultural, Biological, and Environmental Statistics*, Vol. 2, No. 2, pp. 131–156.

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

Kinderman, A. J., Monahan J. F., (1977). Computer generation of random variables using the ratio of uniform deviates. *Association for Computing and Machinery, Inc.*

Liu, Y., Nandram, B., (2020). Sampling methods for the concentration parameter of the Dirichlet process. In *JSM Proceedings, Nonparamtric Section.* Alexandria, VA: American Statistical Association. pp. 1121–1131.

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

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

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

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

Rasmussen, C. E., (2000). The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems* pp. 554–560.

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

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

Yin, J., Nandram, B., (2020a). A Bayesian small area model with Dirichlet processes on responses. *Statistics in Transition, New Series*, 21 (3), pp. 1–19.

Yin, J., Nandram, B., (2020b). A Nonparametric Bayesian Analysis of Response Data with Gaps, Outliers and Ties. *Statistics and Applications, New Series*, 18 (2), pp. 121–141.

Appendix

Logconcavity of the Posterior Density of α

Proof: It is easy to show that the likelihood function for α is

$$p(\alpha \mid k) \propto \frac{\alpha^k}{\prod_{j=1}^{n-1}(j+\alpha)}, \alpha > 0,$$

where k is the number of distinct values for a continuous baseline. For any prior $\pi(\alpha)$, using Bayes' theorem, the posterior density of α is

$$\pi(\alpha \mid k) \propto p(\alpha \mid k)\pi(\alpha).$$

If we make the transformation $\phi = \log(\alpha)$, $p(\alpha \mid k)$ will transform to $p_1(\phi \mid k) = \frac{e^{k\phi}}{\prod_{j=1}^{n-1}(j+e^{\phi})}$ and $\pi(\alpha)$ will transform to $\pi_1(\phi)$ and the Jacobian is e^{ϕ}. We show that if $\pi_1(\phi)$ is logconcave, i.e. $\frac{d^2\ln(\pi_1(\phi))}{d\phi^2} < 0$, then the posterior density on the transformed scale is logconcave. Let

$$\Delta(\phi) = (k+1)\phi - \sum_{j=1}^{n-1} \ln(j+e^{\phi}) + \ln(\pi_1(\phi)).$$

Then,

$$\frac{d\Delta(\phi)}{d\phi} = (k+1) - \sum_{j=1}^{n-1} \frac{e^{\phi}}{j+e^{\phi}} + \frac{d\ln(\pi_1(\phi))}{d\phi}$$

and

$$\frac{d^2\Delta(\phi)}{d\phi^2} = -\sum_{j=1}^{n-1} \frac{je^{\phi}}{(j+e^{\phi})^2} + \frac{d^2\ln(\pi_1(\phi))}{d\phi^2} < 0.$$

Therefore, under the assumption of logconcavity for $\pi_1(\phi)$, the posterior density of α is logconcave.
We mention two useful priors when $\phi = \ln(\alpha)$. The shrinkage prior, $f(2,2)$ distribution, is

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

Another example, the half-Cauchy prior, is

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

Both priors after making the transformation $\phi = \ln(\alpha)$ are logconcave.