The Annals of Statistics
1996, Vol. 24, No. 3, 911–930

NONPARAMETRIC HIERARCHICAL BAYES VIA SEQUENTIAL IMPUTATIONS[1]

By Jun S. Liu

Stanford University

We consider the empirical Bayes estimation of a distribution using binary data via the Dirichlet process. Let $\mathscr{D}(\alpha)$ denote a Dirichlet process with α being a finite measure on $[0, 1]$. Instead of having direct samples from an unknown random distribution F from $\mathscr{D}(\alpha)$, we assume that only indirect binomial data are observable. This paper presents a new interpretation of Lo's formula, and thereby relates the predictive density of the observations based on a Dirichlet process model to likelihoods of much simpler models. As a consequence, the log-likelihood surface, as well as the maximum likelihood estimate of $c = \alpha([0, 1])$, is found when the shape of α is assumed known, together with a formula for the Fisher information evaluated at the estimate. The sequential imputation method of Kong, Liu and Wong is recommended for overcoming computational difficulties commonly encountered in this area. The related approximation formulas are provided. An analysis of the tack data of Beckett and Diaconis, which motivated this study, is supplemented to illustrate our methods.

1. Introduction. The setting for nonparametric problems is usually as follows: we have n iid observations from an unknown probability measure P, and want to make an inference about P. As a Bayesian, one may try to put a prior distribution on a rich class of distributions and hope to find the posterior distribution. Because of the pioneering work of Ferguson (1973, 1974), Blackwell and MacQueen (1973), Doksum (1972), Antoniak (1974) and many others, the choice of a Dirichlet process, $\mathscr{D}(\alpha)$, where α is a finite measure on the sample space, as a prior distribution of the unknown P, has become standard. Many interesting theoretical properties of the Dirichlet process and samples from it have been obtained. For a thorough understanding, see the seminal work of Ferguson (1973), Antoniak (1974) and Korwar and Hollander (1973) among others.

Partly for simplicity, we concentrate on probability measures on $[0, 1]$ in this paper. Let α be a finite measure on the interval $[0, 1]$. A random probability measure P on $[0, 1]$ is said to follow a *Dirichlet process* $\mathscr{D}(\alpha)$ if, for every finite partition $\{B_1, \dots, B_m\}$ of $[0, 1]$ (i.e., measurable, disjoint and exhaustive), the random vector $\{P(B_1), \dots, P(B_m)\}$ follows a Dirichlet distribution with parameter $(\alpha(B_1), \dots, \alpha(B_m))$. We call α the characteristic measure of the Dirichlet process, and define $\|\alpha\|$ by $\|\alpha\| = \int_0^1 \alpha(dx)$. A fundamental fact

Received January 1993; revised November 1995.

[1] Partially supported by NSF Grant DMS-94-04344.

AMS 1991 *subject classifications*. Primary 62G05; secondary 62E25, 65U05.

Key words and phrases. Dirichlet process, empirical Bayes, Gibbs sampler, importance sampling, Pólya urn, sensitivity analysis.

shown by Ferguson (1974) is that:

Suppose $P \sim \mathscr{D}(\alpha)$ and $\zeta_1, \dots, \zeta_n$ is a sample of size n from P. Then the posterior distribution of P given the observations is again a Dirichlet process with a new characteristic measure, that is,

$$P \mid \zeta_1, \dots, \zeta_n \quad \sim \mathscr{D}\left(\alpha + \sum_{i=1}^{n} \delta_{\zeta_i}\right),$$

where δ_x is a probability measure that gives mass 1 to x.

In many situations, however, one may not be able to obtain direct observations from P, but indirect ones whose distribution can be expressed as a convolution of P with a known kernel function [see Lo (1984) for more details]. For convenience, we let F be the cumulative distribution function of P and use F and P interchangeably. Suppose that the random variables y_i, $i = 1, \dots, n$, are binomially distributed with parameters l_i and ζ_i, where l_i is a preassigned number of trials for the ith observation and ζ_i is the corresponding probability of success. Given the set of parameters $\mathbf{z} = (\zeta_1, \dots, \zeta_n)$, all the y_i are independent. Instead of assuming any parametric model for the ζ's, we assume that the ζ_i are iid from the unknown distribution F. To summarize, we have:

1. Given ζ_i, $y_i \sim \text{Bin}(l_i, \zeta_i)$, where the y_i are observed and the l_i are preassigned.
2. Given the unknown distribution F, $\zeta_1, \dots, \zeta_n$ are iid observations on F.
3. F is a priori distributed as $\mathscr{D}(\alpha)$.

This setting is treated in Berry and Christensen (1979), where the focus is on deriving analytic forms of the empirical Bayes estimates for the ζ_i and on computing approximations of these quantities. In a more general setting, Antoniak (1974) shows that the posterior distribution of F is a mixture of Dirichlet processes. Kuo (1986) provides a Monte Carlo method for computing posterior means. Lo (1984) derives the analytic form of Bayes estimators for a density estimation problem that involves convolutions with a known kernel function. Our problem is one special case of his setting.

We assume that the prior measure α is of the form $c\text{Beta}(a, b)$, where $\text{Beta}(a, b)$ represents a beta density with parameters a and b. The parameter c is a positive number representing the weight to put on our prior belief. From an empirical Bayes point of view, we would like to use the data to help determine the parameters a, b and c. But for convenience we will concentrate on c, assuming a and b fixed in advance. In the tack example of Beckett and Diaconis (1994) referred to later in Section 4, we find that the resulting posterior mean of F is sensitive to changes in the prior weight c: the larger the c assumed, the smoother the result. As was shown by Antoniak (1974) and Korwar and Hollander (1973), the weight c is related to the number of clusters, that is, the distinct values one may expect to observe in $(\zeta_1, \dots, \zeta_n)$, a sample of size n from some $F \sim \mathscr{D}(\alpha)$. Roughly, the larger the value of c, the more the number of clusters; and the number of clusters is about $c \log((n + c)/c)$ for large c. Along this line, Korwar and Hollander (1973) provide an estimation method for the weight c when one has direct observations on the ζ_i. Practical methods for dealing with indirect observations in nonparametric Bayes settings only appear recently mainly because of computational difficulties. The recognition of usefulness of Markov chain Monte Carlo methods in facilitating Bayesian inference [see Gelfand and Smith (1990) and Smith and Roberts (1993)] especially contributes to the recent developments in this area. Doss (1994) provides an iterative sampling method for applying nonparametric Bayes to censored data. Escobar (1994) treats the James–Stein problem under a Bayesian nonparametric model using the Gibbs sampler, and mentions the problem of estimating c. The importance of determining the weight c and the shape of α is discussed in more detail in Escobar and West (1995) and West (1992). Especially, they illustrate how to incorporate a beta prior distribution and then how to update for c in a general Gibbs sampling framework. For priors other than the beta distribution or a mixture of the beta distributions, a discretization of the range of c has been recommended.

In this paper, a new interpretation of Lo's formula (his Lemma 2) is presented in Section 2 that relates the predictive density of the data [a terminology from Box (1980)] resulting from imposing the nonparametric Dirichlet process model to the *likelihood* of simpler models. The posterior distribution of $N(\mathbf{z})$, the number of clusters among $(\zeta_1, \dots, \zeta_n)$, is shown to be proportional to these likelihoods. As a consequence, we show that at $c = \hat{c}$, the maximum likelihood estimate of c, $\mathscr{E}\{N(\mathbf{z})|\mathbf{y}\} = \mathscr{E}\{N(\mathbf{z})\}$, where $\mathscr{E}$ represents the expectation with respect to the probability model determined by $\mathscr{D}(\alpha)$. Furthermore, at $c = \hat{c}$, the expected Fisher information of $d = \log(c)$ is found as $\mathrm{Var}\{\mathscr{E}_\alpha(N(\mathbf{z}) \mid y_1, \dots, y_n)\}$. Section 3 is concerned with the computational aspect of the problem. Instead of using the Gibbs sampler, we apply sequential imputations introduced by Kong, Liu and Wong (1994) to obtain the quantities of interest: an approximate likelihood curve of $\log(c)$ and the MLE of c; the posterior distribution of the number of clusters; the posterior mean and variance of each ζ_i and the posterior mean of F. The tack example is treated in Section 4 to illustrate our theory and methods.

2. Some theoretical results on Dirichlet processes. Let y_i be an observation from $\mathrm{Bin}(l_i, \zeta_i)$ where the ζ_i are iid observations on F for $i = 1, \dots, n$ and $F \sim \mathscr{D}(\alpha)$. Throughout the paper, we use $\mathscr{P}(\cdot)$ to denote the probability measure under the Dirichlet process model, and use $\mathscr{E}(\cdot)$ and $\mathrm{Var}(\cdot)$, respectively, to denote the expectation and variance taken under this measure. Let $\mathbf{z} = (\zeta_1, \dots, \zeta_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$. We assume $\alpha = c\mathrm{Beta}(a, b)$.

Let $S = \{1, \dots, n\}$, $\mathbf{P}$ be a partition of S and $|\mathbf{P}|$ be the number of cells in $\mathbf{P}$. If $|\mathbf{P}| = k$, for example, we further use $\mathbf{p}_{(1)}, \dots, \mathbf{p}_{(k)}$ to denote the cells of this partition and use $e_i = |\mathbf{p}_{(i)}|$, $i = 1, \dots, k$, to denote the size of each cell. Therefore, the $\mathbf{p}_{(i)}$ are nonintersecting, nonempty and exhaustive subsets of S without regard to ordering. For any subset $\mathbf{p}_{(i)}$, we use the notation

$$\mathbf{l}_{(i)} = \sum_{j \in \mathbf{p}_{(i)}} l_j, \qquad \mathbf{y}_{(i)} = \sum_{j \in \mathbf{p}_{(i)}} y_j.$$

For any positive integer $n > x$, we denote the Beta function by

$$(1)\quad B_{ab}(n,x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 t^{x+a-1}(1-t)^{n+b-x-1}\,dt = \frac{B(a+x, b+n-x)}{B(a,b)}.$$

An *m-spike model* for the unobservable ζ's is a discrete distribution $G(\zeta)$ with m point masses on the unit interval, that is, with the probability distribution function of the form

$$G(\zeta) = a_1\delta_{z_1}(\zeta) + \cdots + a_m\delta_{z_m}(\zeta).$$

THEOREM 1 (Predictive density). *If the prior characteristic measure $\alpha = c\mathrm{Beta}(a,b)$, then the predictive density of the observations under the Dirichlet process model $\mathscr{D}(\alpha)$ is*

$$(2)\qquad \mathscr{P}(\mathbf{y}) = \frac{(n-1)!}{\prod_{i=1}^n (c+i-1)} \prod_{i=1}^n \binom{l_i}{y_i} \sum_{m=1}^n c^m \mathscr{L}_m(\mathbf{y}),$$

in which each term has the expression

$$\mathscr{L}_m(\mathbf{y}) = \frac{1}{(n-1)!} \sum_{\mathbf{P}:\ |\mathbf{P}|=m} \left\{ \prod_{i=1}^m (e_i - 1)! B_{ab}(\mathbf{l}_{(i)}, \mathbf{y}_{(i)}) \right\}.$$

PROOF. By the Pólya urn argument of Blackwell and MacQueen (1973),

$$\mathscr{P}(\zeta_1,\dots,\zeta_n) = \prod_{i=1}^n \frac{\alpha(\zeta_i) + \sum_{j=1}^{i-1}\delta_{\zeta_j}(\zeta_i)}{\|\alpha\| + i - 1}.$$

Furthermore, it is clear that $\mathscr{P}(y_1,\dots,y_n|\zeta_1,\dots,\zeta_n) = \prod_{i=1}^n \binom{l_i}{y_i}\zeta_i^{y_i}(1-\zeta_i)^{l_i-y_i}$. Letting $g_i(\zeta_i) = \zeta_i^{y_i}(1-\zeta_i)^{l_i-y_i}$ and applying Lemma 2 of Lo (1984), we obtain the result. The factor $\binom{l_i}{y_i}$ is of no interest and will be omitted in the later context. □

THEOREM 2 (Clustering). *Let $N(\mathbf{z})$, where $\mathbf{z} = (\zeta_1,\dots,\zeta_n)$, be the number of distinct values of ζ that are assumed to be drawn from a random distribution function $F \sim \mathscr{D}(\alpha)$. Then the posterior distribution of $N(\mathbf{z})$ is*

$$\mathscr{P}(N(\mathbf{z}) = m \mid \mathbf{y}) = c^m \mathscr{L}_m(\mathbf{y}) / \sum_{j=1}^n c^j \mathscr{L}_j(\mathbf{y}).$$

PROOF. It is not difficult to see that $\mathscr{P}(y_1,\dots,y_n,\ N(\mathbf{z}) = m) \propto c^m\mathscr{L}_m$ because, for fixed $\mathbf{y}$,

$$\mathscr{P}(y_1,\dots,y_n,\zeta_1,\dots,\zeta_n) \propto \prod_{i=1}^n \zeta_i^{y_i}(1-\zeta_i)^{l_i-y_i}\left(\alpha(\zeta_i) + \sum_{j=1}^{i-1}\delta_{\zeta_j}(\zeta_i)\right),$$

and we can integrate out the $\mathbf{z}$ with the constraint that $N(\mathbf{z}) = m$ in the expression. Hence the result follows from Theorem 1. □

Because of the above two theorems, we call $\mathscr{L}_m$, $m = 1, \dots, n$, the *likelihood* of the m-spike model. Another reason is as follows. Suppose we fit the data to an m-spike model

$$G_m(\zeta) = a_1\delta_{z_1}(\zeta) + \cdots + a_m\delta_{z_m}(\zeta), \tag{3}$$

with the constraint that each spike contributes at least one observation [i.e., the model cannot be reduced to an $(m-1)$-spike model]. Then when an improper prior with density $(a_1a_2\cdots a_m)^{-1}$ for the mixing proportion and a Beta(a, b) prior for the z_i are used, the pseudo (since the prior is improper) predictive density for $\mathbf{y}$ in this model is proportional to $\mathscr{L}_m$.

By a result of Antoniak (1974), however, the value of c also reflects the a priori belief about the number of distinct ζ's. How some changes in this belief, as well as in the shape of the characteristic measure α, affect the posterior distribution is of interest.

THEOREM 3 (Sensitivity). *If the continuous characteristic measure α of a Dirichlet process is changed to the continuous measure β, then*

$$\frac{\mathscr{P}_\beta(\mathbf{y}, \mathbf{z})}{\mathscr{P}_\alpha(\mathbf{y}, \mathbf{z})} = \left(\prod_{i=0}^{n-1} \frac{\|\alpha\| + i}{\|\beta\| + i}\right) \frac{\beta(\zeta_1')\cdots\beta(\zeta_{N(\mathbf{z})}')}{\alpha(\zeta_1')\cdots\alpha(\zeta_{N(\mathbf{z})}')},$$

where ζ_i', $i = 1, \dots, N(\mathbf{z})$, are the distinct ζ's among $\zeta_1, \dots, \zeta_n$. Moreover, when $\beta = q\alpha$, where $q > 0$ is a weighting factor, the right-hand side simplifies to $q^{N(\mathbf{z})}\prod_{i=0}^{n-1}\{(\|\alpha\| + i)/(q\|\alpha\| + i)\}$.

PROOF. It is obvious from the hierarchical setting that given $\mathbf{z}$, $\mathbf{y}$ is no longer relevant. The rest of the proof follows easily from the Pólya urn explanation of the Dirichlet process [Blackwell and MacQueen (1973)]. □

The parameter c can be treated as an unknown parameter and estimated by maximum likelihood. To begin with, we can easily write the likelihood function of c:

$$L(c, \mathbf{y}) = \prod_{i=1}^{n}(c + i - 1)^{-1}\sum_{m=1}^{n} c^m \mathscr{L}_m(\mathbf{y}). \tag{4}$$

In the Appendix we prove the following result.

THEOREM 4 (MLE). *Let $\alpha = c\mathrm{Beta}(a, b)$. Then the maximum likelihood estimate of c satisfies the normal equation*

$$\mathscr{E}\{N(\mathbf{z}) \mid y_1, \dots, y_n\} = \mathscr{E}\{N(\mathbf{z})\}.$$

Furthermore, if we let $d = \log(c)$, then the observed and the expected Fisher information of d at $\log(\hat{c})$ are, respectively,

$$\mathbf{i}_{\mathrm{obs}}(d) = \mathrm{Var}\{N(\mathbf{z})\} - \mathrm{Var}\{N(\mathbf{z}) \mid \mathbf{y}\} \quad \textit{and} \quad \mathbf{i}_{\mathrm{exp}}(d) = \mathrm{Var}[E\{N(\mathbf{z}) \mid \mathbf{y}\}].$$

Note that (2) and (4) actually hold for α to be any finite measure (not necessarily of Beta form), although the likelihood $\mathscr{L}_m(\mathbf{y})$ may not have a closed form. Therefore, by reviewing its proof, we find that Theorem 4 holds for general α with a given shape. This result can be regarded as a generalization of Theorem 2.3 of Korwar and Hollander (1973) on the Borel space [0,1] with Lebesgue measure. To see the point, we imagine that $l_i \to \infty$ and that $y_i/l_i = \zeta_i$. Then the above theorem gives $\hat{c}$ determined by

$$N(\mathbf{z}) = c\sum_{i=1}^{n} \frac{1}{c+i-1} \approx c\log\left(\frac{c+n}{c}\right),$$

whereas the Korwar–Hollander estimate, $N(\mathbf{z})/\log(n)$, is approximately $\hat{c}$ for large n.

Even when the l_i are large, or the ζ_i are observed directly, the amount of information for d is of order $O(\log(n))$. More precisely, when the ζ_i are observed directly, $\mathrm{Var}\{N(\mathbf{z}) \mid \mathbf{y}\} = 0$. Hence the maximal possible Fisher information on $d = \log(c)$ is

$$\mathrm{Var}\{N(\mathbf{z})\} = \sum_{m=1}^{n} \frac{c(m-1)}{(c+m-1)^2} \approx c\left\{\log\left(\frac{c+n}{c}\right) - 1\right\}.$$

3. Nonparametric Bayesian computation. We mentioned in Section 1 that the posterior distribution of F is easily updated if all the ζ_i were observed. As was noted by Berry and Christensen (1979) and Lo (1984), however, the situation becomes extremely difficult when one has indirect observations as in our Dirichlet–binomial setting. One helpful way of dealing with the problem is to treat those unobservable ζ's as missing data and use missing data methodology such as imputations to make computation feasible. A standard method to handle Bayesian missing data problems is to approximate the actually incomplete data posterior of the parameter vector by a mixture of complete data posteriors. The multiple complete data sets used in the mixture are created ideally by draws from the posterior distribution of the missing data conditioned on the observed data. A popular way for doing this is, of course, data augmentation, or the Gibbs sampler [see Tanner and Wong (1987) and Gelfand and Smith (1990)], whose usefulness in nonparametric Bayesian problems has been demonstrated by Escobar (1994), Doss (1994), West (1992) and others. The sequential imputation method of Kong, Liu and Wong (1994) is an alternative for creating multiple complete data sets. This procedure does not require iterations and is a variation of importance sampling. Moreover, sensitivity analysis and updating with new data can be done cheaply with the method.

3.1. *General method of sequential imputations.* Let $\boldsymbol{\theta}$ denote the parameter vector of interest with prior distribution $p(\boldsymbol{\theta})$ and let $\mathbf{x}$ denote the complete data in a model where the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{x})$ is assumed to be simple. However, $\mathbf{x}$ is assumed to be only partially observed and can be partitioned as $(\mathbf{y}, \mathbf{z})$, where $\mathbf{y}$ denotes the observed part and $\mathbf{z}$ represents the missing part. Now suppose $\mathbf{y}$ and $\mathbf{z}$ can each be further decomposed into n corresponding components so that

$$\mathbf{x} = (x_1, x_2, x_3, \ldots, x_n) = (y_1, z_1, y_2, z_2, y_3, z_3, \ldots, y_n, z_n), \tag{5}$$

where $x_i = (y_i, z_i)$ for $i = 1, \ldots, n$. We note that

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \int p(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{z}) p(\mathbf{z} \mid \mathbf{y})\, d\mathbf{z}. \tag{6}$$

Hence, if M independent copies of $\mathbf{z}$ are drawn from the conditional distribution $p(\mathbf{z} \mid \mathbf{y})$ and denoted by $\mathbf{z}(1), \mathbf{z}(2), \ldots, \mathbf{z}(M)$, we can approximate $p(\boldsymbol{\theta} \mid \mathbf{y})$ by

$$\frac{1}{M} \sum_{j=1}^{M} p(\boldsymbol{\theta} \mid \mathbf{x}(j)),$$

where $\mathbf{x}(j)$ stands for the augmented complete data set $(\mathbf{y}, \mathbf{z}(j))$ for $j = 1, \ldots, M$. [Note that each $\mathbf{z}(j)$ has n components: $z_1(j), \ldots, z_n(j)$.] However, drawing from $p(\mathbf{z} \mid \mathbf{y})$ directly is usually difficult. The Gibbs sampler mentioned earlier achieves something close to that by using Markov chains. Sequential imputation, however, achieves a similar thing by imputing the z_i sequentially and using importance sampling weights. In general, sequential imputation involves doing the following:

(A) For $t = 1, \cdots, n$, starting from $t = 1$, draw z_t^* from the conditional distribution

$$p(z_t \mid y_1, z_1^*, y_2, z_2^*, \ldots, y_{t-1}, z_{t-1}^*, y_t).$$

Notice that the z^*'s have to be drawn sequentially because each z_t^* is drawn conditioned on the $z_1^*, \ldots, z_{t-1}^*$.

(B) For $t = 2, \ldots, n$, compute

$$p(y_t \mid y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*).$$

After the whole process is finished, we compute

$$w = p(y_1) \prod_{t=2}^{n} p(y_t \mid y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*). \tag{7}$$

Note that (A) and (B) are usually done simultaneously. Both (A) and (B) are computationally simple if the predictive distributions $p(x_t \mid x_1, \ldots, x_{t-1})$, $t = 1, \ldots, n$, are simple. Now suppose (A) and (B) are done repeatedly and independently M times. Let the results be denoted by $\mathbf{z}^*(1), \ldots, \mathbf{z}^*(M)$ and $w(1), \ldots, w(M)$, where $\mathbf{z}^*(j) = (z_1^*(j), \ldots, z_n^*(j))$ for $j = 1, \ldots, M$. We will now estimate the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$ by the weighted mixture

$$\frac{1}{W} \sum_{j=1}^{M} w(j) p(\boldsymbol{\theta} \mid \mathbf{x}^*(j)), \tag{8}$$

where $W = \sum w(j)$ and $\mathbf{x}^*(j)$ denotes the augmented data set $(\mathbf{y}, \mathbf{z}^*(j))$ for $j = 1, \ldots, M$. Refer to Kong, Liu and Wong (1994) for more details.

3.2. *Sequential imputation in nonparametric Bayes problems.* In our Dirichlet–binomial nonparametric setting, the unobservable $\mathbf{z} = (\zeta_1, \dots, \zeta_n)$ can be regarded as a missing value, and the unknown function F as the parameter of interest "$\boldsymbol{\theta}$." Sequential imputation requires that it be easy to sample from the distribution

$$\mathscr{P}(\zeta_t \mid y_1, \dots, y_t, \zeta_1, \dots, \zeta_{t-1})$$

and to compute $\mathscr{P}(y_t \mid y_1, \dots, y_{t-1}, \zeta_1, \dots, \zeta_{t-1})$ to update the importance weights. First, we note that, conditional on $\zeta_1, \dots, \zeta_{t-1}$, F is distributed as $\mathscr{D}(\alpha_{t-1})$ with $\alpha_{t-1} = \alpha + \sum_{i=1}^{t-1} \delta_{\zeta_i}$. This implies that

$$[\zeta_t \mid \zeta_1, \dots, \zeta_{t-1}] \sim \frac{1}{c+t-1}\left(\alpha + \sum_{i=1}^{t-1} \delta_{\zeta_i}\right), \tag{9}$$

which should be interpreted as a probabilistic mixture of α and point masses concentrated at the ζ_i's. It follows that, if the prior measure α is chosen to be $c\text{Beta}(a, b)$,

$$\begin{aligned} &[\zeta_t \mid \zeta_1, \dots, \zeta_{t-1}, y_t] \\ &\quad \sim \left[cB_{ab}(l_t, y_t)\text{Beta}(y_t + a, l_t - y_t + b) + \sum_{i=1}^{t-1} \zeta_i^{y_t}(1-\zeta_i)^{l_t - y_t}\delta_{\zeta_i} \right], \end{aligned} \tag{10}$$

where $B_{ab}(l_t, y_t)$ is defined in (1) and $\{cB_{ab}(l_t, y_t) + \sum_{i=1}^{t-1} \zeta_i^{y_t}(1-\zeta_i)^{l_t-y_t}\}^{-1}$ is the normalizing constant. Note that (10) is a mixture of a Beta distribution and discrete point masses. From (9), we also get

$$\mathscr{P}(y_t \mid \zeta_1, \dots, \zeta_{t-1}) = \frac{c}{c+t-1} B_{ab}(l_t, y_t) + \frac{1}{c+t-1} \sum_{i=1}^{t-1} \zeta_i^{y_t}(1-\zeta_i)^{l_t-y_t},$$

which is the term we need to compute when updating the importance weights.

So, as demonstrated, both steps (A) and (B) of sequential imputation can be easily implemented. We note that a direct application of Gibbs sampling is difficult since drawing the infinite-dimensional parameter F cannot be done cheaply [this step is done approximately in Doss (1994)]. Escobar (1994) describes a Gibbs sampler with the F integrated out, which takes advantage of the simplicity of the predictive distribution (10).

Having sampled ζ_t by means of sequential imputation and computed the associated importance weights M independent times, we have M imputed complete data sets with associated weights:

$$\mathbf{z}(j) = (\zeta_1(j), \dots, \zeta_n(j)) \quad \text{and} \quad w(j) \quad \text{for } j = 1, \dots, M.$$

Let $\mathbf{w} = (w(1), \dots, w(M))$ and $W = w(1) + \cdots + w(M)$. What follows is a summary of what we can do with these imputed $\mathbf{z}$'s.

THEOREM 5 (Sequential imputation). (a) *The posterior distribution of F can be approximated by a weighted mixture of Dirichlet processes*:

$$\mathscr{P}(F \mid \mathbf{y}) \approx \frac{1}{W} \sum_{j=1}^{M} w(j) \mathscr{D}\biggl(\alpha + \sum_{i=1}^{n} \delta_{\zeta_i(j)}\biggr).$$

(b) *The posterior expectation of F, which is also the predictive distribution for a future ζ, can be expressed as a weighted mixture of α and point masses*:

$$\mathscr{E}(F \mid \mathbf{y}) \approx \frac{1}{\|\alpha\| + n} \biggl\{\alpha + \frac{1}{W} \sum_{j=1}^{M} \sum_{i=1}^{n} w(j) \delta_{\zeta_i(j)}\biggr\}.$$

(c) *Posterior means and variances of the ζ's can be approximated by*

$$\mathscr{E}(\zeta_i \mid \mathbf{y}) \approx \frac{1}{W} \sum_{j=1}^{M} w(j) \zeta_i(j)$$

and

$$\mathrm{Var}(\zeta_i \mid \mathbf{y}) \approx \frac{1}{W} \sum_{j=1}^{M} w(j) \{\zeta_i(j)\}^2 - \{\mathscr{E}(\zeta_i \mid \mathbf{y})\}^2.$$

(d) *The posterior distribution of $N(\mathbf{z})$ can be approximated by*

$$\mathscr{P}\{N(\mathbf{z}) = k \mid \mathbf{y}\} \approx \frac{1}{W} \sum_{j=1}^{M} w(j) I_k\{N(\mathbf{z}(j))\},$$

where I_k is an indicator function such that $I_k(x)$ is 1 *if $x = k$ and* 0 *otherwise.*

The proof of the theorem is almost transparent and therefore omitted. It is noted that the approximations of the posterior means and variances of the ζ in (**c**) can be improved by incorporating Rao-Blackwellization [Gelfand and Smith (1990) and Liu, Wong and Kong (1994)], as has been implemented in Escobar (1994) for the Gibbs sampler. Let $\mathbf{z}_{[-t]}$ denote $(\zeta_i,\ i \neq t)$. Then instead of using the $\zeta_t(j)$ directly to estimate $\mathscr{E}(\zeta_t|\mathbf{y})$, for example, we can estimate it by

$$\frac{1}{W} \sum_{j=1}^{M} w(j) \mathscr{E}(\zeta_t \mid \mathbf{y}, \mathbf{z}_{[-t]}(j)).$$

The value of $\mathscr{E}(\zeta_t \mid \mathbf{y}, \mathbf{z}_{[-t]})$ can be found by using a distribution similar to (10). More precisely,

$$\mathscr{E}(\zeta_t \mid \mathbf{y}, \mathbf{z}_{[-t]}) = \frac{c(y_t + a)/(l_t + a + b) + \sum_{i \neq t} \zeta_i^{y_t + 1} (1 - \zeta_i)^{l_t - y_t}}{c B_{ab}(l_t, y_t) + \sum_{i \neq t} \zeta_i^{y_t} (1 - \zeta_i)^{l_t - y_t}}.$$

Similarly, we can compute $\mathrm{Var}(\zeta_t|\mathbf{Y}, \mathbf{z}_{[-t]})$. Then, by variance decomposition,

$$\mathrm{Var}(\zeta_t \mid \mathbf{y}) \approx \frac{1}{W}\sum_{j=1}^{M} w(j)\,\mathrm{Var}(\zeta_t|\mathbf{y}, \mathbf{z}_{[-t]}(j)) + \frac{1}{W}\sum_{j=1}^{M} w(j)\{\mathscr{E}(\zeta_t|\mathbf{y}, \mathbf{z}_{[-t]}(j)) - \mathscr{E}(\zeta_t|\mathbf{y})\}^2.$$

An implementation of the above estimates on the baseball data set of Efron and Morris (1975) showed a significant improvement over the raw approximations in Theorem 5. This improvement is expressed in a Monte Carlo calculation requiring fewer imputations to achieve the desired accuracy. Some of these results are shown in Kong, Liu and Wong (1994).

If we let the total prior mass c diminish to 0, then the posterior mean of F converges, in the limit as $c \to 0$, to $\mathrm{Beta}(Y+a, L-Y+b)$, where $L = l_1+\cdots+l_n$, $Y = y_1 + \cdots + y_n$. Actually,

$$\mathscr{P}(F \mid y_1, \ldots, y_n) \propto \int_0^1 \zeta^{Y+a}(1-\zeta)^{L-Y+b}\mathscr{D}(n\delta_\zeta)\,d\zeta.$$

The posterior means of the ζ's all shrink to the same value $\zeta_0 = (Y+a)/(L+a+b)$. This phenomenon is clear from the Pólya urn viewpoint of Blackwell and MacQueen (1973). When $c = \|\alpha\|$ is close to 0, all the balls drawn from the Polya urn scheme will be of the same "color," and the "color" is a priori distributed as $\mathrm{Beta}(a, b)$ by convention. Hence the model is reduced to a 1-spike model.

Suppose, on the other hand, the total a priori mass $c = \|\alpha\|$ increases to ∞. Then, by the same Pólya urn argument, the posterior mean of F approaches the mixture

$$\frac{1}{n+c}\left[\alpha + \sum_{i=1}^{n}\mathrm{Beta}(y_i + a, l_i - y_i + b)\right].$$

The posterior distribution of ζ_i converges to $\mathrm{Beta}(y_i + a, l_i - y_i + b)$.

PROPOSITION 1 (Reweighting). *If the continuous prior characteristic measure α of a Dirichlet process is changed to the continuous measure β, then the sequential imputation weights for the imputed vector* $\mathbf{z}$ *can be adjusted by*

$$\mathbf{w}_\beta(\mathbf{z}) = \mathbf{w}_\alpha(\mathbf{z})\frac{\mathscr{P}_\beta(\mathbf{y}, \mathbf{z})}{\mathscr{P}_\alpha(\mathbf{y}, \mathbf{z})},$$

where $\mathscr{P}_\alpha(\mathbf{y}, \mathbf{z})$ and $\mathscr{P}_\beta(\mathbf{y}, \mathbf{z})$ can be evaluated by Theorem 3.

The above result is also applicable for sensitivity analysis when using the Gibbs sampler, where we note that $\mathbf{w}_\alpha(\mathbf{z}) \equiv 1$ to begin with. The result can as well be used to conduct empirical Bayes analysis under the Dirichlet process model. Suppose that under a prior $\mathscr{D}(\alpha_0)$ process, the imputation weights are

$w(1), \dots, w(M)$. Then, by subsection 2.4 of Kong, Liu and Wong (1994), the predictive density of the data, when using the prior α_0, can be estimated by

$$\mathscr{P}_{\alpha_0}(\mathbf{y}) \approx \frac{1}{M} \sum_{j=1}^{M} w(j).$$

If, however, we change our prior to be $\mathscr{D}(q\alpha_0)$, where $q > 0$, then the ratio of the predictive density is, by Theorem 3 and the above reweighting proposition,

$$\frac{\mathscr{P}_{q\alpha_0}(\mathbf{y})}{\mathscr{P}_{\alpha_0}(\mathbf{y})} = \int \frac{\mathscr{P}_{q\alpha_0}(\mathbf{y},\mathbf{z})}{\mathscr{P}_{\alpha_0}(\mathbf{y},\mathbf{z})} \mathscr{P}_{\alpha_0}(\mathbf{z} \mid \mathbf{y})\, d\mathbf{z} = \sum_{k=1}^{n} \mathscr{P}\{N(\mathbf{z}) = k \mid \mathbf{Y}\} q^k \prod_{i=0}^{n-1} \frac{\|\alpha_0\| + i}{q\|\alpha_0\| + i}$$

and can be approximated by

$$\frac{1}{W} \sum_{j=1}^{M} q^{N(\mathbf{z}(j))} w(j) \prod_{i=0}^{n-1} \frac{\|\alpha_0\| + i}{q\|\alpha_0\| + i},$$

where $N(\mathbf{z}(j))$ is the number of distinct ζ's in the jth imputed vector $\mathbf{z}(j) = (\zeta_1(j), \dots, \zeta_n(j))$.

Now suppose that sequential imputations are implemented based on a specific prior characteristic measure (not necessarily the best one from a computational point of view), say, $\alpha_0 = c_0 \mathrm{Beta}(a, b)$. In many practical situations we tend to choose the initial $c_0 = 1$. Then let $c = qc_0$, and from the above argument we see that the maximum likelihood estimate of c, that is, the "best" weight to put in front of the prior measure α with shape $\mathrm{Beta}(a, b)$, can be found by maximizing the function

$$\hat{l}(c) = \log\left\{\frac{1}{W} \sum_{j=1}^{M} w(j) \left(\frac{c}{c_0}\right)^{N(\mathbf{z}(j))}\right\} - \sum_{i=0}^{n-1} \log(c + i),$$

which is an approximation of the log-likelihood function $\log\{L(c, \mathbf{y})\}$ introduced in Section 2. The maximizer $\tilde{c}$ of $\hat{l}(c)$ is an approximation of the maximum likelihood estimate $\hat{c}$. Since, as $M \to \infty$, $\hat{l}(c)$ converges almost surely to $\log\{L(c, \mathbf{y})\}$ pointwise on a dense set, by an argument similar to that of Geyer and Thompson (1992), the maximizer of $\hat{l}(c)$ converges to the true maximizer of $\log\{L(c, \mathbf{y})\}$ almost surely, provided that the latter is unique. Letting $d = \log(c)$ and differentiating $\log\{L(\log(d), \mathbf{y})\}$ twice with respect to d, we can easily show that $\log\{L(\log(d), \mathbf{y})\}$ is a concave function in d and therefore has a unique maximum. Hence the maximizer of $\log\{L(c, \mathbf{y})\}$ is also unique.

What initial α_0 should we choose so as to best estimate the weight parameter c? Our experience shows that the larger the initial c_0 is, the smaller the variance of the importance sampling weights for sequential imputations. Therefore, c_0 should be chosen slightly larger than the maximum likelihood estimate of c to obtain a more stable result. However, if c_0 is too large, the actual value of q tends to be too small and its coefficient of variation tends to increase. For instance, it is understood that $c = 0$ and $c = \infty$ are two singular solutions to the normal equation in Theorem 4, which also implies that in practice $\|\alpha_0\|$ should be neither too small nor too large. If we let $\|\alpha_0\| \gg n$, for example, then all the imputed ζ's are distinct. As a consequence, the function $\hat{l}(c)$ is monotone increasing and $\hat{c}$ does not exist.

REMARK. We decide to follow a "rule of thumb" in sampling theory [see details in Kong, Liu and Wong (1994) and Liu (1996)] to assess the efficiency of a general importance sampling plan. The rule suggests that the *efficiency* of an importance sampling plan is measured by the coefficient of variation of the importance weight (i.e., the variance of the renormalized weights). More precisely, a sample of size M drawn from the trial density $f(\mathbf{z})$ can be regarded as "equivalent" to an *effective* sample of size M^* from the true distribution p, where

$$M^* = \frac{M}{1 + \mathrm{var}(w^*)},$$

in which $w^*(\mathbf{z}) = p(\mathbf{z})/f(\mathbf{z})$ and the variance is taken under f. Usually, w^* can only be computed up to a norming constant and is often approximated by renormalization. Other measures of efficiency, for example, the entropy measure, are also possible, but will not be used in our case.

4. Binary data analysis using Dirichlet process.

4.1. *Rolling thumbtacks.* Beckett and Diaconis (1994) generated binary strings from rolls of common thumbtacks. A 1 was recorded if the tack landed point up and a 0 was recorded if the tack landed point down. All tacks started point down. Each tack was flicked or hit with the fingers from where it last rested. Each tack was flicked 9 times. The data, consisting of 320 9-tuples, are reproduced in Table 1. The actual data arose from 16 different tacks, 2 "flickers" and 10 surfaces. For simplicity, we treated the data as though they came from 320 different tacks. We further assume that, conditioned on a certain tack, the results of the 9 different flips are *independent*. In their paper,

TABLE 1
Tack data taken from Beckett and Diaconis (1994)

7	4	6	6	6	6	8	6	5	8	6	3	3	7	8	4	5	5	7	8	5	7	6	5	3	2	7	7	9	6	4	6
4	7	3	7	6	6	6	5	6	6	5	6	5	6	7	9	9	5	6	4	6	4	7	6	8	7	7	2	7	7	4	6
2	4	7	7	2	3	4	4	4	6	8	8	5	6	6	6	5	3	8	6	5	8	6	6	3	5	8	5	5	5	5	6
3	6	8	6	6	6	8	5	6	4	6	8	7	8	9	4	4	4	4	6	7	1	5	6	7	2	3	4	7	5	6	5
2	7	8	6	5	8	4	8	3	8	6	4	7	7	4	5	2	3	7	7	4	5	2	3	7	4	6	8	6	4	6	2
4	4	7	7	6	6	6	8	7	4	4	8	9	4	4	3	6	7	7	5	5	8	5	5	5	6	9	1	7	3	3	5
7	7	6	8	8	8	8	7	5	8	7	8	5	5	8	8	7	4	6	5	9	8	6	8	9	9	8	8	9	5	8	6
3	5	9	8	8	7	6	8	5	9	7	6	5	8	5	8	4	8	8	7	7	5	4	2	4	5	9	8	8	5	7	7
2	6	2	7	6	5	4	4	6	9	3	9	4	4	1	7	4	4	5	9	4	7	7	8	4	6	7	8	7	4	3	5
7	7	4	4	6	4	4	2	9	9	8	6	8	8	4	5	7	5	4	6	8	7	6	6	8	6	9	6	7	6	6	6

Beckett and Diaconis (1994) provide a spectral analysis of the data and reject the ordinary iid model of rolling. Here we would like to review the problem from a hierarchical Bayesian viewpoint. The numbers shown in Table 1 are the number of up's out of 9 flips for each of the 320 binary strings. A histogram of the data is shown in Figure 1.

The binomial–Dirichlet nonparametric Bayes model is applicable. In this example, $l_i = 9$ for all i and $n = 320$. Of course, in general, we do not require that all the l_i's be equal. Clearly, we would not expect that all ζ_i's are the same, but rather assume an unknown distribution F to govern the value of the ζ's. One may think that a histogram as shown in Figure 1 is enough to explain the data. Our later result, however, shows an unusual feature.

Choosing $M = 10,000$, we applied sequential imputations to estimate the posterior processes, with each of 4 prior measures $\alpha = c\text{Beta}(1,1)$, where the weight factor c was taken as 0.1, 1, 5 and 10, respectively. Since F is an infinite-dimensional parameter, there is no easy way of displaying its full posterior distribution. Its posterior mean, which can also be viewed as the predictive distribution of a future tack, is demonstrated to reveal a surprising bimodality. Clearly, this feature is unexpected and cannot be revealed by a regular parametric hierarchical analysis using the Beta–binomial priors. A superficial explanation for this might be that the tack data were produced by two persons with some systematic difference in their flipping.

Plots of the four posterior mean densities of F corresponding to different c's are shown in Figure 2. More precisely, F was approximated by $\sum_j w(j)\mathscr{D}(\alpha(j))/W$, where $\alpha(j) = \alpha + \zeta_1(j) + \cdots + \zeta_n(j)$, $j = 1, \ldots, M$, and

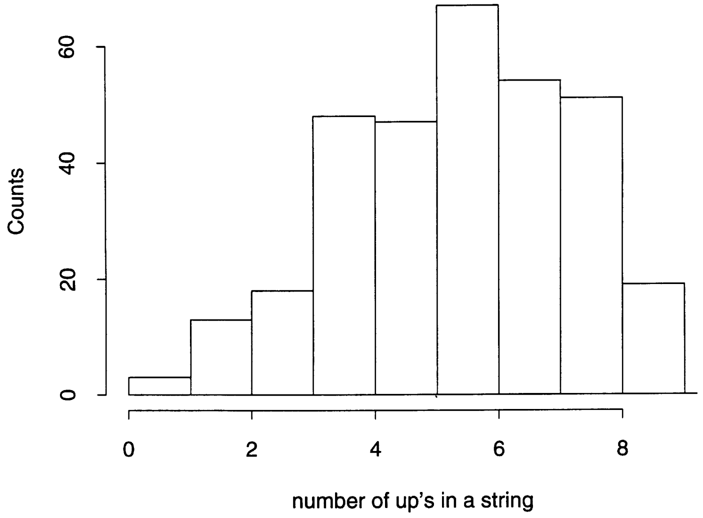

FIG. 1. *Histogram of the tack data produced by Beckett and Diaconis* (1994).

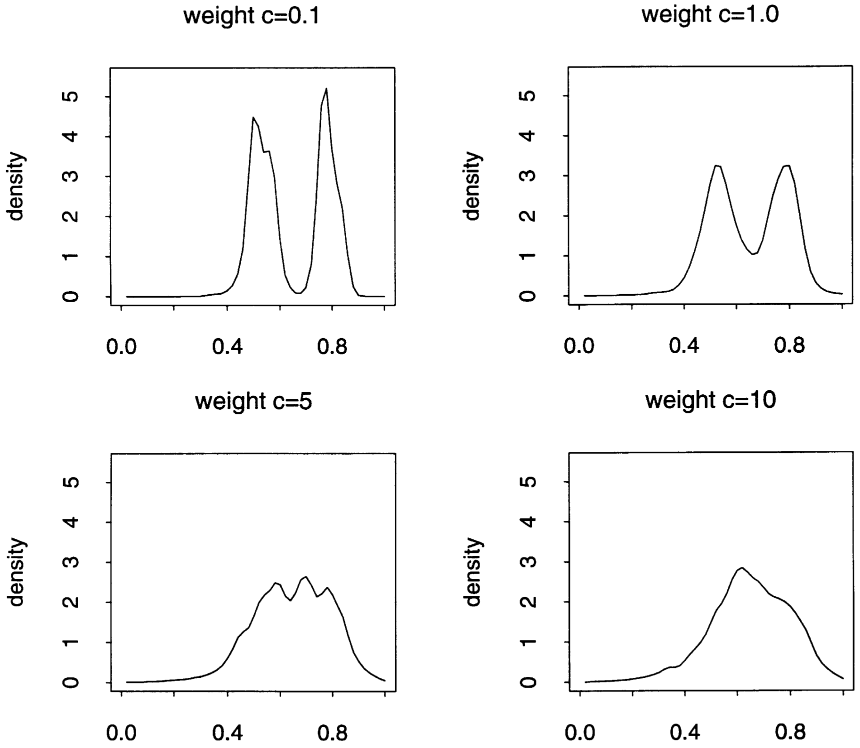

FIG. 2. *Approximated posterior means of F, that is, $\mathscr{E}(F|\mathbf{y})$, based on four different prior weights.*

the curve in Figure 2 is just

$$\frac{1}{W}\sum_{j=1}^{M} w(j)\alpha(j).$$

The smoothness of the curves is due to a Gaussian kernel smoothing of the weighted histogram indicated in the above formula. It is seen that when c is increased, the bimodality of the curve is reduced. The coefficients of variation of the importance weights in the four situations, that is, $c = 0.1$, 1, 5 and 10, were 378, 43, 34 and 33, respectively. With $M = 10,000$, the effective sample sizes in the four situations were about 26, 227, 286 and 300, respectively.

The estimated posterior distribution of $N(\mathbf{z})$, when $c = 1$, is shown in Figure 3, with the posterior mean and variance being 6.342 and 3.495. They are compared with the prior mean

$$\mathscr{E}(N) = \sum_{i=1}^{320}\frac{1}{i} = 6.347$$

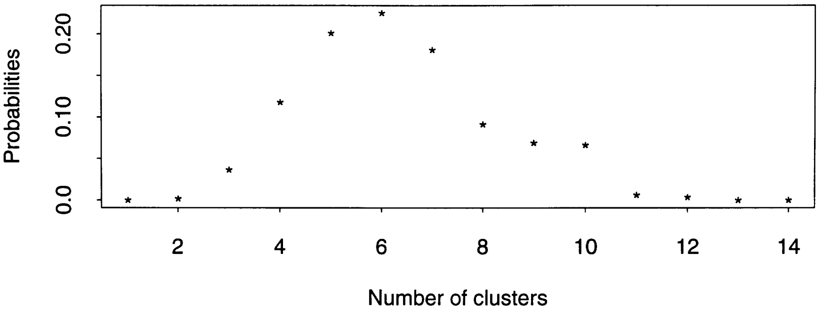

FIG. 3. *Approximation of* $\mathscr{P}(N(\mathbf{z})|\mathbf{y})$.

and the prior variance

$$\operatorname{Var}(N) = \sum_{i=1}^{319} \frac{i}{(i+1)^2} = 4.705.$$

Based on this estimated distribution, we computed the approximate MLE of the weight parameter as $\hat{c} = 1.02$. The estimated variance of $\log(\hat{c})$, that is, the inverse of the observed Fisher information, is as large as 0.826. The log-likelihood surface of c is plotted in Figure 4 based on our sequentially imputed complete data sets and the associated weights. The maximum likelihood solution for c seems to suggest that the bimodal feature is "real." On the other hand, however, the relatively large variance of $d = \log(c)$ at $c = 1.02$ suggests that a better representation of the distribution of ζ may be a weighted mixture of the posterior mean curves over a wide range of different values of c. In other words, the weighting factor c can be subjected to a Bayesian analysis assuming a reasonable prior.

As a comparison, we applied Laird's (1978) method, the so-called NPMLE (nonparametric MLE), to the tack data via an EM algorithm using an m-spike model. Starting with $m = 3, \dots, 10, 13, 30$, we found that it always converged to a 3-spike model with spikes at (0.4385, 0.6013, 0.8144) and mixing weights 0.1906, 0.4402 and 0.3692. Chernoff (1994) reexamined the data by assuming a mixture of two beta distributions for F and obtained some interesting results.

One may have noticed that the expected number of clusters for the ζ's from our nonparametric Bayesian analysis is about 6, while the MLE solution is 3. The mean density of F shown in panel (1, 2) of Figure 2, however, reveals only two modes. To investigate these apparent discrepancies, we decided to fit an m-spike model for F using the Bayes method, which is equivalent to our Dirichlet process model conditioned on $N(\mathbf{z}) = m$. Precisely, we assume that F follows model (3), with unknown a's and z's. A Dirichlet $(1/m, \dots, 1/m)$ prior was used for the a's and independent Beta(1, 1) priors were assumed on all the z's. We ask the same question: what would be the posterior mean density of F? To our surprise, for $m = 2, \dots, 6$, 10 and 12, this density is always bimodal (almost), just like panels (1, 1) and (1, 2) of Figure 2. The two modes

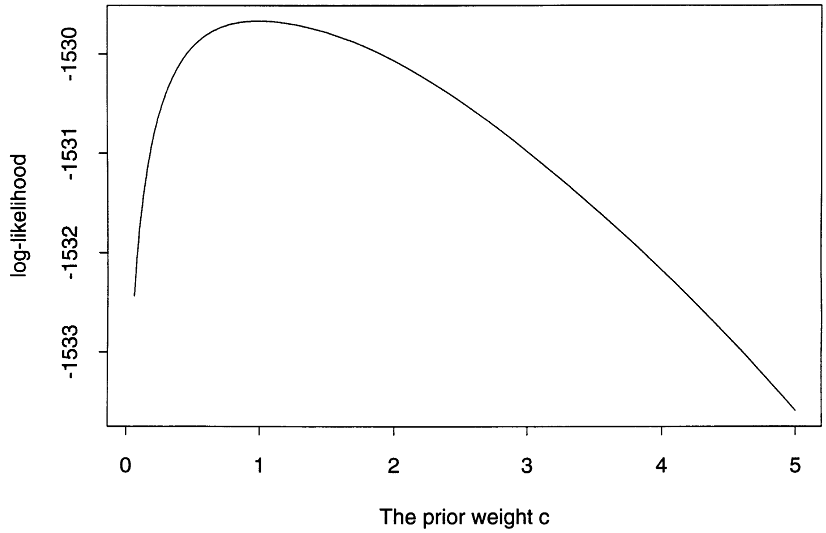

FIG. 4. *Approximate log-likelihood curve for the weight c. The maximum is attained at* $c = 1.02$.

are always around 0.52 and 0.79. When m increases, the magnitude of the two modes decreases and the probability mass between the two modes increases. When $m \geq 6$, a third small mode appears at about 0.65 between two large modes. It persists when $m = 10$ but disappears when $m = 12$. Comparing the fits for $m = 5$, 6 and 10, we find that the density shape seems to be "stabilized" after m is 5 or 6.

These results may provide some insight into the connections between NPMLE and Dirichlet process modeling. First, $N(\mathbf{z}) = 6$ seems to be best supported by the data and, conditioned on which, the mean density of F resembles panel $(1, 2)$ of Figure 2, revealing not only two major modes but also a substantial amount of probability mass between them. A full Dirichlet process model weighted averages over all different $N(\mathbf{z})$'s. Second, the third minor mode in the mean shape of F for the 6-spike model seems to be echoing the result of the NPMLE. Nevertheless, further theoretical studies are called for to completely understand these subtle connections and either to confirm or to invalidate the above explanations.

4.2. *Gibbs sampling and sequential imputations.* Gibbs sampling can also be applied to treat the same problem. Since drawing the infinite-dimensional parameter F is infeasible, one needs to "collapse down" (integrate out) F in the sampler [Escobar (1994)]. General forms of "collapsing" in Bayesian computations are treated in Liu, Wong and Kong (1994) and Liu (1994). A better collapsing scheme for the Dirichlet process problems is proposed by MacEachern (1994), which produces a much more efficient Gibbs sampler. It is not easy to make a general comparison between Gibbs sampling strategy and the sequential imputation method because there are many variations for each method. We only note here that, since sequential imputation is advantageous in updating posterior distributions when new data (incomplete) arrive, Gibbs sampling and sequential imputation can be complementary to each other.

To conclude, we present the following observation in an extreme situation. Suppose we want to draw a sample of size n, $\zeta_1, \dots, \zeta_n$, from an F that follows $\mathscr{D}(\alpha)$ with α being uniform (this corresponds to the situation where all the y's are missing). The goal can be achieved by either running a Gibbs sampler or doing sequential imputation. Each sequence of sequential imputation draws is equivalent to a Pólya urn sequence and is therefore an exact draw from $\mathscr{D}(\alpha)$ that results in full efficiency [Blackwell and MacQueen (1973)]. For the Gibbs sampler, we prove the following theorem in the Appendix.

THEOREM 6. *Let α be uniform on* $[0, 1]$. *A Gibbs sampler with a* random scan *is applied to draw n exchangeable samples $\zeta_1, \dots, \zeta_n$ from a Pólya urn with parameter α. Then the average regeneration time of the chain is $\pi^2 n^2/6$. Here the* random scan *refers to the one that visits each component at random with equal probability.*

APPENDIX

PROOF OF THEOREM 4. By differentiating the log-likelihood function of c and setting it to 0, we get

$$\frac{\sum_{m=1}^{n} m c^{m-1} \mathscr{L}_m}{\sum_{m=1}^{n} c^m \mathscr{L}_m} - \sum_{i=1}^{n} \frac{1}{c+i-1} = 0.$$

By Lemma 2.1 of Korwar and Hollander (1973), it follows that

$$\frac{\sum_{m=1}^{n} m c^{m} \mathscr{L}_m}{\sum_{m=1}^{n} c^m \mathscr{L}_m} = \mathscr{E}\{N(\mathbf{z})\}.$$

The first conclusion then follows from Theorem 2. Furthermore, since $\mathbf{i}_{\text{obs}}(d) = -\partial^2 l(d)/\partial d^2$, where

$$l(d) = \log\left\{\sum_{m=1}^{n} e^{md} \mathscr{L}_m\right\} - \sum_{i=1}^{n} \log(e^d + i - 1) + \text{const.},$$

the second conclusion follows from the observations that

$$\frac{\partial^2}{\partial d^2} \log\left\{\sum_{m=1}^{n} e^{md} \mathscr{L}_m\right\} = \operatorname{Var}\{N(\mathbf{z}) \mid y_1, \dots, y_n\}$$

and

$$\frac{\partial^2}{\partial d^2} \sum_{m=1}^{n} \log(e^d + m - 1) = \sum_{m=1}^{n} \frac{e^d(m-1)}{(e^d + m - 1)^2}.$$

The right-hand side is just $\mathrm{Var}\{N(\mathbf{z})\}$ by Korwar and Hollander (1973). Next,

$$\begin{aligned}\mathbf{i}_{\mathrm{exp}}(d) = \mathscr{E}(\mathbf{i}_{\mathrm{obs}}(d)) &= \mathrm{Var}\{N(\mathbf{z})\} - \mathscr{E}[\mathrm{Var}_{\alpha}\{N(\mathbf{z}) \mid y_1, \dots, y_n\}] \\ &= \mathrm{Var}[\mathscr{E}\{N(\mathbf{z}) \mid y_1, \dots, y_n\}].\end{aligned}$$

Thus the conclusion follows. □

Proof of Theorem 6. For a starting vector $\mathbf{z}^{(0)} = (\zeta_1^{(0)}, \dots, \zeta_n^{(0)})$, we realize that the time when all these "old" values are replaced by the new draws is a regeneration time, that is, the first time T such that the sets $\{\zeta_1^{(T)}, \dots, \zeta_n^{(T)}\}$ and $\{\zeta_1^{(0)}, \dots, \zeta_n^{(0)}\}$ have an empty intersection. This regeneration time does not depend on the actual value of $\mathbf{z}^{(0)}$. Therefore, we can couple our nonstationary chain with a chain started from stationarity so that the both chains have a common regeneration time. Thus, the regeneration time is the time needed for convergence.

Now we calculate this regeneration time. At any stage t, we classify the elements in $\mathbf{z}^{(t)}$ into two groups, one of which is

$$Z_{\mathrm{old}}^{(t)} = \{\zeta_1^{(t)}, \dots, \zeta_n^{(t)}\} \cap \{\zeta_1^{(0)}, \dots, \zeta_n^{(0)}\},$$

and the other is $Z_{\mathrm{new}}^{(t)}$. We denote the number of elements in $Z_{\mathrm{new}}^{(t)}$ as $V^{(t)}$. Then $V^{(t)}$ is a birth–death chain with transition probabilities

$$\begin{aligned}a_k &= P(V^{(t+1)} = k+1 \mid V^{(t)} = k) = \frac{(n-k)(k+1)}{n^2}, \\ b_k &= P(V^{(t+1)} = k-1 \mid V^{(t)} = k) = \frac{k(n-k)}{n^2}.\end{aligned}$$

The reason, for example, for the first probability is that in order to increase the size of Z_{new}, our sampler needs to visit an "old" component first, which has a probability $(n-k)/n$; and then either a draw from α or a draw from $Z_{\mathrm{new}}^{(t)}$ has to happen, the probability of which is $(k+1)/n$. The regeneration time is now equal to the time from $V^{(0)} = 0$ to some T such that $V^{(T)} = n$. When the endpoint n is regarded as an absorbing boundary, we have the recurrence relation

$$D_k = 1 + a_k D_{k+1} + b_k D_{k-1} + (1 - a_k - b_k) D_k, \qquad k = 1, \dots, n-1,$$

and $D_0 = 1 + p_0 D_1 + (1 - p_0) D_0$, $D_n = 0$. Here D_k denotes the expected duration of the chain before hitting n when started from k. It is apparent that $E(T) = D_0$. Furthermore, we find that

$$D_k - D_{k+1} = \frac{n^2}{k+1}\left\{\frac{1}{n-k} + \cdots + \frac{1}{n}\right\}.$$

This together with $D_n = 0$ provides us with

$$D_0 = n^2 \sum_{k=0}^{n-1} \sum_{j=k+1}^{n} \frac{1}{(n-k)j} = n^2\left(1 + \frac{1}{2^2} + \cdots + \frac{1}{n^2}\right) \to \frac{\pi^2}{6} n^2.$$

The second equality can be easily proved by induction. □

Acknowledgments. The author thanks Professors Augustine Kong and Persi Diaconis for providing insightful discussions; Professor Herman Chernoff for reading the manuscript and making thoughtful suggestions; and the Editor, Associate Editor and two referees for many valuable and detailed comments that especially helped to improve the presentations of Sections 2 and 4. Part of the manuscript was prepared when the author was an assistant professor in the Department of Statistics at Harvard University.

REFERENCES

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174.

BECKETT, L. and DIACONIS, P. (1994). Spectral analysis for discrete longitudinal data. *Adv. Math.* **103** 107–128.

BERRY, D. and CHRISTENSEN, R. (1979). Empirical Bayes estimation of a binomial parameter via mixture of Dirichlet processes. *Ann. Statist.* **7** 558–568.

BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355.

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modeling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A* **143** 383–430.

CHERNOFF, H. (1994). Personal communication.

DOKSUM, K. A. (1972). Decision theory for some nonparametric models. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 331–343. Univ. California Press, Berkeley.

DOSS, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Ann. Statist.* **22** 1763–1786.

EFRON, B. and MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311–319.

ESCOBAR, M. D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89** 268–277.

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

FERGUSON, T. S. (1974). Prior distribution on space of probability measures. *Ann. Statist.* **2** 615–629.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

GEYER, C. J. and THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 657–699.

KONG, A., LIU, J. S. and WONG, W. H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.* **89** 278–288.

KORWAR, R. M. and HOLLANDER, M. (1973). Contributions to the theory of Dirichlet processes. *Ann. Statist.* **1** 705–711.

KUO, L. (1986). Computations of mixtures of Dirichlet processes. *SIAM J. Sci. Statist. Comput.* **7** 60–71.

LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.* **73** 805–811.

LIU, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with application to a gene regulation problem. *J. Amer. Statist. Assoc.* **89** 958–966.

LIU, J. S. (1996). Metropolized independent sampling scheme with comparisons to rejection sampling and importance sampling. *Statist. Comput.* **6**. To appear.

LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40.

LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357.

MACEACHERN, S. M. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Comm. Statist. Simulation Comput.* **23** 727–741.

SMITH, A. F. M. and ROBERTS, G. O. (1993). Bayesian computation via the Gibbs samples and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **55** 3–23.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550.

WEST, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. ISDS Discussion Paper 92-A03, Duke Univ.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
E-MAIL: jliu@stat.stanford.edu