The Canadian Journal of Statistics
Vol. 40, No. 2, 2012, Pages 328–344
La revue canadienne de statistique

Nonparametric hierarchical Bayes analysis of binomial data via Bernstein polynomial priors

Tingting ZHANG[1] and Jun S. LIU[2]*

[1] *University of Virginia, Charlottesville, VA, USA*
[2] *Harvard University, Cambridge, MA, USA*

Key words and phrases: Beta-binomial; Bernstein–Dirichlet process; Dirichlet process; posterior consistency.

MSC 2011: Primary 62C10; secondary 62G07.

Abstract: For binomial data analysis, many methods based on empirical Bayes interpretations have been developed, in which a variance-stabilizing transformation and a normality assumption are usually required. To achieve the greatest model flexibility, we conduct nonparametric Bayesian inference for binomial data and employ a special nonparametric Bayesian prior—the Bernstein–Dirichlet process (BDP)—in the hierarchical Bayes model for the data. The BDP is a special Dirichlet process (DP) mixture based on beta distributions, and the posterior distribution resulting from it has a smooth density defined on [0, 1]. We examine two Markov chain Monte Carlo procedures for simulating from the resulting posterior distribution, and compare their convergence rates and computational efficiency. In contrast to existing results for posterior consistency based on direct observations, the posterior consistency of the BDP, given indirect binomial data, is established. We study shrinkage effects and the robustness of the BDP-based posterior estimators in comparison with several other empirical and hierarchical Bayes estimators, and we illustrate through examples that the BDP-based nonparametric Bayesian estimate is more robust to the sample variation and tends to have a smaller estimation error than those based on the DP prior. In certain settings, the new estimator can also beat Stein's estimator, Efron and Morris's limited-translation estimator, and many other existing empirical Bayes estimators. *The Canadian Journal of Statistics* 40: 328–344; 2012 © 2012 Statistical Society of Canada

Résumé: Pour une analyse de données binomiales, plusieurs méthodes ayant des interprétations bayésiennes empiriques ont été développées pour lesquelles une transformation stabilisant la variance et le présupposé de normalité sont habituellement nécessaires. Afin d'obtenir un modèle avec la plus grande flexibilité, nous faisons une inférence bayésienne non paramétrique pour des données binomiales et nous utilisons une densité *a priori* non paramétrique spéciale, un processus Bernstein-Dirichlet (B-D), dans le modèle bayésien hiérarchique des données. Le processus B-D est un cas particulier d'un mélange de processus de Dirichlet (D) basé sur les distributions bêta. La distribution *a priori* résultante possède une densité lisse définie sur [0, 1]. Nous considérons deux procédures de chaînes de Monte-Carlo markoviennes afin de simuler à partir de la densité *a posteriori* résultante et nous comparons leur taux de convergence et leur efficacité de calcul. Contrairement aux résultats déjà existants pour la cohérence *a posteriori* basée sur des observations directes, la cohérence *a posteriori* du processus B-D, étant donné des données binomiales indirectes, est obtenue. Nous comparons les fonctions de rétrécissement et la robustesse des estimateurs *a posteriori* basés sur les processus B-D avec plusieurs autres estimateurs bayésiens empiriques et hiérarchiques. à l'aide d'exemples, nous voyons que l'estimateur bayésien non paramétrique basé sur le processus B-D est plus robuste par rapport à des variations échantillonnales et il tend à avoir une plus petite erreur d'estimation que ceux basés sur des densités *a priori* basées sur les processus D. Dans certains cas, le nouvel estimateur performe mieux que l'estimateur de Stein, l'estimateur à translation limitée d'Efron et Morris et plusieurs autres estimateurs bayésiens empiriques. *La revue canadienne de statistique* 40: 328–344; 2012 © 2012 Société statistique du Canada

Supporting information can be found in the online version of this paper.

* *Author to whom correspondence may be addressed.*
E-mail: jliu@stat.harvard.edu

1. INTRODUCTION

Suppose we have n independent binomial data $y_i \sim \text{Binom}(N_i, \theta_i)$, $i = 1, \ldots, n$, where the N_i's are predetermined numbers of trials and the θ_i's are the unknown probabilities of success. It is of interest to estimate $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and to predict a new binomial probability θ_{n+1}. This classical problem has been addressed by Efron & Morris (1971, 1972, 1975), who introduced an interesting limited-translation estimator under a parametric empirical Bayes framework, and by others (e.g., Berry & Christensen, 1979; Lo, 1984; Kong, Liu, & Wong, 1994; Escobar & West, 1995; Liu, 1996) under a Dirichlet-process-based nonparametric Bayes formulation. Recently, these types of data have been reanalyzed by Brown (2008) using several methods arising from empirical Bayes and hierarchical Bayes interpretations. The empirical Bayes approaches (Efron & Morris, 1975; Brown, 2008) usually conduct a variance-stabilizing transformation of the binomial data as a preliminary step, and assume normality of the transformed data in the analysis. To achieve better model flexibility, we take a nonparametric Bayes approach for inferring $\boldsymbol{\theta}$ in this paper.

Due to the pioneering work by Ferguson (1973, 1974), Lo (1984), and Antoniak (1974), the Dirichlet process (DP) has been widely used as a prior distribution for unknown probability measures and employed in nonparametric Bayesian inference (Escobar, 1994; Escobar & West, 1995; Liu, 1996; MacEachern, 1994; MacEachern, Clyde, & Liu, 1999; see MacEachern & Müller, 2000, for a review). In a typical nonparametric Bayesian procedure, one assumes that the θ_i's are independent and identically distributed (i.i.d.) samples from an unknown distribution F, and F follows a DP, denoted as $F|\alpha \sim \mathcal{D}(\alpha)$ (Liu, 1996), where α is a given finite measure on the sample space Λ on which F is defined, and is called the characteristic measure of the DP. The DP is an almost surely discrete random probability measure (Blackwell & MacQueen, 1973; Ferguson, 1973), and its posterior distribution is also discrete. In addition, the Bayes estimator resulting from the DP, which usually is the posterior mean, tends to have unnatural sharp peaks—even though it is absolutely continuous if α is absolutely continuous (Berry & Christensen, 1979; Liu, 1996; MacEachern, Clyde, & Liu, 1999). A simple and popular extension to remove the constraint to discrete random measures (Müller & Quintana, 2004) is to use a DP mixture (Escobar, 1988; MacEachern, 1994; Escobar & West, 1995), where the distribution F is represented as a convolution of a random measure from the DP and a given smooth density function f:

$$F(x) = \int f(x|\theta)\mathrm{d}G(\theta) \text{ with } G \sim \mathcal{D}(\alpha). \tag{1}$$

DP mixtures of Gaussian densities have been studied by Lo (1984), Escobar & West (1995), and Gasparini (1996). Nonparametric models based on DP mixtures are reviewed in MacEachern & Müller (2000).

Due to the binomial data under study, we focus on the special case where Λ is the unit interval [0, 1]. Then it is natural to use a DP mixture of beta densities in nonparametric Bayesian inference for binomial data. Proposed by Petrone (1999a, b), a new process based on Bernstein polynomials, which is a special DP mixture of beta densities, can be employed for binomial data analysis; we refer to it as the Bernstein–Dirichlet process (BDP) in the following. Given a positive integer k and a function G with support [0, 1], the Bernstein polynomial of order k and G is defined as $B(x; k, G) = \sum_{j=1}^{k} G(\frac{j}{k})\binom{k}{j} x^j (1-x)^{k-j}$, for $x \in [0, 1]$. If G is a cumulative distribution function (CDF) on [0, 1], so is $B(x; k, G)$, and we call it a Bernstein distribution. The BDP is defined as follows.

Definition 1. *Let η be a discrete distribution with support $\{1, 2, \ldots\}$, M a positive constant, and F_0 a given CDF on interval [0, 1]. A probability distribution F is said to follow a BDP with parameters (η, MF_0), denoted as $BD(\eta, MF_0)$, if $F = B(; k, G)$ for $G \sim \mathcal{D}(MF_0)$ and independently $k \sim \eta$.*

The BDP is also called the Bernstein–Dirichlet (BD) prior by Petrone (1999a, b); its posterior consistency was shown by Petrone & Wasserman (2002) when direct observations from F are available. Thus, the BDP is a promising alternative to the DP in nonparametric Bayesian inference for absolutely continuous distributions defined on [0, 1]. In this paper, through theoretical analysis, simulation examples, and real data applications, we study the use of the BDP in the following hierarchical Bayes setting:

$$\begin{aligned} y_i|N_i &\sim \text{Binom}(N_i, \theta_i), i = 1, \ldots, n, \\ \theta_i &\overset{i.i.d.}{\sim} F, i = 1, \ldots, n, \\ F &\sim BD(\eta, MF_0), \end{aligned} \tag{2}$$

where the θ_i's are unobserved.

The paper is organized as follows: Section 2 briefly describes the posterior distribution of F under the framework (2). Section 3 studies posterior consistency of the BDP given binomial data under two different scenarios: (i) the maximum value of the N_i's is fixed, and (ii) both the N_i's and n go to infinity. Section 4 examines computational issues of the problem. Section 5 presents simulation studies to compare predictive densities when the BDP and the DP are, respectively used as priors for F. Using the average of squared error (ASE) as a criterion, we also compare the point estimates of $\boldsymbol{\theta}$ under the above two nonparametric Bayesian settings with those of Stein's estimator, and Efron and Morris's limited-translation estimator. Section 6 applies the new nonparametric Bayesian procedure to real batting-average data from Brown (2008). Comparisons are drawn with empirical Bayes and hierarchical Bayes estimators analyzed by Brown (2008) through this real data analysis.

2. POSTERIOR OF HIERACHICAL BAYES MODELS WITH BD PRIORS

Let $Beta(a, b)$ be the beta CDF with parameters a and b, and $\beta(x; a, b)$ be the associated density evaluated at x. Denote the conditional distribution of X given $Y_1, \ldots, Y_l$ by $[X \mid Y_1, \ldots, Y_l]$.

We name the derivative of the Bernstein distribution $B(; k, G)$ the Bernstein density, which is of the form

$$b(x; k, G) = \sum_{j=1}^{k} W_{j,k}\beta(x; j, k - j + 1), \tag{3}$$

where the weights $W_{j,k} = G(j/k) - G(j - 1/k)$. Since $b(; k, G)$ and $B(; k, G)$ are also uniquely defined by k and $\mathbf{W}_k = (W_{1,k}, \ldots, W_{k,k})$, we use $b(x; k, G)$ and $b(x; k, \mathbf{W}_k)$ interchangeably in the following. The posterior inference for F is focused on the posterior distribution of parameters k and $\mathbf{W}_k$. Since $G \sim \mathcal{D}(MF_0)$ implies that the weights $\mathbf{W}_k$ of the beta densities $\beta(; j, k - j + 1)$, $j = 1, \ldots, k$ in (3) follow a Dirichlet distribution with parameters $(M\alpha_{1,k}, \ldots, M\alpha_{k,k})$, where $\alpha_{j,k} = F_0(j/k) - F_0(j - 1/k)$, $j = 1, \ldots, k$, it is easy to show that the joint posterior distribution of $(k, \mathbf{W}_k)$ given $\boldsymbol{\theta}$ (Petrone, 1999a, b) is proportional to

$$\eta(k)\Gamma(M)\left(\prod_{j=1}^{k}\Gamma(M\alpha_{j,k})\right)^{-1}\prod_{j=1}^{k}W_{j,k}^{M\alpha_{j,k}-1}\prod_{i=1}^{n}b(\theta_i; k, \mathbf{W}_k).$$

 DOI: 10.1002/cjs

Given the indirect binomial data considered here, the joint posterior distribution $f(\mathbf{W}_k, k|\mathbf{y}_n, \mathbf{N}_n)$ is proportional to

$$\eta(k)\frac{\Gamma(M)}{\prod_{j=1}^{k}\Gamma(M\alpha_{j,k})}\prod_{j=1}^{k}W_{j,k}^{M\alpha_{j,k}-1}\prod_{i=1}^{n}\psi_{k,\mathbf{W}_k}(y_i|N_i),$$

where

$$\begin{aligned}\psi_{k,\mathbf{W}_k}(y_i|N_i) &= \int_0^1 \frac{N_i!}{y_i!(N_i-y_i)!}\theta^{y_i}(1-\theta)^{N_i-y_i}b(\theta; k, \mathbf{W}_k)\mathrm{d}\theta \\ &= \sum_{j=1}^{k} W_{j,k}\frac{N_i!}{y_i!(N_i-y_i)!}\frac{Q(k+1, j)}{Q(k+1+N_i, j+y_i)}, \end{aligned} \tag{4}$$

and the function $Q(\cdot, \cdot)$ is defined as

$$Q(k+1, j) = \frac{\Gamma(k+1)}{\Gamma(j)\Gamma(k+1-j)}. \tag{5}$$

The predictive distribution, that is, the posterior mean $E(F|\mathbf{y}_n, \mathbf{N}_n)$, which is of the form $\sum_k \eta(k|\mathbf{y}_n, \mathbf{N}_n)\sum_{j=1}^{k} E(W_{j,k}|\mathbf{y}_n, \mathbf{N}_n, k)\beta(\theta; j, k+1-1)$, is difficult to calculate directly. In real applications, we use Markov chain Monte Carlo (MCMC) simulations to approximate the predictive distribution. Suppose S MCMC samples of $(\mathbf{W}_k, k)$ from their posterior distribution have been produced. Then the predictive density estimate, denoted by $\widehat{b}(\theta)$, is given by

$$\widehat{b}(\theta) = \frac{1}{S}\sum_{s=1}^{S}\sum_{j=1}^{k^{(s)}} W_{j,k^{(s)}}^{(s)}\beta(\theta; j, k^{(s)}+1-j), \tag{6}$$

where $\left\{W_{1,k^{(s)}}^{(s)}, \ldots, W_{k^{(s)},k^{(s)}}^{(s)}, k^{(s)}\right\}$ is the sth sample.

3. POSTERIOR CONSISTENCY OF THE BDP

The asymptotic behavior of general Bayes estimates has been investigated by Freedman (1963, 1965), Schwartz (1965), Diaconis & Freedman (1983, 1986), and many others. The posterior consistency of the DP was established by Barron, Schervish, & Wasserman (1999), and Ghosal, Ghosh, & Ramamoorthi (1999b) proved posterior consistency of the DP mixture of Gaussians. For the BDP, Petrone & Wasserman (2002) proved its posterior consistency and Ghosal (2001) examined its convergence rate for density estimation when direct observations of $\boldsymbol{\theta}$ are available. In our case, the observations are not directly on $\boldsymbol{\theta}$, but on discrete binomial data. Then both the number of trials N_i for each θ_i, $i = 1, \ldots, n$, and the number of observations n would affect posterior estimation. Let p_0 be the common underlying fixed density function of the coordinates of $\boldsymbol{\theta}$. We consider two consistency results here: the posterior consistency for estimating the probabilities of binomial data

$$P_0(y \mid N) = \int_0^1 \theta^y(1-\theta)^{N-y}\frac{N!}{y!(N-y)!}\cdot p_0(\theta)\mathrm{d}\theta, \quad y = 0, \ldots, N,$$

with fixed $N = \max\{N_i, i = 1, \ldots, n\}$ as $n \to \infty$, and the posterior consistency for estimating the underlying density p_0 as both n and the N_i's go to infinity. Using the BDP as the prior for p_0 implies that the prior of $P_0(y|N)$ is of the form $\psi_{k,\mathbf{W}_k}(y|N)$ where $k \sim \eta$ and given k, $\mathbf{W}_k$ follows a Dirichlet distribution with parameters $(M\alpha_{1,k}, \ldots, M\alpha_{k,k})$. Then the posterior distribution of $P_0(y|N)$ is of the same form, with the distribution of k and $\mathbf{W}_k$ being modified by the data. With the fixed maximum value of the N_i's, the weak consistency of the posterior of $P_0(y|N)$ is a direct corollary of the Schwartz Theorem (Schwartz, 1965; the definition of "weak consistency" for discrete random variables can be found in, for example, Ghosal, Ghosh, & Ramamoorthi, 1999a and Barron, Schervish, & Wasserman, 1999). We summarize the result in the following corollary.

Corollary 1. *Suppose that (i) p_0 is continuous on interval (0,1); (ii) the hyper-parameter F_0 has a continuous density on* $[0, 1]$ *with a positive measure on any nonempty open interval in* $[0, 1]$*; (iii) the number of trials N_i follows a discrete distribution ϕ, which is independent of the θ_i, $i = 1, \ldots, n$; (iv) $\phi(N) > 0$ and $\phi(l) = 0$ for every $l > N$; and (v) $\eta(k) > 0$ for $k > 0$. Then the posterior of $\psi_{k,\mathbf{W}_k}(\cdot|l)$ is weakly consistent at $P_0(\cdot|l)$ for any $l \in \{1, \ldots, N\}$.*

Corollary 1 is easily proven using Theorem 2 of Petrone (1999b) and the Schwartz Theorem (Schwartz, 1965). Corollary 1 indicates that for fixed $N = \max\{N_i, i = 1, \ldots, n\}$ as $n \to \infty$, we can always have weakly consistent estimates of $P_0(y|N)$. Thus, we have weakly consistent estimates of the first N moments of the density p_0. In theory, the underlying p_0 is not identifiable with finite N, since the data only contain information about N moments of p_0, and there are infinitely many Bernstein densities having the same N moments as p_0. However, through simulations (more details in Section 5), we found that the posterior given binomial data can be very close to that given $\boldsymbol{\theta}$ under three mild conditions: (i) p_0 is continuously differentiable with bounded second-order derivatives, (ii) the N_i's are moderately large (above 30), and (iii) the prior η assigns small probabilities to very large values of k. This is possibly because, first, under Condition (i) a Bernstein density of a moderately larger order k can approximate p_0 very well (Ghosal, 2001); second, since each y_i/N_i converges to the corresponding θ_i very fast as the N_i increases, the binomial data with moderately large N_i's would carry similar information as that of $\boldsymbol{\theta}$; and third, Condition (iii) leads to small posterior probabilities of Bernstein densities with very large orders, which can possibly have similar N moments as p_0 but distinct functional curves.

It is of greater importance and interest to investigate the posterior consistency of the BDP for p_0 given indirect binomial data. As far as we know, existing results of posterior consistency under nonparametric settings are established for the case where direct observations from the target distribution are available. In the following, notation $A \asymp D$ denotes that A and D are asymptotically of the same order, and $A = O(D)$ denotes that A is asymptotically of order smaller than or equal to that of D. We will show the posterior consistency of the BDP with different convergence rates of the N_i's and n going to infinity: (1) the N_i's are much larger than n; more specifically, we assume $n = O(N^v)$ for some $0 < v < 1$; and (2) the N_i's are a fractional order of n, denoted as $N \asymp n^u$ for some $0 < u < 1$. To simplify the mathematical derivations, we assume that $N_i = N$, $i = 1, \ldots, n$, and that $N \to \infty$.

Definition 2. *The Kullback–Leibler (KL) distance of two distributions with densities p_1 and p_2 on [0, 1] is defined by $K(p_1, p_2) = \int_0^1 p_1(x) \log(p_1(x)/p_2(x))\mathrm{d}x$.*

We say the posterior is weakly consistent at p_0 if for any $\epsilon > 0$ and every ϵ-Kullback–Leibler neighborhood U_ϵ of p_0, the posterior probability of U_ϵ converges to 1 in probability.

Theorem 1. *Suppose (i) p_0 is continuous, first-order differentiable on [0,1], and $\max_{\theta\in[0,1]}\{p_0(\theta), p_0'(\theta)\} \le C$ for some constant $C > 0$; (ii) the hyper-parameter F_0 has a continuous density with a positive measure on any nonempty open interval in* $[0, 1]$*; (iii) $N_i = N$,*

$i = 1, \ldots, n$; (iv) there exists a constant v_1 such that $n = O(N^{v_1})$ and $1/6 > v_1 > 0$; (v) $\eta(k) > 0$ for $k > 0$ and there exists $k_n \to \infty$ such that $k_n \leq N^{v_2}$ for some positive constant $v_2 > 0$ and $2v_2 + 3v_1 < 1/2$, and such that $\sum_{k \geq k_n} \eta(k) \leq \exp\{-r_2 \cdot N^{r_1}\}$ for some $r_1 > v_1$ and $r_2 > 0$. Then the posterior of the BDP is weakly consistent at p_0.

Condition (iv) in Theorem 1 requires that the numbers of trials N_i's go to infinity at a much faster rate than n, so that the posterior based on binomial data is close to that directly based on $\boldsymbol{\theta}$. Condition (v) ensures that the posterior probability of Bernstein densities whose orders are larger than k_n converges to zero in probability. The proof of Theorem 1 is provided in the supplementary file.

When N is much smaller than n, the posterior of the BDP behaves distinctly from that with n being much larger than N. As mentioned in the previous discussion of Corollary 1, there can possibly be an identification problem of p_0 with N being much smaller than n. To address this issue, the key is to limit the set of density functions under consideration. Thus, we impose a stronger condition on the prior η as described below. Nevertheless, we conjecture that the posterior consistency based on the KL neighborhood does not exist for binomial data with N being much smaller than n. In this case, we prove the weak consistency of the BDP using a different metric. The L_2 distance is a standard metric, which is usually used when the set of densities is uniformly bounded. If the densities are uniformly bounded and uniformly bounded away from zero, the L_2 distance is equivalent to the KL distance (Ghosal, Ghosh, & van der Vaart, 2000). For mathematical convenience, we assume that p_0 is bounded away from both zero and infinity, and use the L_2 neighborhood to show posterior consistency in the following theorem.

Theorem 2. *Suppose (i) p_0 is continuous with finite second-order derivatives and bounded away from both zero and infinity on $[0, 1]$, (ii) the hyper-parameter F_0 has a continuous and strictly positive density on $[0, 1]$, (iii) $N_i = N$, $i = 1, \ldots, n$, (iv) there exists a constant $u \in (0, 1/8)$ such that $N \asymp n^u$ as $n \to \infty$, (v) $\eta(k) > 0$ for all the positive integers k and there exists a constant $\varsigma > (1 - u)/u$ such that $\eta(k) \asymp \exp\{-k^\varsigma\}$. Then, for any $\epsilon > 0$ the posterior probability of the ϵ-L_2 neighborhood V_ϵ of p_0 converges to one in probability as $n \to \infty$.*

Conditions (i) and (iv) in Theorem 2 are to ensure enough observations for every binomial value y for $0 \leq y \leq N$, so that there is a uniform convergence of the posterior of binomial probability $P_0(y|N)$ for every y. Under Condition (ii), no region in $[0, 1]$ is priorly weighed too tiny compared to other regions. Condition (v) is used to constrain the posterior probability of the Bernstein densities corresponding to large orders k, which can possibly have similar N moments as p_0 but distinct functional curves. The detailed proof of the theorem is given in the supplementary file. In future research, it will be of interest to investigate the posterior consistency of p_0 with random numbers of trials.

The number of beta components k can be viewed as a kernel-smoothing parameter, and $1/k$ is comparable to a smoothing bandwidth. We have observed that if k is fixed at a much larger value than the real one, the posterior estimate is bumpy and sensitive to the data variation. On the other hand, if k is fixed at a small value, the posterior estimate can be overly smooth. The two theorems previously described show that an appropriate prior on k should be adapted to the given numbers of trials. In practice, we assign a truncated uniform prior to k, that is, $\eta(k) = 1/K$ for $k \leq K$ and $\eta(k) = 0$ for $k > K$ for some fixed constant K, and let the data choose the appropriate smoothness parameter (see simulation studies in Section 5). Petrone & Wasserman (2002) studied the asymptotic behavior of the posterior of the BDP with a truncated prior on k when direct observations are available. Empirically, we found that when the N_i's are large enough (above 30), the posterior of p_0 based on $(\mathbf{y}_n, \mathbf{N}_n)$ is close to that based on direct observations $\boldsymbol{\theta}$. We can follow Petrone and Wasserman's strategy for choosing an appropriate K for large N_i's: If the posterior draws of k concentrate around the boundary $k = K$, we increase the value of K. On

the other hand, for small N_i's, from Corollary 1 we know that to have consistent estimates of $P_0(y|N)$, $y = 0, \ldots, N$, we need at least $K > N$.

4. POSTERIOR SIMULATION VIA MARKOV CHAIN MONTE CARLO

To facilitate simulation of the posterior distribution resulting from the BDP, Petrone (1999a, b) introduces an auxiliary random vector $\mathbf{Z}_n = (Z_1, \ldots, Z_n)$, which is composed of n i.i.d. samples from G for $G \sim \mathcal{D}(MF_0)$. Given $\mathbf{Z}_n$ and k, the θ_i's are mutually independent, and $[\theta_i \mid k, Z_i] = \text{Beta}(j, k+1-j)$ if $Z_i \in (\frac{j-1}{k}, \frac{j}{k}]$, for $j = 1, \ldots, k$. The $\mathbf{Z}_n$ serves as indicators for which beta components the $\boldsymbol{\theta}$ are drawn from. Conditional on $(\mathbf{Z}_n, k)$, $\mathbf{W}_k$ is independent of $(\mathbf{y}_n, \mathbf{N}_n)$ and $[\mathbf{W}_k \mid k, \mathbf{Z}_n] = Dir(M\alpha_{1,k} + n_{1,k}, \ldots, M\alpha_{k,k} + n_{k,k})$, where $n_{j,k} = \#\{Z_i \in ((j-1)/k, j/k], i = 1, \ldots, n\}$, and $Dir(\xi_1, \ldots, \xi_k)$ denotes the Dirichlet distribution with parameters $\xi_1, \ldots, \xi_k$.

For ease of presentation, define a vector of labels of $\boldsymbol{\theta}$: $\mathbf{J}_n^k = (j_{1,k}, \ldots, j_{n,k})$, where $j_{i,k} = \sum_{j=1}^{k} j\delta_{(\frac{j-1}{k}, \frac{j}{k}]}(Z_i)$, $i = 1, \ldots, n$, indicates which interval Z_i falls into. After introducing $\mathbf{Z}_n$, $\mathbf{W}_k$ can be easily integrated out in the posterior. Then the posterior $f(\mathbf{Z}_n, k|\mathbf{y}_n, \mathbf{N}_n)$ is proportional to

$$\eta(k) \prod_{i=1}^{n} \frac{Q(k+1, j_{i,k})}{Q(k+1+N_i, j_{i,k}+y_i)} \frac{M^r}{M^{[n]}} \prod_{i=1}^{r} (n(Z_i') - 1)! f_0(Z_i'),$$

where the function $Q(k+1, j)$ is defined in (5), $Z_1' < \cdots < Z_r'$ are all the distinct values of $(Z_1, \ldots, Z_n)$, $n(Z_i')$ is the number of times Z_i' occurs, and $M^{[n]} = M(M+1)\cdots(M+n-1)$, where by definition $M^{[0]} = 1$.

Our MCMC algorithm is focused on simulating the posterior $f(\mathbf{Z}_n, k|\mathbf{y}_n, \mathbf{N}_n)$. With posterior samples $(\mathbf{Z}_n^{(s)}, k^{(s)})$ for $s = 1, \ldots, S$, since $E(W_{j,k^{(s)}}^{(s)} \mid \mathbf{Z}_n^{(s)}, k^{(s)}) = (M\alpha_{j,k^{(s)}} + n_{j,k^{(s)}})/(M+n)$, we can have a Rao–Blackwellisation version of the Bernstein density estimate (6) as

$$\tilde{b}(\theta) = \frac{1}{S} \sum_{s=1}^{S} \left[\sum_{j=1}^{k^{(s)}} \frac{M\alpha_{j,k^{(s)}} + n_{j,k^{(s)}}}{M+n} \beta(\theta; j, k^{(s)} + 1 - j) \right], \tag{7}$$

where $n_{j,k^{(s)}} = \sharp\{Z_i^{(s)} : Z_i^{(s)} \in ((j-1)/k^{(s)}, j/k^{(s)}]\}$. It is easy to show that (7) always has a smaller variance than (6).

4.1. Procedure I

We use the Gibbs sampler to simulate $(\mathbf{Z}_n, k)$. More specifically, we start with values $(k^{(0)}, \mathbf{Z}_n^{(0)})$ that have a nonzero posterior density, and we iteratively update the vector $(k, \mathbf{Z}_n)$ according to the following steps:

(I.a) Draw the number of beta components k from

$$[k|\mathbf{Z}_n, \mathbf{y}_n, \mathbf{N}_n] \propto \eta(k) \prod_{i=1}^{n} \frac{Q(k+1, j_{i,k})}{Q(k+1+N_i, j_{i,k}+y_i)}.$$

(I.b) For each i, conditional on k and $\mathbf{Z}_{[-i]}$, where the $\mathbf{Z}_{[-i]}$ denotes the vector $(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$, draw Z_i from the following distribution:

$$[Z_i \mid k, \mathbf{Z}_{[-i]}, \mathbf{y}_n, \mathbf{N}_n] = p_{i,0} \cdot f_k(Z_i) + \sum_{v \neq i, v=1}^{n} p_{i,v} \cdot \delta_{Z_v},$$

where the $p_{i,u}$ for $u = 0, \ldots, n$ and $u \neq i$, are probabilities, summing to one, and

$$p_{i,0} \propto \sum_{j=1}^{k} M\alpha_{j,k} \frac{Q(k+1,\, j)}{Q(k+1+N_i,\, j+y_i)},\ p_{i,v} \propto \frac{Q(k+1,\, j_{v,k})}{Q(k+1+N_i,\, j_{v,k}+y_i)}$$

for $v \geq 1$ and $v \neq i$. The $f_k(Z_i)$ is a density function, which is proportional to $f_0(Z_i)Q(k+1,\, j_{i,k})/Q(k+1+N_i,\, j_{i,k}+y_i)$.

Petrone & Veronese (2002) proposed to simulate the posterior $f(\boldsymbol{\theta}, \mathbf{Z}_n, k|\mathbf{y}_n, \mathbf{N}_n)$ instead of integrating $\boldsymbol{\theta}$ out, which is slightly less efficient than the method described above.

4.2. Procedure II on Marginal Distribution

Due to the clustering effect of the DP, direct simulation of $\mathbf{Z}_n$, as in Procedure I, tends to be sticky. One popular approach to address this issue is to integrate out the values of $\mathbf{Z}_n$ and simulate from the resulting marginal distribution of the cluster indicators $\mathbf{I}_n = (I_1, \ldots, I_n)$ of $\mathbf{Z}_n$ (see discussion in Escobar, 1994; MacEachern, 1994; MacEachern, Clyde, & Liu, 1999; Jain & Neal, 2004). The I_i takes an integer value such that if $Z_i = Z_{i'}$ then $I_i = I_{i'}$; otherwise $I_i \neq I_{i'}$ for $i, i' = 1, \ldots, n$. Let r be the number of distinct clusters among $\mathbf{I}_n$, let Δ_c be the set of I_i in the cth cluster with $n_c = |\Delta_c|$, and let $\varphi_{j,c} = P(j_{i,k} = j \text{ for } I_i \in \Delta_c \mid \mathbf{I}_n, k, \mathbf{y}_n, \mathbf{N}_n)$ for $c = 1, \ldots, r$. It is easy to see that

$$\varphi_{j,c} \propto \alpha_{j,k} \prod_{I_i \in \Delta_c} \frac{Q(k+1,\, j)}{Q(k+1+N_i,\, j+y_i)}. \tag{8}$$

Then by marginalizing out $j_{i,k}$, $i = 1, \ldots, n$, conditional on $\mathbf{I}_n$ in $f(\mathbf{Z}_n, k|\mathbf{y}_n, \mathbf{N}_n)$, the joint posterior distribution of $\mathbf{I}_n$ and k is given by

$$f(\mathbf{I}_n, k \mid \mathbf{y}_n, \mathbf{N}_n) \propto \eta(k) \frac{M^r}{M^{[n]}} \prod_{c=1}^{r} \left\{ (n_c - 1)! \Upsilon_{\Delta_c}^k \right\}, \tag{9}$$

where $\Upsilon_{\Delta_c}^k = \sum_{j=1}^{k} \alpha_{j,k} \prod_{I_i \in \Delta_c} Q(k+1,\, j)/Q(k+1+N_i,\, j+y_i)$.

Procedure II is to use a Gibbs sampler to simulate from the posterior $f(\mathbf{I}_n, k|\mathbf{y}_n, \mathbf{N}_n)$. In theory, both procedures converge at a geometric rate and Procedure II converges faster than Procedure I, given the same number of iterations (see the Appendix for theorems on convergence rates of the two procedures). However, in practice, we still recommend Procedure I for two reasons. First, each iteration in Procedure II takes much longer than that in Procedure I. For example, for data generated from Beta(8, 8) with $N_i = 100$ for $i = 1, \ldots, n$ and $n = 100$, the CPU computational time of Procedure II is 50% more than that of Procedure I given the same number of iterations, and the difference is even more pronounced if the elements of $\boldsymbol{\theta}$ are generated from a mixture of beta distributions or a Bernstein density with a large order k. The extra computational time for Procedure II results from time-consuming calculation of conditional probabilities of the I_i's, each of which takes k times more computational time than that of $\mathbf{Z}_n$. Second, we observed that the two procedures in our BDP-based setting lead to almost the same density estimates. We believe this is because calculation of the density estimate only requires us to know which interval each Z_i lies in instead of its accurate value. We found that given $\mathbf{I}_n$, for a cluster containing a moderate number (more than 15) of elements I_i, the probability $\varphi_{j,c}$ tends to concentrate on just one j; that is, the value $\varphi_{j,c}$ for some single j is close to one. Thus, given the same k, the distribution of $\mathbf{W}_k$ conditional on $\mathbf{Z}_n$ is almost the same as that conditional on $\mathbf{I}_n$. Jain & Neal (2004) proposed a split-merge Markov chain sampling algorithm, which can speed up the convergence rate of MCMC

samples of $\mathbf{I}_n$. Still, since their algorithm can not get around the time-consuming calculation of $\Upsilon^k_{\Delta_c}$, it requires much more computational time than Procedure I for each iteration.

4.3. Inferring the Hyper-parameter M from the Data

In the context above, M is treated as a given constant. However, as with the DP, the parameter M can be important in weighing the prior belief versus the data (West, 1992; Escobar & West, 1995; Liu, 1996), and this can influence the inference results significantly.

Suppose $\mathbf{Z}_n$ follows the DP $\mathcal{D}(MF_0)$: $Z_i \overset{i.i.d}{\sim} G$ and $G \sim \mathcal{D}(MF_0)$. It has been shown by Korwar & Hollander (1973) and Antoniak (1974) that the expected number of distinct values of $\mathbf{Z}_n$ is approximately $M \log((M+n)/M)$ for a large n. In our nonparametric hierarchical Bayes setting, where the BDP is used as the prior for F, M regulates the number of distinct beta distributions the $\boldsymbol{\theta}$ are drawn from. A small M around 0 signifies a strong belief that the data are generated from a single beta distribution. This is because for a fixed n, the posterior probability of the number of beta components larger than 1 converges to zero as M goes to zero (Petrone, 1999a, b). We find that if both the sample size n and the trial size N_i are small, fixing M at a small value can lead to a rather undesirable result.

The empirical Bayes approach proposed by Liu (1996) infers M by its MLE, and proceeds as though M is known. We here assign a uniform prior distribution to M and infer M jointly with the other parameters. To facilitate MCMC simulations, we discretize the range of M as $\{0.1, 0.2, \ldots, 9.9, 10\}$. We chose 10 as the upper bound, which is not far from the values $M = 1$ and $M = 2$ suggested by Petrone (1999a, b), so as not to overweigh F_0 in the posterior. We add the following step of simulating M to Procedure I in Section 4.1.

(I.c) full conditional distribution of M given $\mathbf{Z}_n$, $\mathbf{y}_n$ and $\mathbf{N}_n$ is proportional to $M^r/(M(M+1)\cdots(M+n-1))$, where r is the number of clusters of $\mathbf{Z}_n$.

5. SIMULATION STUDIES

We analyze four simulated examples modified from those used in Petrone & Wasserman (2002). We compare the Bernstein density estimate with the density estimate obtained under a DP-based nonparametric Bayes setting, which is referred to as the Dirichlet density estimate in the following. Readers are referred to Kong, Liu, & Wong (1994), Liu (1996), and MacEachern, Clyde, & Liu (1999) for the details of calculating the Dirichlet density estimate.

In addition to the density estimate, point estimates of $\boldsymbol{\theta}$, that is, the estimates of posterior means of $\boldsymbol{\theta}$ based on posterior samples, are also examined in comparison with two well-established point estimators: Stein's estimator (James & Stein, 1961) and Efron & Morris's (1971, 1972, 1975) limited-translation estimator. We denote Stein's estimator by $\hat{\theta}^1$, the limited-translation estimator by $\hat{\theta}^e$ for $e = 0.9$ and 0.8, the MLE by $\hat{\theta}^0$, and the posterior estimates based on the DP and the BDP by $\hat{\theta}^D$ and $\hat{\theta}^B$, respectively. For simplicity, we call $\hat{\theta}^D$ and $\hat{\theta}^B$ the Dirichlet estimator and the Bernstein estimator, respectively. We use the average of squared error ASE $= \sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2/n$ as the criterion to compare the estimators' performance.

We chose the uniform distribution on unit interval [0, 1] to be the hyper-parameter F_0 in the BDP $BD(\eta, MF_0)$ and the DP $\mathcal{D}(MF_0)$. We also chose the uniform prior η on $\{1, 2, \ldots, 200\}$ for k and the uniform prior on $\{0.1, 0.2, \ldots, 9.9, 10\}$ for M. The posterior estimates are based on 20,000 iterations of the Gibbs sampler after 5,000 burn-in time.

For each simulated set of data below, we first generated n i.i.d. θ_i from the prespecified distribution F. Then for each θ_i, we drew y_i from the binomial distribution Binom(N_i, θ_i) given each preassigned number of trials N_i, $i = 1, \ldots, n$.

5.1. Unimodal Beta Density

In this simple example, we let $F = \text{Beta}(8, 8)$, which has a Bernstein density with order $k = 15$ and $W_{7,15} = 1$. The sample size n of the simulated binomial data was set at 100, and the N_i's are integers sampled uniformly between 100 and 200. The Bernstein density estimate based on this data set approximates the underlying truth well, as illustrated in Figure 1(b). As shown in Figure 1(a), the posterior distribution of k has a mode around the true value $k = 15$, but also has a long tail. This is possibly because $\beta(; 8, 8)$ can also be represented as a Bernstein density of order $k > 15$. The posterior distribution of M has its mode at the smallest possible value 0.1, consistent with the underlying single-beta density (graph not shown). Figure 1(c) illustrates the Dirichlet density estimate for the same data when $\mathcal{D}(\text{Beta}(1, 1))$ is used as the prior for F. Because of the discrete property of the DP, the Dirichlet estimate is very bumpy.

The ASE of the five estimators $\hat{\theta}^1$, $\hat{\theta}^{0.9}$, $\hat{\theta}^{0.8}$, $\hat{\theta}^D$ and $\hat{\theta}^B$ are given in the third row of Table 1, indicating very similar performances for all five estimators.

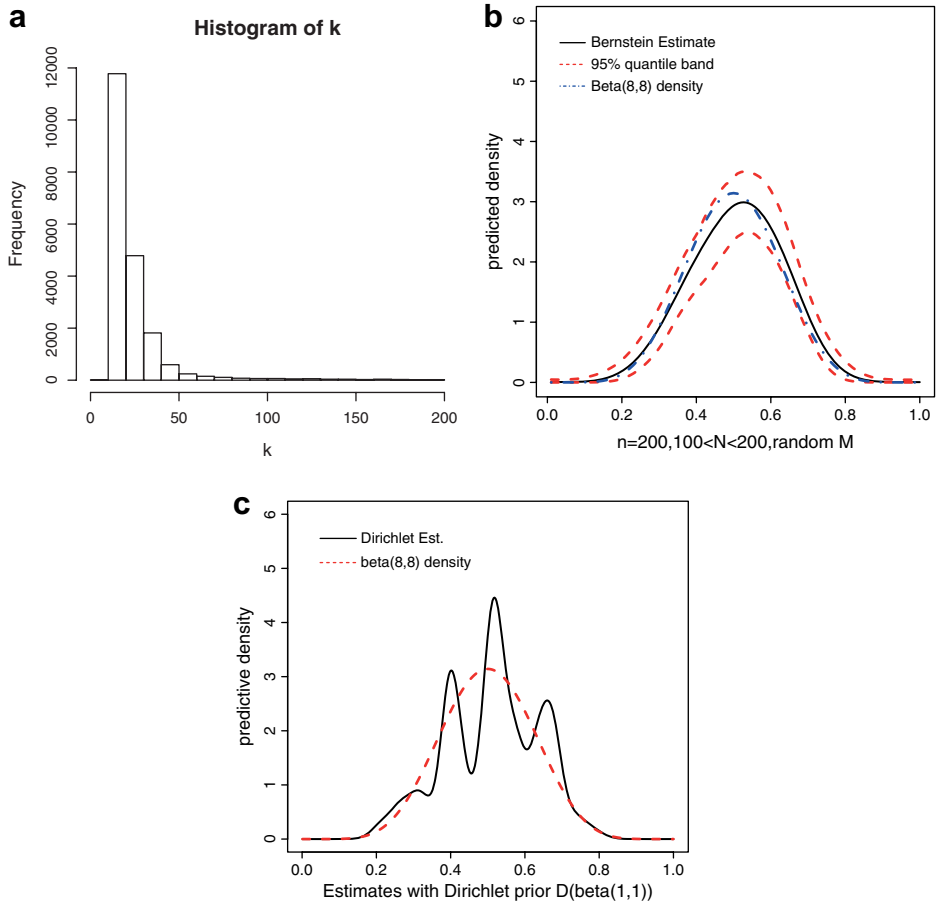

FIGURE 1: **Estimation of binomial data from** Beta**(8,8).** (a) The histogram of 20,000 posterior draws of k. (b) The Bernstein density estimate with its 95% posterior credibility band. (c) The Dirichlet density estimate. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com]

TABLE 1: ASE of five estimators of $\boldsymbol{\theta}$.

Distribution	n	N_i	ASE MLE	ASE $\hat{\theta}^1$	ASE $\hat{\theta}^{0.9}$	ASE $\hat{\theta}^{0.8}$	ASE $\hat{\theta}^D$	ASE $\hat{\theta}^B$
$\beta(;8,8)$	200	[100, 200]	1.68e-3	1.53e-3	1.56e-3	1.57e-3	1.60e-3	1.53e-3
$0.5[\beta(;60,10)+\beta(;10,60)]$	200	100	1.14e-3	1.17e-3	1.17e-3	1.18e-3	7.21e-4	6.80e-4
$Exp(8)$	200	100	7.29e-4	7.75e-4	7.40e-4	7.33e-4	7.56e-4	7.20e-4
$0.5[U(0.25,0.5)+U(0.75,1)]$	100	100	1.42e-3	1.45e-3	1.46e-3	1.46e-3	1.06e-3	1.04e-3

5.2. A Multimodal Distribution

We generated a binomial data set with $n = 200$ and $N_i = 100$ for $i = 1, \ldots, n$ from a mixture distribution $F = (1/2)\text{Beta}(60, 10) + (1/2)\text{Beta}(10, 60)$. Figure 2(a) shows that the posterior distribution of k has a mode around the true value $k = 69$. As illustrated in Figures 2(b) and 2(c), the Bernstein and the Dirichlet density estimates of F are close to its true density. The ASE of the five point estimators are summarized in the fourth row of Table 1. Stein's estimator and the limited-translation estimator, which tend to shrink toward the global mean, performed even worse than the MLE, while the $\hat{\theta}^D$ and $\hat{\theta}^B$ are better adapted to the bimodal feature of the data. We see that $\hat{\theta}^B$ has the smallest ASE, and $\hat{\theta}^D$ gives a similar result.

5.3. Non-Bernstein Densities

In this example, we let F be Exp(8) truncated at 1. Obviously, F does not have a Bernstein density, and is strongly asymmetric. The simulated binomial data is of size $n = 200$ and $N_i = 100$ for $i = 1, \ldots, n$. The Bernstein density estimate shown in Figure 2(d) approximates the true density remarkably well and is much better than the Dirichlet density estimate (graph not shown), even though the true density is not a finite mixture of beta distributions. The ASE of the five estimators are summarized in Table 1. Again, $\hat{\theta}^B$ achieved the smallest ASE.

We also investigated the performance of the Bernstein density estimate when p_0 is discontinuous. We simulated an example where the θ_i's are generated from a mixture of uniform densities $F = 0.5$ Unif $(0.25, 0.5) + 0.5$ Unif $(0.75, 1)$. As shown in the last row in Table 1, the Bernstein estimator can beat all other methods in estimating $\boldsymbol{\theta}$, including the Dirichlet estimator. From simulations, we predict that as long as p_0 has only a small number of discontinuous points, the nonparametric hierarchical Bayes model with the BD prior can still perform well.

6. BATTING-AVERAGE PREDICTIONS FOR THE 2005 SEASON

Brown (2008) analyzed the batting record for each Major League Baseball player over the course of a single season (2005). He used the batting records from the earlier part of the season (i.e., the first 3 months) to estimate the batter's potential ability, θ_i (the probability of the success of hits), and to predict his batting-average performance for the remainder of the season. This data set consists of 567 baseball players. In contrast to previous simulated data sets, the N_i's in this data are highly heterogeneous, ranging from 11 to 338. Brown proposed a variance-stabilizing transformation more appropriate for this heteroscedastic data than the transformation used in

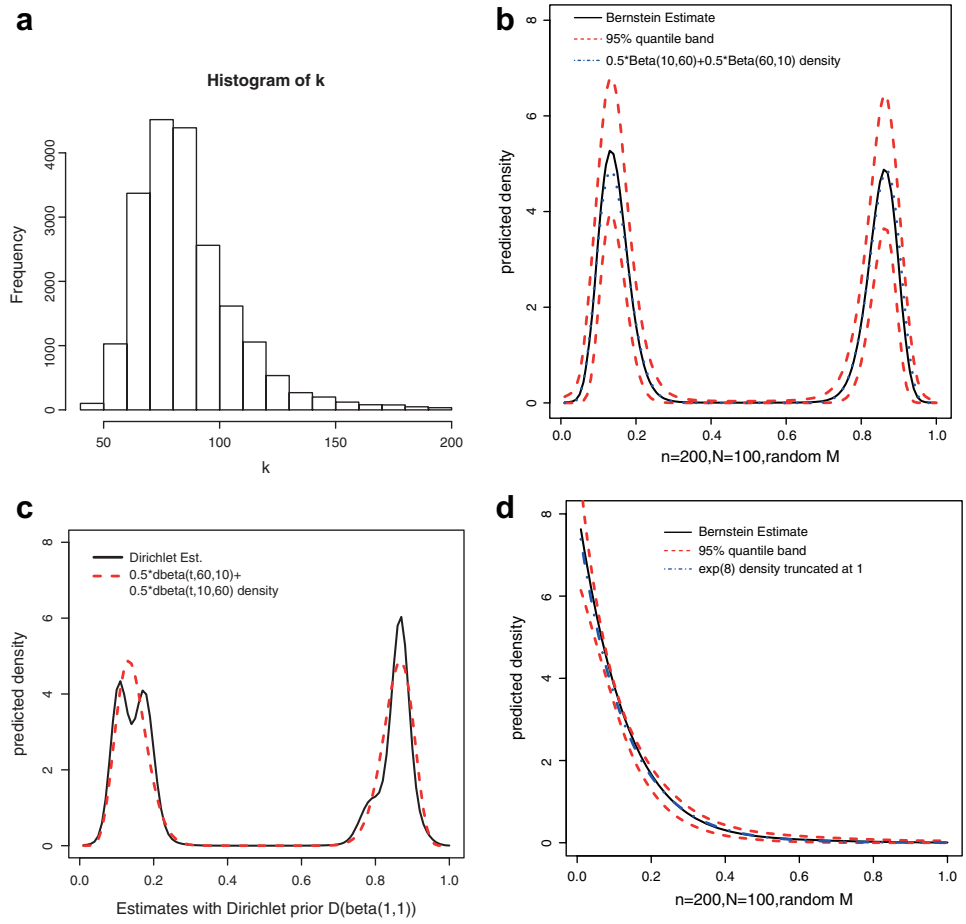

FIGURE 2: **Estimation of binomial data from** $0.5\text{Beta}(60, 10) + 0.5\text{Beta}(10, 60)$. (a) The histogram of 20,000 posterior draws of k. (b) The Bernstein density estimate with its 95% posterior credibility band for the binomial data with $n = 200$ and $N_i = 100$ for $i = 1, \ldots, n$. (c) The Dirichlet density estimate. (d) **Estimation of binomial data from Exp(8) truncated at 1.** The Bernstein density estimate with its 95% posterior credibility band. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com]

Efron & Morris (1975):

$$x_i = \arcsin \sqrt{\frac{y_i + 1/4}{N_i + 1/2}}, \ \mu_i = \arcsin \sqrt{\theta_i}. \tag{10}$$

Then we have approximately $x_i \sim N(\mu_i, \sigma_i^2)$, where $\sigma_i^2 = 1/(4N_i)$.

In addition to the heteroscedasticity, the baseball data also exhibit other special properties: (a) the distribution of the μ_i's cannot be effectively approximated by a normal distribution, as shown in Figure 3(a), or the density of the θ_i's possibly has two modes, as shown in Figure 3(c); and (b) there is a strong correlation between the x_i's and the N_i's. Brown used three criteria, that is, $\widehat{TSE}^*[\varrho]$, $\widehat{TSE}^*_\theta[\hat{\theta}]$ and $\widehat{TWSE}^*[\varrho]$, defined below, to compare and evaluate how the two properties affect the performance of different estimators. To be consistent, we will also

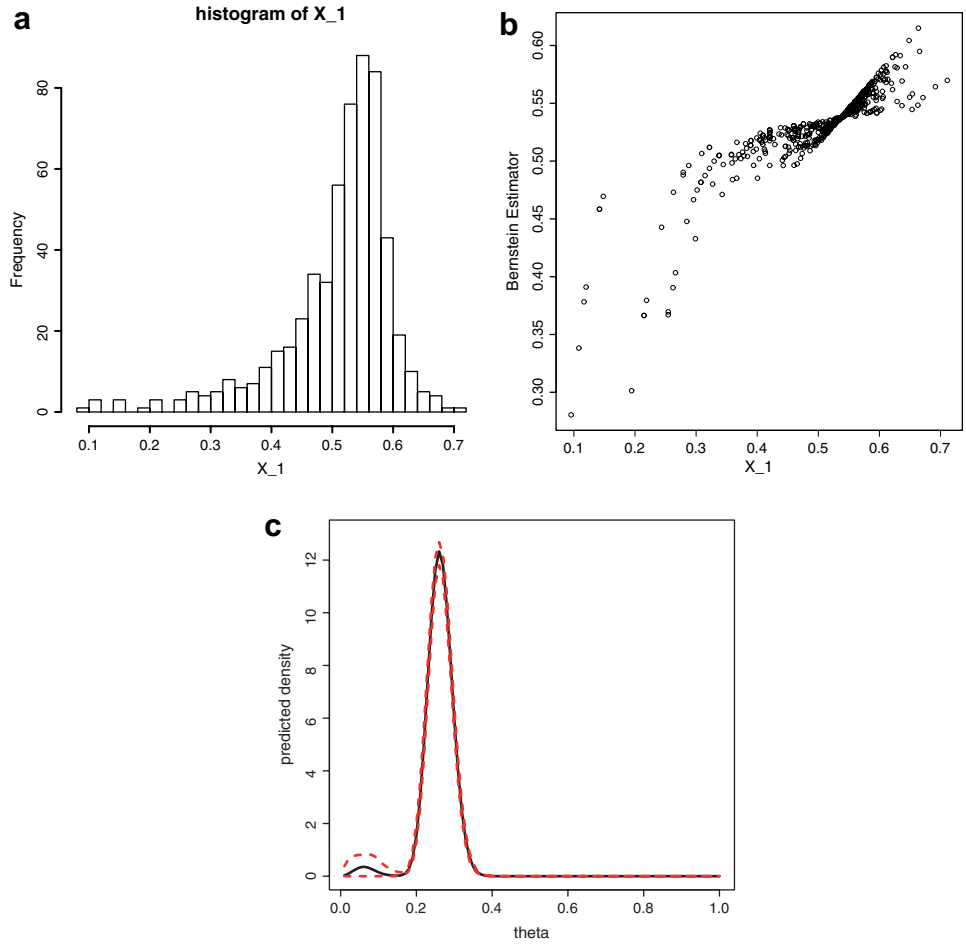

FIGURE 3: **The Batting Average Data of Brown (2008).** (a) Histogram of transformed binomial data. (b) Scatterplot of the Bernstein estimates of μ_i vs. x_i, $i = 1, \ldots, n$. (c) The Bernstein density estimate with its 95% posterior credibility band for the data. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com]

use these three criteria to evaluate our BDP-based estimator. The estimators studied by Brown include (1) the naive estimator, namely the MLE; (2) overall mean; (3) the parametric empirical Bayes method-of-moments estimator, or EB(MM); (4) the parametric empirical Bayes maximum likelihood estimator, or EB(ML); (5) the nonparametric empirical Bayes estimator, which adopts a generalized form of the kernel estimator, or NPEB; (6) the Harmonic Bayes estimator, which employs the harmornic prior in the hierarchical Bayes model; and (7) the James–Stein estimator.

Let $\{y_{ji}, N_{ji}\}$, $j = 1, 2, i = 1, \ldots, n$ denote the records for each half season. Assume that $y_{ji} \sim \text{Binom}(N_{ji}, \theta_i)$. Let ϱ_i be an estimator of μ_i, and let $\varrho^0(X) = X$ be the naive estimator. We derive the corresponding estimator of θ_i as $\hat{\theta}_i = \sin^2 \varrho_i(X)$ according to relationship (10). Since the nonparametric Bayesian approaches estimate θ_i directly, to make it comparable to Brown's estimators of μ_i we transform $\hat{\theta}^B$ to $\varrho^B = \arcsin \sqrt{\hat{\theta}^B}$, which is our Bernstein estimator of μ_i. Brown proposed the estimates of the total squared error of μ_i and θ_i as: $\widehat{\text{TSE}}[\varrho] = \sum_i (X_{2i} -$

TABLE 2: Estimation errors for half-season predictions.

	Naive	Groups' mean	EB(MM)	EB(ML)	NPEB	Harmonic prior	James-Stein	Bernstein	Dirichlet
$\widehat{\text{TSE}}^*[\varrho]$	1	0.852	0.593	0.902	0.508	0.884	0.525	0.663	0.697
$\widehat{\text{TSE}}^*_\theta[\hat{\theta}]$	1	0.887	0.606	0.925	0.509	0.905	0.540	0.683	0.725
$\widehat{\text{TWSE}}^*[\varrho]$	1	1.120	0.626	0.607	0.560	0.600	0.502	0.532	0.597

$\varrho_i)^2 - \sum_i 1/4N_{2i}$, and

$$\widehat{\text{TSE}}_\theta[\hat{\theta}] = \sum_i \left(\frac{y_{2i}}{N_{2i}} - \hat{\theta}_i\right)^2 - \sum_i \frac{1}{N_{2i}} \cdot \frac{y_{2i}}{N_{2i}} \left(1 - \frac{y_{2i}}{N_{2i}}\right).$$

Since all of the methods are compared to the naive estimator, a natural normalization is to divide by the estimated total squared error of the naive estimator. Thus, the normalized estimated squared errors are

$$\widehat{\text{TSE}}^*[\varrho] = \frac{\widehat{\text{TSE}}[\varrho]}{\widehat{\text{TSE}}[\varrho^0]} \qquad \text{and} \qquad \widehat{\text{TSE}}^*_\theta[\hat{\theta}] = \frac{\widehat{\text{TSE}}_\theta[\hat{\theta}]}{\widehat{\text{TSE}}_\theta[\hat{\theta}^0]}.$$

The results are summarized in Table 2. The uses of $\widehat{\text{TSE}}^*[\varrho]$ and $\widehat{\text{TSE}}^*_\theta[\hat{\theta}]$ lead to almost the same results. Our estimator did better than the group mean, EB(ML), and the Harmonic Bayes estimator. However, the Bernstein estimator did not perform the best, possibly because of the strong correlation between x_i and N_i, which violates the general assumption that the θ_i's are i.i.d. given the N_i's. The observed high correlation is due to the fact that better-performing players tend to have much higher numbers of at-bats. The R^2 between the $\{N_{1i}\}$ and $\{X_{1i}\}$ is 0.25. Consequently, the batting-average prediction for batters with small numbers of at-bats is shrunk toward the batting averages of batters with large numbers of at-bats, as shown in Figure 3(b). This explains the poorer performance of our estimator than EB(MM) and NPEB, and also the mediocre performance of two other likelihood-based methods, EB(ML) and the Harmonic Bayes estimator. Readers are referred to Brown (2008) for a discussion of the shrinkage effect of all the other estimators.

To concentrate on predicting the performance of the batters with the most at-bats, Brown has also proposed a weighted squared-error criterion defined as:

$$\widehat{\text{TWSE}}[\varrho] = \sum_i N_{1i}(X_{2i} - \varrho_i)^2 - \sum_i \frac{N_{1i}}{4N_{2i}} \qquad \text{and} \qquad \widehat{\text{TWSE}}^*[\varrho] = \frac{\widehat{\text{TWSE}}[\varrho]}{\widehat{\text{TWSE}}[\varrho_0]}.$$

As shown in the last row of Table 2, the use of $\widehat{\text{TWSE}}^*[\varrho]$ mitigates the effect of correlation, making our estimator the second-best performer, slightly inferior to the James–Stein estimator.

7. DISCUSSION

In this article, we analyzed binomial data by constructing a nonparametric hierarchical Bayes model in which the unobserved random probabilities of success are assigned with the BD prior (Petrone & Veronese, 2002). Under our problem setting, both the number of observations n and the numbers of trials N_i's affect final density estimation of the probabilities of success. We showed

the posterior consistency of estimating the probabilities $P_0(y|N)$ when the maximum value of the N_i's is fixed, and the posterior consistency of estimating p_0 when both the N_i's and n go to infinity. It will be of interest in future research to investigate how the asymptotic results turn out when the N_i's are random and follow a discrete probability. In addition, investigating posterior consistency given indirect observations under general nonparametric Bayesian settings could also be a focus of future research.

We compared the Bernstein density estimate with the Dirichlet density estimate of Liu (1996). The nonparametric BDP-based approach, which incorporates the continuity information of the hidden distribution F, is more robust to data variation and performs significantly better than the Dirichlet density estimate in both simulation studies and the baseball data analysis. In comparison to Stein's estimator, Efron and Morris's limited-translation estimator, and other empirical Bayes estimators, we find that the nonparametric Bayes estimators are more adaptive to the multimodal feature of the data and can achieve a significant gain in efficiency when the second-level hierarchical distribution is indeed multimodal.

ACKNOWLEDGEMENTS

This work was supported by NSF grants DMS-0706989 and DMS-1007762.

BIBLIOGRAPHY

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

Barron, A., Schervish, M. J., & Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27, 536–561.

Berry, D. A. & Christensen, R. (1979). Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *The Annals of Statistics*, 7, 558–568.

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

Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *The Annals of Applied Statistics*, 2, 113–152.

Diaconis, P. & Freedman, D. (1983). On inconsistent Bayes estimates in the discrete case. *The Annals of Statistics*, 11, 1109–1118.

Diaconis, P. & Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, 14, 1–26.

Efron, B. & Morris, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators—Part I: The Bayes case. *Journal of the American Statistical Association*, 66, 807–815.

Efron, B. & Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators—Part II: The empirical Bayes case. *Journal of the American Statistical Association*, 67, 130–139.

Efron, B. & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70, 311–319.

Escobar, M. D. (1988). *Estimating the means of several normal populations by estimating the distribution of the means*, PhD thesis, Yale University.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89, 268–277.

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

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

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2, 615–629.

Freedman, D. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case I. *The Annals of Mathematical Statistics*, 34, 1194–1216.

Freedman, D. (1965). On the asymptotic behavior of Bayes' estimates in the discrete case II. *The Annals of Mathematical Statistics*, 36, 454–456.

Gasparini, M. (1996). Bayesian density estimation via mixtures of Dirichlet processes. *Journal of Nonparametric Statistics*, 6, 355–366.

Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *The Annals of Statistics*, 29, 1264–1280.

Ghosal, S., Ghosh, J. K., & Ramamoorthi, R. V. (1999a). Consistency issues in Bayesian nonparametrics. In *Asymptotics, Nonparametrics, and Time Series: A tribute to Madan Lal Puri*, Ghosh, S., editor. Dekker, New York, pp. 639–668

Ghosal, S., Ghosh, J. K., & Ramamoorthi, R. V. (1999b). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27, 143–158.

Ghosal, S., Ghosh, J. K., & van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28, 500–531.

Jain, S. & Neal, R. M. (2004) A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13, 158–182.

James, W. & Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361–379.

Kong, A., Liu, J. S., & Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89, 278–288.

Korwar, R. M. & Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *The Annals of Probability*, 1, 705–711.

Liu, J. S. (1991) Correlation Structure and Convergence Rate of The Gibbs Sampler. *Ph.D. Thesis*, Department of Statistics, The University of Chicago.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89, 958–966.

Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics*, 24, 911–930.

Liu, J. S., Wong, W. H., & Kong, A. (1995). Covariance Structure and Convergence Rate of the Gibbs Sampler with Various Scans. *Journal of the Royal Statistical Society. Series B*, 57, 157–169.

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

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation*, 23, 727–741.

MacEachern, S. N., Clyde, M., & Liu, J. S. (1999). Sequential importance sampling for nonparametric Bayes models: The next generation. *The Canadian Journal of Statistics*, 27, 251–267.

MacEachern, S. N. & Müller, P. (2000). Efficient MCMC schemes for robust model extensions using encompassing Dirichlet process mixture models. In *Robust Bayesian Analysis*, Ruggeri, F. & Ríos-Insua, D., editors. Springer-Verlag, New York, pp. 295–316.

Müller, P. & Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science*, 19, 95–110.

Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials. *The Canadian Journal of Statistics*, 27, 105–126.

Petrone, S. (1999b). Random Bernstein polynomials. *Scandinavian Journal of Statistics*, 26, 373–393.

Petrone, S. & Veronese, P. (2002). Nonparametric mixture priors based on an exponential random scheme. *Statistical Methods and Applications*, 11, 1–20.

Petrone, S. & Wasserman, L. (2002). Consistency of Bernstein polynomial posteriors. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64, 79–100.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrschein lichkeitstheorie und verwandte Gebiete*, 4, 10–26.

West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. *ISDS Discussion Paper*. ♯92-A03. Duke University.

APPENDIX

Convergence rates of the MCMC samplers

Let $\mathbf{F}_I$ and $\mathbf{F}_{II}$, respectively denote the forward operator for the Gibbs sampler of MCMC Procedure I and II, that is,

$$\mathbf{F}_I : k \to Z_1 \to Z_2 \to \cdots \to Z_n; \quad \mathbf{F}_{II} : k \to I_1 \to I_2 \to \cdots \to I_n.$$

Denote $f_s(\mathbf{Z}_n, k|\mathbf{y}_n, \mathbf{N}_n)$ and $f_s(\mathbf{I}_n, k|\mathbf{y}_n, \mathbf{N}_n)$ the distributions of the sth step MCMC samples of $\mathbf{F}_I$ and $\mathbf{F}_{II}$, $s = 0, 1, \ldots,$ respectively. Let $\pi(k, \mathbf{Z}_n)$ and $\pi(k, \mathbf{I}_n)$ denote the equilibrium distribution.

Theorem 3. *Choose* $(\mathbf{Z}_n^{(0)}, k^{(0)})$ *to be in the support of* $f(\mathbf{Z}_n, k|\mathbf{y_n}, \mathbf{N}_n)$. *For a truncated* η, *the spectral radii of* $\mathbf{F}_I$ *and* $\mathbf{F}_{II}$ *are strictly less than 1, that is, the supremum modulus of the eigenvalues of* $\mathbf{F}_I$ *and* $\mathbf{F}_{II}$ *is smaller than 1. Moreover, the Pearson* χ^2 *distance from* f_s *to the stationary distribution* π *is monotone decreasing at a geometric rate as s increases. The correlations between* $t(\mathbf{Z}_n^{(s)}, k^{(s)})$) *and* $t(\mathbf{Z}_n^{(0)}, k^{(0)})$ *and between* $t(\mathbf{I}_n^{(s)}, k^{(s)})$ *and* $t(\mathbf{I}_n^{(0)}, k^{(0)})$ *converge to 0 at a geometric rate.*

Proof. The result of $\mathbf{F}_{II}$ is obvious, since there are only finite states of $(k, \mathbf{I}_n)$ for a truncated η, and the transition probabilities between any two states are positive. The geometric convergence rate of $\mathbf{F}_I$ can be shown by verifying conditions in Theorem 1 of Liu, Wong, & Kong (1995) and using Lemma 4.1.1. of Liu (1991). ■

Theorem 4. *The norms of the two forward operators are ordered as* $\mathbf{F}_{II} \leq \mathbf{F}_I$.

Proof. Theorem 4 is proven using Theorem 1 of Liu (1994). ■

Received 03 February 2011
Accepted 24 January 2012