The Nested Dirichlet Process [with Comments and Rejoinder]
Author(s): Abel Rodríguez, David B. Dunson, Alan E. Gelfand, Daniel L. Gillen, Wesley O. Johnson, Peter Müller, Luis Nieto-Barajas, Kaushik Ghosh, Pulak Ghosh, Ram C. Tiwari, Steven N. Maceachern and Lancelot F. James
Source: *Journal of the American Statistical Association*, Vol. 103, No. 483 (Sep., 2008), pp. 1131-1154
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: http://www.jstor.org/stable/27640149
Accessed: 22-04-2015 09:06 UTC

The Nested Dirichlet Process

Abel RODRÍGUEZ, David B. DUNSON, and Alan E. GELFAND

In multicenter studies, subjects in different centers may have different outcome distributions. This article is motivated by the problem of nonparametric modeling of these distributions, borrowing information across centers while also allowing centers to be clustered. Starting with a stick-breaking representation of the Dirichlet process (DP), we replace the random atoms with random probability measures drawn from a DP. This results in a nested DP prior, which can be placed on the collection of distributions for the different centers, with centers drawn from the same DP component automatically clustered together. Theoretical properties are discussed, and an efficient Markov chain Monte Carlo algorithm is developed for computation. The methods are illustrated using a simulation study and an application to quality of care in U.S. hospitals.

KEY WORDS: Clustering; Dependent Dirichlet process; Gibbs sampler; Hierarchical model; Nonparametric Bayes; Random probability measure.

1. INTRODUCTION

The Dirichlet process (DP) (Ferguson 1973, 1974) is the most widely used nonparametric model for random distributions in Bayesian statistics, due mainly to the availability of efficient computational techniques (Escobar and West 1995; Lo, Brunner, and Chan 1996; MacEachern and Müller 1998; Neal 2000; Jain and Neal 2000; Ishwaran and James 2001, 2003; Roberts and Papaspiliopoulos 2008; Blei and Jordan 2006). Because the DP puts probability 1 on the space of discrete measures, it typically is not used to model the data directly. Instead, it is more naturally used as a prior for a mixing distribution, resulting in a DP mixture (DPM) model (Lo 1984; Escobar 1994; Escobar and West 1995). Some recent applications of the DP include finance (Kacperczyk, Damien, and Walker 2003), econometrics (Chib and Hamilton 2002; Hirano 2002), epidemiology (Dunson 2005), genetics (Medvedovic and Sivaganesan 2002; Dunson, Herring, and Mulheri-Engel 2007a), medicine (Kottas, Branco, and Gelfand 2002; Bigelow and Dunson 2007), and auditing (Laws and O'Hagan 2002). Although most of these applications focus on problems with exchangeable samples from one unknown distribution, there is growing interest in extending the DP to accommodate *collections* of dependent distributions.

The dependent DP (DDP; MacEachern 1999, 2000) represents an important step in this direction. The DDP induces dependence in a collection of distributions by replacing the elements of the stick-breaking representation (McCloskey 1965; Sethuraman 1994) with stochastic processes. A version of this construction (where dependence is introduced only on the atoms) has been used by DeIorio, Müller, Rosner & MacEachern (2004) to create ANOVA-like models for densities, and by Gelfand, Kottas, and MacEachern (2005) to generate spatial processes that allow for nonnormality and nonstationarity. This latter class of models was extended by Duan, Guindani, and Gelfand (2007) to create generalized spatial DPs (GSDPs) that allow different surface selection at different locations.

Along these lines, another approach to introducing dependence is the hierarchical DP (HDP; Tomlinson 1999; Teh, Jordan, Beal, and Blei 2006). In this setting, multiple group-specific distributions are assumed to be drawn from a common DP whose baseline measure is in turn a draw from another DP. This allows the different distributions to share the same set of atoms but have distinct sets of weights. More recently, Griffin and Steel (2006) proposed an order-dependent DP, in which the weights are allowed to change with the covariates.

An alternative approach is to introduce dependence through linear combinations of realizations of independent DPs. For example, Müller, Quintana, and Rosner (2004), motivated by a similar problem as that of Teh et al. (2006), defined the distribution of each group as the mixture of two independent samples from a DP process: one component that is shared by all groups and one that is idiosyncratic. Dunson (2006) extended this idea to a time setting, and Dunson, Pillai, and Park (2007b) proposed a model for density regression using a kernel-weighted mixture of DPs defined at each value of the covariate.

Our work is motivated by two related problems: clustering probability distributions and simultaneous multilevel clustering in nested settings. As a motivating example, suppose that patient outcomes are measured within different medical centers. The distribution of patients within one specific center can be nonnormal, with mixture models providing a reasonable approximation. In assessing quality of care, it is of interest to cluster centers according to the distribution of patients outcomes, and to identify outlying centers. On the other hand, it is also interesting to simultaneously cluster patients within the centers by borrowing information across centers that present clusters with similar characteristics. This task is different from clustering patients within and across centers, which could be accomplished using the approaches discussed by Teh et al. (2006) and Müller et al. (2004). Moreover, our approach is different from the nested Chinese restaurant process proposed by Blei, Griffiths, Jordan, and Tenenbaum (2004) for the problem of characterizing topic hierarchies within documents. The Chinese restaurant process induces a flexible distribution on words through a tree structure in which the topic on one level is dependent on the distribution of topics at the previous levels. We propose a different type of nested Dirichlet process (NDP) for modeling a collection of dependent distributions using random variables as

Abel Rodriguez is Assistant Professor, Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA 95064 (E-mail: *abel@soe.ucsc.edu*). David B. Dunson is Senior Investigator, Biostatistics Branch, National Institute of Environmental Health Science, Research Triangle Park, NC 27709 (E-mail: *dunson1@niehs.nih.gov*). Alan E. Gelfand is James B. Duke Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708 (E-mail: *alan@isds.duke.edu*). This work is part of the first author's dissertation completed at Duke University and was supported in part by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences. The authors thank Shyamal Peddada, Ju Hyun Park, an associate editor, and two anonymous referees for helpful comments that greatly improved the quality of the article.

In the Public Domain
Journal of the American Statistical Association
September 2008, Vol. 103, No. 483, Theory and Methods
DOI 10.1198/016214508000000553

atoms at the higher level and random distributions as atoms at the lower level.

The article is organized as follows. We start in Section 2 with a short review of the DP. In Section 3 we motivate and define the NDP, explore its theoretical properties, and compare it with other DP extensions. We discuss efficient computational schemes for the NDP in Section 4, and present examples that illustrate the advantages of our methodology in Section 5. Finally, we close with a brief discussion in Section 6.

2. THE DIRICHLET PROCESS

Consider the probability spaces $(\boldsymbol{\Theta}, \mathcal{B}, P)$ and $(\mathbf{P}, \mathcal{C}, Q)$ such that $P \in \mathbf{P}$. Typically, $\boldsymbol{\Theta} \subset \mathbb{R}^d$, $\mathcal{B}$ corresponds to the Borel σ-algebra of subsets of $\mathbb{R}^d$, and $\mathbf{P}$ is the space of probability measures over $(\boldsymbol{\Theta}, \mathcal{B})$, but most of the results mentioned in this section extend to any complete and separable metric space $\boldsymbol{\Theta}$. We refer to $(\boldsymbol{\Theta}, \mathcal{B}, P)$ as the *base space* and to $(\mathbf{P}, \mathcal{C}, Q)$ as the *distributional space*. The DP with base measure H and precision α, denoted as $\mathsf{DP}(\alpha H)$, is a measure Q such that $(P(B_1), \ldots, P(B_k)) \sim \mathsf{Dir}(\alpha H(B_1), \ldots, \alpha H(B_k))$ for any finite and measurable partition $B_1, \ldots, B_k$ of $\boldsymbol{\Theta}$.

The DP can be alternatively characterized in terms of its predictive rule (Blackwell and MacQueen 1973). If $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{n-1})$ is an iid sample from $P \sim \mathsf{DP}(\alpha H)$, then we can integrate out the unknown P and obtain the conditional predictive distribution of a new observation,

$$\boldsymbol{\theta}_n | \boldsymbol{\theta}_{n-1}, \ldots, \boldsymbol{\theta}_1 \sim \frac{\alpha}{\alpha + n - 1} H + \sum_{l=1}^{n-1} \frac{1}{\alpha + n - 1} \delta_{\boldsymbol{\theta}_l},$$

where $\delta_{\boldsymbol{\theta}_l}$ is the Dirac probability measure concentrated at $\boldsymbol{\theta}_l$.

Exchangeability of the draws ensures that the full conditional distribution of any $\boldsymbol{\theta}_l$ has this same form. This result, which relates the DP to a Pólya urn, is the basis for the usual computational tools used to fit DP models (Escobar 1994; Escobar and West 1995; Bush and MacEachern 1996; MacEachern and Müller 1998).

The DP also can be regarded as a type of *stick-breaking prior* (McCloskey 1965; Perman, Pitman, and Yor 1992; Sethuraman 1994; Pitman 1996; Ishwaran and James 2001; Ongaro and Cattaneo 2004). A stick-breaking prior on the space $\mathbf{P}$ has the form

$$P^K(\cdot) = \sum_{k=1}^{K} w_k^* \delta_{\boldsymbol{\theta}_k^*}(\cdot), \qquad \boldsymbol{\theta}_k^* \sim H,$$

$$w_k^* = z_k \prod_{l=1}^{k-1} (1 - z_l), \qquad z_k \sim \begin{cases} \mathsf{beta}(a_k, b_k) & \text{if } k < K \\ \delta_1 & \text{if } k = K, \end{cases}$$

where the number of atoms K can be finite (either known or unknown) or infinite. For example, taking $K = \infty$, $a_k = 1 - a$, and $b_k = b + ka$ for $0 \le a < 1$ and $b > -a$ yields the two-parameter Poisson–Dirichlet process, also known as the Pitman–Yor process (Ishwaran and James 2001), with the choices $a = 0$ and $b = \alpha$, resulting in the DP (Sethuraman 1994).

The stick-breaking representation is probably the most versatile definition of the DP. It has been exploited to generate efficient alternative samplers like the blocked Gibbs sampler (Ishwaran and James 2001), which relies on a finite-sum approximation, and the collapsed Gibbs sampler (Ishwaran and James 2003) and the retrospective sampler (Roberts and Papaspiliopoulos 2008), both of which avoid truncations. It also is the starting point for the definition of many generalizations that allow dependence across a collection of distributions, including the DDP (MacEachern 2000), the πDDP (Griffin and Steel 2006), and the GSDP (Duan et al. 2007).

3. THE NESTED DIRICHLET PROCESS

3.1 Definition and Basic Properties

Suppose that y_{ij}, for $i = 1, \ldots, n_j$ are observations for different subjects within center j, for $j = 1, \ldots, J$. For example, $\mathbf{y}_j = (y_{1j}, \ldots, y_{n_j j})'$ may represent patient outcomes within the jth hospital or hospital-level outcomes within the jth state. Although covariates, $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ are typically available, we initially assume that subjects are exchangeable within centers, with $y_{ij} \overset{\text{iid}}{\sim} F_j$, for $j = 1, \ldots, J$.

In analyzing multicenter data, various customary strategies can be used, the most common being to (a) pool the data from the different centers, (b) analyze the data from the different centers separately, and (c) fit a parametric hierarchical model to borrow information. The first approach is too restrictive, because subjects in different centers may have different distributions, whereas the second approach is inefficient. The third approach parameterizes F_j in terms of the finite-dimensional parameter $\boldsymbol{\eta}_j$ and then borrows information by assuming $\boldsymbol{\eta}_j \overset{\text{iid}}{\sim} F_0$, with F_0 a known distribution (most commonly normal) with possibly unknown parameters (mean, variance). One can potentially cluster centers having similar random effects, $\boldsymbol{\eta}_j$, although clustering may be sensitive to F_0 (Verbeke and Lesaffre 1996). Assuming that F_0 has an arbitrary discrete distribution having k mass points provides more flexible clustering, but the model is still dependent on the choice of k and the specific parametric form for F_j.

Furthermore, clustering based on the random effects has the disadvantage of only borrowing information about aspects of the distribution captured by the parametric model. For example, clustering centers by mean patient outcomes ignores differences in the tails of the distributions. Our motivation is to borrow information and cluster across distributions $\{F_j, j = 1, \ldots, J\}$ nonparametrically to enhance flexibility, and we use a Dirichlet type of specification to enable clustering of random distributions.

Consider a collection of distributions $\{G_1, \ldots, G_J\}$ such that $G_j \sim Q$ with $Q \equiv \mathsf{DP}(\alpha \mathsf{DP}(\beta H))$ and let

$$F_j(\cdot | \boldsymbol{\phi}) = \int_{\boldsymbol{\Theta}} p(\cdot | \boldsymbol{\theta}, \boldsymbol{\phi}) G_j(d\boldsymbol{\theta}). \tag{1}$$

The collection $\{F_1, \ldots, F_J\}$ is said to follow an *NDP mixture*. The definition of the NDP implies that

$$G_j(\cdot) \sim Q \equiv \sum_{k=1}^{\infty} \pi_k^* \delta_{G_k^*(\cdot)} \tag{2}$$

and

$$G_k^*(\cdot) \equiv \sum_{l=1}^{\infty} w_{lk}^* \delta_{\boldsymbol{\theta}_{lk}^*}(\cdot), \tag{3}$$

with $\boldsymbol{\theta}_{lk}^* \sim H$, where H is a probability measure on $(\boldsymbol{\Theta}, \mathcal{B})$, $w_{lk}^* = u_{lk}^* \prod_{s=1}^{l-1} (1 - u_{sk}^*)$, $\pi_k^* = v_k^* \prod_{s=1}^{k-1} (1 - v_s^*)$, $v_k^* \sim \mathsf{beta}(1,$

α), and $u^*_{lk} \sim \text{beta}(1, \beta)$. In (1), $p(\cdot|\boldsymbol{\theta}, \boldsymbol{\phi})$ is a distribution parameterized by the finite-dimensional vectors $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. For example, in the case of a univariate response, if the collection $\{F_1, \ldots, F_J\}$ is assumed to be exchangeable, then an attractive choice would be $\boldsymbol{\theta} = (\mu, \sigma)$ and $p(\cdot|\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathsf{N}(\cdot|\mu, \sigma^2)$, which yields a class that is dense on the space of absolutely continuous distributions (Lo 1984). On the other hand, if a vector $\mathbf{x}$ of covariates is available, then we could opt for a random-effects model, where $\boldsymbol{\theta} = \mu$, $\boldsymbol{\phi} = (\boldsymbol{\gamma}, \sigma^2)$ and $p(\cdot|\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathsf{N}(\cdot|\mu + \mathbf{x}'\boldsymbol{\gamma}, \sigma^2)$, similar in spirit to the models of Mukhopadhyay and Gelfand (1997) and Kleinman and Ibrahim (1998). Extensions to multivariate or discrete outcomes are immediate.

Because a priori, $\mathbb{P}(G_j = G_{j'}|H) = \frac{1}{1+\alpha} > 0$, the model naturally induces clustering in the space of distributions. In addition, for any measurable set $A \in \mathcal{B}$,

$$\mathbb{E}(G_j(A)|H) = H(A) \qquad \text{and}$$

$$\mathbb{V}(G_j(A)|H) = \frac{H(A)(1 - H(A))}{\beta + 1}.$$

Our construction creates a collection of *dependent* random distributions. For a given set $A \in \mathcal{B}$, we have a collection of random variables $\{G_1(A), \ldots, G_J(A)\}$, and (as shown in App. A) the correlation between them is

$$\text{cor}\big(G_j(A), G_{j'}(A)|H\big) = \frac{1}{1+\alpha} = \mathbb{P}(G_j = G_{j'}|H).$$

This result is independent of the set A and provides a natural interpretation for the additional parameter in the NDP. Therefore, hereinafter we refer to it as the prior correlation between distributions, denoted as $\text{cor}(G_j, G_{j'}|H)$. The correlation between draws from the process also can be calculated (see App. A), yielding

$$\text{cor}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j'}) = \begin{cases} \dfrac{1}{(1+\beta)}, & j = j' \\ \dfrac{1}{(1+\alpha)(1+\beta)}, & j \neq j'. \end{cases}$$

This shows that the a priori correlation between observations arising from the same center is larger than the correlation between observations from different centers, which is an appealing feature. Given a specific form for $p(\cdot|\boldsymbol{\theta}_j, \boldsymbol{\phi})$, the previous expression allows us to calculate the prior correlation that the model induces on the observations.

Note that as $\alpha \to \infty$, each distribution in the collection is assigned to a distinct atom of the stick-breaking construction. Therefore, the distributions become a priori independent given the baseline measure H, which agrees with the fact that $\lim_{\alpha\to\infty} \text{cor}(G_j, G_{j'}|H) = 0$. On the other hand, as $\alpha \to 0$, the a priori probability of assigning all of the distributions to the same atom G^* goes to 1, and thus the correlation goes to 1. Thus approaches (1) and (2) for the analysis of multiple centers described earlier are limiting cases of the NDP. Moreover, because $F_j(\cdot) \to p(\cdot|\boldsymbol{\theta}^*_j, \boldsymbol{\phi})$ as $\beta \to 0$, the NDP also encompasses the natural parametric-based clustering [option (3)] as a limiting case.

Because every G^*_k is almost surely discrete, the model simultaneously enables clustering of observations within each center along with clustering the distributions themselves. For example, we can simultaneously group hospitals having the same distribution of patient outcomes while also identifying groups of patients within a hospital having the same outcome distribution. Indeed, centers j and j' are clustered together if $G_j = G_{j'} = G^*_k$ for some k, whereas patients i and i', from hospitals j and j', are clustered together if and only if $G_j = G_{j'} = G^*_k$ and $\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_{i'j'} = \boldsymbol{\theta}^*_{lk}$ for some l.

3.2 Alternative Characterizations of the Nested Dirichlet Process

Just as the DP is a distribution on distributions, the NDP can be characterized as a *distribution on the space of distributions on distributions*. Recall the original definition of the DP (Ferguson 1973, 1974) stated in Section 2. The choice $\boldsymbol{\Theta} \subset \mathbb{R}^n$ for the base space of the DP is merely a practical one, and the aforementioned results extend in general to any complete and separable metric space $\boldsymbol{\Theta}$. In particular, because the space of probability distributions is complete and separable under the weak topology metric, we could have started by taking $(\mathbf{P}, \mathcal{C}, Q)$ (defined before) as our base space and defining a new distributional space $(\mathbf{Q}, \mathcal{D}, S)$ such that $\mathcal{D}$ is the smallest σ-algebra generated by all weakly open sets in $\mathbf{Q}$ and $Q \in \mathbf{Q}$. In this setting $\mathbf{Q}$ is the space of *distributions on probability distributions* on $(\boldsymbol{\Theta}, \mathcal{B})$.

By requiring S to be such that $(Q(C_1), \ldots, Q(C_k)) \sim \mathsf{Dir}(\alpha\nu(C_1), \ldots, \alpha\nu(C_k))$ for any partition $(C_1, \ldots, C_k)$ of $\mathbf{P}$ generated under the weak topology and some α and suitable ν, we have defined a new DP, $S \sim \mathsf{DP}(\alpha\nu)$, this time on an abstract space, that satisfies the usual properties. The NDP is a special case of this formulation in which ν is taken to be a regular $\mathsf{DP}(\beta H)$; therefore, it is an example of a DP in which the baseline measure is a stochastic process generating probability distributions. This justifies the notation $G_j \overset{\text{iid}}{\sim} Q$ with $Q \sim \mathsf{DP}(\alpha\mathsf{DP}(\beta H))$ introduced earlier.

The NDP also can be characterized as a DDP (MacEachern 2000) where the stochastic process generating the elements of the stick-breaking representation corresponds to a Pólya urn (see Rodriguez 2007 for details). Finally, the NDP can be viewed as a way to simultaneously define a prior on a random partition of the collection $\{G_1, \ldots, G_J\}$ (in the style of Quintana and Iglesias 2003) and each of the resulting unique distributions.

3.3 Comparing the Nested Dirichlet Process With Other Nonparametric Models

It is important to note that although both approaches generalize the DP to allow hierarchical data structures, the dependence induced by the NDP is fundamentally different from that induced by the HDP. Figure 1 illustrates these differences. In the HDP, one draw from a Dirichlet process is used as the baseline measure G_0 of the process generating the members of the collection. As discussed by Teh et al. (2006), this implies that $\{G_1, \ldots, G_J\}$ share the same atoms (the atoms of G_0) but assign them different weights. Therefore, $\mathbb{P}(G_j = G_{j'}) = 0$ under the HDP, and clustering occurs only at the level of the observations.

On the other hand, the construction of the NDP implies that two given distributions either share both atoms and weights [making them exactly equal, as G_1 and G_3 in Fig. 1(b)] or do not share any of the features (like G_1 and G_2 in the same panel). This induces clustering on both observations and distributions.

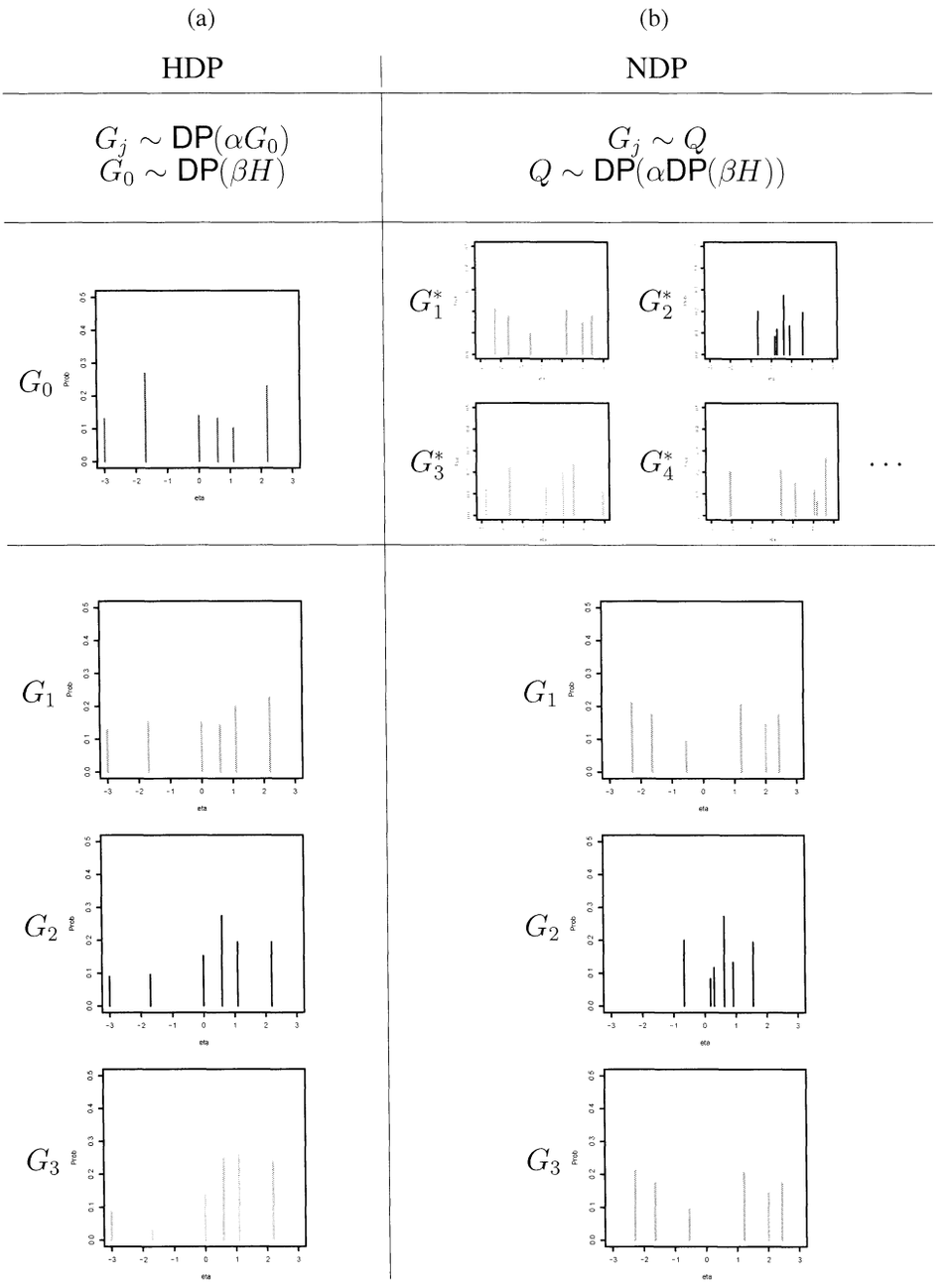

Figure 1. Comparing the HDP (a) and the NDP (b). For the HDP, the distributions $\{G_1, \ldots, G_J\}$ share the same atoms but assign them different weights. For the NDP, the different distributions have either the same atoms with the same weights or completely different atoms and weights.

The NDP is also different from the linear combination models of Müller et al. (2004), which allow for a limited form of clustering across distributions. Müller et al. (2004) represented an unknown distribution G_j as a linear combination $G_j = \epsilon_j H_0 + (1 - \epsilon_j) H_j$, where each H_j is an independent draw from a regular DP. H_0 is called the common component, and the H_j's, for $j \geq 1$, are called the idiosyncratic components. Note that the two distributions G_j and $G_{j'}$ are equal under this model if only if they correspond to the common component in the mixture, that is, $\epsilon_j = \epsilon_{j'} = 0$, implying that $G_j = G_{j'} = H_0$. Thus there is at most one cluster with more than one member.

4. POSTERIOR COMPUTATION

Broadly speaking, there are three strategies for computation in standard DP models: (a) Use the Pólya urn scheme to marginalize out the unknown infinite-dimensional distribution(s) (MacEachern 1994; Escobar and West 1995; MacEachern and Müller 1998); (b) use a truncation approximation to the stick-breaking representation of the process and then resort to methods for computation in finite-mixture models (Ishwaran and Zarepour 2002; Ishwaran and James 2001); and (c) use reversible-jump Markov chain Monte Carlo (RJMCMC) algorithms for finite mixtures with an unknown number of components (Dahl 2003; Green and Richardson 2001; Jain and Neal

2000). In the sequel we use the auxiliary variables $\zeta_j = k$ and $\xi_{ij} = l$ if $G_j = G_k^*$ and $\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_{lk}^*$ to indicate membership to the distributional and observational clusters.

Samplers for the NDP based on Pólya urns are in general infeasible. Although sampling ξ_{ij} given $(\zeta_1, \ldots, \zeta_J)$ using a Pólya urn scheme is straightforward, sampling ζ_j requires evaluation of the predictive distributions $p(\mathbf{y}_j|H)$ or $p(\mathbf{y}_j|\{\mathbf{y}_s|\zeta_s = k\})$ (both of which are finite mixtures with a number of terms that grows exponentially with n_j), or $p(\mathbf{y}_j|G_s^*)$ (the evaluation of which requires an infinite sum). The supplemental material at *http://www.amstat.org/PUBLICATIONS/jasa/* provides details.

Algorithms using RJMCMC in the NDP are likely to encounter similar problems, with the added disadvantage of low acceptance probabilities due to the large number of parameters that need to be proposed at the same time, with no obvious way to construct efficient proposals. Thus we focus on samplers based on truncation approximations, which are obtained by replacing the infinite sums in (2) and (3) by finite sums of K and L elements. It can be shown that the total variation distance between the prior predictive distributions generated by the NDP and its truncation has a strictly decreasing bound as L and K go to infinity. In addition, the posterior distribution for $\{\boldsymbol{\theta}_{ij}\}_{i\le n_j, j\le J}$ under the truncation converges in distribution to the posterior distribution under the NDP prior. Details on these results closely follow the results of Ishwaran and James (2001, 2002) and are given in Appendix B. The supplemental material also presents some numerical evidence suggesting reasonable truncation values as a function of sample size, which for the numerical examples discussed in this article were fixed to $K = 35$ and $L = 55$. Computation proceeds through the following steps:

1. Sample the center indicators ζ_j for $j = 1, \ldots, J$ from a multinomial distribution with probabilities
$$\mathbb{P}(\zeta_j = k|\cdots) = q_k^j \propto \pi_k^* \prod_{i=1}^{I_j} \sum_{l=1}^{L} w_{lk}\, p(y_{ij}|\boldsymbol{\theta}_{lk}^*, \boldsymbol{\phi}).$$
2. Sample the group indicators ξ_{ij} for $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$ from another multinomial distribution with probabilities
$$\mathbb{P}(\xi_{ij} = l|\cdots) = b_{ij}^l \propto w_{l,\zeta_j}^*\, p(y_{ij}|\boldsymbol{\theta}_{l\zeta_j}^*, \boldsymbol{\phi}).$$
3. Sample π_k^* by generating
$$(u_k^*|\cdots) \sim \text{beta}\left(1 + m_k, \alpha + \sum_{s=k+1}^{K} m_s\right), \quad k = 1, \ldots, K-1,$$
$$u_K^* = 1,$$
where m_k is the number of distributions assigned to component k, and constructing $\pi_k^* = u_k^* \prod_{s=1}^{k-1}(1 - u_s^*)$.
4. Sample w_{lk}^* by generating
$$(v_{lk}^*|\cdots) \sim \text{beta}\left(1 + n_{lk}, \beta + \sum_{s=l+1}^{L} n_{ls}\right), \quad l = 1, \ldots, L-1,$$
$$v_{Lk}^* = 1,$$
where n_{lk} is the number of observations assigned to atom l of distribution k, and constructing $w_{lk}^* = v_{lk}^* \times \prod_{s=1}^{l-1}(1 - v_{sk}^*)$.
5. Sample $\boldsymbol{\theta}_{lk}^*$ from
$$p(\boldsymbol{\theta}_{lk}^*|\cdots) \propto \left[\prod_{\{i,j|\zeta_j=k,\xi_{ij}=l\}} p(y_{ij}|\boldsymbol{\theta}_{lk}^*, \boldsymbol{\phi})\right] h(\boldsymbol{\theta}_{lk}^*),$$
where $h(\boldsymbol{\theta}_{lk}^*)$ is the density corresponding to the baseline measure H. If no observation is assigned to a specific cluster, then the parameters are drawn from the prior distribution $h(\boldsymbol{\theta}_{lk}^*)$. If the prior is conjugate to the likelihood, then sampling is greatly simplified; however, nonconjugate priors can be accommodated using rejection sampling or Metropolis–Hastings steps.
6. Sample the concentration parameters α and β from
$$p(\alpha|\cdots) \propto \alpha^{K-1} \exp\left\{\alpha \sum_{k=1}^{K-1} \log(1 - u_k^*)\right\} p(\alpha)$$
and
$$p(\beta|\cdots) \propto \beta^{K(L-1)} \exp\left\{\beta \sum_{l=1}^{L-1}\sum_{k=1}^{K} \log(1 - v_{lk}^*)\right\} p(\beta).$$
If conditionally conjugate priors $\alpha \sim \mathsf{G}(a_\alpha, b_\alpha)$ and $\beta \sim \mathsf{G}(a_\beta, b_\beta)$ are chosen, then
$$(\alpha|\cdots) \sim \mathsf{G}\left(a_\alpha + (K-1), b_\alpha - \sum_{k=1}^{K-1} \log(1 - u_k^*)\right)$$
and
$$(\beta|\cdots) \sim \mathsf{G}\left(a_\beta + K(L-1), b_\beta - \sum_{l=1}^{L-1}\sum_{k=1}^{K} \log(1 - v_{lk}^*)\right).$$
Note that the accuracy of the truncation depends on the values of α and β. Thus the hyperparameters (a_α, b_α) and (a_β, b_β) should be chosen to give little prior probability to values of α and β larger than those used to calculate the truncation level.
7. Sample ϕ from its full conditional distribution,
$$p(\boldsymbol{\phi}|\cdots) \propto \left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} p\left(y_{ij}|\boldsymbol{\theta}_{\xi_{ij},\zeta_j}^*, \boldsymbol{\phi}\right)\right] p(\boldsymbol{\phi}).$$

Besides its simplicity of implementation, an additional advantage of this truncation scheme is that implementation in parallel computing environments is straightforward, which can be especially useful for large sample sizes.

5. ILLUSTRATIONS

In Section 5.1 we present a simulation study in which we focus on the problem of clustering distributions but do not attempt to interpret the clusters induced by the model on the observations. The problem of nested clustering is discussed in Section 5.2 in the context of a real data set.

5.1 Simulation Study

In this section we present a simulation study designed to (a) illustrate the ability of the model to discriminate among distribution functions and (b) show the ability of the NDP to borrow information and provide more accurate density estimates. The setup of the study is as follows: J samples of size n are obtained from four mixtures of four Gaussian components defined in Table 1 and plotted in Figure 2. These distributions have been chosen to exemplify functions that are hard to distinguish; T1 and T2 are asymmetric and composed of the same two Gaussian components that have been weighted differently, whereas T3 and T4 share three distributions located symmetrically around the origin, differing only in an additional bump that T4 presents on the right tail.

The values of J and n were varied across the study to assess the influence of sample size on the discriminating capability of the model. The precision parameters α and β were both fixed to 1, and a normal inverse-gamma distribution, $\mathsf{NIG}(0, .01, 3, 1)$, was chosen as the baseline measure H, implying that, a priori, $\mathbb{E}(\mu|\sigma^2) = 0$, $\mathbb{V}(\mu|\sigma^2) = 100\sigma^2$, $\mathbb{E}(\sigma^2) = 1$, and $\mathbb{V}(\sigma^2) = 3$. The algorithm described in Section 4 was used to obtain samples of the posterior distribution under the NDP. Based on the empirical analysis presented in the supplemental material (available at *http://www.amstat.org/PUBLICATIONS/jasa/*), truncation levels were chosen as $K = 35$ and $L = 55$. All results shown below are based on 50,000 samples obtained after a burn-in period of 5,000 iterations.

Visualization of high-dimensional clustering structures is a hard task. A commonly used summary looks at the set of $J(J-1)/2$ possible pairs of populations and for each pair obtains the probability that the two of them fall in the same cluster. Estimates of these probabilities are easily obtained from the output of our MCMC algorithm and can be effectively displayed using heat maps, like those shown in Figure 3. To simplify interpretation of the plot, samples from the same mixture distribution are adjacent. Other possible summaries are discussed in Section 5.2.

For small values of n, the NDP is able to roughly separate T1 and T2 from T3 and T4, but cannot discriminate between T1 and T2 or between T3 and T4. This is not really surprising, because the method is designed to induce clustering. Therefore, when differences are highly uncertain, the method prefers to create fewer, rather than more, clusters. But as n increases, the model is able to distinguish between distributions and to correctly identify both the number of groups and the membership of the distributions. It is particularly interesting that the model finds it easier to discriminate between distributions that differ just in one atom rather than in weights. On the other hand, as J increases, the model is capable of discovering the underlying groups of distributions, but the uncertainty on the membership is not reduced without increasing n.

Table 1. Parameters for the true distributions $p_T = \sum_i w_i \mathsf{N}(\mu_i, \sigma_i^2)$ used in the simulation study

Distribution	Comp 1			Comp 2			Comp 3			Comp 4		
	w	μ	σ^2	w	μ	σ^2	w	μ	σ^2	w	μ	σ^2
T1	.75	0	1.0	.25	3.0	2.0						
T2	.55	0	1.0	.45	3.0	2.0						
T3	.40	0	1.0	.30	−2.0	2.0	.30	2.0	2.0			
T4	.39	0	1.0	.29	−2.0	2.0	.29	2.0	2.0	.03	10.0	1.0

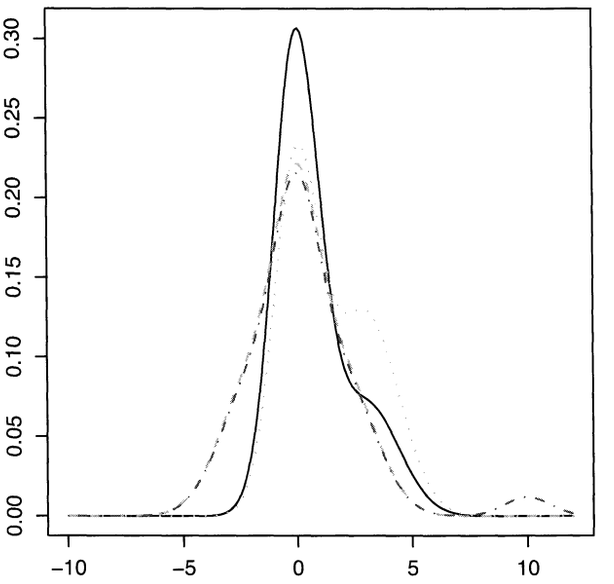

Figure 2. True distributions used in the simulation study (——, T1; · · · · ·, T2; – – –, T3; – · – ·, T4).

Figure 4 shows density estimates obtained for sample 1 of the example where $J = 20$ and $n = 100$. Panel (a) shows the density estimate obtained from the NDP (which borrows information across all samples), whereas panel (b) was obtained by fitting a regular DPM model with the same precision parameter $\beta = 1$ and baseline measure. We note that although the NDP borrows information across samples that actually come from a slightly different data-generating mechanism, the estimate is more accurate; it not only captures the small mode to the right more clearly, but also emphasizes the importance of the main mode. Indeed, the Kullback–Leibler divergence of the density estimate relative to the true distribution for the estimate of T1 under the NDP is .011, whereas under the regular DPM it is .017.

5.2 Health Care Quality in United States

Data on quality of care in hospitals across the United States and associated territories are publicly available from the Department of Health and Human Services at *http://www.hhs.gov/*. Twenty measures are recorded for each hospital, comprising such aspects as proper and timely application of medication, treatment, and discharge instructions. In what follows we focus on one specific measure: the proportion of patients given the most appropriate initial antibiotic(s), transformed through the logit function. Four covariates are available for each center: type of hospital (either acute care or critical access), ownership (nine possible levels, including government at different levels, proprietary, and different types of voluntary nonprofit hospitals), whether the hospital provides emergency services (yes or no), and whether it has an accreditation (yes or no). Location, in the form of the ZIP code, also is available. Hospitals with fewer than 30 patients treated and territories with fewer than

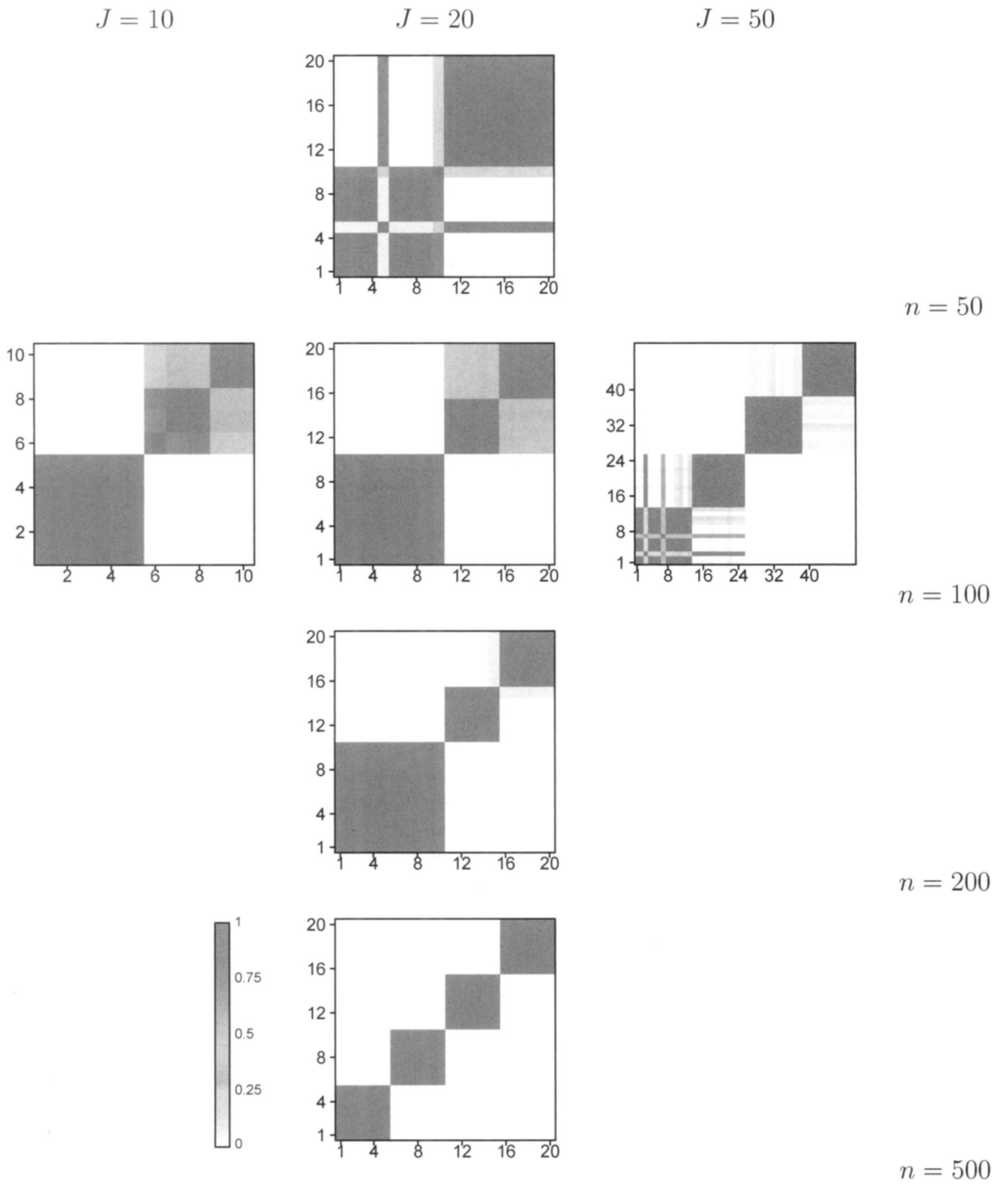

Figure 3. Pairwise probabilities of joint classification for the simulation study.

4 hospitals were judged to be misrepresentative and removed from the sample, yielding a final sample size of 3,077 hospitals in 51 territories (the 50 states plus the District of Columbia). The number of hospitals per state varies widely, with, for example, 5 in Delaware, 10 in Alaska, 13 in Idaho, 164 in Florida, 205 in Texas, and 254 in California. The number of patients per hospital varies between 30 and 1,175, with quartiles at 76, 130, and 197 patients. Because the values tend to be large, we perform our analysis on the observed proportion without adjusting for sample sizes.

We wish to study differences in quality of care across states after adjusting for the effect of the available covariates. Specifically, we are interested in clustering states according to their quality, rather than obtaining smoothed quality estimates. Indeed, differences in quality of care are probably due to a combination of state policies and practice standards, and clustering patterns can be used to identify such factors. Therefore, there is no reason to assume a priori that physically neighboring states have similar outcomes.

To motivate the use of the NDP, we consider first a simple preliminary analysis of the data. To adjust for the covariates, an ANOVA model containing only main effects is fitted to the data. Of these effects, only the presence of an emergency service and the ownership seem to affect the quality of the hospital

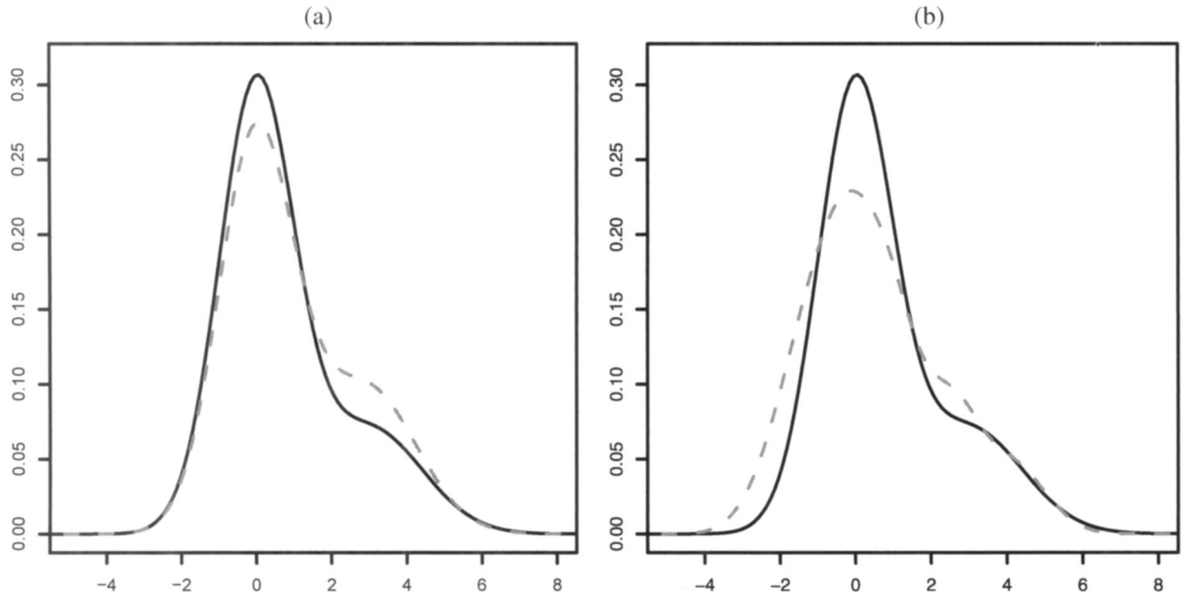

Figure 4. True (——) and estimated (– – – –) densities for distribution 1 of the simulation with $J = 20$ and $n = 100$. (a) An estimate based on the NDP, which borrows information across all samples. (b) An estimate based only on sample 1.

(p values .011 and 1.916×10^{-8}). Residual plots for this model show some deviation from homocedasticity and normality (see Fig. 5), but, given the large sample size, it is unlikely that this has any impact on the significance of the covariates.

It is clear from Figure 6 that residual distributions vary across states. At this point, one possible course of action is to assume normality within each state and cluster states according to the mean and/or variance of its residual distribution. But the illustrative density estimates in Figure 7 (obtained using Gaussian kernels and a bandwidth chosen using the rule of thumb described in Silverman 1986) show that state-specific residual distributions can be highly nonnormal and that changes across states can go beyond location and scale changes to affect the entire shape of the distribution. Invoking asymptotic arguments at this point is inappropriate, because sample sizes are small and we are concerned about the shape of the distribution (rather than the parameters), for which no central limit theorem can be invoked. Figure 7 also shows that states located in very different geographical areas (e.g., California and Minnesota or Florida and North Carolina) can have similar error distributions.

To improve the analysis, we resort to a Bayesian formulation of the main-effects ANOVA and use the NDP to model the

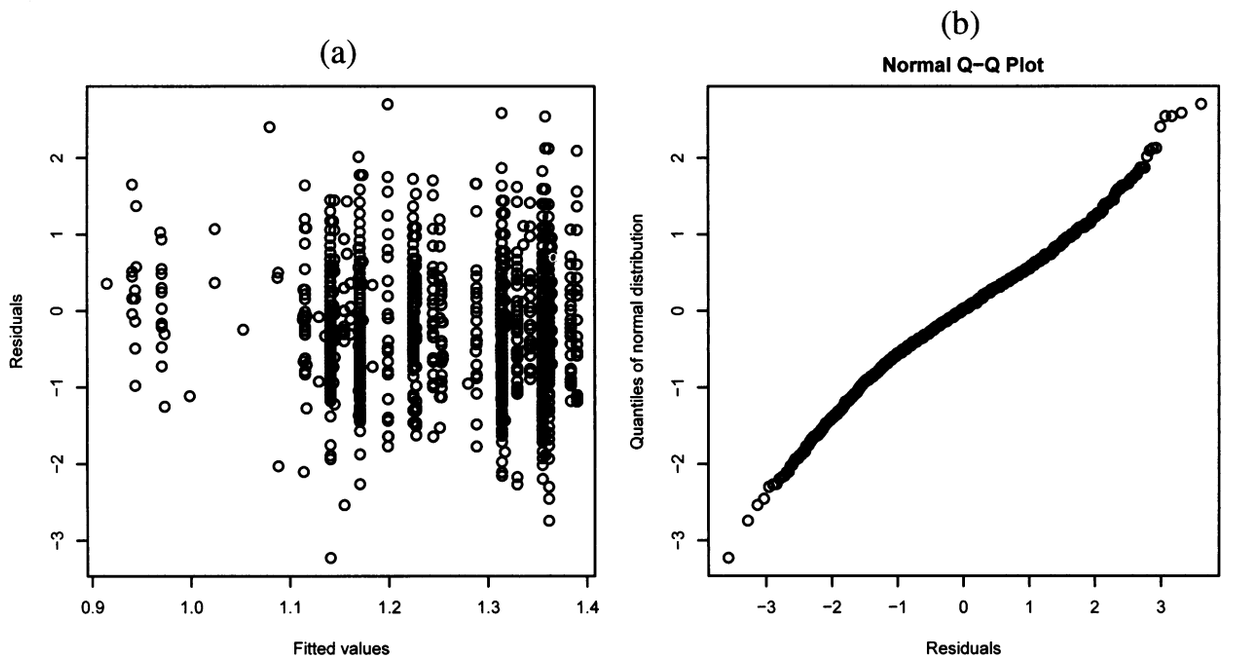

Figure 5. Residual plots for the ANOVA model on the initial antibiotic data. (a) Residuals versus fitted values. (b) Quantile–quantile plot.

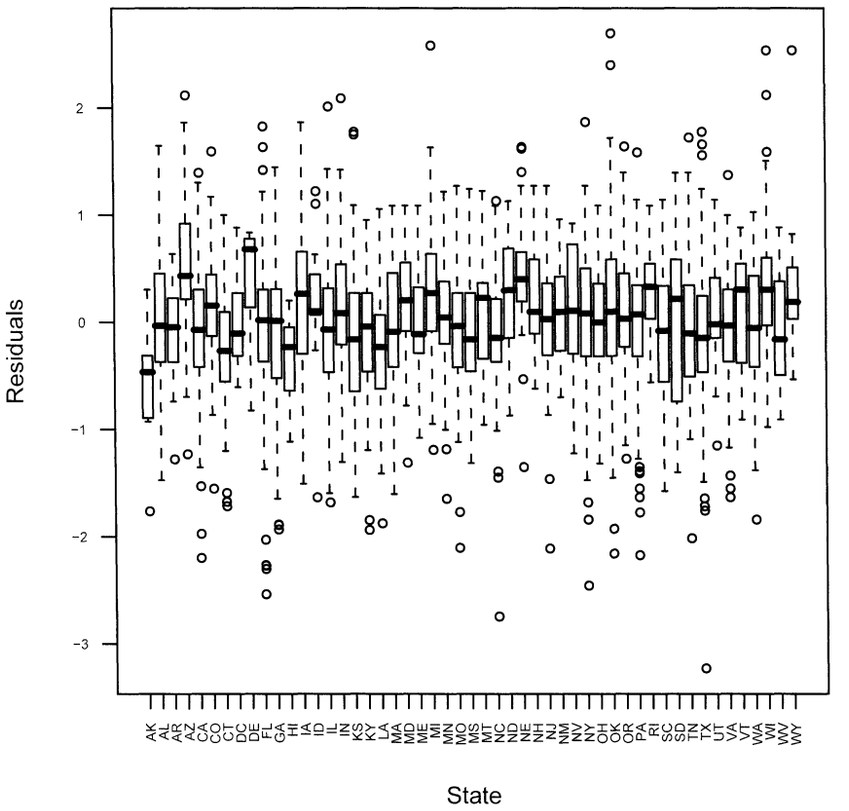

Figure 6. State-specific residual boxplots for the ANOVA model on the initial antibiotic data.

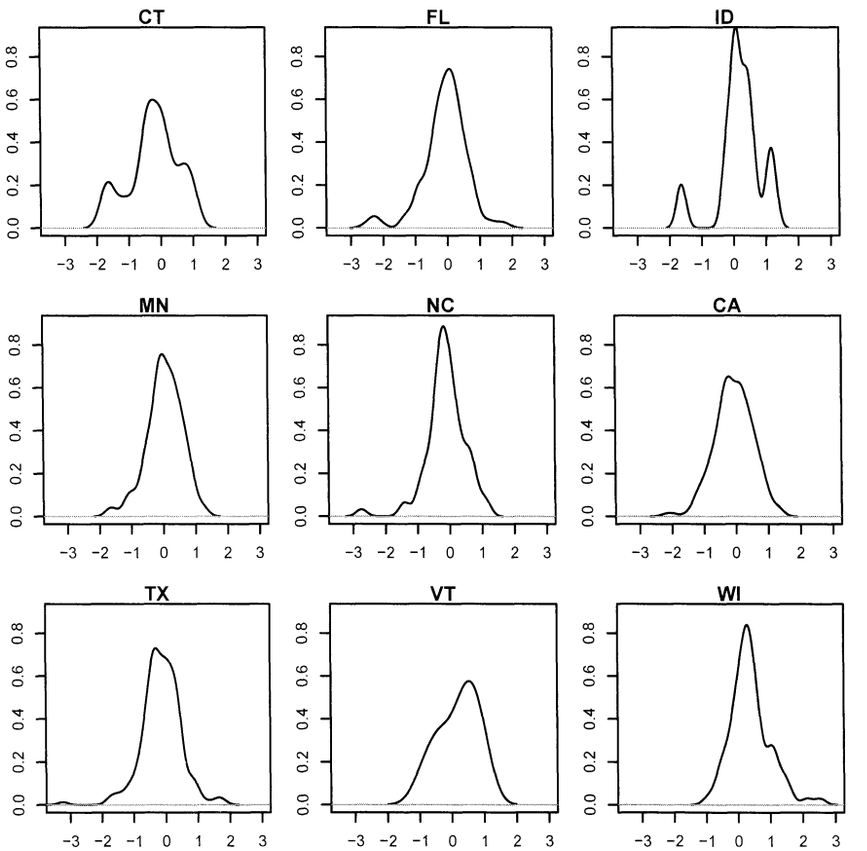

Figure 7. Density estimates for the residual distribution in selected states. Note that the distributions seem clearly nonnormal, and that their shape can have important variations, making any parametric assumption difficult to support.

state-specific error distributions. Specifically, if we let y_{ij} be the response of hospital i in state j after subtraction of the global mean, then

$$y_{ij} = \mu_{ij} + \mathbf{x}_{ij}\boldsymbol{\gamma} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathsf{N}(0, \sigma_{ij}),$$
$$(\mu_{ij}, \sigma_{ij}^2) \sim G_j, \{G_1, \ldots, G_J\} \sim \mathsf{DP}(\alpha \mathsf{DP}(\beta, H)),$$

where $\mathbf{x}_{ij}$ is the vector of covariates associated with the hospital. Thus here, $\boldsymbol{\theta}_{ij} = (\mu_{ij}, \sigma_{ij}^2)$ and $\boldsymbol{\phi} = \boldsymbol{\gamma}$. Prior specification is simplified by centering the observations. We choose $H = \mathsf{NIG}(0, .01, 3, 3)$, which implies that $\mathbb{E}(\mu|\sigma^2) = 0$, $\mathbb{V}(\mu|\sigma^2) = 100\sigma^2$, $\mathbb{E}(\sigma^2) = 1$, and $\mathbb{V}(\sigma^2) = 3$. This choice reflects the natural scale (logit) of the data, which on a Gaussian linear model would be expected to have mean 0 and variance close to unit after adjusting for covariates. We use a standard reference (flat) prior on $\boldsymbol{\gamma}$. Finally, we set $\alpha, \beta \sim \mathsf{G}(3, 3)$ a priori, implying that $\mathbb{E}(\alpha) = \mathbb{E}(\beta) = 1$ (a common choice in the literature) and $\mathbb{P}(\alpha > 3) = \mathbb{P}(\beta > 3) \approx .006$. Note that this choice implies that $\mathbb{P}(\text{cor}(G_j, G_{j'}) > .25) \approx .994$.

Posterior computation is straightforward using the algorithm presented in Section 4. As described there, the model is a regular ANOVA with known variance conditional on $\boldsymbol{\theta}$, and the full conditional posterior distribution of $\boldsymbol{\gamma}$ following a normal distribution. On the other hand, conditional on $\boldsymbol{\gamma}$, we can use the NDP sampler on the pseudo-observations $z_{ij} = y_{ij} - \mathbf{x}_{ij}\boldsymbol{\gamma}$. The results that follow are based on 50,000 iterations obtained after a burn-in period of 5,000 iterations. As with the simulation study, we choose $K = 35$ and $L = 55$ as the truncation levels. The results seem to be robust to reasonable changes in prior specification, and there is no evidence of lack of convergence from visual inspection of trace plots.

The posterior distribution on the number of distinct G's shows strong evidence in favor of either two or three components (posterior probabilities .616 and .363) and little support for one, four, or five distributions (posterior probabilities 0, .02, and .001). As with the simulated example, we visualize the matrix of pairwise probabilities using a heat map, as shown in Figure 8. To make sense of the plot, we first reorder the states using an algorithm that borrows ideas from hierarchical clustering (see the supplemental materials for details).

This heat map provides additional insight into the clustering structure. It shows three well-defined groups: (a) a large homogenous clusters of 31 members (lower left corner of the plot), (b) a small homogenous cluster of 6 states (upper right corner), and (c) an heterogeneous group comprising the remaining 15 states, which are not clear members of any of the 2 previous clusters and do not seem to form a coherent cluster among themselves.

Several different approaches can be used to choose one specific partition of the set of states. One appealing option is to

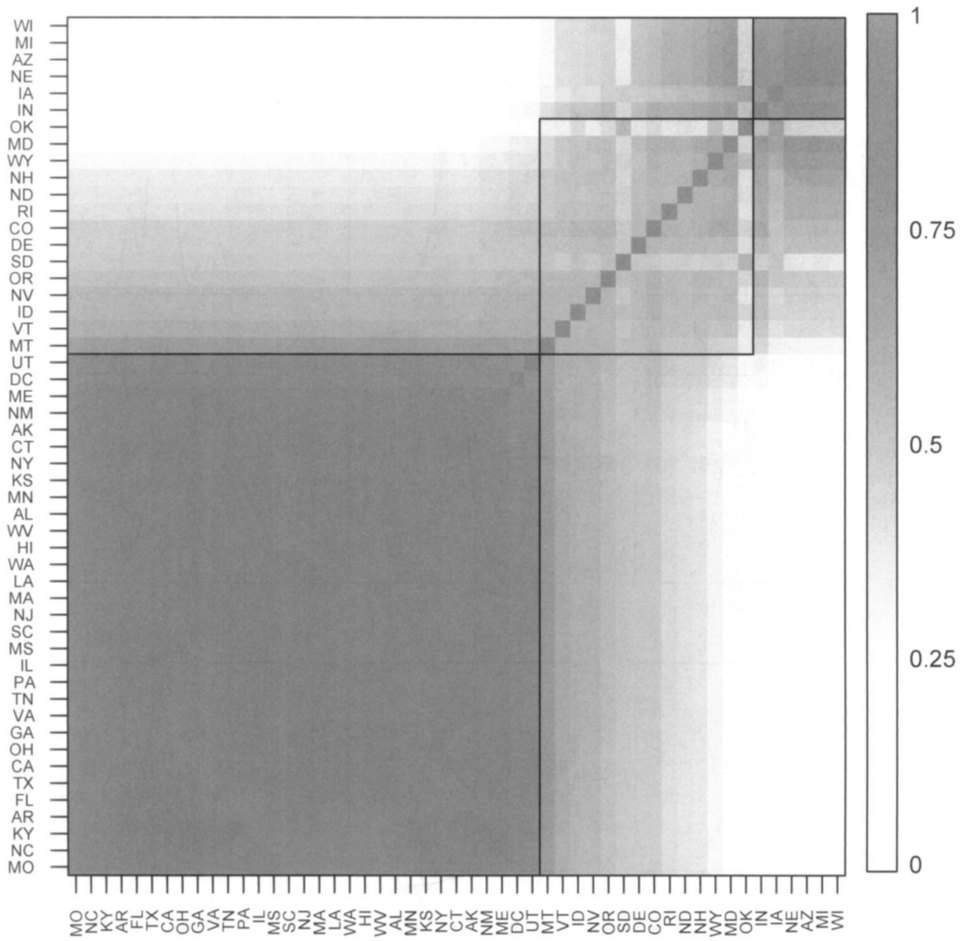

Figure 8. Residual plots for the ANOVA model on the initial antibiotic data.

choose $\hat{\mathbf{p}}$ such that it minimizes a given loss functions. Following Binder (1978, 1981) and Lau and Green (2006), we chose the label-invariant loss function

$$\Psi(\mathbf{p}, \hat{\mathbf{p}}) = \sum_{\{(j,j'):j<j'\le J\}} \left(a\mathbf{1}_{(\zeta_j=\zeta_{j'},\hat{\zeta}_j\neq\hat{\zeta}_{j'})} + b\mathbf{1}_{(\zeta_j\neq\zeta_{j'},\hat{\zeta}_j=\hat{\zeta}_{j'})}\right), \tag{4}$$

where $\mathbf{p}$ denotes the true (unknown) partition of states, $\mathbf{1}_A$ is the indicator function on the set A, ζ_j and $\hat{\zeta}_j$ denote the true and estimated clustering indicators induced by $\mathbf{p}$ and $\hat{\mathbf{p}}$, and a and b are pairwise misclassification penalties. Minimizing the posterior expected loss under Ψ is equivalent to picking a partition $\hat{\mathbf{p}}$ such that the function

$$\sum_{\{(j,j'):j<j'\le J\}} \mathbf{1}_{(\hat{\zeta}_j=\hat{\zeta}_{j'})}(\rho_{jj'} - \tau)$$

is maximized, where $\rho_{jj'}$ is the probability of joint classification for states i and j (the values depicted in Fig. 8) and $\tau = b/(a+b) \in [0, 1]$. For $\tau = 0$, the optimal partition places all states in a single cluster. On the other hand, if $\tau = 1$, then the optimal allocation creates individual clusters for each state. Intermediate values of τ correspond to a compromise between both types of errors. Thus for $\tau = .3$, the optimal partition divides the 51 states into 2 groups, a small group comprising 8 states (AZ, IA,IN, MD, MI, NE, OK, and WI) and a large group comprising all of the remaining states. For $\tau = .5$, the optimal allocation corresponds to three clusters: a very small cluster comprising only OK and SD; an intermediate cluster comprising AZ, CO, DE, IA, IN, MD, MI, ND, NE, NH, RI, WI, and WY (note the similarities with $\tau = .3$); and a large cluster comprising the remaining states. Finally, for $\tau = .75$, the optimal clustering agrees with the one depicted in Figure 8, with 2 tight groups and 14 single-state clusters. The posterior probabilities for each of these partitions estimated from the MCMC are 4×10^{-5}, 0, and 0. In contrast, the most frequent configuration sampled by the model (posterior probability 7×10^{-4}, much larger but still rather small) divides the sample into 2 groups, a small group with 17 states (AZ, CO, IA, ID, IN, MD, MI, ND, NE, NH, NV, OK, OR, RI, SD, WI, and WY), and another group with the remaining states.

We also can study the clustering of hospitals within states, but meaningful interpretations must be done conditionally on the state-level partition. As an illustration, consider conditioning in the optimal clustering suggested by taking $\tau = .75$. The small cluster (comprising AZ, IA, IN, MI, NE, and WI) is composed of 1 group (posterior probability .81) or 2 groups (posterior probability .18) of hospitals, whereas the large cluster (comprising 31 states, including TX and NC) is composed of 2 (probability .89) or 3 (probability .10) different groups of hospitals. This shows that low-/high-quality groups of hospitals can be identified within each group of states and that state-specific distributions are nonnormal as expected.

Indeed, Figure 9 shows posterior predictive density estimates for the residuals of four representative states: North Carolina (cluster 1), Wisconsin (cluster 2), and South Dakota and Oklahoma, which belong to the third group. North Carolina (and, in general, the states in group 1) presents a lower mean and a heavier-than-Gaussian left tail, indicating that each of those states contains some underperforming hospitals and few or no over performing hospitals. The situation for Wisconsin and cluster 2 is reversed; these seem to be states with a higher average performance, quite a few hospitals with an excellent record in the application of antibiotics, and few or no low-quality hospitals. Finally, South Dakota and Oklahoma exhibit a mixed behavior, showing evidence of both underperforming and overperforming hospitals. Note that these density estimates are much smoother than those shown in Figure 7. This is not surprising for three reasons: (a) As discussed by Escobar and West (1995), location-scale mixtures act as adaptive-bandwidth kernel estimators; (b) we are borrowing information across states; and (c) our estimates average over a large number of alternative models, which induces smoothness. All of these features tend to produce smoother estimates than those obtained from standard kernel density estimates.

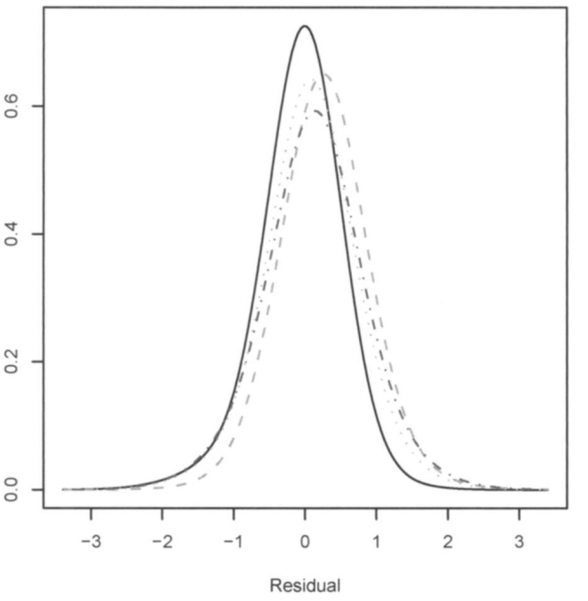

Figure 9. Mean predictive density for four representative states: North Carolina (NC; ——), Wisconsin (WI; – – ·), South Dakota (SD; · · · · ·), and Oklahoma (OK; · – · -).

It is interesting to contrast these results with those obtained from a similar model that uses the HDP instead of the NDP to induce dependence among residual distributions. Although density estimates (not shown) for the different states appear similar to those shown in Figure 9, the HDP does not provide an equivalent to Figure 8, because it clusters only hospitals, not states. Indeed, the HDP-based model divides the 3,077 hospitals in roughly 3 groups, which we can easily label as average (for the largest, central group), underperformers, and overperformers (both containing a relatively small number of observations). The density estimates are then obtained by weighting these groups differentially for each state.

6. DISCUSSION

We have formulated a novel extension of the DP for a family of a priori exchangeable distributions that allows us to simultaneously cluster groups and observations within groups. Moreover, the groups are clustered by their entire distribution, rather than by particular features of the distribution. We demonstrated the flexibility of the model through both a simulation study and

an application where the NDP is used to jointly model the random effect and error distribution of an ANOVA model.

One natural generalization of the NDP is to replace the $\mathsf{beta}(1,\alpha)$ and $\mathsf{beta}(1,\beta)$ stick-breaking densities with more general forms. In the setting of stick-breaking priors for a single random probability measure, Ishwaran and James (2001) considered general $\mathsf{beta}(a_k, b_k)$ forms, with the DP corresponding to the special case where $a_k = 1$ and $b_k = \alpha$. Similarly, by using $\mathsf{beta}(a_k, b_k)$ and $\mathsf{beta}(c_k, d_k)$, we can obtain a rich class of nested stick-breaking priors that encompasses the NDP as a particular case.

Including hyperparameters in the baseline measure H is another straightforward extension. We note that, conditional on H, the distinct atoms $\{G_k^*\}_{k=1}^{\infty}$ are assumed to be independent. Therefore, including hyperparameters in H allows us to parametrically borrow information across the distinct distributions.

APPENDIX A: CORRELATION IN THE NESTED DIRICHLET PRIOR

We start by calculating the correlation between distributions. In the first place,

$$\begin{aligned}\mathbb{E}(G_j(B)G_k(B)) &= \mathbb{E}\big(G_j(B)G_k(B)|G_j = G_k\big)\mathbb{P}(G_j = G_k)\\ &\quad + \mathbb{E}\big(G_j(B)G_k(B)|G_j \neq G_k\big)\mathbb{P}(G_j \neq G_k)\\ &= \mathbb{E}(G_j^2(B))\frac{1}{\alpha+1} + \mathbb{E}(G_j(B))\mathbb{E}(G_k(B))\frac{\alpha}{\alpha+1}\\ &= \frac{H(B)(1-H(B))}{(\alpha+1)(\beta+1)} + H^2(B).\end{aligned}$$

Finally,

$$\begin{aligned}\operatorname{cor}(G_j(B), G_k(B)) &= \frac{\mathbb{E}(G_j(B)G_k(B)) - \mathbb{E}(G_j(B))\mathbb{E}(G_k(B))}{\sqrt{\mathbb{V}(G_j(B)\mathbb{V}(G_k(B)}}\\ &= \frac{\frac{H(B)(1-H(B))}{(\alpha+1)(\beta+1)} + H^2(B) - H^2(B)}{\frac{H(B)(1-H(B))}{(\beta+1)}}\\ &= \frac{1}{\alpha+1}.\end{aligned}$$

For the correlation between samples of the NDP, note that for the NDP and if $j = j'$, then

$$\begin{aligned}\operatorname{cov}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j'}) &= \operatorname{cov}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j}|\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_{i'j} = \boldsymbol{\theta}_{lk}^*)\mathbb{P}(\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_{i'j} = \boldsymbol{\theta}_{lk}^*)\\ &\quad + \operatorname{cov}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j}|\boldsymbol{\theta}_{ij} \neq \boldsymbol{\theta}_{i'j})\mathbb{P}(\boldsymbol{\theta}_{ij} \neq \boldsymbol{\theta}_{i'j})\\ &= \frac{1}{1+\beta}\mathbb{V}(\boldsymbol{\theta}_{lk}^*).\end{aligned}$$

Because the $\boldsymbol{\theta}_{lk}^*$'s are iid for all l and k, it follows that $\operatorname{cor}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j}) = \frac{1}{1+\beta}$. On the other hand, if $j \neq j'$, then

$$\begin{aligned}\operatorname{cov}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j'}) &= \operatorname{cov}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j'}|G_j = G_{j'} = G_k^*, \boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_{i'j'} = \boldsymbol{\theta}_{lk}^*)\\ &\quad \times \mathbb{P}(G_j = G_{j'} = G_k^*, \boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_{i'j'})\\ &\quad + \operatorname{cov}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j'}|G_j \neq G_{j'} \text{ or } \boldsymbol{\theta}_{ij} \neq \boldsymbol{\theta}_{i'j'})\\ &\quad \times \mathbb{P}(G_j \neq G_{j'} \text{ or } \boldsymbol{\theta}_{ij} \neq \boldsymbol{\theta}_{i'j'})\\ &= \frac{1}{(1+\alpha)(1+\beta)}\mathbb{V}(\boldsymbol{\theta}_{lk}^*).\end{aligned}$$

APPENDIX B: TRUNCATIONS

Here we consider finite-mixture versions of the NDP. Finite mixtures are usually simpler to understand, and considering them can help provide insight into the more complicated, infinite-dimensional models. In addition, they provide useful approximations that can be used for model fitting.

Definition B.1. An LK truncation of an $\mathsf{DP}(\alpha\mathsf{DP}(\beta, H))$ is defined by the finite-mixture model

$$\begin{aligned}G_j^K(\cdot) &\sim \sum_{k=1}^{K}\pi_k^*\delta_{G_k^{L*}(\cdot)}; \qquad \pi_k^* = v_k^*\prod_{s=1}^{l-1}(1-v_s^*);\\ G_k^{L*}(\cdot) &= \sum_{l=1}^{L} w_{lk}^*\delta_{\boldsymbol{\theta}_{lk}^*}(\cdot); \qquad w_{lk}^* = u_{lk}^*\prod_{s=1}^{l-1}(1-u_{sk}^*);\\ v_k^* &\sim \mathsf{beta}(1,\alpha), \qquad k = 1,\ldots,K-1; \qquad v_K^* = 1;\\ u_{lk}^* &\sim \mathsf{beta}(1,\beta), \qquad l = 1,\ldots,L-1; \qquad u_{Lk}^* = 1;\\ \boldsymbol{\theta}_{lk}^* &\sim H.\end{aligned}$$

We refer to this model as a bottom-level truncation or $\mathsf{NDP}^{L\infty}$ if $K = \infty$ and $L < \infty$, whereas if $K < \infty$ and $L = \infty$, we refer to it as a top-level truncation or $\mathsf{NDP}^{\infty K}$. Finally, if both L and K are finite, then we have a two-level truncation, or NDP^{LK}.

The total variation distance between an NDP and its truncation approximations can be shown to have decreasing bounds as $L, K \to \infty$. For simplicity, we consider the case where $n_j = n\ \forall j$.

Theorem B.1. Assume that samples of n observations have been collected for each of J distributions and are contained in vector $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_J')$. Also, let

$$P^{\infty\infty}(\boldsymbol{\theta}) = \int\int P(\boldsymbol{\theta}|G_j)P^{\infty}(dG_j|Q)P^{\infty}(dQ)$$

and

$$P^{LK}(\boldsymbol{\theta}) = \int\int P(\boldsymbol{\theta}|G_j)P^{L}(dG_j|Q)P^{K}(dQ)$$

be the prior distribution of the model parameters under the NDP model and its corresponding LK truncation after integrating out the random distributions, and let $P^{\infty\infty}(\mathbf{y})$ and $P^{LK}(\mathbf{y})$ be the prior predictive distribution of the observations derived from these priors. Then

$$\begin{aligned}\int |P^{LK}(\mathbf{y}) - P^{\infty\infty}(\mathbf{y})|\,d\mathbf{y} &\le \int |P^{LK}(\boldsymbol{\theta}) - P^{\infty\infty}(\boldsymbol{\theta})|\\ &\le \epsilon^{LK}(\alpha,\beta),\end{aligned}$$

where

$$\epsilon^{LK}(\alpha,\beta) = \begin{cases} 4\left(1 - \left[1 - \left(\frac{\alpha}{1+\alpha}\right)^{K-1}\right]^{J}\right) \\ \quad \text{if } L = \infty,\ K < \infty \\ 4\left(1 - \left[1 - \left(\frac{\beta}{\beta+1}\right)^{L-1}\right]^{nJ}\right) \\ \quad \text{if } L < \infty,\ K = \infty \\ 4\left(1 - \left[1 - \left(\frac{\alpha}{1+\alpha}\right)^{K-1}\right]^{J} \right. \\ \qquad \left. \times \left[1 - \left(\frac{\beta}{\beta+1}\right)^{L-1}\right]^{nJ}\right) \\ \quad \text{if } L < \infty,\ K < \infty. \end{cases}$$

The proof of this theorem closely follows theorem 1 and corollary 1 of Ishwaran and James (2002) and theorem 2 of Ishwaran and James (2001) and is included in the supplemental material available at *http://www.amstat.org/PUBLICATIONS/jasa/*. Note that the bounds approach zero in the limit, so the truncation approximations and its predictive distribution converge in total variation (and thus also in distribution) to the NDP. Furthermore, the bounds are strictly decreasing in both L and K. As a consequence of this observation, we have the following corollary.

Corollary B.1. The posterior distribution under an LK truncation and the corresponding NDP converge in distribution as both L and $K \to \infty$.

The proof is a simple consequence of the Bayes theorem,

$$\begin{aligned}\lim_{K,L\to\infty} p^{LK}(\boldsymbol{\theta}|\mathbf{y}) &= \lim_{K,L\to\infty} \frac{p(\mathbf{y}|\boldsymbol{\theta})\, p^{LK}(\boldsymbol{\theta})}{p^{LK}(\mathbf{y})} \\ &= \frac{p(\mathbf{y}|\boldsymbol{\theta}) \lim_{K,L\to\infty} p^{LK}(\boldsymbol{\theta})}{\lim_{K,L\to\infty} p^{LK}(\mathbf{y})} \\ &= p^{\infty\infty}(\boldsymbol{\theta}|\mathbf{y}).\end{aligned}$$

It is straightforward to extend the previous results and show that $\lim_{L\to\infty} \mathsf{NDP}^{LK} = \mathsf{NDP}^{\infty K}$ and $\lim_{K\to\infty} \mathsf{NDP}^{LK} = \mathsf{NDP}^{L\infty}$ in distribution.

[Received September 2006. Revised September 2007.]

REFERENCES

Bigelow, J. L., and Dunson, D. B. (2007), "Posterior Simulation Across Nonparametric Models for Functional Clustering," *Journal of the Royal Statistical Society*, Ser. B, to appear.

Binder, D. A. (1978), "Bayesian Cluster Analysis," *Biometrika*, 65, 31–38.

——— (1981), "Approximations to Bayesian Clustering Rules," *Biometrika*, 68, 275–285.

Blackwell, D., and MacQueen, J. B. (1973), "Ferguson Distribution via Pólya Urn Schemes," *The Annals of Statistics*, 1, 353–355.

Blei, D. M., and Jordan, M. I. (2006), "Variational Inference for Dirichlet Process Mixtures," *Bayesian Analysis*, 1, 121–144.

Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B. (2004), "Hierarchical Topic Models and the Nested Chinese Restaurant Process," in *Advances in Neural Information Processing Systems 16*, Cambridge, MA: MIT Press.

Bush, C. A., and MacEachern, S. N. (1996), "A Semiparametric Bayesian Model for Randomised Block Designs," *Biometrika*, 83, 275–285.

Chib, S., and Hamilton, B. H. (2002), "Semiparametric Bayes Analysis of Longitudinal Data Treatment Models," *Journal of Econometrics*, 110, 67–89.

Dahl, D. (2003), "An Improved Merge-Split Sampler for Conjugate Dirichlet Process Mixture Models," technical report, University of Wisconsin, Dept. of Statistics.

DeIorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004), "An ANOVA Model for Dependent Random Measures," *Journal of the American Statistical Association*, 99, 205–215.

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007), "Generalized Spatial Dirichlet Process Models," *Biometrika*, 94, 809–825.

Dunson, D. (2005), "Bayesian Semiparametric Isotonic Regression for Count Data," *Journal of the American Statistical Association*, 100, 618–627.

——— (2006), "Bayesian Dynamic Modeling of Latent Trait Distributions," *Biostatistics*, 7, 551–568.

Dunson, D. B., Herring, A. H., and Mulheri-Engel, S. A. (2007a), "Bayesian Selection and Clustering of Polymorphisms in Functionally-Related Genes," *Journal of the American Statistical Association*, 103, 534–546.

Dunson, D. B., Pillai, N., and Park, J.-H. (2007b), "Bayesian Density Regression," *Journal of the Royal Statistical Society*, Ser. B, 69, 163–183.

Escobar, M. D. (1994), "Estimating Normal Means With a Dirichlet Process Prior," *Journal of the American Statistical Association*, 89, 268–277.

Escobar, M. D., and 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.

——— (1974), "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 2, 615–629.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005), "Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing," *Journal of the American Statistical Association*, 100, 1021–1035.

Green, P., and Richardson, S. (2001), "Modelling Heterogeneity With and Without the Dirichlet Process," *Scandinavian Journal of Statistics*, 28, 355–375.

Griffin, J. E., and Steel, M. F. J. (2006), "Order-Based Dependent Dirichlet Processes," *Journal of the American Statistical Association*, 101, 179–194.

Hirano, K. (2002), "Semiparametric Bayesian Inference in Autoregressive Panel Data Models," *Econometrica*, 70, 781–799.

Ishwaran, H., and James, L. F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161–173.

——— (2002), "Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information," *Journal of Computational and Graphical Statistics*, 11, 508–532.

——— (2003), "Some Further Developments for Stick-Breaking Priors: Finite and Infinite Clustering and Classification," *Sankyā*, 65, 577–592.

Ishwaran, H., and Zarepour, M. (2002), "Dirichlet Prior Sieves in Finite Normal Mixtures," *Statistica Sinica*, 12, 941–963.

Jain, S., and Neal, R. M. (2000), "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model," technical report, University of Toronto, Dept. of Statistics.

Kacperczyk, M., Damien, P., and Walker, S. G. (2003), "A New Class of Bayesian Semiparametric Models With Applications to Option Pricing," technical report, University of Michigan Business School.

Kleinman, K., and Ibrahim, J. (1998), "A Semi-Parametric Bayesian Approach to Generalized Linear Mixed Models," *Statistics in Medicine*, 17, 2579–2596.

Kottas, A., Branco, M. D., and Gelfand, A. E. (2002), "A Nonparametric Bayesian Modeling Approach for Cytogenetic Dosimetry," *Biometrics*, 58, 593–600.

Lau, J. W., and Green, P. (2006), "Bayesian Model–Based Clustering Procedures," technical report, University of Bristol, Dept. of Mathematics.

Laws, D. J., and O'Hagan, A. (2002), "A Hierarchical Bayes Model for Multilocation Auditing," *Journal of the Royal Statistical Society*, Ser. D, 51, 431–450.

Lo, A. Y. (1984), "On a Class of Bayesian Nonparametric Estimates, I: Density Estimates," *The Annals of Statistics*, 12, 351–357.

Lo, A. Y., Brunner, L. J., and Chan, A. T. (1996), "Weighted Chinese Restaurant Processes and Bayesian Mixture Models," technical report, Hong Kong University of Science and Technology, ISMT Dept.

MacEachern, S. N. (1994), "Estimating Normal Means With a Conjugate-Style Dirichlet Process Prior," *Commnunications in Statistics, Part B—Simulation and Computation*, 23, 727–741.

——— (1999), "Dependent Nonparametric Processes," in *Proceedings of the Section on Bayesian Statistical Science*, American Statistical Association, pp. 50–55.

——— (2000), "Dependent Dirichlet Processes," technical report, Ohio State University, Dept. of Statistics.

MacEachern, S. N., and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–238.

McCloskey, J. W. (1965), "A Model for the Distribution of Individuals by Species in an Environment," unpublished doctoral thesis, Michigan State University.

Medvedovic, M., and Sivaganesan, S. (2002), "Bayesian Infinite Mixture Model-Based Clustering of Gene Expression Profiles," *Bioinformatics*, 18, 1194–1206.

Mukhopadhyay, S., and Gelfand, A. (1997), "Dirichlet Process Mixed Generalized Linear Models," *Journal of the American Statistical Association*, 92, 633–639.

Müller, P., Quintana, F., and Rosner, G. (2004), "Hierarchical Meta-Analysis Over Related Non-Parametric Bayesian Models," *Journal of the Royal Statistical Society*, Ser. B, 66, 735–749.

Neal, R. M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

Ongaro, A., and Cattaneo, C. (2004), "Discrete Random Probability Measures: A General Framework for Nonparametric Bayesian Inference," *Statistics and Probability Letters*, 67, 33–45.

Perman, M., Pitman, J., and Yor, M. (1992), "Size-Biased Sampling of Poisson Point Processes and Excursions," *Probability Theory and Related Fields*, 92, 145–158.

Pitman, J. (1996), "Some Developments of the Blackwell–Macqueen Urn Scheme," in *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, eds. T. S. Ferguson, L. S. Shapeley, and J. B. MacQueen, Hayward, CA: IMS, pp. 245–268.

Quintana, F., and Iglesias, P. L. (2003), "Bayesian Clustering and Product Partition Models," *Journal of the Royal Statistical Society*, Ser. B, 65, 557–574.

Roberts, G., and Papaspiliopoulos, O. (2008), "Retrospective Markov Chain Monte Carlo Methods for Dirichlet Process Hierarchical Models," *Biometrika*, 95, 169–186.
Rodriguez, A. (2007), "Some Advances in Bayesian Nonparametric Modeling," unpublished doctoral thesis, Duke University, Institute of Statistics and Decision Sciences.
Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.
Silverman, B. (1986), *Density Estimation*, London: Chapman & Hall.
Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006), "Sharing Clusters Among Related Groups: Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566–1581.
Tomlinson, G. (1999), "Analysis of Densities," unpublished doctoral thesis, University of Toronto, Dept. of Mathematics.
Verbeke, G., and Lesaffre, E. (1996), "A Linear Mixed-Effects Model With Heterogeneity in the Random-Effects Population," *Journal of the American Statistical Association*, 91, 217–221.

Comment

Daniel L. GILLEN and Wesley O. JOHNSON

1. INTRODUCTION

We commend Rodriguez, Dunson, and Gelfand (RDG) for presenting a novel and fascinating nonparametric approach to inference for data that are nested (e.g., hospital within state) and where two kinds of clustering are are anticipated (e.g., among states and among hospitals within clusters of states). In contrast, the HDP is implemented through a single draw from a DP followed by iid draws from a DP with that as its base, implying that every state forms a unique cluster of size 1. The contrast between these approaches is quite interesting. RDG focus mainly on use of the NDP as a "top-level" clustering tool (e.g., for clustering states). We begin our discussion by also considering the interpretation of regression effects in the NDP setting as well as other common approaches to the analysis of two-stage clustered data. In Section 3 we seek to reinforce the construction of the NDP, further emphasizing the difference between the HDP and the NDP. Finally, we conclude by considering possible extensions of the NDP to other inferential settings.

2. INTERPRETATION OF REGRESSION EFFECTS

RDG consider the problem of analyzing multicenter data and draw motivation from the deficiencies of what they consider the most common analytic approaches. The authors mention three approaches to modeling, including (A) pooling the data across centers, (B) stratifying by center and analyzing each separately, and (C) using a parametric hierarchical model to borrow information from "neighboring" centers. An alternative approach not mentioned is the use of generalized estimating equations (GEEs) (cf. Zeger and Liang 1986; Prentice 1988; Prentice and Zhao 1991) to provide inferences that account for clustering (hospitals within each state form natural clusters) without having to provide a full probability model for the data. Under mild regularity conditions, GEE methods as proposed by Zeger and Liang (1986) based on an independence working correlation structure and combined with the use of robust variance estimators lead to consistent estimators and asymptotically valid inference (Pepe and Anderson 1994). Inefficiency due to misspecification of the working correlation structure (relative to a fully parametric procedure assuming the correct probability model) is often compensated for by the robustness of the GEE methodology, which has made it a common tool used in analysis of correlated data. To the best of our knowledge, there is no modification of a GEE approach that handles all of the clustering issues raised by RDG, because the GEE methodology is designed not for cluster identification, but rather for parameter estimation and inference. To analyze data in which cluster membership is unknown in the GEE framework, one likely would need to use a two-stage modeling strategy, first using a clustering algorithm to define group membership and then implementing GEE for parameter estimation. We also are not aware of any Bayesian approaches that correspond to the GEE approach. The Bayesian approach relies on a full probability model for the data, which is not specified in the GEE approach.

RDG focus mainly on clustering aspects and consider only regression effects in their illustration, regarding them more as confounding factors requiring an adjusted analysis. Inevitably, the NDP model will be used in conjunction with additional modeling of covariate effects. RDG's model for their example is semiparametric and assumes that the effects of covariates on the mean response will be the same from one state to the next. In Section 4 we discuss the possibility of a fully nonparametric approach in which regression effects are allowed to vary from (top level) cluster to cluster.

In those cases where regression effects are important, in the premodeling stage questions can be asked about scientific goals for interpreting regression effects associated with the analysis of multicenter data. A standard question is whether it is most relevant to discern the importance of (a) marginal covariate effects across the population, (b) covariate effects conditional on state, or (c) the characteristics defined by any random effects in the model. Approach A, along with GEE, estimates marginal covariate effects, whereas approaches B and C, as well as the proposed NDP, estimate conditional effects. Although this distinction is of no consequence under a standard linear model, various authors have discussed the ramifications of the difference between marginal and conditional effects for nonlinear models (e.g., Neuhaus, Kalbfleisch, and Hauck 1991; Pendergast et al. 1996; Heagerty and Zeger 2000). Clearly, when deciding among the analytic approaches that might be considered, it is of utmost importance to first decide on the target of inference and then choose the methodology that best addresses the

Daniel L. Gillen is Associate Professor (E-mail: *dgillen@uci.edu*) and Wesley O. Johnson is Professor (E-mail: *wjohnson@ics.uci.edu*), Department of Statistics, University of California Irvine, Irvine, CA 92697-1250.

Journal of the American Statistical Association
September 2008, Vol. 103, No. 483, Theory and Methods
DOI 10.1198/016214508000000571

scientific question of interest. In the context of the example presented by RDG, if the end goal is to modify public health policy, then marginal covariate effects may be of interest, whereas conditional effects are important for individual patients and clinicians.

3. REVISITING THE DISTINCTION BETWEEN THE NESTED DIRICHLET PROCESS AND THE HIERARCHICAL DIRICHLET PROCESS

The major contribution of the NDP paradigm is the ability to allow observations to cluster both across "states" and within clusters. RDG point out that none of approaches A, B, or C (or GEE) has the capability of borrowing information in this way. A natural comparison is with the HDP, as RDG have done. Such a comparison is key to highlighting the NDP's flexibility. To further illustrate the differences between the HDP and NDP, we consider the two frameworks side by side in Figure 1, borrowing from RDG's notation. As RDG note, the clear distinction between the two procedures comes at the top level of clustering. With the HDP, a single discrete realization of G_0 is sampled, followed by iid draws from a DP with that as its base, thereby leading to distinct G_j's with probability 1, which implies that $\Pr[G_j = G'_j] = 0$ for all $j \neq j'$. Thus every state forms a unique cluster of size 1. In contrast, because the NDP involves selecting J iid draws from $\{G_k^* : k = 1, 2, \ldots\}$ according to the discrete distribution $\{\pi_k^* : k = 1, 2, \ldots\}$, it is possible to get repeats and thus to obtain $G_j = G'_j$, implying that centers may cluster together along with patients within clustered centers. The realized distribution of outcomes from all hospitals that are in the same cluster of hospitals with realization G_k^* is then the simple mixture of kernels $\int p(\cdot|\theta, \phi) G_k^*(d\theta) = \sum_l w_{lk^*} p(\cdot|\theta_{lk}^*, \phi)$, where point masses are the support of G_k^* and the mixing weights are the corresponding probabilities. By RDG's approximation, this is a finite sum, but if the total mass concentration β were small, then the number of terms in this sum with appreciable weight would be expected to be small. This leads to the possibility of detecting subgroups of hospitals that might correspond to a bump in either tail of the distribution. Hospitals associated with a bump in the upper (lower) tail would correspond to better (worse) outcomes, and perhaps why this is the case could be discovered by investigating what these hospitals have in common that might explain this bump. Thus, by using the stick-breaking representation of the DP, RDG have provided an elegant approach to highlighting the flexibility of the NDP in relation to the HDP. They also have provided a new tool for guiding postanalysis exploration.

4. POSSIBLE EXTENSIONS

RDG have provided a flexible way to borrow information through clustering by using the NDP. Clearly, many possible extensions lie ahead. A conceptually simple extension can model covariate information nonparametrically rather than semiparametrically. This would be accomplished, using the notation in the article, by letting $(\mu_{ij}, \gamma_{ij}, \sigma_{ij}) \sim G_j$ in section 5.2. Thus instead of a DPM that mixes on the intercept and scale, this approach also mixes on the regression slopes. This model then generalizes the hierarchical parametric random effects regression model where slopes and intercepts would be distinct (albeit correlated) within each state. It allows for each cluster of states to have its own regression model (actually a mixture of regression models) relating covariates to response. This model embeds a DDP regression model in the NDP (De Iorio, Müller, Rosner, and MacEachern 2004; De Iorio, Johnson, Müller, and Rosner 2008). Conditional on the clustering of states, outcomes within clusters will behave as realizations from a DDP regression rather than as realizations from a simple DP.

The covariate vector also could be partitioned into two parts, one with random coefficients embedded into the NDP as described here and the other treated as having the same effect across clusters of states, as in RDG's approach. Another variation of this theme would incorporate the possibility of a random function into the regression structure to handle the correlation structure associated with longitudinal data (cf. Zhang, Lin, Raz, and Sowers 1998). It is also clear that, at least conceptually, the NDP structure can be easily adapted to generalized linear model notation and thus in all likelihood can be used to, for example, cluster binary outcomes associated with hospitals within states and, of course, in the more general context of individuals within centers. Of course, the devil is in the details; perhaps this is easier said than done.

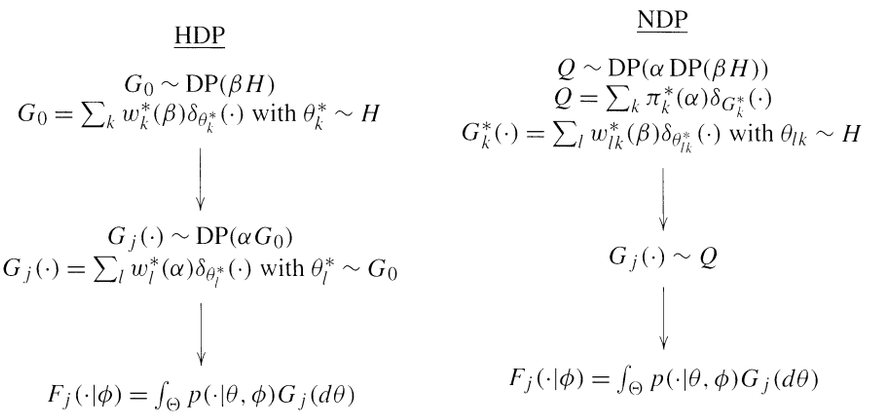

Figure 1. Contrasting the HDP and NDP paradigms.

Another obvious avenue is to extend the work here to the setting of censored survival data, where the analysis of multicenter data has focused primarily on frailty models (e.g., Clayton 1978; Lee, Wei, and Amato 1992; Gustafson 1997) and marginal models (e.g., Lin and Wei 1989; Wei, Lin, and Weissfeld 1989). There are many recent Bayesian semiparametric approaches to survival analysis, including those of Kuo and Mallick (1997), Kottas and Gelfand (2001), and Hanson and Johnson (2002), to name only a few. Extensions of this work that allow for NDP priors will make it possible to cluster hospitals and patients within clusters of hospitals based on their survivability. Clusters of hospitals with notably better survival prospects can be identified and studied, as can clusters of patients within clusters that are found to have appreciably better or worse survival. This would greatly enhance the flexibility of current methods for analyzing multivariate failure time data, and we look forward with excitement to the extensions of the NDP as presented by RDG.

ADDITIONAL REFERENCES

Clayton, D. G. (1978), "A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence," *Biometrika*, 65, 141–152.

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2008), "Bayesian Nonparametric Non-Proportional Hazards Survival Modeling," *Biometrics*, in press.

Gustafson, P. (1997), "Large Hierarchical Bayesian Analysis of Multivariate Survival Data," *Biometrics*, 53, 230–242.

Hanson, T., and Johnson, W. O. (2002), "Modeling Regression Error With a Mixture of Polya Trees," *Journal of the American Statistical Association*, 97, 1020–1033.

Heagerty, P. J., and Zeger, S. L. (2000), "Marginalized Multilevel Models and Likelihood Inference" (with discussion), *Statistical Science*, 15, 1–26.

Kottas, A., and Gelfand, A. E. (2001), "Bayesian Semiparametric Median Regression Modeling," *Journal of the American Statistical Association*, 96, 1458–1468.

Kuo, L., and Mallick, B. (1997), "Bayesian Semiparametric Inference for the Accelerated Failure-Time Model," *Canadian Journal of Statistics*, 25, 457–472.

Lee, E. W., Wei, L. J., and Amato, D. A. (1992), "Cox-Type Regression Analysis for Large Numbers of Small Groups of Correlated Failure Time Observations," in *Survival Analysis: State of the Art*, eds. J. P. Klein and P. K. Goel, Boston, MA: Kluwer Academic Publishers, pp. 237–247.

Lin, D. Y., and Wei, L. J. (1989), "The Robust Inference for the Cox Proportional Hazards Model," *Journal of the American Statistical Association*, 84, 1074–1078.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1991), "A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data," *International Statistical Review*, 59, 25–35.

Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., and Fisher, M. R. (1996), "A Survey of Methods for Analyzing Clustered Binary Response Data," *International Statistical Review*, 64, 89–118.

Pepe, M. S., and Anderson, G. L. (1994), "A Cautionary Note on Inference for Marginal Regression Models With Longitudinal Data and General Correlated Response Data," *Communications in Statistics, Part B—Simulation and Computation*, 23, 939–951.

Prentice, R. L. (1988), "Correlated Binary Regression With Covariates Specific to Each Binary Observation," *Biometrics*, 44, 1033–1048.

Prentice, R. L., and Zhao, L. P. (1991), "Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses," *Biometrics*, 47, 825–839.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989), "Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions," *Journal of the American Statistical Association*, 84, 1065–1073.

Zeger, S. L., and Liang, K.-Y. (1986), "Longitudinal Data Analysis for Discrete and Continuous Outcomes," *Biometrics*, 42, 121–130.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998), "Semiparametric Stochastic Mixed Models for Longitudinal Data," *Journal of the American Statistical Association*, 93, 710–719.

Comment

Peter MÜLLER and Luis NIETO-BARAJAS

Rodríguez, Dunson, and Gelfand (RDG) propose a probability model for a family of random probability models. The motivating application is inference for quality of care measurements in hospitals nested within states in the United States. Let G_j denote the distribution of quality of care measurements, θ_{ij}, for hospitals, $i = 1, \ldots, n_j$, in state j (ignoring for the moment an additional smoothing kernel and regression used in the article). States are clustered by defining a random partition of the state indices $\{1, \ldots, 50\}$. For the kth cluster of states a random distribution $G_k^\star(\theta)$ is generated by a DP prior, and $G_j \equiv G_k^\star$ for all states in the cluster. The random partition of states into clusters is defined by the Pólya urn scheme implied by another DP with total mass α.

We congratulate the authors for introducing an interesting new nonparametric Bayesian probability model. The authors correctly cite several recent articles that propose probability models for dependent random probability distributions by defining dependence on the locations or the weights in the stick-breaking representation of the DP. In contrast, RDG use clustering of the G_j (i.e., the members of the family of random distribution) to induce the desired correlation of $(G_j(A), G_{j'}(A))$. The underlying structure of related random distributions is common in many biomedical problems and often is ignored or simplified for technical convenience. We agree with RDG on the need for more flexible models to address such inference problems and feel that the proposed approach addresses this need. We have a few points to add.

RDG present their approach as a distribution on the space of distributions on distributions. But we see a distribution only on distributions, that is, an element *in* the space of distributions on distributions. This is easily verified by noting that the argument of G_j in eq. (2) is θ, not a random distribution. In other words, if the location of the parentheses in eq. (2) were changed to $G_j(\bullet) = \sum \pi_k^\star \delta_{G_j^\star}(\bullet)$, then the argument $\bullet$ would stand for a set of random measures and G_j in fact would be a distribution on the space of distributions on distributions, as stated.

Peter Müller is Professor, Department of Biostatistics, University of Texas M. D. Anderson Cancer Center, Houston, TX 77030 (E-mail: *pm@odin.mdacc.tmc.edu*). Luis Nieto-Barajas is Professor, Department of Biostatistics, University of Texas M. D. Anderson Cancer Center, Houston, TX 77030 and ITAM, Mexico City, Mexico (E-mail: *lnieto@itam.mx*).

Journal of the American Statistical Association
September 2008, Vol. 103, No. 483, Theory and Methods
DOI 10.1198/016214508000000599

Such a super-process has been defined by, for example, Nieto-Barajas and Walker (2007). Details of notation apart, we also feel that from a data analysis perspective, the model is better described as random clustering of a set of random distributions. This notion is supported by the illustrative examples in the article. In both examples RDG report inference for each random G_j, but not for the joint distribution of the random probability measures. In the latter case, they report summaries of co-clustering probabilities. The nature of the proposed model as a random partition model is further highlighted by the following observation. In the example, the conjugate choice of the kernel $p(\cdot \mid \theta, \phi)$ and base measure H allow us to analytically integrate out θ conditional on given values of the membership indicators ξ and ζ, leaving a probability model on the random partitions only.

We appreciate the comments about limitations of sharing information across related random measures by a regression on the level of hyperparameters that index the random measures only. For example, in the case of DP priors for random measures $G_j \sim \mathrm{DP}(\alpha G_{j0})$, linking the random measures by, for example, assuming that $G_{j0} = \mathrm{N}(a + bx_j, \sigma^2)$ restricts the nature of the possible borrowing of strength. But some of the early work on dependent DP models based on such constructions should be credited, perhaps to Cifarelli and Regazzini (1978), Muliere and Petrone (1993), or Mira and Petrone (1996).

Finally, we cannot see why posterior simulation requires the use of truncations and the implied approximations. Consider the two sets of latent cluster membership indicators, ξ_{ij} and ζ_j, defined in section 4. Let $S_k = \{(i, j) \text{ with } \zeta_j = k\}$. The conditional prior $p(\xi_{ij}, (i, j) \in S_k \mid \zeta)$ is the usual Pólya urn scheme that describes the random partition induced by a DP with total mass parameter β, and similarly for $p(\zeta_j, j = 1, \ldots, J)$ for total mass parameter α. Conditional on ζ, updating ξ reduces to the usual inference for a DP random measure, as RDG comment in the supplemental materials. As RDG point out, evaluating the marginal $p(\theta \mid \zeta)$ involves a prohibitive computational effort. But this would not seem to be necessary. Using the simple closed-form expressions for $p(\zeta)$ and $p(\xi \mid \zeta)$, we can evaluate the joint prior of any proposed set of parameters (θ is marginalized out analytically). This allows construction of a Metropolis–Hastings proposal to update ζ, proceeding similarly to the approach proposed by Jain and Neal (2004). The proposal must be a joint proposal for (ζ, ξ). For example, one could consider split and merge moves of the following type. Let $m_k = |\{j : \zeta_j = k\}|$ denote the size of the kth cluster of states, and let $\tilde{x}$ denote the value of the quantity x in the proposed move. Select an index, $\tilde{j} \in \{1, \ldots, J\}$. If $\tilde{j}$ is a singleton (i.e., $m_{\zeta_{\tilde{j}}} = 1$), then randomly select another cluster $\tilde{k} \neq \zeta_{\tilde{j}}$ and propose $\tilde{\zeta}_{\tilde{j}} = \tilde{k}$, and add the distinct point masses in $\{\theta_{i\tilde{j}}\}$ to the set of unique point masses generated from $G_{\tilde{k}}$. The latter is done by setting $\tilde{S}_{\tilde{k}} = S_{\tilde{k}} \cup \{(i, j), j = \tilde{j}\}$. If $\tilde{j}$ is not a singleton, then propose the opposite move. The prior, the likelihood (marginalizing with respect to θ) and the proposal probability can all be easily evaluated, thus facilitating the practical implementation of such a transition probability. The many housekeeping details make it sound more difficult than it actually is. Even with a different top-level sampling model that might preclude analytic marginalization with respect to θ, one could implement such moves without any complications beyond what a traditional DP mixture would require. In particular, the described Metropolis–Hastings move for ζ_j would not require reversible jump.

In summary, we congratulate RDG on a stimulating and very interesting discussion that highlights the flexibility of nonparametric Bayesian inference. We welcome the NDP as a happy new member in the rapidly growing xyz-DP family.

ADDITIONAL REFERENCES

Cifarelli, D., and Regazzini, E. (1978), "Problemi Statistici Non-Parametrici in Condizioni di Scambialbilita Parziale e Impiego di Medie Associative," technical report, Quaderni Istituto Matematica Finanziaria, Torino.

Jain, S., and 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.

Mira, A., and Petrone, S. (1996), "Bayesian Hierarchical Nonparametric Inference for Change-Point Problems," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 693–703.

Muliere, P., and Petrone, S. (1993), "A Bayesian Predictive Approach to Sequential Search for an Optimal Dose: Parametric and Nonparametric Models," *Journal of the Italian Statistical Society*, 2, 349–364.

Nieto-Barajas, L. E., and Walker, S. G. (2007), "Gibbs and Autoregressive Markov Processes," *Statistics and Probability Letters*, 77, 1479–1485.

Comment

Kaushik GHOSH, Pulak GHOSH, and Ram C. TIWARI

First, we would like to take this opportunity to congratulate Rodriguez, Dunson, and Gelfand (henceforth RDG) for a very interesting article. The article provides a novel and important contribution to the nonparametric Bayesian literature, and the methodology presented therein will prove useful in many disciplines. In our discussion of the article, we first emphasize its usefulness by providing an application of the NDP to prediction of cancer mortality. We then comment on some other aspects of the model.

Kaushik Ghosh is Assistant Professor, Department of Mathematical Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154-4020 (E-mail: *kaushik.ghosh@unlv.edu*). Pulak Ghosh is Associate Professor, Department of Biostatistics and Winship Cancer Institute, Emory University, Atlanta, GA 30322 (E-mail: *pulakghosh@gmail.com*). Ram C. Tiwari is Associate Director, Office of Biostatistics, Center for Drug Evaluation & Research, FDA, Silver Spring, MD 20993-0002 (E-mail: *ram.tiwari@fda.hhs.gov*). Ram C. Tiwari's research was conducted while employed by the National Cancer Institute. Pulak Ghosh's research was supported in part by National Institutes of Health contract 263-MQ-514250.

In the Public Domain
Journal of the American Statistical Association
September 2008, Vol. 103, No. 483, Theory and Methods
DOI 10.1198/016214508000000562

1. APPLICATION TO PREDICTION OF CANCER MORTALITY

Ghosh and Tiwari (2007) used a semiparametric Bayesian method to predict U.S. cancer mortality counts at the national level. The method used aggregated mortality data available from the previous years to obtain a 3-year-ahead prediction for the current year. One natural question to ask is how one would proceed when more detailed data (e.g., state-level mortality data) are available. It would seem that a better job could be done by using the state-specific data to obtain predictions for the individual states and then aggregating them to get the overall prediction for the United States, instead of predicting the United States as a whole by ignoring state information. In doing so, it would be natural to combine the information from "similar" states to get a better prediction; however, the selection of those states needs to be data-driven. A nice application of NDP to address this problem is described below:

Let d_{sj} denote the number of deaths from a common cancer in state s at time j ($s = 1, \ldots, S$; $j = 1, \ldots, J$). Following Ghosh and Tiwari (2007), we assume the local quadratic model,

$$d_{s,j+1} | (\mathbf{d}_{sj}, \boldsymbol{\beta}_{sj}, \boldsymbol{\gamma}_{sj}, \sigma^2) \overset{\text{indep.}}{\sim} \mathrm{N}(d_{sj} + \beta_{sj} + \gamma_{sj}, \sigma^2), \quad (1)$$

$$\beta_{s,j+1} | (\boldsymbol{\beta}_{sj}, \boldsymbol{\gamma}_{sj}) \overset{\text{indep.}}{\sim} \mathrm{N}(\beta_{sj} + 2\gamma_{sj}, \kappa\sigma^2), \quad (2)$$

$$\gamma_{sj} \overset{\text{iid}}{\sim} G_s, \quad (3)$$

$$G_1, \ldots, G_S \overset{\text{iid}}{\sim} \mathrm{nDP}(\alpha, \eta, H), \quad (4)$$

$$H \sim \mathrm{N}(\zeta, \rho^2), \quad (5)$$

where γ_{sj} is the instantaneous "acceleration" of mortality counts in state s at time j and $\mathbf{d}_{sj} = (d_{s1}, \ldots, d_{sj})'$ is the vector of mortality counts in state s up to time j. The vectors $\boldsymbol{\beta}_{sj}$ and $\boldsymbol{\gamma}_{sj}$ are defined similarly. This approach to modeling amounts to having states with the same acceleration distribution cluster together. We also follow Ghosh and Tiwari (2007) in assigning prior distributions. In particular, $\beta_{s0} \sim \mathrm{N}(0, 100)$, $\kappa \sim \mathrm{IG}(11.0, 10.0)$, $\zeta \sim \mathrm{N}(0, 1)$, and $\rho^2 \sim \mathrm{IG}(10.2, 10.1)$. Finally, $\alpha, \eta \overset{\text{indep.}}{\sim} \mathrm{gamma}(.1, .1)$.

The model was easily implemented in WinBUGS using the finite truncation approximation of NDP suggested in the article. Table 1 presents the results of 3-year-ahead predictions of lung cancer mortality of males for the years 2002–2004, for five selected states. Thus the 2002 predictions are based on mortality data from 1969–1999, the 2003 predictions are based on data from 1969–2000, and so on. The predictions are based on mortality data obtained from the National Center for Health Statistics (NCHS), which is also available from the National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) program. The two sets of numbers in the "predicted" column were obtained using NDP for the unparenthesized version and standard DP for the parenthesized version.

Using the NDP has resulted in sharing of information among states with similar mortality profiles, thereby improving the resulting predictions. The logarithm of mean predicted squared error is 14.79 using NDP and 18.18 using standard DP. Clearly, using NDP results in better predictions, most likely due to sharing of information among states that have clustered together.

In fact, for prediction of the 2004 figures, the 51 states (including the District of Columbia) clustered into 13 groups, as shown in Table 2. The cluster structure was obtained using the method of Dahl (2006). An in-depth analysis of the prediction of cancer mortality counts and rates for the current calendar year using NDP described here will be presented in a separate article.

Table 2. Clustering of the 50 states and the District of Columbia based on lung cancer mortality in males, 1969–2001

Cluster	Members
1	AL, AR, DC, KS, ME, OH, OR, SC, VT, VA
2	AK, CA, IL, KY, NC
3	AZ, CT, IA, NY, WV
4	CO, ND, OK
5	DE, MI
6	FL, NV, WI
7	GA, MD, MN, TX
8	HI, MS
9	ID, MO, NM
10	IN, NH, PA, SD, TN, WA
11	LA, MT, NE, RI
12	MA, UT
13	NJ, WY

2. OTHER COMMENTS

2.1 Precision Parameters

The clustering of the groups (states) is governed by the precision parameter α and within-group clustering is governed by the precision parameter η in $\mathrm{NDP}(\alpha, \eta, H)$. In many cases this is too strong a restriction to impose, because each group need not have the same η for within-group clustering. For example, in the analysis presented here, we would have liked to assign

Table 1. Three-year-ahead predictions for lung cancer mortality in U.S. males, 2002–2004

	2002		2003		2004	
States	Observed	Predicted	Observed	Predicted	Observed	Predicted
California	7,451	$8{,}129_{(9{,}130)}$	7,259	$7{,}732_{(8.711)}$	7,150	$7{,}531_{(8{,}413)}$
Utah	233	$225.3_{(202)}$	253	$229_{(198.7)}$	278	$249_{(210)}$
Michigan	3,218	$3{,}154_{(3{,}371)}$	3,195	$3{,}179_{(3{,}220)}$	3,290	$3{,}324_{(3{,}490)}$
Georgia	2,639	$2{,}674_{(2{,}891)}$	2,521	$2{,}619_{(2{,}832)}$	2,672	$2{,}741_{(2{,}806)}$
New York	5,180	$5{,}214_{(5{,}293)}$	5,093	$5{,}191_{(5{,}276)}$	4,980	$5{,}061_{(5{,}208)}$

NOTE: The numbers in parentheses were obtained using the standard DP.

different η's to the different states. We feel that this aspect of generalization of NDP will be quite useful for practical purposes and merits further study.

2.2 Clustering

As RDG point out, NDP provides a methodology to cluster groups and observations within groups simultaneously. This is done under the a priori assumption that the distributions (of the groups) are exchangeable, however. For example, the clusters given in Table 2 were obtained based on the distributions of the "acceleration" parameters of the year-to-year mortality counts, assuming exchangeability between states. This may not be realistic, when, for instance, one has additional information that must be incorporated into the prior structure. For example, California and Oregon are geographically and demographically closer than, say, California and Kentucky, but the latter are clustered together based on the observed data. The proposed NDP framework does not allow us to incorporate such information a priori.

2.3 Computations

The current NDP implementation is based on a finite truncation approach to the stick-breaking representation of DPs first presented by Sethuraman and Tiwari (1982). The advantage of this representation is that it can be easily implemented using standard software like WinBUGS, even in the presence of nonconjugate priors. The procedure has been shown to closely approximate the true distribution when K and L are large. It is interesting to note that $K = 35$ and $L = 55$ provide a good approximation as long as $n \leq 500$ and $J \leq 50$. (The truncation points do not depend on the number of groups and the number of observations per group.)

2.4 Convergence

Suppose that we have a sequence of NDP's given by NDP(α_r, η_s, H), where $\alpha_r \to 0$ as $r \to \infty$ and $\eta_s \to 0$ as $s \to \infty$. Using the results of Sethuraman and Tiwari (1982), we can examine the convergence of the process NDP(α_r, η_s, H), its posterior, and Bayes estimators of functionals of the posterior.

ADDITIONAL REFERENCES

Dahl, D. (2006), "Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model," in *Bayesian Inference for Gene Expression and Proteomics*, eds. K.-A. Do, P. Müller, and M. Vannucci, New York: Cambridge University Press, pp. 201–218.

Ghosh, K., and Tiwari, R. C. (2007), "Prediction of U.S. Cancer Mortality Counts Using Semiparametric Bayesian Techniques," *Journal of the American Statistical Association*, 102, 7–15.

Sethuraman, J., and Tiwari, R. C. (1982), "Convergence of Dirichlet Measures and the Interpretation of Their Parameter," in *Statistical Decision Theory and Related Topics III*, Vol. 2, eds. S. S. Gupta and J. O. Berger, New York: Academic Press, pp. 305–315.

Comment

Steven N. MACEACHERN

Rodriguez, Dunson, and Gelfand have written an interesting article that provides a neat answer to a recurring question: Given a set of observations from each of a number of distributions, can the distributions be clustered in some flexible, sensible, data-based, and non–ad hoc fashion? The authors answer in the affirmative, constructing a coherent model and addressing clustering inferentially. Because the model is coherent, the full range of Bayesian inferences can be made.

The coherent model, the nested Dirichlet process (NDP), fits nicely in the array of nonparametric (or semiparametric) Bayesian procedures being energetically developed. Like many processes, the NDP is perhaps most naturally constructed as a Dirichlet process (DP) or dependent Dirichlet process (DDP). To do so, one merely notes that the distribution G_k^* is defined by a countable collection of random variables $(u_{lk}^*, \theta_{lk}^*, l = 1, 2, \ldots)$ and applies the usual Dirichlet methods to a countable, rather than a finite, vector. It is through the lens of the DDP that I look at the NDP.

The DDP and its extension to more general forms (e.g., MacEachern, Kottas, and Gelfand 2001) were developed to provide a latent structure for the hierarchical Bayesian model with greater support than the DP. The latent structure replaces a simpler structure, such as a DP, with a collection of (dependent) nonparametric distributions. These distributions are indexed by what we call a covariate and formally comprise a distribution-valued stochastic process. This allows one to incorporate covariates in a direct fashion, specifying the marginal prior distribution at each value of the covariate and also the joint prior distribution as a function of the covariate. In many instances, the latent DDP is a DP (depending on which definition of the DP is used); however, its use for modeling tends to be quite different.

Like any latent modeling, DDP modeling requires a connection between the latent structure and the observed data. The key is how this connection is made. A direct, general recipe (connected to the DDP in MacEachern 2007) is to use a selector surface that is an integer-valued stochastic process with index set equal to the covariate space. Marginally, at a specific covariate value, the distribution over the integers matches the distribution over components of the countable mixture given by the DDP at that covariate value.

The selector surface determines a range of joint behaviors for the observable data. Assume that the observations can be partitioned into conditionally independent groups. Each group of observations is associated with a selector surface. Focus on

Steven N. MacEachern is Professor, Department of Statistics, Ohio State University, Columbus, OH 43210-1247 (E-mail: *snm@stat.ohio-state.edu*).

Journal of the American Statistical Association
September 2008, Vol. 103, No. 483, Theory and Methods
DOI 10.1198/016214508000000607

a single group. Evaluating the selector surface determines the components of the mixture with which individual observations are associated; the integer value of the selector surface at a covariate value gives the component of the mixture for the observation with that covariate value. At one extreme, the group may consist of a single observation and so be connected to only one value of the covariate. In this case the marginal distribution of the selector surface at the (single) covariate value determines the distribution of the observation and the joint behavior of the selector surface is unimportant. Traditional regression modeling (MacEachern 2001) is an example of this case. Alternatively, the group may consist of more than one observation, with different observations associated with different covariate values. With a single-p DDP, the dependence in the group can be as simple as selecting a single mixture component for all observations. (In the single-p DDP, the mixture weights do not vary with covariate value.) To achieve this, the realization of the selector surface must be constant across all values of the covariate. Gelfand, Kottas, and MacEachern (2005) described this use of the DDP as a spatial Dirichlet process. There are many possibilities between the extremes of a constant selector surface and an arbitrarily jointly defined selector surface. The properties of the selector surface play important roles in determining properties of the observable vectors.

The NDP model gives rise to distributions that are either identical or are conditionally independent. If distribution G_j is drawn from atom l of the process and distribution $G_{j'}$ is drawn from atom l', with $l \neq l'$, then the distributions are conditionally independent draws from a Dirichlet process with base measure βH. If instead, $l = l'$, then $G_j = G_{j'}$. This is a strong assumption, because in many contexts we believe that all of the F_j's (and thus all of the G_j's) differ. In such settings, the primary question may be whether a pair of distributions differ greatly or differ only slightly.

We are all comfortable with approximations in modeling. As such, an important question becomes: "Under what conditions is the difference between similarity and identity likely to be have an impact on clustering?" One would expect this problem to become greater with large sample sizes (large n_j), where the data have the ability to distinguish relatively minor differences in distributions. In a similar vein, dimensionality of the response also is important. High-dimensional distributions are nearly guaranteed to differ, and individual observations are unlikely to be close to one another.

The inferential approach to clustering, with use of the tuning parameter, has scope for remediating this problem to some extent; however, in many applications there will be substantial variation among the n_j. In my experience with the National Marrow Donor Program, the number of patients treated varies considerably by hospital. My impression is that such variation is common in the patients within centers setting. Where there is substantial variation among the n_j's, and thus differing amounts of information about identity/nonidentity of pairs of distributions, I suspect that tuning cannot fully adjust the inference. To me, it seems more natural to adjust the model.

DDP modeling suggests that the model be adjusted by breaking the "deterministic" dependence across levels of the covariate and replacing it with a weaker dependence. The covariate in the NDP is j, assuming values in the set $\{1, \ldots, J\}$. In the DDP modification, the "atoms" θ^*_{lk} are replaced with processes θ^*_{lkj}, where j indexes the covariate. In this context, j lives in a discrete space with no suggestion that particular pairs of j's should be treated differently than other pairs of j's. Thus $\theta^*_{lk1}, \ldots, \theta^*_{lkJ}$ naturally would be treated as exchangeable. For scalar, normally distributed θ^*_{lk}'s, a natural replacement would be multivariate normal vectors with a common off-diagonal covariance. In other contexts, j might live in a covariate space for which the strength of relationship between distributions decays as distance increases. In such cases closer covariate values would be modeled as having stronger dependence between their θ^*_{lkj} values. This modification also allows one to embed the J distributions in a continuous covariate space. The values of θ^*_{lk} at an unobserved covariate value would have distributions depending on their position in the covariate space.

Modification of the NDP by changing the θ^*_{lk} leads to replacement of the latent distributions in (3) with $G^*_{kj}(\cdot) = \sum_{l=1}^{\infty} w^*_{lk} \delta_{\theta^*_{lkj}}(\cdot)$. As in the NDP, if the distributions G_j and $G_{j'}$ are from different atoms, then they are conditionally independent draws from a DP. But if they are from the same atom, then the distributions are not identical, but merely similar, sharing the same mixing weights w^*_{lk}. The atoms in the mixture will be similar, with $\theta^*_{lkj} \approx \theta^*_{lkj'}$, and so the distributions of observables also will be similar. If desired, the mixing weights also can be allowed to vary with j. The similarity or difference in the G_{kj}'s feeds through to the distributions $\{F_1, \ldots, F_J\}$ through the convolution in (1).

There is a long tradition in our discipline of assessing whether distributions are identical or different. Old-style, heavily assumptionized ANOVA, where equal means implies equal distributions, falls under this heading. The authors focus on this problem, but with much more sophisticated models that capture important features of the distributions.

Traditionally, and with enough assumptions, the identity of distributions hinges on whether or not a particular parameter (say an additive effect for a treatment mean) is 0. With multivariate observables, such parameters also may represent conditional independence between subsets of variables. Interest then focuses on whether the same conditional independence holds across different centers or whether the conditional independences differ. Conditional independence plays an important role in our understanding of collections of variates, and much of the work on causal inference focuses on a description of conditional independence. Starting from this point, the NDP can be adjusted to address this sort of inferential question in a fully nonparametric context.

To impose conditional independence between components of the observable vector, replace $\mathrm{DP}(\beta H)$ in the NDP model with the product of two independent DPs, as is done in, for example, Bush's (1994) development of the nonparametric Bayesian mixed model and used by Bush, Lee, and MacEachern (2007) in the context of multiple comparisons. Thus, extending the authors' notation, we may replace $\mathrm{DP}(\beta H)$ with $\mathrm{DP}(\beta_1 H_1)\,\mathrm{DP}(\beta_2 H_2)$, where the vector θ is partitioned into two components, θ_1 and θ_2. A distribution G^*_k is determined by the pair of component distributions, $G^*_{1k}(\cdot) = \sum_{l=1}^{\infty} w^*_{1lk} \delta_{\theta^*_{1lk}}(\cdot)$ and $G^*_{2k}(\cdot) = \sum_{l=1}^{\infty} w^*_{2lk} \delta_{\theta^*_{2lk}}(\cdot)$. The mass assigned by G^*_k to a vector θ is the product of the masses assigned to its components by G^*_{1k} and G^*_{2k}. The resulting distribution enforces conditional

independence of θ_1 and θ_2. If similarity rather than identity is desired, then θ^*_{1lk} and θ^*_{2lk} can be indexed by the covariate, as described earlier. The distribution is determined by a countable collection of random vectors with the iid structure that makes the DP/DDP so appealing. Conditional independence of θ_1 and θ_2 is easily extended to conditional independence of a pair of observable components of a vector by enforcing dependence on θ_1 alone and θ_2 alone on the respective components.

To address the question of whether a particular split of θ has or does not have conditional independence, the NDP can be tweaked. The $\mathrm{DP}(\beta H)$ is replaced by a structure that draws two distributions for each G^*_k, one coming from the $\mathrm{DP}(\beta H)$ and the other coming from the product $\mathrm{DP}(\beta_1 H_1)\,\mathrm{DP}(\beta_2 H_2)$. Thus G^*_k is replaced by a pair of distributions: The first will result in dependent components θ_1 and θ_2, whereas the second will result in independent components. Data at a particular center will be attached to only one of these distributions. The choice can be formalized either as an unobserved covariate or in terms of a selector. Clustering would then focus on this covariate/selector. The resulting analysis would identify groups of distributions (groups of centers) for which the same conditional independences hold. Clearly, the notion of conditional independence given earlier extends to more than two conditionally independent components of θ; it also extends to cases where the portions of θ that go into θ_1 and θ_2 are unknown.

Variations on this theme allow one to align the models with core modeling concepts. For example, the traditional distinction between explanatory and response variates might lead one to match portions of the two types of distributions. G^*_{1k}, the explanatory portion, could be identical for both the dependent and independent versions of G^*_k, whereas G^*_{2k}, the response portion, would either allow dependence of θ_2 on θ_1 or enforce independence. Data from a particular center would be tied to either the first version of G^*_k or the second version of G^*_k.

ADDITIONAL REFERENCES

Bush, C. A. (1994), "Semi-Parametric Bayesian Linear Models," unpublished doctoral dissertation, Ohio State University, Dept. of Statistics.

Bush, C. A., Lee, J., and MacEachern, S. N. (2007), "Minimally Informative Nonparametric Bayesian Analysis, With Application to Multiple Comparisons," technical report, Ohio State University, Dept. of Statistics.

MacEachern, S. N. (2001), "Decision-Theoretic Aspects of Dependent Nonparametric Processes," in *Bayesian Methods With Applications to Science, Policy and Official Statistics*, ed. E. I. George, pp. 551–560, available at *http://www.stat.cmu.edu/ISBA*.

——— (2007), Discussion of "Bayesian Nonparametric Modelling for Spatial Data Using Dirichlet Processes," by A. E. Gelfand, M. Guindini, and S. Petrone, in *Bayesian Statistics 8*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxford, U.K.: Oxford University Press, pp. 196–198.

MacEachern, S. N., Kottas, A., and Gelfand, A. E. (2001), "Spatial Nonparametric Bayesian Models," in *Proceedings of the Joint Statistical Meetings*, American Statistical Association.

Comment

Lancelot F. JAMES

It is my pleasure to comment on this article by Abel Rodríguez, David Dunson, and Alan Gelfand. The authors present an interesting application of random probability measures on quite abstract spaces. As one who is often asked why discuss Polish spaces and the like, I am happy that the authors have provided a concrete and practically useful set of examples. I am most pleased to see yet another interesting and creative application of somewhat nonstandard Bayesian nonparametric ideas.

I take the path here of expanding on the basic idea of the authors. First, I offer an equivalent (albeit slightly more general) description of the NDP. Then I discuss, and pose a question on, the usage of a finite-dimensional Dirichlet process. Finally, I discuss generalizations of the NDP that still embodies the idea of the article under discussion. In particular, I end with a description of a hybrid NDP–HDP model showing that these two relatively new ideas can be combined. I also want to note that although can one obtain some nice simplification by using stick-breaking representations, doing so is not a necessity. One can do other things that allow for the use of wider classes of processes. I refrain from elaborating on this last point, however.

1. THE NESTED DIRICHLET PROCESS IN TWO STAGES

The procedure considered by the authors is in fact quite simple to describe. Following the authors' exposition, let us consider $j = 1, \ldots, J$ objects or classes, each associated with samples of size n_j, $j = 1, \ldots, J$. The NDP can be viewed as a repetition of a two-stage procedure where in stage I these J objects are randomly assigned to, say, $N(J)$ nonempty classes out of a possible $N \le \infty$ classes. Once these classes are created, one is left with $N(J)$ independent DP hierarchical mixture models where in stage II, standard MCMC or other computational procedures can be applied (see, e.g., Ishwaran and James 2004, sec. 5).

The NDP rests on the introduction of the random probability measures

$$G_j(\cdot) \sim Q \equiv \sum_{k=1}^{N} \pi^*_k \delta_{G^*_k(\cdot)},$$

where $G^*_k(\cdot)$ are iid Dirichlet processes with total mass parameter β and such that

$$\mathbb{E}[G^*_k(\cdot)] = H(\cdot).$$

Lancelot F. James is Professor, Department of Informations Systems, Business Statistics and Operations Management, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR (E-mail: *lancelot@ust.hk*). This work was supported by HKSAR grants RGC-HKUST 600907, SBI06/07.BM14, and RGC-HKUST 6159/02P.

Journal of the American Statistical Association
September 2008, Vol. 103, No. 483, Theory and Methods
DOI 10.1198/016214508000000580

The sequence (π_k^*) corresponds to the stick-breaking weights for a DP with total mass parameter α as described by the authors. Here I allow N to be possibly less that ∞, whereas the authors are concerned primarily with $N = \infty$, which would allow applications to many interesting finite mixture models and also agrees with the process used in the MCMC procedure described in the article.

I use an equivalent description, which can be deduced from the discussion of Ishwaran and James (2001), where

$$G_j(\cdot) \stackrel{d}{=} G^*_{\zeta_j}(\cdot)$$

and $(\zeta_1, \ldots, \zeta_J)$ are the classification variables such that, conditional on the sequence (π_k^*),

$$\zeta_j \stackrel{\text{iid}}{\sim} \sum_{k=1}^{N} \pi_k^* \delta_k(\cdot).$$

2. FINITE–DIMENSIONAL DIRICHLET VERSUS STICKBREAKING

In previous work, it has been suggested that one could use a finite-dimensional Dirichlet vector $(D_1, \ldots, D_n)$ with parameters $(\alpha/N, \ldots, \alpha/N)$ in place of the truncated stick-breaking weights $(\pi_1^*, \ldots, \pi_N^*)$. I wonder whether the authors have attempted to use these variables as substitutes in their MCMC procedures? In general, I am quite curious as to how this procedure might perform. It is indeed unfortunate that for the weights $(D_1, \ldots, D_N)$, we do not have the precise error bounds that we can get from the stick-breaking representation. But in practice, Ishwaran and James (2001) noted very little difference between the two choices of weights for a rather moderate choice of N. Furthermore, Ishwaran and Zarepour (2002) and Ishwaran, James, and Sun (2001) have provided some interesting analyses related to sieve models and model selection problems involving finite-dimensional Dirichlet vectors. An interesting comparison of the usage of finite-dimensional Dirichlet vectors and the stick-breaking weights has been given by Kurihara, Welling, and Teh (2007).

3. A FIRST EXTENSION: MORE GENERAL RANDOM ATOMS

In previous work (Ishwaran and James 2003), we discussed the idea of constructing stick-breaking type processes that would assign non-iid distributions to each class and explored how this might be relevant in classification problems. The framework of the NDP can be easily modified to achieve this. In fact, the NDP can be viewed as a special case of a class of nested random probability measures where the (π_k^*) could be replaced by another sequence of consistent random probabilities and the $(G_k^*(\cdot))$ could be replaced by independent random probability measures with various laws. For concreteness, one can imagine constructing each $G_k^*(\cdot)$ to be an independent two-parameter Poisson Dirichlet process, otherwise known as a Pitman–Yor process (so named in Ishwaran and James 2001), with parameters $0 \le \gamma_k < 1$ and $\theta_k > -\gamma_k$, for $k = 1, \ldots, N$, and otherwise depending on H in the usual way. Let us say that the law of $G_k^*(\cdot)$ is $\mathcal{PY}(\gamma_k, \theta_k, H)$, that is,

$$G_k^*(\cdot) \sim \mathcal{PY}(\gamma_k, \theta_k, H).$$

When $\gamma_k = 0$, the random probability measures $G_k^*(\cdot)$ reduce to independent DPs with total mass parameter θ_k. As is now quite well known, the general Pitman–Yor class of random probability measures has many desirable features that lend themselves easily to practical usage. But perhaps the most interesting aspect is that it creates classes with possibly quite different clustering behavior than the DP. Similar to the DP, this clustering behavior is produced by drawing, say, $\tilde{n}_l$ exchangeable values that, conditional on $G_l^*(\cdot)$, are iid $G_l^*(\cdot)$, and clustering their indexes according to the ties in the sample.

To be specific, suppose that indexes l and k were picked through the $(\zeta_1, \ldots, \zeta_J)$, forming classes with respective sizes

$$\tilde{n}_l = \sum_{j=1}^{J} n_j \mathbb{I}(\zeta_j = l) \qquad \text{and} \qquad \tilde{n}_k = \sum_{j=1}^{J} n_j \mathbb{I}(\zeta_j = k),$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Then if γ_k and γ_l were positive, we could sample exchangeable random vectors of size $\tilde{n}_l$ and $\tilde{n}_k$ obtained from conditionally iid samples from $G_l^*(\cdot)$ and $G_k^*(\cdot)$. These samples would be such that the number of distinct clusters would exhibit power law behavior of order $\tilde{n}_k^{\gamma_k}$ and $\tilde{n}_l^{\gamma_l}$. Specifically, the powers are γ_l and γ_k. If, on the other hand, $\gamma_k = 0$, then the number of distinct clusters would be $\theta_k \log(\tilde{n}_k)$, which is the known logarithmic behavior of samples drawn from a DP. Naturally, in the nested scheme this behavior holds for a given sequence of $(\zeta_1, \ldots, \zeta_J)$.

The ability to produce quite different clustering behavior (particularly power law behavior) is an important aspect of the Pitman–Yor process that has not been exploited much in the applied literature. But recently Goldwater, Griffiths, and Johnson (2006) and Teh (2006) have argued that the power law behavior, in terms of numbers of distinct words, induced by general Pitman–Yor processes is appropriate when applied to various natural language models. In addition, they have argued that the logarithmic behavior produced by the DP appears to be quite inappropriate for such models.

Incidentally, the covariance structure for these $G_j(\cdot)$ is given by

$$\operatorname{cov}\bigl(G^*_{\zeta_j}(A) G^*_{\zeta_i}(A)\bigr) = \left(\sum_{k=1}^{N} \mathbb{E}[(\pi_k^*)^2] \frac{1-\gamma_k}{1+\theta_k}\right) H(A)[1 - H(A)].$$

4. HYBRID NESTED DIRICHLET PROCESS–HIERARCHICAL DIRICHLET PROCESS PROCESSES

As I mentioned earlier, the NDP can be considered a two-stage procedure in which stage I is a classification procedure through random variables $(\zeta_1, \ldots, \zeta_J)$ and stage II can be practically any procedure involving analysis on the given classes involving Dirichlet or more general random probability measures. This suggests that stage II can even consist of the HDP of Teh, Jordan, Beal, and Blei (2006) or in fact the hierarchical Pitman–Yor processes used by Teh (2006) in a quite interesting application to language modeling (as mentioned in the previous section). Without worrying too much about details, here I describe a (rough) variation of Teh's (2006) formulation. Imagine

that after stage I, one has randomly picked $1, \ldots, N(J)$ distinct classes that are associated with the $N(J)$ unique values among $(G^*_{\zeta_1}(\cdot), \ldots, G^*_{\zeta_J}(\cdot))$. Each of the $G^*_{\zeta_j}(\cdot)$ can produce an exchangeable sequence of size, say, $\tilde{n}_{\zeta_j}$, which would, under a general Pitman–Yor process, exhibit power law behavior as described in the previous setting and furthermore is of the type exploited by Teh (2006). To get a hierarchical structure closer to that of Teh (2006) or the simpler HDP, one would simply choose H to have a $\mathcal{PY}(\gamma_0, \theta_0, H_0)$ law, where H_0 is some probability measure. Note that in Teh's (2006) application, H_0 is a discrete distribution on words, in which case it makes sense to allow H_0 to depend on $N(J)$. Alternatively, more generally, if there are $(w_1, \ldots, w_J)$ objects, this distribution would be over the $N(J)$ distinct values in $(w_{\zeta_1}, \ldots, w_{\zeta_J})$, call this distribution $H_{0,N(J)}$. So within the HDP-type steps, we have distributions

$$(G^*_{\zeta_j}(\cdot)|\zeta_j, H) \overset{ind}{\sim} \mathcal{PY}\big(\gamma_{\zeta_j}, \theta_{\zeta_j}, H\big)$$

and

$$(H|N(J)) \sim \mathcal{PY}\big(\gamma_0, \theta_0, H_{0,N(J)}\big).$$

Furthermore, although I did say that for a fixed configuration of $(\zeta_1, \ldots, \zeta_J)$, this procedure is similar to that of Teh (2006), I am not claiming that they are exactly the same. Teh's procedure involves using a hierarchical Pitman–Yor process within each class, which can be naturally implemented here as well. On the other hand, without going into details, if θ_k is set to zero for all k, then the foregoing scheme is equivalent in distribution to Teh's scheme if we choose γ_k based on knowledge of the lengths of the *contexts* in each class. But the NDP idea dictates that every sampling of $(\zeta_1, \ldots, \zeta_J)$ would result in a new configuration of the HDP-type models just described.

5. CONCLUDING REMARKS

The sketch here represents only a few possibilities for using the two-stage type procedure suggested by the NDP. I thank the authors for stimulating my interest along these lines. I look forward to any comments from the authors, as well as more innovative concrete applications by others using these and other ideas from Bayesian nonparametrics.

ADDITIONAL REFERENCES

Goldwater, S., Griffiths, T. L., and Johnson, M. (2006), "Interpolating Between Types and Tokens by Estimating Power-Law Generators," in *Advances in Neural Information Processing Systems*, Vol. 18, eds. Y. Weiss, B. Schölkopf, and J. Platt, pp. 459–466.

Ishwaran, H., and James, L. F. (2004), "Computational Methods for Multiplicative Intensity Models Using Weighted Gamma Processes: Proportional Hazards, Marked Point Processes and Panel Count Data," *Journal of the American Statistical Association*, 99, 175–190.

Ishwaran, H., James, L. F., and Sun, J. (2001), "Bayesian Model Selection in Finite Mixtures by Marginal Density Decompositions," *Journal of the American Statistical Association*, 96, 1316–1332.

Kurihara, K., Welling, M., and Teh, Y. W. (2007), "Collapsed Variational Dirichlet Process Mixture Models," *International Joint Conference on Artificial Intelligence*.

Teh, Y. W. (2006), *A Hierarchical Bayesian Language Model Based on Pitman–Yor Processes*, Coling/ACL.

Rejoinder

Abel RODRÍGUEZ, David B. DUNSON, and Alan E. GELFAND

We are very grateful to the editors and all of the discussants for their valuable suggestions and positive comments. Most of these comments were focused on generalizations and additional applications of the NDP, with some helpful suggestions on computational approaches. We are delighted that there are so many possibilities for additional work in this area.

1. APPLICATIONS

The application of the NDP to the problem of prediction of lung cancer mortality presented by Ghosh, Ghosh, and Tiwari is quite interesting, and we are thrilled that the NDP can be implemented so easily in WinBUGS. Perhaps the WinBUGS code can be made publicly accessible. We agree that it would be useful in certain settings to allow for a different precision parameter β_k for each G^*_k, providing more flexibility in characterizing variability across groups in the number of clusters. One can then assume that the β_k's are drawn from a common hyperprior, such as a gamma, to allow borrowing of information.

Gillen and Johnson make an important point that the NDP can be used not only for random intercepts, but also, much more broadly, for borrowing of information and multilevel clustering of random intercepts and slopes. They mention GEE approaches as a potential competitor. But in our own experience, GEE approaches can have poor performance in small to moderate samples when the covariance structure is badly misspecified. In addition, our motivation in developing the NDP was not to obtain an approach for flexibly characterizing nuisance dependence in multilevel studies, but instead to carefully study differences among groups in the distribution of the response variables without imposing parametric assumptions. The NDP also can be used much more broadly as a component within hierarchical models for borrowing information and clustering in complex data. For example, Ni, Paisley, Carin, and Dunson (2008) used the NDP within a model for analyzing and sorting large sequential databases, with variational Bayes methods used for inference. In other work (Rodriguez, Dunson, and Gelfand 2008), we instead used the NDP for functional data.

Abel Rodriguez is Assistant Professor, Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA 95064 (E-mail: *abel@soe.ucsc.edu*). David B. Dunson is Senior Investigator, Biostatistics Branch, National Institute of Environmental Health Science, Research Triangle Park, NC 27709 (E-mail: *dunson1@niehs.nih.gov*). Alan E. Gelfand is James B. Duke Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708 (E-mail: *alan@isds.duke.edu*).

In the Public Domain
Journal of the American Statistical Association
September 2008, Vol. 103, No. 483, Theory and Methods
DOI 10.1198/016214508000000616

2. MODIFICATIONS AND GENERALIZATIONS

There are a number of generalizations of the NDP formulation that are well motivated by applications. Ghosh, Ghosh, and Tiwari mentioned that in the cancer mortality application, it would be appealing to include information on spatial location in the clustering process. In fact, the NDP could be combined with the kernel stick-breaking process (KSBP) of Dunson and Park (2008) to allow clustering to depend on spatial location, time, or predictors. To clarify, this would involve modifying expression (2) to let

$$G_j(\cdot) \sim \sum_{k=1}^{\infty} \pi_k^*(\mathbf{x}_j)\delta_{G_k^*(\cdot)},$$

where $\mathbf{x}_j$ are features of the jth group (e.g., spatial location), and the weights $\pi_k^*(\mathbf{x})$ are defined as functions of the features through the KSBP, which generalizes the stick-breaking formulation of the DP to include feature dependence through kernels placed at random locations.

James and MacEachern proposed various alternative generalizations. James suggested replacing the DP components with more flexible two-parameter Poisson–Dirichlet processes in order to induce a more flexible prior on the clustering process. One advantage of the DP formulation is that clusters are introduced slowly as the number of observations increases, with allocation to existing clustering increasingly favored. This tends to lead to a sparse formulation that is conservative in adding new clusters. However, in certain applications, such as natural language modeling, it may be necessary to allow a more rapid introduction clusters to avoid problems with underfitting. Another interesting variation is to replace the iid realizations $\{G_k^*\}$ with dependent distributions, for example, arising from an HDP. This leads to a more parsimonious representation of the data. In addition, a formulation of this type has the advantage of allowing clusters of hospitals to be comparable across groups of states.

MacEachern notes that it may be unrealistic in most applications to assume that two distributions, G_j and $G_{j'}$, can be exactly equal, so one may expect less clustering as the sample size per group increases. We agree that exact clustering in distributions serves as an approximation, but expect that two distributions that are very close but not strictly identical will tend to be clustered unless the sample size is extremely large. We note that the standard DP also induces global clustering, motivating a new literature on local partition processes (Dunson, Xue, and Carin 2008; Petrone, Guindani, and Gelfand 2008; Dunson 2008). One simple approach to relax the assumption of $F_j = F_{j'}$ for two groups in the same cluster is to incorporate a small contamination so that ϵ probability is allocated to group-specific atoms. MacEachern provides a number of useful alternative strategies.

We conclude this section by noting that our application considers clustering of states and hospitals within states. Suppose that we were to add another level, say patients within hospitals. Retaining our generic notation, we would now have $\boldsymbol{\theta}_{ijk}$ with patients, indexed by i, within hospitals, indexed by j, within states, indexed by k. Suppose that we drew $\boldsymbol{\theta}_{ijk}$ from G_{jk}, where the $G_{jk} \sim \mathrm{DP}(\delta G_k)$ with the G_k from an NDP, that is, G_k iid from Q with $Q \sim \mathrm{DP}(\alpha\,\mathrm{DP}(\beta H))$. This specification would allow clustering of patients within hospitals and clustering of states but not of hospitals. Suppose that instead we assumed that the G_{jk}'s came from NDPs indexed by k (i.e., $G_{jk} \sim Q_k$), where we had iid $Q_k \sim \mathrm{DP}(\alpha\,\mathrm{DP}(\beta H))$. Now we could cluster patients and hospitals, but not states. In both cases we would be subject to the limitations of the HDP as described in section 3.3. To enable clustering at all three levels, we need iid $Q_k \sim \Lambda$, where $\Lambda \sim \mathrm{DP}(\delta\,\mathrm{DP}(\alpha\,\mathrm{DP}(\beta H)))$. We have a NDP nested within an NDP.

3. COMPUTATION

Our proposed approach relies on truncations of a stick-breaking formulation, but certainly other possibilities exist. James suggests exploring an alternative based on finite-dimensional Dirichlet distributions. We agree that the two approximations should offer similar performance when the truncation bound is chosen conservatively; however, it can be shown that for the standard DP, the truncated stick-breaking approximation converges faster to the DP compared with the finite-dimensional Dirichlet approximation (Paisley, Carin, and Dunson, manuscript in preparation). We would expect this result to hold for the NDP as well. Potentially, a finite approximation can be avoided by relying on a slice sampler that generalizes the algorithm of Walker (2007), although we have not yet implemented this approach. Müller and Nieto-Barajas note that although marginal samplers cannot be efficiently implemented for the NDP, the Polya urn representation still can be used to construct reversible-jump MCMC samplers, which typically have better mixing properties than collapsed samplers. This is certainly a direction that we plan to explore in the near future. A major advantage of the truncation approach is simplicity, allowing for straightforward implementation in WinBUGS, as noted by Ghosh, Ghosh, and Tiwari.

4. CONCLUSION

The use of nonparametric Bayes methods in applications has dramatically increased over the past several years, particularly in biomedical and machine learning applications. We are currently exploring the use of the NDP for multitask learning problems where interest focuses on flexible borrowing of information across data from different sources. In this fast-moving area, we anticipate many other possibilities, stimulated by the need for flexible models to tease out structure in large, complex data sets.

ADDITIONAL REFERENCES

Dunson, D. B. (2008), "Local Partition Processes," discussion paper, Duke University, Dept. of Statistical Science.

Dunson, D. B., and Park, J. H. (2008), "Kernel Stick-Breaking Processes," *Biometrika*, 95, 307–323.

Dunson, D. B., Xue, Y., and Carin, L. (2008), "The Matrix Stick-Breaking Process: Flexible Bayes Meta Analysis," *Journal of the American Statistical Association*, 103, 317–327.

Ni, K., Paisley, J., Carin, L., and Dunson, D. B. (2008), "Multi-Task Learning for Analyzing and Sorting Large Databases of Sequential Data," *IEEE Transactions on Signal Processing*, 56, 3918–3931.

Petrone, S., Guindani, M., and Gelfand, A. E. (2008), "The Hybrid Functional Dirichlet Process," *Journal of the Royal Statistical Society*, Ser. B, forthcoming.

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "Bayesian Nonparametric Functional Data Analysis Through Density Estimation," *Biometrika*, forthcoming.

Walker, S. G. (2007), "Sampling the Dirichlet Mixture Model With Slices," *Communications in Statistics, Part B—Simulation and Computation*, 36, 45–54.