

THE NESTED DIRICHLET PROCESS

ABEL RODRÍGUEZ, DAVID B. DUNSON, AND ALAN E. GELFAND

ABSTRACT. 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 Dirichlet process (nDP) 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 MCMC algorithm is developed for computation. The methods are illustrated using a simulation study and an application to quality of care in US hospitals.

1. INTRODUCTION

The Dirichlet Process (DP) (Ferguson, 1973, 1974) is the most widely used non-parametric model for random distributions in Bayesian statistics, due mainly to the availability of efficient computational

¹Abel Rodriguez is Ph.D. candidate, Institute of Statistics and Decision Sciences, Duke University, Box 90251, Durham, NC 27708, abel@isds.duke.edu. David B. Dunson is Senior Investigator, Biostatistics Branch, National Institute of Environmental Health Science, P.O. Box 12233, RTP, NC 27709, dunson1@niehs.nih.gov. Alan E. Gelfand is James B. Duke professor, Institute of Statistics and Decision Sciences, Duke University, Box 90251, Durham, NC 27708, alan@isds.duke.edu. This work was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

Key words and phrases. Clustering; Dependent Dirichlet process; Gibbs sampler; Hierarchical model; Nonparametric Bayes; Random probability measure.

The authors would like to thank Shyamal Peddada and Ju Hyun Park for helpful comments.

techniques (Escobar and West, 1995; Bush and MacEachern, 1996; MacEachern and Müller, 1998; Neal, 2000; Green and Richardson, 2001; Jain and Neal, 2000). Since the Dirichlet Process puts probability one on the space of discrete measures, it is typically not used to model the data directly. Instead, it is more naturally employed as a prior for a mixing distribution, resulting in a DP mixture (DPM) model (Antoniak, 1974; Escobar, 1994; Escobar and West, 1995). Some recent applications of the Dirichlet Process include finance (Kacperczyk et al., 2003), econometrics (Chib and Hamilton, 2002; Hirano, 2002), epidemiology (Dunson, 2005), genetics (Medvedovic and Sivaganesan, 2002; Dunson et al., 2005), medicine (Kottas et al., 2002; Bigelow and Dunson, 2005) 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 Dirichlet Process to accommodate multiple dependent distributions.

The dependent Dirichlet process (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 (Sethuraman, 1994) with stochastic processes. The “single p ” version of this construction (where dependence is introduced only on the atoms) has been employed by DeIorio et al. (2004) to create ANOVA-like models for densities, and by Gelfand et al. (2005) to generate spatial processes that allow for non-normality and non-stationarity. This last class of models is extended in Duan et al. (2005) to create generalized spatial Dirichlet processes (GSDP) that allow different surface selection at different locations.

Along these lines, another approach to introduce dependence is the hierarchical Dirichlet process (HDP) (Teh et al., 2004). In this setting, multiple group-specific distributions are assumed to be drawn from a common Dirichlet Process whose baseline measure is in turn a draw from another Dirichlet process. This allows the different distributions to share the same set of atoms but have distinct sets

of weights. More recently, Griffin and Steel (2006) propose an order-dependent Dirichlet Process, where the weights are allowed to change with the covariates.

An alternative approach is to introduce dependence through linear combinations of realizations of independent Dirichlet processes. For example, Müller et al. (2004), motivated by a similar problem to Teh et al. (2004), define 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 et al. (2004) propose a model for density regression using a kernel-weighted mixture of Dirichlet Processes 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 non-normal, 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, and to do so 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 in Teh et al. (2004) and Müller et al. (2004).

In developing methods for characterizing topic hierarchies within documents, Blei et al. (2004) proposed a nested Chinese restaurant process. This approach 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 for modeling *a collection* of dependent distributions.

The paper is organized as follows. We start in section 2 with a short review of the Dirichlet process. Section 3 motivates and defines the nested Dirichlet process (nDP), and explores its theoretical

properties. Truncations of the nDP and their application in deriving efficient computational schemes is discussed in sections 4 and 5. Sections 6 and 7 present examples that illustrate the advantages of our methodology. Finally, we close in section 8 with a brief discussion.

2. THE DIRICHLET PROCESS

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

The Dirichlet process can be alternatively characterized in terms of its predictive rule (Blackwell and MacQueen, 1973). If $(\theta_1, \dots, \theta_{n-1})$ is an iid sample from $P \sim \text{DP}(\alpha H)$, we can integrate out the unknown P and obtain the conditional predictive distribution of a new observation,

$$\theta_n | \theta_{n-1}, \dots, \theta_1 \sim \frac{\alpha}{\alpha + n - 1} H + \sum_{l=1}^{n-1} \frac{1}{\alpha + n - 1} \delta_{\theta_l}$$

where δ_{θ_l} is the Dirac probability measure concentrated at θ_l .

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

The Dirichlet process can also be regarded as a type of *stick-breaking prior* (Sethuraman, 1994; Pitman, 1996; Ishwaran and James, 2001; Ongaro and Cattaneo, 2004). A stick-breaking prior on the

space \mathbf{P} has the form

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

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

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 $0 \leq a < 1$ and $b > -a$ yields the two-parameter Poisson-Dirichlet Process, also known as Pitman-Yor Process (Pitman, 1996), with the choice $a = 0$ and $b = \alpha$ resulting in the Dirichlet Process (Sethuraman, 1994).

The stick-breaking representation is probably the most versatile definition of the Dirichlet Process. 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 Retrospective sampler (Roberts and Papaspiliopoulos, 2004), which does not require truncation. It is also the starting point for the definition of many generalizations that allow dependence across a collection of distributions, including the DDP (MacEachern, 2000), the π DDP (Griffin and Steel, 2006) and the GSDP (Duan et al., 2005).

3. THE NESTED DIRICHLET PROCESS

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

In analyzing multi-center data, there are a number of customary strategies, with the most common being (1) pool the data from the different centers; (2) analyze the data from the different centers separately; and (3) fit a parametric hierarchical model to borrow information. The first approach is too restrictive, as subjects in different centers may have different distributions, while the second approach is inefficient. The third approach parameterizes F_j in terms of the finite-dimensional parameter θ_j , and then borrows information by assuming $\theta_j \stackrel{iid}{\sim} F_0$, with F_0 a known distribution (most commonly normal), possibly having unknown parameters (mean, variance). One can potentially cluster centers having similar random effects, θ_j , though 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 $\{F_j, j = 1, \dots, J\}$ nonparametrically to enhance flexibility, and we use a Dirichlet type of specification to enable clustering of random distributions.

In what follows, a collection of distributions $\{F_1, \dots, F_J\}$ is said to follow a *Nested Dirichlet Processes Mixture* if

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

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

$$(3) \quad G_k^*(\cdot) = \sum_{l=1}^{\infty} w_{lk}^* \delta_{\theta_{lk}^*(\cdot)}$$

with $\theta_{lk}^* \sim H$, $w_{lk}^* = u_{lk}^* \prod_{s=1}^{l-1} (1 - u_{sk}^*)$, $\pi_k^* = v_k^* \prod_{s=1}^{k-1} (1 - v_s^*)$, $v_k^* \sim \text{beta}(1, \alpha)$ and $u_{lk}^* \sim \text{beta}(1, \beta)$. In expression 1, $p(\cdot | \theta, \phi)$ is a distribution parameterized by the finite dimensional vectors θ and ϕ , whose specific choice depends on the application at hand. For example, in the case of a univariate response, if the collection $\{F_1, \dots, F_J\}$ is assumed exchangeable, an attractive choice would be $\theta = (\mu, \sigma)$ and $p(\cdot | \theta, \phi) = \text{N}(\cdot | \mu, \sigma^2)$, 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, we could opt for a random effects model where $\theta = \mu$, $\phi = (\gamma, \sigma^2)$ and $p(\cdot | \theta, \phi) = \text{N}(\cdot | \mu + \mathbf{x}'\gamma, \sigma^2)$ in a similar spirit to Mukhopadhyay and Gelfand (1997) and Kleinman and Ibrahim (1998). Extensions to multivariate or discrete outcomes are immediate using the standard Bayesian machinery.

The collection $\{G_1, \dots, G_J\}$, used as the mixing distribution, is said to follow a *Nested Dirichlet Process* with parameters α , β and H , and is denoted $\text{nDP}(\alpha, \beta, H)$. In a more concise notation, the model for our clustering problem can be rewritten as

$$y_{ij} \sim p(y_{ij} | \theta_{ij}) \quad \theta_{ij} \sim G_j \quad \{G_1, \dots, G_J\} \sim \text{nDP}(\alpha, \beta, H)$$

Since $\mathbb{P}(G_j = G_{j'}) = \frac{1}{1+\alpha} > 0$, the model naturally induces clustering in the space of distributions. Also, the stick breaking construction of G_k^* ensures that marginally, $G_j \sim \text{DP}(\beta H)$ for every j . Therefore, for any measurable set $A \in \mathcal{B}(\Theta)$

$$\mathbb{E}(G_j(A)) = H(A) \quad \text{and} \quad \mathbb{V}(G_j(A)) = \frac{H(A)(1 - H(A))}{\beta + 1}$$

In understanding the prior correlation between two distributions G_j and $G_{j'}$, it is natural to consider $\text{Cor}(G_j, G_{j'})$. For the nDP, this value can be calculated for any non-atomic H by exploiting the stick-breaking construction of the process (see appendix A), yielding

$$\text{Cor}(G_j, G_{j'}) = \text{Cor}(G_j(A), G_{j'}(A)) = \frac{1}{1 + \alpha} = \mathbb{P}(G_j = G_{j'})$$

This expression, which provides a natural interpretation for the additional parameter in the nDP, will be referred to as the correlation between distributions since it is independent of the set A . The correlation between draws from the process can also be calculated (see again appendix A), yielding

$$\mathbb{C}or(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j'}) = \begin{cases} \frac{1}{(1+\beta)} & j = j' \\ \frac{1}{(1+\alpha)(1+\beta)} & j \neq j' \end{cases}$$

showing 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 $\alpha \rightarrow \infty$, 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 $\lim_{\alpha \rightarrow \infty} \mathbb{C}or(G_j, G_{j'}) = 0$. On the other hand, as $\alpha \rightarrow 0$ the a priori probability of assigning all the distributions to the same atom G^* goes to 1, and thus the correlation goes to 1. Hence, approaches (1) and (2) for the analysis of multiple centers described above are limiting cases of the nDP. Moreover, since $F_j(\cdot) \rightarrow p(\cdot|\boldsymbol{\theta}_j^*, \boldsymbol{\phi})$ as $\beta \rightarrow 0$, the nDP also encompasses the natural parametric-based clustering (option (3) above) as a limiting case.

Since 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 , while patients i and i' , respectively from hospitals j and j' , are clustered together if and only if $G_j = G_{j'} = G_k^*$ and $\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_{i'j'} = \boldsymbol{\theta}_{ik}^*$ for some

l . In the sequel we use indicator $\zeta_j = k$ and $\xi_{ij} = l$ if $G_j = G_k^*$ and $\theta_{ij} = \theta_{lk}^*$ to denote membership to the distributional and observational clusters respectively.

3.2. Alternative characterizations of the nDP. Just as the Dirichlet Process 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 Dirichlet Process (Ferguson, 1973, 1974) stated in section 2. The choice $\Theta \subset \mathbb{R}^n$ for the base space of the Dirichlet Process is merely a practical one, and the results mentioned above extend in general to any complete and separable metric space Θ . In particular, since the space 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 and $Q \in \mathbf{Q}$. In this setting, \mathbf{Q} is the space of *distributions on probability distributions* on (Θ, \mathcal{B}) .

By requiring S to be such that $(Q(C_1), \dots, Q(C_k)) \sim \text{DP}(\alpha\nu(C_1), \dots, \alpha\nu(C_k))$ for any partition (C_1, \dots, C_k) of \mathbf{P} generated under the weak topology and some α and suitable ν , we have defined a new Dirichlet Process $S \sim \text{DP}(\alpha\nu)$, this time on an abstract space, that satisfies the usual properties. The nested Dirichlet process is a special case of this formulation in which ν is taken to be a regular $\text{DP}(\beta H)$. Therefore, the nDP is an example of a DP where the baseline measure is an stochastic process generating probability distributions. An alternative notation for the nDP corresponds to $G_j \stackrel{iid}{\sim} Q$ with $Q \sim \text{DP}(\alpha \text{DP}(\beta H))$.

The nDP can also be characterized as a dependent Dirichlet process (MacEachern, 2000) where the stochastic process generating the elements of the stick-breaking representation corresponds to a Dirichlet process.

4. TRUNCATIONS

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

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

$$\begin{aligned}
 G_j^K(\cdot) &\sim \sum_{k=1}^K \pi_k^* \delta_{G_k^{L*}(\cdot)} \\
 G_k^{L*}(\cdot) &= \sum_{l=1}^L w_{lk}^* \delta_{\theta_{lk}^*}(\cdot) \quad \pi_k^* = v_k^* \prod_{s=1}^{l-1} (1 - v_s^*) \text{ with } v_K^* = 1 \\
 & \quad v_k^* \sim \text{beta}(1, \alpha) \quad k = 1, \dots, K - 1 \\
 \theta_{lk}^* &\sim H \quad w_{lk}^* = u_{lk}^* \prod_{s=1}^{l-1} (1 - u_{sk}^*) \text{ with } u_{Lk}^* = 1 \\
 & \quad u_{lk}^* \sim \text{beta}(1, \beta) \quad l = 1, \dots, L - 1
 \end{aligned}$$

We refer to this model as a bottom-level truncation or $\text{nDP}^{L\infty}(\alpha, \beta, H)$ if $K = \infty$ and $L < \infty$, whereas if $K < \infty$ and $L = \infty$ we refer to it as a top-level truncation or $\text{nDP}^{\infty K}(\alpha, \beta, H)$. Finally, if both L and K are finite we have a two-level truncation or $\text{nDP}^{LK}(\alpha, \beta, H)$.

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

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

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

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

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

$$\int |P^{LK}(\mathbf{y}) - P^{\infty\infty}(\mathbf{y})| d\mathbf{y} \leq \int |P^{LK}(\boldsymbol{\theta}) - P^{\infty\infty}(\boldsymbol{\theta})| \leq \epsilon^{LK}(\alpha, \beta)$$

where

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

The proof of this theorem is presented in appendix B. Note that the bounds approach zero in the limit, so the truncation approximations and its predictive distribution converge in total variation (and therefore in distribution) to the nDP. Even more, the bounds are strictly decreasing in both L and K . As a consequence of this observation we have the following corollary.

Corollary 1. *The posterior distribution under a LK truncation and the corresponding nDP converge in distribution as both $L, K \rightarrow \infty$.*

The proof is presented in appendix C. It is straightforward to extend the previous results and show that $\lim_{L \rightarrow \infty} \text{nDP}^{LK} = \text{nDP}^{\infty K}$ and $\lim_{K \rightarrow \infty} \text{nDP}^{LK} = \text{nDP}^{L\infty}$ in distribution.

In order to better understand the influence of the truncation levels on the accuracy of the approximation we show in Figure 1 the error bounds for a $nDP(3, 3, H)$ in various sample size settings. The value $\alpha = \beta = 3$ in this simulation, which will typically lead to a relatively large number of components in the mixtures, was chosen as a worst case scenario since the bounds are strictly decreasing in both α and β .

The first three examples have a total of 5,000 observations, which have been split in different ways. Note that, as the number of groups J increases, K needs to be increased to maintain accuracy. The fourth example has the same number of observations per group as the first, but double the number of groups. In every case, increasing K over 35 seems to have little effect on the error bound. These results suggest that for moderately large sample sizes ($n \leq 500$ and $J \leq 50$), and typical values of the concentration parameters α and β , a choice of $K = 35$ and $L = 55$ seems to provide an adequate approximation.

5. POSTERIOR COMPUTATION

Broadly speaking, there are three strategies for computation in standard DP models: (1) Employ the Pólya urn scheme to marginalize out the unknown infinite-dimensional distribution(s) (MacEachern, 1994; Escobar and West, 1995; MacEachern and Müller, 1998), (2) Employ 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 (3) Use reversible-jump MCMC (RJMCMC) algorithms for finite mixtures with an unknown number of components (Dahl, 2003; Green and Richardson, 2001; Jain and Neal, 2000). In this section, we explore the use of these strategies to construct efficient algorithms for inference in the nDP setting. In the sequel, let $\zeta_j = k$ and $\xi_{ij} = l$ iff $G_j = G_k^*$ and $\theta_{ij} = \theta_{l\zeta_j}^*$.

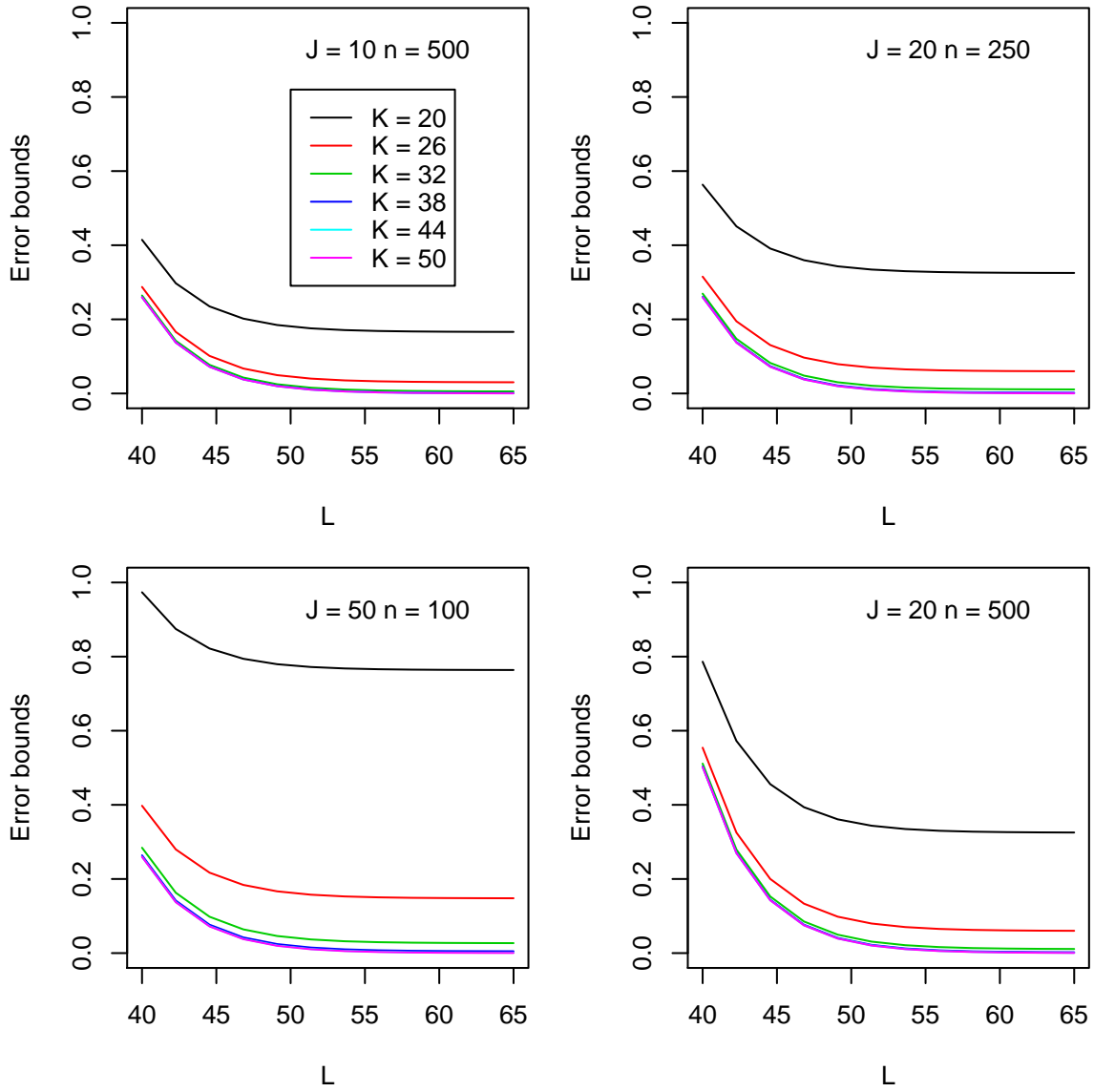


FIGURE 1. Approximate error bounds for the LK truncation of a $nDP(3, 3, H)$. Top left corner corresponds to $n = 500$ and $J = 10$, top right to $n = 250$ and $J = 20$, bottom left to $n = 100$ and $J = 50$ and bottom right to $n = 500$ and $J = 20$.

Implementations of the nDP based on (1) are, in general, infeasible. Although sampling ξ_{ij} given $(\zeta_1, \dots, \zeta_J)$ using a Pólya urn scheme is straightforward, sampling ζ_j requires the 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 the conditional $p(\mathbf{y}_j|G_s^*)$ (whose evaluation requires an infinite sum since $G_s^* \sim \text{DP}(\beta H)$). Algorithms using RJMCMC in the nDP are likely to run into 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, without any obvious way to construct efficient proposals. Hence, we focus on combinations of truncation approximations.

5.1. Sampling by double truncation. The obvious starting place is to consider a two-level truncation of the process using values of K and L elicited from plots like those shown in Figure 1.

Once adequate values of K and L have been chosen, computation proceeds through the following steps:

- (1) Sample the center indicators ζ_j for $j = 1, \dots, J$ from a multinomial distribution with probabilities

$$\mathbb{P}(\zeta_j = k | \dots) = q_k^j \propto w_k^* \prod_{i=1}^{I_j} \sum_{l=1}^L \pi_{lk} p(y_{ij} | \boldsymbol{\theta}_{lk}^*)$$

- (2) Sample the group indicators ξ_{ij} for $j = 1, \dots, J$ and $i = 1, \dots, n_j$ from another multinomial distribution with probabilities

$$\mathbb{P}(\xi_{ij} = l | \dots) = b_{ij}^l \propto \pi_{l\zeta_j}^* p(y_{ij} | \boldsymbol{\theta}_{l\zeta_j}^*)$$

- (3) Sample π_k^* by generating

$$(u_k^* | \dots) \sim \text{beta} \left(1 + m_k, \alpha + \sum_{s=k+1}^K m_s \right) \quad k = 1, \dots, K-1 \quad u_K^* = 1$$

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

$$(v_{lk}^* | \dots) \sim \text{beta} \left(1 + n_{lk}, \beta + \sum_{s=l+1}^L n_{ls} \right) \quad l = 1, \dots, L-1, \quad v_{Lk}^* = 1$$

where n_{lk} is the number of observations assigned to atom l of distribution k , and constructing $w_{lk}^* = v_{lk}^* \prod_{s=1}^{l-1} (1 - v_{sk}^*)$

(5) Sample θ_{lk}^* from

$$p(\theta_{lk}^* | \dots) \propto \left[\prod_{\{i,j | \zeta_j = k, \xi_{ij} = l\}} p(y_{ij} | \theta_{lk}^*) \right] p(\theta_{lk}^*)$$

Note that if no observation is assigned to a specific cluster, then the parameters are drawn from the prior distribution (baseline measure) $p(\theta_{lk}^*)$. Also, if the prior is conjugate to the likelihood then sampling is greatly simplified. However, non-conjugate priors can be accommodated using rejection sampling or Metropolis-Hastings steps.

(6) Sample the concentration parameters α and β from

$$p(\alpha | \dots) \propto \alpha^{K-1} \exp \left\{ \alpha \sum_{k=1}^{K-1} \log(1 - u_k^*) \right\} p(\alpha)$$

$$p(\beta | \dots) \propto \beta^{K(L-1)} \exp \left\{ \beta \sum_{l=1}^{L-1} \sum_{k=1}^K \log(1 - v_{lk}^*) \right\} p(\beta)$$

If conditionally conjugate priors $\alpha \sim \text{G}(a_\alpha, b_\alpha)$ and $\beta \sim \text{G}(a_\beta, b_\beta)$ are chosen then,

$$(\alpha | \dots) \sim \text{G} \left(a_\alpha + (K-1), b_\alpha - \sum_{k=1}^{K-1} \log(1 - u_k^*) \right)$$

$$(\beta | \dots) \sim \text{G} \left(a_\beta + K(L-1), b_\beta - \sum_{l=1}^{L-1} \sum_{k=1}^K \log(1 - v_{lk}^*) \right)$$

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.

Besides the simplicity of its implementations, 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. Note that the most computationally expensive steps are (1), (2) and (5). However, $(\zeta_j | \dots)$ and $(\zeta_{j'} | \dots)$ are independent, as well as $(\xi_{ij} | \dots)$ and $(\xi_{i'j'} | \dots)$, and $(\theta_{lk}^* | \dots)$ and $(\theta_{l'k'}^* | \dots)$. Hence, steps (1), (2) and (5) can be divided into subprocesses that can be run in parallel.

5.2. One-level truncation. In order to compute predictive probabilities needed to sample the center indicators, only the top-level truncation is strictly necessary. If this level is truncated, ζ_1, \dots, ζ_J can be sampled using a regular Pólya Urn scheme avoiding the need for the second truncation. However, even a prior $p(\theta_{ij})$ conjugate to the likelihood $p(y_{ij} | \theta_{ij})$ does not imply a conjugate model on the distributional level. Hence, Pólya urn methods for *non-conjugate distributions* (MacEachern and Müller, 1998; Neal, 2000) need to be employed in this setup, greatly reducing the computational advantages of the Pólya urn over truncations.

6. SIMULATION STUDY

In this section we present a simulation study designed to provide insight into the discriminating capability of the nDP, as well as its ability to provide more accurate density estimates by borrowing strength across centers. The set up of the study is as follows: J samples of size N are obtained from four mixtures of four Gaussians defined in table 1 and plotted in Figure 2. These distributions have been chosen to reflect situations that are conceptually hard: T1 and T2 are asymmetric and composed of the same two Gaussian components which have been weighted differently, while T3 and T4 share three distributions located symmetrically around the origin, differing only in an additional bump that T4 presents on the right tail.

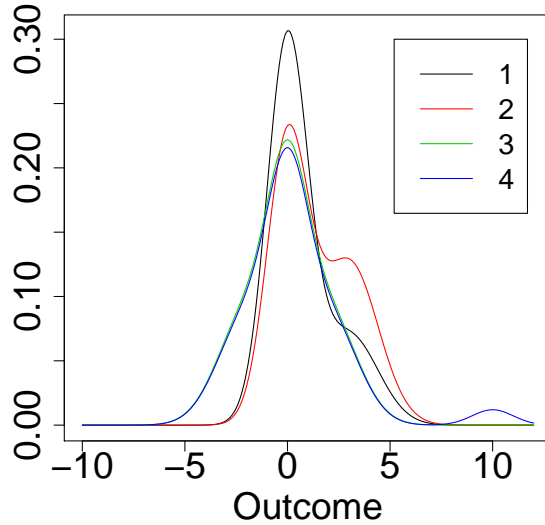


FIGURE 2. True distributions used in the the simulation study.

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	0.75	0.0	1.0	0.25	3.0	2.0	-	-	-	-	-	-
T2	0.55	0.0	1.0	0.45	3.0	2.0	-	-	-	-	-	-
T3	0.40	0.0	1.0	0.30	-2.0	2.0	0.30	2.0	2.0	-	-	-
T4	0.39	0.0	1.0	0.29	-2.0	2.0	0.29	2.0	2.0	0.03	10.0	1.0

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

The value of J and N was varied across the study in order to assess the influence of the sample size on the discriminating capability of the model. To simplify interpretation of the results, the same number of samples were contiguously obtained from each of the true distributions. The precision parameters α and β were both fixed to 1 and a Normal Inverse-Gamma distribution $\text{NIG}(0, 0.01, 3, 1)$ 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 5.1 was used to obtain samples of the posterior distribution under the nDP. 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 summary commonly employed is the set of marginal probabilities of any two pairs belonging to the same cluster. We visualize these pairwise probabilities in our study using the symmetric heatmaps presented in Figure 3. Although multiple runs were performed, we present only one representative random sample for each combination of values N and J . Since the diagonal of the matrix represents the probability of a distribution being classified with itself, it takes the value 1.

For small values of N , the nDP is able to roughly separate T1 and T2 from T3 and T4, but not to discriminate between T1 and T2 or T3 and T4. This is not really surprising: the method is designed to induce clustering. Therefore, when differences are highly uncertain, it prefers to create less rather than more clusters. However, as N increases, the model is able to distinguish between distributions and 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 .

In Figure 4 we show density estimates obtained for sample 1 of the example $J = 20$, $N = 100$. The left panel shows the one obtained from the nDP (which borrows information across all samples), while the right panel 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-generation mechanism, the estimate is more accurate: it not only captures the small mode to the right more clearly, but it also emphasizes the importance

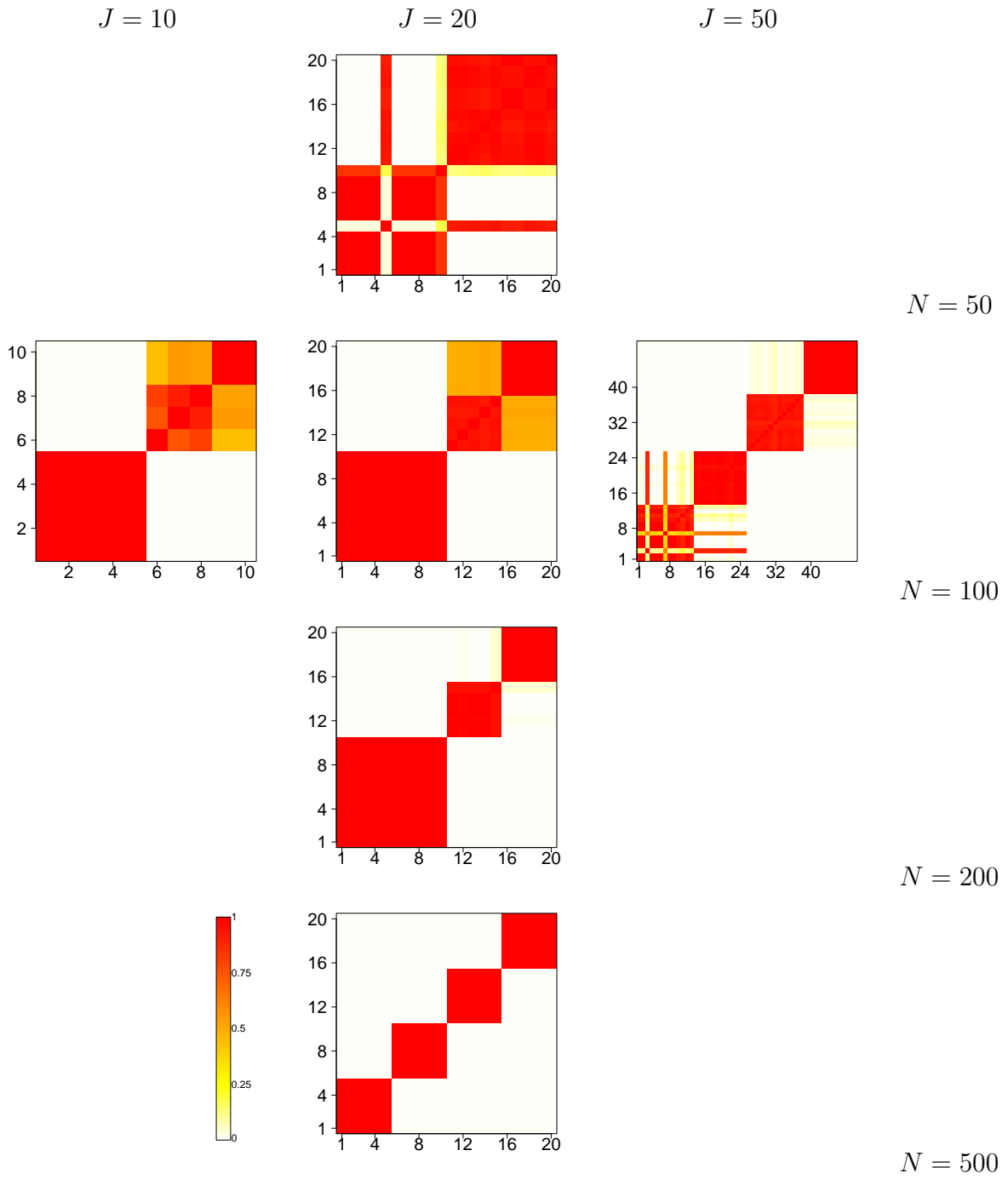


FIGURE 3. Pairwise probabilities of joint classification for the simulation study

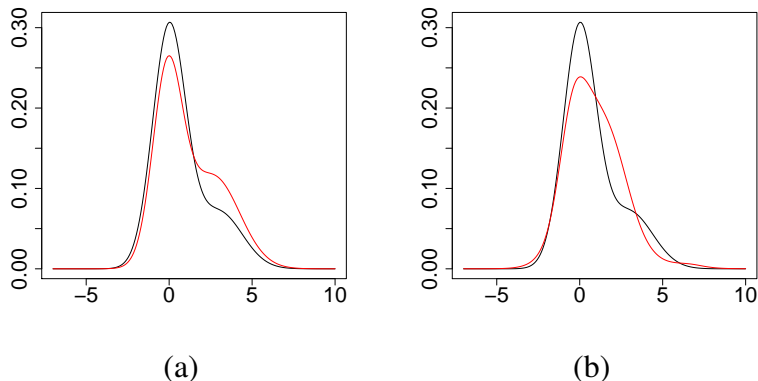


FIGURE 4. True (black) and estimated (red) densities for distribution 1 of the simulation with $J = 20$ and $N = 100$. Panel (a) corresponds to an estimate based on the nDP, which borrows information across all samples, while Panel (b) corresponds to an estimate based only on sample 1.

of the main mode. Indeed the entropy of the density estimate relative to the true distribution for the estimate of T1 under the nDP is 0.011, while under the regular DPM it was 0.017.

7. AN APPLICATION: HEALTH CARE QUALITY IN UNITED STATES

Data on quality of care in hospitals across the United States and associated territories is made publicly available by the Department of Health and Human Services at the website <http://www.hhs.gov/>. Twenty measures are recorded for each hospital, comprising aspects like proper and timely application of medication, treatment and discharge instructions. In what follows we focus on one specific measure: the proportion of patients that were 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 non-profit 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, is also available. Hospitals with less than 30 patients treated and territories with less than 4 hospitals were judged misrepresentative and removed from the sample, yielding a final sample size of 3077 hospitals in 51 territories (the 50 states plus the District of Columbia). Number of hospitals per state varies widely, with 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 1175, with quartiles at 76, 130 and 197 patients. Since the value tends 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 getting 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.

In order 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 was 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 (p -values 0.011 and 1.916×10^{-8}). Residual plots for this model show some deviation from homoscedasticity and normality (see Figure 5), but given the large sample size it is unlikely that this has any impact on the results so far.

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. However, the density estimates in Figure 7 (obtained using Gaussian kernels with a bandwidth chosen with the rule of thumb described in Silverman (1986)) show that state-specific residual distributions can be highly non-normal and that changes across states

can go beyond location and scale changes to affect the whole shape of the distribution. Invoking asymptotic arguments at this point is not viable since sample sizes are small and we are dealing with 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 can have similar error distributions, like California and Minnesota or Florida and North Carolina.

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

$$\begin{aligned} y_{ij} &= \mu_{ij} + \mathbf{x}_{ij}\boldsymbol{\gamma} + \epsilon_{ij} & \epsilon_{ij} &\sim \mathbf{N}(0, \sigma_{ij}) \\ (\mu_{ij}, \sigma_{ij}^2) &\sim G_j & \{G_1, \dots, G_J\} &\sim \text{nDP}(\alpha, \beta, H) \end{aligned}$$

where \mathbf{x}_{ij} is the vector of covariates associated with the hospital. Prior elicitation is simplified by centering the observations. We pick $H = \text{NIG}(0, 0.01, 3, 1)$, which implies $\mathbb{E}(\mu|\sigma^2) = 0$, $\mathbb{V}(\mu|\sigma^2) = 100\sigma^2$, $\mathbb{E}(\sigma^2) = 1$ and $\mathbb{V}(\sigma^2) = 3$. We use a standard reference (flat) prior on $\boldsymbol{\gamma}$. Finally, we set $\alpha, \beta \sim \text{G}(3, 1)$ a priori, implying that $\mathbb{E}(\alpha) = \mathbb{E}(\beta) = 1$ (a common choice in the literature) and $\mathbb{P}(\alpha > 3) = \mathbb{P}(\beta > 3) \approx 0.006$. Note that this choice implies that $\mathbb{P}(\text{Cor}(G_j, G_{j'}) > 0.25) \approx 0.994$.

Posterior computation is a straightforward extension of the algorithm presented in section 5.1. Conditional on μ_{ij} and σ_{ij} the model is a regular ANOVA with known variance, with the 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 $z_{ij} = y_{ij} - \mathbf{x}_{ij}\boldsymbol{\gamma}$. Results below are based on 50,000 iterations obtained after a burn-in period of 5,000 iterations. According to the simulation study presented in section 4, we choose $K = 35$ and $L = 55$ as the truncation levels. Results seem to be robust to

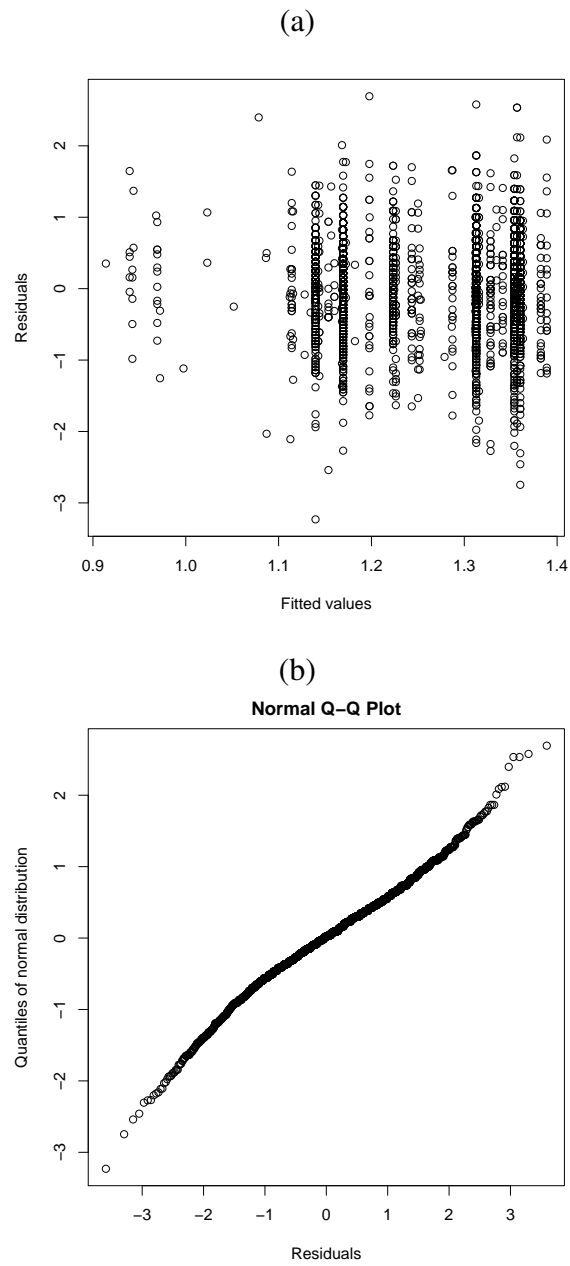


FIGURE 5. Residual plots for the ANOVA model on the initial antibiotic data: (a) Residuals vs. fitted values, (b) Quantile-quantile plot

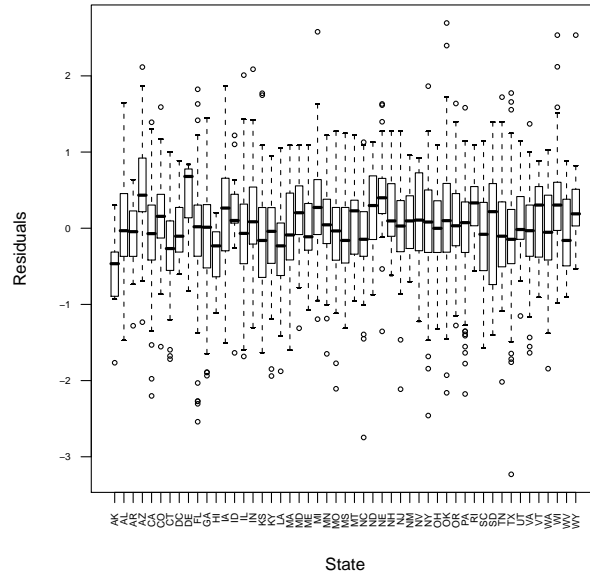


FIGURE 6. State-specific residual boxplots for the ANOVA model on the initial antibiotic data.

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 distributions shows strong evidence in favor of either 2 or 3 components (posterior probabilities 0.616 and 0.363 respectively), and little support for either 1, 4 or 5 distributions (posterior probabilities 0.00, 0.02 and 0.001 respectively). As with the simulated example, we visualize the matrix of pairwise probabilities using a heatmap, which is shown in Figure 8. In order to make sense of the plot, we first reorder the states using an algorithm inspired by those used for microarray analysis.

This heatmap provides additional insight into the clustering structure. It shows three well defined groups: (1) a large clusters of 31 members (lower left corner of the plot); (2) a small cluster of 5 states (upper right corner); and (3) the remaining 15 states, which are not clear members of any of the two previous clusters and do not seem to form a coherent cluster among themselves. Indeed, these states

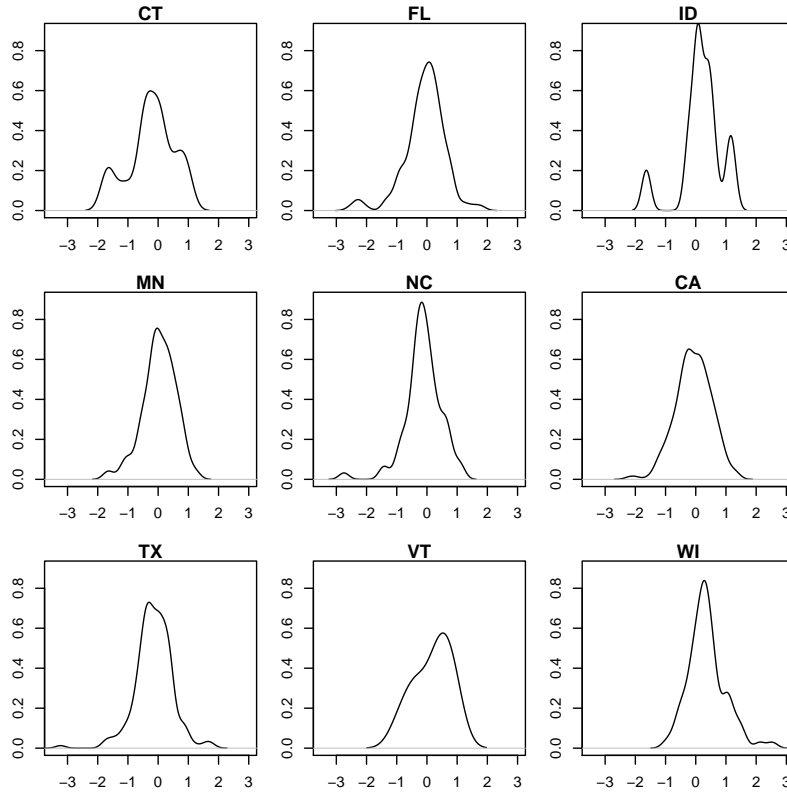


FIGURE 7. Density estimates for the residual distribution in selected states. Note that distributions seem clearly non-normal and that their shape can have important variations, making any parametric assumption hard to support.

seem to be independently classified in both clusters with relatively large probability, while sometimes forming a separate cluster altogether. This indicates a sort of intermediate behavior not completely consistent with any of the first two clusters. We note that, although the third group is made up mostly of states with a small number of hospitals, the phenomenon does not seem to be an artifact of the method.

Figure 9 shows posterior density estimates for 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

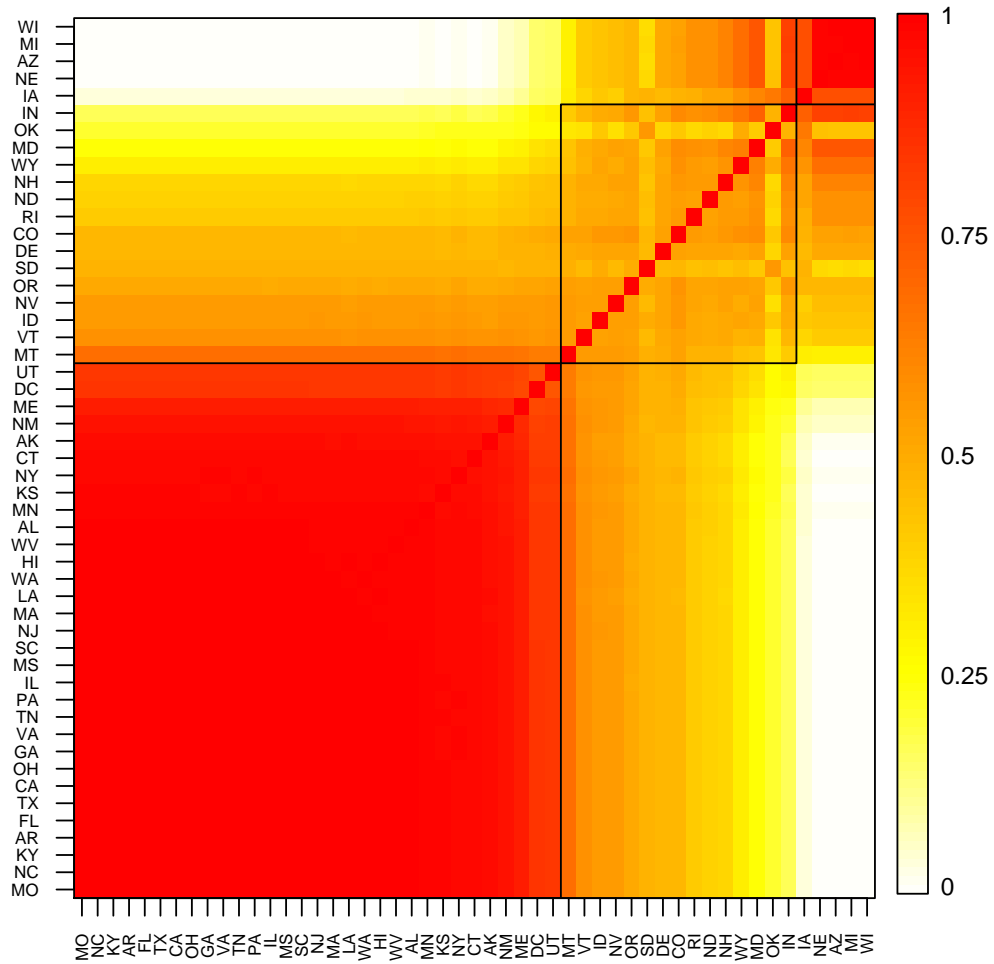


FIGURE 8. Residual plots for the ANOVA model on the initial antibiotic data.

left tail, indicating that each of those states contains some underperforming hospitals and few or none overperforming hospitals. The situation for Wisconsin and cluster 2 is reversed: these seem to be states with a higher average performance, quite a few hospitals that have an excellent record in the application of antibiotics and few or no low-quality hospitals. Finally, South Dakota and Oklahoma present a mixed behavior, showing evidence for both under and overperforming hospitals.

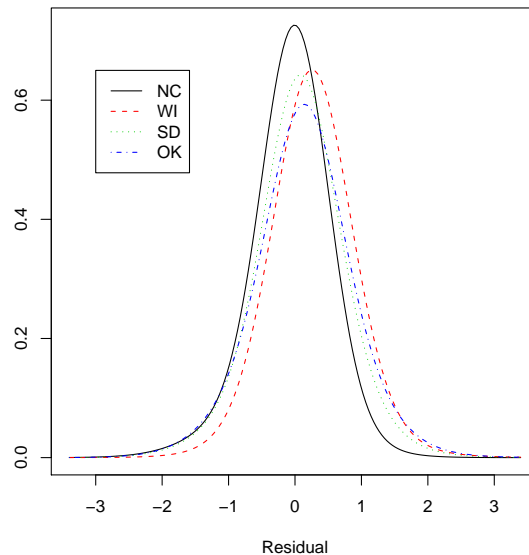


FIGURE 9. Mean predictive density for four representative states: North Carolina (NC), Wisconsin (WI), South Dakota (SD) and Oklahoma (OK).

8. DISCUSSION

We have formulated a novel extension of the Dirichlet process 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 it. After examining some of the theoretical properties of the model, we describe a computationally efficient implementation and demonstrate 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. We also offer heatmaps to summarize the clustering structure generated by the model. Attractively, while being non-parametric, the nDP encompasses a number of typical parametric and non-parametric models as limiting cases.

One natural generalization of the nDP is to replace the $\text{beta}(1, \alpha)$ and $\text{beta}(1, \beta)$ 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 $\text{beta}(a_k, b_k)$ forms, with the DP corresponding to the special case $a_k = 1, b_k = \alpha$. Similarly, by using $\text{beta}(a_k, b_k)$ and $\text{beta}(c_k, d_k)$ respectively, 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 NDP

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

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

Finally

$$\begin{aligned} \text{Cov}(G_j(B), G_k(B)) &= \frac{H(B)(1 - H(B))}{(\alpha + 1)(\beta + 1)} + H^2(B) - H^2(B) \\ &= \frac{H(B)(1 - H(B))}{(\alpha + 1)(\beta + 1)} \end{aligned}$$

and

$$\text{Cor}(G_j(B), G_k(B)) = \frac{\text{Cov}(G_j(B), G_k(B))}{\sqrt{\mathbb{V}(G_j(B))\mathbb{V}(G_k(B))}} = \frac{1}{\alpha + 1}$$

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

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

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

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

APPENDIX B. PROOF OF THEOREM 1

Let $P^{\infty\infty}(\boldsymbol{\theta})$ be the joint probability measure induced by the nDP for the vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_J)$, and let $P^{LK}(\boldsymbol{\theta})$ be the corresponding joint measure under the LK truncation. Then

$$\begin{aligned} \int |P^{LK}(\mathbf{y}) - P^{\infty\infty}(\mathbf{y})| d\mathbf{y} &\leq \int \int p(\mathbf{y} | \boldsymbol{\theta}) |P^{LK}(d\boldsymbol{\theta}) - P^{\infty\infty}(d\boldsymbol{\theta})| d\mathbf{y} \\ &= \int |P^{LK}(d\boldsymbol{\theta}) - P^{\infty\infty}(d\boldsymbol{\theta})| \\ &= 2 \sup_{A \in \Theta} |P^{LK}(A) - P^{\infty\infty}(A)| \end{aligned}$$

where the last equality is due to Scheffe's lemma. This means that the total variation distance between the true and approximated marginal densities can be bounded by the total variation distance between

the priors. Now,

$$\begin{aligned} \sup_{A \in \Theta} |P^{LK}(A) - P^{\infty\infty}(A)| &\leq 2(1 - \mathbb{P}[\zeta_j \leq K - 1 \forall j, \xi_{ij} \leq L - 1 \forall i, j]) \\ &= 2(1 - \mathbb{P}[\zeta_j \leq K - 1 \forall j] \times \\ &\quad \mathbb{P}[\xi_{ij} \leq L - 1 \forall i, j | \zeta_j \leq K - 1 \forall j]) \end{aligned}$$

Consider first the case $L = \infty$ and $K < \infty$. Then

$$\mathbb{P}[\xi_{ij} \leq L - 1 \forall i, j | \zeta_j \leq K - 1 \forall j] = 1$$

and

$$\begin{aligned} \mathbb{P}[\zeta_j \leq K - 1 \forall j] &= \mathbb{E} \left\{ \left[\sum_{s=1}^{K-1} \pi_s^* \right]^J \right\} \\ &\geq \left[\sum_{s=1}^{K-1} \mathbb{E}(\pi_s^*) \right]^J \end{aligned}$$

by Jensen's inequality. Now,

$$\mathbb{E}(\pi_s^*) = \frac{1}{1 + \alpha} \left(\frac{\alpha}{1 + \alpha} \right)^{s-1} \Rightarrow \sum_{s=1}^{K-1} \mathbb{E}(\pi_s^*) = 1 - \left(\frac{\alpha}{1 + \alpha} \right)^{K-1}$$

And therefore

$$\mathbb{P}[\zeta_j \leq K - 1 \forall j] \geq \left[1 - \left(\frac{\alpha}{1 + \alpha} \right)^{K-1} \right]^J$$

If $L < \infty$ and $K = \infty$. Then

$$\mathbb{P}[\zeta_j \leq K - 1 \forall j] = 1$$

and

$$\begin{aligned} \mathbb{P}[\xi_{ij} \leq L-1 \forall i, j | \zeta_j \leq K-1 \forall j] &= \mathbb{P}[\xi_{ij} \leq L-1 \forall i, j] \\ &= \sum_{(m_1, \dots, m_J) \in C_J} \mathbb{P}[\xi_{ij} \leq L-1 \forall i, j | (m_1, \dots, m_J)] \times \\ &\quad \mathbb{P}[(m_1, \dots, m_J)] \end{aligned}$$

where $(m_1, \dots, m_J) \in C_J$ is an assignment of J distributions to atoms $\{G_k^*\}_{k=1}^\infty$ such that there are m_1 distinct distributions appearing only once, m_2 that occur exactly twice and so on, and C_J is the set of all such possible assignments. From Antoniak (1974)

$$\mathbb{P}[(m_1, \dots, m_J)] = \frac{J! \Gamma(J + \alpha)}{\prod_{j=1}^J m_j! j^{m_j}} \frac{\alpha^{\sum_{j=1}^J m_j}}{\Gamma(\alpha)}$$

and since $\{G_k^*\}_{k=1}^K$ are in turn independent samples from a DP,

$$\begin{aligned} \mathbb{P}[\xi_{ij} \leq L-1 \forall i, j | (m_1, \dots, m_J)] &= \prod_{j=1}^J \left\{ \mathbb{E} \left[\left(\sum_{l=1}^{L-1} w_{l1}^* \right)^{jn} \right] \right\}^{m_j} \\ &\geq \prod_{j=1}^J \left[\left(\sum_{l=1}^{L-1} \mathbb{E}(w_{l1}^*) \right)^{jn} \right]^{m_j} \\ &= \left[1 - \left(\frac{\beta}{\beta+1} \right)^{L-1} \right]^{n \sum_{j=1}^J j m_j} \\ &= \left[1 - \left(\frac{\beta}{\beta+1} \right)^{L-1} \right]^{nJ} \end{aligned}$$

since $\sum_{j=1}^J j m_j = J$ for any configuration (m_1, \dots, m_J) . Therefore,

$$\begin{aligned} \mathbb{P}[\xi_{ij} \leq L - 1 \forall i, j | \zeta_j \leq K - 1 \forall j] &\geq \left[1 - \left(\frac{\beta}{\beta + 1} \right)^{L-1} \right]^{nJ} \sum_{(m_1, \dots, m_J) \in C_J} \mathbb{P}[(m_1, \dots, m_J)] \\ &= \left[1 - \left(\frac{\beta}{\beta + 1} \right)^{L-1} \right]^{nJ} \end{aligned}$$

Finally, the case $K < \infty$ and $L < \infty$ combines both results. As before

$$\mathbb{P}[\zeta_j \leq K - 1 \forall j] \geq \left[1 - \left(\frac{\alpha}{1 + \alpha} \right)^{K-1} \right]^J$$

Since K is finite, the expressions of Antoniak (1974) cannot be used in this case. However, we do not need an explicit expression for $\mathbb{P}[(m_1, \dots, m_J)]$ since we only need its sum, which is 1. Therefore

$$\mathbb{P}[\xi_{ij} \leq L - 1 \forall i, j | \zeta_j \leq K - 1 \forall j] \geq \left[1 - \left(\frac{\beta}{\beta + 1} \right)^{L-1} \right]^{nJ}$$

as before.

APPENDIX C. PROOF OF COROLLARY 1

By Bayes theorem,

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

REFERENCES

- [1] Antoniak, C. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics* **2**, 1152–1174.

- [2] Bigelow, J. L. and D. B. Dunson (2005). Semiparametric classification in hierarchical functional data analysis. Technical report, Institute of Statistics and Decision Sciences, Duke University.
- [3] Blackwell, D. and J. B. MacQueen (1973). Ferguson distribution via pólya urn schemes. *The Annals of Statistics* **1**, 353–355.
- [4] Blei, D. M., T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum (2004). Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems 16*.
- [5] Bush, C. A. and S. N. MacEachern (1996). A semiparametric bayesian model for randomised block designs. *Biometrika* **83**, 275–285.
- [6] Chib, S. and B. H. Hamilton (2002). Semiparametric bayes analysis of longitudinal data treatment models. *Journal of Econometrics* **110**, 67–89.
- [7] Dahl, D. (2003). An improved merge-split sampler for conjugate dirichlet process mixture models. Technical report, Department of Statistics, University of Wisconsin.
- [8] DeIorio, M., P. Müller, G. L. Rosner, and S. N. MacEachern (2004). An anova model for dependent random measures. *Journal of the American Statistical Association* **99**, 205–215.
- [9] Duan, J., M. Guindani, and A. Gelfand (2005). Generalized spatial dirichlet process models. Technical report, Institute of Statistics and Decision Sciences, Duke University.
- [10] Dunson, D. (2005). Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association* **100**, 618–627.
- [11] Dunson, D. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics*. In press.
- [12] Dunson, D. B., A. H. Herring, and S. A. Mulheri-Engel (2005). Bayesian selection and clustering of polymorphisms in functionally-related genes. Technical report, Institute of Statistics and Decision Sciences, Duke University.
- [13] Dunson, D. B., N. Pillai, and J.-H. Park (2004). Bayesian density regression. Technical report, Institute of Statistics and Decision Sciences, Duke University.

- [14] Escobar, M. D. (1994). Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association* **89**, 268–277.
- [15] Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of American Statistical Association* **90**, 577–588.
- [16] Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.
- [17] Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2**, 615–629.
- [18] Gelfand, A. E., A. Kottas, and S. N. MacEachern (2005). Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association* **100**, 1021–1035.
- [19] Green, P. and S. Richardson (2001). Modelling heterogeneity with and without the dirichlet process. *Scandinavian Journal of Statistics* **28**, 355–375.
- [20] Griffin, J. E. and M. F. J. Steel (2006). Order-based dependent dirichlet processes. *Journal of the American Statistical Association* **101**, 179–194.
- [21] Hirano, K. (2002). Semiparametric bayesian inference in autoregressive panel data models. *Econometrica* **70**, 781–799.
- [22] Ishwaran, H. and L. James (2002). Approximate dirichlet process computing in finite normal mixtures: Smoothing and prior information. *Journal of Computational and Graphical Statistics* **11**, 508–532.
- [23] Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.
- [24] Ishwaran, H. and M. Zarepour (2002). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica* **12**, 941–963.

- [25] Jain, S. and R. M. Neal (2000). A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. Technical report, Department of Statistics, University of Toronto.
- [26] Kacperczyk, M., P. Damien, and S. G. Walker (2003). A new class of bayesian semiparametric models with applications to option pricing. Technical report, University of Michigan Bussiness School.
- [27] Kleinman, K. and J. Ibrahim (1998). A semi-parametric bayesian approach to generalized linear mixed models. *Statistics in Medicine* **17**, 2579–2596.
- [28] Kottas, A., M. D. Branco, and A. E. Gelfand (2002). A nonparametric bayesian modeling approach for cytogenetic dosimetry. *Biometrics* **58**, 593–600.
- [29] Laws, D. J. and A. O’Hagan (2002). A hierarchical bayes model for multilocation auditing. *Journal of the Royal Statistical Society, Series D* **51**, 431–450.
- [30] Lo, A. (1984). On a class of bayesian nonparametric estimates: I. density estimates. *Annals of Statistics* **12**, 351–357.
- [31] MacEachern, S. N. (1994). Estimating normal means with a conjugate style dirichlet process prior. *Communciations in Statistics, Part B - Simulation and Computation* **23**, 727–741.
- [32] MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55.
- [33] MacEachern, S. N. (2000). Dependent dirichlet processes. Technical report, Ohio State University, Department of Statistics.
- [34] MacEachern, S. N. and P. Müller (1998). Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223–238.
- [35] Medvedovic, M. and S. Sivaganesan (2002). Bayesian infinite mixture model-based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–1206.
- [36] Mukhopadhyay, S. and A. Gelfand (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association* **92**, 633–639.

- [37] Müller, P., F. Quintana, and G. Rosner (2004). Hierarchical meta-analysis over related non-parametric bayesian models. *Journal of Royal Statistical Society, Series B* **66**, 735–749.
- [38] Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–.
- [39] Ongaro, A. and C. Cattaneo (2004). Discrete random probability measures: a general framework for nonparametric bayesian inference. *Statistics and Probability Letters* **67**, 33–45.
- [40] Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. In T. S. Ferguson, L. S. Shapeley, and J. B. MacQueen (Eds.), *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, pp. 245–268. Hayward, CA:IMS.
- [41] Roberts, G. and O. Papaspiliopoulos (2004). Retrospective markov chain monte carlo. Technical report, Department of Mathematics and Statistics, Lancaster University.
- [42] Sethuraman, J. (1994). A constructive definition of dirichelt priors. *Statistica Sinica* **4**, 639–650.
- [43] Silverman, B. (1986). *Density Estimation*. Chapman and Hall, London.
- [44] Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2004). Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in Neural Information Processing Systems 17*.
- [45] Verbeke, G. and E. Lesaffre (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91**, 217–221.