Generalized Species Sampling Priors With Latent Beta Reinforcements

Edoardo M. Airoldi, Thiago Costa, Federico Bassetti, Fabrizio Leisen, and Michele Guindani

Many popular Bayesian nonparametric priors can be characterized in terms of exchangeable species sampling sequences. However, in some applications, exchangeability may not be appropriate. We introduce a novel and probabilistically coherent family of nonexchangeable species sampling sequences characterized by a tractable predictive probability function with weights driven by a sequence of independent Beta random variables. We compare their theoretical clustering properties with those of the Dirichlet process and the two parameters Poisson–Dirichlet process. The proposed construction provides a complete characterization of the joint process, differently from existing work. We then propose the use of such process as prior distribution in a hierarchical Bayes' modeling framework, and we describe a Markov chain Monte Carlo sampler for posterior inference. We evaluate the performance of the prior and the robustness of the resulting inference in a simulation study, providing a comparison with popular Dirichlet process mixtures and hidden Markov models. Finally, we develop an application to the detection of chromosomal aberrations in breast cancer by leveraging array comparative genomic hybridization (CGH) data. Supplementary materials for this article are available online.

KEY WORDS: Bayesian nonparametrics; Cancer; Genomics; MCMC; Predictive probability functions; Random partitions.

1. INTRODUCTION

Bayesian nonparametric priors have become increasingly popular in applied statistical modeling in the last few years. Examples of their wide area of applications range from variable selection in genetics (Kim, Tadesse, and Vannucci 2006) to linguistics (Teh 2006a; Wallach, Sutton, and McCallum 2008), psychology (Navarro et al. 2006), human learning (Griffiths et al. 2008), image segmentation (Sudderth and Jordan 2008) and applications to the neurosciences (Jbabdi, Woolrich, and Behrens 2009). See also Hjort et al. (2010). The increased interest in nonparametric Bayesian approaches is motivated by a number of attractive inferential properties. For example, Bayesian nonparametric priors are often used as flexible models to describe the heterogeneity of the population of interest, as they implicitly induce a clustering of the observations into homogenous groups. Such a clustering can be seen as a realization of a random partition scheme and can often be characterized in terms of a species sampling (SS) allocation rule. More formally, an SS sequence is a sequence of random variables $X_1, X_2, \ldots$, characterized by the predictive probability functions,

$$P\{X_{n+1} \in \cdot \mid X_1, \ldots, X_n\} = \sum_{j=1}^{n} q_{n,j}\delta_{X_j}(\cdot)$$
$$+ q_{n,n+1}G_0(\cdot), \qquad (1)$$

where $\delta_x(\cdot)$ denotes a point mass at x, $q_{n,j}$ $(j = 1, \ldots, n+1)$ are nonnegative functions of $(X_1, \ldots, X_n)$, or weights, such that $\sum_{j=1}^{n+1} q_{n,j} = 1$, and G_0 is a nonatomic probability measure (Pitman 1996). Collecting the unique values of X_j, (1) can be rewritten as

$$P\{X_{n+1} \in \cdot \mid X_1, \ldots, X_n\} = \sum_{j=1}^{K_n} q_j^*\delta_{X_j^*}(\cdot) + q_{K_n+1}^*G_0(\cdot), \quad (2)$$

where K_n is the (random) number of distinct values, say $(X_1^*, \ldots X_{K_n}^*)$, in the vector $X(n) = (X_1, \ldots, X_n)$ and q_j^* are suitable nonnegative weights. In particular, an exchangeable SS sequence is characterized by weights q_j^* that depend only on $\mathbf{n}_n = (n_{1n}, \ldots, n_{K_n n})$, where n_{jn} is the frequency of X_j^* in $X(n)$ (Fortini, Ladelli, and Regazzini 2000; Hansen and Pitman 2000; Lee et al. 2013). The most well-known example of predictive rules of type (1) is the Blackwell–MacQueen sampling rule, which implicitly defines a Dirichlet process (DP, Blackwell and MacQueen 1973; Ishwaran and Zarepour 2003). The predictive rule characterizing a DP with mass parameter θ and base measure $G_0(\cdot)$, $DP(\theta, G_0)$, sets $q_{n,j} = \frac{1}{n+\theta}$ and $q_{n,n+1} = \frac{\theta}{n+\theta}$ in (1).

Whenever the weights $q_j^*(\mathbf{n}_n)$ and $q_{K_n+1}^*(\mathbf{n}_n)$ do not depend only on $\mathbf{n}_n$, the sequence $(X_1, X_2, \ldots)$ is not exchangeable. Models with nonexchangeable random partitions have recently appeared in the literature, for example, to allow for partitions that depend on covariates. Park and Dunson (2010) derived a generalized product partition model (GPPM) in which the partition process is predictor-dependent. Their GPPM generalizes the

DP clustering mechanism to relax the exchangeability assumption through the incorporation of predictors, implicitly defining a generalized Pólya urn scheme. Müller and Quintana (2010) defined a product partition model that includes a regression on covariates, allowing units with similar covariates to have greater probability of being clustered together. Arguably, the previous models provide an implicit modification of the predictive rule (1) where the weights can be seen as function of some external predictor. Alternatively, other authors model the weights $q_j(\mathbf{n}_n)$ explicitly, for instance, by specifying the weights as a function of the distance between data points (Dahl, Day, and Tsai 2008; Blei and Frazier 2011). However, the general properties of the random partitions generated by such processes have not been specifically addressed.

In this article, we introduce a novel and probabilistically coherent family of nonexchangeable species sampling sequences, where the weights are specified sequentially and do not depend on the cluster sizes, but instead they depend on the realizations of a set of latent variables. Working within this family, we propose a simple characterization of the weights in the predictive probability function as a product of independent Beta random variables. This strategy leads to a well-defined random allocation scheme for the observables. The resulting process, which we call Beta-GOS process, is a special case of a generalized Ottawa sequence (GOS), recently introduced by Bassetti, Crimaldi, and Leisen (2010).

In Section 2, we discuss the properties of the Beta-GOS process, with particular regard to the clustering induced on the observables. In Section 3, we study the asymptotic distribution of the (random) number of distinct values in the sequence, K_n, for some natural specifications of the weights, and we compare those results with the well-known asymptotic results characterizing the DP and the two-parameters Poisson–Dirichlet process. In many applications, nonparametric processes are often used within hierarchical models to specify the prior distribution of some parameters of the distribution of the observables. For example, this is a popular use for mixtures of DP. Similarly, the Beta-GOS process can also be used to define a prior in a hierarchical model. In Section 4, we outline a basic hierarchical model based on the Beta-GOS process and we outline the basic steps of a Markov chain Monte Carlo (MCMC) sampler for posterior inference. In Section 5, we design a set of simulations to compare the behavior of the Beta-GOS model with that of DP mixtures and hidden Markov models (HMM) in terms of cluster estimation. Our results suggest that the Beta-GOS can be seen as a robust alternative to the DP when exchangeability would be hardly justified in practice, but still there is a need to describe the heterogeneity of our observations by virtue of an unsupervised clustering of the data. The Beta-GOS also provides an alternative to customary HMM, especially when the number of states is unknown or the Markovian structure is expected to vary with time.

In Section 6, we analyze two published datasets of genomic and transcriptional aberrations (Chin et al. 2006; Curtis et al. 2012). Bayesian models for array comparative genomic hybridization (CGH) data have been recently investigated by Guha, Li, and Neuberg (2008), DeSantis et al. (2009), Baladandayuthapani et al. (2010), Du et al. (2010), Cardin et al. (2011), and Yau et al. (2011), among others. Guha et al. proposed a four state

homogenous Bayesian HMM to detect copy number amplifications and deletions and partition tumor DNA into regions (clones) of relatively stable copy number. DeSantis et al. extended this approach and developed a supervised Bayesian latent class approach for classification of the clones that relies on a heterogenous HMM to account for local dependence in the intensity ratios. In a heterogenous HMM, the transition probabilities between states depend on each single clone or the distance between adjacent clones (Marioni et al. 2006). Du et al. proposed a sticky hierarchical DP-HMM (Fox et al. 2011; Teh et al. 2006b) to infer the number of states in an HMM, while also imposing state persistence. Yau et al. (2011) also proposed a nonparametric Bayesian HMM, but used instead a DP mixture to model the likelihood in each state. With respect to those proposals, we also assume that the number of states is unknown, as it is typical in a Bayesian nonparametric setting, but we do not need a parameter to explicitly account for state persistence. This is because the Beta-GOS model is "nonhomogenous" by design, as the weights in the species sampling mechanisms adapt to take into account the local dependence in the clones' intensities. We show that the Beta-GOS is able to identify clones that have been linked to breast cancer pathophysiologies in the medical literature.

We conclude with some final remarks in Section 7. Technical details and proofs of theorems and lemmas are provided in the Appendix.

2. THE BETA-GOS PROCESS

As anticipated, the Beta-GOS process is defined by a modification of the predictive rule that characterizes the species sampling mechanism (1), where the weights are a product of independent Beta random variables. More in general, we start considering a sequence of random variables $(X_n)_{n \geq 1}$ characterized by the predictive distributions

$$P\{X_{n+1} \in \cdot \mid X(n), W(n)\} = \sum_{j=1}^{n} p_{n,j}\delta_{X_j}(\cdot) + r_n G_0(\cdot), \quad (3)$$

where $W(n) = (W_1, \ldots, W_n)$ is a vector of independent random variables W_k taking values in [0, 1], and the weights are defined by

$$p_{n,j} := (1 - W_j)\prod_{i=j+1}^{n} W_i, \qquad r_n := \prod_{i=1}^{n} W_i. \quad (4)$$

The prediction rule (3) defines a special case of a GOS, introduced in Bassetti, Crimaldi, and Leisen (2010), a type of generalized Pólya urn sequences where the reinforcement is randomly determined by the realizations of a latent process (see also Guha 2010, for an alternative proposal). Except from a few special cases, the X_i's in a GOS are not exchangeable. However, it can be shown that these sequences maintain some properties typical of exchangeable sequences. Most notably, any GOS is *conditionally identically distributed* (CID), that is, for all $n > 0$, the X_{n+j}'s, $j \geq 1$, are identically distributed, conditionally on $(X_1, \ldots, X_n, W_1, \ldots, W_n)$. Hence, the X_i's are also marginally identically distributed. Note that a CID sequence is not necessarily stationary. If a CID sequence is also stationary, then it is exchangeable. Finally, although no

representation theorem is known for CID sequences, it can be shown that given any bounded and measurable function f, the predictive mean $E[f(X_{n+1})|X_1, \ldots, X_n]$ and the empirical mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ converge to the same limit as n goes to infinity. For details, see Berti, Pratelli, and Rigo (2004), where CID sequences have been first introduced. The predictive rule (3) reduces to known cases with a suitable choice of the latent W_n's; for instance if $W_n := (\theta + n - 1)/(\theta + n)$, then (3) coincides with the Blackwell–MacQueen sampling rule characterizing a DP(θ, G_0).

In this article, we propose $(W_n)_{n \geq 1}$ be a sequence of independent Beta(α_n, β_n) random variables and we call the resulting $(X_1, X_2, \ldots)$ a Beta-GOS sequence. The choice of Beta latent variables allows for a flexible specification of the species sampling weights, while retaining a simple and interpretable model together with computational simplicity (see later sections). The allocation rule can also be described in terms of a preferential attachment scheme, where each observation is attached to any of the preceding by means of a "geometric-type" assignment. In this scheme, every individual X_i is characterized by a random weight (or "mark"), $1 - W_i$. We can interpret each individual mark as an individual specific attractivity index, as it determines the probability that the next observation will be clustered with X_i. More precisely, the first individual is assigned a random value (or "tag") X_1, according to G_0. Now, suppose we have $X_1, \ldots, X_n$ together with their marks up to time n, $(1 - W_1, \ldots, 1 - W_n)$. Then, the $(n + 1)$th individual will be assigned the same tag as X_n with probability $1 - W_n$; the probability of pairing X_{n+1} to X_{n-1} will be $W_n(1 - W_{n-1})$, and so forth. In general, $p_{n,j}$ will be the product of the repulsions W_i for the latest $n - j$ subjects and the jth attractivity $1 - W_j$. Summarizing, X_{n+1} will result in a new tag (i.e., $X_{n+1} \sim G_0$) with probability r_n, or will be clustered together with a previously observed tag, say X_k^*, with probability $\sum_{j:X_j=X_k^*} p_{n,j}$. In the next section, we discuss the clustering behavior induced by different specifications of the Beta weights in more detail.

3. CLUSTERING BEHAVIOR OF THE BETA-GOS

The predictive rule (3) implicitly defines a random partition of the set $\{1, \ldots, n\}$ into K_n blocks. In probability theory, K_n is also referred to as the *length* of the partition. Knowledge of the behavior of K_n is useful to understand the clustering structure implied by (3). For instance, for a DP(θ, G_0), it is well known that $K_n/\log(n)$ converges almost surely to a constant, indeed the mass parameter θ. This asymptotic behavior is sometimes described as a "self-averaging" property of the partition (Aoki 2008). From a practical point of view, since $K_n/\log(n)$ converges to a constant, then in the limit K_n is essentially $\theta \log(n)$; thus, for modeling purposes it suffices to consider only the first moment of K_n. In the case of the two parameter Poisson–Dirichlet process, the length of the partition K_n (suitably rescaled) converges instead to a random variable. More precisely, for a PD(α, θ), with $0 < \alpha < 1$, $\theta > -\alpha$, then K_n/n^α converges a.s. to a strictly positive random variable (see Theorem 3.8 in Pitman 2006). Therefore, the PD sequence is nonself-averaging. When the limit of K_n is essentially a random variable, extra care is needed in the prior assessment of the parameters of the nonparametric prior, since the clustering

behavior is ultimately governed by the whole distribution of the limit random variable. For the Beta-GOS process, we focus on the following two cases:

(i) $\alpha_n = a > 0$ and $\beta_n = b > 0$ for all $n \geq 1$;
(ii) $\alpha_n = \theta - 1 + n$ $(\theta > 0)$ and $\beta_n \geq 1$ for all $n \geq 1$.

Then, we can prove the following.

Proposition 1. Let K_n be the length of the partition induced by a Beta-GOS, with G_0 nonatomic and $W_n \sim$ Beta(α_n, β_n) $(n \geq 1)$.

(a) If $\alpha_n = n + \theta - 1$, $\beta_n = 1$, for given $\theta > 0$, $K_n/\log(n)$ converges in distribution to a Gamma$(\theta, 1)$ random variable.
(b) If $\alpha_n = n + \theta - 1$, $\beta_n = \beta$, $(\theta > 0, \beta > 1)$ or if $\alpha_n = a$, $\beta_n = b$, $(a > 0, b > 0)$, then K_n converges almost surely to a finite random variable K_∞. In particular, if $\alpha_n = a$, $\beta_n = b$, then

$$E[e^{-tK_\infty}] = e^{-t} \sum_{m \geq 0} (e^{-t} - 1)^m \prod_{j=1}^{m} \frac{(a)^{(j)}}{(a+b)^{(j)} - (a)^{(j)}},$$

where $(t)^{(m)} = t(t+1)\ldots(t+m-1)$ and

$$E[(K_\infty - 1)\ldots(K_\infty - m)] = m! \prod_{j=1}^{m} \frac{(a)^{(j)}}{(a+b)^{(j)} - (a)^{(j)}}.$$

The proof is detailed in the Appendices, where we also provide a general formula for the probability distribution, the kth moment and the generating function of K_n. The result in Proposition 1(a) represents a case of a quite natural (nonexchangeable) partition model for which the length K_n scales as $\log(n)$ but is not self-averaging. When $\alpha_n = a$, $\beta_n = b$, according to Proposition 1(b), the convergence of K_n to a finite random variable naturally implies the creation of a few big clusters, as n increases. Instead, for $\alpha_n = n + \theta - 1$, $\beta_n = 1$, the mean length of the partition depends on the value of θ, since a bigger number of clusters is associated on average with greater values of θ. However, as θ increases so does the asymptotic variability of K_n; therefore, in this case, a Beta-GOS process can be used to represent uncertainty on K_n (by the lack of the self-averaging property of the process). By means of simulations, we have also confirmed that, for small values of θ, the partition of n elements is skewed, that is, it is characterized by a small number of big clusters as well as a few small clusters. As θ increases, the sizes of the clusters decrease accordingly, the observations being grouped into clusters of relatively fewer elements. This is similar to what happens for the DP, and indeed in this case the parameter θ could be interpreted as a mass parameter for the Beta-GOS.

The parameters of the W_i's can be chosen to model the autocorrelation expected a priori in the dynamics of the sequence. The probability of a tie may decrease with n and atoms that have been observed at farthest times may have a greater probability to be selected if they have also been observed more recently. Such considerations may be helpful to guide prior assessment of the Beta hyperparameters. For given $n \geq 1$, taking

expectations with respect to the weights W_i's we obtain

$$E[r_n] = \prod_{j=1}^{n} \frac{\alpha_j}{\alpha_j + \beta_j}, \qquad E[p_{n,k}] = \frac{\beta_k}{\alpha_k + \beta_k} \prod_{j=k+1}^{n} \frac{\alpha_j}{\alpha_j + \beta_j}$$
$$k = 1, \ldots, n. \quad (5)$$

Under (a), it follows that $E[r_n] = (a/(a+b))^n$ and $E[p_{n,k}] = (a/(a+b))^{n-k}(b/(a+b))$; hence, the probabilities of ties depend only on the lag $n-k$ and decrease exponentially as a function of $n-k$. Under (b),

$$E[r_n] = \prod_{j=1}^{n} \frac{j+\theta-1}{j+\theta-1+\beta} = \frac{\Gamma(\theta+n)\Gamma(\theta+\beta)}{\Gamma(\theta+\beta+n)\Gamma(\theta)}$$

$$E[p_{n,k}] = \frac{\beta}{k+\theta-1+\beta} \prod_{j=k+1}^{n} \frac{j+\theta-1}{j+\theta-1+\beta}$$
$$= \beta \frac{\Gamma(\theta+n)\Gamma(\theta-1+\beta+k)}{\Gamma(\theta+\beta+n)\Gamma(\theta+k)} \quad k = 1, \ldots, n.$$

Thus, for $n, k \to +\infty$, $E[r_n] \sim \frac{1}{n^\beta}$ and $E[p_{n,k}] \sim \frac{k^{\beta-1}}{n^\beta}$. For example, if $\theta = 1$ and $\beta = 2$, then $\alpha_j = j$ and $\beta_j = 2$ and $E[r_n] = \frac{2}{(1+n)(2+n)}$, $E[p_{n,k}] = \frac{2(k+1)}{(n+1)(n+2)}$, $k = 1, \ldots, n$, so that the weights decrease linearly as a function of the lag $n-k$. If $\alpha_j = \theta - 1 + j$ ($\theta > 0$) and $\beta_j = 1$, then $E[r_n] = \frac{\theta}{\theta+n}$ and $E[p_{n,k}] = \frac{1}{\theta+n}$, $k = 1, \ldots, n$, that is, any observation has the same weight. This latter specification leads to an expression similar to that in the Blackwell–McQueen Pólya Urn characterization of the DP; however, this identity is true only in expectation, and the clustering behavior of the DP and Beta-GOS process with $\alpha_j = j + \theta - 1$ and $\beta_j = 1$ may be quite different, as it is evident from Proposition 1.

In practice, the determination of the parameters of the Beta distributions is not trivial, and may be problem-dependent, especially given the sensitivity of the clustering behavior to the values of α_j and β_j. As a general rule, following what it is usually done with DP priors, one should consider eliciting the parameters on the basis of the expected number of clusters $E(K_n) = 1 + \sum_{j=1}^{n-1} E[r_j]$. For example, one should set $\alpha_j = a$ and $\beta_j = b$ to represent a short memory process, and the values of a, b can be chosen based on the asymptotic relationship $E(K_n) \approx \frac{a+b}{b}$. We further suggest to choose $b = 1$, or anyway $b < a$, to encourage a priori low autocorrelation of the sequence, since then $E(p_{n,n}) < 0.5$. As a matter of fact, we implemented those suggestions in the application to the array CGH data presented in Section 6, where biological considerations lead to further expect the true number of states to be around 4. On the other hand, one should set $\alpha_j = j + \theta - 1$, $\beta_j = 1$ to represent a long memory process, and then choose θ based on $E(K_n) = \sum_{j=0}^{n-1} \frac{\theta}{\theta+j} \sim \theta \log(n)$, for large n. The latter, single-parameter, formulation should be the default choice in those applications where prior information on the expected number of clusters is unavailable otherwise. Alternative strategies are possible. For example, one could consider second moments, or otherwise require further constraints on the expected autocorrelation of the sequence. However, we leave the exploration of those possibilities to future work. See also the discussion at the end of Section 4.2.

Finally, we note the functional form of (4) may initially suggest a relationship with the stick-breaking characterization of the DP. However, the stick-breaking construction characterizes the representation of the DP as a random measure, not the corresponding predictive probability function. Furthermore, the sequence generated by a DP is exchangeable, whereas a Beta-GOS in general is not and includes the DP as a special case. As a matter of fact, if one would like to stress the "stick-breaking" analogy anyway, one should more properly interpret (3) in terms of an *inverse* stick-breaking, since each $p_{n,j}$, which defines the probability of a tie, say $X_{n+1} = X_j$, does not depend on the W_i's observed before time j, $j = 1, \ldots, n$, whereas the probability of choosing a new tag depends only on the part of the stick that is left at time n. This is evident if we consider the alternative characterization of (3) with $p_{n,j} = W_j \prod_{i=j+1}^{n}(1 - W_i)$ and $r_n(W_1, \ldots, W_n) = \prod_{i=1}^{n}(1 - W_i)$, $W_i \sim \text{Beta}(\beta_i, \alpha_i)$ and choose $\beta_i = 1$ and $\alpha_i = \theta$ as in the DP. Then, $p_{n,j} = W_j \prod_{i=j+1}^{n}(1 - W_i)$, $j = 1, \ldots, n$. For $n = 3$, $p_{3,1} = W_1(1 - W_2)(1 - W_3)$, $p_{3,2} = W_2(1 - W_3)$, $p_{3,3} = W_3$. By contrast, in a DP each piece of the unitary stick is defined from what is left by the previous ones.

4. A BETA-GOS HIERARCHICAL MODEL

In this section, we show how the Beta-GOS process could be used as a prior in a hierarchical model, and we discuss a straightforward MCMC sampling algorithm for posterior inference.

4.1 The Hierarchical Model

Beta-GOS processes can be used to model dependencies between nonexchangeable observations. Let $Y = (Y_1, \ldots, Y_m)^T$ be a vector of observations, for example, a time series. Then, following a Bayesian approach, we can assume that the data can be described by a hierarchical model as

$$Y_i | \mu_i \overset{\text{ind.}}{\sim} p(y_i | \mu_i), \quad i = 1, \ldots, m, \quad (6)$$

for some probability density $p(\cdot | \mu_i)$, where the vector $(\mu_1, \ldots, \mu_m)^T$ is a realization of a Beta-GOS process with parameters α_i, β_i, $i = 1, \ldots, m$, and base measure G_0, which we succinctly denote as

$$\mu_1, \ldots, \mu_m \sim \text{Beta-GOS}(\boldsymbol{\alpha}_m, \boldsymbol{\beta}_m, G_0), \quad (7)$$

that is, is a sample from a random distribution characterized by the predictive rule (3), for some $W_i \sim \text{Beta}(\alpha_i, \beta_i)$, $i = 1, \ldots, m$. As noted in Section 2, any Beta-GOS defines a CID sequence. In particular, marginally $\mu_i \sim G_0$, $i = 1, \ldots, m$. Therefore, G_0 can be regarded as a centering distribution, as in DP mixture models: G_0 can represent a vague parametric prior assumption on the distribution of the parameters of interest. The hierarchical model may be extended by putting hyperpriors on the remaining parameters of the model, including the parameters of the Beta-GOS ($\boldsymbol{\alpha}_m, \boldsymbol{\beta}_m, G_0$), although here we focus on the characterization of the behavior of the Beta-GOS for fixed choices of the Beta parameters.

We conclude this section by noting that the sequence $Y_1, Y_2, \ldots$, defined through (6) and (7), with joint density

$$\int \prod_{i=1}^{m} p(y_i | \mu_i) \pi(d\mu_1, \ldots, d\mu_m), \quad m \geq 0,$$

and $\pi(\cdot) \equiv$ Beta-GOS($\boldsymbol{\alpha}_m, \boldsymbol{\beta}_m, G_0$), is also a CID sequence. Therefore, although not exchangeable, the Y_{n+j}'s, $j \geq 1$ are conditionally identically distributed given $(Y_1, \ldots, Y_n, \mu_1, \ldots, \mu_n)$. For a proof of this statement, see Proposition 2 in the Appendices.

4.2 MCMC Posterior Sampling

Posterior inference for the model (6)–(7) entails learning about the vector of random effects μ_i and their clustering structures. As the posterior is not available in closed form, we need to revert to MCMC sampling. In this section, we describe a Gibbs sampler that relies on sampling the subsequent cluster assignments of the observations $Y_1, \ldots, Y_m$ according to the rule (3). To do this, the partition structure will be described by introducing a sequence of labels $(C_n)_{n \geq 1}$ recording the pairing of each observation according to (3), that is, which other data point, among those with index $j < i$, the ith observation has been matched to. Hence, here the label C_i is not a simple indicator of the cluster membership, as it is typical in most MCMC algorithms devised for the DP, although cluster membership can be easily retrieved by analyzing the sequence of pairings. In what follows, C_i will be sometimes referred to as the ith pairing label. In particular, if the ith observation is not paired to any of those preceding, $C_i = i$; in this case, the ith point consists of a draw from the base distribution G_0, and thus generates a new cluster. This slightly different representation of data points in terms of data-pairing labels, instead of cluster-assignment labels, turns useful to develop an MCMC sampling scheme for nonexchangeable processes, as it has been thoroughly discussed in Blei and Frazier (2011), who have shown that such representation allows for larger moves in the state space of the posterior and faster mixing of the sampler. It is easy to see that the pairing sequence $(C_n)_{n \geq 1}$ assigns $C_1 = 1$ and has distribution

$$P\{C_n = i | C_1, \ldots, C_{n-1}, W\} = P\{C_n = i | W_1, \ldots, W_{n-1}\}$$
$$= r_{n-1}\mathbb{I}\{i = n\} + p_{n-1,i}\mathbb{I}\{i \neq n\}, \quad (8)$$

for $i = 1, \ldots, n$, where $\mathbb{I}(\cdot)$ denotes, as usual, the indicator function, such that, given a set A, $\mathbb{I}(i \in A) = 1$ if $i \in A$ and 0 otherwise. As mentioned, the clustering configuration is a by-product of the representation in terms of data-pairing labels. If two observations are connected by a sequence of interim pairings, then they are in the same cluster. Given $C = (C_1, \ldots, C_m, \ldots)$, let $\Pi(C)$ denote the partition on $\mathbb{N}$ generated by C. Accordingly, if $(\mu_k^*)_{k \geq 1}$ is a sequence of independent random variables with common distribution G_0, we set $\mu_i = \mu_k^*$ if i belongs to $\Pi(C)_k$, that is, the kth block (cluster) of $\Pi(C)$. For any m and any $i \leq m$, let $C(m) = (C_1, \ldots, C_m)$, $C_{-i} = (C_1, \ldots, C_{i-1}, C_{i+1}, \ldots, C_m)$; analogously, let $W(m) = (W_1, \ldots, W_m)$, and $W_{-i} = (W_1, \ldots, W_{i-1}, W_{i+1}, \ldots, W_m)$. Then, the full conditional for the pairing indicators C_i's is

$$P\{C_i = j | C_{-i}, Y(m), W(m)\} \propto P\{C_i = j, Y(m) | C_{-i}, W(m)\}$$
$$= P\{Y(m) | C_i = j, C_{-i}, W(m)\}P\{C_i = j | C_{-i}, W(m)\}. \quad (9)$$

The second term in (9) is the prior predictive rule (8), whereas

$$P\{Y(m) | C_i = j, C_{-i}, W(m)\} = \prod_{k=1}^{|\Pi(C_{-i}, j)|} \int \prod_{l \in \Pi(C_{-i}, j)_k}$$
$$p(Y_l | \mu_j^*)G_0(d\mu_j^*),$$

where $\Pi(C_{-i}, j)$ denotes the partition generated by $(C_1, \ldots, C_{i-1}, j, C_{i+1}, \ldots, C_m)$. If G_0 and $p(y|\mu)$ are conjugate, the latter integral has a closed form solution. The nonconjugate case could be handled by appropriately adapting the algorithms of MacEachern and Müller (1998) and Neal (2000). Instead, we believe that split and merge moves as the ones considered in Jain and Neal (2007) and Dahl (2005) are more problematic to implement given the implied exchangeability of the clustering assignments in those algorithms. As far as the full conditional for the latent variables W_i's, we can show that $W_i | C(m), W_{-i}, Y(m) \sim$ Beta(A_i, B_i), where $A_i = \alpha_i + \sum_{j=i+1}^{m} \mathbb{I}\{C_j < i \text{ or } C_j = j\}$, and $B_i = \beta_i + \sum_{j=i+1}^{m} \mathbb{I}\{C_j = i\}$; hence, they depend only on the clustering configurations and not on the values of W_{-i}.

Then, consider the set of cluster centroids μ_i^*'s. The algorithm described so far allows faster mixing of the chain by integrating over the distribution of the μ_i^*. However, in case we were interested on inference on the vector $(\mu_1, \ldots, \mu_m)$, it is possible to sample the unique values at each iteration of the Gibbs sampler, from

$$P\{\mu_j^* | C(m), W(m), Y(m)\} \propto \prod_{i \in \Pi_j(m)} p(Y_i | \mu_j^*)G_0(d\mu_j^*), \quad (10)$$

where $\Pi_j(m)$ denotes the partition set of the observations such that $\mu_i = \mu_j^*$, $i = 1, \ldots, m$. Again, if $p(y|\mu)$ and G_0 are conjugate, the full conditional of μ_j^* is available in closed form, otherwise we can update μ_j^* by standard Metropolis–Hastings algorithms (Neal 2000).

In addition, we note that if a prior distribution for the Beta hyperparameters $\boldsymbol{\alpha}_m$ and $\boldsymbol{\beta}_m$, say $\pi(\boldsymbol{\alpha}_m, \boldsymbol{\beta}_m)$, were to be specified, one could implement a Metropolis–Hasting scheme to learn about their posterior distribution, since it can be shown that

$$P(\boldsymbol{\alpha}_m, \boldsymbol{\beta}_m | C(m), Y(m)) \propto \pi(\boldsymbol{\alpha}_m, \boldsymbol{\beta}_m) \prod_{i=1}^{m} \frac{B(A_i, B_i)}{B(\alpha_i, \beta_i)}, \quad (11)$$

where A_i and B_i are defined as above and $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x + y)$ denotes the Beta function. Equation (11) is an adaptation of well-known results for the DP (Escobar and West 1995) to the Beta-GOS process. A thorough study of the efficiency of this algorithm, however, as well as the choice of adequate proposal distributions is beyond the scope of this work and will be pursued elsewhere.

5. A SIMULATION STUDY

In this section, we provide a full specification for model (6)–(7) and test the properties of the Beta-GOS prior on a set of simulated examples; more specifically, we develop some comparison with the DP and popular HMM.

5.1 Model Specifications

Throughout this section, model (6)–(7) will be specified as follows. First, we assume a Gaussian distribution for the observables, $Y_i \sim$ Normal(μ_i, τ^2). The base measure G_0 is also assumed to be normal, Normal(μ_0, σ_0^2), and $\tau^2 \sim$ Inv-Gamma(a_0, b_0). The parameters of the latent Beta reinforcements, $W_i \sim$ Beta(α_i, β_i), are separately indicated in each simulation and are chosen to allow for a range of prior beliefs on the clustering behavior of the process (see Section 3). Details

Table 1. Summary statistics for the simulation study in Section 5.2. The table compares the Beta-GOS and a Dirichlet process model under different specifications of hyperparameters when the data-generating process is Beta-GOS

Data-generating process	Beta-GOS $\alpha_n = n, \; \beta_n = 1$		Beta-GOS $\alpha_n = 3, \; \beta_n = 1$	
Model-fitting method	Beta-GOS $\alpha_n = n, \; \beta_n = 1$	Dir. Proc. $\theta = 1$	Beta-GOS $\alpha_n = n, \; \beta_n = 1$	Dir. Proc. $\theta = 1$
Number of clusters				
Ground truth	5.24 ± 3.88		4.14 ± 1.81	
Estimation	4.30 ± 2.67	4.51 ± 2.62	3.61 ± 1.49	3.96 ± 1.72
Accuracy of cluster assignment	0.97 ± 0.06	0.96 ± 0.08	0.99 ± 0.01	0.99 ± 0.02
Predictive bias	4.13 ± 7.18	4.34 ± 7.27	0.67 ± 2.61	1.29 ± 3.93

of the MCMC algorithm for posterior inference and parameter estimation in the Beta-GOS model are given in Appendix A.

5.2 Model Fitting and Parameter Estimation

A first simulation study considers an ideal setting. We generate 1000 samples of 101 observations each from the Beta-GOS model (6)–(7), with (a) $\alpha_n = n, \beta_n = 1$ and (b) $\alpha_n = 3, \beta_n = 1$. The first 100 points are used for fitting purpose, whereas the 101st point is used to assess goodness-of-fit. Without much loss of generality, we fix $\mu_0 = 0$ and $\sigma_0 = 10$. We mimic the typical scale observed in the data analyzed in Section 6 and set $\tau = 0.25$ to distinguish the sample variability from the variability of the base measure. We fit the data using a Beta-GOS hierarchical model, with default Beta hyperparameters $\alpha_n = n, \beta_n = 1$, and study how well we can recover the basic characteristics of the data under such specification. We assume $\tau^2 \sim$ Inv-Gamma(2.004, 0.0063) in the model fitting. This choice of the Inverse-Gamma hyperparameters allows τ^2 to have mean around 0.25^2 and relatively large variance. In addition, we fit a DP mixture model with concentration parameter $\theta = 1$, which on the basis of Proposition 1(a) can be seen as compatible with the parameters used in our model. The mixture of DP model is fit to data using the R package "DPpackage" (Jara 2007). In this framework, the DP provides a convenient comparison; however, we should stress that, in general, the underlying exchangeability assumption may not be appropriate to fully capture the dependency structure of the data-generating process.

The results of this simple simulation study are summarized in Table 1. Table 1 reports summary statistics aimed at providing synthetic measures of the goodness-of-fit, namely, the estimated number of clusters and the accuracy of cluster assignments, together with a measure of predictive bias. Following the machine learning terminology for classification performance metrics, we call accuracy the ratio of the correct cluster assignments with respect to the total of assignments. We compute the predictive bias as follows: for each sample, and each MCMC output, we predict a new observation on the basis of the estimated parameters and the clustering configurations provided by the algorithm, say Y_{pred}. The prediction is compared with the original value, Y_{101}. The predictive bias is simply the average of $|Y_{101} - Y_{\text{pred}}|$, and can be regarded as a measure of how well the model can predict future observations. Nearly all data points were assigned to the correct clusters. The Beta-GOS appears to compare fa-

vorably in terms of predictive bias, especially when the data incorporate a stronger dependency structure. Most of the error is intrinsic to the data-generating process. As typical of most Bayesian nonparametric models, including the DP, the ability of the model and estimation algorithms to recover the ground truth may be affected by the choice of the relative magnitudes of the hyperparameters σ_0^2 and τ^2. The online supplementary materials contain additional results for several specifications of the data-generating mechanism as well as several choices of the hyperparameters for model fitting, confirming the above remarks.

5.3 Fitting Mixture of Gaussians

A second simulation study is designed to assess the robustness of the Beta-GOS framework to model misspecifications, that is, we fit the proposed nonexchangeable model to exchangeable data. The DP process provides a sensible baseline for this study. More in detail, we first generate 1000 datasets (101 observations each) from a Normal mixture model with five components. The components' means are sampled from a Normal($\mu_0 = 0, \sigma_0 = 10$), whereas their standard deviation is set either to $\tau = 0.25$ or $\tau = 0.5$ to provide some insight into the robustness of the results to different levels of noise. The vector of mixture components' weights is chosen as $\pi = (0.2, 0.35, 0.15, 0.1, 0.2)^T$. We fit the data with a DP ($\theta = 1$), and a Beta-GOS process, with (a) $\alpha_n = \beta_n = 1$, and (b) $\alpha_n = n, \beta_n = 1$. Case (a) corresponds to a process with short autocorrelation expected a priori and, asymptotically, a finite number of clusters, whereas case (b) assumes that the rescaled number of clusters, $K_n / \log(n)$, converges to a Gamma(1, 1), and $E[K_n] \sim \log(n)$. The choice of hyperparameters for the Inverse-Gamma on τ^2 sets the mean around the true value and allows for a relatively large variance. The results of the simulations are shown in Table 2. Overall, the Beta-GOS framework is quite robust to model misspecifications. For the mixture of Gaussians data, accuracy of cluster assignments was high (94%) that is better or comparable to that of the DP; correspondingly, parameters' estimates were close to the true parameter values. For all processes, the accuracy decreases slightly with increasing level of noise. In Figure 1, we report the posterior distribution of the number of clusters for the three processes, for the case $\tau = 0.25$. In accordance with the findings of Proposition 1, we can see that if $\alpha_n = \beta_n = 1$ the distribution is more concentrated around the mean and fewer clusters are generated in the fit.

Table 2. Summary statistics for the simulation study in Section 5.3. The table compares the Beta-GOS and a Dirichlet process model under different specifications of hyperparameters when the data-generating process is a mixture of five Gaussian components

	Data-generating process: Gaussian mixture - 5 Gaussians					
	$\tau = 0.25$			$\tau = 0.5$		
True sample variability	Beta-GOS		Dir. Proc.	Beta-GOS		Dir. Proc.
Model-fitting method	$\alpha_n = n,\ \beta_n = 1$	$\alpha_n = \beta_n = 1$	$\theta = 1$	$\alpha_n = n,\ \beta_n = 1$	$\alpha_n = \beta_n = 1$	$\theta = 1$
Estimated number of clusters	4.95 ± 0.97	4.71 ± 0.76	5.52 ± 1.48	4.70 ± 1.32	4.19 ± 0.99	5.28 ± 1.90
Accuracy of cluster assignment	0.94 ± 0.09	0.93 ± 0.09	0.93 ± 0.09	0.86 ± 0.11	0.84 ± 0.12	0.85 ± 0.13
Predictive bias	8.86 ± 9.02	8.73 ± 9.02	8.84 ± 8.99	8.53 ± 8.61	8.39 ± 8.37	8.55 ± 8.61
Estimated sample variability	0.25 ± 0.01	0.25 ± 0.01	0.27 ± 0.05	0.56 ± 0.68	0.69 ± 1.19	0.62 ± 0.19

Finally, we note that in our simulations, posterior inference for the Beta-GOS process seemed only minimally affected by the two different specifications of the parameters of the Beta weights. This consideration confirms the suggestion that using $\alpha_n = n + \theta - 1$, $\beta_n = 1$ represents a default choice in many applications, where there is no a priori information to guide parameter choice. In this case, θ can be chosen or estimated similarly as what is routinely done for mixtures of DPs. The online supplementary materials contain additional results for several specifications of the model hyperparameters, overall confirming the above remarks.

5.4 Fitting Hidden Semi-Markov Models

A third simulation study is designed to assess the robustness of the Beta-GOS framework to a misspecification of a different nature. In many problems (e.g., change-point detection), HMM are used as computationally convenient substitutes for temporal processes that are known to be more complex than implied by first-order Markovian dynamics. Here, we generate nonexchangeable sequences from a hidden semi-Markov process (HSMM; Ferguson 1980; Yu 2010) and study how the Beta-GOS process performs in fitting this type of data. HSMMs are an extension of the popular HMM where the time spent in each state (state occupancy or sojourn time) is given by an explicit (discrete) distribution. A geometric state occupancy distribution characterizes ordinary HMM. Therefore, HSMMs have also been referred to as "HMM with explicit duration" (Mitchell,

Harper, and Jamieson 1995; Dewar, Wiggins, and Wood 2012) or "variable-duration HMM" (Rabiner 1989).

We generate 1000 datasets (1000 observations each) using a HSMM with four states and a negative binomial distribution for the state occupancy distribution. More specifically, we parameterize the negative binomial in terms of its mean and an ancillary parameter, which is directly related to the amount of overdispersion of the distribution (Airoldi et al. 2006; Hilbe 2011). If the data are not overdispersed, the negative binomial reduces to the Poisson, and the ancillary parameter is zero. For the simulations presented here, we consider a NegBin(15, 0.15), which corresponds to assuming a large overdispersion (17.25). We also consider $\tau = 0.25$ and $\tau = 0.5$ to explore robustness to different levels of noise. We fit the data by means of a Beta-GOS model with Beta hyperparameters defined by: (a) $\alpha_n = n$, $\beta_n = 1$; (b) $\alpha_n = 5$, $\beta_n = 1$; (c) $\alpha_n = 1$, $\beta_n = 1$. Based on Proposition 1, those choices correspond to assuming different clustering behaviors; in particular, different expected number of clusters a priori. We then compare the Beta-GOS with the fit resulting from HMM, assuming 3, 4, and 5 states, respectively. Results from the simulations are reported in Table 3, where the HMM was implemented using the R package "RHmm" (Taramasco and Bauer 2012). Table 3 shows that the Beta-GOS is a viable alternative to HMM, as it can provide more accurate inference than a single HMM where the number of states is fixed a priori. As expected, higher levels of noise decrease the accuracy of the estimates, but the reduction affects the fit of the Beta-GOS and HMM similarly. Furthermore, the fit obtained with the Beta-GOS appears quite robust to the different choices of the hyperparameters.

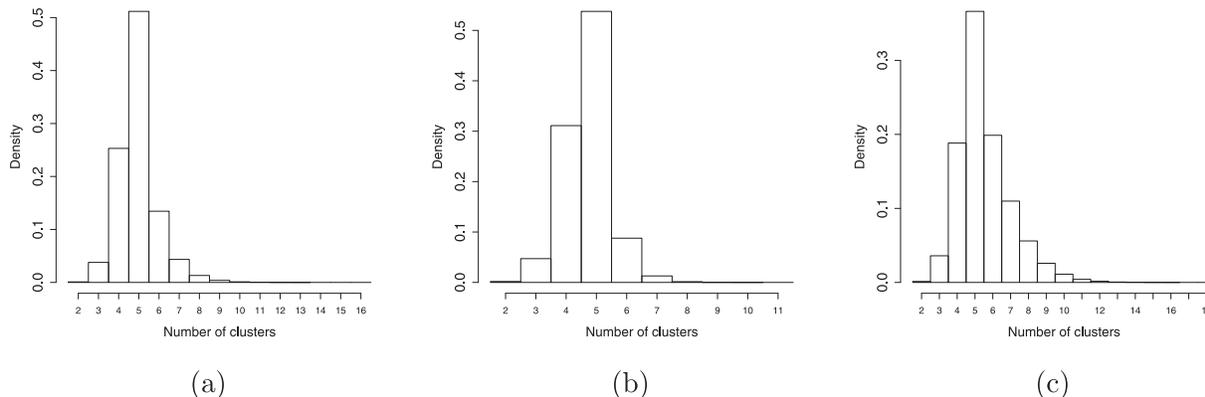
(a)   (b)   (c)

Figure 1. Posterior distribution of the number of clusters in the simulation of Section 5.3 ($\tau = 0.25$). Case (a) corresponds to a Beta-GOS($\alpha_n = n$, $\beta_n = 1$), case (b) to a Beta-GOS($\alpha_n = \beta_n = 1$), and case (c) to a Dirichlet process with parameter $\theta = 1$.

Table 3. Summary statistics for the simulation studies described in Section 5.4. The table compares the Beta-GOS and a hidden Markov model under different specifications of hyperparameters. The data-generating process assumes a hidden semi-Markov with state occupancy distribution NegBin(15, 0.15) and two levels of the sampling noise $\tau = 0.25$ and $\tau = 0.5$

| | (i) Data-generating process: Hidden semi-Markov model (HSMM) with four states and NegBin(15, 0.15) | | | | | |
| | Beta-GOS | | | HMM | | |
Model-fitting method	$\alpha_n = n;\ \beta_n = 1$	$\alpha_n = 5;\ \beta_n = 1$	$\alpha_n = 1;\ \beta_n = 1$	3 States	4 States	5 States
			$\tau = 0.25$			
Estimated number of clusters	3.89 ± 0.53	4.04 ± 0.59	4.09 ± 0.63	2.98 ± 0.13	3.94 ± 0.29	4.87 ± 0.55
Accuracy of cluster assignment	0.95 ± 0.10	0.97 ± 0.07	0.96 ± 0.08	0.72 ± 0.09	0.84 ± 0.13	0.90 ± 0.13
			$\tau = 0.5$			
Estimated number of clusters	3.69 ± 0.81	3.89 ± 0.96	4.06 ± 0.97	2.99 ± 0.12	3.96 ± 0.25	4.90 ± 0.48
Accuracy of cluster assignment	0.86 ± 0.14	0.90 ± 0.12	0.90 ± 0.12	0.71 ± 0.11	0.83 ± 0.12	0.88 ± 0.13

Figure 2 illustrates the clustering induced by the Beta-Gos and a four-state HMM for a subset of the data generated in two specific simulation replicates. The middle column illustrates the allocation, respectively, from a Beta-Gos($\alpha_n = 1$, $\beta_n = 1$) (top) and a Beta-Gos($\alpha_n = n$, $\beta_n = 1$)(bottom), whereas column (c) illustrates the clustering attained by the HMM. Caution is necessary to avoid overinterpreting the results in the figure. Overall, the segmentation plots suggest similarity in the allocations induced by the Beta-GOS and the HMM. In some instances, the Beta-GOS fit seems to allow shorter stretches of contiguous identical states, as illustrated in the top row of Figure 2. On the other hand, when data are characterized by elevated intracluster variability, as in the bottom row of Figure 2, both the Beta-Gos and the HMM could fail to attain a fair representation of the true clustering structure of the data. Our practical experience suggests that the issue is more prominent for the "default" Beta-Gos($\alpha_n = n$, $\beta_n = 1$) than for the "informative" Beta-Gos($\alpha_n = a$, $\beta_n = b$) formulations. This is in accordance with the discussion in Section 3 and, in particular, with the consideration that a Beta-Gos($\alpha_n = n$, $\beta_n = 1$) should represent a long memory process where all previous observations are ex-

pected to contribute the same weight in (3). The supplementary materials contain results for a wider range of parameter settings, as well as different data-generating mechanisms, confirming the results noted above.

6. QUANTIFYING CHROMOSOMAL ABERRATIONS IN BREAST CANCER

We first apply the Beta-GOS to a classic dataset that has been used to link patterns of chromosomal aberrations to breast cancer pathophysiologies in the medical literature (Chin et al. 2006). The raw data measure genome copy number gains and losses over 145 primary breast tumor samples, across the 23 chromosomes, obtained using BAC array CGH. More precisely, the measurements consist of $\log_2$ intensity ratios obtained from the comparison of cancer and normal female genomic DNA labeled with distinct fluorescent dyes and co-hybridized on a microarray in the presence of Cot-1 DNA to suppress unspecific hybridization of repeat sequences (see Redon, Fitzgeral, and Carter 2009). The analysis of array CGH data presents some challenges, because data are typically very noisy and spatially
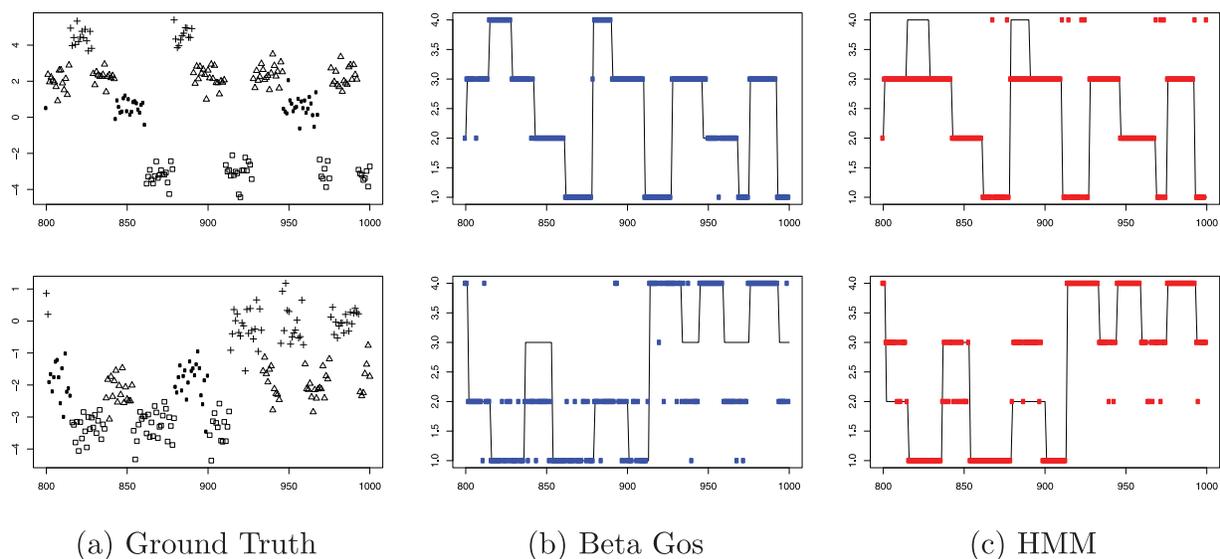

(a) Ground Truth   (b) Beta Gos   (c) HMM

Figure 2. Illustrative segmentation-type plots for the simulation study in Section 5.4. Column (a): subset of data for two replicates. Column (b) top: an example of allocation for a Beta-Gos($\alpha_n = 1$, $\beta_n = 1$) plotted versus the truth (black line); column (b) bottom considers a Beta-Gos($\alpha_n = n$, $\beta_n = 1$). Column (c) illustrates the fitting by an HMM with four states.
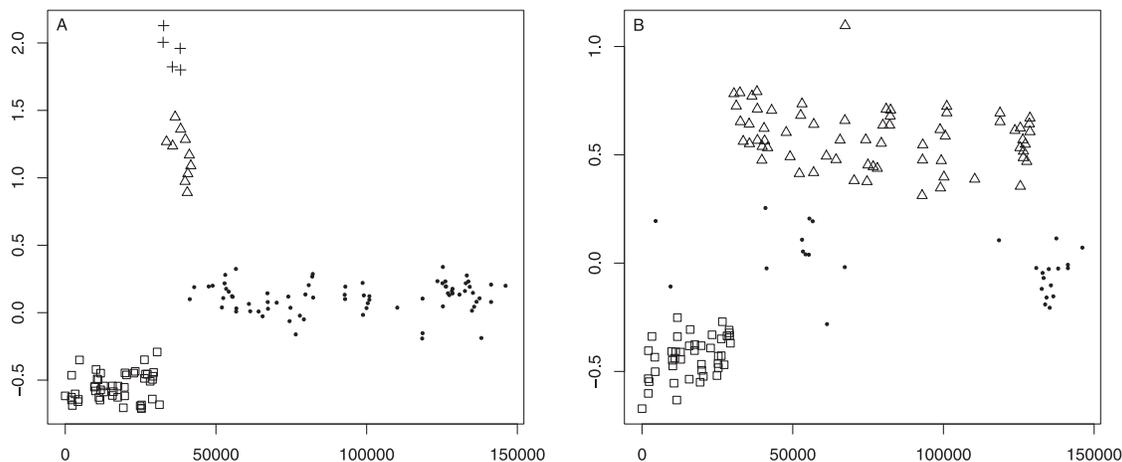
Figure 3. Model fit overview: array CGH gains and losses on chromosome 8 for two samples of breast tumors in the dataset in Chin et al. (2006). Points with different shapes denote different clusters.

correlated. More specifically, copy numbers gains or losses at a region are often associated with an increased probability of gains and losses at a neighboring region. We use the Beta-GOS model developed in the previous sections to analyze and cluster clones with similar level of amplification/deletion, for each breast tumor sample and each chromosome in the dataset. For array CGH data, it is typical to distinguish regions with a normal amount of chromosomal material, from regions with single copy loss (deletion), single copy gain, and amplifications (multiple copy gains). Therefore, we present here the results of the analysis where the latent Beta hyperparameters are set to $\alpha_n = 3$ and $\beta_n = 1$, corresponding to $E(K_n) = 4$ states for large n (see Section 3). We have also considered $\alpha_n = n$ and $\beta_n = 1$, with no remarkable differences in the results. We complete the specification of model (6)–(7) with a vague base distribution, Normal(0, 10), and a vague inverse gamma distribution for τ centered around $\tau = 0.1$. This choice of τ is motivated by the typical scale of the array CGH data and is in accordance with similar choices in the literature (see, e.g., Guha, Li, and Neuberg 2008).

Figure 3 exemplifies the fit to chromosome 8 on two tumor samples. The model is able to identify regions of reduced copy number variation and high amplification. Note how contiguous clones tend to be clustered together, in a pattern typical of these chromosomal aberrations. Figure 4 replicates fig. 1 in Chin et al. (2006) and shows the frequencies of genome copy number gains and losses among all 145 samples plotted as a function of genome location. To identify a copy number aberration for this plot, for each chromosome and sample, at each iteration we consider the cluster with lowest absolute mean and order the other clusters accordingly. The lowest absolute mean is chosen to identify the copy neutral state. Following Guha, Li, and Neuberg (2008), any other cluster is identified as a copy number gain or loss if its mean, say $\hat{\mu}_{(j)}$, is farther than a specified threshold from the minimum absolute mean, say $\hat{\mu}_{(1)}$, that is, if $\hat{\mu}_{(j)} - \hat{\mu}_{(1)} > \epsilon$. We experimented with a range of choices of ϵ in the range [0.05, 0.15] and used $\epsilon = 0.1$ for the current analysis. Furthermore, if the mean of a cluster is above the mean of all declared gains plus two standard deviations, all genes in that cluster are considered high-level amplifications. We
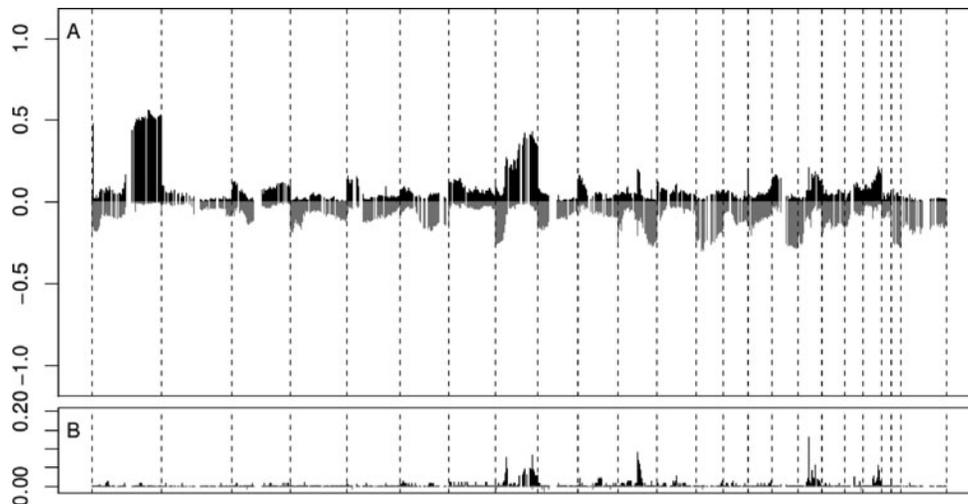

Figure 4. (a) Frequencies of genome copy number gains and losses plotted as a function of genomic location. (b) Frequency of tumors showing high-level amplification. The dashed vertical lines separate the 23 chromosomes.

Table 4. False discovery rate analysis for clones with high-level amplification previously identified by Chin et al. (2006). The individual amplicons are reported together with the locations of the flanking clones on the array platform

Amplicon	Flanking clone (left)	Flanking clone (right)	Kb start	Kb end	FDR q-value
8p11-12	RP11-258M15	RP11-73M19	33579	43001	0.021
8q24	RP11-65D17	RP11-94M13	127186	132829	0.021
11q13-14	CTD-2080I19	RP11-256P19	68482	71659	0.022
11q13-14	RP11-102M18	RP11-215H8	73337	78686	0.024
12q13-14	BAL12B2624	RP11-92P22	67191	74053	0.011
17q11-12	RP11-58O8	RP11-87N6	34027	38681	0.017
17q21-24	RP11-234J24	RP11-84E24	45775	70598	0.017
20q13	RMC20B4135	RP11-278I13	51669	53455	0.021
20q13	GS-32I19	RP11-94A18	55630	59444	0.017

identify a clone with an aberration (or high-level amplification) if it is such in more than 70% of the MCMC iterations; then, we compute the frequency of aberrations and high-level amplifications among all 145 samples, which are reported, respectively, at the top and bottom of Figure 4. As expected, the clusters identified by the model tend to be localized in space all over the genome. This feature may be facilitated by the increasingly low reinforcement of faraway clones embedded in the Beta-GOS, and corresponds to the understanding that clones that live at adjacent locations on a chromosome can be either amplified or deleted together due to the recombination process.

Finally, we considered some regions of chromosomes 8, 11, 17, and 20 that have been identified by Chin et al. (2006) and shown to correlate in their analysis to increased gene expression. We adapt the procedure described in Newton et al. (2004) to compute a region-based measure of the false discovery rate (FDR) and determine the q-values for neutral-state and aberration regions estimated in our analysis. The q-value is the FDR analog of the p-value, as it measures the minimum FDR threshold at which we may determine that a region corresponds to significant copy number gains or losses (Storey 2003; Storey, Dai, and Leek 2007). More specifically, after conducting a clone-based test as described in the previous paragraph, we identify regions of interest by taking into account the strings of consecutive calls. These regions then constitute the units of the subsequent cluster-based FDR analysis. Alternatively, the regions of interest could be prespecified on the basis of the information available in the literature. The optimality of the type of procedures here described for cluster-based FDR is discussed in Sun et al. (in press). See also Heller et al. (2006), Müller, Parmigiani, and Rice (2007), and Ji, Lu, and Mills (2008). In Table 4, we report the q-values from a set of candidate oncogenes in well-known regions of recurrent amplification (notably, 8p12, 8q24, 11q13-14, 12q13-14, 17q21-24, and 20q13). Our findings confirm the previous detections of chromosomal aberrations in the same locations.

Next, we apply our methodology to the analysis of a modern large-scale CGH array dataset (Curtis et al. 2012). More specifically, here we consider one sample from the data published by Curtis et al. (2012). We fit the Beta-GOS model to the entire sequence of 969,700 probes matched to genomics locations using a priority queue on the Harvard Odyssey cluster. Model fit took about 24 hr. We also fit an HMM and an HSMM with negative-binomial run lengths times, both set to have three

states (Yau et al. 2011), to the same sample. The parameters of both models were estimated using standard techniques (Rabiner 1989; Guedon 2011). The estimates for the negative-binomial run lengths, shared across states for simplicity, were $\hat{r} = 10$ and $\hat{\pi} \approx 0.25$, leading to a mean run length of 30 probes with a standard deviation of 10.

For validation purposes, we accessed a list of 152 consensus genomic locations where chromosomal aberrations were found in breast cancer tumor samples. This list is included in the data files associated with The Genome Cancer Atlas (TCGA) project, partly curated by the Broad Institute and hosted by the NIH. The 152 consensus genomic locations range in size from 5 to 49 probes. This list provides a list of locations, which have been reported as likely altered in terms of DNA content in a number of publications, using multiple types of datasets and analyses. Therefore, the list is independent of the specifics of any particular technique, and it can be used as a reference for evaluating the comparative performance of our Beta-GOS model with HMM and HSMMs. For each method, we declared a success at detecting a chromosomal aberration (either deletion or amplification) at any of the 152 consensus genomic locations if the method correctly labeled at least 80% of the probes associated with any given consensus genomic location. This choice was necessary since locations span multiple probes. According to this simple measure of performance, the Beta-GOS correctly labeled 133 locations (or 87.50%) of the 152 consensus genomic locations as having a chromosomal aberration, versus 94 locations (or 61.84%) using the HMM, and 118 locations (or 77.63%) using the HSMM. Of course, caution should be taken against overinterpreting the results of a single illustrative example. However, the results from the simulation studies and data analysis all concur to suggests that the Beta-GOS is a flexible model and that can be usefully employed in detecting chromosomal aberrations in array CGH data, since it can account for long-range dependencies in the sequence and achieve improved accuracy with respect to competing HMM-based approaches.

7. CONCLUDING REMARKS

Starting from the characterization of species sampling sequences in terms of their predictive probability functions, we have considered predictive rules where the weights are functions of latent Beta random variables. The resulting Beta-GOS process defines a novel and probabilistically coherent Bayesian

nonparametric model for nonexchangeable data. We have discussed the clustering behavior of the Beta-GOS processes for some specifications of the latent Beta densities and illustrated their use as priors in a hierarchical model setting. Finally, we have analyzed the performance of this modeling framework by means of a set of simulation studies. The results outlined in Section 6 illustrate how the proposed Beta-GOS model can be a useful tool for the analysis of CGH array data. In medical applications, for instance, it might be used to complement tumor subtype definition, or to suggest candidate genes for follow-up clinical studies. We expect our approach will be useful in other applications where hidden Markov and semi-Markov model are currently considered as standard, for example, in text segmentation and speech processing (e.g., Rabiner 1989; Blei and Moreno 2001; Chien and Furui 2005; Ren et al. 2010; Yau and Holmes 2013; Fox et al. 2014).

Recently, Teh et al. (2006b), Fox et al. (2011), and Yau et al. (2011) had developed flexible and effective hierarchical Bayesian nonparametric extensions of HMM that allow posterior inference over the number of states. The Beta-GOS model provides an alternative, nonexchangeable, Bayesian nonparametric formalism to model heterogeneity across nonexchangeable observations that are sequentially ordered, by enabling clustering in a number of unknown states. Since the Beta-GOS model does not rely on the estimation of a single transition matrix across time points, as in the HMM, we do not need to consider an explicit parameter to account for state persistence, as in Fox et al. (2011), or assume a distribution for the sojourn times as in HSMMs. Indeed, since the predictive weights depend on the sequence of observations itself, the Beta-GOS seems particularly convenient when the underlying generative process is nonstationary, for example, as a possible alternative to more complicated nonhomogenous HMMs. Monteiro, Assunçao, and Loschi (2011) discussed a similar issue in a product partition model framework and explicitly assume that the observations in a cluster have their distributions indexed by different parameters. Our approach is different, for example, we do not need to explicitly model the dependence structure within the clusters.

Arguably, the major obstacle we can foresee in the wider applicability of this type of models relies in the specification of the prior hyperparameters in the latent Beta distributions. Some specific suggestions have been provided in Section 3. However, in cases where there is not enough prior information to advise differently, our experience suggests that the default choice of the hyperparameters outlined in Proposition 1(a) not only reduces the problem to the choice of a single parameter as it is usual in DP mixture models, but may also suffice for inferential purposes. Alternatively, one could assume a prior distribution on the parameters of the Beta latent variables and conduct posterior inference by means of MCMC methods, as briefly discussed in Section 4. Nevertheless, in specific applications the optimal modeling of the latent Beta densities requires further study and will be pursued elsewhere. In addition, the proposed approach inherits the general computational limitations of nonparametric Bayesian methods. For example, a full MCMC algorithm for posterior inference may be unfeasible for genomic sequences with several millions of reads. Scalable algorithms may facilitate fast inference in those settings (e.g., Colella et al. 2007).

Finally, we believe that the flexibility of the latent specification and the possibility to tie the clustering implied by the generalized Pólya urn scheme directly to a set of latent random variables gives an opportunity to further modeling the complex relationships typical of heterogenous datasets. For example, further developments may substitute the general latent Beta specification with a probit/logistic specification, and define a generalized Pólya urn scheme in the aims of Rodriguez and Dunson (2011) that allows the clustering at each observation to be dependent on a set of (possibly sequentially recorded) covariates or curves. Similarly, we can imagine using multivariate generalized Pólya urn schemes of the sort we describe in this article to model time-dependent parameters in time series, which may be important to identify time-varying structures or regime changes at the base of phenomena like the so-called financial contagion, that is, the co-movement of asset prices across global markets after large shocks (see, e.g., Liu, Windle, and Scott 2013).

APPENDIX A: DETAILS OF POSTERIOR MCMC SAMPLING FOR THE BETA-GOS MODEL

Here, we provide the details of the MCMC sampling algorithm described in Section 4.2 for the special case of a Normal sampling distribution and a Normal (or Normal-Inverse-Gamma) base measure.

A.1 Full Conditionals for the Gibbs Sampler

At each iteration of Gibbs sampler, we sample from the full conditionals of C_n and W_n, for $n = 1, \ldots, N$. Here we derive the analytical form of these distributions, for the Beta-GOS model specified in Section 5. Recall that the full conditional distribution for C_n is

$$P\{C_n = i | C_{-n}, W(N), Y(N), \tau^2\}$$
$$\propto P\{Y(N) | C_n = i, C_{-n}, W(N), \tau^2\} \cdot P\{C_n = i | C_{-n}, W(N)\},$$

where the factor on the right is given by (8) and (3), and the left factor is obtained by integration,

$$P\{Y(N) \mid C_n = i, C_{-n}, W(N), \tau^2\} = P\{Y(N) \mid C_n = i, C_{-n}, \tau^2\}$$

$$= \int P\{Y(N), \mu \mid C_n = i, C_{-n}, \tau^2\} d\mu$$

$$= \prod_{j=1}^{J} \int \prod_{l \in \Pi_j} p(Y_l | \mu_j^*) G_0(d\mu_j^*)$$

$$\propto \prod_{j=1}^{J} \frac{1}{(\sqrt{2\pi} \, \tau)^{|\Pi_j|}} \exp \left\{ - \frac{\sum_{l \in \Pi_j} y_l^2}{2\tau^2} - \frac{\mu_0^2}{2\sigma_0^2} \right.$$

$$\left. + \frac{1}{2} \frac{\left(\frac{\mu_0}{\sigma_0^2} + \sum_{l \in \Pi_j} \frac{y_l}{\tau^2} \right)^2}{\frac{1}{\sigma_0^2} + \frac{|\Pi_j|}{\tau^2}} \right\} \frac{1}{\sqrt{\frac{|\Pi_j|\sigma_0^2}{\tau^2} + 1}},$$

where Π_j is the set of indices of data points in cluster j, and J is the number of clusters at that iteration. Note that the latent reinforcements $W(N)$ are used to define the cluster assignments through the data-pairing labels $C(N)$. Conditionally on the data-pairing labels $C(N)$, the data $Y(N)$ is independent of the latent reinforcements $W(N)$.

The full conditional for W_n, denoted by $P(W_n | C(N), W_{-n}, Y(N))$, is Beta distributed with updated parameters A_n, B_n, defined as in (8).

A.2 Inference on the Cluster Centroids of the Beta-GOS Process

For the purpose of computational efficiency, it is generally preferable to sample the random partitions integrating out with respect to the

parameters of the Beta-GOS process, as described in Section 4.2 and in Appendix A.1. If the sampling distribution and the base measure are conjugate, this usually results in improved mixing of the chain. However, in many cases, it may be required to draw inferences on the cluster centroids themselves. As usual with mixtures of DP, inference on the cluster centroids can be easily conducted (even ex-post) from the clustering configurations at each iteration. Therefore, we do not have to sample the centroids within each Gibbs iteration, but if the need be, we can easily resample them at the end of each iteration, or at the end of the sampler from the stored output.

A.3 Inference on the Cluster and Global Variances

Let the variance of the sampling distribution be τ^2. We assume $\tau^2 \sim \text{IGamma}(a_0, b_0)$. The posterior distribution of the variance in each cluster j, is given by

$$\tau_j^2 \mid \mu_j^*, Y_i, i \in \Pi_j \sim \text{IGamma}\left(a_0 + \frac{|\Pi_j|}{2}, b_0 + \frac{1}{2}\sum_{i \in \Pi_j}\left(Y_i - \mu_j^*\right)^2\right),$$

Note that, in case of need and for computational efficiency, we could use these also quantities to obtain a global estimate for the sampling variance at each iteration, in an MCMC-EM step, as $\hat{\tau}^2 = \sum_{j=1}^{J}\frac{(|\Pi_j|-1)\tau_j^2}{N-J}$. This may turn useful, for example, for parallelization purposes, as in the simulations of Section 5.

A.4 Inference on the Cluster Means

In the normal–normal model described in Section 5, the posterior distribution of μ_j^* given data Y_i in the jth cluster can be evaluated at each iteration as

$$P(\mu_j^* \mid \tau_j^2, Y_i, i \in \Pi_j) \sim N\left(\frac{\frac{\mu_0}{\sigma_0^2} + \frac{|\Pi_j|\overline{Y_j}}{\tau_j^2}}{\frac{1}{\sigma_0^2} + \frac{|\Pi_j|}{\tau_j^2}}, \left(\frac{1}{\sigma_0^2} + \frac{|\Pi_j|}{\tau_j^2}\right)^{-1}\right).$$

for $j = 1, \ldots, J$, where $\bar{Y}_j$ is the jth cluster specific mean. Note that we have assumed a common sampling variance τ^2; the modification of the previous formula to take into account a cluster specific variance is of course straightforward.

APPENDIX B: DETAILS OF THE PROOFS AND ADDITIONAL THEORETICAL RESULTS

BS.5 Generalized Ottawa Sequence and Its Moments

According to Bassetti, Crimaldi, and Leisen (2010), a sequence $(X_n)_{n\geq 1}$ of random variables taking values in a Polish space is a GOS if there exists a sequence $(W_n)_{n\geq 1}$ (of random variables) such that the following conditions are satisfied: (i) the law of X_1 is G_0; (ii) for $n \geq 1$, X_{n+1} and the subsequence $(W_{n+j})_{j\geq 1}$ are conditionally independent given the filtration $\mathcal{F}_n := \sigma(X_1, \ldots, X_n, W_1, \ldots, W_n)$; (iii) the predictive distribution of X_{n+1} given $\mathcal{F}_n$ is given by (3) where the r_n's are strictly positive functions, $r_n(W_1, \ldots, W_n)$, of the vector of latent variables, such that

$$r_n(W_1, \ldots, W_n) \geq r_{n+1}(W_1, \ldots, W_n, W_{n+1}), \tag{B.1}$$

almost surely, with $r_0 = 1$, and the weights $p_{n,i} = p_{n,i}(W_1, \ldots, W_n)$ are

$$p_{n,i} = \frac{r_n(r_{i-1} - r_i)}{r_{i-1}r_i} \qquad i = 1, \ldots, n. \tag{B.2}$$

The predictive distribution (3)–(4) corresponds to choice $r_n(W_1, \ldots, W_n) = \prod_{i=1}^{n} W_i$, where $(W_n)_{n\geq 1}$ is a sequence of independent random variables.

We conclude this section by providing a general result for the kth moment and for the moment generating function of the length K_n of

a GOS. Suppose that the sequence $(X_n)_{n\geq 1}$ is a GOS, with G_0 diffuse, and let $U_j = K_j - K_{j-1}$ with $K_0 = 0$. Then, $K_n = \sum_{j=1}^{n} U_j$ and the joint distribution of $U_1, \ldots, U_n$ conditionally on $r_1, \ldots, r_{n-1}$, is

$$P\{U_1 = 1, \ldots, U_n = e_n | r_1, \ldots, r_{n-1}\} = \prod_{i=2}^{n} r_{i-1}^{e_i}(1 - r_{i-1})^{1-e_i},$$

for every vector $(e_2, \ldots, e_n)$ in $\{0, 1\}^{n-1}$, since $P(U_1 = 1) = 1$ by definition. Since $K_1 = U_1 = 1$, it follows that, for every $k \geq 1$,

$$P\{K_{n+1} = k + 1\} = P\left\{\sum_{j=2}^{n+1} U_j = k\right\} = \sum_{\underline{e}} E\left[\prod_{i=1}^{n} r_{i-1}^{e_i}(1 - r_{i-1})^{1-e_i}\right],$$

where the summation is extended over all sequences $\underline{e} = (e_1, \ldots, e_n)$ in $\{0, 1\}^n$ such that $\sum_{i=1}^{n} e_i = k$. Moreover, for every $k \geq 1$ and $n \geq 2$, it is easy to see that

$$E[(K_{n+1} - 1)^k] = E\left[\left(\sum_{j=2}^{n+1} U_j\right)^k\right] = \sum_{m=1}^{k\wedge n} m!S(k, m)\phi_{n,m}, \quad \text{(B.3)}$$

where $k \wedge n = \min(k, n)$,

$$\phi_{n,m} := \sum_{1\leq l_1 < l_2 < \cdots < l_m \leq n} E[r_{l_1}, \ldots, r_{l_m}]. \tag{B.4}$$

and $S(k, m) := \frac{k!}{m!}\sum_{\{n_i > 0: \sum_{i=1}^{m} n_i = k\}}\frac{1}{n_1!\ldots n_m!}$ is the Stirling number of second kind. Hence, $E[(K_{n+1} - 1)^k]$ depends recursively on functions $\phi_{n-1,m}, m = 1, \ldots, k$. It may be interesting to note that, using the well-known relation between factorial moments and ordinary moments (see, e.g., Example 2.3 in Charalambides 2005), from (B.3) one gets, for any $m \leq n$,

$$E[(K_{n+1} - 1)_{(m)}] = m!\phi_{n,m}, \tag{B.5}$$

where $(t)_{(m)} = t(t - 1)\ldots(t - m + 1)$ is the falling factorial. Moreover, since

$$\sum_{k\geq m}(-t)^k\frac{S(k, m)}{k!} = \frac{(e^{-t} - 1)^m}{m!},$$

see, for example, Theorem 2.3 in Charalambides (2005), it follows that the moment-generating function of K_{n+1} is given by

$$\begin{aligned}M_{n+1}(t) :&= E[e^{-tK_{n+1}}] = e^{-t}E[e^{-t(K_{n+1}-1)}]\\&= e^{-t}\sum_{k\geq 0}\frac{(-t)^k}{k!}E[(K_{n+1} - 1)^k] = e^{-t} + e^{-t}\\&\sum_{k\geq 1}\sum_{m=1}^{k\wedge n}\frac{(-t)^k m!}{k!}S(k, m)\phi_{n,m}\\&= e^{-t} + e^{-t}\sum_{m=1}^{n} m!\,\phi_{n,m}\sum_{k\geq m}\frac{(-t)^k}{k!}S(k, m)\\&= e^{-t}\sum_{m=0}^{n}(e^{-t} - 1)^m\phi_{n,m}\end{aligned} \tag{B.6}$$

with $\phi_{n,0} := 1$.

B.6 Proof of Proposition 1

If we consider Equation (B.4) with $(W_i)_{i\geq 1}$ independent random variables taking values in $[0, 1]$, then

$$\phi_{n,m} = \sum_{1\leq l_1 < l_2 < \cdots < l_m \leq n}\prod_{j=1}^{m}\prod_{i=l_{j-1}+1}^{l_j} E\left[W_i^{m+1-j}\right], \tag{B.7}$$

where $l_0 := 0$. We need some preliminary results.

Lemma B.2. If $W_i \sim \text{Beta}(i + \theta - 1, 1)$, for given $\theta > 0$, then

$$\phi_{n,m} = \frac{\Gamma(\theta + m)}{\Gamma(\theta)} \sum_{j_1=m}^{n} \sum_{j_2=m}^{j_1} \sum_{j_3=m}^{j_2} \cdots$$

$$\sum_{j_m=m}^{j_{m-1}} \frac{1}{(j_1 + \theta)(j_2 + \theta) \cdots (j_m + \theta)}. \quad (B.8)$$

In particular, as n goes to $+\infty$,

$$E[K_n^k] = \frac{\Gamma(\theta + k)}{\Gamma(\theta)} \log^k(n)[1 + o(1)]. \quad (B.9)$$

Let us start by proving (B.8). First, note that since W_i is a $\text{Beta}(i + \theta - 1, 1)$ random variable then, for $1 \le j \le m$, $E[W_i^{m+1-j}] = \frac{i+\theta-1}{i+\theta+m-j}$. Hence, by (B.7),

$$\phi_{n,m} = \sum_{1 \le l_1 < l_2 < \cdots < l_m \le n} \prod_{j=1}^{m} \prod_{i=l_{j-1}+1}^{l_j} \frac{i + \theta - 1}{i + \theta + m - j}, \quad (B.10)$$

which, after some algebra, returns (B.8). To prove the second part of Lemma B.1 we need to introduce additional notation. For $\theta > 0, k \ge 1$, $m \ge 2$ and $n \ge k$, set

$$\Psi_{k,\theta}(n, m) := \sum_{j_1=k}^{n} \sum_{j_2=k}^{j_1} \sum_{j_3=k}^{j_2} \cdots \sum_{j_m=k}^{j_{m-1}} \frac{m!}{(j_1 + \theta)(j_2 + \theta) \cdots (j_m + \theta)},$$

$$\Psi_{k,\theta}(n, 1) := \sum_{j_1=k}^{n} \frac{1}{(j_1 + \theta)}.$$

Note that $m! \phi_{n,m} = \Psi_{m,\theta}(n, m) \Gamma(\theta + m) / \Gamma(\theta)$. For all $k \ge 1, m \ge 1$, and $n \ge k$, set $Q_{k,\theta}(m, n) := \Psi_{k,\theta}(n, m) - \log^m(n + \theta)$. Now formula (B.9) in Lemma B.1 follows easily from (B.3) and the next result.

Lemma B.3. For $\theta > 0, k \ge 1$ and $m \ge 1$, there is a constant $C_{k,\theta}(m)$ such that

$$|Q_{k,\theta}(m, n)| \le C_{k,\theta}(m) \log^{m-1}(n + \theta) \quad \text{for every } n \ge k. \quad (A.12)$$

Let $k \ge 1$ and $\theta > 0$. For $m \ge 1$ and $n \ge k$ set

$$S_{k,\theta}(m, n) := \sum_{j=k}^{n} \frac{m \log^{m-1}(j + \theta)}{j + \theta},$$

and

$$R_{k,\theta}(m, n) := S_{k,\theta}(m, n) - \log^m(n + \theta) = \sum_{j=k}^{n} \frac{m \log^{m-1}(j + \theta)}{j + \theta}$$

$$- \log^m(n + \theta). \quad (B.12)$$

We claim that, for any $m \ge 1$, there is a constant $C_m^* = C_{m,\theta,k}^*$ such that

$$|R_{k,\theta}(m, n)| \le C_m^*, \quad \text{for all } n \ge k. \quad (B.13)$$

Now observe that $\Psi_{k,\theta}(n, 1) = S_{k,\theta}(1, n)$. Hence, (B.13) proves (B.11) for $m = 1$ and every $k \ge 1$ and $\theta > 0$. By induction suppose that (B.11) is true for $m = 1, \ldots, M - 1$. Note that, for $m \ge 2$,

$$\Psi_{k,\theta}(n, m) = \sum_{j_1=k}^{n} \frac{m}{j_1 + \theta} \Psi_{k,\theta}(j_1, m - 1),$$

hence, by induction hypothesis, for every $\theta > 0, k \ge 1$ and $n \ge k$,

$$\Psi_{k,\theta}(n, M) = \sum_{j_1=k}^{n} \frac{M}{j_1 + \theta} \left[\log^{M-1}(j_1 + \theta) + Q_{k,\theta}(M - 1, j_1) \right].$$

Using (B.12) one gets

$$\Psi_{k,\theta}(n, M) = \log^M(n + \theta) + R_{k,\theta}(M, n) + \sum_{j_1=k}^{n} \frac{M}{j_1 + \theta} Q_{k,\theta}(M - 1, j_1).$$

Hence, using (B.13) and the induction hypothesis, one can write

$$|Q_{k,\theta}(M, n)| \le |R_{k,\theta}(M, n)| + \sum_{j_1=k}^{n} \frac{M}{j_1 + \theta} |Q_{k,\theta}(M - 1, j_1)|$$

$$\le C_{M,\theta,k}^* + \frac{M C_{k,\theta}(M - 1)}{M - 1} \sum_{j_1=k}^{n} \frac{M - 1}{j_1 + \theta} \log^{M-2}(j_1 + \theta)$$

$$\le C_{M,\theta,k}^* + \frac{M C_{k,\theta}(M - 1)}{M - 1} [\log^{M-1}(n + \theta) + |R_{k,\theta}(M - 1, n)|]$$

$$\le C_{M,\theta,k}^* + \frac{M C_{k,\theta}(M - 1)}{M - 1} [\log^{M-1}(n + \theta) + C_{M-1,\theta,k}^*],$$

which proves (B.11) for $m = M$. To complete the proof let us prove (B.13). Observe that $x \mapsto \frac{\log^{m-1}(x+\theta)}{x+\theta}$ is a nonincreasing function on $[x_0, +\infty)$ for a suitable $x_0 = x_0(k, \theta, m)$. Assume, without real loss of generality, that $k \ge x_0 + 1$. Note that, in this case,

$$\int_{k}^{n+1} \frac{m \log^{m-1}(x + \theta)}{x + \theta} dx \le S_{k,\theta}(m, n) \le \int_{k-1}^{n} \frac{m \log^{m-1}(x + \theta)}{x + \theta} dx.$$

Hence,

$$\log^m(n + 1 + \theta) - \log^m(k + \theta) \le S_{k,\theta}(m, n) \le \log^m(n + \theta)$$

$$- \log^m(k - 1 + \theta),$$

which gives

$$\log^m(n + \theta) - \log^m(k + \theta) \le S_{k,\theta}(m, n) \le \log^m(n + \theta),$$

and then

$$|S_{k,\theta}(m, n) - \log^m(n + \theta)| \le \log^m(k + \theta).$$

Proof of Proposition 1(a). It follows immediately from (B.9) and a classical result concerning the convergence in distribution when the moments converge. Indeed, $E\left[\left(\frac{K_n}{\log n}\right)^k\right]$ converges to $\frac{\Gamma(\theta+k)}{\Gamma(\theta)}$ that is the kth moment of a $\Gamma(\theta, 1)$ random variable.

Proof of Proposition 1(b). The first part of the statement of Proposition 1(b) follows from Proposition 2.1 in Bassetti, Crimaldi, and Leisen (2010) if one shows that $E[\sum_{i=1}^{\infty} r_i] < \infty$. For $\alpha_n = a$ and $\beta_n = b$ one gets $E[r_n] = a^n/(a + b)^n$ and the thesis follows. When $\alpha_n = n + \theta - 1$ and $\beta_n = \beta$, as explained in Section 3, $E[r_n] \sim n^{-\beta}$ and the thesis follows since $\beta > 1$. It remains to prove the assertion concerning the moment-generating function and the factorial moments of K_∞.

If $\alpha_n = a$ and $\beta_n = b$, (B.7) becomes

$$\phi_{n,m} = \sum_{1 \le l_1 < l_2 < \cdots < l_m \le n} \prod_{j=1}^{m} \left(E[W_1^{m+1-j}] \right)^{l_j - l_{j-1}},$$

$$= \sum_{1 \le l_1 < l_2 < \cdots < l_m \le n} \prod_{j=1}^{m} \left(\prod_{i=1}^{m+1-j} \gamma_i \right)^{l_j - l_{j-1}}$$

since $E[W_1^m] = \prod_{i=1}^{m} \gamma_i$ for $\gamma_i = (a + i - 1)/(a + b + i - 1)$. Taking the limit for $n \to +\infty$, we get

$$\lim_n \phi_{n,m} = \sum_{1 \le l_1 < l_2 < \cdots < l_m} \prod_{j=1}^{m} \left(\prod_{i=1}^{m+1-j} \gamma_i \right)^{l_j - l_{j-1}}$$

$$= \sum_{k_1 \ge 1} \sum_{k_2 \ge 1} \cdots \sum_{k_m \ge 1} \prod_{j=1}^{m} \left(\prod_{i=1}^{m+1-j} \gamma_i \right)^{k_j}$$

$$= \prod_{j=1}^{m} \sum_{k_j \ge 1} \left(\prod_{i=1}^{m+1-j} \gamma_i \right)^{k_j} = \prod_{j=1}^{m} \frac{\gamma_1 \cdots \gamma_j}{1 - \gamma_1 \cdots \gamma_j}$$

and then

$$\lim_n \phi_{n,m} = \prod_{j=1}^{m} \frac{(a)^{(j)}}{(a+b)^{(j)} - (a)^{(j)}},$$

where $(t)^{(j)} = t(t+1)\ldots(t+j-1)$ is the rising factorial. Combining this fact with (B.6), it follows that, in this case,

$$E[e^{-tK_\infty}] = e^{-t} \sum_{m \geq 0} (e^{-t} - 1)^m \prod_{j=1}^{m} \frac{(a)^{(j)}}{(a+b)^{(j)} - (a)^{(j)}}.$$

In addition, (B.3)–(B.5) give

$$E\left[\frac{(K_\infty - 1)_m}{m!}\right] = \prod_{j=1}^{m} \frac{(a)^{(j)}}{(a+b)^{(j)} - (a)^{(j)}},$$

$$E[(K_\infty - 1)^k] = \sum_{m=1}^{k} m! S(k,m) \prod_{j=1}^{m} \frac{(a)^{(j)}}{(a+b)^{(j)} - (a)^{(j)}}.$$

B.7 Conditional Identity in Distribution of the Beta-GOS Hierarchical Model

Proposition 2. The sequence $(Y_n)_n$ defined by formula (6)–(7) is conditionally identically distributed with respect to the filtration $\mathcal{H}_n = \sigma(Y(n), W(n), \mu(n))$.

Proof. Let $\mathcal{G}_n = \sigma(W(n), \mu(n))$ and $\mathcal{H}_n = \sigma(W(n), \mu(n), Y(n))$. We have to prove that for every real, bounded and measurable function g

$$E(g(Y_{n+j})|\mathcal{H}_n) = E(g(Y_{n+1})|\mathcal{H}_n). \tag{B.14}$$

Now, for every $j > 0$

$$\mathcal{L}(Y_{n+j} \mid \mathcal{H}_n, \mu_{n+j}) = \mathcal{L}(Y_{n+j} \mid \mu_{n+j}) = p(\cdot \mid \mu_{n+j}) \tag{B.15}$$

and for every j and n

$$\mathcal{L}(\mu_{n+j} \mid \mathcal{H}_n) = \mathcal{L}(\mu_{n+j} \mid \mathcal{G}_n). \tag{B.16}$$

As already recalled, $(\mu_n)_n$ is CID with respect to $\mathcal{G}_n = \sigma(W(n), \mu(n))$. This means that for every real, bounded and measurable function f

$$E(f(\mu_{n+j})|\mathcal{G}_n) = E(f(\mu_{n+1})|\mathcal{G}_n) \tag{B.17}$$

for all $j \geq 1$, see Berti, Pratelli, and Rigo (2004). Thanks to (B.16), equality (B.17) also holds with respect to the sigma-field $\mathcal{H}_n$. Indeed,

$$E(f(\mu_{n+j})|\mathcal{H}_n) = E(f(\mu_{n+j})|\mathcal{G}_n) = E(f(\mu_{n+1})|\mathcal{G}_n) = E(f(\mu_{n+1})|\mathcal{H}_n)$$

(B.15) implies that

$$E(g(Y_{n+j})|\mathcal{H}_n, \mu_{n+j}) = E(g(Y_{n+j})|\mu_{n+j})$$
$$= \int g(y)p(dy \mid \mu_{n+j}) \tag{B.18}$$

(B.17) and (B.18) allow to prove the thesis. Indeed,

$$E(g(Y_{n+j})|\mathcal{H}_n) = E(E(g(Y_{n+j})|\mathcal{H}_n, \mu_{n+j}|\mathcal{H}_n)$$
$$= E\left(\int g(y)p(dy \mid \mu_{n+j})|\mathcal{H}_n\right)$$
$$= E\left(\int g(y)p(dy \mid \mu_{n+1})|\mathcal{H}_n\right) = E(E(g(Y_{n+1})|\mathcal{H}_n, \mu_{n+1}|\mathcal{H}_n)$$
$$= E(g(Y_{n+1})|\mathcal{H}_n).$$

$\square$

SUPPLEMENTARY MATERIALS

The online supplementary materials contain additional results for various models discussed throughout the article.

REFERENCES

Airoldi, E. M., Anderson, A., Fienberg, S. E., and Skinner, K. K. (2006), "Who Wrote Ronald Reagan's Radio Addresses?," *Bayesian Analysis*, 1, 289–320. [1472]

Aoki, M. (2008), "Thermodynamic Limits of Macroeconomic or Financial Models: One- and Two-Parameter Poisson-Dirichlet Models," *Journal of Economic Dynamics and Control*, 32, 66–84. [1468]

Baladandayuthapani, V., Ji, Y., Nieto-Barajas, L. E., and Morris, J. S. (2010), "Bayesian Random Segmentation Models to Identify Shared Copy Number Aberrations for Array CGH Data," *Journal of the American Statistical Association*, 105, 1358–1375. [1467]

Bassetti, F., Crimaldi, I., and Leisen, F. (2010), "Conditionally Identically Distributed Species Sampling Sequences," *Advances in Applied Probability*, 42, 433–459. [1467,1477,1478]

Berti, P., Pratelli, L., and Rigo, P. (2004), "Limit Theorems for a Class of Identically Distributed Random Variables," *Annals of Probability*, 32, 2029–2052. [1468,1479]

Blackwell, D., and MacQueen, J. B. (1973), "Ferguson Distributions via Pólya Urn Schemes," *The Annals of Statistics*, 1, 353–355. [1466]

Blei, D., and Frazier, P. (2011), "Distance Dependent Chinese Restaurant Processes," *Journal of Machine Learning Research*, 12, 2461–2488. [1467,1470]

Blei, D. M., and Moreno, P. (2001), "Topic Segmentation With an Aspect Hidden Markov Model," *Proceedings of the 24th Annual International ACM SIGIR Conference*, pp. 343–348. [1476]

Cardin, N., Holmes, C., The Wellcome Trust Case Control Consortium, Donnelly, P., and Marchini, J. (2011), "Bayesian Hierarchical Mixture Modeling to Assign Copy Number From a Targeted CNV Array," *Genetic Epidemiology*, 35, 536–548. [1467]

Charalambides, C. A. (2005), *Combinatorial Methods in Discrete Distributions, Wiley Series in Probability and Statistics*, Hoboken, NJ: Wiley-Interscience, Wiley. [1477]

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W. L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Kingsley, C., Dairkee, S., Meng, Z., Chew, K., Pinkel, D., Jain, A., Ljung, B. M., Esserman, L., Albertson, D. G., Waldman, F. M., and Gray, J. W. (2006), "Genomic and Transcriptional Aberrations Linked to Breast Cancer Pathophysiologies," *Cancer Cell*, 10, 529–541. [1467,1473,1474,1475]

Chien, J. T., and Furui, S. (2005), "Predictive Hidden Markov Model Selection for Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, 13, 377–387. [1476]

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., Basset, A. S., Seller, A., Holmes, C. C., and Ragoussis, J. (2007), "QuantiSNP: An Object Bayes Hidden-Markov Model to Detect and Accurately Map Copy Number Variation Using SNP Genotyping Data," *Nucleic Acids Research*, 35, 2013–2025. [1476]

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., METABRIC Group , Langerâd, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Bârresen-Dale, A.-L., Brenton, J. D., Tavaré, S., Caldas, C., and Aparicio, S. (2012), "The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Re-Veals Novel Subgroups," *Nature*, 486, 346–352. [1467,1475]

Dahl, D. B. (2005), "Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models," technical report. [1470]

Dahl, D. B., Day, R., and Tsai, J. W. (2008), "Distance-Based Probability Distribution on Set Partitions with Applications to Protein Structure Prediction," technical report. [1467]

DeSantis, S. M., Houseman, E. A., Coull, B. A., Louis, D. N., Mohapatra, G., and Betensky, R. A. (2009), "A Latent Class Model With Hidden Markov Dependence for Array CGH Data," *Biometrics*, 65, 1296–1305. [1467]

Dewar, M., Wiggins, C., and Wood, F. (2012), "Inference in Hidden Markov Models With Explicit State Duration Distributions," *Signal Processing Letters, IEEE*, 19, 235–238. [1472]

Du, L., Chen, M., Lucas, J., and Carin, L. (2010), "Sticky Hidden Markov Modeling of Comparative Genomic Hybridization," *IEEE Transactions on Signal Processing*, 58, 5353–5368. [1467]

Escobar, M., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588. [1470]

Ferguson, J. D. (1980), "Variable Duration Models for Speech," *Proceedings of the Symposium on the Applications of Hidden Markov Models to Text and Speech*, pp. 143–179. [1472]

Fortini, S., Ladelli, L., and Regazzini, E. (2000), "Exchangeability, Predictive Distributions and Parametric Models," *Sankhya,* Series A, 62, 86–109. [1466]

Fox, E. B., Hughes, M. C., Sudderth, E. B., and Jordan, M. I. (2014), "Joint Modeling of Multiple Time Series via the Beta Process With Application to Motion Capture Segmentation," *Annals of Applied Statistics*, 8, 1281–1313. [1476]

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011), "A Sticky HDP-HMM With Application to Speaker Diarization," *Annals of Applied Statistics*, 5, 1020–1056. [1467,1476]

Griffiths, T. L., Sanborn, A. N., Canini, K. R., and Navarro, D. J. (2008), "Categorization as Nonparametric Bayesian Density Estimation," in *The Probabilistic Mind: Prospects for Rational Models of Cognition*, eds. M. Oaksford, and N. Chater, Oxford: Oxford University Press, pp. 303–328. [1466]

Guedon, Y. (2011), "Estimating Hidden Semi-Markov Chains From Discrete Sequences," *Journal of Computational and Graphical Statistics*, 12, 604–639. [1475]

Guha, S. (2010), "Posterior Simulation in Countable Mixture Models for Large Datasets," *Journal of the American Statistical Association*, 105, 775–786. [1467]

Guha, S., Li, Y., and Neuberg, D. (2008), "Bayesian Hidden Markov Modeling of Array CGH Data," *Journal of the American Statistical Association*, 103, 485–497. [1467,1474]

Hansen, B., and Pitman, J. (2000), "Prediction Rules for Exchangeable Sequences Related to Species Sampling," *Statistics and Probability Letters*, 46, 251–256. [1466]

Heller, R., Stanley, D., Yekutieli, D., Rubin, N., and Benjamini, Y. (2006), "Cluster-Based Analysis of fMRI Data," *Neuroimage*, 33, 599–608. [1475]

Hilbe, J. M. (2011), *Negative Binomial Regression*, Cambridge: Cambridge University Press. [1472]

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010), *Bayesian Nonparametrics*, Cambridge: Cambridge University Press. [1466]

Ishwaran, H., and Zarepour, M. (2003), "Random Probability Measures via Pólya Sequences: Revisiting the Blackwell-MacQueen Urn Scheme." Available at *http://arxiv.org/abs/math/0309041*. [1466]

Jain, S., and Neal, R. M. (2007), "Splitting and Merging Components of a Nonconjugate Dirichlet Process Mixture Model," *Bayesian Analysis*, 3, 445–472. [1470]

Jara, A. (2007), "Applied Bayesian Non- and Semi-Parametric Inference using DP package," *Rnews*, 7, 17–26. [1471]

Jbabdi, S., Woolrich, M. W., and Behrens, T. E. J. (2009), "Multiple-Subjects Connectivity-Based parcellation Using Hierarchical Dirichlet Process Mixture Models," *NeuroImage*, 44, 373–384. [1466]

Ji, Y., Lu, Y., and Mills, G. (2008), "Bayesian Models Based on Test Statistics for Multiple Hypothesis Testing Problems," *Bioinformatics*, 24, 943–939. [1475]

Kim, S., Tadesse, M. G., and Vannucci, M. (2006), "Variable Selection in Clustering via Dirichlet Process Mixture Models," *Biometrika*, 93, 877–893. [1466]

Lee, J., Quintana, F., Müller, P., and Trippa, L. (2013), "Defining Predictive Probability Functions for Species Sampling Models," *Statistical Science*, 28, 209–222. [1466]

Liu, Z., Windle, J., and Scott, J. C. (2013), "Case Studies in Bayesian Screening for Time-Varying Model Structure," in *Bayesian Theory and Applications*, eds. P. Damien, P. Dellaportaas, N. G. Ploson, and D. A. Stephens, Oxford: Oxford University Press, pp. 337–442. [1476]

MacEachern, S. N., and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–238. [1470]

Marioni, J. C., Thorne, N. P., Tavare, S., and Radvanyi, F. (2006), "BioHMM: A Heterogeneous Hidden Markov Model for Segmenting Array CGH Data," *Bioinformatics*, 22, 1144–1146. [1467]

Mitchell, C., Harper, M., and Jamieson, L. (1995), "On the Complexity of Explicit Duration Hmms," *IEEE Transactions on Speech and Audio Processing*, 3, 213–217. [1472]

Monteiro, J. V., Assunçao, R. M., and Loschi, R. H. (2011), "Product Partition Models With Correlated Parameter," *Bayesian Analysis*, 6, 691–726. [1476]

Müller, P., Parmigiani, G., and Rice, K. (2007), "FDR and Bayesian Multiple Comparisons Rules," in *Bayesian Statistics 8*, eds. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, Oxford: Oxford University Press, 349–370. [1475]

Müller, P., and Quintana, F. (2010), "Random Partition Models With Regression on Covariates," *Journal of Statistical Planning and Inference*, 140, 2801–2808. [1467]

Navarro, D. J., Griffiths, T. L., Steyvers, M., and Lee, M. D. (2006), "Modeling Individual Differences Using Dirichlet Processes," *Journal of Mathematical Psychology*, 50, 101–122. [1466]

Neal, R. M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265. [1470]

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004), "Detecting Differential Gene Expression With a Semiparametric Hierarchical Mixture Method," *Biostatistics*, 5, 155–176. [1475]

Park, J. H., and Dunson, D. B. (2010), "Bayesian Generalized Product Partition Model," *Statistica Sinica*, 20, 1203–1226. [1466]

Pitman, J. (1996), "Some Developments of the Blackwell-MacQueen Urn Scheme," in *Statistics, Probability and Game Theory; Papers in honor of David Blackwell* (*Lecture Notes-Monograph Series,* Vol. 30), eds. T. S. Ferguson et al., Hayward, CA: Institute of Mathematical Statistics, pp. 245–267. [1466]

——— (2006), *Combinatorial Stochastic Processes. Ecole d'Eté Probabilités de Saint-Flour XXXII 2002, Lecture Notes in Mathematics*, Berlin/Heidelberg: Springer. [1468]

Rabiner, L. R. (1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, 77, 257–287. [1472,1475]

Redon, R., Fitzgeral, T., and Carter, N. P. (2009), "Comparative Genomic Hybridization: DNA Labeling, Hybridization and Detection," *Methods in Molecular Biology*, 529, 267–278. [1473]

Ren, L., Dunson, D., Lindroth, S., and Carin, L. (2010), "Dynamic Nonparametric Bayesian Models for Analysis of Music," *Journal of the American Statistical Association*, 105, 458–472. [1476]

Rodriguez, A., and Dunson, D. B. (2011), "Nonparametric Bayesian Models Through Probit Stick-Breaking Processes," *Bayesian Analysis*, 6, 145–178. [1476]

Storey, J. (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the *q*-Value," *The Annals of Statistics*, 31, 2013–2035. [1475]

Storey, J., Dai, J., and Leek, J. (2007), "The Optimal Discovery Procedure for Large-Scale Significance Testing, With Applications to Comparative Microarray Experiments," *Biostatistics*, 8, 414–432. [1475]

Sudderth, E. B., and Jordan, M. I. (2008), "Shared Segmentation of Natural Scenes Using Dependent Pitman-Yor Processes," in *Advances in Neural Information Processing Systems*, pp. 1585–1592. [1466]

Sun, W., Reich, B., Cai, T., Guindani, M., and Schwartzman, A. (in press), "False Discovery Control in Large Scale Spatial Multiple Testing," *Journal of the Royal Statistical Society,* Series B. [1475]

Taramasco, O., and Bauer, S. (2012), RHmm: Hidden Markov Models Simulations and Estimations R Package, version 2.0.3. [1472]

Teh, Y. W. (2006a), "A Hierarchical Bayesian Language Model Based on Pitman-Yor Processes," in *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, Morristown, NJ: Association for Computational Linguistics, pp. 985–992. [1466]

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006b), "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566–1581. [1467,1476]

Wallach, H., Sutton, C., and McCallum, A. (2008), "Bayesian Modeling of Dependency Trees Using Hierarchical Pitman-Yor Priors," in *Proceedings of the Workshop on Prior Knowledge for Text and Language* (held in conjunction with ICML/UAI/COLT), Helsinki, Finland, pp. 15–20. [1466]

Yau, C., and Holes, C. C. (2013), "A Decision-Theoretic Approach for Segmental Classification," *Annals of Applied Statistics*, 7, 1814–1835. [1476]

Yau, C., Papaspiliopoulos, O., Roberts, G. O., and Holmes, C. (2011), "Bayesian Nonparametric Hidden Markov Models with Application to the Analysis of copy-Number-Variation in Mammalian Genomes," *Journal of the Royal Statistical Society,* Series B, 73, 37–57. [1467,1475,1476]

Yu, S.-Z. (2010), "Hidden Semi-Markov Models," *Artificial Intelligence*, 174, 215–243. [1472]