Species sampling priors for modeling dependence: an application to the detection of chromosomal aberrations

Federico Bassetti,[*] Fabrizio Leisen,[†]
Edoardo Airoldi[‡] and Michele Guindani[§]

April 2, 2015

Abstract

We discuss a class of Bayesian nonparametric priors that can be used to model local dependence in a sequence of observations. Many popular Bayesian nonparametric priors can be characterized in terms of exchangeable species sampling sequences. However, in some applications, common exchangeability assumptions may not be appropriate. We discuss a generalization of species sampling sequences, where the weights in the predictive probability functions are allowed to depend on a sequence of independent (not necessarily identically distributed) latent random variables. More specifically, we consider conditionally identically distributed (CID) Pitman-Yor sequences and the Beta-GOS sequences recently introduced by Airoldi et al. (2014). We show how those processes can be used as a prior distribution in a hierarchical Bayes modeling framework, and, in particular,

[*] Universitá di Pavia, Dipartimento di Matematica, via Ferrata, 1 27100 Pavia, Italy, federico.bassetti@unipv.it

[†] University of Kent, School of Mathematics, Statistics and Actuarial Sciences, Cornwallis building, CT2 7NF, Canterbury, Kent, UK,fabrizio.leisen@gmail.com

[‡] Harvard University, Department of Statistics, 1 Oxford Street, Cambridge, MA 02138, USA, airoldi@fas.harvard.edu

[§] University of Texas, MD Anderson Cancer Center, Department of Biostatistics, Houston, TX, USA,mguindani@mdanderson.org

how the Beta-GOS can provide a reasonable alternative to the use of nonhomogenous Hidden Markov models, further allowing unsupervised clustering of the observations in an unknown number of states. The usefulness of the approach in biostatistical applications is discussed and explicitly shown for the detection of chromosomal aberrations in breast cancer.

1 Introduction

Due to their clustering properties, Bayesian nonparametric methods have been widely employed for the analysis of various types of data in genetics, e.g. for identifying disease subtypes and isolating discriminating genes, proteins or samples (see, e.g., Kim et al., 2006; Guindani et al., 2009; Lee et al., 2013). In order to take into account measurement characteristics (e.g., continuos support, long tails, skewness, multimodality or overdispersion of the frequency distribution), it is often convenient to employ a hierarchical model specification. At the top level of the hierarchy, observations are assumed to be conditionally independent given some "latent" process, i.e. the sampling distribution is

$$y_i|\theta_i \overset{ind}{\sim} p(y_i|\theta_i) \quad i = 1, 2, \dots \tag{1}$$

where $p(\cdot|\theta_i)$ denotes a probability density function or probability mass function, dependent on the values of a set of parameters θ_i. The distribution of the θ_i's is then assumed to follow a process that captures relevant features of the data. Let p denote the unknown distribution of the model parameters, and Q be a prior probability measure for p. Then, the hierarchical model specification can be concisely described as follows,

$$\begin{aligned} \theta_1, \theta_2, \dots |p &\sim p \\ p &\sim Q. \end{aligned} \tag{2}$$

Model (1)–(2) schematically encompasses both popular Dirichlet Process mixtures (Lo, 1984) as well as Dependent Dirichlet Process mixtures (MacEachern, 1999). The prior process p can often be represented by means of a sequence of predictive distributions that typically encode exchangeability assumptions on the model parameters and the data (see Section 2). In some applications, however, the usual exchangeability assumptions may be hardly justified. For example, if $\theta_1, \theta_2, \dots$ represent a process in time (space), then the model should properly account for the dependence relations among nearby time points (neighboring locations).

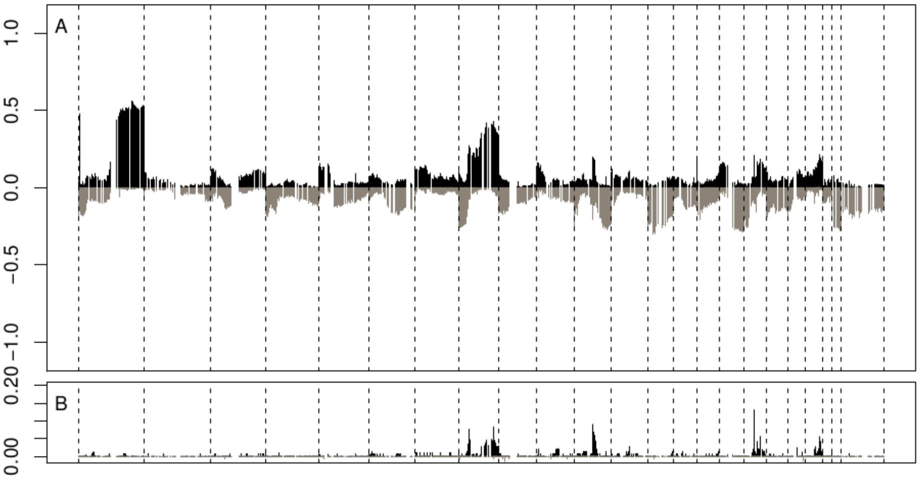

Figure 1: (a) Frequencies of genome copy number gains and losses plotted as a function of genomic location. (b) Frequency of tumors showing high-level amplification. The dashed vertical lines separate the 23 chromosomes.

To illustrate the point, in Figure 1 (a) we consider the frequency of genome copy number abnormalities, as estimated from data obtained in a classical study of the genetic determinants of breast cancer pathophysiologies (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 Comparative Genomic Hybridization (CGH). Regions of relative gains or losses are identified by measuring the fluorescence ratio 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 et al., 2009). The reference DNA is assumed to have two copies of each chromosome. If the test sample has no copy number aberrations, the log2 of the intensity ratio is theoretically equal to zero.

Array CGH data are typically very noisy and spatially correlated. More specifically, copy number gains or losses at a region are often associated to an increased probability of gains and losses at a neighboring region. Bayesian models for array CGH data have been recently investigated by Guha et al. (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. propose 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. extend this approach and propose a supervised Bayesian latent class approach for classification of the clones, which relies on a heterogenous hidden Markov model to account for local dependence in the intensity ratios. In a heterogeneous hidden Markov model, the transition probabilities between states depend on each single clone or the the distance between adjacent clones (Marioni et al., 2006). Using a Bayesian nonparametric approach, Du et al. propose a sticky Hierarchical DP-HMM (Fox et al., 2011; Teh et al., 2006) to infer the number of states in an HMM, while also imposing state persistence. Yau et al. (2011) also propose a nonparametric Bayesian HMM, but use instead a DP mixture to model the likelihood in each state.

In this chapter, we present an alternative approach, which flexibly models the evolution of the parameters $\theta_1, \theta_2, \ldots$ by means of a general class of non-exchangeable species sampling sequences. As it is typical in a Bayesian nonparametric setting, we allow clustering of the observations, further assuming that the number of states is unknown and can be inferred from the data. Furthermore, in finite HMMs, the distribution of state durations are necessarily restricted to a geometric form, so that departures from this assumption, e.g. state persistence, must be appropriately accounted for in the modeling (Yu, 2010; Fox et al., 2011; Johnson and Willsky, 2013). The species sampling priors, which we discuss in the next Section, model "non-homogenous" assumptions in the state durations more flexibly, since the weights in the species sampling rule can adapt to take into account local dependences in the data.

2 Species Sampling Sequences: Basics and Extensions

In this Section, we review basic definitions and properties of species sampling (SS) sequences, and also discuss their generalizations to a class of random sequences that are appealing for modeling non-exchangeable observations.

More specifically, for defining SS-sequences, we refer to the hierarchical formulaton (2), and charaterize the sequence of random variables $\theta_1, \theta_2, \ldots$ by means of the sequence of predictive probability functions,

$$P\{\theta_{n+1} \in \cdot \,|\theta_1, \ldots, \theta_n\} = \sum_{i=1}^{n} q_{n,i}\delta_{\theta_i}(\cdot) + q_{n,n+1}G_0(\cdot), \tag{3}$$

where $\delta_x(\cdot)$ denotes a point mass at x, and G_0 is a non-atomic probability measure (base measure, Pitman, 1996). The weights $q_{n,i}$, $i = 1,\dots,n+1$, are non–negative functions of $(\theta_1,\dots,\theta_n)$, such that $\sum_{i=1}^{n+1} q_{n,i} = 1$, and define the probability that the sampled value of θ_{n+1} coincides with one of the previous values in the sequence or is a new draw from the base measure. In (3), it is implicitly assumed that $\theta_1 \sim G_0$. If $q_{n,n+1} < 1$, there's a positive probability of ties among the θ_i's, that is some of the θ_i's will share a common value. We can collect the unique values in a vector $(\theta_1^*,\dots\theta_{K_n}^*)$, where K_n indicates the (random) number of distinct values in the subsequence $\theta(n) = (\theta_1,\dots,\theta_n)$. Alternatively, we can say that (3) implicitly defines a random partition $\Pi^{(n)} = \{\Pi_1^{(n)},\dots,\Pi_{K_n}^{(n)}\}$ of the set $\{1,\dots,n\}$ into K_n blocks, where $i \in \Pi_j^{(n)}$ if and only if $\theta_i = \theta_j^*$.
If the probability of a tie, $P(\theta_{n+1} = \theta_j^* | \theta(n))$, depends only on the cardinality of each block, i.e. the frequency $n_{jn} = |\Pi_j^{(n)}|$ of each value θ_j^* in $\theta(n)$, $j = 1,\dots,K_n$, then the sequence $\theta_1, \theta_2, \dots$ is exchangeable. The result characterizes all exchangeable SS-sequencies (see Fortini et al., 2000; Hansen and Pitman, 2000; Lee et al., 2008, for more details). The most notable example of exchangeable SS-sequences is the Blackwell MacQueen sampling rule, which defines a Dirichlet Process (see Blackwell and MacQueen, 1973; Ishwaran and Zarepour, 2003). Let p be a DP with mass parameter γ and base measure $G_0(\cdot)$, denoted as $p \sim DP(\gamma, G_0)$. Then, the corresponding sequence of predictive probability function is the well-known Blackwell MacQueen sampling rule, which sets $q_{n,i} = \frac{1}{n+\gamma}$ and $q_{n,n+1} = \frac{\gamma}{n+\gamma}$ in (3).

The dependence of the weights only on the sequence $\theta(n)$ may be seen as a limiting feature in some applications, e.g. whenever one could contemplate that additional covariate information might affect the clustering of the observations. For example, Park and Dunson (2010) propose a generalized product partition model (GPPM) in which the clustering process is predictor-dependent. Their GPPM relax the exchangeability assumption through the incorporation of predictors, implicitly defining a generalized Pólya urn scheme. Similarly, Müller and Quintana (2010) define a product partition model that includes a regression on covariates, which allows units with similar covariates to have greater probability of being clustered together.

Here, we consider a generalization of the predictive rule (3), where the weights are allowed to depend on a sequence of independent (not necessarily identically distributed) latent random variables $W_1, W_2, \dots$ More specifically, we consider a sequence $(\theta_n)_{n\geq 1}$ characterized by the following predictive distributions,

$$P\{\theta_{n+1} \in \cdot \, | \theta(n), W(n)\} = \sum_{i=1}^{n} p_{n,i} \delta_{\theta_i}(\cdot) + r_n G_0(\cdot), \qquad (4)$$

where $W(n) = (W_1, \ldots, W_n)$ and the weights $p_{n,i}$ are strictly positive functions of the partitions $\Pi^{(n)}$ and the random variables $W(n)$, i.e. $p_{n,i} = p_{n,i}(\Pi^{(n)}, W(n)) > 0$, with $\sum_{i=1}^{n} p_{n,i} < 1$ and $r_n := 1 - \sum_{i=1}^{n} p_{n,i}$.

The specific choice of the weights $p_{n,i}$'s determines the clustering behavior of the sequence $(\theta_n)_n$. In this chapter, we focus on the general class of *conditionally identically distributed* (CID) sequences (Berti et al., 2004). This class generalizes the notion of exchangeable sequences, while still preserving some of their important characteristics. Formally, a sequence $(\theta_n)_{n\geq 1}$ is CID with respect to a filtration $\mathscr{G} = (\mathscr{G}_n)_{n\geq 0}$, whenever for each $n \geq 0$ all the random variables θ_{n+i}, with $i \geq 1$, are identically distributed conditionally on $\mathscr{G}_n$. In the definition it is assumed that $\mathscr{G}$ contains the natural filtration of $(\theta_i)_{i\geq 1}$. It is clear that every exchangeable sequence is a CID sequence with respect to its natural filtration, but a CID sequence does not necessarily need to be exchangeable nor stationary. Indeed, if a CID sequence is stationary then it is also exchangeable. A remarkable property of CID sequences is that the θ_i's are marginally identically distributed. No representation theorem is known for CID sequences. However, it can be shown that given any bounded and measurable function f, the predictive mean $E[f(\theta_{n+1})|\theta_1, ..., \theta_n]$ and the empirical mean $\frac{1}{n}\sum_{i=1}^{n} f(\theta_i)$ converge to the same limit as n goes to infinity. For details, we refer to Berti et al. (2004). Finally, if the sequence of observations $(Y_1, Y_2, \ldots)$ follows the hierarchical model (1) and the latent process $(\theta_1, \theta_2, \ldots)$ is a CID sequence, then it can be shown that the sequence of observations Y_i's also forms a CID sequence. This result has been proved in Airoldi et al. (2014) specifically for the Beta-GOS prior (see below); however, the proof can be easily extended to a general CID sequence.

Two interesting types of CID sampling sequences are the following:

a) **CID Pitman-Yor sequences.** A Pitman-Yor process (Pitman, 2006), is an exchangeable sequence characterized by the following predictive probability functions,

$$P\{\theta_{n+1} \in \cdot \, | \theta_1, \ldots, \theta_n\} = \textstyle\sum_{j=1}^{K_n} \displaystyle\frac{n_{jn} - \alpha}{\gamma + n} \delta_{\theta_j^*}(\cdot) + \frac{\gamma + \alpha K_n}{\gamma + n} G_0(\cdot), \qquad (5)$$

for $\gamma > 0$ and $\alpha \in [0,1]$, as a function of the partition $\Pi^{(n)} = \{\Pi_1^{(n)}, \ldots, \Pi_{K_n}^{(n)}\}$ of the set $\{1, \ldots, n\}$ into K_n blocks. When $\alpha = 0$, the sequence (5) defines a Dirichlet Process, $DP(\theta, G_0)$. There exists a generalization of the classical Pitman-Yor process (5) as a CID sequence. More specifically, the CID generalization assumes that the weights in (4) are functions of a sequence of random variables $W(n)$, with weights $p_{n,i}(\Pi^{(n)}, W(n)) = (W_i - \alpha/n_{k_i n})/(\gamma + \sum_{j=1}^{n} W_j)$ and $r_n(\Pi^{(n)}, W(n)) = (\gamma + \alpha K_n)/(\gamma + \sum_{j=1}^{n} W_j)$ where $n_{k_i n}$ denotes the cardinality of the block in $\Pi^{(n)}$ that contains observation i. Then,

$$P\{\theta_{n+1} \in \cdot \,|\theta(n), W(n)\} = \sum_{j=1}^{K_n} \frac{\left(\sum_{i \in \Pi_j^{(n)}} W_i\right) - \alpha}{\gamma + \sum_{i=1}^{n} W_i} \delta_{\theta_j^*}(\cdot) + \frac{\gamma + \alpha K_n}{\gamma + \sum_{i=1}^{n} W_i} G_0(\cdot), \tag{6}$$

which reduces to (5) if $W_n = 1$. Similarly to the Chinese Restaurant Process (CRP) representation of the Dirichlet Process, equation (6) has an intuitive illustration in terms of the seating allocation at a restaurant. In this representation, each customer enters the restaurant with a distinctive "mark" (the random variables W_i's). When customers enter the restaurant, they have the possibility to start a new table (with probability dependent on the parameter γ) or join a table already occupied by other customers. In the CID version, the "attractiveness" of a table depends on $\sum_{i \in \Pi_j^{(n)}} W_i$ in (6), i.e. the sum of the individual marks for each customer already seating at the table. In other words, the process takes into account possible additional variability in the "seating plan" due to individual random effects.

In terms of clustering, the asymptotic behavior of the CID version of the DP, obtained by setting $\alpha = 0$ in (6), is similar to that of the classical DP: if the W_i's are i.i.d. with finite variance and mean $E[W_i] = m$, then $K_n/\log(n)$ converges almost surely to γ/m (see Bassetti et al., 2010, Example 5.8). The situation is less simple for the case in which $\alpha \neq 0$.

b) **Beta-GOS sequences.** An alternative specification of (4) considers weights obtained as a product of independent Beta random variables. More specifically, Airoldi et al. (2014) assume that the random variables $(W_i)_{i \geq 1}$ are draws from independent Beta(α_i, β_i) distributions, and then set $p_{n,i} = (1 - W_i)\prod_{j=i+1}^{n} W_j$ and $r_n = \prod_{j=1}^{n} W_j$ in (4). The resulting sequence $(\theta_1, \theta_2, \ldots)$ defines the so-called *Beta-GOS* sequence, a particular case of a Generalized Ottawa Sequence (GOS) in the class of CID sequences (see, for details, Bassetti et al., 2010). The choice of Beta latent variables allows for a flexible specification of the species sampling weights, while it still retains simplicity and interpretability of the sequence allocation scheme. As a matter of fact, this allocation rule can also be described in terms of a preferential attachment scheme, similarly to the CID Pitman-Yor sequences. Also in this scheme, each customer, θ_i, is characterized by a random weight (or "mark"), $1-W_i$, and can join the table where any of the previous customer is sitting by means of a "geometric-type" assignment scheme. More precisely, suppose we have customers $\theta_1,\ldots,\theta_n$ together with their marks up to time n, $(1-W_1,\ldots,1-W_n)$. Then, the $(n+1)$-th individual will be assigned to the same table as the previous customer, θ_n, with probability $1-W_n$; the probability of pairing θ_{n+1} to θ_{n-1} will be $W_n(1-W_{n-1})$, and so forth. In general, in this representation, W_i will represent a "repulsion" score associated to customer i. Thus, each weight $p_{n,i}$ will be represented by the product of the W_j's associated to the latest $n-j$ subjects and the "mark" or "attractiveness" score of customer i, $1-W_i$. Summarizing, customer θ_{n+1} will occupy a new table (i.e., $\theta_{n+1}\sim G_0$) with probability r_n, or instead they will join one of the previously occupied tables, say table j, with probability $\sum_{i:\theta_i=\theta_j^*} p_{n,i}$. Of course, the seating assignment and the clustering behavior of the sequence is determined by the specification of the parameters α_i and β_i in the distribution of the W_i's. We briefly discuss the issue in the next Section, where we review some asymptotic results and their interpretation in terms of clustering of the sequence, for a set of parameter specifications.

In the next Sections, we will focus specifically on the use of the Beta-GOS sequences for modeling latent dependence in Bayesian hierarchical models and we will discuss their application to the detection of chromosomal aberrations in array CGH data.

3 A Beta-GOS Hierarchical Model

In this Section, we focus on the Beta-GOS sequences and discuss how they can be used to define a prior in a hierarchical model (for a broader discussion, see Airoldi et al., 2014). Although the discussion pertains specifically to the Beta-GOS process, the basic modeling idea naturally extends to the CID Pitman-Yor sequences and the general CID sequences. We then discuss the prior specification of the parameters of the Beta random variables in the Beta-GOS. Finally, we briefly present the MCMC sampling algorithm for conducting posterior inference with this type of models.

Similarly to the hierarchical Bayesian specification in (1)-(2), we can assume that at the highest level of the hierarchy the sampling distribution is specified as

$$Y_i|\theta_i \overset{ind.}{\sim} f(y_i|\theta_i), \quad i = 1,\dots,n, \tag{7}$$

where the vector $(\theta_1,\dots,\theta_n)^T$ is a realization of a Beta-GOS process characterized by auxiliary random variables $W_i \sim Be(\alpha_i,\beta_i)$, $i = 1,\dots,n$, and base measure G_0. We can succinctly denote the Beta-GOS prior as

$$\theta_1,\dots,\theta_m \sim \text{Beta-GOS}(\boldsymbol{\alpha}_n,\boldsymbol{\beta}_n,G_0), \tag{8}$$

where $\boldsymbol{\alpha}_n = (\alpha_1,\dots\alpha_n)$ and $\boldsymbol{\beta}_n = (\beta_1,\dots,\beta_n)$. As discussed in Section 2, the Beta-GOS is a particular case of a CID sequence. Hence, in particular, marginally $\theta_i \sim G_0$, $i = 1,\dots,n$. Therefore, the base G_0 can be regarded as a centering distribution, as it is typical in DP mixture models: G_0 represents a vague parametric prior assumption on the distribution of the parameters of interest. The hierarchical model may be extended by putting hyper-priors on the remaining parameters of the model, including the hyper-parameters of the base measure G_0 as well as the vectors $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}_n$.

The parameters of the Beta random variables control the asymptotic behavior of the sequence, and the clustering properties of the prior. For example, if we set $\alpha_i = i+\gamma-1, \beta_i = 1$, for given $\gamma > 0$, then $K_n/\log(n)$ converges in distribution to a $Gamma(\gamma,1)$ random variable. As a comparison, for a $DP(\gamma,G_0)$, it is well-known that $K_n/\log(n)$ converges almost surely to γ. If we set $\alpha_i = a, \beta_i = b$, for some $a,b > 0$, then K_n converges almost surely to a finite random variable. This result naturally implies that the resulting partition is characterized by a few big clusters, as n increases. We refer to Airoldi et al. (2014) for further details and proofs. In addition, the parameters α_i and β_i implicitly 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. More specifically, setting $\alpha_i = \gamma - 1 + j$ $(\gamma > 0)$ and $\beta_i = 1$ implies that $E[r_n] = \gamma/(\gamma+n)$ and $E[p_{n,i}] = 1/(\gamma+n)$, $i = 1,\dots,n$. This specification can be seen as a feature of a process with a long memory, since all the previous observations have the same weight on average. For $\alpha_i = a, \beta_i = b$, $E[r_n] = (a/(a+b))^n$ and $E[p_{n,i}] = (a/(a+b))^{n-i}(b/(a+b))$. Hence, the probabilities of ties d decrease exponentially as a function of the lag $n-i$, describing a short memory process. 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_i and β_i. As a general rule, following what it is usually done with Dirichlet processes priors, we suggest to elicit the parameters on the basis of the expected number of clusters *a priori*, i.e. $E(K_n) = 1 + \sum_{j=1}^{n-1} E[r_j]$. For example, one could set $\alpha_i = a$ and $\beta_i = b$ 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, in Section 5 we will follow the previous guidelines in the application to the detection of chromosomal aberrations, since biological considerations lead to expect the true number of states to be around 4. On the other hand, the single-parameter specification $\alpha_j = j + \gamma - 1$, $\beta_j = 1$ should be the default choice in those applications where prior information on the expected number of clusters is more vague, and the choice of the parameter γ should be based on $E(K_n) = \sum_{j=0}^{n-1} \frac{\gamma}{\gamma+j} \sim \gamma \log(n)$, for large n.

3.1 MCMC Posterior Sampling

Posterior inference for the model (7)-(8) entails learning about the clustering and corresponding estimates of the parameters θ_i. In this Section, we describe a Gibbs sampler scheme. The basic idea is to describe the partition $\Pi^{(n)}$ by introducing a sequence of labels C_i, $i = 1, \ldots, n$ which record the pairing of observation i with one of the previous observations, $j < i$. 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 Dirichlet process, 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 i-th pairing label. In particular, if the i-th observation is not paired to any of the preceding ones, we set $C_i = i$. Then, θ_i is a draw from the base distribution G_0, and thus it 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 non-exchangeable processes, as described in Blei and Frazier (2011) and Airoldi et al. (2014). It is easy to see that the pairing sequence $(C_n)_{n\geq 1}$ assigns $C_1 = 1$ and has full conditional distribution

$$\begin{aligned} P\{C_n = i | C_1, \ldots, C_{n-1}, W\} &= P\{C_n = i | W_1, \ldots, W_{n-1}\} \\ &= r_{n-1} \mathbb{I}\{i = n\} + p_{n-1,i} \mathbb{I}\{i \neq n\}, \end{aligned} \tag{9}$$

for $i = 1, \ldots, n$, where $\mathbb{I}(\cdot)$ denotes the indicator function, such that, given a set A, $\mathbb{I}(A) = 1$ if A is true and 0 otherwise. 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 $C(n) = (C_1, \ldots, C_n)$, then we denote by $\Pi(C(n))$ the partition generated by the pairings $C(n)$, i.e. $\Pi^{(n)}$. For any n and any $i \leq n$, let $C_{-i} = (C_1, \ldots, C_{i-1}, C_{i+1}, \ldots, C_n)$; analogously, let $W(n) = (W_1, \ldots, W_n)$, and $W_{-i} = (W_1, \ldots, W_{i-1}, W_{i+1}, \ldots, W_n)$. Then, the full conditional for the pairing indicators C_i's is

$$\begin{aligned} P\{C_i = j | C_{-i}, Y(n), W(n)\} &\propto P\{C_i = j, Y(n) | C_{-i}, W(n)\} \\ &= P\{Y(n) | C_i = j, C_{-i}, W(n)\} P\{C_i = j | C_{-i}, W(n)\}. \end{aligned} \tag{10}$$

The second term in (10) is the prior predictive rule (9), whereas

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

where $\Pi(C_{-i}, j)$ denotes the partition generated by $(C_1, \ldots, C_{i-1}, j, C_{i+1}, \ldots, C_n)$. If G_0 and $f(y|\theta)$ are conjugate, the latter integral has a closed form solution. The non-conjugate case could be handled by appropriately adapting the algorithms of MacEachern and Müller (1998) and Neal (2000). As far as the full conditional for the latent variables W_i's, we can show that $W_i | C(n), W_{-i}, Y(n) \sim \text{Beta}(A_i, B_i)$, where $A_i = \alpha_i + \sum_{j=i+1}^{n} \mathbb{I}\{C_j < i \text{ or } C_j = j\}$, and $B_i = \beta_i + \sum_{j=i+1}^{n} \mathbb{I}\{C_j = i\}$; hence, they depend only on the clustering configurations and not on the values of W_{-i}.

Then, let's consider the set of cluster centroids θ_i^*'s. The algorithm above allows faster mixing of the chain by integrating over the distribution of the θ_i^*. However, in case inference on the vector $(\theta_1, \ldots, \theta_m)$ is of interest, it is possible to sample the unique cluster values at each iteration, as

$$\theta_j^* | C(n), W(n), Y(n) \propto \prod_{i \in \Pi_j(n)} p(Y_i | \theta_j^*) G_0(d\theta_j^*), \tag{11}$$

where $\Pi_j(n)$ denotes the partition set of those observations with $\theta_i = \theta_j^*$, $i = 1, \ldots, n$. Again, if $f(y|\theta)$ and G_0 are conjugate, the full conditional of θ_j^* is available in closed form, otherwise we can update θ_j^* by standard Metropolis Hastings algorithms (Neal, 2000).

Finally, we note that if $\pi(\boldsymbol{\alpha}_n, \boldsymbol{\beta}_n)$ is a prior distribution for the Beta hyper-parameters $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}_n$, one could implement a Metropolis Hasting scheme to learn about their posterior distribution, since

$$\boldsymbol{\alpha}_n, \boldsymbol{\beta}_n | C(n), Y(n) \propto \pi(\boldsymbol{\alpha}_n, \boldsymbol{\beta}_n) \prod_{i=1}^{n} \frac{B(A_i, B_i)}{B(\alpha_i, \beta_i)}, \tag{12}$$

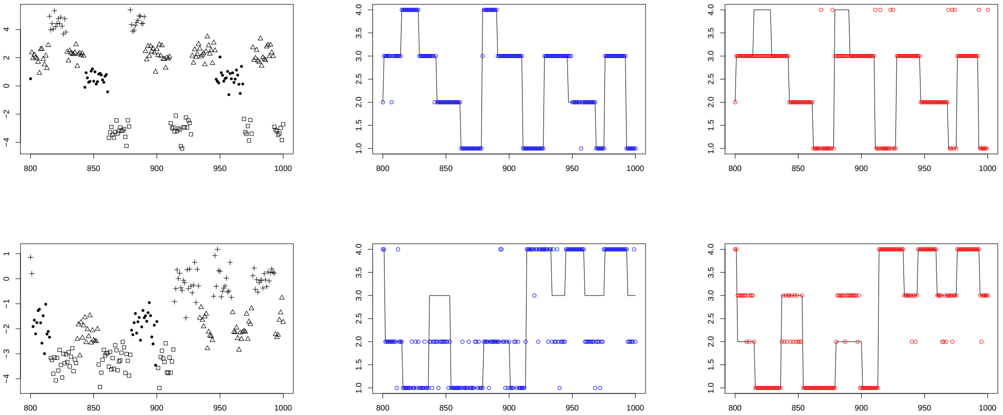

Figure 2: Illustrative segmentation-type plots for the simulation study in Section 4. Right column: subset of data for two replicates. Center column *top*: an example of allocation for a Beta-GOS($\alpha_i = 1, \beta_i = 1$) plotted *vs* the truth (black line); *bottom* considers a Beta-GOS($\alpha_i = i, \beta_i = 1$). Left column illustrates the fitting by a HMM with 4 states.

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 (12) is an adaptation of well known results for the Dirichlet Process (Escobar and West, 1995).

4 A Comparison with Hidden Semi-Markov Models

In many problems (e.g. change point detection), hidden Markov Models are used as computationally convenient substitutes for temporal processes that are known to be more complex than what could be implied by first order Markovian dynamics. Here, we generate non-exchangeable 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. Hidden semi-Markov processes are an extension of the popular hidden Markov model 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 hidden Markov models. Therefore, hidden semi-Markov process have also been referred to as "hidden Markov

Models with explicit duration" (Mitchell et al., 1995; Dewar et al., 2012) or "variable-duration hidden Markov Models" (Rabiner, 1989).

We generate 1,000 datasets (1000 observations each) using a hidden semi-Markov process with four states and a negative binomial distribution for the state occupancy distribution. More specifically, we parametrize the negative binomial in terms of its mean and an ancillary parameter, which is directly related to the amount of overdispersion of the distribution (Hilbe, 2011; Airoldi et al., 2006). 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 $\text{NegBin}(15, 0.15)$, which corresponds to assuming a large overdispersion (17.25). We also consider $\tau = 0.5$ for the noise. We fit the data by means of a Beta-GOS model with Beta hyper-parameters defined by: a) $\alpha_i = i, \beta_i = 1$; b) $\alpha_i = 5, \beta_i = 1$; c) $\alpha_i = 1, \beta_i = 1$, $i = 1, \ldots, n$. 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 hidden Markov models, assuming 3, 4 and 5 states, respectively. Results from the simulations are reported in Table 4, where the HMM was implemented using the R package "RHmm" (Taramasco and Bauer, 2012). Table 4 shows that the Beta-GOS is a viable alternative to HMM, as it can provide more accurate inference than a single hidden Markov model where the number of states is fixed a priori. The fit obtained with the Beta-GOS appears quite robust to the different choices of the hyper-parameters. Figure 2 illustrates the clustering induced by the Beta-GOS and a 4-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($\alpha_i = 1, \beta_i = 1$) (*top*) and a Beta-GOS($\alpha_i = i, \beta_i = 1$) (*bottom*), whereas column (c) illustrates the clustering attained by the HMM. 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 intra-claster 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($\alpha_i = i, \beta_i = 1$) than for the "informative" Beta-GOS($\alpha_i = a, \beta_i = b$) formulations. This is in accordance with the discussion in Section 3 and, in particular, with the consideration that a Beta-GOS($\alpha_i = i, \beta_i = 1$) should represent a long memory process.

i) Data Generating Process: Hidden Semi Markov Model (HSMM) with 4 states and $\mathrm{NegBin}(15, 0.15)$						
Model Fitting Method	**Beta-GOS**			**HMM**		
	$\alpha_n = n; \beta_n = 1$	$\alpha_n = 5; \beta_n = 1$	$\alpha_n = 1; \beta_n = 1$	3 States	4 States	5 States
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 3: Summary statistics for the simulation studies described in Section 4. The table compares the Beta-GOS and a hidden Markov model under different specifications of hyper-parameters. The data generating process assumes a hidden semi-Markov with state occupancy distribution $\mathrm{NegBin}(15, 0.15)$ and two levels of the sampling noise $\tau = 0.25$ and $\tau = 0.5$.

5 Application to the Analysis of Array CGH Data

We apply the Beta-GOS model (7)–(8) to the analysis of the array CGH data from Chin et al. (2006) wich we presented in Section 1. More specifically, we consider the raw log2 intensity ratio measurements and seek to identify and cluster clones with similar levels 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 hyper-parameters are set to $\alpha_i = 3$ and $\beta_i = 1$, corresponding to $E(K_n) = 4$ states for large n. We have also considered $\alpha_n = n$ and $\beta_n = 1$, with no remarkable differences in the results. We complete the specification of model (7)–(8) with a vague base distribution, $\mathrm{Normal}(0, 10)$, and a vague inverse gamma distribution for τ centered around $\tau = 0.1$. This choice of τ is motivated by the typical scale of array CGH data and is in accordance with similar choices in the literature (see, for example Guha et al., 2008).

Figure 4 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 1 shows the frequencies of genome copy number gains and losses among all 145 samples plotted as a function of genome location. In order 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

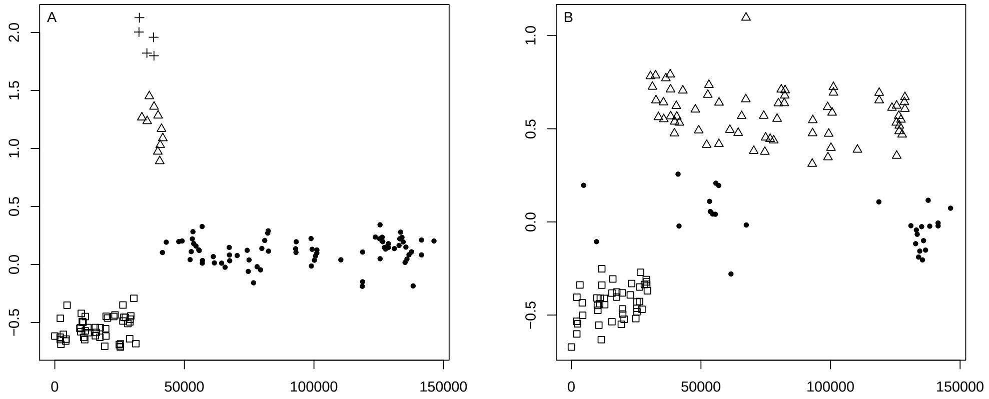

Figure 4: Model fit overview: Array CGH gains and losses on chromosome 8 for two samples of breast tumors in the dataset in (Chin et al., 2006). Points with different shapes denote different clusters.

neutral state. Following Guha et al. (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 $\hat{\mu}_{(1)}$, i.e. if $\hat{\mu}_{(j)} - \hat{\mu}_{(1)} > \varepsilon$. We experimented with choices of ε in the range $[0.05, 0.15]$, but we report here only the results for $\varepsilon = 0.1$. 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 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 the values reported, respectively, at the top and bottom of Figure 1. 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 far away 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 have been shown to correlate to increased gene expression in their analysis. We adapt the procedure

Table 1: 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

described in Newton et al. (2004) to compute a region-based measure of the false discovery rate (FDR) and determine the q-values for the neutral-state and aberration regions estimated from our model. The q-value is the FDR analogue 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, 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 pre-specified 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., 2015. See also Heller et al., 2006, Müller et al., 2007 and Ji et al., 2008). In Table 1 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 also lead to detect chromosomal aberrations in the same locations reported by Chin et al. (2006).

6 Final Remarks

We have discussed a set of generalizations of the predictive rules that charaterize the species sampling mechanism underlying many commonly used

Bayesian Nonparametric priors, such as the Dirichlet process and the Pitman Yor process. Those generalizations allow the clustering of the observations in the sequence to depend on latent random variables or “marks”, which are associated to each observation. Although the resulting sequence is in general not exchangeable, the framework provides a flexible way to model latent and local dependence in the observations.

We illustrated this feature in an application to a study of chromosomal aberrations in breast cancer. Although it’s known that copy number gains and losses are spatially correlated, the extent of such correlation varies along the genome. Homogeneous Hidden Markov models have been widely employed to model copy number data (Guha et al., 2008), but it’s been recognized that such models may not completely capture local dependence in the intensity ratios, which results in location-dependent transition probabilities and corresponding locally varying state persistence properties of the aberrations (DeSantis et al., 2009; Du et al., 2010; Fox et al., 2011). By considering species sampling sequences where the weights are modeled as functions of latent Beta random variables, we have defined a Beta-GOS process prior that provides an alternative Bayesian nonparametric formalism to model heterogeneity and local spatial dependence across observations that are sequentially ordered. In particular, since the Beta-GOS model does not rely on the estimation of a single transition matrix across time points, as in a homogenous 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 assumed in Hidden Semi-Markov models. Indeed, since the predictive weights depend on the sequence of observations itself, the use of such prior appears to be particularly convenient when the underlying generative process is non-stationary, e.g. as a possible alternative to more complicated non-homogeneous HMMs. In addition, our modeling approach enables unsupervised clustering of the observations in an unknown number of states, as it is typical of Bayesian nonparametric priors.

The previous considerations remain valid also for the CID Pitman-Yor sequences we presented in Section 2 and can be extended to other types of conditionally identically distributed sequences characterized by the predictive rule (4). 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 provides an opportunity to flexibly model the complex relationships typical of many heterogenous datasets encountered in biostatistics. For example, the approach may be helpful for modeling individual random effects in longitudinal studies. In functional data analysis, these priors could be used to detect change points in a curve.

Further developments may substitute the latent variable specification with a probit/logistic specification, and define a generalized Pólya Urn scheme that allows the clustering at each observation to be dependent on a set of individual covariates, possibly varying with time.

References

Airoldi, E., T. Costa, F. Bassetti, F. Leisen, and M. Guindani (2014). Generalized Species Sampling Priors With Latent Beta Reinforcements. *Journal of the American Statistical Association 109*, 1466–1480.

Airoldi, E. M., A. Anderson, S. Fienberg, and K. Skinner (2006). Who wrote Ronald Reagan's radio addresses? *Bayesian Anal. 1*, 289–320.

Baladandayuthapani, V., Y. Ji, R. Talluri, L. E. Nieto-Barajas, and J. S. Morris (2010). Bayesian random segmentation models to identify shared copy number aberrations for array cgh data. *Journal of the American Statistical Association 105*(492), 1358–1375.

Bassetti, F., I. Crimaldi, and F. Leisen (2010). Conditionally identically distributed species sampling sequences. *Adv. in Appl. Probab 42*, 433–459.

Berti, P., L. Pratelli, and R. P. (2004). Limit Theorems for a Class of Identically Distributed Random Variables. *Ann. Probab. 32*(3), 2029–2052.

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

Blei, D. and P. Frazier (2011). Distance dependent Chinese restaurant processes. *Journal of Machine Learning Reseach 12*, 2461–2488.

Cardin, N., C. Holmes, T. W. T. C. C. Consortium, P. Donnelly, and J. Marchini (2011). Bayesian hierarchical mixture modeling to assign copy number from a targeted cnv array. *Genetic Epidemiology 35*(6), 536–548.

Chin, K., S. DeVries, J. Fridlyand, P. T. Spellman, R. Roydasgupta, W.-L. Kuo, A. Lapuk, R. M. Neve, Z. Qian, T. Ryder, F. Chen, H. Feiler, T. Tokuyasu, C. Kingsley, S. Dairkee, Z. Meng, K. Chew, D. Pinkel, A. Jain, B. M. Ljung, L. Esserman, D. G. Albertson, F. M. Waldman, and J. W. Gray (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell 10*(6), 529 – 541.

DeSantis, S. M., E. A. Houseman, B. A. Coull, D. N. Louis, G. Mohapatra, and R. A. Betensky (2009). A latent class model with hidden markov dependence for array cgh data. *Biometrics 65*(4), 1296–1305.

Dewar, M., C. Wiggins, and F. Wood (2012). Inference in Hidden Markov Models with Explicit State Duration Distributions. *Signal Processing Letters, IEEE 19*(4), 235–238.

Du, L., M. Chen, J. Lucas, and L. Carlin (2010). Sticky hidden Markov modelling of comparative genomic hybridization. *IEEE TRANSACTIONS ON SIGNAL PROCESSING 58(10)*, 5353–5368.

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

Ferguson, J. D. (1980). Variable duration models for speech. In *Proceedings of the Symposium on the Applications of Hidden Markov Models to Text and Speech*, pp. 143–179.

Fortini, S., L. Ladelli, and E. Regazzini (2000). Exchangeability, predictive distributions and parametric models. *Sankhya 62*(1), 86–109.

Fox, E., E. Sudderth, M. Jordan, and A. Willsky (2011). A sticky HDP-HMM with application to speaker diarization. *Annals of Applied Statistics 5(2A)*, 1020–1056.

Guha, S., Y. Li, and D. Neuberg (2008). Bayesian hidden Markov modelling of array cgh data. *JASA 103*, 485–497.

Guindani, M., P. Müller, and S. Zhang (2009). A Bayesian discovery procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(5), 905–925.

Hansen, B. and J. Pitman (2000). Prediction rules for exchangeable sequences related to species sampling. *Statist. Probab. Lett. 46*(251–256).

Heller, R., D. Stanley, D. Yekutieli, N. Rubin, and Y. Benjamini (2006). Cluster-based analysis of fmri data. *Neuroimage 33*, 599–608.

Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press.

Ishwaran, H. and M. Zarepour (2003). Random probability measures via Pólya sequences: revisiting the Blackwell-MacQueen urn scheme. Technical report, Arxiv.org.

Ji, Y., Y. Lu, and G. Mills (2008). Bayesian models based on test statistics for multiple hypothesis testing problems. *Bioinformatics 24*, 943–949.

Johnson, M. J. and A. S. Willsky (2013). Bayesian nonparametric hidden semi-markov models. *J. Mach. Learn. Res. 14*(1), 673–701.

Kim, S., M. G. Tadesse, and M. Vannucci (2006). Variable selection in clustering via dirichlet process mixture models. *Biometrika 93*(4), 877–893.

Lee, J., P. Müller, Y. Zhu, and Y. Ji (2013). A nonparametric bayesian model for local clustering with application to proteomics. *Journal of the American Statistical Association 108*(503), 775–788.

Lee, J., F. Quintana, P. Müller, and L. Trippa (2008). Defining Predictive Probability Functions for Species Sampling Models. *Statist.Sci. 2*, 209–222.

Lo, A. (1984). On a class of bayesian nonparametric estimates: I density estimates. *Ann. Statist. 12 (1)*, 351–357.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *Proceedings of the Section on Bayesian Statistical Science*.

MacEachern, S. N. and P. Müller (1998). Estimating mixtures of Dirichlet process models. *Journal of Computational and Graphical Statistics 7*, 223–238.

Marioni, J. C., N. P. Thorne, and S. Tavaré (2006). Biohmm: a heterogeneous hidden markov model for segmenting array cgh data. *Bioinformatics 22*(9), 1144–1146.

Mitchell, C., M. Harper, and L. Jamieson (1995). On the complexity of explicit duration hmm's. *Speech and Audio Processing, IEEE Transactions on 3*(3), 213–217.

Müller, P., G. Parmigiani, and K. Rice (2007). FDR and Bayesian multiple comparisons rules. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West (Eds.), *Bayesian Statistics 8*. Oxford, UK: Oxford University Press.

Müller, P. and F. Quintana (2010). Random partition models with regression on covariates. *Journal of Statistical Planning and Inference 140*(10), 2801–2808.

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

Newton, M. A., A. Noueiry, D. Sarkar, and P. Ahlquist (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics 5*, 155—176.

Park, J. and D. Dunson (2010). Bayesian generalized product partition model. *Statistica Sinica 20*(1203–1226).

Pitman, J. (1996). *Some developments of the Blackwell-MacQueen urn scheme*, Volume 30, pp. 245–267. Lecture Notes-Monograph Series, Institute of Mathematical Statistics, Hayward, California.

Pitman, J. (2006). *Combinatorial Stochastic Processes.* Lecture Notes in Mathematics. Ecole d'Eté Probabilités de Saint-Flour XXXII, 2002'.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–286.

Redon, R., T. Fitzgerald, and N. Carter (2009). Comparative genomic hybridization: Dna labeling, hybridization and detection. In M. Dufva (Ed.), *DNA Microarrays for Biomedical Research*, Volume 529 of *Methods in Molecular Biology*, pp. 267–278. Humana Press.

Storey, J. D. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics 31*, 2013–2035.

Storey, J. D. (2007). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics 8*, 414–432.

Sun, W., B. J. Reich, T. Tony Cai, M. Guindani, and A. Schwartzman (2015). False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society Series B 77*, 59–83.

Taramasco, O. and S. Bauer (2012). Rhmm: Hidden markov models simulations and estimations. Technical report, CRAN.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association 101*(476), 1566–1581.

Yau, C., O. Papaspiliopoulos, G. O. Roberts, and C. Holmes (2011). Bayesian non-parametric hidden markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(1), 37–57.

Yu, S.-Z. (2010). Hidden semi-markov models. *Artificial Intelligence 174*(2), 215 – 243. Special Review Issue.