Stat Methods Appl (2018) 27:175–206
https://doi.org/10.1007/s10260-017-0405-z

ORIGINAL PAPER

Nonparametric Bayesian inference in applications

Peter Müeller[1] · **Fernando A. Quintana**[2] · **Garritt Page**[3]

Accepted: 19 September 2017 / Published online: 6 October 2017

Abstract Nonparametric Bayesian (BNP) inference is concerned with inference for infinite dimensional parameters, including unknown distributions, families of distributions, random mean functions and more. Better computational resources and increased use of massive automated or semi-automated data collection makes BNP models more and more common. We briefly review some of the main classes of models, with an emphasis on how they arise from applied research questions, and focus in more depth only on BNP models for spatial inference as a good example of a class of inference problems where BNP models can successfully address limitations of parametric inference.

Keywords Nonparametric inference · Bayesian inference · Dirichlet process · Polya tree

1 Introduction

Nonparametric Bayesian (BNP) inference is concerned with inference for infinite dimensional parameters, including unknown distributions, families of distributions, random mean functions, random partitions, feature allocations and more. One of the first prior probability models for a random distribution in the context of statistical inference is Ferguson's (1973) construction of the Dirichlet process. Since then, and especially over the past 20 years, related literature has exploded, mainly still focused on

✉ Peter Müeller
pmueller@math.utexas.edu

1 University of Texas at Austin, Austin, TX, USA

2 Pontificia Universidad Católica de Chile, Santiago, Chile

3 Brigham Young University, Provo, UT, USA

inference for random distributions and related problems. The construction has proven incredibly useful across many application areas, including in particular bioinformatics, biostatistics, ecology, uncertainty quantification and many more. Many studies in bioinformatics and biostatistics naturally give rise to BNP inference problems due to rapidly increasing data volumes that allow more flexible inference, increasingly more complex study designs and inference problems that often rely on details of a distribution rather than only means, medians or other low dimensional summaries.

In this paper we review some of the related literature by means of discussing some basic inference problems. We introduce and motivate the inference problems with typical examples from the recent literature. The examples include inference that arises from some quite complicated scientific research problems from which we extract relatively simple stylized statistical inference problems. We introduce appropriate BNP models and show how inference under these models can address the desired inference questions. We select only one area, spatial inference problems, as an example for a more extensive discussion and literature review.

A rapidly increasing related literature makes it impossible and futile to attempt a complete literature review of all BNP in the limited scope of this short article. For more extensive reviews of BNP see, for example, Hjort et al. (2010), Walker (2013), Phadia (2013), Müller and Mitra (2013), Ghoshal and Vaart (2017) and Müller et al. (2015).

2 Inference for an unknown distribution

Example 1 (Progel) Xu et al. (2017) construct a clinical trial design for patients who experience air leaks after pulmonary resection. The trial compares standard procedure versus the use of a novel hydrogel sealant (Progel). Progel is hoped to reduce the probability of air leaks and/or accelerate time until leak resolution. Figure 1a shows historical data under control (black and white histogram), together with hypothetical data under Progel (red shaded histogram). Progel is expected to shift the distribution towards shorter resolution times, including some probability that air leaks might not develop at all. Common study designs to compare the two treatments would be based on a comparison of the averages. However, short resolution times are clinically far more preferable than larger times. Therefore a reduction of 1 day is far more important for short resolution times than the same reduction starting from a much larger resolution time. This prompts us to use a weighted average of resolution times as a design criterion, with appropriately higher weights for quick resolution times. For the meaningful use of this weighted mean criterion it is important to use a flexible probability model for resolution times. The model should allow inference about detail features of the distribution, such as more probability mass for short resolution times (even if the mean event time were to remain almost unchanged). The following considerations lead us to use a BNP prior for the probability distribution of resolution times under Progel (F_1) and control (F_0).

Figure 1bcd summarizes inference. Panel (b) shows the two distributions under Progel ($F_1(y)$, red) and under control ($F_0(y)$, black). Panel (c) shows weights $v(y)$ that were elicited from the clinicians to represent relative preferences for different

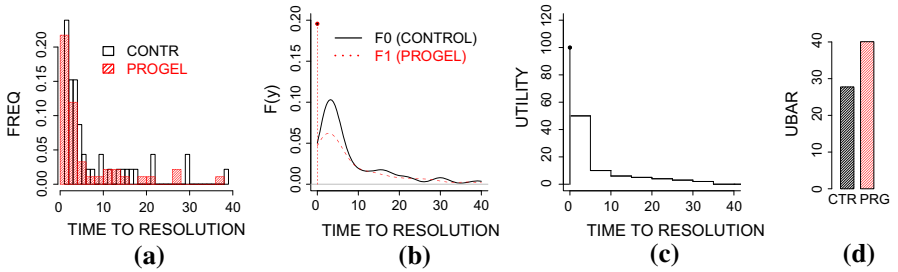

Fig. 1 **a** Resolution times y_{ji} under control (black and white histogram) and under Progel (hypothetical; red). Weighting the random resolution times under each treatment (**b**) with elicited utilities (**c**), we get a weighted average utility $\overline{V}_0$ and $\overline{V}_1$ **d** for each treatment

resolution times. The right panel shows the weighted average, $\overline{V}_j = \int v(y) dF_j(y)$, $j = 0, 1$. Let $\boldsymbol{y}$ denote the observed resolution times across all patients. Xu et al. (2017) propose a clinical trial design that calls for the evaluation of the posterior probability

$$\pi(\boldsymbol{y}) = p(\overline{V}_1 > \overline{V}_0 + \delta \mid \boldsymbol{y}),$$

interpretable as the probability of Progel being at least δ units superior to control in terms of expected weighted mean time to resolution (note that by the decreasing nature of the weights $v(y_i)$ a larger $\overline{V}_1$ implies shorter resolution times and is more desirable). The probability $\pi(\boldsymbol{y})$ is evaluated under a Bayesian inference model. Let y_{ji} denote the resolution time for the i-th patient treated under Progel ($j = 1$) or control ($j = 0$), $i = 1, \ldots, n_j$. The sampling model is straightforward,

$$p(\boldsymbol{y} \mid F_0, F_1) = \prod_j \prod_{i=1}^{n_j} F_j(y_{ji}). \tag{1}$$

To proceed with Bayesian inference we complete the model with a prior for F_j.

Proceeding with Bayesian inference in the previous example requires a prior probability model for the unknown distribution F_j. A common way to proceed is to assume that F_j is a member of some parametric family, for example, a Weibull model, with shape parameter k and scale parameter λ. That is $F_j \in \{p_\theta;\ \theta = (k, \lambda),\ k > 0,\ \lambda > 0\}$. We could then specify a prior for F_j as $p(\lambda, k)$ for $\theta = (\lambda, k)$. Under this prior and the sampling model (1) the posterior probability $\pi(\boldsymbol{y})$ is well defined and easily derived.

In the application to the Progel trial the restriction of F_j to a Weibull family is too restrictive. For example, it would not allow us to learn about an increased probability of $y_{ji} = 0$, that is, immediate resolution of air leaks. Similarly, it only allows to add some probability mass to early resolution times by changing the shape of the distribution across the entire positive line. If instead we wish to treat F_j itself as the unknown quantity, that is, consider F_j as an infinite dimensional unknown parameter, then we need a prior probability model $p(F_j)$ for a random distribution. Prior probability models for infinite dimensional parameters, such as the unknown F_j here, are known as Bayesian nonparametric (BNP) models. One could argue that "massive parametric Bayes" would be a more appropriate term; but "nonparametric Bayes" is traditional,

perhaps because inference tends to appear similar to genuinely nonparametric classical inference.

2.1 Dirichlet process and DP mixture

One of the earliest BNP models for a random distribution F is the Dirichlet process (DP) prior introduced in Ferguson (1973). There are many alternative defining properties that can be used to characterize the DP. Among them is the so called stick-breaking representation due to Sethuraman (1994). Letting δ_x denote a point mass at x we define a random probability measure

$$G = \sum_{h=1}^{\infty} w_h \delta_{m_h} \tag{2}$$

with atoms of probability mass w_h at location m_h. If the locations are an i.i.d. sample from a given base measure $G^\star$ and the weights are generated by i.i.d. beta fractions v_h,

$$m_h \overset{\text{iid}}{\sim} G^\star \quad \text{and} \quad w_h = v_h \prod_{\ell<h}(1-v_\ell) \quad \text{with} \quad v_h \overset{\text{iid}}{\sim} \text{Be}(1,\alpha),$$

then the random probability measure G is said to follow a DP with base measure $G^\star$ and total mass α. We write $G \sim \text{DP}(G^\star, \alpha)$ or, short, $G \sim \text{DP}(\alpha G^\star)$ with an unnormalized base measure $\alpha G^\star$. The construction is known as "stick-breaking" because the prior on the weights w_h can be visualized as iteratively breaking beta-distributed fractions v_h of a stick of initial length 1.0. For later reference we note that implicit in the definition is the almost surely discrete nature of a DP random distribution.

The previous constructive definition of the DP is characterized by two quantities or "parameters", α and $G^\star$. Ferguson's definition of the DP implies that for any measurable set B in the support of $G^\star$ we have that if $G \sim \text{DP}(\alpha G^\star)$ then $G(B) \sim \text{Be}(\alpha G^\star(B), \alpha(1-G^\star(B)))$, and it follows that

$$E(G(B)) = G^\star(B) \qquad \text{and} \qquad \text{Var}(G(B)) = \frac{G^\star(B)(1-G^\star(B))}{1+\alpha}. \tag{3}$$

Because of (3), we see that $G^\star$ serves as a mean measure for any random $G \sim \text{DP}(\alpha G^\star)$, and this is why $G^\star$ is usually referred to as *baseline* or *centering* distribution. On the other hand α is a positive quantity referred to as *total mass parameter*. From (3), α can be also interpreted as a prior precision parameter, controlling the variability of $G(B)$ around $G^\star(B)$.

The DP has been extended in many different ways. See, e.g., the discussion in Blasi et al. (2015). One simple generalization is the two-parameter DP, also known as Pitman-Yor (PY) process (Pitman and Yor 1997), for which the stick-breaking weights are constructed as above but with $v_h \sim \text{Be}(1-\sigma, \alpha+h\sigma)$ for $h \geq 1$. Here, $0 \leq \sigma < 1$ and $\alpha > -\sigma$. The DP is recovered when $\sigma = 0$. Both DP and PY are special cases of stick-breaking processes (Ishwaran and James 2001), which consider $v_h \sim \text{Be}(a_h, b_h)$, with $a_h, b_h > 0$ for all $h \geq 1$ in the stick-breaking construction. Some care is needed to ensure that $P\left(\sum_{h\geq 1} w_h = 1\right) = 1$, which is formally required

for the random probability measure to be well defined. Fortunately, this is easy to check: the condition $\sum_{h\geq 1} \log(1 - E(v_h)) = \sum_{h\geq 1} \log(\frac{b_h}{a_h+b_h}) = -\infty$ is necessary and sufficient for the stick-breaking prior to be well defined.

For many applications, including the Progel trial in Example 1, the use of discrete distributions is awkward. Let $N(y \mid m, \sigma^2)$ indicate a $N(m, \sigma^2)$ random variable y, or, by a slight abuse of notation, a normal kernel in y centered at m with variance σ^2. The DP prior is easily extended to a random continuous distribution by convolution with an additional kernel, say, a normal kernel to define

$$F(y) = \sum_{h=1}^{\infty} w_h N(y \mid m_h, \sigma^2), \tag{4}$$

or $F(y) = \int N(y \mid m, \sigma^2)\, dG(m)$ with a DP prior $G \sim \text{DP}(G^\star, \alpha)$. The model is known as a DP mixture (DPM), an idea originally introduced in Lo (1984). We write $F \sim \text{DPM}(G^\star, \alpha, \varphi)$ for a DPM with a kernel $\varphi(y \mid m)$. And the kernel typically includes some additional hyperparameters, such as the variance σ^2 in the earlier normal mixture of DP. Model (4) is in fact used as prior for F_j in the analysis of the Progel data in Example 1 (Xu et al. 2017).

The DPM model is often written as a hierarchical model. The sampling model $y_i \sim F$ and DPM prior with base measure $G^\star$, total mass α and normal kernel is equivalent to the hierarchical model

$$y_i \mid \mu_i \sim N(\mu_i, \sigma^2), \quad \mu_i \mid G \sim G \quad \text{and} \quad G \sim \text{DP}(G^\star, \alpha), \tag{5}$$

with latent variables $\mu_i \sim G$. The hierarchical model representation is computationally convenient. And it also highlights another important feature of the DPM model. As samples of a discrete random probability measure G the μ_i include many ties. Let $\mu_1^\star, \ldots, \mu_K^\star$ denote the $K \leq n$ unique values and let $s_i = k$ if $\mu_i = \mu_k^\star$. Interpreting $\boldsymbol{s} = (s_1, \ldots, s_n)$ as cluster membership indicators we see how the DPM prior implies a prior on clustering the experimental units. That is, $[n] = \{1, \ldots, n\}$ is partitioned into clusters $S_k = \{i : \mu_i = \mu_k^\star\}$ with $\bigcup_k S_k = [n]$ and $S_k \cap S_{k'} = \emptyset$ for $k \neq k'$. We will return to this interpretation of the DPM as a prior on clustering later, in Sect. 4.

One of the attractions of the DP and DPM prior is easy computation. Posterior inference under (5) can easily be implemented by Markov chain Monte Carlo posterior simulation, exploiting a closed form expression for the posterior distribution on $\boldsymbol{\mu}$ after analytically marginalizing with respect to G. For a summary of related results and more see, for example, the excellent summary by Ghoshal (2010). Inference for many DP-based models is implemented in the R package `DPpackage` (Jara et al. 2011).

2.2 Pólya tree

The DP prior arises as special case of several other more general BNP priors. In particular, the DP is a special case of a Pólya tree (PT) prior (Lavine 1992, 1994). The PT is essentially a random histogram. Without loss of generality we explain it for

a random probability measure F over the unit interval. Splitting the unit interval into $B_0 = [0, \frac{1}{2}]$ and $B_1 = (\frac{1}{2}, 1]$ we define two bins. Let $Y_0 = F(B_0)$ and $Y_1 = F(B_1)$ denote the corresponding probabilities under F. We define a prior for Y_0, Y_1 as $Y_0 \sim \text{Be}(a_0, b_0)$ and $Y_1 = 1 - Y_0$. This defines a (trivial) random histogram F_1 over the two bins defined by B_0 and B_1. Next we refine the random histogram F_1 by splitting B_0 into $B_{00} = [0, \frac{1}{4}]$ and $B_{01} = (\frac{1}{4}, \frac{1}{2}]$ and similarly for $B_1 = B_{10} \cup B_{11}$. Defining $F(B_{e_1 e_2})$ for the four level 2 partitioning subsets $B_{00}, \ldots, B_{11}$ we need to be careful to respect the already generated $F(B_{e_1})$ for the level 1 subsets B_0, B_1. This is easiest achieved by generating random splitting probabilities $Y_{00} = F(B_{00} \mid B_0) \sim \text{Be}(a_{00}, a_{01})$ and $Y_{01} = 1 - Y_{00}$, and similarly for Y_{10}, Y_{11}. We continue like this to define in general

$$Y_{\boldsymbol{e}0} = F(B_{\boldsymbol{e}0} \mid B_{\boldsymbol{e}}) \sim \text{Be}(a_{\boldsymbol{e}0}, a_{\boldsymbol{e}1}) \quad \text{and} \quad Y_{\boldsymbol{e}1} = 1 - Y_{\boldsymbol{e}0} \tag{6}$$

for $\boldsymbol{e} = e_1 \cdots e_m$ being a sequence of m binary indices e_ℓ. The conditional splitting probabilities $Y_{\boldsymbol{e}0}$ imply

$$F(B_{\boldsymbol{e}}) = \prod_{\ell=1}^{m} Y_{e_1 \cdots e_\ell}$$

for any level m subset $B_{\boldsymbol{e}}$. The random probabilities $F(B_{\boldsymbol{e}})$ define a random histogram over a grid defined by the 2^m partitioning subsets $B_{\boldsymbol{e}}$ at level m. This constructive definition of a random probability measure F defines a Pólya tree. It is characterized by a nested sequence of partitions $\Pi = \{\Pi_1, \Pi_2, \ldots\}$ with $\Pi_m = \{B_{\boldsymbol{e}};\ \boldsymbol{e} = e_1 \cdots e_m\}$, and a sequence $\mathscr{A} = \{a_{\boldsymbol{e}}\}$ of beta parameters. We write $F \sim \text{PT}(\Pi, \mathscr{A})$. The DP is a special case of the PT when $a_{\boldsymbol{e}} = a_{\boldsymbol{e}0} + a_{\boldsymbol{e}1}$. In that case the PT prior is invariant with respect to the choice of Π and any nested partition sequence could be used. In general, however, Π as well as $\mathscr{A}$ need to be specified. Some default choices are discussed below.

The main attraction of the PT prior is that, in contrast to the DP prior, it allows to generate continuous distributions. In fact, $a_{\boldsymbol{e}} = cm^2$ for $\boldsymbol{e} = e_1 \cdots e_m$ is a sufficient condition for $F \sim \text{PT}(\Pi, \mathscr{A})$ being a.s. continuous. A PT prior can easily be centered at a given distribution $F^\star$ on the real line by the following construction. Let $Q(u) = F^{\star -1}(u)$ denote the u-quantile of $F^\star$, with $Q(0) = -\infty$ and $Q(1) = \infty$. Let $B_0 = (Q(0), Q(\frac{1}{2})]$, $B_{01} = (Q(0), Q(\frac{1}{4})]$ etc. We write $F \sim \text{PT}(F^\star, \mathscr{A})$. If a is chosen such that $a_{\boldsymbol{e}0} = a_{\boldsymbol{e}1}$, then it is easy to show $E(F(B_{\boldsymbol{e}})) = F^\star(B_{\boldsymbol{e}})$. Alternatively, the same prior centering can be obtained by appropriate choice of $a_{\boldsymbol{e}}$ for any nested partition sequence Π if $a_{\boldsymbol{e}0} = cF^\star(B_{\boldsymbol{e}0})$. We write $F \sim \text{PT}(\Pi, F^\star)$.

Finally, the PT prior is conjugate under i.i.d. sampling. Consider the sampling model $y_i \sim F$, $i = 1, \ldots, n$, together with a PT prior on the unknown distribution, $F \sim \text{PT}(F^\star, \mathscr{A})$. Let $n_{\boldsymbol{e}} = \sum_i I(y_i \in B_{\boldsymbol{e}})$ count the number of data points in subset $B_{\boldsymbol{e}}$. Then $p(F \mid \boldsymbol{y}) = \text{PT}(F^\star, \mathscr{A}^\star)$ is again a PT, with updated PT parameters $a^\star_{\boldsymbol{e}} = a_{\boldsymbol{e}} + n_{\boldsymbol{e}}$

The PT also allows a closed form expression for the marginal distribution of a random sample from F. Berger and Guglielmi (2001) exploit this feature to develop a Bayesian test of fit by testing a parametric model F_θ versus a PT non-parametric

alternative centered around F_θ. The fact that the PT is able to model continuous distributions is critical for this application.

2.3 NRMI

Another generalization of the DP prior is the class of normalized random measures with independent increments (NRMI). The DP can be constructed as a normalized gamma process (Ferguson 1973). The gamma process is a particular example of a much wider class of models known as completely random measures (CRM) (Kingman 1993, chapter 8). Consider any non-intersecting measurable subsets $A_1, \ldots, A_k$ of the desired sample space X. The defining property of a CRM $\widetilde{G}$ is that the random measures $\widetilde{G}(A_1), \ldots, \widetilde{G}(A_k)$ be mutually independent. For example, the gamma process is a CRM with $\widetilde{G}(A_j) \sim \text{Ga}(\alpha G_0(A), 1)$, for a probability measure G_0 and $\alpha > 0$. Normalizing $\widetilde{G}$ by $G(A) = \widetilde{G}(A)/\widetilde{G}(X)$ defines a $\text{DP}(\alpha G_0)$ prior.

Replacing the gamma process by any other CRM defines the class of NRMI priors (Regazzini et al. 2003). A recent review of NRMI's appears in Lijoi and Prünster (2010). Besides the DP prior, another interesting example is the normalized generalized gamma process (NGG), discussed in Lijoi et al. (2007). We write $G \sim \text{NGG}(\alpha, \kappa, \gamma, G_0)$. The NGG is indexed by a total mass parameter $\alpha > 0$, two more scalar parameters $\kappa \geq 0$ and $\gamma \in [0, 1)$ and a base probability measure G_0. In fact, the DP is a special case of the NGG with $\kappa = 1$ and $\gamma = 0$.

Like for any CRM, a realization from the generalized gamma process (before normalization) can be generated using the following constructive definition based on a Poisson process over $\Re^+ \times X$, with Poisson intensity $\nu(\tilde{w}, m)$ (Kingman 1993; section 8.2). The arguments of $\nu(\cdot)$ are already labeled in anticipation of the following construction. The choice of $\nu(\cdot)$ determines different CRM's. For the generalized gamma process use

$$\nu(\tilde{w}, m) = \rho(\tilde{w})\, \alpha G_0(m) \text{ with } \rho(\tilde{w}) = e^{-\kappa \tilde{w}} \tilde{w}^{-(1+\gamma)} / \Gamma(1-\gamma).$$

Let $(\tilde{w}_h, m_h)$, $h = 1, \ldots,$ denote a realization of a Poisson process with intensity $\nu(\cdot, \cdot)$. Then $\widetilde{G} = \sum \tilde{w}_h \delta_{m_h}$ is a realization of the CRM and the normalized random probability measure $G \propto \widetilde{G}$ is a realization of the NGG, $G \sim \text{NGG}(\alpha, \kappa, \gamma, G_0)$. In summary, we define

$$\widetilde{G} = \sum \tilde{w}_h \delta_{m_h} \quad \text{and} \quad G = \sum w_h \delta_{m_h} \text{ with } w_h = \tilde{w}_h / \widetilde{G}(X). \tag{7}$$

Similar to the DPM model, mixtures with respect to an NGG prior are useful as priors for continuous random probability measures (Barrios et al. 2013).

$$y_i \mid \boldsymbol{\theta} \sim p(y_i \mid \theta_i), \quad \theta_i \mid G \sim G, \quad G \sim \text{NGG}(\alpha, \kappa, \gamma, G_0). \tag{8}$$

The discussion in Barrios et al. (2013) is more general, allowing for any other NRMI, but the NGG is a sufficiently rich model for most purposes. However, in comparison with the DP prior the additional flexibility of the NGG is important for modeling

(Blasi et al. 2015). As pointed out in Argiento et al. (2016), the γ parameter provides a richer reinforcement mechanism in the predictive distribution of the sample, and better control on the implied distribution of the number of clusters K. In fact, γ can be used to tune the variance of K, which in the DP case depends only on α.

Most importantly, posterior inference under NRMI mixtures is still easily implemented. Barrios et al. (2013) as well as Favaro and Teh (2013) and Argiento et al. (2010) outline specific MCMC algorithms. Both are based on a representation of the posterior distribution for NRMI's discussed in James et al. (2009), under independent sampling, $\theta_i \sim G$ as in (8) and an NRMI prior for G. And most conveniently, the R package `BNPdensity` implements inference for NRMI mixtures as described in Barrios et al. (2013).

2.4 Survival analysis

Many applications and motivating problems for BNP priors of random distributions occur in the context of survival analysis. Hanson and Jara (2013) present a review focused on DPs and PT's and their dependent extensions, representing the most commonly used BNP models for event time data. But many other BNP priors beyond the DP and PT have been proposed, including the neutral to the right processes of Ferguson and Phadia (1979), the extended Gamma processes of Dykstra and Laud (1981), the Beta processes of Hjort (1990), the beta-Stacy processes of Walker and Muliere (1997), the Lévy driven Markov processes of Nieto-Barajas and Walker (2002) and many others. See further review and discussion in Müller and Quintana (2004).

Perhaps the main motivation for the use of BNP priors for time-to-event distributions is the fact that for important event times one is naturally interested in details of the unknown distribution beyond summaries like mean and median. The unknown distribution of resolution times in Example 1 is a typical example. The details matter. In particular, it is desirable to allow additional probability mass for quick resolution times. Another example is inference for semicompeting risks in the following example.

Example 2 (Semicompeting risks) Xu et al. (2016) consider inference for semicompeting risks in a study of patients with recurrent malignant brain tumors. The study records progression times P_i and overall survival D_i for patients $i = 1, \ldots, n$. The term "semicompeting risks" refers to the implied asymmetric censoring. Death censors progression, i.e., $P_i > D_i$ can not be observed, but not vice versa. Patients are randomized to one of two treatments, $z_i \in \{0, 1\}$, and the aim is to compare treatments on the basis of time to progression. A meaningful comparison has to adjust for the competing risk of death, lest the more letal treatment might seem favorable with long progression times. Let (P_i^0, D_i^0) and (P_i^1, D_i^1) denote hypothetical outcomes under the two treatments. They are "hypothetical" since for each patient only one of the pairs is observed, $P_i = P_i^{z_i}$ and $D_i = D_i^{z_i}$. The outcomes under $(1 - z_i)$ remain hypothetical or "counterfactual". In addition, death censors progression. We set up a comparison using the odds for progression conditional on a patient being alive under both treatments. Dropping the $_i$ index for the moment, we have

$$h(t) = \frac{p(P^1 \le t \mid D^1 \ge t, D^0 \ge t)}{p(P^0 \le t \mid D^1 \ge t, D^0 \ge t)}. \tag{9}$$

Bayesian inference for the odds $h(t)$ requires a prior probability model for the unknown joint distribution of (P^0, D^0, P^1, D^1). We implement this as a bivariate BNP model for $F_0(P^0, D^0)$ and another one for $F_1(P^1, D^1)$. By a slight abuse of notation we will use $F_z(\cdot)$ to generically denote probabilities and marginal distributions under F_z, with the arguments clarifying the specific use. Let $\Phi(\cdot)$ denote a standard normal c.d.f. The two models are stitched together by a Gaussian copula for (D^0, D^1). Letting $A_z = \Phi^{-1}\{F_z(D^z)\}$, we assume $(A_0, A_1) \sim N(0, \rho I)$, implying $p(A_0 \mid A_1) = N(\rho A_1, 1 - \rho^2)$, and similarly for $p(A_1 \mid A_0)$. We can use the assumption to augment $F_1(P^1, D^1)$ to $F_1^\star(P^1, D^1, A_0) = F_1(P^1, D^1)p(A_0 \mid A_1 = \Phi^{-1}F_1(D^1))$ and similarly for $F_0^\star(\cdot)$.

For fixed ρ, the inference target $h(t)$ then becomes a functional of (F_0, F_1), as

$$h(t;\ F_0, F_1) = \frac{F_1^\star(P^1 \le t, D^1 \ge t, A_0 \ge F_0(D^0))}{F_0^\star(P^0 \le t, D^0 \ge t, A_1 \ge F_1(D^1))}.$$

Under a prior probability model for F_z, $z = 0, 1$, and data (Z_i, P_i, D_i), $i = 1, \ldots, n$ (for the moment ignoring censoring), we obtain a joint posterior distribution for F_0 and F_1 and can report posterior inference on $h(t;\ F_0, F_1)$ (for fixed ρ).

Inference in the above application hinges then on a flexible prior for $F_z(P^z, D^z)$. Note that the earlier discussion of BNP priors for random probability measures was never restricted to scalar random variables. In fact, Xu et al. (2016) assume DPM of normal priors for F_z, independently across $z = 0, 1$. Let $\widetilde{F}_z(P^z \le t, D^z \le s, P^z \le D^z)$, $t \le s$, denote the subdistribution implied by F_z, that is, the part of the joint distribution that is indeed observable (progression can only be observed before death). The evaluation of the functional $h(t;\ F_0, F_1)$ requires only $\widetilde{F}$, making it identifiable. The bivariate normal copula is an elegant and easy way to link the two bivariate distributions without changing anything in the marginal prior for each, F_0 and F_1. The data provides no information on this link. We therefore have to fix it by selecting a specific ρ. Figure 2 shows inference on $h(t)$ in three simulations, using different values of ρ. In all three simulations we generated data for $n = 200$ hypothetical patients under an assumed simulation truth. Inference under the BNP model recovers the simulation truth. The flexible BNP model is critical to adapt to the diverse simulation truths in the different simulations.

3 Regression

One of the most basic statistical inference problems is regression. In its most general form regression can be described as sampling from a family of probability measures indexed by covariates. That is, let $\mathscr{F} = \{F_x;\ x \in X\}$ denote a family of probability

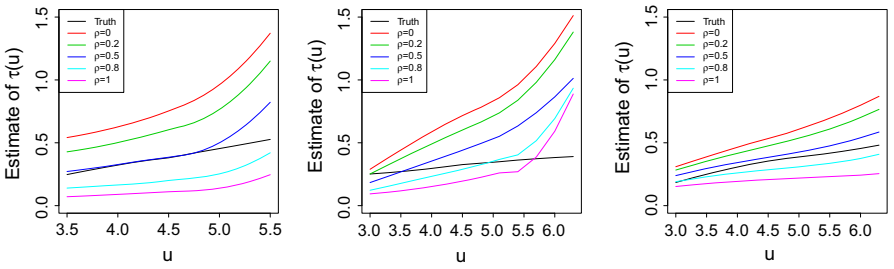

Fig. 2 Posterior inference on $h(t)$ versus u in three simulations using $\rho = 0.3, 0.5, 0.8$. The black, red, and blue lines represent the medians when $\rho = 0.3, 0.5, 0.8$, respectively, the green lines represent the simulation truths, the dashed lines represent 25% quantiles and 75% quantiles

measures indexed by covariates x. Regression of an outcome y_i on covariates x_i is then stated as

$$y_i \mid x_i = x, \mathscr{F} \sim F_x.$$

Proceeding with Bayesian inference we complete the sampling model with a prior on the unknown quantities, in this case $\mathscr{F}$. If $F_x = N(\beta' x, \sigma^2)$ the problem reduces to a normal linear regression problem, using a family $\mathscr{F}$ indexed by the regression coefficients β (assuming for the moment known residual variance σ^2).

3.1 Partially nonparametric regression

Abandoning the restriction to a parametric family $\mathscr{F}$ we end up with BNP regression. Depending on the level of generalization we naturally get three cases of BNP regression: (i) nonparametric residual distribution; (ii) nonparametric mean function; and (iii) fully nonparametric regression. Under the first case we assume

$$y_i = \beta' x + \epsilon_i \quad \text{with} \quad \epsilon_i \sim F$$

for an unknown residual distribution F. The model is completed with a BNP prior on F. To maintain the interpretation of ϵ_i as residuals it is useful to center F at zero. A version of this construction is developed in Walker and Mallick (1999) who propose a PT prior model for a residual distribution (in the context of a regression of log event times y_i on risk factors x_i). A limitation of the model is the dependence of inference on the partition boundaries from the PT construction. Hanson and Johnson (2002) address this problem by using mixture of PT priors, still centered at median 0. Alternative constructions are introduced in Schörgendorfer et al. (2013) for dichotomized continuous outcomes, in Hanson and Johnson (2004) using a DP prior, and in Kottas and Gelfand (2001) using a DP scale mixture of uniforms to represent unimodal residual distribution F.

Under the second case, that is, nonparametric mean regression, we assume

$$y_i = f(x_i) + \epsilon_i \quad \text{with} \quad \epsilon_i \sim N(0, \sigma^2)$$

and unknown f. The model is completed with a BNP prior on the mean function f. Widely used priors expand the function f in terms of some basis, as $f(x) =$

$\sum_k c_k \psi_k(x)$, and then proceed with a prior on the coefficients c_k with respect to this basis. A typical example are wavelet bases, that is, an orthonormal function basis based on shifted and scaled versions ψ_{jk} of an underlying wavelet function $\psi(\cdot)$ (Chipman et al. 1997; Vidakovic 1998; Clyde and George 2000). Another example are splines (Fahrmeir et al. 2004; Baladandayuthapani et al. 2005). Nonparametric regression with splines is implemented, for example, in the R package `BayesX` (Brezger et al. 2005). Another widely used class of nonparametric priors for a mean function $f(x)$ are Gaussian process (GP) priors. A GP specifies a prior on f by assuming a multivariate normal for f evaluated at any finite set of covariate values x_i,

$$(f(x_1), \ldots, f(x_n)) \sim N(0, S)$$

with the (i, j) element of S given by a covariance function $C(x_i, x_j)$. The zero mean could be replaced by a non-zero fixed mean function $\mu(x)$ evaluated at $x_1, \ldots, x_n$. We write $f \sim \text{GP}(\mu(\cdot), C(\cdot, \cdot))$. Bayesian inference under GP priors can be very computationally intensive, essentially due to the $(n \times n)$ covariance matrix S, with typically all non-zero correlation. Many properties and applications of GPs can be found in Rasmussen and Williams (2006), including various computational approaches. One particular solution to the computational complexity problem is proposed by Gramacy and Lee (2008) who develop treed GP priors which avoid high-dimensional matrix factorization by partitioning the covariate space. The approach is implemented in the R package `tgp`.

Finally, under the third case, that is, fully nonparametric regression, let $\mathscr{F} = \{F_x, \ x \in X\}$ denote a family of random probability measures F_x, indexed by x in some covariate space X. We assume

$$y_i \mid x_i, \mathscr{F} \sim F_{x_i}.$$

The model is completed with a BNP prior on the family $\mathscr{F}$. In the next section we discuss some examples for this case in more details.

3.2 Fully nonparametric regression

Example 3 (Survival regression) Xu et al. (2016) discuss Bayesian inference for dynamic treatment regimes (DTR), that is, a treatment that proceeds in stages. The particular example is inference for data from a leukemia study. The first decision is the assignment of the initial induction therapy (A). In response to A the tumor can develop resistance, show complete response, or the patient can die before either. In the case of resistance the treating physician decides on a salvage therapy (S1). In the case of complete response, the next state could either be death or progression. In the latter case the physician decides again on (a different) salvage therapy (S2). The scheme is illustrated as a flow diagram in Fig. 3a. Details of the scheme are not important for the upcoming discussion beyond the notion that it is a sequence of decisions (A,S1, S2) and that it involves several transition times (time to resistance T^R, time to complete response T^C, time from complete response to progression T^{CP}, etc.). The problem is

that the initial treatment assignment (A) is randomized, but the later salvage treatment choices (S1 and S2) are not. This lack of randomization complicates a comparison of alternative treatment regimes. For example, if physicians were to always choose a particular S1 for patients with poor prognosis, one would be mislead to underestimate the effect of this S1 treatment.

To allow for a meaningful comparison we use statistical post-processing to adjust for this lack of randomization. Treatments are compared on the basis of overall survival, that is, the sum of all transition times starting with the initial treatment assignment until final death. Transition times include T^R, T^D, T^{RD} etc., as indicated in the figure. Overall survival is the sum of all transition times that a patient goes through. If we had a good model for each transition time, then we could compare treatments on the basis of predicted mean overall survival under alternative treatment assignments. The model for each transition time would be a survival regression based on patient baseline covariates as well as earlier transition times. The model would allow us to impute (hypothetical) overall survival under all possible treatment assignments. Of course, only one treatment assignment is observed, but we could carry out a comparison of treatment plans on the basis of predicted mean survival under all possible treatment plans. This implements the desired model-based adjustment for the lack of randomization by replacing a comparison of treatment specific average outcomes, which suffices under randomization, by a comparison of predicted mean overall survival based on the model. The approach can be characterized as model-based causal inference. The main limitation of the approach is that the models for the transition times need to be very flexible, lest inference will be hopelessly biased by parametric assumptions and extrapolation. This is where BNP comes into play.

3.2.1 The dependent DP (DDP)

Xu et al. (2016) implement the required BNP survival regression for the transition times T^R, T^D, T^C etc. and carry out the comparison based on inference under these BNP survival regressions. In a simulation study it turns out that this BNP model-based adjustment for lack of randomization performs well and compares favorably with currently used frequentist methods. The essential reason is that BNP models are "always right," in the sense of allocating prior probability in a neighborhood of any true model.

Let $y_i = \log(T_i)$ denote one of the transition times (on a log scale) for the i-th patient in the example. The model in Xu et al. (2016) implies for each patient a DPM model for $p(y_i)$. That is, $y_i \sim F = \sum_h w_h N(m_h, \sigma^2)$ with a DP prior on the mixing measure, as before. Next let x_i denote covariates including patient baseline covariates and the current history (prior transition times for the same patient). We introduce a regression on x_i by replacing m_h with a stochastic process $m_h(x_i)$ indexed by x_i. In particular, we use a GP prior,

$$\{m_h(x)\}_x \sim \text{GP}(0, C(\cdot, \cdot)).$$

independent across h where $C(\cdot, \cdot)$ denotes a covariance function. In the application we use the widely adopted (Williams 1997) squared exponential $C(x_i, x_j) =$

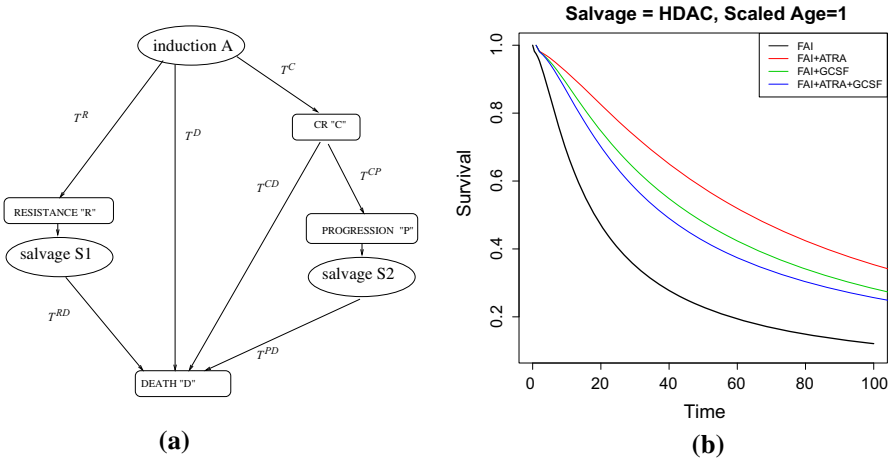

Fig. 3 **a** A flow diagram of the DTR. **b** The survival regression for the transition times T^{PD} (from progression to death) as estimated survival functions for different covariates x_i. The panel shows the posterior estimated survival functions of patient at scaled age 1 and poor cytogenetic abnormality assigned to a particular salvage therapy S_1, S_2 and for four induction therapies A respectively. Black, red, green and blue curves represent the four different choices for A

$\exp\{-\sum_{m=1}^{M}(x_{im}-x_{jm})^2\}+J^2$. This construction replaces F by F_x with covariate-specific location parameters $m_h(x)$. The marginal prior for a particular F_x is still a DPM prior. In particular, the $m_h(x)$ are independent across h. However, we introduced dependence across x by assuming the GP prior for $m_h(x)$ across x. Figure 3 shows the survival regression for one of the transition times in the previous example.

In summary we have defined a prior for $\mathscr{F} = \{F_x,\ x \in X\}$ that is such that $F_x \sim$ DPM, marginally, but F_x are dependent via the dependence of the point masses $m_h(x)$ across x. This construction is known as the dependent DP (DDP). The DDP was first introduced by MacEachern (1999). Since then a vast literature has appeared related to many extensions and applications of the basic model. The earlier example is one of a huge number of successful applications of this scheme. A good recent review appears in Foti and Williamson (2015), including related models that generalize the DDP.

The construction of dependent F_x, $x \in X$, that we just introduced built on the stick-breaking representation (2). We introduced a specific variation by making the locations $m_h(x)$ dependent across x. In general, the dependence could be introduced on the weights v_h and/or the locations m_h.

3.2.2 Dependent NRMI's

An entirely alternative approach starts with the representation (7) of the DP or any NRMI as a normalized CRM (Lijoi et al. 2014). The basic idea is simple. The CRM $\widetilde{G}$ can be pieced together as a sum of two independent CRM's, $\widetilde{G} = \widetilde{G}_1 + \widetilde{G}_0$. Consider another CRM $\widetilde{H}$, decomposed as $\widetilde{H} = \widetilde{G}_2 + \widetilde{G}_0$. By sharing the common component $\widetilde{G}_0$ the two CRM's and thus the corresponding NRMI's $G \propto \widetilde{G}$ and $H \propto \widetilde{H}$ are dependent. Finally, $\widetilde{G}_j$, $j = 0, 1, 2$ can be chosen such that $\widetilde{G}$ is any desired CRM.

Recall that a CRM $\widetilde{G}$ can be generated by a Poisson process on $\Re^+ \times X$ with intensity $\nu(\tilde{w}, m) = \rho(\tilde{w})\alpha G_0(m)$. Defining Poisson processes with intensities ν_0 and ν_1 such that $\nu_0 + \nu_1 = \nu_0 + \nu_2 = \nu$ achieves the desired decomposition of CRM $\widetilde{G}$ and $\widetilde{H}$ such that they share a common term that induces dependence and such that $\widetilde{G}$ and $\widetilde{H}$ follow marginally any desired CRM. Camerlenghi (2015) further develops the notion of introducing dependence across NRMI's based on a suitable construction of dependent CRM's. In addition to the additive construction of Lijoi et al. (2014) he introduces two more methods based on elegant generalizations of the hierarchical DP (HDP) of Teh et al. (2006) and a second one based on the nested DP (NDP) of Rodríguez et al. (2008) to arbitrary CRM's. It would be interesting to compare the relative strengths of this construction of dependent random measures versus the DDP construction that we introduced previously. To our knowledge no such systematic comparison exist.

4 Partitions, feature allocations and trait allocations

4.1 Random partitions

Example 4 (Nested clustering) Lee et al. (2013) analyze a data set with expression levels of $G = 55$ proteins that were selected from two cell signaling pathways (PI3K and MAPK) for $n = 256$ samples from breast cancer patients. Breast cancer is known to be a very heterogeneous disease. At least three subtypes of breast cancer are distinguished, known as HR+, HER2+ and TN. The classification into these three subtypes is traditional and widely used, but also known to have limitations. The idea of this study is to cluster patients on the basis of these proteins and explore how results might refine and revise the traditional classification.

The main inference target is a partition of the n samples into more homogeneous subsets. The notion of subsets of proteins corresponding to different biologic functions motivates us to expect more than one way of clustering patients, depending on which subset of proteins one focuses on. We thus refine the inference target to finding first a partition of the G proteins $[G] = \{1, \ldots, G\}$ into subsets $S_1, \ldots, S_K$, and then, nested within each protein cluster S_k, a nested partition of samples $[n] = \{1, \ldots, n\}$ into $R_{k1}, \ldots, R_{kL_k}$. This explicitly allows for a different clustering of samples with respect to different biologic processes (as represented by the protein clusters)

Proceeding with Bayesian inference in Example 4 requires a prior probability model for partitions and nested partitions. We first focus on the earlier only. Let $\rho_G = \{S_1, \ldots, S_K\}$ denote a partition of $[G] = \{1, \ldots, G\}$. That is, $S_k \subseteq [G]$, $S_k \cap S_j = \emptyset$ for $k \neq j$ and $\bigcup S_k = [G]$. A partition can alternatively be recorded as a set of cluster membership indicators s_g with $s_g = k$ if $g \in S_k$. The problem then becomes to define a prior probability model $p(\boldsymbol{s})$ for $\boldsymbol{s} = (s_1, \ldots, s_G)$.

Recall the DPM model (5) and the notation for unique values $\mu_k^\star$ and cluster membership indicators s_i. Using $\boldsymbol{\mu}^\star$ and $\boldsymbol{s}$ the DPM model can equivalently be written as

$$y_i \mid s_i = k, \mu_k^\star \sim N(\mu_k^\star, \sigma^2), \qquad \mu_k^\star \mid \boldsymbol{s} \sim G^\star,$$

independently across $i = 1, \ldots, n$ and $k = 1, \ldots, K$. The dependence on $\boldsymbol{s}$ in the last distribution is only indirectly through the number K of clusters or unique values. Let $n_k = |S_k|$ denote the size of the k-th cluster, that is, the number of $\mu_i = \mu_k^\star$. Also, we number clusters by appearance, that is, $s_1 = 1$ and $s_i \leq \max\{s_1, \ldots, s_{i-1}\} + 1$. The DP prior on G in (5) implies

$$p_{DP}(\boldsymbol{s} \mid \alpha) = \frac{\alpha^K \left\{\prod_{k=1}^K (n_k - 1)!\right\}}{\left\{\prod_i (\alpha + i - 1)\right\}}. \tag{10}$$

The prior $p(\boldsymbol{s})$ is known as the Pólya urn. The formula in (10) is also related to the Ewens sampling formula (Ewens 1972). Following the recent discussion in Crane (2016), the formula was derived in the context of allele frequencies in a certain population. Each gene has a certain allelic type and any given sample of size n of genes can be represented by its allelic partition $(m_1, \ldots, m_n)$, where m_j is the number of alleles appearing exactly j times, for $j = 1, \ldots, n$. The Ewens sampling formula with parameter $\alpha > 0$ establishes that

$$p(m_1, \ldots, m_n \mid \alpha) = \frac{n!}{\prod_{j=1}^n (\alpha + j - 1)} \prod_{k=1}^n \frac{\alpha^{n_k}}{k^{m_k}\, m_k!}, \tag{11}$$

for any collection of nonnegative integers $m_1, \ldots, m_n$ such that $\sum_{j=1}^n j \cdot m_j = n$. Labeling now the genes in the sample as $1, \ldots, n$, and assigning each an allelic type, the allelic partition can be alternatively represented as a partition into subsets where all genes in a given subset are of the same allelic type. A probability distribution on partitions can then be generated by first drawing an allelic partition $(m_1, \ldots, m_n)$ according to (11) and then randomly selecting one partition out of the collection of partitions that share the same allelic distribution. The result of this is exactly (10). See further discussion in Crane (2016).

The closed form expression for $p(\boldsymbol{s})$ in the right-hand side of (10) is very useful for the implementation of posterior MCMC simulation for DPM models. However, here we use the random partition $p_{DP}(\boldsymbol{s} \mid \alpha)$ itself as prior for the partition of proteins $\{1, \ldots, G\}$ in Example 4 (without any reference to a random distribution G or samples μ_i). The use of $p_{DP}(\cdot)$ as a prior for random partitions is a very common model choice, including in Bayesian biostatistics and bioinformatics.

In Example 4 we require a second, nested, random partition model for clustering patients with respect to every protein cluster. Denote with $\{R_{k1}, \ldots, R_{kL_k}\}$ a partition of $[n]$ that describes the clustering of samples with respect to proteins in cluster S_k. Alternatively we can again use cluster membership indicators r_{ki} with $r_{ki} = \ell$ if $i \in R_{k\ell}$. As prior probability model for $\boldsymbol{r}$ we use again the same p_{DP} prior. That is $\boldsymbol{r}_k = (r_{k1}, \ldots, r_{kn}) \sim p_{PD}(\boldsymbol{r}_k \mid \beta)$. Conditional on $\boldsymbol{s}$ and $\boldsymbol{r}_k$, $k = 1, \ldots, K$, we then introduce protein and cluster-specific parameters $\boldsymbol{\theta}_{g\ell}$ that index the sampling model for the recorded protein activations for proteins g and samples in (the nested) cluster $R_{k\ell}$. Inference under this model is shown in Fig. 4. Importantly, sampling model parameters vary by proteins. Protein clusters are defined by the shared nested partition

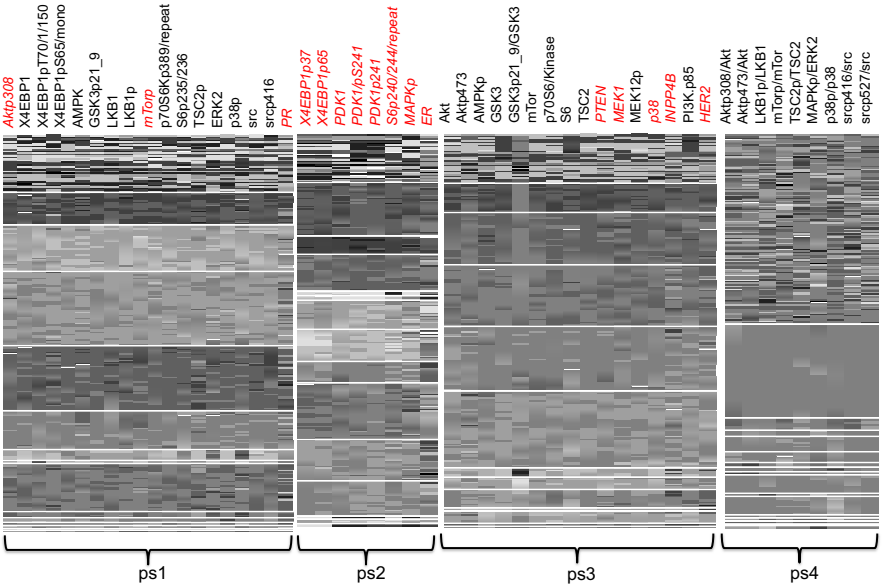

Fig. 4 Inference for the nested clustering of protein activation data. Heatmaps of four identified protein sets (ps) by the NoB-LoC model. Notation "ps" refers to "protein set". The samples are rearranged according to the estimated cluster memberships within each of active protein sets 1, 2, 3 and 4. White horizontal lines show the sample clusters. The first block of top rows for each active protein set are the inactive samples for that protein set. The protein names in red are specifically discussed in Lee et al. (2013)

$\boldsymbol{r}_k$, not by shared parameters, implying for example, that one protein $g_1 \in S_k$ might be deactivated for all samples in $R_{k\ell}$, while another protein $g_2 \in S_k$ in the same protein cluster might be activated. The important feature is that activation and de-activation are synchronized to apply for the same subset of samples. Lee et al. (2013) refer to this model as the "NoB-loc model" (non-parametric Bayesian local clustering).

Related nested partitions are generated by the nested DP (Rodríguez et al. 2008) and a similar construction in the enriched DP (Wade et al. 2011). However, in both constructions the nested clustering of samples would be protein specific, say $\boldsymbol{r}_g$, and all proteins $g \in S_k$ in the same protein cluster share a common *prior* $p(\boldsymbol{r}_g) = G_k$ (rather than a common nested partition $\boldsymbol{r}_k$). Rodríguez and Ghosh (2012) define a very similar nested partition. But clusters are defined by common protein cluster-specific sampling model parameters, rather than by shared nested partitions.

4.2 Random feature allocation

A random partition of $\{1, \ldots, n\}$ can be represented as a binary $(n \times K)$ matrix $\boldsymbol{Z}$ with $Z_{ik} = 1$ if $s_i = k$. The number of columns, K, is the random size of the partition and the matrix $\boldsymbol{Z}$ is constrained to have row sums equal 1, and non-zero columns (an all 0 column would correspond to an empty cluster). A prior $p(\boldsymbol{s})$ implies a prior $p(\boldsymbol{Z})$ on the random binary matrix $\boldsymbol{Z}$, including the random number of columns and subject to the unit row constraint. Some problems call for a generalization to random binary matrices $\boldsymbol{Z}$ without the unit row constraint. Interpreting $Z_{ik} = 1$ as an indicator for

unit i being a member in subset S_k this corresponds to a model for i being a member in a random number of subsets, allowing membership in multiple (or no) subsets. For technical reasons we restrict to membership in finitely many subsets. A prior $\boldsymbol{Z}$ for such binary matrices with a random number K of columns is known as feature allocation model. The random matrix $\boldsymbol{Z}$ is subject to no non-zero columns.

Example 5 (Tumor heterogeneity) Lee et al. (2015) discuss inference for tumor heterogeneity. For a fixed set of S genomic loci (that is, particular DNA base pairs) the data records the count N_i of reads from a solid tumor sample that are mapped to locus i and the count y_i of reads out of these N_i that carry a variant allele (relative to some reference genome). Let $\widehat{p}_i = y_i/N_i$ denote the variant allele fraction (VAF) at locus i. For a homogeneous cell population $\widehat{p}_i$ would be either 0 or 1. For simplicity we assume in this description that a read must be either reference or variant allele, ignoring the fact that multiple variant alleles are possible and ignoring the diploid nature of human data. However, for tumor data the VAF's are usually any fractions, $0 \leq \widehat{p}_i \leq 1$. This can be explained by the tumor being composed of a mix of underlying homogeneous cell subpopulations ("subclones"), $k = 1, \ldots, K$. The problem then becomes (i) to characterize each subclone in terms of a set of loci that carry the variant allele (SNV, "single nucleotide variation"); (ii) to represent the observed heterogeneous cell population as a mixture of these K hypothetical homogeneous subpopulations. We use a random binary $(n \times K)$ matrix $\boldsymbol{Z}$ with a random number of columns and all non-zero columns to identify the subclones, with $Z_{ik} = 1$ indicating that subclone k has a variant allele at locus i. That is, each column of $\boldsymbol{Z}$ characterizes one of the (random number) K subclones. Next we introduce a vector $\boldsymbol{w} = (w_1, \ldots, w_K)$ of weights that represent the (unknown) fraction of subclone k in the sample. For a given $\boldsymbol{Z}$ and $\boldsymbol{w}$ the relative frequency of a variant allele in the sample at locus i should then be $p_i = \sum_{k=1}^{K} w_k Z_{ik}$. That is all. We are now ready to write down a statistical inference model for tumor heterogeneity. Let $\boldsymbol{\theta} = (\boldsymbol{Z}, \boldsymbol{w})$ denote the unknown parameters. We assume a binomial sampling model for the observed VAF's, $y_i \sim \text{Bin}(N_i, p_i)$ with $p_i = \sum_{k=1}^{K} w_k Z_{ik}$, and thus a joint likelihood

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) \propto \prod_i p_i^{y_i} (1 - p_i)^{N_i - y_i} \tag{12}$$

as a likelihood for $\boldsymbol{\theta}$. The model is completed with a prior for $\boldsymbol{w}$ and $\boldsymbol{Z}$. We assume a symmetric Dirichlet distribution, $\boldsymbol{w} \sim \text{Dir}(a, \ldots, a)$, leaving only the prior $\boldsymbol{Z}$ to be determined.

The random binary matrix $\boldsymbol{Z}$ represents a feature allocation. We complete model (12) with a feature allocation prior on $\boldsymbol{Z}$. The most commonly used feature allocation model $p(Z)$ is the *Indian Buffet Process* (IBP) (Griffiths and Ghahramani 2006). Let n_k denote the column sum in the k-th column, let K_i denote the i-th row sum, and let $h_n = \sum_{i=1}^{n} 1/i$. A random $(n \times K)$ binary matrix $\boldsymbol{Z}$ is said to follow an Indian buffet process (IBP) if

$$p(\boldsymbol{Z}) = \frac{\alpha^K}{\prod_{i=1}^{n} K_i^+!} e^{-\alpha h_n} \prod_{k=1}^{K} \frac{(n - n_k)!(n_k - 1)!}{n!}.$$

The model includes a random number of columns K and $\boldsymbol{Z}$ is constrained to indexing features by appearance. That is, if i_k is the first row with $Z_{ik} = 1$, the rows are ordered by $i_k \leq i_{k+1}$. The name arises from a generative model that uses a metaphor with rows i corresponding to customers entering an Indian buffet restaurant and columns k indexing dishes. Let $m_{ik} = \sum_{\ell=1}^{i} Z_{\ell k}$ denote the number of customers prior to customer $i+1$ who picked dish k. The probability of customer $i+1$ selecting one of the earlier dishes again is proportional to m_{ik}.

Similar to the DP there is also a stick-breaking construction (Teh et al. 2007):

$$Z_{ik} \mid w_k \overset{\text{ind}}{\sim} \text{Ber}(w_k), \quad w_k = \prod_{j=1}^{k} v_j, \quad v_j \mid \alpha \overset{\text{iid}}{\sim} \text{Be}(\alpha, 1), \tag{13}$$

$k = 1, 2, \ldots$. The binary matrix Z is obtained after removing all columns that contain only zeros, which leaves us with a random $n \times K_n$ matrix with a random number of columns K_n. And if desired, the columns of $\boldsymbol{Z}$ can be rearranged for an order constraint. Yet another alternative representation of the IBP is in terms of CRM's. Thibaux and Jordan (2007) construct the IBP based on a beta process (BP) and a Bernoulli process (BeP).

A good review of random feature allocation models appears in Broderick et al. (2013) who introduce an exchangeable feature probability function (EFPF) to characterize a feature allocation model similar to the exchangeable partition probability function (EPPF) for random partition models. Similar to how feature allocation models generalize partitions by allowing membership in multiple sets, Campbell et al. (2016) introduce random trait allocation models to generalize feature allocation by allowing multiple memberships in the same set. That is, instead of a random binary matrix $\boldsymbol{Z}$, they define a random matrix of counts, with counts indicating the level of multiple memberships.

5 Bayesian nonparametric models for spatial data

For spatially referenced data investigators are often interested in characterizing how a response variable changes across locations (i.e., geographical coordinates). The general idea is that observations taken at locations that are close to each other in space will display more similarity than observations taken at locations that are farther apart. For a particular example see Example 6, below, where the outcome are results from high school math exams across 1215 schools around the greater Santiago area in Chile. Formally, let $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ represent a set of spatial coordinates, typically either 2- or 3-dimensional vectors each. Assume that there is a corresponding set of p-dimensional observations $\{\boldsymbol{y}_i : i = 1, \ldots, n\}$ at each location and that we are interested in developing models for spatial dependence. A crucial part of this effort consists then in defining suitable covariance functions that capture the desired dependence, and that assist in implementing predictive inference for data to be recorded at new locations.

5.1 Models based on the Dirichlet process

Perhaps the simplest way of directly incorporating spatial information in a BNP model is to augment the original response $\boldsymbol{y}_i$ with spatial location $\boldsymbol{s}_i$ (i.e., build $\tilde{\boldsymbol{y}}_i = (\boldsymbol{y}_i, \boldsymbol{s}_i)$), and proceed with inference under a DPM model for the augmented response $\tilde{\boldsymbol{y}}_i$ (Müller et al. 1996). The implied conditional distribution of the response given spatial coordinates allows for the desired predictions at new locations. However, this approach does not explicitly model spatial dependence in a BNP model. Gelfand et al. (2005) proposed one of the first BNP approaches for spatial inference for point-referenced data, that is, data that may be observed continuously over a certain spatial domain of interest $\mathscr{D} \subset \mathscr{R}^d$. The approach requires that replicates at each of the locations are available. Although true replication at each location is not common, temporal replication (which is quite common) can serve as a surrogate. Denoting by $\boldsymbol{y}_t = (y_1, \ldots, y_n)$, the tth replicate for $t = 1, \ldots, T$, and $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ the complete dataset, Gelfand et al. (2005) assume

$$\boldsymbol{y}_t \mid \boldsymbol{\beta}, \boldsymbol{\mu}, \tau, \overset{\text{ind}}{\sim} N(\boldsymbol{x}_t'\boldsymbol{\beta} + \boldsymbol{\mu}_t, \tau^2 \boldsymbol{I}), \qquad t = 1, \ldots, T,$$

where $\boldsymbol{\mu} = \{\mu_t : t = 1, \ldots, T\}$, $\boldsymbol{x}_t$ is a set of covariates recorded together with the tth replicate, and a prior structure given by

$$\boldsymbol{\mu}_t \mid G \overset{\text{iid}}{\sim} G, \qquad G \sim \text{DP}(G^\star, M),$$

where $G^\star$ is a multivariate normal that arises from a Gaussian process realized at the observed locations, with zero mean and covariance function $\sigma^2 \boldsymbol{H}_\phi(\cdot, \cdot)$. They complete the model specification with conjugate-style priors for τ^2, $\boldsymbol{\beta}$, σ^2, a gamma prior for M and a suitable prior for ϕ. Marginalizing with respect to $\boldsymbol{\mu}_t$ the sampling model reduces to a mixture

$$\boldsymbol{y}_t \mid \boldsymbol{\beta}, \tau, G \overset{\text{ind}}{\sim} \int N(\boldsymbol{x}_t'\boldsymbol{\beta} + \boldsymbol{\mu}, \tau^2 \boldsymbol{I})\, dG(\boldsymbol{\mu}), \qquad G \sim \text{DP}(G^\star, M).$$

By the DP representation described in Sethuraman (1994), the sampling model can be alternatively expressed as

$$\boldsymbol{y}_t \mid \boldsymbol{\beta}, \{\boldsymbol{\theta}_h\}_{h \geq 1}, \tau \overset{\text{ind}}{\sim} \sum_{h=1}^{\infty} w_h N(\boldsymbol{x}_t'\boldsymbol{\beta} + \boldsymbol{\theta}_h, \tau^2 \boldsymbol{I}),$$

where $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \overset{\text{iid}}{\sim} G^\star$ are independent multivariate normal realizations, that is, surfaces over $\boldsymbol{s}$, and the $\{w_h\}$ follow the usual stick-breaking construction i.e. $w_1 = V_1$, and $w_h = V_h \prod_{i=1}^{h-1}(1 - V_i)$ for $h \geq 2$, with $V_i \overset{\text{ind}}{\sim} \text{Beta}(a_i, b_i)$ (Ishwaran and James 2001). From this model, the induced covariance structure conditional on G turns out to be:

$$\text{Cov}(\boldsymbol{y} \mid G, \boldsymbol{\beta}, \tau^2) = \tau^2 \boldsymbol{I} + \sum_{h \geq 1} w_h \boldsymbol{\theta}_h \boldsymbol{\theta}_h' - \left\{ \sum_{h \geq 1} w_h \boldsymbol{\theta}_h \right\} \left\{ \sum_{h \geq 1} w_h \boldsymbol{\theta}_h \right\}'.$$

A related model, that includes spatial as well site-specific correlation, was proposed by Reich et al. (2013) for the analysis of periodontal disease data.

One limitation of the model by Gelfand et al. (2005) is that for any set of n locations, the model assumes the same set of stick-breaking probabilities. Relatedly, surface selection is global. That is, a particular repeat observation corresponds to the same surface across the entire space. To allow such localized surface selection Duan et al. (2007) introduce a random distribution for spatial effects where the selected surface can change with locations, and the joint selection of surfaces can also vary with the choice of locations. Their proposal thus extends the model proposed by Gelfand et al. (2005), while still retaining the property that the marginal distribution at each location follows a regular DP. Their basic proposal consists of extending the definition of stick-breaking weights, using instead a collection of joint probabilities $\{p_{i_1,\ldots,i_n}\}$ where $i_j = i(\boldsymbol{s}_j)$ for $j = 1, \ldots, n$ denotes the location-specific selected surface. Duan et al. (2007) choose the joint probabilities that constitute the weights in their construction to satisfy a continuity property in the sense that the random laws corresponding to nearby locations $\boldsymbol{s}_{i_1}$ and $\boldsymbol{s}_{i_2}$ are similar. Furthermore, they provide details for a particular construction of the spatially varying weights that can be seen as a multivariate stick-breaking process, with Gaussian thresholding. See additional details and applications in Duan et al. (2007). Motivated by functional data analysis, Petrone et al. (2009) developed a more parsimonious approach to spatial surface estimation relative to Duan et al. (2007) and therefore in a sense simplified Duan et al. (2007)'s method.

DP priors have also been used to specify other aspects of nonparametric spatial models. Reich and Fuentes (2012) consider the problem of modeling the spatial covariance function, a fundamental component of any spatial data analysis. Assume a model of the form $y(\boldsymbol{s}) = x(\boldsymbol{s})'\boldsymbol{\beta} + \mu(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$, where again, $\boldsymbol{s}$ represents the spatial locations, $x(\boldsymbol{s})$ is a location-specific vector of covariates, $\mu(\boldsymbol{s})$ is a spatial process, and $\epsilon(\boldsymbol{s}) \overset{\text{iid}}{\sim} N(0, \sigma^2)$ is a residual term. Under the assumption of a stationary covariance function, i.e. $\text{Cov}(\mu(\boldsymbol{s}), \mu(\boldsymbol{s} + \boldsymbol{r})) = C(\boldsymbol{r})$, the key aspect is to model $C(\boldsymbol{r})$. By Bochner's theorem the stationary covariance function $C(\boldsymbol{r})$ is non-negative definite if and only if

$$C(\boldsymbol{r}) = \int_{\mathbb{R}^2} \cos(\boldsymbol{\omega}'\boldsymbol{r})\, F(d\boldsymbol{\omega}),$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2) \in \mathbb{R}^2$ is a frequency, and F is a positive and finite measure on $\mathbb{R}^2$. Assuming $F(d\boldsymbol{\omega}) = \tau^2 f(\boldsymbol{\omega})$, where f is the *spectral density*, it follows that the so-represented spatial process is real-valued if and only if $f(\boldsymbol{\omega}) = f(-\boldsymbol{\omega})$. The spectral representation theorem thus formulates the spatial process as a convolution of trigonometric basis functions and stochastic processes in the frequency domain, with independent increments. Reich and Fuentes (2012) model the spectral density f as

$$f(\boldsymbol{\omega}) = \frac{1}{2}\sum_{j\geq 1} w_j \left\{\delta_{\boldsymbol{\alpha}_j}(\boldsymbol{\omega}) + \delta_{-\boldsymbol{\alpha}_j}(\boldsymbol{\omega})\right\}, \tag{14}$$

where $\{w_j : j \geq 1\}$ arise as the stick-breaking weights in the DP representation, and $\boldsymbol{\alpha}_j \overset{\text{iid}}{\sim} f_\theta$, where f_θ is a suitable parametric centering density with hyperparameters θ. Model (14) defines a *symmetric DP*. It is related to the invariant DP of Dalal (1979), a skewed extension of which was proposed in Iglesias et al. (2009). It follows (Reich and Fuentes 2012) that the covariance function induced by (14) is given by

$$\text{Cov}(\mu(\boldsymbol{s}), \mu(\boldsymbol{s}+\boldsymbol{r})) = \tau^2 \sum_{j\geq 1} w_j \cos(\boldsymbol{\alpha}_j'\boldsymbol{r}).$$

As usual, the limitations implied by the discreteness of (14) can be solved by mixing this representation with a convenient kernel density, and this is the approach taken by Reich and Fuentes (2012). Specifically, letting $g_\gamma(\boldsymbol{\omega} \mid \boldsymbol{\alpha})$ be a density with location $\boldsymbol{\alpha}$, parameters γ, and such that $g_\gamma(\boldsymbol{\omega} \mid \boldsymbol{\alpha}) = g_\gamma(-\boldsymbol{\omega} \mid -\boldsymbol{\alpha})$, they assume

$$f(\boldsymbol{\omega}) = \frac{1}{2}\sum_{j\geq 1} w_j \left\{g_{\gamma_j}(\boldsymbol{\omega}_j \mid \boldsymbol{\alpha}_j) + g_{\gamma_j}(\boldsymbol{\omega}_j \mid -\boldsymbol{\alpha}_j)\right\}.$$

See further details in Reich and Fuentes (2012). Additional uses of the DP to model spatial dependence include the spatial DP copula model for spatial extremes of Fuentes et al. (2013).

5.2 Models beyond the DP

Nonparametric spatial constructions based on models beyond the DP have also been considered in the literature. To analyze hurricane surface wind fields, Reich and Fuentes (2007) consider a similar data model as in the previous section i.e.

$$y_i = \mu_i + x_i\beta + \epsilon_i,$$

but now introduce spatial dependance through the weights of stick-breaking prior distribution. Specifically, they model $\mu_i \sim F(\boldsymbol{s}_i)$ and $F(\boldsymbol{s}_i) = \sum_{h=1}^m w_h(\boldsymbol{s}_i)\delta(\theta_h)$ where $w_h(\boldsymbol{s}_i) = V_h(s_i)\prod_{\ell=1}^{h-1}(1 - V_\ell(\boldsymbol{s}_i))$ and $V_h(\boldsymbol{s}_i) = V_h k_h(\boldsymbol{s}_i)$ for all $h \geq 1$. Here $\{k_h(\boldsymbol{s}_i)\}$ is a collection of spatial kernel functions that formalize the spatial dependence and $V_h \overset{\text{ind}}{\sim} \text{Beta}(a_h, b_h)$ as before. Reich and Fuentes (2007) provide several examples of kernel functions, which are defined by considering knots $\boldsymbol{\psi}_h = (\psi_{h1}, \psi_{h2})$, that serve as centers, with a spread controlled by bandwidth parameters $\boldsymbol{\varepsilon}_h = (\varepsilon_{h1}, \varepsilon_{h2})$ so that k_h decays as the distance to the knot $\boldsymbol{\psi}_h$ increases, relative to the bandwidth $\boldsymbol{\varepsilon}_h$.

A related multivariate spatial model was developed by Fuentes and Reich (2013), where the atoms are defined in terms of a multivariate Gaussian process, and the weights follow a similar construction as in Reich and Fuentes (2007).

Sudderth and Jordan (2009) apply a spatial version of a Pitman-Yor process (Pitman and Yor 1997) to the problem of unsupervised segmentation and discovery of visual object categories from image databases. They use Gaussian processes to discover spatially contiguous segments which respect image boundaries, thus encouraging nearby pixels to belong to the same cluster a priori.

Papageorgiou et al. (2015) consider a model for spatially indexed data of mixed type. Their construction relies on a probit-stick breaking mixture model representation (Rodríguez and Dunson 2011) of responses. Specifically, denote by $\boldsymbol{y}_i$ the mixed type responses, by $\boldsymbol{w}_i$ the confounders, and by $\boldsymbol{x}_i$ the covariates for area i. Papageorgiou et al. (2015) assumed a model with a truncated mixture

$$p_i(\boldsymbol{y}_i, \boldsymbol{w}_i \mid \boldsymbol{x}_i) = \sum_{h=1}^{T} \pi_{hi}\, p(\boldsymbol{y}_i, \boldsymbol{w}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_h), \tag{15}$$

where the mixture weights arise from a Gaussian Markov random field (Rue and Held 2005) and are defined through

$$\pi_{hi} = \Phi(\eta_{hi}) \prod_{\ell<h} \{1 - \Phi(\eta_{\ell i})\},$$

with $\eta_{hi} = \alpha + u_{hi}/\phi$, and the Gaussian Markov random field realizations $\boldsymbol{u}_h = (u_{h1}, \ldots, u_{hn})$ are generated as independent draws from the multivariate normal distribution $N_n(\boldsymbol{0}, \boldsymbol{Q}_\lambda^{-1})$. In addition, they assume $\boldsymbol{Q}_\lambda = \lambda A + \boldsymbol{I}_n$ where the adjacency matrix $\boldsymbol{A} = (a_{ij})$ is given by $a_{ii} =$ the number of neighbors of area i, $a_{ij} = -1$ if locations i and j are neighbors, and $a_{ij} = 0$ otherwise. The spatial dependence is then conveyed by the random field assumption, and the neighboring structure implied by the choice of $\boldsymbol{Q}_\lambda$. In addition, the responses $\boldsymbol{y}_i$ are regarded as manifestations of some latent variables $\tilde{\boldsymbol{y}}_i$, say. Responses and latent variables are linked by a thresholding defined through suitable cutoffs. For instance, if y_{ik} is a Binomial response corresponding to N_{ik} trials, they let

$$y_{ik} = \sum_{q=0}^{N_{ik}} q\, I\{c_{ik,q-1} < \tilde{y}_{ik} < c_{ik,q}\},$$

where $c_{ik,-1} = -\infty$ and, for $\ell \geq 1$, $c_{ik,\ell} = c_\ell(r_{ik}) = \Phi^{-1}\{G(\ell;\ N_{ik}, r_{ik})\}$, and where $G(\cdot;\ N_{ik}, r_{ik})$ is the Binomial CDF. This way, $p(\cdot)$ in (15) takes the form

$$p_i(\boldsymbol{y}_i, \boldsymbol{w}_i \mid \boldsymbol{x}_i\boldsymbol{\theta}_i) = \int \cdots \int N(\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{w}}_i \mid \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}}_i)\, d\tilde{\boldsymbol{y}}_i\, d\tilde{\boldsymbol{w}}_i.$$

See further details, and an assessment of the association between birth outcomes and exposure to ambient air pollution, in Papageorgiou et al. (2015).

More recently, Jo et al. (2015) and Jo et al. (2017) proposed a spatial density estimation procedure based on a *proper* dependent species sampling model (SSM). A proper SSM (Pitman 1996) consists of discrete random probability measure of the

form $F(\cdot) = \sum_{h\geq 1} w_h \delta_{\theta_h}(\cdot)$, where the atoms $\{\theta_h\}_{h\geq 1}$ are independent of the weights $\{w_h\}_{h\geq 1}$, with $\theta_h \overset{\text{iid}}{\sim} F_0$ and $P(\sum_{h\geq 1} w_h = 1) = 1$. Obviously, this includes the DP, and more generally, the class of stick-breaking priors as special cases. Lee et al. (2013) studied a particular SSM construction based on normalizing a sequence of independent random variables $\{r_h\}_{h\geq 1}$ such that $P(\sum_{h\geq 1} r_h < \infty) = 1$, by setting

$$w_h = \frac{r_h}{\sum_{\ell\geq 1} r_\ell}, \qquad h \geq 1. \tag{16}$$

For the normalization to make sense, it suffices to impose $\sum_{h\geq 1} E(r_h) < \infty$ and $\sum_{h\geq 1} Var(r_h) < \infty$ (Lee et al. 2013).

Jo et al. (2017) define a prior distribution for a spatially dependent collection of random probability measures $\{G_i : i = 1, \ldots, n\}$ at each of corresponding locations $\{\boldsymbol{s}_i : i = 1, \ldots, n\}$ by assuming

$$G_i(\cdot) = \sum_{h\geq 1} w_{i,h} \delta_{\theta_h}(\cdot),$$

where the weights follow the construction in (16) and are defined as

$$r_{i,h} = \exp(u_{i,h}), \qquad 1 \leq i \leq n, \quad h \geq 1, \tag{17}$$

and the $\{u_{i,h}\}$ arise from a Gaussian conditional autoregressive (CAR) model (Banerjee et al. 2014, chapter 3). Jo et al. (2017) consider the special case of a prior for which

$$u_{i,h} \mid u_{\ell,h}, \ell \neq i \sim N\left(m_{i,h} - \sum_{\ell:\, \ell\neq i} \xi_{i,\ell}(u_{\ell,h} - m_{\ell,h}), \tau^2\right), \qquad 1 \leq i \leq n,$$

with $\tau^2 > 0$, $m_{i,h} = \log\left(1 - (1 + e^{b-ah})^{-1}\right)$ for $1 \leq i \leq n$ with $a, b > 0$ (see Lee et al. 2013, for an explanation of why these choices produce a valid process) and where the $\{\xi_{i,\ell}\}$ variables are defined so as to induce two possible types of neighborhood and/or covariance structure. See further details in Jo et al. (2017).

5.3 Models based on random partitions

An alternative way to introduce spatial dependence is based on clustering spatial locations with cluster-specific parameters in the sampling models. Page and Quintana (2016) develop such constructions based on product partition models (PPMs). PPMs were introduced in Hartigan (1990) and are characterized by a distribution on partitions, called the *product distribution*. Let $y_1, \ldots, y_n$ denote the responses for n individuals, with corresponding parameters $\theta_1, \ldots, \theta_n$. Assume conditional independence, i.e. $p(\boldsymbol{y}^n \mid \boldsymbol{\theta}^n) = \prod_{i=1}^n p(y_i \mid \theta_i)$ where $\boldsymbol{y}^n$ and $\boldsymbol{\theta}^n$ are the complete set of observations and parameters, respectively.

The PPM involves a distribution on the set $\mathscr{P}_n$ of all possible partitions ρ_n of $[n] = \{1, \ldots, n\}$, where $n \geq 1$. A partition ρ_n with k_n subsets can be generically described as $\rho_n = (S_1, \ldots, S_{k_n})$ where the S_j are mutually exclusive nonempty and exhaustive subsets of $[n]$. For any subset $S \subset [n]$ define its cohesion function $c(S)$ as a nonnegative quantity with higher values indicating greater belief in that the elements of S will be together in the resulting partition. The PPM assumes a prior on ρ_n (i.e. the product distribution) as

$$p(\rho_n = (S_1, \ldots, S_{k_n})) \propto \prod_{j=1}^{k_n} c(S_j), \tag{18}$$

with normalization constant $\sum_{\rho_n \in \mathscr{P}_n} \prod_{j=1}^{k_n} c(S_j)$. On top of (18), the PPM assumes that the parameters in $\boldsymbol{\theta}^n$ follow the partition structure in ρ_n in the sense that $i, i' \in S_j$ implies that $\theta_i = \theta_{i'}$, i.e., there are ties in the parameters. Denoting then by $\theta_1^*, \ldots, \theta_{k_n}^*$ the unique values among $\boldsymbol{\theta}^n$, the PPM assumes

$$p(\boldsymbol{y}^n \mid \boldsymbol{\theta}^n, \rho_n) = \prod_{j=1}^{k_n} \prod_{i \in S_j} p(y_i \mid \theta_j^*). \tag{19}$$

See further connections between PPMs and nonparametric models in Quintana and Iglesias (2003) and Quintana (2006).

Hegarty and Barry (2008) consider a modified PPM for spatial inference in the context of disease mapping. Specifically, they aimed at identifying areas of high or low risk among a set of areas $A_1 \ldots, A_n$. They considered the cohesion function as $c(S) = \beta^{\ell(s)}$, where $\ell(S) = \sum_{A_i \in S} \ell(A_i)$ and where $\ell(A_i)$ denotes the number of neighbors (i.e. areas that share a border) of A_i not in S. This definition encourages maps with few fragmented components. See further details in Hegarty and Barry (2008).

Müller et al. (2011) proposed another modification of the PPM to introduce dependence on covariates. See also Müller and Quintana (2010). This consisted in modifying the product distribution (18) by adding a *similarity function* $g(\boldsymbol{x}_j^*)$ that measures how homogeneous the covariate values in the cluster-specific subset $\boldsymbol{x}_j^* = \{x_i : i \in S_j\}$ are, with higher values of $g(\boldsymbol{x}_j^*)$ indicating an increased preference for the corresponding subset. Then, the cohesion function in (18) is replaced by $c(S_j)g(\boldsymbol{x}_j^*)$. Page and Quintana (2016) exploited this idea by proposing and studying several ways to define spatial similarities, where $\boldsymbol{s}_i$ represents the spatial coordinates for the ith location.

An alternative random partition model that incorporates spatial information is the spatial distance dependent Chinese restaurant process of Ghosh et al. (2011). The authors develop a methodology that produces a non-exchangeable distribution on spatially dependent partitions. This is done by weighting cluster membership probabilities with a distance dependent decay function.

5.4 Standardized math testing in Chile

Example 6 (Standardized math testing in Chile) Chile's Ministry of Education has established a national large-scale standardized test called SIMCE (Sistema de Medición de la Calidad de la Educación, System Measurement of Quality of Education). The test is administered at the end of the 4th and 8th grades in elementary school and at the end of the 11th grade in high school and attempts to assess among other things mathematics skill. The results of the exam are now a key component of Chilean educational policies and can also be used by parents as a tool in decision making regarding their children's education.

Here we analyze results of the 4th grade math SIMCE test taken in 2011. To simplify the discussion, instead of analyzing individual test scores we compute school specific averages. All schools in Chile were recorded. Here we focus on the 1215 schools in the greater Santiago area. Figure 5 shows a spatial plot of the average SIMCE score for each institution. Schools in the north east part of the city tend to have higher math SIMCE scores compared to those in the south and west. As in many large cities, available socio-economic resources tend to cluster spatially in Santiago with areas in the north east part of the city generally recognized as more affluent, thus providing reason to believe that spatial structure exists in SIMCE scores.

The goal is to understand spatial variation of school performance, exploiting spatial smoothing as appropriate. Spatial dependence is expected to vary across the region due to neighborhood boundaries, topography, infrastructure etc. Inference should include grouping and clustering of schools which is useful for possible policy implications.

We use the discussed methods for inference with the Chilean math testing scores. Based on availability of software, we consider the spatial stick-breaking (SSB) of Reich and Fuentes (2007), the spatial product partition model (sPPM) of Page and Quintana (2016) and a Dirichlet process mixture (sDPM) that models the math SIMCE score and spatial coordinates jointly (see Müller et al. 1996). For the sPPM we consider the following two cohesions

$$C_1(S_h, \boldsymbol{s}_h^\star) = \begin{cases} \dfrac{M \times \Gamma(|S_h|)}{\Gamma(\alpha \mathscr{D}_h)\mathbb{I}[\mathscr{D}_h \geq 1] + (\mathscr{D}_h)\mathbb{I}[\mathscr{D}_h < 1]} & \text{if } |S_h| > 1 \\ M & \text{if } |S_h| = 1. \end{cases} \tag{20}$$

$$C_2(S_h, \boldsymbol{s}_h^\star) = M \times \Gamma(|S_h|) \times \int \prod_{i \in S_h} q(\boldsymbol{s}_i|\boldsymbol{\xi}_h) q(\boldsymbol{\xi}_h|\boldsymbol{s}_h^\star) d\boldsymbol{\xi}_h. \tag{21}$$

The first is constructed to explicitly model spatial dependence where $\boldsymbol{s}_h^\star = \{\boldsymbol{s}_i : i \in S_h\}$ and $\mathscr{D}_h = \sum_{i \in S_h} d(\boldsymbol{s}_i, \bar{\boldsymbol{s}}_h)$ denotes the sum of all Euclidean centroid distances and $\bar{\boldsymbol{s}}_h$ denotes the centroid of cluster S_h whose coordinates are calculated using $\bar{s}_{hk} = 1/n_h \sum_{i \in S_h} s_{ik}$ for $k = 1, 2$ and $n_h = |S_h|$. The later cohesion is motivated by ideas found in Müller et al. (2011). $M \times \Gamma(|S_h|)$ is included so that partitions with a small number of large clusters are favored over a large number of small clusters. M regulates the number of clusters similar to the dispersion parameter of a DP. For C_1

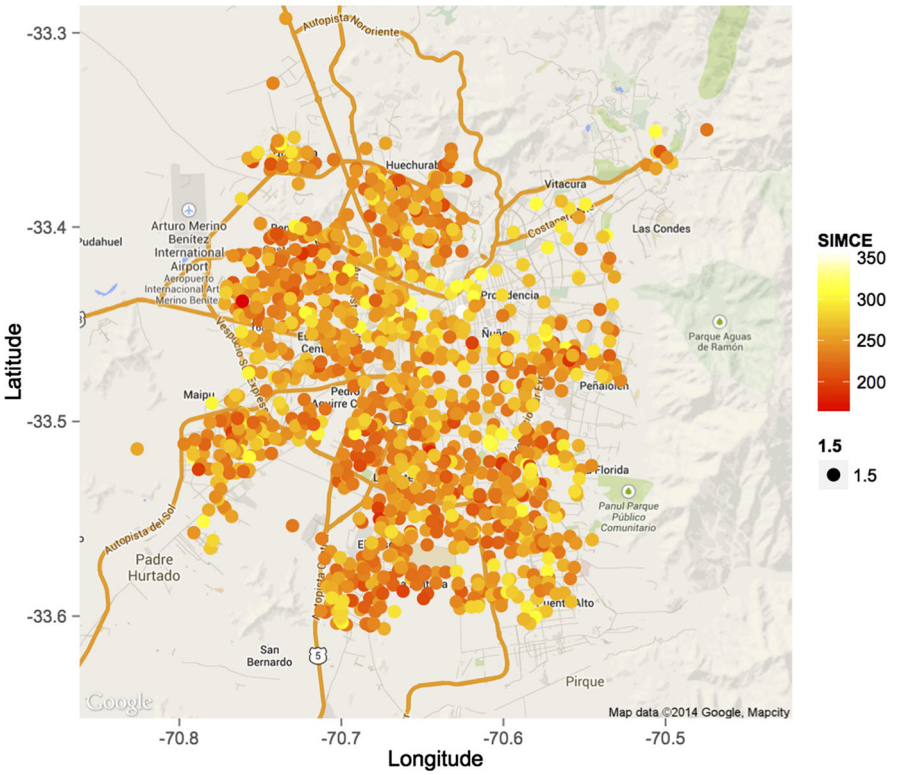

Fig. 5 SIMCE math scores plotted by location. Each point corresponds to a school's location and color to the institution's average SIMCE math score

we set $M = 0.0001$ and $\alpha = 2$ and for C_2 we set $M = 1$. For more details regarding motivation behind C_1 and C_2 and values employed for tuning parameters see Page and Quintana (2016).

For the SSB procedure we employed a Gaussian kernel, fixed the bandwidth at 0.1 (which penalizes distance more heavily) and set $M = 1$ (which favors more clusters). The sDPM model was fit using the default values provided in the function `DPcdensity` found in the `DPpackage` (Jara et al. 2011).

Upon introducing cluster labels the precise model posed for the sPPM and SSB is

$$y(\boldsymbol{s}_i)|c_i, \mu^*_{c_i}(\boldsymbol{s}_i), \sigma^2 \overset{ind}{\sim} N(\mu^*_{c_i}(\boldsymbol{s}_i), \sigma^2),\ \sigma \sim UN(0, 10),\ \beta \sim N(0, 10^2)$$
$$\mu^*_h(\boldsymbol{s}_i) \overset{iid}{\sim} N(\mu_0, \sigma_0^2),\ \mu_0 \sim N(0, 10^2),\ \sigma_0 \sim UN(0, 10)$$

where for the sPPM we have $\{c_i\}_{i=1}^n \sim sPPM$ and for the SSB $\{c_i\}_{i=1}^n \sim SSB$. The exact model used for the DPM is

Table 1 Model fit results from the four procedures

Method	Cluster	MSPE	Method	Cluster	MSPE
sPPM C_1	41.63	566.23	SSB	19.23	591.06
sPPM C_2	33.87	565.64	sDPM	12.82	568.65

The second column reports the posterior expected number of clusters. The last column is mean square prediction error

$$(\boldsymbol{y}_i, \boldsymbol{s}_i) \mid G \sim \int N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}) dG(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad G \sim DP(M, G_0)$$

where $G_0 = N_3(\boldsymbol{\theta}, (1/k_0)\boldsymbol{\Sigma}) \times IW(\nu, \psi)$. For more details see Jara et al. (2011). Each procedure was fit to data by collecting 1000 MCMC iterates after discarding the first 10,000 as burn-in and thinning by 10. To facilitate prior value selection the Math SIMCE scores were standardized to have mean zero and unit standard deviation.

To compare model results from the four procedures we focus on out of sample prediction and spatial clustering. We assess out of sample prediction in two ways. First we divided the 1215 schools into 600 training observations and 615 testing observations and used the 615 testing observations to assess predictive accuracy. Second we created a fine grid contained in the convex hull of the school locations to compare spatial structure in the predictive maps. Out of sample prediction results are provided in Table 1 and predictive maps can be found in the left column of Fig. 6.

The column of Table 1 titled "Cluster" is the posterior expected number of clusters. Values under the column "MSPE" report the average squared prediction error for the 615 testing observations. It appears that overall, sPPM using C_2 produced the lowest out of sample prediction error. Focusing on Fig. 6, it appears that the SSB smooths the most as the range of predictive values is the smallest and spatial structure is the least present. The sDPM appears to smooth the least as the range of predictive values is the greatest. The spatial structure between the two sPPM procedures are very similar (only one shown). The predictive map of the sDPM displays more peaks and valleys, which we interpret as further evidence that it smooths the least.

Regarding spatial clustering, we use Dahl (2006)'s least squares method to summarize the posterior MCMC simulation output by a single estimate. Results are shown in right column of Fig. 6. Posterior clustering under the SSB model is the least intuitively appealing, in the sense that clusters are not geographically constrained. Clustering under the sDPM produces a partition that is practically disjoint spatially. The estimated partition under the sPPM model (only C_2 shown) provides an attractive compromise between the SSB and sDPM.

6 Conclusion

We reviewed some applications of nonparametric Bayesian methods and approaches. By focusing on specific application examples and spatial inference we avoided duplication and overlap with many other recent reviews of BNP. This focus on specific

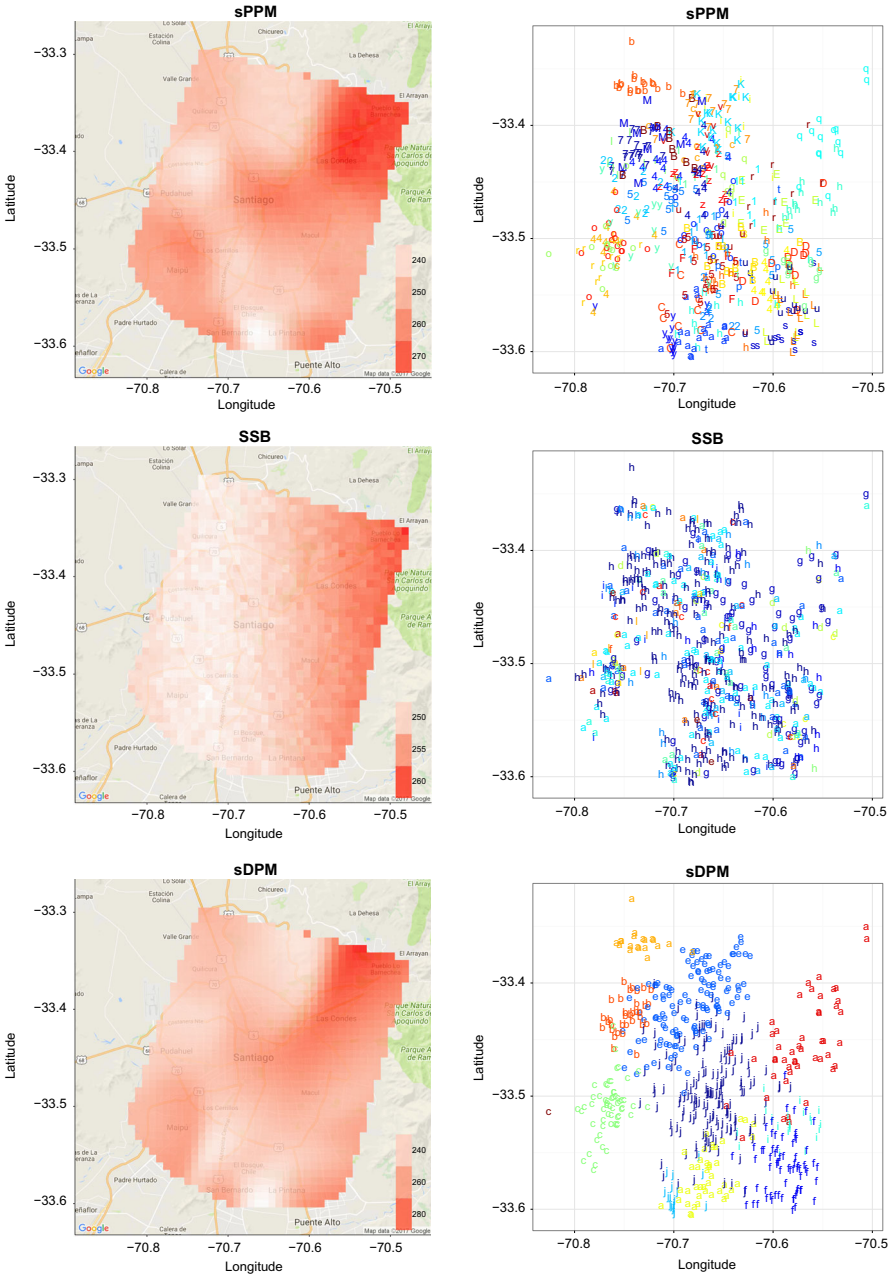

Fig. 6 Predictive maps (left column) and estimated partition of schools (right panels). The partitions are determined using the algorithm of Dahl (2006). Each color/symbol combination corresponds to a cluster

examples and (in some more detail) spatial problems excluded discussion of many otherwise important recent contributions to BNP. In particular, we did not review asymptotic results for BNP models. A systematic review of related results is forthcoming in Ghoshal and Vaart (2017). Also, by focusing on applications and methods we did not discuss many of the important computational problems related to implementing inference under BNP methods. Naturally, many of the computational methods remain very specific to particular models and applications, precluding a general review.

Acknowledgements Peter Müeller was partially funded by Grant NIH R01 CA132891-06A1. Fernando A. Quintana was partially funded by Grant FONDECYT 1141057.

References

Argiento R, Bianchini I, Guglielmi A (2016) A blocked Gibbs sampler for NGG-mixture models via a priori truncation. Stat Comput 26(3):641–661

Argiento R, Guglielmi A, Pievatolo A (2010) Bayesian density estimation and model selection using nonparametric hierarchical mixtures. Comput Stat Data Anal 54(4):816–832

Baladandayuthapani V, Mallick BK, Carroll R (2005) Spatially adaptive Bayesian penalized regression splines (P-splines). J Comput Graph Stat 14:378–394

Banerjee S, Carlin BP, Gelfand AE (2014) Hierarchical modeling and analysis for spatial data, 2nd edn. Chapman and Hall/CRC, Boca Raton

Barrios E, Lijoi A, Nieto-Barajas LE, Prünster I (2013) Modeling with normalized random measure mixture models. Stat Sci 28(3):313–334

Berger J, Guglielmi A (2001) Bayesian testing of a parametric model versus nonparametric alternatives. J Am Stat Assoc 96:174–184

Brezger A, Kneib T, Lang S (2005) BayesX: analyzing Bayesian structural additive regression models. J Stat Softw 14(1):1–22

Broderick T, Pitman J, Jordan MI (2013) Feature allocations, probability functions, and paintboxes. Bayesian Anal 8(4):801–836

Camerlenghi F (2015) Hierarchical and nested random probability measures with statistical applications, PhD thesis, Universita degli Studi di Pavia, Pavia

Campbell T, Cai D, Broderick T (2016) Exchangeable trait allocations. ArXiv e-prints

Chipman HA, Kolaczyk ED, McCulloch RE (1997) Adaptive Bayesian wavelet shrinkage. J Am Stat Assocn 92:1413–1421

Clyde M, George E (2000) Flexible empirical Bayes estimation for wavelets. J R Stat Soc Ser B 62:681–698

Crane H (2016) The ubiquitous Ewens sampling formula. Stat Sci Rev J Inst Math Stat 31(1):1–19

Dahl DB (2006) Model-based clustering for expression data via a Dirichlet process mixture model. In: Vannucci M, Do KA, Müller P (eds) Bayesian inference for gene expression and proteomics. Cambridge University Press, Cambridge, pp 201–218

Dalal SR (1979) Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions. Stoch Process Their Appl 9:99–108

De Blasi P, Favaro S, Lijoi A, Mena R, Prünster I, Ruggiero M (2015) Are Gibbs-type priors the most natural generalization of the Dirichlet process? IEEE Trans Pattern Anal Mach Intell 37:212–229

Duan JA, Guindani M, Gelfand AE (2007) Generalized spatial Dirichlet process models. Biometrika 94(4):809–825

Dykstra RL, Laud P (1981) A Bayesian nonparametric approach to reliability. Ann Stat 9:356–367

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol Int J 3:87–112; erratum, ibid. 3 (1972), 240; erratum, ibid. 3 (1972), 376

Fahrmeir L, Kneib T, Lang S (2004) Penalized structured additive regression for space-time data: a Bayesian perspective. Stat Sin 14:731–761

Favaro S, Teh YW (2013) MCMC for normalized random measure mixture models. Stat Sci 28(3):335–359

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 1:209–230

Ferguson TS, Phadia EG (1979) Bayesian nonparametric estimation based on censored data. Ann Stat 7(1):163–186

Foti NJ, Williamson SA (2015) A survey of non-exchangeable priors for Bayesian nonparametric models. IEEE Trans Pattern Anal Mach Intell 37:359–371
Fuentes M, Henry J, Reich B (2013) Nonparametric spatial models for extremes: application to extreme temperature data. Extremes 16(1):75–101
Fuentes M, Reich B (2013) Multivariate spatial nonparametric modelling via kernel processes mixing. Stat Sin 23(1):75–97
Gelfand AE, Kottas A, MacEachern SN (2005) Bayesian nonparametric spatial modeling with Dirichlet process mixing. J Am Stat Assoc 100:1021–1035
Ghosh S, Ungureanu AB, Sudderth EB, Blei DM (2011) Spatial distance dependent Chinese restaurant processes for image segmentation. In: Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ (eds) Advances in neural information processing systems, vol 24. Curran Associates, New York, pp 1476–1484
Ghoshal S (2010) The Dirichlet process, related priors and posterior asymptotics. In: Hjort et al. (2010), pp 22–34
Ghoshal S, van der Vaart A (2017) Fundamentals of nonparametric Bayesian inference. Cambridge University Press, Cambridge
Gramacy RB, Lee HKH (2008) Bayesian treed Gaussian process models with an application to computer modeling. J Am Stat Assoc 103:1119–1130
Griffiths TL, Ghahramani Z (2006) Infinite latent feature models and the Indian buffet process. In: Weiss Y, Schölkopf B, Platt J (eds) Advances in neural information processing systems, vol 18. MIT Press, Cambridge, pp 475–482
Hanson TE, Jara A (2013) Surviving fully Bayesian nonparametric regression models. In: Damien P, Dellaportas P, Polson NG, Stephens DA (eds) Bayesian theory and applications. Oxford University Press, Oxford, pp 593–615
Hanson T, Johnson WO (2002) Modeling regression error with a mixture of Polya trees. J Am Stat Assoc 97:1020–1033
Hanson T, Johnson WO (2004) A Bayesian semiparametric AFT model for interval-censored data. J Comput Graph Stat 13:341–361
Hartigan JA (1990) Partition models. Commun Stat Theory Methods 19(8):2745–2756
Hegarty A, Barry D (2008) Bayesian disease mapping using product partition models. Stat Med 27(19):3868–3893
Hjort NL (1990) Nonparametric Bayes estimators based on beta processes in models for life history data. Ann Stat 18:1259–1294
Hjort NL, Holmes C, Müller P, Walker SG (eds) (2010) Bayesian nonparametrics. Cambridge University Press, Cambridge
Iglesias PL, Orellana Y, Quintana FA (2009) Nonparametric Bayesian modelling using skewed Dirichlet processes. J Stat Plan Inference 139(3):1203–1214
Ishwaran H, James LF (2001) Gibbs sampling methods for stick-breaking priors. J Am Stat Assoc 96(453):161–173
James LF, Lijoi A, Prünster I (2009) Posterior analysis for normalized random measures with independent increments. Scand J Stat 36(1):76–97
Jara A, Hanson T, Quintana F, Müller P, Rosner G (2011) DPpackage: Bayesian semi- and nonparametric modeling in R. J Stat Softw 40(5):1–30
Jo S, Lee J, Müller P, Quintana FA, Trippa L (2017) Dependent species sampling models for spatial density estimation. Bayesian Anal 12(2):379–406
Jo S, Lee J, Page G, Quintana FA, Trippa L, Müller P (2015) Spatial species sampling and product partition models. In: Mitra R, Müller P (eds) Nonparametric Bayesian inference in biostatistics. Springer, New York, pp 359–375
Kingman JFC (1993) Poisson processes. Oxford University Press, Oxford
Kottas A, Gelfand AE (2001) Bayesian semiparametric median regression modeling. J Am Stat Assoc 96:1458–1468
Lavine M (1992) Some aspects of Polya tree distributions for statistical modeling. Ann Stat 20:1222–1235
Lavine M (1994) More aspects of Polya tree distributions for statistical modelling. Ann Stat 22:1161–1176
Lee J, Müller P, Gulukota K, Ji Y (2015) A Bayesian feature allocation model for tumor heterogeneity. Ann Appl Stat 9(2):621–639
Lee J, Müller P, Zhu Y, Ji Y (2013) A nonparametric Bayesian model for local clustering with application to proteomics. Stat Sci 28:209–22

Lee J, Quintana F, Müller P, Trippa L (2013) Defining predictive probability functions for species sampling models. Stat Sci 28(2):209–222
Lijoi A, Mena RH, Prünster I (2007) Controlling the reinforcement in Bayesian non-parametric mixture models. J R Stat Soc Ser B (Statistical Methodology) 69(4):715–740
Lijoi A, Nipoti B, Prünster I (2014) Bayesian inference with dependent normalized completely random measures. Bernoulli 20(3):1260–1291
Lijoi A, Prünster I (2010) Models beyond the Dirichlet process. In: Hjort et al. (2010), pp 80–136
Lo AY (1984) On a class of Bayesian nonparametric estimates I: density estimates. Ann Stat 12:351–357
MacEachern S (1999) Dependent nonparametric processes. In: ASA proceedings of the section on Bayesian Statistical Science. ASA, Alexandria, VA
Müller P, Erkanli A, West M (1996) Bayesian curve fitting using multivariate normal mixtures. Biometrika 83:67–79
Müller P, Mitra R (2013) Bayesian nonparametric inference–why and how. Bayesian Anal 8(2):269–302
Müller P, Quintana F (2010) Random partition models with regression on covariates. J Stat Plan Inference 140(10):2801–2808
Müller P, Quintana FA (2004) Nonparametric Bayesian data analysis. Stat Sci Rev J Inst Math Stat 19(1):95–110
Müller P, Quintana F, Jara A, Hanson T (2015) Nonparametric Bayesian data analysis. Springer, New York
Müller P, Quintana F, Rosner GL (2011) A product partition model with regression on covariates. J Comput Graph Stat 20(1):260–278. Supplementary material available online
Nieto-Barajas L, Walker SG (2002) Markov beta and gamma processes for modelling hazard rates. Scand J Stat 29:413–424
Page GL, Quintana FA (2016) Spatial product partition models. Bayesian Anal 11(1):265–298
Papageorgiou G, Richardson S, Best N (2015) Bayesian non-parametric models for spatially indexed data of mixed type. J R Stat Soc Ser B Stat Methodol 77(5):973–999
Petrone S, Guindani M, Gelfand AE (2009) Hybrid Dirichlet mixture models for functional data. J R Stat Soc Ser B 71(4):755–782
Phadia EG (2013) Prior processes and their applications. Springer, New York
Pitman J (1996) Some developments of the Blackwell-MacQueen Urn scheme. In: Ferguson TS, Shapeley LS, MacQueen JB (eds) Statistics, probability and game theory. Papers in Honor of David Blackwell, Haywar, California, IMS Lecture Notes - Monograph Series, pp 245–268
Pitman J, Yor M (1997) The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann Probab 25(2):855–900
Quintana FA (2006) A predictive view of Bayesian clustering. J Stat Plan Inference 136(8):2407–2429
Quintana FA, Iglesias PL (2003) Bayesian clustering and product partition models. J R Stat Soc Ser B 65:557–574
Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. In: Adaptive computation and machine learning. MIT Press, Cambridge, MA
Regazzini E, Lijoi A, Prünster I (2003) Distributional results for means of normalized random measures with independent increments. Ann Stat 31(2):560–585
Reich BJ, Bandyopadhyay D, Bondell HD (2013) A nonparametric spatial model for periodontal data with nonrandom missingness. J Am Stat Assoc 108(503):820–831
Reich BJ, Fuentes M (2007) A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. Ann Appl Stat 1:249–264
Reich BJ, Fuentes M (2012) Nonparametric Bayesian models for a spatial covariance. Stat Methodol 9(1–2):265–274
Rodríguez A, Dunson DB (2011) Nonparametric Bayesian models through probit stick-breaking processes. Bayesian Anal 6(1):145–177
Rodríguez A, Dunson DB, Gelfand AE (2008) The nested Dirichlet process, with discussion. J Am Stat Assoc 103:1131–1144
Rodríguez A, Ghosh K (2012) Modeling relational data using nested infinite relational models, Technical report. Department of Applied Mathematics and Statistics, University of California, Santa Cruz
Rue H, Held L (2005) Gaussian Markov random fields. Monographs on statistics and applied probability, vol 104. Chapman & Hall/CRC, Boca Raton
Schörgendorfer A, Branscum A, Hanson T (2013) A Bayesian goodness of fit test and semiparametric generalization of logistic regression with measurement data. Biometrics 69:508–519
Sethuraman J (1994) A constructive definition of Dirichlet priors. Stat Sin 4(2):639–650

Sudderth EB, Jordan MI (2009) Shared segmentation of natural scenes using dependent Pitman-Yor processes. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) Advances in neural information processing systems, vol 21. Curran Associates, New York, pp 1585–1592

Teh YW, Görür D, Ghahramani Z (2007) Stick-breaking construction for the Indian buffet process. In: Proceedings of the 11th conference on artificial intelligence and statistics

Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Sharing clusters among related groups: hierarchical Dirichlet processes. J Am Stat Assoc 101:1566–1581

Thibaux R, Jordan M (2007) Hierarchical beta processes and the Indian buffet process. In: Proceedings of the 11th conference on artificial intelligence and statistics (AISTAT), Puerto Rico

Vidakovic B (1998) Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. J Am Stat Assoc 93:173–179

Wade S, Mongelluzzo S, Petrone S (2011) An enriched conjugate prior for Bayesian nonparametric inference. Bayesian Anal 6(3):359–385

Walker S (2013) Bayesian nonparametrics. In: Damien P, Dellaportas P, Polson NG, Stephens DA (eds) Bayesian theory and applications. Oxford University Press, Oxford, pp 249–270

Walker S, Mallick B (1999) A Bayesian semiparametric accelerated failure time model. Biometrics 55:477–483

Walker S, Muliere P (1997) Beta-Stacy processes and a generalization of the Pólya-urn scheme. Ann Stat 25:1762–1780

Williams CKI (1997) Prediction with Gaussian processes: from linear regression to linear prediction and beyond. In: Learning and inference in graphical models. Kluwer, pp 599–621

Xu Y, Müller P, Wahed AS, Thall PF (2016) Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. J Am Stat Assoc 111:921–950

Xu Y, Scharpstein D, Müller P, Daniels M (2016) A Bayesian nonparametric approach for semi-competing risks, Technical report. Johns Hopkins University

Xu Y, Thall PF, Müller P, Reza MJA (2017) A decision-theoretic comparison of treatments to resolve air leaks after lung surgery based on nonparametric modeling. Bayesian Anal 12(3):639–652