Bayesian Nonparametric Inference – Why and How

Peter Müller and Riten Mitra

Abstract

We review inference under models with nonparametric Bayesian (BNP) priors. The discussion follows a set of examples for some common inference problems. The examples are chosen to highlight problems that are challenging for standard parametric inference. We discuss inference for density estimation, clustering, regression and for mixed effects models with random effects distributions. While we focus on arguing for the need for the flexibility of BNP models, we also review some of the more commonly used BNP models, thus hopefully answering a bit of both questions, why and how to use BNP.

1 Introduction

All models are wrong, but some are useful (Box 1979). Most statisticians and scientists would agree with this statement. In particular, it is convenient to restrict inference to a family of models that can be indexed with a finite dimensional set of parameters. Under the Bayesian paradigm inference builds on the posterior distribution of these parameters given the observed data. In anticipation of the upcoming generalization we refer to such inference as parametric Bayes. However, it can be dangerous to forget the simplification implied by this process. There are problems where inference under the simplified model can lead to misleading decisions and inference. We discuss a class of statistical inference approaches that relax this framework by allowing for a richer and larger class of models. This is achieved by considering infinite dimensional families of probability models. Priors on such families are known as nonparametric Bayesian (BNP) priors.

For example, consider a density estimation problem, with observed data $y_i \sim G$, $i = 1, \ldots, n$. Inference under the Bayesian paradigm requires a completion of the model with a prior for the unknown distribution G. Unless G is restricted to some finite dimensional parametric family this leads to a BNP model with a prior $p(G)$, that is a probability model for the infinite dimensional G. A related application of BNP priors on random probability measures is for random effects distributions in mixed effects models. Such generalizations

of parametric models are important when the default choice of multivariate normal random effects distribution might understate uncertainties and miss some important structure. Another important class of BNP priors are priors on unknown functions, for example as prior $p(f)$ for the unknown mean function $f(x)$ in a regression model $y_i = f(x_i) + \epsilon_i$.

In this article we review some common BNP priors. Our argument for BNP inference rests on a set of examples that highlight typical situations where parametric inference might run into limitations, and BNP can offer a way out. Examples include a false sense of posterior precision in extrapolating beyond the range of the data, the restriction of a density estimate to a unimodal family of distributions and more. One common theme is the honest representation of uncertainties. Restriction to a parametric family can mislead investigators into an inappropriate illusion of posterior certainty. Honest quantification of uncertainty is less important when the goal is to report posterior means $E(G \mid \boldsymbol{y})$, but could be critical if either the primary inference goal is to characterize this uncertainty or the goal is prediction, if the probability model is part of a decision problem, or if the nonparametric model is part of a larger encompassing model. Some of these issues are highlighted in the upcoming examples. For each example we briefly review appropriate methods, but without any attempt at an exhaustive review of BNP methods and models. For a more exhaustive discussion of BNP models see, for example, recent discussions in Hjort et al. (2010), Müller and Rodriguez (2013), Walker et al. (1999), Müller and Quintana (2004), and Walker (2013).

2 Density Estimation

2.1 Dirichlet Process (Mixture) Models

Example 1 (T-cell diversity) *Guindani et al. (2012) estimate an unknown distribution F for count data y_i. Assuming $y_i \sim F$, $i = 1, \ldots, n$, i.i.d., the problem can be characterized as inference for the unknown F. The data are shown in Table 1. The application is to inference for T-cell diversity. Different types of T-cells are observed with counts y_i. T-cells are white blood cells and are a critical part of the immune system. In particular, investigators are interested in estimating $F(0)$, for the following reason. The experiment generates a random sample of T-cells from the population of all T-cells that are present in a probe. The sample is recorded by tabulating the counts y_i for all observed T-cell types, $i = 1, \ldots, n$. However, some rare but present T-cell types, $i = n+1, \ldots, N$, might not be recorded, simply by sampling variation, that is when $y_i = 0$ for a rare T-cell type. Naturally, zero counts are censored by the nature of the experiment. Inference for $F(0)$ would allow us to impute the number of not observed zero counts and thus inference on the total number of T-cell types. The latter is an*

Table 1: Clonal size distribution for one of the experiments reported in Guindani et al. (2012, Table 2). For example, there are $y_1 = 37$ T-cell receptor types that were observed once in the data, $y_2 = 11$ that were observed twice, etc. The number y_0 of T-cell receptors that were not observed in the sample is censored.

y_i	0	1	2	3	4	other
frequency	–	37	11	5	2	0

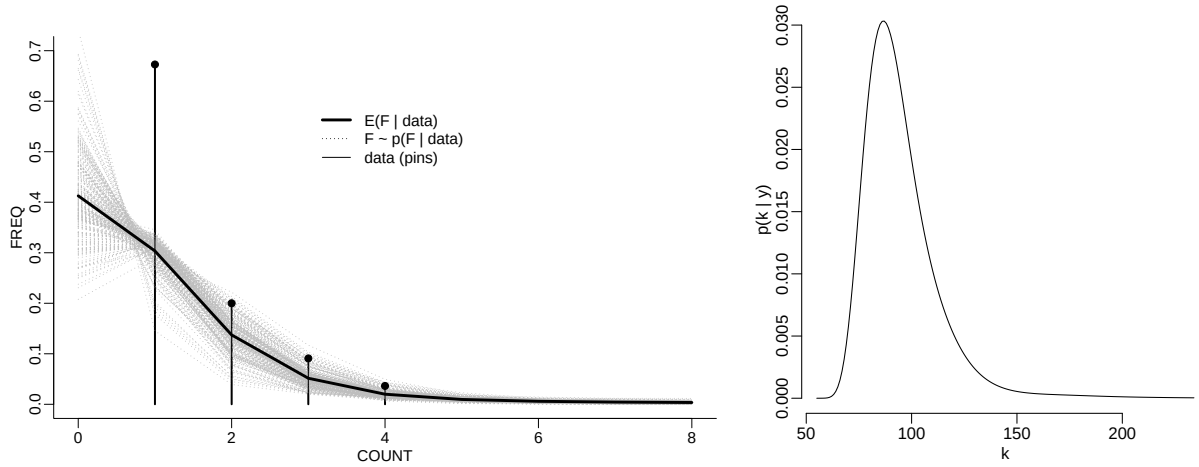
important characteristic of the strength of the immune system.

Table 1 shows the observed data y_i for one of the experiments reported in Guindani et al. (2012, Table 2). There are $n = 55$ distinct T-cell receptor sequences. The total number of recorded T-cell receptor sequences is $\sum_{i=1}^{4} i \cdot y_i = 82$.

Figure 1 shows the empirical distribution $\widehat{F}(y_i)$ together with a BNP estimate $E(F \mid y)$. Inferene on $F(\cdot)$ allows imputation of $N - n$, the number of zero-censored T-cells. A parametric model $F_\theta(y)$, like a simple Poisson model or a finite mixture of Poisson models would report misleadingly precise inference for θ – and thus $F_\theta(0)$ – based on the likelihood $p(\boldsymbol{y} \mid \theta) = \prod_{i=1}^{n} F_\theta(y_i)/(1 - F_\theta(0))$. Guindani et al. (2012) use instead a Dirichlet process (DP) mixture of Poisson model for F. We discuss details below. Figure 1b shows the posterior distribution $p(N \mid y)$ under the same model that was used for the posterior inference in Figure 1a.

The DP prior (Ferguson, 1973) is arguably the most commonly used BNP prior. We write $G \sim \mathrm{DP}(\alpha, G^\star)$ for a DP prior on a random probability measure G. The model uses two parameters, the total mass parameter α and the base measure $G^\star$. The base measure specifies the mean, $E(G) = G^\star$. The total mass parameter determines, among other implications, the uncertainty of G. Consider any (measureable) set A. Then the probability $G(A)$ under G is a beta distributed random variable with $G(A) \sim \mathrm{Be}\left[\alpha G^\star(A), \alpha(1 - G^\star(A))\right]$. Similarly, for any partition $\{A_1, A_2, \ldots, A_K\}$ of the sample space S, i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$ and $\bigcup_{k=1}^{K} A_k = S$, the vector of random probabilities $(G(A_1), \ldots, G(A_K))$ follows a Dirichlet distribution, $p(G(A_1), \ldots, G(A_K)) \sim \mathrm{Dir}(\alpha G^\star(A_1), \ldots, \alpha G^\star(A_K))$. This property is a defining characteristic. Alternatively the DP prior can be defined as follows. Let $\delta_x(\cdot)$ denote a point mass at x. Then G is a discrete probability measure

$$G(\theta) = \sum_{h=1}^{\infty} \pi_h \delta_{\widetilde{\theta}_h} \tag{1}$$



(a) $p(G \mid y)$ (b) $p(N \mid y)$

Figure 1: T-cell diversity. Figure (a) shows the data (as pin plot) and a posterior sample $F \sim p(F \mid y)$ under a DP mixture prior (grey curves) and the posterior estimate $\overline{F} = E(F \mid y)$ (black curve). The plotted curves connect the point masses $F(i)$ and $\overline{F}(i)$ for better display (the connection itself is meaningless). Panel (b) shows the implied posterior $p(N \mid y)$ on the total number of T-cell types.

with $\widetilde{\theta}_h \sim G^\star$, i.i.d., and $\pi_h = v_h \prod_{\ell<h}(1 - v_\ell)$ for $v_h \sim \text{Be}(1, \alpha)$, i.i.d. The constructive definition (1) is known as the stick-breaking representation of the DP prior (Sethurman, 1994). For a recent discussion of the DP prior and basic properties see for example Ghoshal (2010). An excellent review of several alternative constructions of the DP prior is included in Lijoi and Prünster (2010).

Implicit in this constructive definition is the fact that a DP random measure is a.s. discrete and can be written as a sum of point masses $\widetilde{\theta}_h$. In many applications the a.s. discreteness of the DP is awkward. For example, in a density estimation problem, $y_i \sim G(\cdot)$, $i = 1, \ldots, n$, it would be inappropriate to assume $G \sim \text{DP}$ if the distribution were actually known to be absolutely continuous. A simple model extension fixes the awkward discreteness by assuming $y_i \sim F$ and

$$F(y) = \int p(y \mid \theta)\, dG(\theta) \text{ with } G \sim \text{DP}(\alpha, G_0). \tag{2}$$

In words, the unknown distribution is written as a mixture with respect to a mixing measure with DP prior. Here $p(y \mid \theta)$ is some model indexed by θ. The model is known as the DP mixture (DPM) model. If desired, a continuous distribution $p(y \mid \theta)$ creates a continuous probability measure F. Often the mixture is written as an equivalent hierarchical model, by

introducing latent variables θ_i:

$$\begin{aligned} p(y_i \mid \theta_i) &\sim p(y_i \mid \theta_i) \\ \theta_i &\sim G \\ G &\sim \mathrm{DP}(\alpha, G^\star). \end{aligned} \tag{3}$$

Marginalizing with respect to θ_i, model (3) reduces again to $y_i \sim \int p(y \mid \theta)\, dG(\theta)$, i.i.d., as desired.

Example 1 (ctd.) *Let Poi$(y;\ \theta)$ denote a Poisson model with parameter θ. In Guindani et al. (2012) we use a DPM model with $p(y \mid \theta) = Poi(y;\ \theta)$. Here the motivation for the DPM is the flexibility compared to a simpler parametric family. Also, the latent variables θ_i that appear in the hierarchical model* (3) *are attractive for this application to inference for T-cell diversity. The latent θ_i become interpretable as mean abundance of T-cell type i. The use of the BNP model for $F(\cdot)$ addressed several key problems in this inference problem. The BNP model allowed the critical extrapolation to $F(0)$ without relying on a particular parametric form of the extrapolation. And equally important, the extrapolation is based on a coherent probability model. The latter is important for the derived inference about N. In Figure 1a, the grey curves illustrate the posterior distribution $p(F \mid y)$. The implied histogram of $F(y)$ at $y = 0$ estimates $p(F(0) \mid \boldsymbol{y})$ and it implies in turn the posterior distribution $p(N \mid \boldsymbol{y})$ for the primary inference target that is shown in Figure 1b. Implementing the same inference in a parametric model would be challenging.*

In the context of inference for SAGE (serial analysis of gene expression) data Morris et al. (2003) use parametric inference for similar data. However, in their problem estimation of N is not the primary inference target. Their main aim is to estimate the unknown prevelance of the different species, equivalent to estimating $F(i)$, $i \geq 1$ in the earlier description.

One of the attractions of DP (mixture) models is easy computation, including the availability of R packages. For example, posterior inference for DP mixture models and many extensions is implemented in the R package *DPpackage* (Jara et al., 2011).

2.2 Polya Tree Priors

Many alternatives to the DP(M) prior for a random probability measure G have been proposed in the BNP literature. Especially for univariate and low-dimensional distributions the Polya urn prior (Lavine, 1992, 1994; Mauldin et al., 1992) is attractive. It requires no additional mixture to create absolutely continous probability measures.

The construction is straightforward. Without loss of generality assume that we wish to construct a random probability measure $G(y)$ on the unit interval, $0 \leq y \leq 1$. Essentially we construct a random histogram. We start with the simplest histogram, with only two bins by splitting the sample space into two subintervals B_0 and B_1 and assigning random probability

$$Y_0 \equiv G(B_0) \sim \text{Be}(a_0, a_1).$$

and $Y_1 = G(B_1) = 1 - Y_0$ to the two intervals, using a beta prior to generate the random probability Y_0. Next we refine the histogram by splitting B_0 in turn into two subintervals $B_0 = B_{00} \cup B_{01}$ and similar for $B_1 = B_{10} \cup B_{11}$. We use the random splitting probabilities $Y_{00} \equiv G(B_{00} \mid B_0) \sim \text{Be}(a_{00}, a_{01})$ and $Y_{10} \equiv G(B_{10} \mid B_1) \sim \text{Be}(a_{10}, a_{11})$. Again let $Y_{01} = 1 - Y_{00}$ etc. Let $\epsilon = \varepsilon_1 \cdots \varepsilon_m$ denote a length m binary sequence. After m refinements we have a partition $\{B_{\varepsilon_1 \cdots \varepsilon_m};\ \varepsilon_j \in \{0, 1\}\}$ of the sample space with

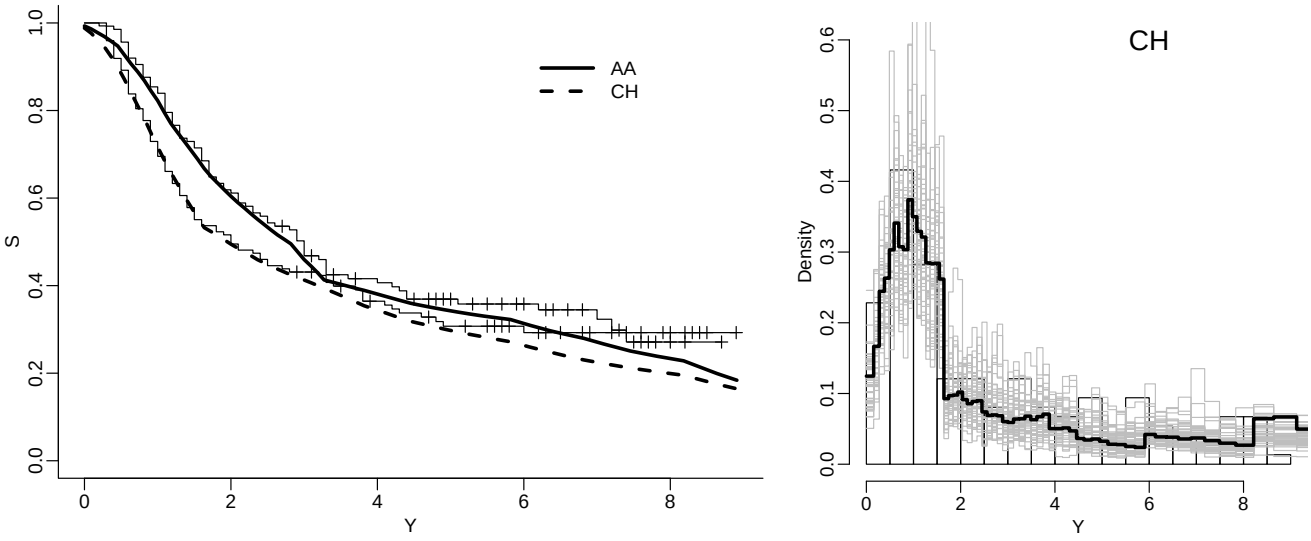
$$G(B_\epsilon) = \prod_{j=1}^{m} Y_{\varepsilon_1 \cdots \varepsilon_j}$$

In summary, the Polya tree prior is indexed by the nested sequence of partitions $\Pi = \{B_\epsilon\}$ and the beta parameters $\mathcal{A} = \{a_\epsilon\}$. We write $G \sim \text{PT}(\Pi, \mathcal{A})$. One of the attractions of the PT prior is the easy prior centering at a desired prior mean $G^\star$. Let q_a denote the quantile with $G^\star\{[0, q_a]\} = a$. Fix $B_0 = [0, q_{1/2})$, $B_1 = [q_{1/2}, 1]$, $B_{00} = [0, q_{1/4}), B_{01} = [q_{1/4}, q_{1/2}), \ldots, B_{11} = [q_{3/4}, 1], B_{000} = [0, q_{1/8})$, etc. In other words, we fix Π as the dyadic quantiles of the desired $G^\star$. If additionally $\alpha_\epsilon = c_m$ is constant across all level m subsets, then $E(G) = G^\star$, as desired. Alternatively, for arbitrary Π, fixing $a_{\epsilon x} = c_m G^\star(B_{\epsilon x})/G^\star(B_\epsilon)$, $x = 0, 1$ also implies $E(G) = G^\star$. With a slight abuse of notation we write $G \sim \text{PT}(G^\star, \mathcal{A})$ and $G \sim \text{PT}(\Pi, G^\star)$ to indicate that the partition sequence or the beta parameters are fixed to achieve a prior mean $G^\star$. A popular choice for c_m is $c_m = cm^2$, which guarantees an absolutely continous random probability measure G (Lavine, 1992). On the other hand, with $a_\epsilon = \alpha G^\star(B_\epsilon)$, and thus $a_\epsilon = a_{\epsilon 0} + a_{\epsilon 1}$, the PT reduces to a $\text{DP}(\alpha, G^\star)$ prior with an a.s. discrete random probability measure G.

Example 2 (Prostate cancer study) *Zhang et al. (2010) use a PT prior to model the distribution of time to progression (TTP) for prostate cancer patients. The data are available in the on-line supplementary materials for this paper (TTP, treatment indicator and censoring status). The application also includes a regression on a longitudinal covariate and a possible cure rate. For the moment we only focus on the PT prior for the survival time. The study includes two treatment arms, androgen ablation (AA) and androgen ablation plus three 8-week cycles of chemotherapy (CH). Zhang et al. (2010) used a PT prior to*

model time to progression y_i for $n_1 = 137$ patients under CH and for $n_2 = 149$ under AA treatment. Let G_1 and G_2 denote the distribution of time to progression under CH and AA treatment, respectively. We assume $G_j \sim PT(G^\star_\beta, \mathcal{A})$ with $a_\epsilon = cm^2$ and centering measure $G^\star_\beta = Weib(\tau, \beta)$, a Weibull distribution with $\tau = 4.52$ and $\beta = 1.23$.

Figure 2a shows the data as a Kaplan-Meier plot together with the posterior estimated survival functions. Inference for G_1 and G_2 is under $G_j \sim PT(G^\star_\beta, \mathcal{A})$, independently across G_1 and G_2, for fixed $G^\star_\beta$. Inference in Zhang et al. (2010) is based on a larger model that



(a) $p(G_j \mid y)$ as survival function

(b) $p(G_2 \mid y)$ as p.d.f.

Figure 2: The horizontal axis indicates years after treatment. In panel (a), the step function shows the KM estimates (with censoring times marked as +). The solid line, and the dashed line are estimates based on the BNP model. Panel (b) shows the posterior $p(G_2 \mid \boldsymbol{y})$ and the posterior mean $E(G_2 \mid \boldsymbol{y})$ (thick line).

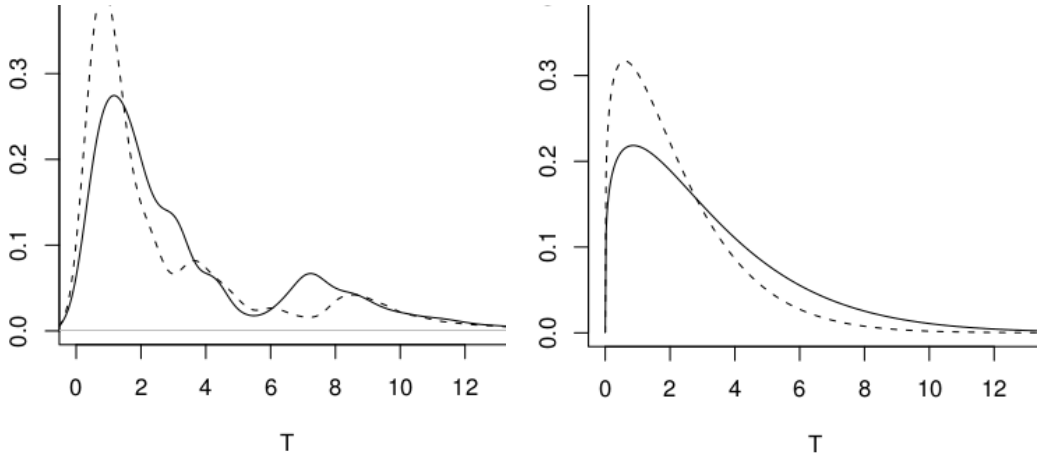
includes the PT prior as a submodel for the TTP event times. Additionally, the model adds the possibility of patients being cured of the disease, i.e., the model replaces i.i.d. sampling of TTP's $T_{ji} \sim G_j$ by a hierarchical model with $p(w_{ji} = 1) = p_j$ and $p(T_{ji} \mid w_{ji} = 0) = G_j$ where w_{ji} is an indicator for a patient under treatment j being cured, and p_j is the cured fraction under treatment j. Also, the model includes an additional regression on a longitudinal covariate $y_{ji} = (y_{jik},\ k = 1, \ldots, n_{ji})$ (prostate specific antigen, PSA). For the implementation of inference on these two additional model features it is important that posterior inference on G_j remain flexible and be fully model-based. In particular, inference on the tails of G_j immediately impacts inference on the cured fractions, as it speaks to the

possibility of possible (latent) later TTP beyond the censoring time. The full description of uncertainties is equally important for the regression on longitudinal PSA measurements. The imputed G_j is used to impute latent TTP values for susceptible (susceptible) patients. Imputed large TTP's could easily become influential outliers in the regression problem. Figure 3 shows inference on G_j, now also including the cured fraction and the regression on PSA. See Zhang et al. (2010) for details on the implementation of the regression model. The secondary mode around $T = 8$ is interesting from a clinical perspective, but would be almost impossible to find with parametric inference. It was not revealed by the initial Kaplan-Meier plot. A parametric model can only accomodate such features if the possibility of a second mode were anticipated in the model construction. But this is not the case here.

A minor concern with inference under the PT prior in some applications is the dependence on the chosen partition sequence. Figure 2b shows inference for G_1, represented by its probability density function. The partition boundaries are clearly visible in the inference. This is due to the fact that the PT prior assumes independent conditional splitting probabilities Y_ϵ, independent across m and across the level m partitioning subsets. The same independence persists a posteriori. There is no notion of smoothing inference on the splitting probabilities across partitioning subsets. This awkward dependence on the boundaries of the partitioning subsets can easily be mitigated by defining a mixture of PTs (MPT), mixing with respect to the parameters of the centering measure $G^\star_\eta$. Let η denote the parameters of the centering measure $G^\star_\eta$. In the example, $\eta = (\tau, \beta)$. We augment the model by adding a hierarchical prior $p(\beta)$, leaving τ fixed. This leads to an MPT model, $G_j(\cdot) = \int \mathrm{PT}(G_j;\ G^\star_\eta, \mathcal{A})\ dp(\eta)$. Here $\mathrm{PT}(G;\ G^\star, \mathcal{A})$ indicates a PT prior for the random probability measure G, with the nested partition sequence defined by dyadic quantiles of $G^\star$ and beta parameters $\mathcal{A}$. Such MPT constructions were introduced in Hanson (2006) and in a variation in Paddock et al. (2003). In the prostate cancer data, the estimated survival curve remains practically unchanged from Figure 2a. The p.d.f. is smoothed (not shown).

Branscum et al. (2008) report another interesting use of PT priors. They implement inference for ROC (receiver operating characteristic) curves based on two independent PT priors for the distribution of scores under the true positive (G_1) and true negative population (G_0), respectively. In this application uncertainty about G_j is critical. A commonly reported summary of the ROC curves is the area under the curve (AUC), which can be expressed as $\mathrm{AUC} = p(X > Y)$ for $X \sim G_1$ and $Y \sim G_0$. Complete description of all uncertainties in the two probability models is critical for the estimate of the ROC curve. A fortiori, it is critical for a fair characterization of uncertainties about the ROC curve. The latter becomes important, for example, in biomarker trials (Pepe et al., 2001). Uncertainty about

the ROC curve in an earlier still exploratory trial is used to determine the sample size for a later prospective validation trial. A complete description of uncertainties is critical in such applications.



(a) $E(G_j \mid y)$ for $j = 1, 2$ (b) Parametric m.l.e., G_θ for $\theta = \widehat{\theta}_j$.

Figure 3: Prostate cancer study. Inference from Zhang et al. (2010) on G_j. The model included a cured fraction and a regression on a longitudinal covariate (prostate specific antigen) in addition to $y_{ji} \sim G_j$.

Finally, a brief note on computation. Posterior updating for a PT prior is straightforward. It is implemented in the R package *DPpackage* (Jara et al., 2011). The definition of a PT prior for a multivariate random probability measure requires a clever definition of the nested partition sequence and can become cumbersome in higher dimensions. Hanson and Johnson (2002) proposes a practicable construction for multivariate PT construction centered at a multivariate normal model.

2.3 More Random Probability Measures

Many alternative priors $p(G)$ for random probability measures have been proposed. Many can be characterized as natural generalizations or simplifications of the DP prior. Ishwaran and James (2001) propose generalizations and variations based on the stick-breaking definition (1). The finite DP is constructed by truncating (1) after K terms, with $v_K = 1$. The truncated DP is particularly attractive for posterior computation. Ishwaran and James (2001) show a bound on the approximation error that arises when using inference under the a truncated DP to approxiamte inference under the corresponding DP prior. The beta

priors for v_h in (1) can be replaced by any alternative $v_h \sim \text{Be}(a_h, b_h)$, without complicating posterior simulations. In particular, $v_h \sim \text{Be}(1-a, b+ha)$ defines the Pitman Yor process with parameters a, b (Pitman and Yor, 1997).

Alternatively we could focus on other defining properties of the DP to motivate generalizations. For example, the DP can be defined 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 measureable subsets $A_1, \ldots, A_k$ of the desired sample space. The defining property of the CRM μ is that the $\mu(A_j)$ be mutually independent. The gamma process is a CRM with $\mu(A_j) \sim \text{Ga}(M\mu_0(A), 1)$, mutually independent, for a probability measure μ_0 and $M > 0$. Normalizing μ by $G(A) = \mu(A)/\mu(S)$ defines a DP prior with base measure proportional to μ_0. Replacing the gamma process by any other CRM defines alternative BNP priors for random probability measures.

Such priors are known as normalized random measures with independent increments (NRMI) and were first described in Regazzini et al. (2003) and include a large number of BNP priors. A recent review of NRMI's appears in Lijoi and Prünster (2010). Besides the DP prior other examples are the normalized inverse Gaussian (NIG) of Lijoi et al. (2005) and the normalized generalized gamma process (NGGP), discussed in Lijoi et al. (2007). The construction of the NIG in many ways parallels the DP prior. Besides the definition as a CRM, a NIG process G can also be characterized by a normalized inverse Gaussian distribution (Lijoi et al., 2005) for the joint distribution of random probabilities $(G(A_1), \ldots, G(A_k))$, and like for the DP the probabilities for cluster arrangements defined by ties under i.i.d. sampling are available in closed form. For the DP, we will still consider this distribution in more detail in the next section. The NIG, as well as the DP are special cases of the NGGP.

Two recent papers (Barrios et al., 2013; Favaro and Teh, 2013) describe practicable implementions of posterior simulation for mixtures with respect to arbitrary NRMIs, based on a characterization of posterior inference in NRMIs discussed in James et al. (2009) who characterize $p(G \mid \boldsymbol{y})$ under i.i.d. sampling $y_i \sim G$, $i = 1, \ldots, n$, from a random probability measure G with NRMI prior. Both describe algorithms specifically for the NGGP. Both use conditioning on the same latent variable U that is introduced as part of the description in James et al. (2009). Favaro and Teh (2013) describe what can be characterized as a modified version of the Pólya urn. The Pólya urn defines the marginal distribution of $(y_1, \ldots, y_n)$ under the DP prior, after marginalizing with respect to G. We shall discuss the marginal model under the DP in more detail in the following sesction. Barrios et al. (2013) describe

an approach that includes sampling of the random probability measure. This is particularly useful when desired inference summaries require imputation of the unknown probability measure. The methods of Barrios et al. (2013) are implemented in the R package *BNPdensity*, which is available in the CRAN package repository (http://cran.r-project.org/).

3 Clustering

3.1 DP Partitions

The DP mixture prior (3) and variations are arguably the most popular BNP priors for random probability measures. The popularity is mainly due to perhaps two reasons. One is computational simplicity. In model (3) it is possible to analytically marginalize with respect to G, leaving a model in θ_i only. This greatly simplifies posterior inference. The second, and related, reason is the implied clustering. As samples from a discrete probability measure G the latent θ_i include many ties. One can use the ties to define clusters. Let $\theta_k^\star$, $k = 1, \ldots, K$ denote the $K \leq n$ unique values among the θ_i, $i = 1, \ldots, n$. Then $S_k = \{i : \ \theta_i = \theta_k^\star\}$, $k = 1, \ldots, K$, defines a random partition of the experimental units $\{1, \ldots, n\}$. Let $\rho_n = \{S_1, \ldots, S_K\}$ denote the random partition. Sometimes it is more convenient to use an alternative description of the partition in terms of cluster membership indicators $\boldsymbol{s} = (s_1, \ldots, s_n)$ with $s_i = k$ if $i \in S_k$. We add the convention that clusters are labeled in the order of appearance, in particular $s_1 = 1$. One of the attractions of the DP prior is the simple nature of the implied prior $p(\rho_n)$. Let $n_k = |S_k|$ denote the size of the $k-$th cluster. Then

$$p(\rho_n) \propto \alpha^K \prod_{k=1}^{K} (n_k - 1)! \tag{4}$$

implying in particular the following complete conditional prior. We write $\boldsymbol{s}^-$ for $\boldsymbol{s}$ without s_i, n_k^- for the size of S_k without unit i, etc.

$$p(s_i = k \mid \boldsymbol{s}^-) \propto \begin{cases} n_k^- & \text{for } k = 1, \ldots, K^- \\ \alpha & \text{for } k = K^- + 1. \end{cases} \tag{5}$$

Here $s_i = K^- + 1$ indicates that i forms a new $(K^- + 1)$-st singleton cluster of its own. The probability model (5) is known as the Pólya urn.

Many applications of the popular DPM exploit the implied prior $p(\rho_n)$ in (4). Often the random probability measure G itself is not of interest. The model is only introduced for the side effect of creating a random partition. In such applications the use of the DP prior can be

questioned, as the prior $p(\rho_n)$ includes several often inappropriate features. The cluster sizes are a priori geometrically ordered, with one large cluster and geometrically smaller clusters, including many singleton clusters. However, this is less of a concern when either prediction is the focus or only major clusters are interpreted.

BNP inference on ρ_n offers some advantages over parametric alternatives. A parametric model might induce clustering of the experimental units, for example, by specifying a mixture model with J terms, $y_i \sim \sum_{j=1}^{J} w_j p_j(y_i \mid \theta_j)$. Replacing the mixture model by a hierarchical model, $p(y_i \mid s_i = j, \theta_j) = p_j(y_i \mid \theta_j)$ and $p(s_i = j) = w_j$ with latent indicator variables s_i implicitely defines a random partition by interpreting the indicators s_i as cluster membership indicators. Such random partition models are known as model based clustering or mixture of experts models (when the weights are allowed include a regression on covariates). In contrast to the nonparametric prior, the parametric model requires to specify the size J of the mixture, either by fixing it or by extending the hierarchical model with a hyperprior on J.

Recall the definition of BNP models as probability models on infinite dimensional random elements. However, there are only finitely many partitions ρ_n, leaving the question why random partition models should be considered BNP models. Traditionally they are. Besides tradition, perhaps another reason is a one-to-one correspondence between an exchangeable random partition and a discrete probability measure (Pitman, 1996, Proposition 13). An exchangeable random partition $p(\rho_n)$ can always be thought of as arising from the configuration of ties under i.i.d. sampling from a discrete probability measure.

3.2 Hierarchical Extensions

An interesting class of extensions of the basic DP model define hierarchical models and other extensions to multiple random probability measures. One of the earlier extensions was the hierarchical DP (HDP) of Teh et al. (2006), who define a prior for random probability measures G_j, $j = 1, \ldots, J$, with $G_j \mid G^\star \sim \mathrm{DP}(M, G^\star)$, independently. By completing the model with a prior on the common base measure, $G^\star \sim \mathrm{DP}(B, H)$, they define a joint probability model for $(G_1, \ldots, G_J)$. Importantly, the discrete nature of the $G^\star$ as a DP random measure itself introduces positive probabilities for ties in the atoms of the random G_j, and thus the possibility of ties among samples $\theta_{ij} \sim G_j$, $i = 1, \ldots, n_j$, and $j = 1, \ldots, J$. We could again use these ties to define a random partition. Let $\{\theta_k^{\star\star},\ k = 1, \ldots, K\}$ denote the unique values among the θ_{ij} and define clusters $S_k = \{(ji) :\ \theta_{ij} = \theta_k^{\star\star}\}$. This defines random clusters of experimental units across j. In summary, the HDP generates random probability measures G_j that share the same atoms across j. However, the random distributions G_j

are different. The common atoms have different weights under each G_j. This distinguishes the HDP from the related nested DP (NDP) of Rodríguez et al. (2008). The NDP allows for some of the G_j to be identical. While the HDP uses a common discrete base measure $G^\star$ to generate the atoms in the G_j's, the NDP uses a common discrete prior $Q(G_j)$ for the distributions G_j themselves, thus allowing $p(G_j = G_{j'}) > 0$ for $j \neq j'$. The prior for Q is a DP prior whose base measure has to generate random probability measures which serve as the atoms of Q. Another instance of a DP prior is used for this purpose. In summary, $G_j \sim Q$ and $Q \sim \mathrm{DP}(M, \mathrm{DP}(\alpha, G^\star))$. Another related extension of the DP is the enriched DP of Wade et al. (2011).

3.3 More Random Partitions

Several alternatives to DP priors for random partitions have been discussed in the literature. The special feature of the DP prior is the simplicity of (5). While any discrete random probability measure gives rise to a prior $p(\rho_n)$, few are as simple as (5). The already mentioned Pitman-Yor process (Pitman and Yor, 1997) implies very similar conditionals for s_i, with

$$p(s_i = k \mid \boldsymbol{s}^-) \propto \begin{cases} n_k^- - \beta & k = 1, \ldots, K^- \\ \alpha + \beta K^- & k = K^- + 1 \end{cases}$$

where $0 < \beta < 1$ and $\alpha > -\beta$. See also the discussion in Ishwaran and James (2001). Similarly, any NRMI defines a prior $p(\rho_n)$. For the NGGP Lijoi et al. (2007) give explicit expressions for $p(s_i = k \mid \boldsymbol{s}^-)$. They discuss the larger family of Gibbs-type priors as a class of priors $p(\rho_n)$ that include the one implied under the NGGP as a special case. While the simple nature of (5) is computationally attractive, it can be criticized for lack of structure. For example, the conditional probabilities for cluster membership depend only on the sizes n_k^- of the clusters, not on the number of clusters or the distribution of cluster sizes. For a related discussion see also Quintana (2006) and Lee et al. (2013).

For alternative constructions of $p(\rho_n)$ we could focus on the form of (4) as a product over functions $c(S_k) = \alpha(n_k - 1)!$ that depend on only one cluster at a time. Random partition models of the form $p(\rho_n) \propto \prod_{k=1}^K c(S_k)$ for some function $c(S_k)$ are known as product partition models (PPM) (Hartigan, 1990). Together with a sampling model that assumes independence across clusters the posterior $p(\rho_n \mid \boldsymbol{y})$ is again of the same form.

Müller et al. (2011) define a variation of the PPM by explicitely including covariates. Let x_i denote covariates, let y_i denote responses, and let $x_k^\star = \{x_i;\ i \in S_k\}$ denote covariates arranged by clusters. The goal is to a priori favor partitions with clusters that are more homogeneous in x. Posterior predictive inference allows then to define regression based on

clustering. We define a function $g(x_k^\star) \geq 0$ such that $g(x^\star)$ is large for a set of covariates $x_k^\star$ that are judged to be very similar, and small when $x^\star$ includes a diverse set of covariate values. The definition of $g(\cdot)$ is problem-specific. For example, for a categorical covariate $x_i \in \{1, \ldots, Q\}$, let m_k denote the number of unique values x_i in cluster k. and we could use $g(x_k^\star) = 1/m_k$. A cluster with all equal x_i has the highest similarity. Müller et al. (2011) define the PPMx model

$$p(\rho_n \mid x) \propto \prod_{k=1}^{K} c(S_k) g(x_k^\star).$$

Example 3 (Sarcoma trial) *Leon-Novelo et al. (2012) consider clustering of different sarcoma types. Table 2 shows data from a phase II clinical trial with sarcoma patients. Sarcoma*

Table 2: Number of patients n_i and number of responses y_i for sarcoma subtypes $i = 1, \ldots, n$.

Intermediate Prognosis		
subtype	n_i	y_i
Leiomyosarcoma	28	6
Liposarcoma	29	7
MFH	29	3
Osteosarcoma	26	5

Intermediate (ctd.)		
subtype	n_i	y_i
Synovial	20	3
Angiosarcoma	15	2
MPNST	5	1
Fibrosarcoma	12	1

Good Prognosis		
subtype	n_i	y_i
Ewing's	13	0
Rhabdo	2	0

is a rare type of cancer affecting connective or supportive tissues and soft tissue (e.g., cartilage and fat). There are many subtypes of sarcomas, reflecting the definition of sarcomas as cancers of a large variety of tissues. Some subtypes are very rare, making it attractive to pool patients across subtypes. Leon-Novelo et al. (2012) propose to pool patients on the basis of a random partition of subtypes. Keeping the clustering of subtypes random acknowledges the uncertainty about the different nature of the subtypes. However, in setting up a prior probability model for the random partition of subtypes, not all subtypes are exchangeable. For example, some are known to have better prognosis than others. Leon-Novelo et al. (2012) exploit this information. Let $x_i \in \{-1, 0, 1\}$ denote an indicator of poor, intermediate or good prognosis for subtype i. We define a prior model $p(\rho_n \mid x_1, \ldots, x_n)$ with increased probability of including any two subtypes of equal prognosis in the same cluster. Let $Q = 3$ denote the number of different prognosis types, and let m_{kq} denote the number of $x_i = q$ for all $i \in S_k$ and $n_k = \sum_q m_{kq}$ the size of the $k-th$ cluster. We use the similarity function $d(x_k^\star) = \prod_{q=1}^{Q} m_{kq}!/(Q-1)!\,/(n_k + Q - 1)!$ The particular choice is motivated mainly

by computational convenience.[1] *It allows a particularly simple posterior MCMC scheme. Müller et al. (2011) argue that with this choice of* $d(x_k^\star)$ *the posterior distribution on* ρ_n *is identical to the posterior in a DPM model under a model augmentation, and thus any MCMC scheme for a DPM models can be used. The important feature, however, is the increased probability for homogeneous clusters. Let* $Bin(y;\ n, \pi)$ *denote a binomial probability distribution for the random variable* y *with binomial sample size* n *and success probability* π*. Let* $\boldsymbol{\pi}^\star = (\pi_k^\star,\ k = 1, \ldots, K)$ *denote cluster specific success rates. Conditional on* ρ_n *and* $\boldsymbol{\pi}^\star$ *we assume a sampling model* $p(y_i \mid s_i = k, \pi_k^\star) = Bin(y_i;\ n_i, \pi_k^\star)$*. The probability model is completed with a conjugate prior for the cluster-specific success rates* $\pi_k^\star$*. Let* $\pi_i = \pi_{s_i}^\star$ *denote the success probability for sarcoma type* i*. Figure 4 shows posterior means and 90% credible*

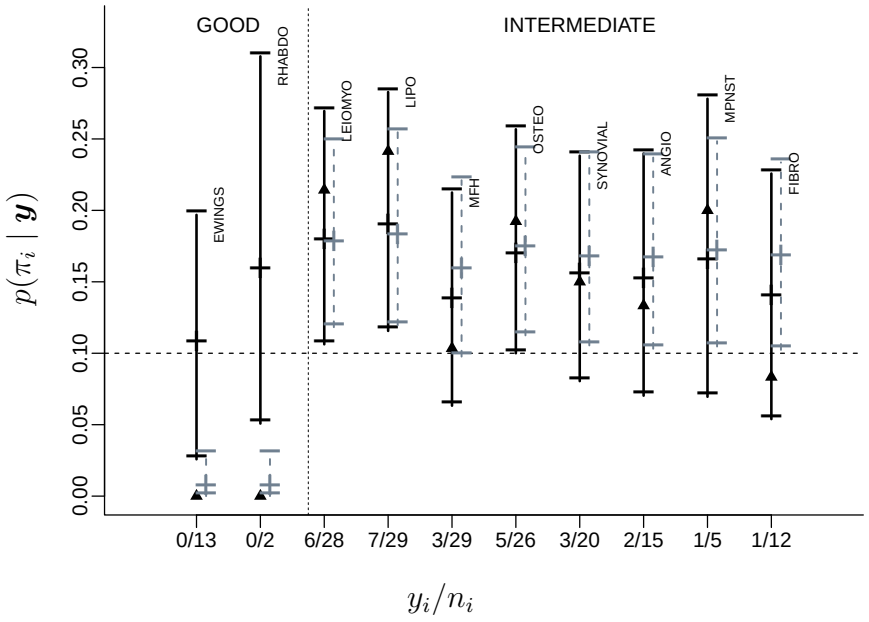


Figure 4: Central 90% posterior credible intervals of success probabilities π_i for each sarcoma subtype under the BNP model (black lines) and under a comparable parametric model (grey lines). The central marks ("+") are the posterior means, the triangles are the m.l.e.'s.

intervals for π_i *by sarcoma type, compared with inference under an alternative partially exchangeable model, i.e., a model with separate submodels for* $x_i = -1, 0$ *and* 1*. Notice how the BNP model strikes a balance between the separate analysis of a partially exchangeable model and the other extreme which would pool all subtypes. In summary, the use of the BNP model*

[1] Leon-Novelo et al. (2012) use $d(x_k^\star)$ without the $(Q-1)! = 2$ factor, which, however, in the light of (4) is equivalent to simply rescaling α by 2.

here allowed to borrow strength across the related subpopulations while acknowledging that it might not be appropriate to pool all.

A practical problem related to posterior inference for random partitions is the problem of summarizing $p(\rho_n \mid \boldsymbol{y})$. Many authors report posterior probabilities of co-clustering. Let $d_{ij} = I(s_i = s_j)$ and define $D_{ij} = p(d_{ij} = 1 \mid \boldsymbol{y})$. Dahl (2006) went a step further and introduced a method to obtain a point estimate of the random clusters based on least-square distance from the matrix of posterior pairwise co-clustering probabilities. Quintana and Iglesias (2003) address the problem of summarizing $p(\rho_n \mid \boldsymbol{y})$ as a formal decision problem.

3.4 Feature Allocation Models

In many applications the strict mutual exclusive nature of the cluster sets in a partition is not appropriate. For example, in an application to find sets of proteins that correspond to some common biologic processes, one would want to allow for some proteins to be included in multiple sets, i.e., to be involved in more than one process. Such structure can be modeled by feature allocation models. For example, the Indian buffet process (IBP) (Griffiths and Ghahramani, 2006) defines a prior for a binary random matrix whose entries can be interpreted as membeship of proteins (rows) in protein sets (columns). Ghahramani et al. (2007) review some applications of the IBP. An excellent recent discussion of such models and how they generalize random partition models appears in Broderick et al. (2013).

4 Regression

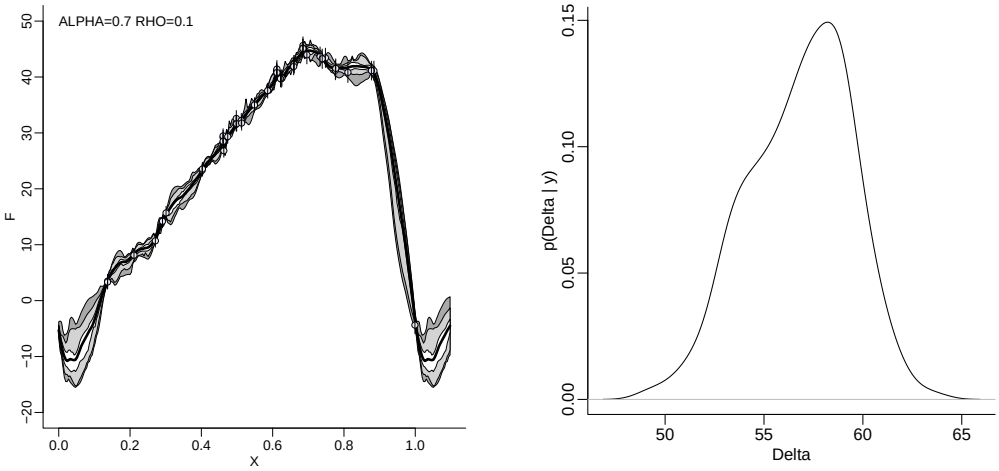
4.1 Nonparametric Residuals

Consider a generic regression problem $y_i = f(x_i) + \epsilon_i$ with responses y_i, covariates x_i and residuals $\epsilon_i \sim p(\epsilon_i)$ for experimental units $i = 1, \ldots, n$. In a parametric regression problem we assume that the regression mean function $f(\cdot)$ and the residual distribution $p(\cdot)$ are indexed by a finite dimensional parameter vector, $f(x) = f_\theta(x)$ and $p(\epsilon) = p_\theta(\epsilon)$. Sometimes a parametric model is too restrictive and we need nonparametric extensions. The earlier stylized description of a regression problem suggests three directions of such model extensions. We could relax the parametric assumption on $f(\cdot)$, or go nonparametric on the residual distribution $p_\theta(\cdot)$, or on both. We refer to the first as BNP regression with a nonparametric mean function, the second as a nonparametric residual model and the combination as a fully nonparametric BNP regression or density regression.

Hanson and Johnson (2002) discuss an elegant implementation of a nonparametric residual model. Assuming $\epsilon_i \sim G$ and a nonparametric prior $G \sim p(G)$ reduces the problem to essentially the earlier discussed density estimation problem. The only difference being that now the i.i.d. draws from G are the latent residuals ϵ_i. In principle, any model that was used for density estimation could be used. However, there is a minor complication. To maintain the interpretation of ϵ_i as residuals and to avoid identifiability concerns, it is desireable to center the random G at zero, for example, with $E(G) = 0$ or median 0. Hanson and Johnson (2002) cleverly avoid this complication by using a PT prior. The PT prior allows simple centering of G by fixing $B_0 = (-\infty, 0]$ and $Y_0 = \frac{1}{2}$, thus fixing the median at 0.

4.2 Nonparametric Mean Function

Example 4 (Cepheid data) *Barnes et al. (2003) discuss an application of BNP regression to analyze data from Cepheid variable stars. Cepheid variable stars serve as mile stones,*



(a) Phase-velocity curve $f(x)$ (b) $p(\max(f) - \min(f) \mid \boldsymbol{y})$

Figure 5: Phase-velocity curve $f(x)$ for T. Moncertois. Panel (a) shows the posterior estimated phase-velocity curve $E(f \mid y)$ (thick central line), and pointwise central HPD 50% (light grey) and 95% (dark grey) intervals for $f(x)$. The circles shows the data points. Inference is under a BNP model $p(f)$ using a basis expansion of f with wavelets. Panel (b) shows posterior inference on the range $\Delta = \max(f) - min(f)$.

or standard candles, to establish distances to galaxies in the universe. This is because the luminosities for these stars are highly correlated with their pulsation periods, allowing indirect

measurement of a Cepheid star's luminosity (light output), which in turn can be related to the observed brightness to infer the distance. Calibration of the luminosty-period relation involves a non-linear relationship that includes among others the integral ΔR of radial velocity with respect to phase. Figure 5a plots radial velocity versus phase for the Cepheid variable star T Moncerotis. The circles indicate the observed data points. The short vertical line segments show the (known) measurement error standard deviation. The periodic nature of the data adds a constraint $f(0) = f(1)$ for the phase-velocity curve $f(x)$. The sparse data around $x = 0$ makes it difficult to determine the regression mean function around $x = 0$. The many data points in other parts of the curve mislead a parametric model to believe in precisely estimated parameters, including the critical interpolation around $x = 0$. We therefore consider a non-parametric regression.

A convenient way to define a BNP prior for an unknown regression mean function is the use of a basis representation. Let $\{\varphi_j\}$ denote a basis, for example, for square integrable functions. Any function of the desired function space can be represented as

$$f(\cdot) = \sum_h d_h \varphi_h(\cdot), \tag{6}$$

i.e., functions are indexed by the coefficients d_h with respect to the chosen basis. Putting a prior probability model on $\{d_h\}$ implicitely defines a prior on f. Wavelets (Vidakovic, 1998) provide a computationally very attractive basis. The (super) fast wavelet transform allows quick and computationally efficient mapping between a function f and the coefficients. The basis functions $\varphi_h(\cdot)$ are shifted and scaled versions of a mother wavelet, $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$, $j \geq J_0$, together with shifted versions of the scaling function $\phi_{J_0 k}(x) = 2^{J_0/2}\phi(2^{J_0}x - k)$, $k \in Z$, i.e.,

$$f(\cdot) = \sum_k c_{J_0 k}\phi_{J_0 k}(x) + \sum_{j \geq J_0}\sum_k d_{jk}\psi_{jk}(\cdot), \tag{7}$$

The coefficients $\boldsymbol{d} = (c_{J_0 k}, d_{jk}\ j \geq J_0,\ k \in Z)$ parametrize the function. The ψ_{jk} and $\phi_{J_0 k}$ form an orthnormal basis. The choice of J_0 is formally arbitrary. Consider $J > J_0$. The mapping between c_{Jk} and $(c_{J_0 k}, d_{jk}, j = J_0, \ldots, J)$ is carried out by an iterative algorithm known as the pyramid scheme. Let $\boldsymbol{f} = (f_1, \ldots, f_{2^J})$ denote the function evaluated on a regular grid. In view of the normalization property, $||\phi_{Jk}|| = 1$, scaling coefficients at a high level of detail J are aproximately proportional to the represented function, $c_{Jk} \approx 2^{-J/2} f_k$. Thus for large J the mapping between c_{Jk} and $(c_{J_0 k}, d_{jk}, j = J_0, \ldots, J)$ effectively becomes a mapping between $\boldsymbol{f}$ and the coefficients and defines the super fast one-to-one map between f and the coefficients that we mentioned before. The nature of the basis functions ψ_{jk} as shifted and scaled versions of the mother wavelet allows an interpretation of d_{jk} as representing a

signal at scale j and shift k. This interpretation suggests a prior probabilty model with increasing probability of zero coefficients, increasing with level of detail j. Barnes et al. (2003) use $p(d_{jk} = 0) = 1 - \alpha^{j+1}$, and a multivariate normal prior for the non-zero d_{jk} and $c_{J_0 k}$, conditional on keeping the zero coefficients. And the periodic nature of the function, with $f(0) = f(1)$, adds another constraint. Chipman et al. (1997), Clyde et al. (1998) and Vidakovic (1998) discuss Bayesian inference in similar models assuming equally spaced data, i.e., covariate values x_i are on a regular grid. Non-equally spaced data do not add significant computational complications.

Example 4 (ctd.) *Figure 5a shows* $\overline{f} = E(f \mid \boldsymbol{y})$ *under a BNP regresssion model based on* (7) *with* $p(d_{jk} = 0) = 1 - \alpha^{j+1}$ *and a multivariate normal dependent prior on* c_J. *The primary inference target here is the range* $\Delta = \max(f) - \min(f)$ *whose posterior uncertainty is mostly determined by the uncertainty in* $f(\cdot)$ *around* $x = 0$ *and thus* ΔR. *Figure 5b shows the implied* $p(\Delta \mid \boldsymbol{y})$.

Morris and Carroll (2006) build functional mixed effects models with hierarchical extensions of (7) across multiple functions. Wavelets are not the only popular basis used for nonparametric regression models. Many alternative basis functions are used. For example, Baladandayuthapani et al. (2005) represent a random function using P-splines.

Gaussian process priors. Besides basis representations like (6), another commonly used BNP prior is the Gaussian process (GP) prior. A random function $f(x)$ with $x \in \Re^d$ has a GP prior if for any finite set of points $x_i \in \Re^d$, $i = 1, \ldots, n$, the function evaluated at those points is a multivariate normal random vector. Let $f^\star(x)$, $x \in \Re^d$ denote a given function and let $r(x_1, x_2)$ for $x_j \in \Re^d$ denote a covariance function, i.e., the $(n \times n)$ matrix R with $R_{ij} = r(x_i, x_j)$ is positive definite for any set of distinct $x_i \in \Re^d$. We write $f \sim \mathrm{GP}\left(f^\star(x), r(x, y)\right)$ if

$$(f(x_1), \ldots, f(x_n))' \sim N\left((f^\star(x_1), \ldots, f^\star(x_n))', R\right).$$

Assuming normal residuals, the posterior distribution for $\boldsymbol{f} = (f(x_1), \ldots, f(x_n))$ is multivariate normal again. Similarly, $f(x)$ at new locations x_{n+i} that were not recorded in the data is characterized by multivariate normal distributions again. See O'Hagan (1978) for an early discussion of GP priors, and Kennedy and O'Hagan (2001) for a discussion of Bayesian inference for GP models in the context of modeling output from computer simulations.

4.3 Fully Nonparametric Regression

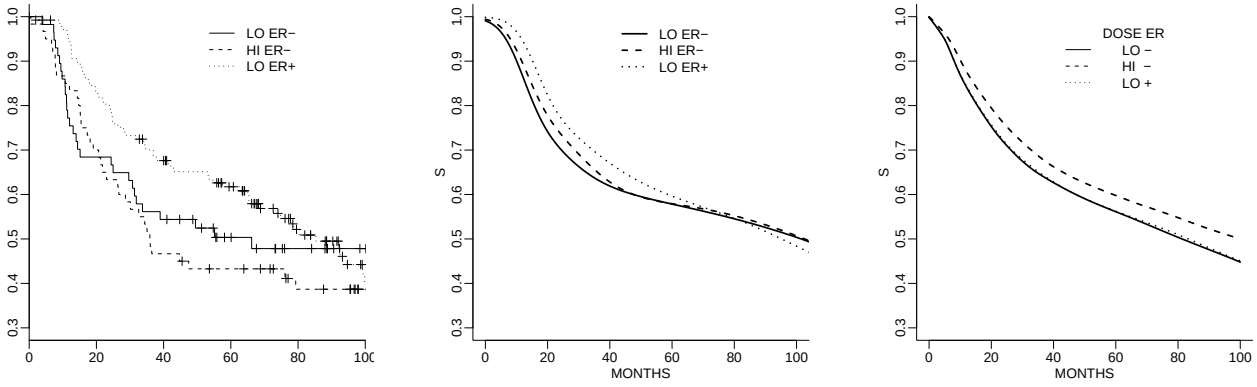
Regression can be characterized as inference for a family of probability models on y that are indexed by the covariate x, i.e., $y \mid x \sim G_x(y)$ and a BNP prior $p(\mathcal{G})$ on $\mathcal{G} = \{G_x(y),\ x \in X\}$. In BNP regression with a nonparametric mean function the model G_x is implied by a parametric residual distribution and a BNP prior for the mean function $f(\cdot)$. In contrast, in fully nonparametric regression the BNP prior is put on the family $\mathcal{G}$.

Example 5 (Breast cancer study) *We illustrate fully nonparametric regression with surival regression in a cancer clinical trial. The trial is described in Rosner (2005). The data record the event-free survival time t_i in months for $n = 761$ women. A subset of $n_0 = 400$ observations are censored. Researchers are interested in determining whether high doses of the treatment are more effective for treating the cancer compared to lower doses. We consider two categorical and one continuous covariate, and one interaction variable: treatment dose (TRT) (-1 = low, 1 = high), estrogen receptor (ER) status (-1 = negative, 1 = positive), the size of the tumour (standardized to zero mean and unit variance), and a dose/ER interaction (1 if a patient receives high treatment dose and has positive ER status and 0 otherwise). This defines a vector x_i of covariates for each patient. The desired inference is to learn about $G_x(y) = p(y \mid x)$, in particular, the comparison with respect to TRT. Figure 6a shows the data as a Kaplan-Meier plot.*

De Iorio et al. (2009) implement inference using a dependent Dirichlet process model (DDP). The DDP was proposed by MacEachern (1999) as a clever extension of the DP prior for one random probability measure G to the desired prior $p(G_x;\ x \in X)$ for a family of random probability measures. The construction builds on the stick-breaking representation (1) of the DP. Consider a DP prior for G_x,

$$G_x(\theta) = \sum_{h=1}^{\infty} \pi_h \delta_{\widetilde{\theta}_h(x)}(\theta) \tag{8}$$

with $\widetilde{\theta}_h(x) \sim G_x^\star$ independent *across* h and $\pi_h = v_h \prod_{\ell<h}(1 - v_\ell)$ with $v_h \sim \text{Be}(1, M)$, i.i.d. Definition (8) ensures that $G_x \sim \text{DP}(G_x^\star, M)$, marginally. The key observation is that we can introduce dependence of $\widetilde{\theta}_h(x)$ *across* x. That is all. By defining a dependent prior on $\{\widetilde{\theta}_h(x)\}_{x\in X}$ we create dependent random probability measures G_x. As a default choice MacEachern (1999) proposes a Gaussian process prior on $\{\widetilde{\theta}_h(x)\}_{x\in X}$. Depending on the nature of the covariate space X other models can be useful too. The DDP model (8) is sometimes referred to as variable location DDP. Alternatively the weights $\pi_h(x)$ and/or both, weights and locations, could be indexed with x, leading to variable weight and variable weight and location DDP.



(a) data (KM) (b) ANOVA DDP (c) AFT

Figure 6: Cancer clinical trial. Panel (a) shows the data as a KM plot arranged by dose and ER status. Posterior survivor functions under the ANOVA DDP model (panel b) and alternatively under the AFT median regression model (panel c). In both plots, the solid line refers to low treatment dose and negative ER status. The dashed line corresponds to high treatment dose and negative ER status, while the long dashed line shows the survival for a patient in the low dose group but with positive ER status.

Example 5 (ctd.) *De Iorio et al. (2009) define inference for a set* $\{G_x;\ x \in X\}$ *indexed by a covariate vector* x *that combines two binary covariates and one continous covariate. In that case, a convenient model for dependent probability distributions on* $(\widetilde{\theta}_h(x);\ x \in X)$ *is a simple ANOVA model for the categorical covariates. Adding a continuous covariate (tumor size) defines an ANCOVA model. Figure 6b shows inference under the ANOVA DDP model. For comparison, Figure 6c show inference for the same data under a semiparametric accelerated failure time (AFT) median regression model with a mixture of Polya trees on the error distribution. The model is described in Hanson and Johnson (2002). The PT is centered at a Weibull model. Panels (b) and (c) report inference for a patient with average tumor size (this explains the discrepancy with the KM plot). The BNP model recovers a hint of crossing survival functions.*

A construction similar to the DDP is introduced in Gelfand et al. (2005) who define a spatial DP mixture by considering a DP prior with a base measure $G^\star$ which itself is a GP, indexed with a spatial covariate x, say $x \in \Re^2$. In other words, a realizations of the spatial DP is a random field $(\theta(x), x \in \Re^2)$. Focusing on one location x we see that the spatial DP induces a random probability measure for $\theta(x)$, call it G_x. However, the spatial DP

defines a stick-breaking mixture of GP realizations. I.e., $\theta(\cdot) \sim \sum \pi_h \delta_{\widetilde{\theta}_h(\cdot)}$. For example, one observation $(\theta(x_1), \theta(x_2))$ at a pair of spatial locations is based on one realization of the base measure GP. In contrast, under the DDP a pair of realizations $\theta(x_1), \theta(x_2)$, could be based on two realizations of the base measure GP (with the possibility of a tie only because of the discrete nature of the distributions). Under the spatial DP, a sampling model for observed data might still add an additional regression.

Many other variations of the DDP have been proposed, including matrix stick-breaking (Dunson et al., 2008) and the kernel stick-breaking of Dunson and Park (2007). Matrix stick-breaking introduces dependence for a set of random probability measures that are arranged in a matrix, i.e., indexed by two categorical indices, say $\{G_{ij} = \sum \pi_{ijh} \delta_{\widetilde{\theta}_h}\}$. In contrast to the common weights π_h in the basic DDP model (8) the model uses varying weights and common locations $\widetilde{\theta}_h$. The construction starts with stick-breaking as in (1), but then assumes $v_{ijh} = U_{ih} W_{jh}$ with the independent beta priors on U_{ih} and W_{jh}. Similarly, kernel stick-breaking introduces dependence across random probability measures $G_x = \sum \pi_{xh} \delta_{\widetilde{\theta}_h}$ by replacing v_{xh} in the stick breaking construction by $V_h K(x, \Gamma_h)$, where V_h is common across all x, $K(x, m)$ is a kernel centered at m and Γ_h are kernel locations. The intention of the construction is to create π_{xh} that are a continuous function of x. The specific nature of π_{xh} as a function of x is hidden in the kernel.

In the recent literature several alternatives to the DDP have been proposed. Dunson et al. (2007) propose density regression as a locally weighted mixture of a fixed set of independent random probability measures. The weights are written as functions of the covariates. A similar model is the already mentioned kernel stick-breaking process of Dunson and Park (2007). Trippa et al. (2011) define a dependent Polya tree model by replacing the random splitting probabilities Y_ϵ by a stochastic process $(Y_\epsilon(x))_{x \in X}$, maintaining the marginal beta distribution for any x. Jara and Hanson (2011) propose a similar construction, but with the random splitting probabilties $Y_\epsilon(x)$ defined by a transformation, for example a logistic transformation, of a GP. The constructed family of dependent random probability measures is known as dependent tail-free processes (DTFP). A special case is the linear dependent tailfree process (LDTFP) that is also discussed in Hanson and Jara (2013).

5 Random Effects Distributions

5.1 Mixed Effects Models

An important application of nonparametric approaches arises in modeling random effects distributions in hierarchical mixed effects models. Often little is known about the specific

form of the random effects distribution. Assuming a specific parametric form is typically motivated by technical convenience rather than by genuine prior beliefs. Although inference about the random effects distribution itself is rarely of interest, it can have implications on the inference of interest, especially when the random effects model is part of a larger model. Thus it is important to allow for population heterogeneity, outliers, skewness etc.

In this context of a mixed effects model with random effects θ_i a BNP model can be used to allow for more general random effects distribution $G(\theta_i)$. Let $y_{ij} = \theta_i + \boldsymbol{\beta}' x_{ij} + \epsilon_{ij}$ denote a randomized block ANOVA with residuals $\epsilon_{ij} \sim N(0, \sigma^2)$, fixed effects $\boldsymbol{\beta}$ and random effects θ_i for blocks of experimental units, $i = 1, \ldots, I$. For technically convenient posterior analysis one could assume a normal random effects distribution $\theta_i \sim N(0, \tau^2)$ and conditionally conjugate priors $p(\beta, \sigma^2, \tau^2)$. While the prior for the fixed effects might be based on substantive prior information, the choice of the random effects distribution is rarely based on actual prior knowledge. The relaxation of the convenient, but often arbitrary distributional assumption for the random effects distribution is a typical application of BNP models. A non-parametric Bayesian model can relax the assumption without losing interpretability and without substantial loss of computational efficiency. Many non-parametric Bayes models allow us to center the prior model $p(G)$ around some parametric model p_η, indexed by hyperparameters η. For example, we could center a prior $p(G)$ for a random effects distribution G around a $N(0, \tau^2)$ model with hyperparameter $\eta = \tau$. The construction allows us to think of the non-parametric model as a natural extension of the fully parametric model.

In the nonparametric extension the random effects distribution itself becomes an unknown quantity. We replace the normal random effects distribution with $\theta_i \sim G,\ G \sim p(G)$ with a BNP prior $p(G)$ for the unknown G. For later reference we state the full mixed effects model

$$\begin{aligned} p(y_{ij} \mid \beta, \theta_i),\ j = 1, \ldots, n_i \\ p(\theta_i \mid G) = G \text{ and } G \sim p(G). \end{aligned} \tag{9}$$

Here the sampling model $p(y_{ij} \mid \beta, \theta_i)$ could, for example be an ANOVA model. The non-parametric prior $p(G)$ is a prior for a density estimation with the (latent) random effects θ_i. We could use any prior that was discussed earlier. The only difference is that now the latent θ_i replace the oberved data in the straightforward density estimation problem. There is one complication. The nature of $G(\cdot)$ as a random effects distribution requires centering at 0 to ensure identifiability. Often, this detail is ignored, or only mitigated by setting up the prior with a prior mean $G^\star = E(G)$ such that $G^\star$ is centered around 0. However, centering the prior mean does not imply centering of G. A non-zero mean of G could be confounded with corresponding fixed effects. Li et al. (2011) propose a clever postprocessing step of MCMC output to allow the use of DPM models including MCMC without any constraint.

Bush and MacEachern (1996) propose a DP prior for $\theta_i \sim G$, $G \sim \text{DP}(G^\star, M)$. Kleinman and Ibrahim (1998) propose the same approach in a more general framework for a linear model with random effects. They discuss an application to longitudinal random effects models. Müller and Rosner (1997) use DP mixture of normals to avoid the awkward discreteness of the implied random effects distribution. Also, the additional convolution with a normal kernel greatly simplifies posterior simulation for sampling distributions beyond the normal linear model. Mukhopadhyay and Gelfand (1997) implement the same approach in generalized linear models with linear predictor $\theta_i + x_i'\beta$ and a DP mixture model for the random effect θ_i. In Wang and Taylor (2001) random effects θ_i are entire longitudinal paths for each subject in the study. They use integrated Ornstein-Uhlenbeck stochastic process priors for θ_i.

Example 6 (Phage display experiment) *Leon-Novelo et al. (2013) discuss an application of BNP priors for random effects distributions that includes a decision problem on top of the inference problem. The BNP prior matters. The decision hinges on a full description of uncertainties in the random effects distribution. Leon-Novelo et al. (2013) analyze count data from a phage display experiment with three stages. The data come from three consecutive human subjects who met the formal criteria for brain-based determination of death. The primary aim of the experiment is to identify peptides that bind with high affinity to particular tissue (bone-marrow, fat, muscle, prostate and skin). Bacteriophages, for short phages, are viruses. They provide a convenient mechanism to study the preferential binding of peptides to tissues, essentially because it is possible to experimentally manipulate the phages to display various peptides on the surface of the viral particle. See Leon-Novelo et al. (2013) for a more detail description of the experimental setup and the study. The data are tripeptide counts by tissue and stage. The experiment is set up in such a way that peptides that preferentially bind to a particular organ should be recorded with systematically increasing counts over the three stages. The inference goal is to select from a large list of peptide and tissue pairs those with significant increase over stages. Figure 7 shows the data. Let $i = 1, \ldots, n$, index all $n = 2763$ recorded tripeptide/tissue pairs. Each line connects the three counts y_{i1}, y_{i2}, y_{i3} for one tripeptide/tissue pair. Of course, even if there were no true preferential binding, and all counts were on average constant across stages, one would expect about $1/4$ of the observed counts to be increasing across the 3 stages. The decision problem is to select pairs with significant increase and report them for preferential binding. Let $Poi(\lambda)$ denote a Poisson distribution. We assume $y_{i1} \sim Poi(\mu_i)$, $y_{i2} \sim Poi(\mu_i\beta_i)$ and $y_{i3} \sim Poi(\mu_i\delta_i)$ for random effects $(\mu_i, \beta_i, \delta_i)$. The event of increasing mean counts becomes $A_i = \{1 \leq \beta_i \leq \delta_i\}$. We use a random effects distribution $(\beta_i, \delta_i) \sim G$ with BNP prior $p(G)$. Figure 8a shows $E(G \mid \boldsymbol{y})$.*

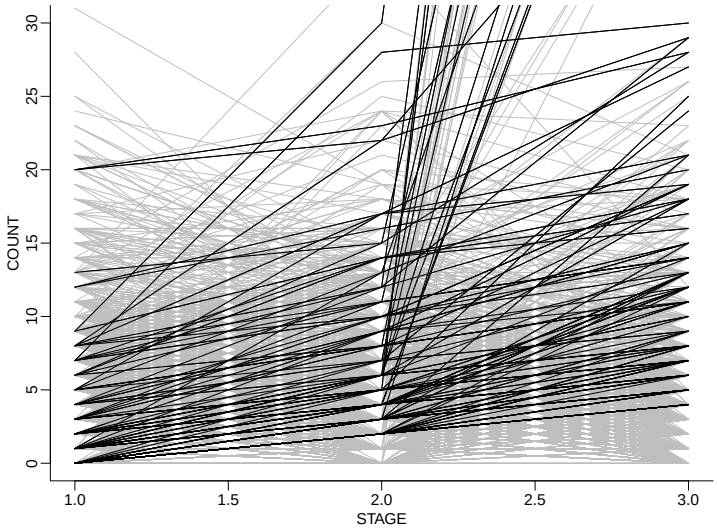
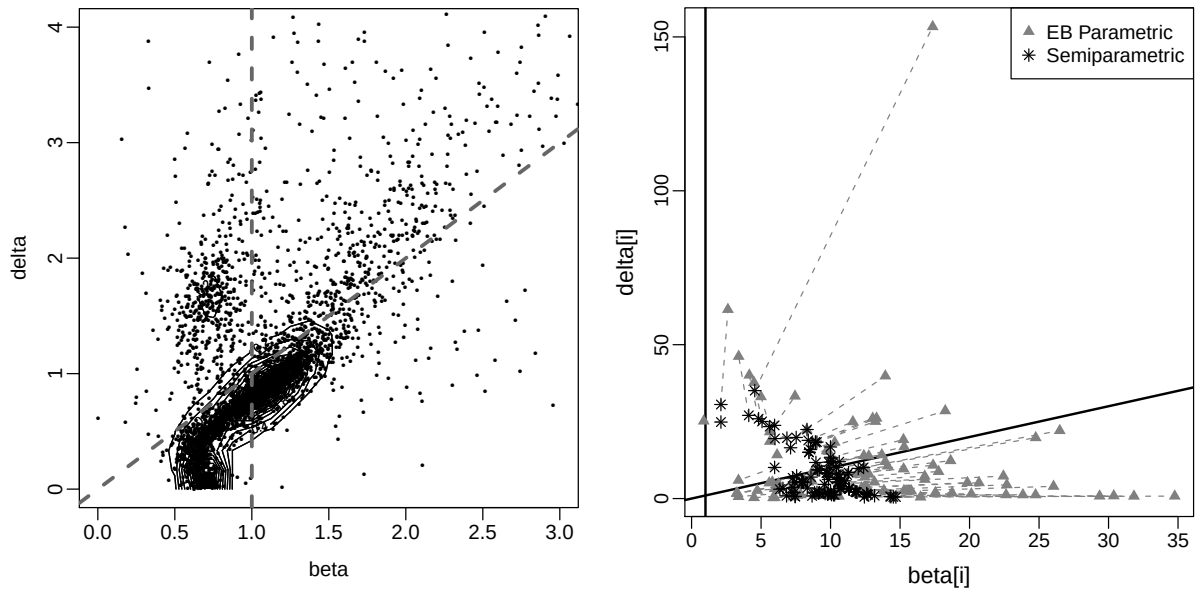


Figure 7: Observed counts y_{i1}, y_{i2}, y_{i3} over stages 1 through 3. Each line connects the three counts for one tripeptide-tissue pair. Tripeptide/tissue pairs with increasing counts $y_{i1}+1 < y_{i2}$ and $y_{i2}+1 < y_{i3}$ are plotted in black (adding the increment 1 to avoid a cluttered display). Others are plotted in grey.

The event $A = \{0 < \beta < \delta\}$ is in the right upper quadrant, between the two lines. The main inference summary are the posterior probabilities for increasing mean counts, $p_i = p(A_i \mid \boldsymbol{y})$. Thresholding p_i defines a decision rule $\delta_i = I(p_i > c)$ for reporting preferential binding for the tripeptide/tissue pair i. Leon-Novelo et al. (2013) use a bound on posterior expected false discovery rate to fix the threshold c. Figure 8b highlights the importance of the BNP model here. The figure reports $(\overline{\beta}_i, \overline{\delta}_i) = E(\beta_i, \delta_i \mid \boldsymbol{y})$ under two alternative models, the described BNP model (marked as "semiparametric" in the figure) and a corresponding parametric model ("EB"). Results in (b) are for a – different – simulated data set. Short line segments connect $(\overline{\beta}_i, \overline{\delta}_i)$ under the two models for each tripeptide/tissue pair i. The corrections are substantial, impacting the posterior probabilities p_i and thus change the decisions δ_i for many pairs.

5.2 Multiple Subpopulations and Classification

The use of BNP priors for random effects distributions becomes particularly useful when the model includes subpopulations, say $v = 1, \ldots, V$ with separate, but related random effects



(a) $E(F \mid \boldsymbol{y})$ (b) $E(\beta_i, \delta_i \mid \boldsymbol{y})$

Figure 8: Estimated random effects distribution (panel a) and posterior estimated random effects $E(\beta_i, \delta_i \mid \boldsymbol{y})$ (panel b, marked with "*") versus posterior means under a similar parametric model ("+"). Results in (b) are under a – different – simulated data set.

distributions G_v. We let $\mathcal{G} = (G_v,\ v = 1, \ldots, V)$ and augment (9) to

$$\begin{aligned} y_{vij} &\sim p(y_{vij} \mid \beta, \theta_{vi}), \quad j = 1, \ldots, n_{vi} \\ \theta_{iv} \mid \mathcal{G} &\sim G_v,\ i = 1, \ldots, n_i, \\ \mathcal{G} &\sim p(\mathcal{G}). \end{aligned} \tag{10}$$

Here $p(\mathcal{G})$ is a BNP prior for a family or random probability measures, for example the DDP model introduced in (8).

Example 7 (Hormone data) *De la Cruz et al. (2007) analyze hormone data for 173 pregnancies. The data report repeat measurements on the pregnancy hormone β-HCG for 173 young women during the first 80 days of gestational age. Figure 9 shows the data. The data include* $n_0 = 124$ *normal pregnancies and* $n_1 = 49$ *pregnancies that were classified as abnormal. The goal is to predict normal or abnormal pregnancy for a future woman on the basis of the longitudinal data as it accrues over time. See Figure 10c shows the desired inference. The figure plots the posterior probability of a normal pregnancy against the number of hormone measurements for two hypothetical future women, one with a normal pregnancy and one with an abnormal pregnancy. Let* $y_i = (y_{i1}, \ldots, y_{in_i})$ *denote the β-HCG repeat measurments for*

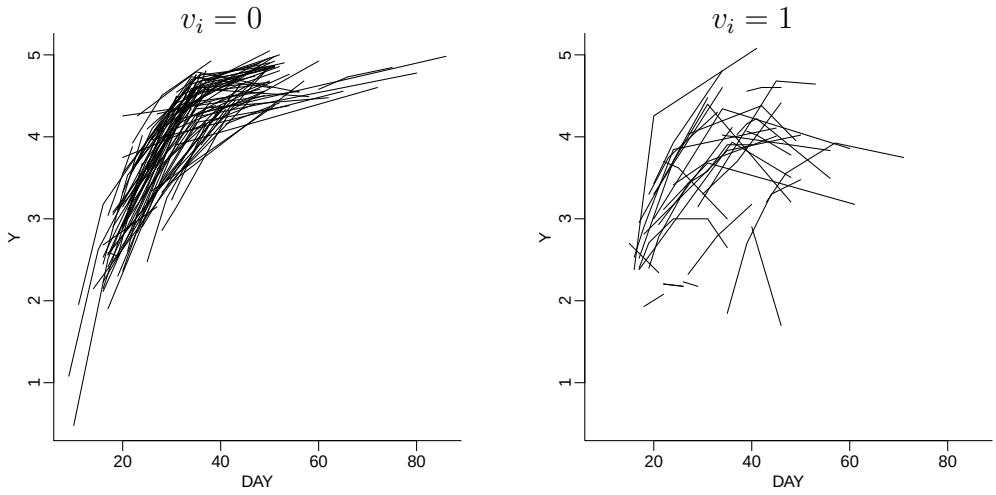


Figure 9: Hormone data. Observed repeat β-HCG measurement for normal (left panel) and abnormal (right) pregnancies.

the i-th woman, recorded at times t_{ij}, $j = 1, \ldots, n_{ij}$. *Let* $v_i \in \{0, 1\}$ *denote an indicator for abnormal pregnancy. The longitudinal data are modeled as a non-linear mixed-effects model*

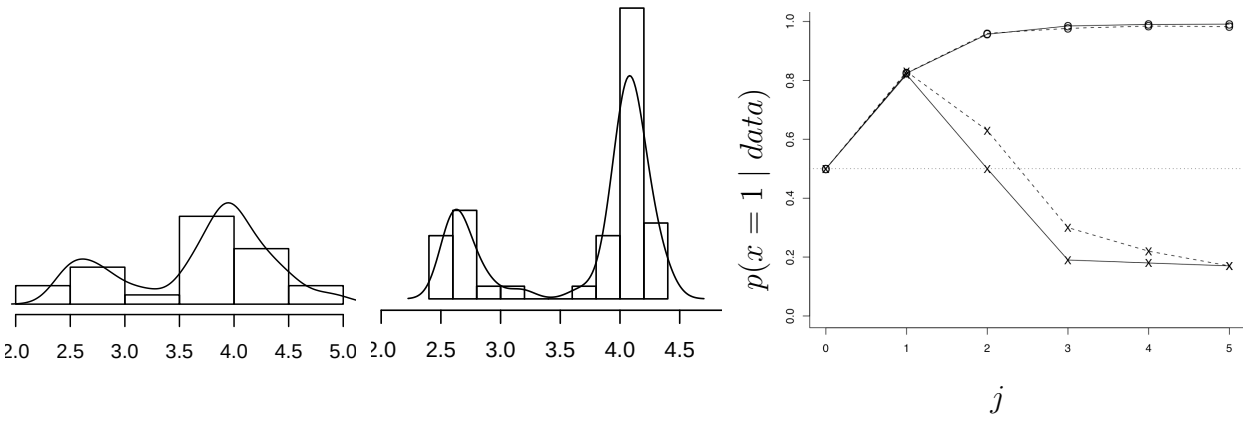
$$p(y_{ij} \mid v_i = v, \beta_v, \sigma_v^2, \theta_i) = N(m_{ij}, \sigma_v^2) \text{ with } m_{ij} = \theta_i \left[1 + \exp\left\{-(t_i - \beta_{1v})/\beta_{2v}\right\}\right]^{-1},$$

i.e., a logistic regression with coefficients β_v *and scaled by random effects* θ_i *and with a normal residuals. Both,* β_v, σ_v^2 *are specific to each group,* $v = 0$ *and* $v = 1$. *Let* $\phi = (\beta_v, \sigma_v^2,\ v = 0, 1)$. *The model includes a patient-specific random effect* θ_i *with* $\theta_i \mid v_i = x \sim G_v(\theta_i)$. *We assume a BNP prior* $p(G_0, G_1)$. *We use an ANOVA DDP prior on* $G_v = \sum w_h \delta_{\widetilde{\theta}_h(v)}$, $v = 0, 1$. *The binary nature of* $v \in \{0, 1\}$ *makes the model particularly simple, with* $\widetilde{\theta}_h(x) = m_h + a_{hv} + \epsilon_{hv}$, *where* $a_{h0} = 0$ *and* $\epsilon_{hv} \sim N(0, \tau^2)$. *The model is completed with a bivariate normal prior* $G^\star(m_h, a_{h1})$ *and conditionally conjugate priors for* ϕ. *Figure 10ab shows the estimated random effects distributions.*

A simple augmentation of model (10) allows to use the same model for classification. First we change indexing of experimental units i to run $i = 1, \ldots, n$ across all subgroups, and add $v_i \in \{1, \ldots, V\}$ as an indicator for unit (patient) i being in group v. Then add a prior $p(v_i = v) = \pi_v$, to get

$$p(v_{n+1} = v \mid y_{n+1,1}, \ldots, y_{n+1,j}, \boldsymbol{\pi}, \boldsymbol{y}) \propto \pi_v \, E\left[\int p(y_{n+1} \mid \theta_{n+1}, \phi)\, dG_v(\theta_{n+1}) \mid \boldsymbol{y}\right], \qquad (11)$$

for $v = 1, \ldots, V$. The expectation in square brackets is with respect to the posterior probability model on $\mathcal{G}, \phi$ given the observed data $(y_i, v_i;\ i = 1, \ldots, n)$.



(a) $E(G_0 \mid y)$ (b) $E(G_1 \mid y)$

Figure 10: Hormone data. Estimated random effects distributions $E(G_v \mid y)$, $v = 0, 1$.

Example 7 (ctd.) *Figure 10c shows the posterior probability* $p(v_{n+1} = 1 \mid y_{n+1,1}, \ldots, y_{n+1,j}, \boldsymbol{y})$ *for two hypothetical future patient* $i = n + 1$*, plotted against* $j = 0, 1, \ldots, 6$*, as repeat observations accrue. The evaluation of the classification rule in* (11) *makes use of* $p(G_1, G_0 \mid \boldsymbol{y})$*. The BNP model matters.*

6 Asymptotics

With sufficiently large data, the posterior distribution should be concentrated more and more tightly around the true parameter θ_0. This property is known as posterior consistency. Posterior consistency statements are results about probabilities under repeat experimentation under some unknown truth θ_0, i.e., results about frequentist probabilities. A lot of recent BNP research is concerned with such asymptotic results. In the world of Bayesian nonparametrics, the true parameter is typically an infinite dimensional object. It could be a probability density function, a c.d.f., the spectral density of a time series etc. We therefore consider distances in function spaces. Some commonly studied metrics in posterior consistency are the Hellinger distance, the Kullback-Leibler metric, and the the L_1 metric. Neighborhoods defined by the L_1 metric are known as strong neighborhoods. A weak neighborhood V of a function f_0 is a set indexed by ϵ and a finite set of bounded continuous functions $\phi_1 \ldots \phi_k$ such that $V = \{f : |\int \phi_i f - \int \phi_i f_0| < \epsilon \;\; i = 1 \ldots k\}$. We say that a measure f_0 is in the support of a prior $p(f)$ if every weak neighborhood of f_0 has positive p measure. For the rest of the discussion we assume that the goal is inference for an unknown distribution F_0, and the data are i.i.d. observations, $x_i \sim F_0$, $i = 1, \ldots, n$. We say that a

model exhibits posterior consistency with respect to a particular topology (strong or weak) if $p(U \mid x_1 \dots x_n) \to 1$ a.s.-F_0 for all neighborhoods U of F_0 corresponding to that topology.

Freedman (1963) proposed tail-free distributions as a class of priors for which posterior consistency holds. Consider a nested sequence of partitions (π_m) of the sample space, $\pi_1 = \{B_0, B_1\}$, $\pi_2 = \{B_{00}, B_{01}, B_{10}, B_{11}\}$ etc., such that π_{m+1} is a refinement of π_m, i.e., $B_\epsilon = B_{\epsilon 0} \cup B_{\epsilon 1}$, where $\epsilon = e_1 e_2 \cdots e_m$ is an m-digit binary index. A prior $p(G)$ is called tail-free with respect to a nested sequence of partitions if $\{G(B_0)\}, \{G(B_{00} \mid B_0), G(B_{01} \mid B_0)\}$, etc. are independent across partitions. Two important priors that exhibit consistency due to a tail-free property are the DP and the PT priors. However the tail-free property is not common and can be destroyed when the process is convoluted with a mixing measure. This concern generated a need to formalize good priors in terms of consistency theorems that impose general sufficient conditions on the true density F_0 and the prior $p(F)$. Schwartz's theorem (1965) is the first important step in this direction and forms a strong basis for a lot of subsequent work.

Schwartz (1965) proposed an important condition for consistency. The prior should put positive mass on all Kullback Leibler (KL) neighborhoods of the true density. This is referred to as Schwartz's prior positivity condition or the KL property of the prior. A second condition is the existence of a sequence of uniformly exponentially consistent tests of $H_0 : F = F_0$ vs. $H_1 : F \in U^c$ for every neighborhood U of F_0. Together, these conditions ensure consistency. The second condition is readily met for weak topology. Thus, as a corollary, prior positivity becomes a sufficient condition for weak consistency.

We review some results on weak consistency of DPM's of normal models. Let $\phi(x;\ \theta, h)$ denote a normal p.d.f. with location θ and scale h. We consider DPM models of the form $F(x) = \int \phi(x;\ \theta, h)\, dG(\theta)$ with $G \sim \mathrm{DP}(G^\star, M)$. The model is completed with a prior $p(h)$. We refer to such DPM models as DP location mixture of normals, for short "location mixtures." Ghosal et al. (1999) proves prior positivity, and hence weak consistency for a location mixture. The sufficient conditions are that the true density itself is a convolution, i.e $F_0 = \int \phi(x;\ \theta, h_0) dP_0(\theta)$ where P_0 is compactly supported and belongs to the weak support of the DP prior and h_0 is in the support of $p(h)$. In the same paper, the result is extended to DPM location-scale mixture of normals $F(x) = \int \phi(x;\ \theta, h)\, dG(\theta, h)$, for short "location-scale mixtures."

To establish strong consistency of DPM of normal models additional techniques, like constructions of sieves are required. Using such constructions, Ghosal et al. (1999) proves strong consistency for location mixture priors when the true F_0 is in the KL support of the prior, subject to some conditions on $p(h)$ and the tail of the base measure $G^\star$ of the DP

prior. These conditions are satisfied for a normal base measure for θ and an inverse gamma prior $p(h^2)$. Lijoi et al. (2005) improved upon these results by replacing the exponential tail condition by $\int |\theta| G^\star(\theta) < \infty$. Ghosal and van der Vaart (2001) established a convergence rate of $log(n)^k/\sqrt{n}$ for strong consistency in location-scale mixtures, where k depended on the tail behavior of the base measure. The result assumed that the true densities are DPMs with compactly supported mixing measure and that h is in a bounded interval. Such densities are known as super-smooth. Ghosal and van der Vaart (2007) generalize the result to the larger class of twice differentiable true densities. They assume location mixtures, with the prior $p_n(h)$ on the scale changing with sample size. A rate, lower than that in Ghosal and van der Vaart (2001), but equal to an optimal rate of a kernel estimator is obtained in this setting. Tokdar (2006) established both strong and weak consistency for a large class of true densities F_0 satisfying $\int |x|^\eta F_0(x) < \infty$ for some $\eta > 0$. This class includes heavy tailed distributions like the t density. The priors are location-scale mixtures with some regularity conditions on the tail of the base measure $G^\star$, which are shown to be satisfied for normal and inverse gamma base measures.

Although most arguments use sieves and Schwarz's framework, there are some alternative approaches too. Walker and Hjort (2002) and Walker (2004, 2003) use the martingale property of marginal densities as a unifying tool. For recent reviews of consistency and convergence rates in DPM models, see Walker et al. (2007) and Ghoshal (2010).

Some recent work considers posterior consistency for models beyond DP priors. Jang et al. (2010) showed that in the class of Pitman-Yor process priors, DP priors are the only ones with posterior consistency. Gaussian processes are another important class of priors with well known consistency results. For example, assume a regression setting with binary outcomes y_i where the success probabilities $p(y_i = 1 \mid x_i)$ are a smooth unknown function $f(x_i)$ of a set of covariates x_i. Let $h(\theta)$ denote an inverse logit link (or any other monotone mapping from $\Re$ to the unit interval) and define a prior $p(f)$ by assuming $f(x) = h[\theta(x)]$ for a Gaussian process $\theta(x) \sim$ GP. Ghosal and Roy (2006) discussed posterior consistency for such models. More general results on consistency and rates of convergence for a large class of GP priors (e.g Brownian motion) are shown in van der Vaart and van Zanten (2008).

7 Conclusion

We tried to motivate BNP inference by a discussion of some important inference problems and examples that highlight the limitations of parametric inference. The statement is meant in reference to a standard, default parametric model. Naturally, in each of these examples

one could achieve similar inference with sufficiently complicated parametric models like a finite mixture etc. However, inference under such models is usually nothing easier than under the BNP model. For example, inference with a finite mixture model gives rise to all the same complications as a nonparametric mixture, such as the DPM model.

We have not discussed two important aspects of BNP inference. Inference for many models quickly runs into computation intensive posterior inference problems. We did not discuss many such details. Also, a large part of the recent BNP literature is concerned with asymptotic properties of BNP inference, which we only briefly summarized in this review. For an excellent recent review of posterior asymptotics in DP and related models see Ghoshal (2010).

Finally, we owe a comment about the term "nonparametric." We started out by defining BNP as probabilty models for infinite dimensional random quantities like curves or densities. It might be more fittingly called "massively parametric Bayes". The label nonparametric has been used because inference under BNP models often looks similar to (genuinely) nonparametric classical inference.

Acknowledgments and Data

Both authors were partially funded by grant NIH/CA075981.
The data for examples 1 through 4 are available on-line.

References

Baladandayuthapani, V., B. K. Mallick, and R. J. Carroll (2005). Spatially adaptive Bayesian penalized regression splines (P-splines). *Journal of Computational and Graphical Statistics 14*(2), 378–394.

Barnes, T. G., W. H. Jefferys, J. O. Berger, P. Müller, K. Orr, and R. Rodríguez (2003). A Bayesian Analysis of the Cepheid Distance Scale. *The Astrophysical Journal 592*(1), 539.

Barrios, E. J., A. Lijoi, L. E. Nieto-Barajas, and I. Prünster (2013). Modeling with normalized random measure mixture models. *Statistical Science*, to appear.

Branscum, A. J., W. O. Johnson, T. E. Hanson, and I. A. Gardner (2008). Bayesian semiparametric ROC curve estimation and disease diagnosis. *Statistics in Medicine 27*(13), 2474–2496.

Broderick, T., J. Pitman, and M. I. Jordan (2013, January). Feature allocations, probability functions, and paintboxes. *ArXiv e-prints*.

Bush, C. A. and S. N. MacEachern (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika 83*, 275–285.

Chipman, H., E. Kolaczyk, and R. McCulloch (1997). Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association 92*, 440.

Clyde, M., G. Parmigiani, and B. Vidakovic (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika 85*, 391–402.

Dahl, D. B. (2006). Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model. In M. Vannucci, K.-A. Do, and P. Müller (Eds.), *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press.

De Iorio, M., W. O. Johnson, P. Müller, and G. L. Rosner (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics 65*(3), 762–771.

De la Cruz, R., F. A. Quintana, and P. Müller (2007). Semiparametric Bayesian classification with longitudinal markers. *Applied Statistics 56*(2), 119–137.

Dunson, D. B. and J.-H. Park (2007). Kernel stick-breaking processes. *Biometrika* **95**, 307–323.

Dunson, D. B., N. Pillai, and J.-H. Park (2007). Bayesian density regression. *Journal of the Royal Statistical Society, Series B: Statistical Methodology 69*(2), 163–183.

Dunson, D. B., Y. Xue, and L. Carin (2008). The matrix stick-breaking process: Flexible bayes meta-analysis. *Journal of the American Statistical Association 103*(481), 317–327.

Favaro, S. and Y. W. Teh (2013). MCMC for normalized random measure mixture models. Technical report, Oxford University, Department of Statistics.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*, 209–230.

Freedman, D. (1963). On the asymptotic behavior of bayes' estimates in the discrete case. *The Annals of Mathematical Statistics 34*(4), 1386–1403.

Gelfand, A. E., A. Kottas, and S. N. MacEachern (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* **100**, 1021–1035.

Ghahramani, Z., T. Griffiths, and P. Sollich (2007). Bayesian nonparametric latent feature models. In J. M. Bernardo, M. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, and A. F. M. Smith (Eds.), *Bayesian Statistics 8*, pp. 201226. Oxford University Press.

Ghosal, S., J. Ghosh, and R. Ramamoorthi (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics 27*(1), 143–158.

Ghosal, S. and A. Roy (2006). Posterior consistency of Gaussian process prior for nonparametric binary regression. *The Annals of Statistics 34*(5), 2413–2429.

Ghosal, S. and A. van der Vaart (2001). Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *The Annals of Statistics 29*(5), 1233–1263.

Ghosal, S. and A. van der Vaart (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics 35*(2), 697–723.

Ghoshal, S. (2010). The Dirichlet process, related priors and posterior asymptotics. See Hjort et al. (2010), pp. 22–34.

Griffiths, T. and Z. Ghahramani (2006). Infinite latent feature models and the Indian buffet process. In Y. Weiss, B. Schölkopf, and J. Platt (Eds.), *Advances in Neural Information Processing Systems 18*, pp. 475–482. MIT Press.

Guindani, M., N. Sepúlveda, C. D. Paulino, and P. Müller (2012). A Bayesian semiparametric approach for the differential analysis of sequence counts data. Technical report, M.D. Anderson Cancer Center.

Hanson, T. and A. Jara (2013). Surviving fully Bayesian nonparametric regression. In P. Damien, P. Dellaportas, N. G. Polson, and D. A. Stephens (Eds.), *Bayesian Theory and Applications*, pp. 593–618. Oxford University Press.

Hanson, T. and W. Johnson (2002). Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association 97*, 1020–1033.

Hanson, T. E. (2006). Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association 101*(476), 1548–1565.

Hartigan, J. A. (1990). Partition models. *Communications in Statistics: Theory and Methods 19*, 2745–2756.

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

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*, 161–173.

James, L. F., A. Lijoi, and I. Prünster (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics 36*(1), 76–97.

Jang, G. H., J. Lee, and S. Lee (2010). Posterior consistency of species sampling priors. *Statistica Sinica 20*(2), 581.

Jara, A., T. Hanson, F. Quintana, P. Müller, and G. Rosner (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software 40*(5), 1–30.

Jara, A. and T. E. Hanson (2011). A class of mixtures of dependent tail-free processes. *Biometrika 98*(3), 553–566.

Kennedy, M. C. and A. O'Hagan (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 63*(3), pp. 425–464.

Kingman, J. F. C. (1993). *Poisson Processes*. Oxford University Press.

Kleinman, K. and J. Ibrahim (1998). A semi-parametric Bayesian approach to the random effects model. *Biometrics 54*, 921–938.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics 20*, 1222–1235.

Lavine, M. (1994). More aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics 22*, 1161–1176.

Lee, J., F. Quintana, P. Mueller, and L. Trippa (2013). Defining predictive probability functions for species sampling models. *Statistical Science*, to appear.

Leon-Novelo, L., B. Bekele, P. Müller, F. Quintana, and K. Wathen (2012). Borrowing strength with non-exchangeable priors over subpopulations. *Biometrics 68*, 550–558.

Leon-Novelo, L. G., P. Müller, W. Arap, M. Kolonin, J. Sun, R. Pasqualini, and K.-A. Do (2013). Semiparametric Bayesian inference for phage display data. *Biometrics*, to appear.

Li, Y., P. Müller, and X. Lin (2011). Center-adjusted inference for a nonparametric Bayesian random effect distribution. *Statistica Sinica 21*(3), 1201–23.

Lijoi, A., R. H. Mena, and I. Prünster (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association 100*(472), pp. 1278–1291.

Lijoi, A., R. H. Mena, and I. Prünster (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(4), 715–740.

Lijoi, A. and I. Prünster (2010). Models beyond the Dirichlet process. See Hjort et al. (2010), pp. 80–136.

Lijoi, A., I. Prünster, and S. G. Walker (2005). On consistency of nonparametric normal mixtures for Bayesian density estimation. *Journal of the American Statistical Association 100*(472), 1292–1296.

MacEachern, S. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA. American Statistical Association.

Mauldin, R. D., W. D. Sudderth, and S. C. Williams (1992). Polya trees and random distributions. *The Annals of Statistics 20*, 1203–1221.

Morris, J., K. Baggerly, and K. Coombes (2003). Bayesian shrinkage estimators of the relative abundance of mRNA transcripts using SAGE. *Biometrics 59*, 476–486.

Morris, J. S. and R. J. Carroll (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology 68*(2), 179–199.

Mukhopadhyay, S. and A. Gelfand (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association 92*, 633–639.

Müller, P., F. Quintana, and G. Rosner (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics 20*, 260–278.

Müller, P. and F. A. Quintana (2004). Nonparametric Bayesian data analysis. *Statistical Science 19*, 95–110.

Müller, P. and A. Rodriguez (2013). *Nonparametric Bayesian Inference*. IMS-CBMS Lecture Notes. IMS.

Müller, P. and G. Rosner (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association 92*, 1279–1292.

O'Hagan, T. (1978). Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological) 40*(1), pp. 1–42.

Paddock, S. M., F. Ruggeri, M. Lavine, and M. West (2003). Randomized Polya tree models for nonparametric Bayesian inference. *Statistica Sinica 13*(2), 443–460.

Pepe, M. S., R. Etzioni, Z. Feng, J. D. Potter, M. L. Thompson, M. Thornquist, M. Winget, and Y. Yasui (2001). Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute 93*(14), 1054–1061.

Pitman, J. (1996). Some Developments of the Blackwell-MacQueen Urn Scheme. In T. S. Ferguson, L. S. Shapeley, and J. B. MacQueen (Eds.), *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, pp. 245–268. Haywar, California: IMS Lecture Notes - Monograph Series.

Pitman, J. and M. Yor (1997). The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator. *The Annals of Probability 25*, 855–900.

Quintana, F. A. (2006). A predictive view of Bayesian clustering. *J. Statist. Planning and Inference 136*, 2407–2429.

Quintana, F. A. and P. L. Iglesias (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(2), 557–574.

Regazzini, E., A. Lijoi, and I. Prünster (2003). Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics 31*(2), 560–585.

Rodríguez, A., D. B. Dunson, and A. E. Gelfand (2008). The nested Dirichlet process, with discussion. *Journal of American Statistical Association 103*, 1131–1144.

Rosner, G. L. (2005). Bayesian monitoring of clinical trials with failure-time endpoints. *Biometrics 61*, 239–245.

Schwartz, L. (1965). On bayes procedures. *Probability Theory and Related Fields 4*(1), 10–26.

Sethurman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*, 639–650.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Sharing clusters among related groups: Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**, 1566–1581.

Tokdar, S. T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics 68*, 90–110.

Trippa, L., P. Müller, and W. Johnson (2011). The multivariate beta process and an extension of the Polya tree model. *Biometrika 98*(1), 17–34.

van der Vaart, A. and J. van Zanten (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics 36*(3), 1435–1463.

Vidakovic, B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *Journal of the American Statistical Association 93*, 173–179.

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

Walker, S. (2003). On sufficient conditions for Bayesian consistency. *Biometrika 90*(2), 482–488.

Walker, S. (2004). New approaches to Bayesian consistency. *The Annals of Statistics 32*(5), 2028–2043.

Walker, S. (2013). Bayeisan nonparametrics. In P. Damien, P. Dellaportas, N. G. Polson, and D. A. Stephens (Eds.), *Bayesian Theory and Applications*, pp. 249–270. Oxford University Press.

Walker, S., P. Damien, P. Laud, and A. Smith (1999). Bayesian nonparametric inference for distributions and related functions (with discussion). *Journal of the Royal Statistical Society, Series B 61*, 485–527.

Walker, S. and N. Hjort (2002). On Bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(4), 811–821.

Walker, S., A. Lijoi, and I. Prünster (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *The Annals of Statistics 35*(2), 738–746.

Wang, Y. and J. M. Taylor (2001). Jointly modeling longitudinal and event time data with applcation to acquired immunodeficiency syndrome. *Journal of the American Statistical Association 96*, 895–903.

Zhang, S., P. Müller, and K.-A. Do (2010). A Bayesian semiparametric survival model with longitudinal markers. *Biometrics 66*(2), 435–443.