Chapter 1
Bayesian Nonparametric Models

Peter Müller and Riten Mitra

Abstract We briefly review some of the nonparametric Bayesians models that are most widely used in biostatistics and bioinformatics. We define the Dirichlet process, Dirichlet process mixtures, the Polya tree, the dependent Dirichlet process and the Gaussian process prior. These few models and variations cover a major part of the models that are used in the literature. The discussion includes references to variations of the basic models that are defined in the chapters of this volume.

1.1 Nonparametric Bayesian Inference in Biostatistics and Bioinformatics

The increased complexity of biomedical inference problems requires ever more sophisticated and flexible approaches to statistical inference. The challenges include in particular massive data, high-dimensional sets of potential covariates, highly structured stochastic systems, and complicated decision problems. Some of these challenges can be naturally addressed with a class of inference approaches known as nonparametric Bayesian (BNP) methods. A technical definition of BNP models is that they are probability models on infinite dimensional probability spaces. This includes priors on random probability measures, random mean functions, and more.

BNP methods relax the sometimes restrictive assumptions of traditional parametric methods. A parametric model is indexed by an unknown finite dimensional

P. Müller (✉)
The University of Texas at Austin, 1, University Station, C1200, Austin, TX 78712, USA
e-mail: pmueller@math.utexas.edu

R. Mitra
University of Louisville, Louisville, KY, USA
e-mail: riten82@gmail.com

R. Mitra, P. Müller (eds.), *Nonparametric Bayesian Inference in Biostatistics*,
Frontiers in Probability and the Statistical Sciences,
DOI 10.1007/978-3-319-19518-6_1

parameter vector θ. Bayesian inference proceeds by assuming a prior probability model $p(\theta)$ which is updated with the relevant sampling model $p(y \mid \theta)$ for the observed data y.

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. If G is restricted to be in a family $\{G_\theta,\ \theta \in \Re^d\}$, then the prior is specified as a prior probability model $p(\theta)$ for the d-dimensional parameter vector θ. In contrast, if G is not restricted to a finite dimensional parametric family, then the prior model $p(G)$ becomes a probability model for the infinite dimensional G.

A very common related use 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 might understate uncertainties and miss some important structures. 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 $y_i = f(x_i) + \varepsilon_i$.

The chapters in this volume discuss important research problems in biostatistics and bioinformatics that are naturally addressed by BNP methods. Each chapter introduces and defines the BNP methods and models that are used to address the specific problem. In this introductory chapter we briefly introduce and review some of the most commonly used BNP priors. Posterior inference in many of these models gives rise to challenging computational problems. We review some of most commonly used computational methods and include some references. The brief review in this introduction includes the ubiquitous Dirichlet process (DP) model, the DP mixture model (DPM), the dependent DP (DDP) model, the Polya tree (PT) prior, and the Gaussian process (GP) prior. These models and their variations are the workhorses of BNP inference in biostatistics. The next chapter in this volume discusses some typical examples by reviewing BNP methods in some important applications.

For a more exhaustive discussion of BNP models, see, for example, recent discussions in Hjort et al. (2010), Müller and Rodríguez (2013), Walker et al. (1999), Müller and Quintana (2004), Walker (2013) and Müller et al. (2015).

1.2 Dirichlet Process

Let $\delta_x(\cdot)$ denote a point mass at x. The DP prior (Ferguson 1973) is a probability model for a random distribution G,

$$G = \sum_{h=1}^{\infty} w_h \delta_{m_h}, \tag{1.1}$$

with independent locations $m_h \sim G_0$, i.i.d., and weights that are constructed as $w_h = v_h \prod_{\ell<h}(1-v_\ell)$ with independent beta fractions $v_h \sim \text{Be}(1,M)$, i.i.d. (Sethurman 1994). The prior on w_h is known as the stick-breaking process. It can be described as breaking off fractions v_h of a stick of initially unit length. The DP prior is characterized by the base measure G_0 that generates the locations of the atoms m_h and the total mass parameter M that determines the distribution of the beta fractions v_h. We write $G \sim \text{DP}(M,G_0)$. Implied in the constructive definition of the stick breaking construction is an important property of DP random measures. A DP random measure $G \sim \text{DP}(M,G_0)$ is discrete with probability one.

The DP is a conjugate prior under i.i.d. sampling. That is, assume $x_i \mid G \sim G$, i.i.d., $i = 1,\ldots,n$ and $G \sim \text{DP}(M,G_0)$. Let $F_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ denote the empirical distribution. Then $p(G \mid \boldsymbol{x}) = \text{DP}(M+n,G_1)$ with $G_1 \propto MG_0 + nF_n$. An interesting limiting case occurs for $M \to 0$, when the posterior on G is entirely determined by the empirical distribution. This leads to a construction known as the Bayesian bootstrap, which is discussed in Chap. 16 (Inácio de Carvalho et al. 2015).

One of the reasons for the wide use of the DP prior is ease of computation for posterior inference in models based on the DP. In particular, the DP prior implies a particularly simple predictive probability function $p(x_n \mid x_1,\ldots,x_{n-1})$. Under i.i.d. sampling from a DP random measure the marginal distribution $p(x_1,\ldots,x_n) = \int \prod_{i=1}^n G(x_i)\, dp(G)$ reduces to a simple expression which is easiest characterized as $p(x_1,\ldots,x_n) = \prod_{i=1}^n p(x_i \mid x_1,\ldots,x_{i-1})$ with increasing conditionals

$$p(x_i \mid x_1,\ldots,x_{i-1}) \propto MG_0(x_i) + \sum_{\ell=1}^{i-1} \delta_{x_\ell}. \tag{1.2}$$

With probability $\pi_0 = M/(i-1+M)$ the sample x_i is a new draw from G_0, and with probability $1/(i-1+M)$ the new sample is tied with a previous sample x_ℓ. The conditional distribution (1.2) is also known as the Polya urn. We will return to it below. Let $\boldsymbol{x}_{-i} = \boldsymbol{x} \setminus \{x_i\}$. For later reference we note that by symmetry the conditional distribution $p(x_i \mid \boldsymbol{x}_{-i})$ takes the same form.

1.2.1 DP Mixture

The discrete nature of a DP random measure is awkward in many applications and is therefore often avoided by using an additional convolution with a continuous kernel. Let $k(x_i \mid \theta)$ denote a continuous kernel, for example a Gaussian kernel. Without loss of generality we assume in the remaining discussion $k(x_i \mid \theta) = N(x_i \mid \theta, s)$ (with fixed s). The DP mixture (DPM) model assumes $G = \int N(x_i \mid \theta, s) dF(\theta)$, with $F \sim \text{DP}(M,F_0)$. We write $G \sim \text{DPM}(M,G_0,k)$. It is often convenient to rewrite the mixture as an equivalent hierarchical model. Instead of $y_i \sim G$ and $G \sim \text{DPM}(M,G_0,k)$ we write

$$y_i \mid \theta_i \sim N(\theta_i, s) \text{ and } \theta_i \sim F \tag{1.3}$$

with $F \sim \text{DP}(M, G_0)$. The DPM model is one of the most widely used BNP priors for random distributions. In this volume we find it, for example, in Chap. 11 (Zhou and Hanson 2015) to construct a semiparametric version of an accelerated failure time model; in Chap. 16 (Inácio de Carvalho et al. 2015) as a prior for the distribution of test outcomes to develop inference on ROC curves; in Chap. 21 (Daniels and Linero 2015) for longitudinal outcomes under different missingness patterns; and many more.

Consider again the θ_i in (1.3). As a sample from the discrete random measure F, the newly introduced latent variables θ_i include many ties. Let $\boldsymbol{\theta}^\star = \{\theta_1^\star, \ldots, \theta_k^\star\}$ denote the $k \leq n$ unique values and let $S_j = \{i : \ \theta_i = \theta_j^\star\}$ denote the indices $[n] \equiv \{1, \ldots, n\}$ arranged by the configuration of ties. Then $\rho_n \equiv \{S_1, \ldots, S_k\}$ defines a partition of $[n]$. Since the θ_i were random, as a consequence the partition is random. That is, the DP mixture model (1.3) induces a random partition $p(\rho_n)$. At first glance this seems like a coincidental detail of the model. However, many applications of the DPM model exploit exactly this feature. It features in many chapters in this volume. The implied prior $p(\rho_n)$ on the random partition is also known as Chinese restaurant process (CRP). It is used, for example, in Chap. 3 (Zhang et al. 2015).

Sometimes it is convenient to index the partition ρ_n alternatively by an equivalent set of cluster membership indicators. Let s_i denote indicators with $s_i = j$ if $i \in S_j$, that is when $\theta_i = \theta_j^\star$. Let $n_j = |S_j|$ denote the size of the jth cluster, $n_j^- = |S_j \setminus \{i\}|$ and let k^- denote the number of unique values θ_ℓ in $\boldsymbol{\theta}_{-i}$. Then we can rewrite (1.2) as

$$s_i \mid \boldsymbol{s}_{-i} = \begin{cases} j & \text{with prob } \frac{n_j^-}{n-1+M}, \quad j = 1, \ldots, k^- \\ k^- + 1 & \text{with prob } \frac{M}{n-1+M} \end{cases} \tag{1.4}$$

The attraction of model (1.3) is the ease of posterior simulation. Consider a generic model $y_i \sim G$ with DPM prior (1.3) and similar to k^- and n_j^- let $\theta_j^{\star -}$ denote the jth unique value among $\boldsymbol{\theta}_{-i}$. Then (1.4) implies

$$\theta_i \mid \boldsymbol{y}, \boldsymbol{\theta}_{-i} = \begin{cases} \theta_j^{\star -} & \text{with prob.} \propto n_j^- \, p(y_i \mid \theta_j^{\star -}) \\ \sim H_1 & \text{with prob.} \propto M \int p(y_i \mid \theta) \, dG_0(\theta) \end{cases} \tag{1.5}$$

with $H_1(\theta) \propto p(y_i \mid \theta) G_0(\theta)$. If $p(y_i \mid \theta)$ and $G_0(\theta)$ are chosen as a conjugate pair of sampling model and prior, then generating from (1.5) is straightforward. In the general case, the evaluation of $h_0 \equiv \int p(y_i \mid \theta) \, dG_0(\theta)$ can be computationally challenging. Several MCMC algorithms have been proposed to circumvent the evaluation of an analytically intractable integral h_0 (Neal 2000). For a recent review of the DP and related models, see, for example, Ghosal (2010).

1.2.2 Generalizations of the DP

Many generalizations of the DP prior have been proposed in the literature. One example is the Poisson-Dirichlet (PD) process that is used in Chap. 9 (Guha et al. 2015). The PD arises by replacing the $\text{Be}(1, M)$ prior on the fractions v_h in the

stick breaking construction by $\text{Be}(1-a, b+ha)$ priors. Other generalizations are specifically focused on the implied random partition model, like the generalized Ottawa sequence introduced in Chap. 5 (Bassetti et al. 2015) or the hierarchical DP (HDP) model in Chap. 7 (Iorio et al. 2015). The latter defines a prior on a family of random probability measures $\{G_j; j = 1, \ldots, j\}$.

1.3 Dependent Dirichlet Process

Many problems involve a family of unknown random probability measures $\mathscr{F} = \{F_x; x \in X\}$. For example, in a mixed effects model that includes data from several related studies, F_j might be the random effects distribution for patients in study j. More generally, a formalization of non-parametric regression could assume

$$y_i \mid \boldsymbol{x}_i = x, \mathscr{F} \sim F_x \tag{1.6}$$

$i = 1, \ldots, n$. That is, we denote by F_x the sampling model for the response of a subject with covariates $\boldsymbol{x}_i = x$. If we are willing to assume $F_x = N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$, then the problem reduces to parametric inference on the finite dimensional parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$. In other words, we restrict $\mathscr{F}$ to the family of probability measures indexed by $\boldsymbol{\theta}$. In the absence of such restrictions Bayesian inference in (1.6) requires a prior probability model $p(\mathscr{F})$ that allows for dependence and borrowing of strength across x, short of the strict parametric assumption, but still more than in a model with independent, separate priors on each F_x.

One of the most popular models in the recent literature for a family of random probability measures $\mathscr{F}$ is the dependent DP (DDP) and variations of it. The model was first introduced in MacEachern (1999). The idea is simple. We continue to use

$$F_x = \sum_{h=1}^{\infty} w_h \delta_{m_{xh}}, \tag{1.7}$$

with independent locations m_{xh}, i.i.d. across h and weights that are constructed with independent beta fractions as before, in (1.7). The only addition is that we now introduce dependence on the point masses m_{xh} *across* x. For example, we could assume that $(m_{xh},\ x \in X)$ is a realization of a Gaussian process indexed by x. In the simplest implementation the weights w_h are shared across all x, as implied in the notation w_h without a subindex for x.

Similar to the DP mixture model, the DDP model (1.7) is often combined with a continuous kernel, for example a normal kernel to define

$$G_x(y) = \int N(y \mid \theta, \sigma^2)\, dF_x(\theta) = \sum_{h=1}^{\infty} w_h N(y \mid m_{xh}, \sigma^2). \tag{1.8}$$

with a DDP prior on $\{F_x,\, x \in X\}$. Here $N(y \mid m, s^2)$ denotes a normal kernel in y. We refer to (1.8) as a DDP mixture of normals. For categorical covariates $x \in X$ the dependent probability model for $(m_{xh},\, x \in X)$ could be defined, for example, as an ANOVA model. This defines the ANOVA DDP proposed in DeIorio et al. (2002). A version of the same, with a general linear model in place of the ANOVA model is the linear dependent DP (LDDP) (Jara et al. 2010).

1.3.1 Variations of the DDP

The DDP prior and variations of it are used in several chapters in this volume. Chapter 12 (Jara et al. 2015) uses an LDDP to implement survival regression. Chapter 20 (Karabatsos and Walker 2015) constructs a variation of a DDP by introducing the dependence on covariates in (1.7) by a probit regression in the weights w_h, rather than the atoms m_h.

1.4 Polya Tree

The Polya tree (PT) prior (Lavine 1992, 1994) is an attractive alternative BNP prior for a random probability measure. The PT prior is essentially a random histogram. Without loss of generality, assume that we wish to define a random probability measure G on the unit interval $[0,1]$. We could start with a random histogram with two bins $\{B_0, B_1\}$, say over $B_0 = [0, 0.5)$ and $B_1 = [0.5, 1]$. Let $Y_0 = G(B_0)$ and $Y_1 = 1 - Y_0$ denote the (random) probabilities of B_0 and B_1. Next we refine the histogram by splitting the bins into $B_0 = B_{00} \cup B_{01}$ with $B_{00} = [0, 0.25)$, etc. Let $Y_{00} = G(B_{00} \mid B_0)$, $Y_{10} = G(B_{10} \mid B_1)$, $Y_{01} = 1 - Y_{00}$, and $Y_{11} = 1 - Y_{10}$. We continue refining the histogram to 2^m bins, $m = 1, 2, \ldots$ by repeating similar binary splits. The process creates a sequence $\Pi = \{\Pi_m,\, m = 1, 2, \ldots\}$ of nested binary partitions $\Pi_m = \{B_{e_1 \cdots e_m}\}$ with $e_j \in \{0, 1\}$. The PT defines a prior on G by assuming

$$Y_{\varepsilon 0} \sim \mathrm{Be}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}),$$

independently across ε and $Y_{\varepsilon 1} = 1 - Y_{\varepsilon 0}$. The nested partitions Π together with the beta parameters $\mathscr{A} = \{\alpha_\varepsilon\}$ characterize the PT prior. We write $G \sim \mathrm{PT}(\Pi, \mathscr{A})$.

One of the attractions of the PT prior is the ease of centering the model. Let $0.e_1 \cdots e_m = \sum_j e_j 2^{-j}$ denote the number with binary digits $\varepsilon = e_1, \ldots, e_m$ and let q_ε denote the corresponding quantile of a fixed probability measure G_0. That is, for example, q_1, q_{01}, q_{10} are the median and the first and third quartile of G_0. Next define B_ε to denote the corresponding partitioning subsets and let $\prod$ denote the nested partition sequence with partitioning subsets B_ε. If $G \sim \mathrm{PT}(\Pi, \mathscr{A})$ with $\alpha_{\varepsilon 0} = \alpha_{\varepsilon 1}$, then $E(G) = G_0$. We write

$$G \sim \mathrm{PT}(G_0, \mathscr{A}).$$

A particularly attractive choice is $\alpha_{e_1 \cdots e_m} = c2^m$ which can be shown to imply a continuous random probability measure G. We write $G \sim \text{PT}(G_0, c)$. Alternatively, for an arbitrary nested partitioning sequence Π, define $\mathscr{A}$ by $\alpha_{\varepsilon e_m} = c\, G_0(B_{\varepsilon e_m} \mid B_\varepsilon)$ and assume $G \sim \text{PT}(\Pi, \mathscr{A})$. Then again $E(G) = G_0$. We write

$$G \sim \text{PT}(\Pi, G_0).$$

For a recent review of the PT prior, see, for example, Müller et al. (2015 Chapter 3). PT priors are used, for example, in Chap. 11 (Zhou and Hanson 2015) to construct a semi-parametric accelerated failure time model.

1.5 Gaussian Process

Gaussian Process (GP) priors are widely used in machine learning, medical imaging, ecology, and various disease risk models. A GP is a stochastic process $\{Y(s); s \in S\}$ that extends (finite dimensional) multivariate Gaussians to infinite dimensions. Here Y(.) is a function-valued random variable while S denotes the domain (typically $\Re^e$) of the function. The domain S and thus $Y(.)$ can have very different interpretation and meaning depending upon specific applications. For example, in Chap. 17 (Reich and Fuentes 2015), and typically in the context of spatial models, S refers to all location points in a given region. For machine learning applications, it can be the set of all possible input stimuli. It could even represent the time domain for recording neuronal activity as in the case study provided in Chap. 13 (Shahbaba et al. 2015). S is usually endowed with its own specific metric, e.g. the Euclidean distance in spatial applications. The problem of analyzing the random function $Y(.)$ or predicting its value Y(**s**) at a specific point **s** can be formulated within the framework of non-parametric regression, where the values in S play the role of covariates and $Y(.)$ is the regression function to be estimated. A prior on the random function $Y()$ would simply refer to the probability law of the stochastic process.

We formally characterize a GP as a stochastic process with mean function $m(.)$ and covariance function $k(\cdot, \cdot)$ if every finite sub-collection of this process, $[Y(s_1), Y(s_2) \ldots Y(s_n)]$ is multivariate Gaussian

$$[Y(s_1), \ldots, Y(s_n)] \sim N(\mu, \Sigma) \text{ with } \mu = [m(s_1), \ldots, m(s_n)] \text{ and } \Sigma_{ij} = k(s_i, s_j).$$

We write $Y \sim \text{GP}(m, k)$. The covariance function is sometimes also referred to as the kernel of the GP. The prior on the random $Y(.)$, thus defined, is called a GP prior. Simply put, a GP extends finite multivariate Gaussian models to infinite dimensions. It can be shown that such an extension is possible using Kolmogorov's consistency theorem. Naturally, the infinite process inherits many attractive properties of its finite version. For example, no restrictions are required for the mean function m. However, since all finite dimensional subsets are required to be Gaussian, a condition of positive semi-definiteness is implied on V for any finite subset of S.

A variety of different families of valid kernels are in common use. Some popular choices include squared exponential (SE), polynomial, neural network, Ornstein-Uhlenbeck (OU), Matern, etc. Each of these families typically has a number of free hyper-parameters. Choosing a covariance function for a particular application thus comprises both, the setting of hyper-parameters within a family, and sometimes the comparison across different families through model-selection techniques. Alternatively, flexible and non-parametric covariance functions can be built by exploiting the spectral representation of a GP. Chapter 17 (Reich and Fuentes 2015) introduces such general priors for spatial covariances by applying the DP and the DPM priors to the coefficients of the spectral density.

In general, all covariance functions formally encode some notion of similarity between a pair of random observations based on the distance between corresponding elements of S. Consider, for example, the SE kernel given by $k(s,t) = exp(-||s-t||^2/2\tau^2)$. The functional form suggests that observations corresponding to proximal points are highly correlated, with the correlation dropping off exponentially with the distance between the points.

Posterior inference and prediction with GP priors is made immensely easy by using the analytical results for multivariate Gaussians. For this, it is enough to observe that the collection of new and observed variables is a finite subset of the GP and their joint density is a multivariate Gaussian. Hence, the posterior predictive distribution, obtained by conditioning on the observed data, appears as another multivariate normal. The infinite dimension of the prior, while providing substantial modeling flexibility, poses no concern for inference and computation. These properties turn out to be critical for several analytical manipulations with the GP prior.

However a known computational bottleneck is the inversion of $(n \times n)$ matrices that appear in the analytical results, thus making the computational complexity cubic in the number of data points. For large datasets (n $>$ 10,000) this is prohibitive (in both time and space) for any inference, Bayesian or otherwise. So a number of computational methods [e.g., reduced rank matrix approximations (Fine et al. 2001; Smola and Schökopf 2000)] have been developed. Another approach is to exploit structures of special classes of covariance functions for exact computation. These methods are iterative and the computation scales linearly with the size of the data (Johannesson and Cressie 2004). Cressie and Johannesson (2008) extended this approach to a flexible class of covariance functions. The computational complexity also increases drastically in multivariate settings with several spatially dependent response variables. Banerjee et al. (2008) used induced predictive process models as a clever strategy for dimension reduction and to reduce computational cost in this context. An alternative solution to the computational problem is the treed Gaussian process of Gramacy and Lee (2008). The approach proceeds by first partitioning the covariate space into a number of smaller regions, similar to a classification and regression tree (CART). Next, independent GP's are fit to each subregion. The overall inversion of a large matrix is replaced by a number of smaller, computationally feasible inversions. Posterior inference is efficiently handled in the `tgp` package for R.

An excellent reference on Gaussian process models for regression is Rasmussen and Williams (2005).

1.6 Conclusion

In this brief review we only introduced some of the most popular BNP models and variations. Some of the chapters use models beyond this selection. Chapter 4 (Ji et al. 2015) uses an Indian buffet process as a prior probability model for a feature allocation problem. Feature allocation generalizes random clusters, that is, non-overlapping subsets, to families of possibly overlapping subsets. Chapter 10 (Nieto-Barajas 2015) introduces several alternative models, including, for example, the normalized generalized gamma (NGG) process. The same NGG process appears in Chap. 6. Some chapters define random functions based on spline bases, including Chap. 14 (Telesca 2015) and Chap. 11 (Zhou and Hanson 2015). Finally, Chap. 8 (Ni et al. 2015) discusses prior probability models for random networks.

The next chapter, Chap. 2 continues this review by discussing some typical applications of basic BNP models.

References

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(4), 825–848.

Bassetti, F., Leisen, F., Airoldi, E., and Guindani, M. (2015). Species sampling priors for modeling dependence: an application to the detection of chromosomal aberrations. In Mitra and Müller (2015).

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(1), 209–226.

Daniels, M. J. and Linero, A. R. (2015). Bayesian nonparametrics for missing data in longitudinal clinical trials. In Mitra and Müller (2015).

DeIorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2002). ANOVA DDP models: A review. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, and B. Yu, editors, *Nonlinear Estimation and Classification*, page 467. Springer-Verlag.

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

Fine, S., Scheinberg, K., Cristianini, N., Shawe-taylor, J., and Williamson, B. (2001). Efficient svm training using low-rank kernel representations. *Journal of Machine Learning Research*, **2**, 243–264.

Ghosal, S. (2010). The Dirichlet process, related priors and posterior asymptotics. In Hjort et al. (2010), pages 22–34.

Gramacy, R. B. and Lee, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, **103**, 1119—1130.

Guha, S., Banerjee, S., Gu, C., and Baladandayuthapani, V. (2015). Nonparametric variable selection, clustering and prediction for large biological datasets. In Mitra and Müller (2015).

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

Inácio de Carvalho, V., Jara, A., and de Carvalho, M. (2015). Bayesian nonparametric approaches for ROC curve inference. In Mitra and Müller (2015).

Iorio, M. D., Favaro, S., and Teh, Y. W. (2015). Bayesian inference on population structure: from parametric to nonparametric modeling. In Mitra and Müller (2015).

Jara, A., Lesaffre, E., De Iorio, M., and Quintana, F. A. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *The Annals of Applied Statistics*, **4**, 2126–2149.

Jara, A., García-Zattera, M. J., and st Komárek, A. (2015). Fully nonparametric regression modelling of misclassified censored time-to-event data. In Mitra and Müller (2015).

Ji, Y., Sengupta, S., Lee, J., Müller, P., and Gulutoka, K. (2015). Estimating latent cell subpopulations with Bayesian feature allocation models. In Mitra and Müller (2015).

Johannesson, G. and Cressie, N. (2004). Variance-covariance modeling and estimation for multi-resolution spatial models. In *geoENV IV – Geostatistics for Environmental Applications*, pages 319–330. Springer.

Karabatsos, G. and Walker, S. G. (2015). A Bayesian nonparametric causal model for regression discontinuity designs. In Mitra and Müller (2015).

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.

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

Mitra, R. and Müller, P., editors (2015). *Nonparametric Bayesian Methods in Biostatistics and Bioinformatics*. Springer-Verlag.

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

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

Müller, P., Quintana, F., Jara, A., and Hanson, T. (2015). *Nonparametric Bayesian Data Analysis*. Springer Verlag.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.

Ni, Y., Marchetti, G. M., Baladandayuthapani, V., and Stingo, F. C. (2015). Bayesian approaches for large biological networks. In Mitra and Müller (2015).

Nieto-Barajas, L. E. (2015). Markov processes in survival analysis. In Mitra and Müller (2015).

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Reich, B. J. and Fuentes, M. (2015). Spatial Bayesian nonparametric methods. In Mitra and Müller (2015).

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

Shahbaba, B., Behseta, S., and Vandenberg-Rodes, A. (2015). Neuronal spike train analysis using gaussian process models. In Mitra and Müller (2015).

Smola, A. J. and Schökopf, B. (2000). Sparse greedy matrix approximation for machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 911–918, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Telesca, D. (2015). Bayesian analysis of curves shape variation through registration and regression. In Mitra and Müller (2015).

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

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

Zhang, Z., Pati, D., and Srivastava, A. (2015). Bayesian shape clustering. In Mitra and Müller (2015).

Zhou, H. and Hanson, T. (2015). Bayesian spatial survival models. In Mitra and Müller (2015).