Journal of Statistical Research *ISSN 0256 - 422 X*
2016, Vol. 48-50, No. 2, pp. 1-19

A SHORT TUTORIAL ON BAYESIAN NONPARAMETRICS

PETER MÜLLER
Department of Statistics and Data Science, University of Texas, Austin, TX, USA
Email: pmueller@math.utexas.edu

YANXUN XU
Department of Applied Mathematics & Statistics, Johns Hopkins University, Baltimore, MD, USA
Email: yxu.stat@gmail.com

ALEJANDRO JARA
Department of Statistics, Pontificia Universidad Católica de Chile, Chile
Email: atjara@uc.cl

SUMMARY

Bayesian nonparametric (BNP) models are prior models for infinite-dimensional parameters, such as an unknown probability measure F or an unknown regression mean function f. We review some of the most widely used BNP priors, including the Dirichlet process (DP), DP mixture, the Polya tree (PT), and Gaussian process (GP) priors. We discuss how these models are used in typical inference problems. The examples include `R` code using available packages for inference under BNP priors.

Keywords and phrases: Bayes; nonparametrics; stochastic processes

AMS Classification: 62G05, 62G07, 62G08, 62F15

1 Introduction

Statistical models are almost never right. All models involve certain parametric and structural assumptions. Bayesian nonparametric inference is an increasingly widely used approach to mitigate the dependence on such assumptions. Technically, Bayesian nonparametric (BNP) models can be defined as probability models on infinite-dimensional parameter spaces, usually devised for random distributions or random mean functions. Typical examples are the Dirichlet process (DP) and the Polya tree (PT) priors for random distributions, or Gaussian process (GP) priors for random functions.

In this review we introduce some of the most widely used models and methods, with an emphasis on practical implementation. Recent more comprehensive reviews of BNP inference appear in Walker et al. (1999), Hjort (2003), Müller and Quintana (2004), Hjort et al. (2010), Walker (2013), Phadia (2013), or Müller and Mitra (2013). An in-depth discussion of asymptotic properties can be found in the forthcoming book by Ghoshal and van der Vaart (2017). A recent more applied discussion of BNP, similar in style to this review, appears in Müller et al. (2015).

2 Density Estimation - Random Probability Measures

One could argue that density estimation is the simplest statistical inference problem. Given data $x_i \sim F$, i.i.d., $i = 1, \ldots, n$, we wish to estimate F. Letting $\boldsymbol{x} = (x_1, \ldots, x_n)$, this defines a sampling model

$$p(\boldsymbol{x} \mid F) = \prod_{i=1}^{n} F(x_i). \tag{2.1}$$

Choosing a Bayesian approach we need to complete the model by adding a prior probability model on all unknown quantities that appear in the sampling model, in this case F. We could now assume that F is a member of some parametric family, like $F \in \{F_\theta,\ \theta \in \Theta\}$, for example with $\boldsymbol{\theta} = (\mu, \sigma^2)$ and $F_{\boldsymbol{\theta}} = N(\mu, \sigma^2)$. In that case we indirectly put a prior on F by assuming a prior $p(\theta)$ and the problem reduces to traditional parametric inference. We would report the posterior distribution $p(\theta \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \theta) p(\theta)$.

Often, however, investigators are not willing to make such a sweeping assumption, and prefer instead to treat F itself as the unknown quantity. In that case Bayesian inference requires to complete (2.1) with a prior probability model $p(F)$ for the unknown distribution. Prior probability models for infinite dimensional quantities, such as the probability measure F in this case, are known as BNP models.

2.1 Dirichlet process (DP) prior

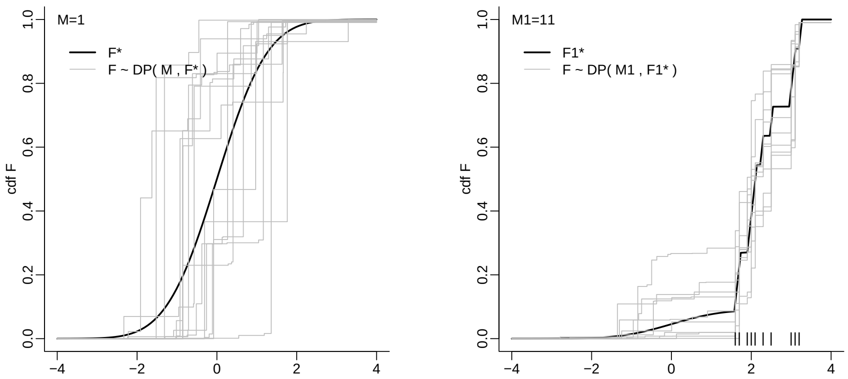
The first discussion of priors on random probability measures in the context of statistical inference was Ferguson (1973), who introduces the Dirichlet process (DP) prior. Let δ_x denote a unit point mass at x. The idea is very simple. We define a random probability measure

$$F = \sum_{h=1}^{\infty} w_h \delta_{m_h} \tag{2.2}$$

by generating $m_h \sim F^\star$, i.i.d. and generating the w_h as beta-distributed fractions by $w_h = v_h \prod_{\ell<h}(1 - v_h)$ with $v_h \sim \text{Be}(1, M)$, i.i.d. In words, w_h is a v_h fraction of whatever probability mass is left of an initial total probability 1.0. We say the random distribution F follows a DP with base measure $F^\star$ and total mass M, and write

$$F \sim \text{DP}(M, F^\star).$$

The base measure has an interpretation as prior mean. Consider any event A and the probability $F(A)$. Since F is random, the probability $F(A)$ becomes a random variable itself. It is easy to show $E\{F(A)\} = F^\star(A)$. Here, the expectation is with respect to the random F, that is, with respect to the w_h and m_h in (2.2). The total mass parameter has an interpretation as precision parameter. In fact, one can show $F(A) \sim \text{Be}\{MF^\star(A), M(1 - F^\star(A))\}$. That is, the random probability is a beta random variable. Considering the expression for the variance of a beta random variable we see that uncertainty decreases with larger M, leaving it interpretable as a precision parameter. Figure 1a shows an example of $F \sim \text{DP}(M, F^\star)$ with $M = 1$ and a standard normal $F^\star$. The random



(a) $F \sim \text{DP}(M, F^\star)$ (b) $F \mid \boldsymbol{x} \sim \text{DP}(M_1, F_1^\star)$

Figure 1: The left panel shows draws from a DP prior for a random probability measure F. The thick line shows the prior mean $F^\star$. The many thin lines show 10 random draws $F \sim \text{DP}(M, F^\star)$. Distributions are shown as cumulative distribution functions (cdf). The right panel shows the posterior DP, $F \mid \boldsymbol{x} \sim \text{DP}(M_1, F_1^\star)$, conditional on data (shown as tick marks on the x-axis).

distributions are shown as c.d.f's. This is convenient since F is a.s. discrete, as is already implicit in the notation used in (2.2).

One of the reasons for the wide use of the DP prior is its conjugacy under i.i.d. sampling. Assume $x_i \mid F \sim F$, i.i.d., $i = 1, \ldots, n$, as in (2.1), together with a DP prior on F, i.e., $F \sim \text{DP}(M, F^\star)$. Then $p(F \mid \boldsymbol{x})$ is again a DP. Let $\hat{F}_n = \frac{1}{n}\sum \delta_{x_i}$ denote the empirical distribution. Then

$$F \mid \boldsymbol{x} \sim \text{DP}(M_1, F_1^\star) \text{ with } M_1 = M + n,\ F_1^\star = (MF^\star + n\hat{F}_n)/(M+n). \tag{2.3}$$

Figure 1b shows random draws from a posterior DP, conditional on observing data $\boldsymbol{x} = (x_1, \ldots, x_n)$.

2.2 Dirichlet process mixture (DPM)

The discrete nature of F under a DP prior is awkward for many applications and usually makes it unsuitable as a prior for F in the density estimation problem (2.1). This is the case in the following example.

Example 2.1 (Old Faithful geyser)**.** Azzalini and Bowman (1990) analyze a data set concerning eruptions of the Old Faithful geyser in Yellowstone National Park in Wyoming. The data record eruption durations and intervals between subsequent eruptions, collected continuously from August 1st until August 15th, 1985. Of the original 299 observations we removed 78 observations that were taken at night and only recorded durations as "short", "medium", or "long". Let x_i, $i = 1, \ldots, n$ denote the remaining $n = 221$ eruption durations. Figure 2a shows a histogram of the data. The data look decidedly non-normal. The data are available, for example, in the `R` package `DPpackage`

(Jara et al., 2011), as `faithful$eruptions`. Assuming $x_i \sim F$ we wish to make inference on F.

The DP prior (2.2) is easily extended to a prior model for continuous distributions by convoluting with a continuous kernel. Let $N(y; \mu, \sigma^2)$ indicate a normal distributed r.v. y, and by a slight abuse of notation a normal kernel in y, centered at μ and variance σ^2. We generalize (2.2) to

$$F = \sum_{h=1}^{\infty} w_h N(y; m_h, \sigma^2) = \int N(y; m, \sigma^2)\, dG(m) \tag{2.4}$$

with $G = \sum w_h \delta_{m_h} \sim \text{DP}(M, G^\star)$. The normal kernel could be replaced by any other continuous kernel $\varphi(y; m)$. The model is known as DP mixture. We write

$$F \sim \text{DPM}(M, G^\star, \varphi).$$

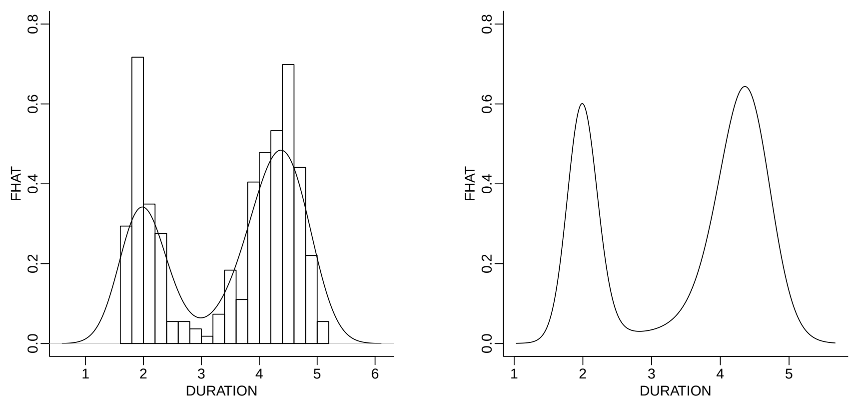
Often the kernel includes some additional hyperparameters, like σ^2 above. DPM models were introduced in Ferguson (1983), Lo (1984), Escobar (1988, 1994), and Escobar and West (1995). Inference under the DPM model is implemented in `DPpackage` as the function `DPdensity(.)`. We briefly show the code to estimate F in Example 2.1, using a DPM prior. See the documentation of `DPpackage` and Jara et al. (2011) for details on the parameters and settings. Figure 2b shows the estimated distribution $\overline{F} = E(F \mid \boldsymbol{x})$ for example 2.1.

```
require("DPpackage")                              ## cran.r-project.org/
y <- round(faithful$eruptions, digits=2)          # data
state <- NULL                                     # Initial state
mcmc <- list(nburn=10,nsave=1000,nskip=10,ndisplay=100) # MCMC parameters
prior1<-list(alpha=1,m1=rep(0,1),                 # prior
      psiinv1=diag(0.5,1),nu1=4,tau1=1,tau2=100)
fit1 <-DPdensity(y=y,prior=prior1,mcmc=mcmc,      # fit the model
          state=state,status=TRUE)
plot(fit1,ask=FALSE)                               # Plot the estimated density
cbind(fit1$x1,fit1$dens)                           # Extracting Fhat
plot(fit1, ask=T,output="param", nfigr=2, nfigc=2) # plot pars
```

Model-based clustering with DP mixtures. For later reference we state two more equivalent ways of writing the DPM model (2.4). First, the integral in $F = \int N(y; m, \sigma^2)\, dG(m)$ can be replaced by a hierarchical model by way of introducing latent variables μ_i. Assume $y_i \mid F \sim F$. We can equivalently write

$$\begin{aligned} y_i \mid \mu_i &\sim N(\mu_i, \sigma^2) \\ \mu_i \mid G &\sim G, \end{aligned} \tag{2.5}$$

$i = 1, \ldots, n$ and $G \sim \text{DP}(M, G^\star)$. Marginalizing with respect to the newly introduced latent variables μ_i we get back to $y_i \sim \int N(y; m, \sigma^2)\, dG(m)$, as before.



(a) data x_i (b) estimate $\overline{F} = E(F \mid \boldsymbol{x})$

Figure 2: The left panel shows the data x_i, together with a kernel density estimate. The right panel shows the estimate $\overline{F} = E(F \mid \boldsymbol{x})$ under the DPM prior (using `plot(fit1)` in the included code fragment).

As a sample from a discrete probability measure G the μ_i can include ties. Let $\{\theta_1^\star, \ldots, \theta_K^\star\}$ denote the $K \leq n$ unique values, let $s_i = k$ if $\mu_i = \mu_k^\star$ and let $n_k = |\{i: \ s_i = k\}|$. We use the convention of labeling $\theta_k^\star$ by appearance, that is, $s_1 = 1$ and $s_i \leq \max\{s_\ell, \ \ell < i\} + 1$. We can then alternatively rewrite the DPM model as

$$\begin{aligned} y_i \mid s_i = k, \boldsymbol{\mu}^\star & \sim N(\mu_k^\star, \sigma^2) \\ \mu_k^\star & \sim G^\star, \end{aligned} \tag{2.6}$$

$i = 1, \ldots, n$ and $k = 1, \ldots, K$, independently. It can be shown that (2.5) implies

$$p(\boldsymbol{s}) \propto \alpha^K \prod_{k=1}^{K} (n_k - 1)!. \tag{2.7}$$

The indicators s_i can be interpreted as cluster membership indicators for clusters $S_k = \{i: \ s_i = k\}$. This makes (2.7) a random partition of $[n] = \{1, \ldots, n\}$ into subsets $S_1, \ldots, S_K$. The prior (2.7) is also known as the Polya urn prior or Chinese restaurant process. In other words, the DPM model includes inference on a partition $\boldsymbol{s}$ of experimental units (by unique $\mu_k^\star$). What might seem like a coincidental property of the DPM model is in fact often the main inference target. Often the DPM model is explicitly used for model-based clustering of the experimental units $i = 1, \ldots, n$. *A posteriori*, $p(\boldsymbol{s} \mid \boldsymbol{x})$ summarizes inference about the unknown partition of the experimental units $\{1, \ldots, n\}$. We will not further explore this in the upcoming discussion. For a more extensive review see, for example, Müller et al. (2015, chapter 8).

2.3 Polya tree

The DP can be characterized as a special case of several other more general models. One is the Polya tree (PT) prior. The PT specifies essentially a random histogram. Without loss of generality, assume that G is a random probability measure on the unit interval $[0, 1]$. The PT defines G as a random histogram over $[0, 1]$. Start with the simplest possible histogram with two bins, $B_0 = [0, \frac{1}{2})$ and $B_1 = [\frac{1}{2}, 1]$. Let $Y_0 = G(B_0)$ and $Y_1 = G(B_1)$. We assume $Y_0 \sim \text{Be}(a_0, a_1)$ and $Y_1 = 1 - Y_0$. This defines the random probability measure G at the very coarse level of this partition $[0, 1] = B_0 \cup B_1$. Next we refine the histogram by splitting B_0 into $B_{00} = [0, \frac{1}{4})$ and $B_{01} = [\frac{1}{4}, \frac{1}{2})$ and similarly for $B_1 = B_{10} \cup B_{11}$. Defining $G(B_{e_1 e_2})$, $e_m \in \{0, 1\}$, we need to be careful to respect the already defined $G(B_{e_1})$. This is easiest done by defining conditional probabilities $Y_{00} = G(B_{00} \mid B_0)$ etc. Continuing like this we define

$$Y_{\boldsymbol{e},0} = G(B_{\boldsymbol{e}0} \mid B_{\boldsymbol{e}}) \sim \text{Be}(a_{\boldsymbol{e}0}, a_{\boldsymbol{e}1}) \tag{2.8}$$

for any length m binary sequence $\boldsymbol{e} = e_1 \cdots e_m$. The construction implies

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

for any partitioning subset $B_{\boldsymbol{e}}$. That is all! In summary the PT prior is determined by a nested sequence of partitions $\Pi = \{\Pi_1, \Pi_2, \ldots\}$ with $\Pi_m = \{B_{e_1 \cdots e_m}\}$, $m = 1, 2, \ldots$, and a sequence of beta parameters $\mathcal{A} = \{a_{\boldsymbol{e}};\ \boldsymbol{e} = e_1 \cdots e_m\}$. We write

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

See Lavine (1992, 1994) for an extensive discussion. The special case with $a_{\boldsymbol{e}} = a_{\boldsymbol{e}0} + a_{\boldsymbol{e}1}$, that is, the beta coefficients adding up over different level partitions, reduces to the DP. In general the nested partition sequence Π and $\mathcal{A}$ need to specified. However, there are convenient default choices. For example, if $a_{\boldsymbol{e}} = cm^2$ for $\boldsymbol{e} = e_1 \cdots e_m$ and any $c > 0$, then the PT prior generates a.s. continuous distributions. And Π can be chosen by specifying a desired prior mean, say $G^\star$ by using dyadic quantiles as the boundaries of the partitioning subsets $B_{\boldsymbol{e}}$. For example, for a distribution $G^\star$ on the real line, let Q_1, Md, and Q_3 denote the 1st quartile, median, and 3rd quartile and define $B_0 = (-\infty, \text{Md}]$, $B_1 = (\text{Md}, \infty)$, $B_{00} = (-\infty, Q_1]$, $B_{01} = (Q_1, \text{Md}]$, etc. Together with symmetric beta parameters, $a_{\boldsymbol{e}0} = a_{\boldsymbol{e}1}$, this implies $E(G(A)) = G^\star(A)$. We write

$$G \sim \text{PT}(G^\star, \mathcal{A}).$$

Example 2.2 (Galaxy data)**.** Roeder (1990) analyzes a data set with radial velocities (km/second) for 82 galaxies (Postman et al., 1986). The galaxies are located in six well-separated conic sections of the Corona Borealis region. Figure 3 shows a histogram of the data and the estimated density $\overline{F} = E(F \mid \boldsymbol{x})$ under a PT prior on F. Inference was implemented using the function `PTdensity` in `DPpackage` . See the code fragment below. See the package documentation for details on the function.

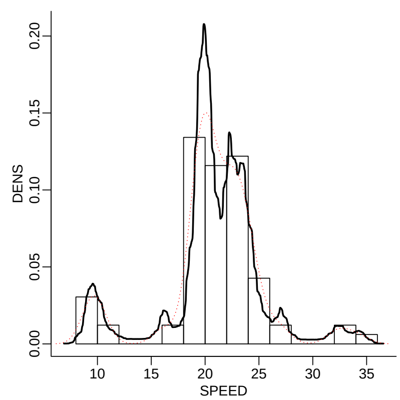


Figure 3: Estimated $\overline{F} = E(F \mid \boldsymbol{x})$ under a PT prior $F \sim \text{PT}(\Pi, \mathcal{A})$. For reference the histogram shows the data and the thin red line shows a kernel density estimate. Inference includes mixing with respect to c in $a_{\boldsymbol{e}} = cm^2$.

```
require(DPpackage)                                  ## cran.r-project.org/
data(galaxy)                                        ## Data
speeds<-galaxy$speed/1000
state <- NULL                                       ## Initial state
mcmc <- list(nburn=2000,nsave=5000,nskip=49,ndisplay=500,
             tune1=0.03,tune2=0.25,tune3=1.8)       ## MCMC parameters
prior<-list(a0=1,b0=0.01,M=6,m0=21,S0=100,sigma=20) ## Prior information
fit1 <- PTdensity(y=speeds,                         ## Fitting the model
  ngrid=1000,prior=prior,mcmc=mcmc, state=state,status=TRUE)
plot(fit1$x1, fit1$dens,                            ## estimated density
     xlab="SPEED",ylab="DENS", bty="l",type="l", lwd=2)
hist(speeds, nclass=12, add=T,prob=T)               ## add the data
dens <- density(speeds)                             ## add kernel density estimate
lines(dens$x,dens$y,type="l",col=2,lty=3)
```

3 Regression

Regression analysis assumes that a response y_i is generated from some underlying probability model F_{x_i} that is indexed by covariates x_i. In other words, we assume a family of probability models $\mathcal{F} = \{F_x;\ x \in X\}$, indexed by covariates x. For a particular observation y_i the assumed sampling model is the one indexed by the corresponding covariate x_i. If F_x is described by a finite dimensional parameter vectors, for example, $F_x = N(\beta' \boldsymbol{x}_i, \sigma^2)$, then inference reduces to learning about the parameter vector $\boldsymbol{\theta} = (\beta, \sigma^2)$, with the sampling model defined by

$$y_i = f_{\boldsymbol{\theta}}(x_i) + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma^2) \tag{3.1}$$

with some parametrized function $f_{\boldsymbol{\theta}}(x_i)$, such as $f_{\boldsymbol{\theta}}(x_i) = \beta' x_i$. A prior probability model on $\mathcal{F}$ is defined by assuming a prior $p(\boldsymbol{\theta})$ on the parameter vector and we are back to usual parametric inference.
In many problems, however, investigators are not able to restrict F_x to a parametric family. This leads to BNP to relax the mean function, the residual distribution or both in (3.1). Under this description it becomes natural to distinguish three types of BNP regression.

3.1 Partially nonparametric regression

Nonparametric residual distribution. Parametric mean function and unknown residual distribution

$$y_i = f_{\boldsymbol{\theta}}(x_i) + \epsilon_i, \text{ with } \epsilon_i \sim F,$$

with some BNP prior $p(F)$ on the residual distribution. This approach is explored, for example, in Hanson and Johnson (2002) who use a mixture of PT priors for $p(F)$. For a meaningful interpretation of F as a residual distribution it is important to restrict F to zero mean or median. One attraction of the PT prior is that it easy to restrict to zero median. Recall the earlier construction of the PT prior, and restrict the first level partition $B_0 \cup B_1$ to using a partition boundary at 0, and fix $Y_0 \equiv 0.5$. This restriction ensures a zero median.

Nonparametric mean function. Parametric residual distribution with nonparametric mean function,

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

and BNP prior $p(f)$ on f. A widely used prior for a random mean function f is the Gaussian process prior. A GP specifies a prior on f by assuming a multivariate normal for f evaluated at any finite set of covariate values x_i,

$$(f(x_1), \ldots, f(x_n)) \sim N(\boldsymbol{m}, S).$$

Here the (i, j) element of S is given by a covariance function $C(x_i, x_j)$ and the mean $\boldsymbol{m}$ is a mean function $\mu(x)$ evaluated at $x_1, \ldots, x_n$. We write

$$f \sim \text{GP}(\mu(\cdot), C(\cdot, \cdot)).$$

Bayesian inference under GP priors can be computationally intensive, essentially due to the $(n \times n)$ covariance matrix S, with typically all non-zero correlation. One solution is proposed by Gramacy and Lee (2008) who develop treed GP priors which avoid high-dimensional matrix factorization by partitioning the covariate space. The approach is implemented in the R package `tgp`.

3.2 Fully nonparametric regression

In the third case neither mean function nor residual distribution are restricted to a parametric form, leaving $\mathcal{F} = \{F_x;\ x \in X\}$ as the unknown quantity and assuming

$$y_i \mid \boldsymbol{x}_i = \boldsymbol{x} \sim F_{\boldsymbol{x}}.$$

To proceed with Bayesian inference we need to complete the inference model with a prior $p(\mathcal{F})$ on a set of random probability measures indexed by x.

Dependent DP. The by far most popular such prior is the dependent DP (DDP) (MacEachern, 1999). The construction is actually quite simple. Recall the stick breaking representation (2.2) of the DP prior,

$$F_x = \sum_h w_{xh} \delta_{m_{xh}}, \tag{3.2}$$

$x \in X$. We have slightly modified the stick-breaking representation for the upcoming discussion by adding a second index $_x$ on weights w_{xh} and locations m_{xh}. The DP construction involved then the independent beta fractions to generate w_{xh} and i.i.d. m_{xh}. Importantly, independence is across h. Across x we are free to introduce any construction. That is exactly the idea of the DDP. We define m_{xh} as a realization of a stochastic process $\{\mu_h(x)\}_x$, indexed by x. For example, this could be a GP over x. There is one realization $\{\mu_h(x)\}$ for each h, and they are independent across h. In the simplest DDP construction $w_{xh} = w_h$ are shared across all x. This is all. The same description in other words: For each x we generate a DP random measure F_x, including independence of the point mass locations m_{xh} across h. For different x_1 and x_2, the point masses $m_{x_1 h}, m_{x_2 h}$ for the same h are dependent. We introduce this dependence using a GP prior for $\mu_h(x) = m_{xh}$. The weights are generated as before by independent beta distributed random fractions of a unit total probability mass. We write

$$\{F_x;\ x \in X\} \sim \text{DDP}(M, GP(\mu(\cdot), C(\cdot,\cdot)) \tag{3.3}$$

for a DDP with GP prior to introduce the dependence across x on the point masses. Other variations of the DDP introduce dependence on weights w_{xh} and/or locations m_{xh}. But the basic principle remains the same. Convoluting F_x in (3.3) with an additional normal kernel to obtain continuous random probability measures we get

$$G_x = \sum w_{xh} N(\mu_h(x), \sigma^2) = \int N(m, \sigma^2)\ dF_x(m),$$

with $\{F_x\} \sim$ DDP.

ANOVA-DDP (LDDP). A particularly simple version of the DDP arises when we replace the GP prior for the dependent (across x) locations by a simple linear model, that is, $\mu_h(\boldsymbol{x}) = \beta_h' \boldsymbol{x}$ with $\beta_h \sim G^\star$ (De Iorio et al., 2009). De Iorio et al. (2009) refer to the model as DDP-ANOVA, having in mind the case when $\boldsymbol{x}$ indicates categorical factors. Already including the convolution with the normal kernel Jara and Hanson (2011) refer to the model as LDDP (linear dependent DP). We write

$$\{F_x\} \sim \text{ANOVA-DDP}(M, G^\star, X, \sigma^2),$$

where X is the design matrix with i-th row $\boldsymbol{x}_i$ (or some function of $\boldsymbol{x}_i$). For an application of the DDP-ANOVA model specifically for survival analysis see De Iorio et al. (2009). Below is an example using the `R` package `ddpanova`, available from `www.math.utexas.edu/users/pmueller/prog.html` (as "ANOVA-DDP univariate").

Example 3.1 (Oral cancer)**.** We use a dataset from Klein and Moeschberger (2003, Section 1.11). The data report survival times y_i for $n = 80$ oral cancer patients. Samples are classified as one of

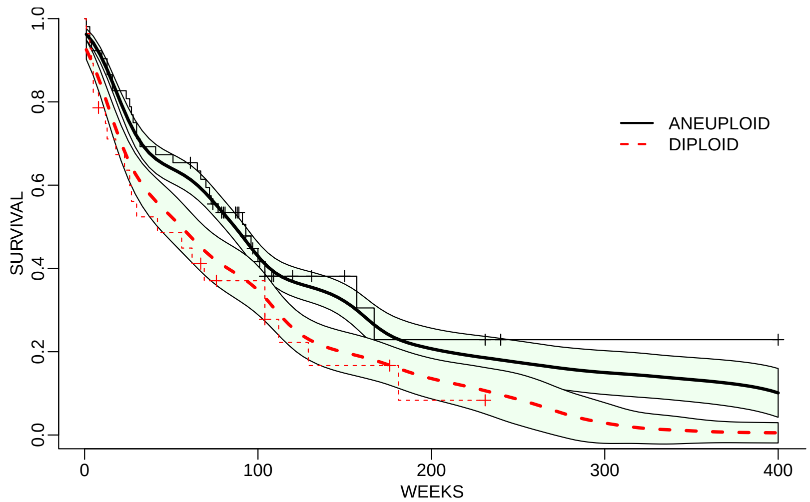


Figure 4: Estimated survival function by tumor type (solid black and dashed red curves). The grey shaded bands around the estimated survival functions show pointwise ± 1.0 posterior standard deviation bounds. The piecewise constant lines plot the Kaplan-Meier estimates.

two types, aneuploid ($x_i = 1$) versus diploid ($x_i = 0$). We use the R function `ddpsurvival()` from the package `ddpanova` to estimate an ANOVA DDP model for survival. The only covariate is an indicator for type. Posterior inference is shown in Figure 4.

```
require(KMsurv)                        ## from R CRAN
require(ddpanova)                      ## from www.math.utexas.edu/pmueller/prog
## tongue data from Section 1.11, Klein & Moeschberger (2003)
data(tongue); attach(tongue)           ## data
Y <- cbind(time,delta,time)
D = cbind(1, ifelse(type==2,1,-1)) ## design matrix
D0= cbind(1, c(-1,1))                  ## design matrix for prediction
ddpsurvival(Y,D,n.iter=3000,d0=D0,S.init=100,S.prior=0) ## fit
pp <- post.pred()                      ## posterior predictive
matplot(pp$ygrid,t(pp$Sy),type="l")
fit <- survfit(Surv(time,delta)~type)  ## add KM plot
lines(fit,col=1:2,bty="l",lty=1:2)
```

Inference under the LDDP, that is, DDP-ANOVA with an additional normal kernel, is also implemented in the function `LDDPsurvival` in `DPpackage` . The implementation includes the possibility of interval censored observations, like in the following example.

Example 3.2 (Breast retraction data.)**.** Hanson and Johnson (2004) analyze data on the time to cosmetic deterioration of the breast for women with stage 1 breast cancer who have undergone a lumpectomy (Beadle et al., 1984). Women were assigned to one of two treatments, A ($x_i = 0$) or B ($x_i = 1$). There are $n_B = 46$ patients under A and $n_A = 48$ patients under B. The outcome is time y_i to moderate or severe breast retraction. Event times are interval censored, with interval endpoints occurring at clinic visits. We fit the data using the ANOVA DDP model. The only predictor is the treatment indicator x_i. Below is the `R` code to implement inference in `DPpackage` . See the `DPpackage` documentation for the meaning of the hyperparamers in `prior`. Inference summaries are shown in Figure 5.

```
require(DPpackage)                                   ## cran.r-project.org/
data(deterioration); attach(deterioration)
ymat <- cbind(left,right)                            ## data
zpred <- rbind(c(1,0),c(1,1))     ## design matrix for posterior predictive
S0=diag(100,2) m0=rep(0,2) psiinv=diag(1,2)          ## Prior
prior <- list(a0=10, b0=1, nu=4, m0=m0, S0=S0, psiinv=psiinv,
              tau1=6.01, taus1=6.01, taus2=2.01)
state <- NULL                                        ## initial state
mcmc <- list(nburn=5000, nsave=5000, nskip=3, ndisplay=100) ## MCMC pars
fit1 <- LDDPsurvival(ymat~trt,prior=prior,           ## fit model
        mcmc=mcmc,state=state,status=TRUE, grid=seq(0.01,70,1),zpred=zpred)

plot(fit1$grid,fit1$survp.h[1,],type="l",            ## x0=(1,0)
     xlab="TIME",ylab="SURVIVAL",lty=2,lwd=1,ylim=c(0,1),bty="l")
lines(fit1$grid,fit1$survp.l[1,],lty=2,lwd=1)
lines(fit1$grid,fit1$survp.m[1,],lty=1,lwd=3)
lines(fit1$grid,fit1$survp.h[2,],lty=2,lwd=2,col=2)    ## Add: x0=(1,1)
lines(fit1$grid,fit1$survp.l[2,],lty=2,lwd=2,col=2)
lines(fit1$grid,fit1$survp.m[2,],lty=1,lwd=3,col=2)
```

Recent literature includes almost endless variations of similar constructions. Some examples are the order based DDP of Griffin and Steel (2006), the probit stick-breaking model (PSBP) of Chung and Dunson (2008) and the weighted mixture of DPs (WMDP) of Dunson et al. (2007). The order based DDP introduces the desired dependence across F_x by permuting the weights in a systematic fashion as x changes. The PSBP parametrization uses a representation like (1), but with covariate-dependent weights w_{xh} and common point masses m_h. The weights are explicitly parametrized as a regression on x. The WMDP assumes that the random distributions F_x are weighted mixtures of independent random probability distributions F^o_ℓ. The weights are functions of the covariates.

Dependence by additive constructions. Müller et al. (2004) consider a variation of the DDP mixture of normal model for the special case when $x \in \{1, 2, \ldots, k\}$ indexes k related studies. We define $p(\mathcal{F})$ by assuming an additive decomposition of the mixing measure G_x for F_x, as

$$G_x = \epsilon H_0 + (1-\epsilon) H_x \text{ and } H_j \sim \text{DP}(M, H^\star),$$

independently across $j = 0, 1, \ldots, k$. The construction has a natural interpretation when the F_x are distributions for patient-specific random effects in related studies $x = 1, \ldots, k$. The model reflects heterogeneity of patient populations, with H_0 representing a subpopulation that is common across studies and H_x representing patient subpopulations specific to each study. A similar construction,

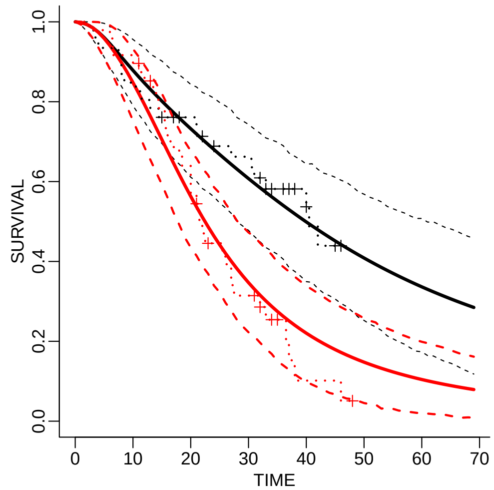


Figure 5: Estimated survival curves for $x = (1, 0)$ (black) and $x = (1, 1)$ (red) and pointwise 95% HPD intervals (dashed lines). For comparison the dotted line shows a Kaplan Meier estimate.

but in much more generality and in such a way that G_x is again a well known process is introduced in Lijoi et al. (2014).

3.3 Conditional regression

We introduced fully nonparametric regression using BNP priors on $\mathcal{F} = \{F_x;\ x \in X\}$. An alternative approach reduces regression to density estimation, using the following model augmentation. Note that in the earlier construction we used $\boldsymbol{x}_i$ only to select one of the models in $\mathcal{F}$. There was no notion of a probability model for $\boldsymbol{x}_i$. But for a moment pretend that $\boldsymbol{x}_i$ were also random. In observational studies this is a reasonable assumption. Let $\tilde{\boldsymbol{y}}_i = (y_i, \boldsymbol{x}_i)$ denote an augmented outcome vector and assume

$$\tilde{\boldsymbol{y}}_i \sim F \tag{3.4}$$

$i = 1, \ldots, n$, i.i.d., and complete the inference model with a BNP prior $p(F)$ on F. The problem is now reduced to a density estimation problem on F. We proceed as before, in Section 2. The implied conditional distribution under F, that is $F(y \mid \boldsymbol{x}) \propto F(y, \boldsymbol{x})$, as a function of y for fixed $\boldsymbol{x}$, solves the original regression problem. The conditional $F(y \mid \boldsymbol{x})$ is the desired model F_x. Note that here and elsewhere we use generic notation F for a probability model, using the arguments to clarify the specific use (joint, conditional etc.).

Müller et al. (1996) and Park and Dunson (2010) propose this approach using a DP mixture model for inference on the unknown joint distribution $F(y, \boldsymbol{x})$. The implied regression mean function is

$$f(\boldsymbol{x} \mid F) = E_F(y \mid \boldsymbol{x})$$

with the expectation being with respect to y (under the implied conditional $F(y \mid \boldsymbol{x})$). The posterior estimated mean function becomes $\bar{f}(\boldsymbol{x}) = E\{f(\boldsymbol{x} \mid F) \mid data\}$ with the additional expectation being with respect to the posterior on F. The mean function $f(\boldsymbol{x} \mid F)$ under this approach takes the form of a locally weighted linear regression line, similar to traditional kernel regression in classical nonparametric inference. In words, this is the case, because a (DP) mixture of normal model for $(y_i, \boldsymbol{x}_i)$ implies a locally weighted mixture of linear regressions for $p(y \mid x, data)$ for a future observation. For a detail statement, consider a DP mixture of normal kernels, mixing with respect to location and scale. Write the DPM as a hierarchical model as in (2.5),

$$\begin{aligned} (x_i, y_i \mid \mu_i, \Sigma_i) &\sim \mathrm{N}(\mu_i, \Sigma_j) \\ \theta_i \equiv (\mu_i, \Sigma_i) \mid G \sim G \quad &\text{and} \quad G \sim \mathrm{DP}(M, G_0). \end{aligned} \tag{3.5}$$

Let $\theta_k^\star = (\mu_k^\star, \Sigma_k^\star)$, $j = 1, \ldots, K$, denote the unique values of θ_i, $i = 1, \ldots, n$, with multiplicities n_k. Let $g(y \mid x, \theta_k^\star)$ denote the conditional normal density in y given x under the multivariate normal $\mathrm{N}(\mu_k^\star, \Sigma_k^\star)$ and let $s(x \mid \theta_k^\star)$ denote the marginal normal density in x under $\mathrm{N}(\mu_k^\star, \Sigma_k^\star)$. Similarly, let $g_0(y \mid x)$ and $s_0(x)$ denote the implied conditional and marginal when $\theta^\star$ is generated from $G^\star(\theta^\star)$, i..e., $g_0(y \mid x) = \int g(y \mid x, \theta)\, \mathrm{d}G^\star(\theta)$ and $s_0(x) = \int s(x \mid \theta)\, \mathrm{d}G^\star(\theta)$. Now consider a future observation θ_{n+1} and write (x, y) as short for (x_{n+1}, y_{n+1}). We get the predictive distribution

$$p(y \mid x, \theta_1^\star, \ldots, \theta_k^\star) \propto M\, s_0(x) g_0(y \mid x) + \sum_{k=1}^{K} n_k\, s(x \mid \theta_k^\star)\, g(y \mid x, \theta_k^\star). \tag{3.6}$$

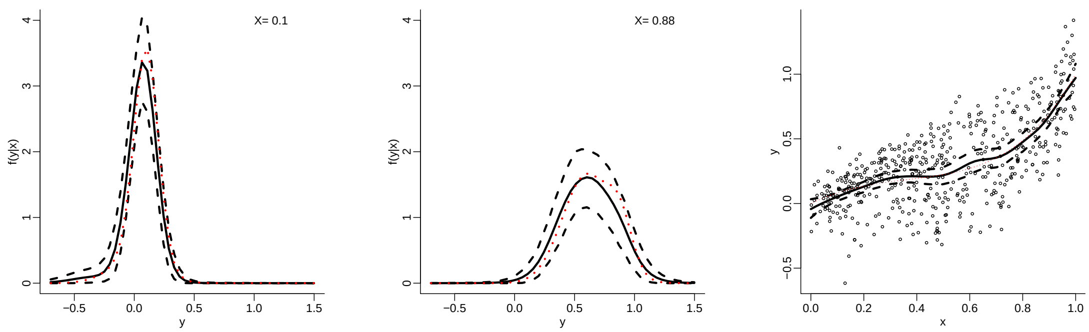
The predictive $p(y \mid x, \theta_1^\star, \ldots, \theta_K^\star)$ takes the form of a locally weighted mixture of linear regressions, each regression line being indexed by a unique $\theta_k^\star$, and the weights being the normal kernels $n_k s(x \mid \theta_k^\star)$. Plus one term corresponding to the base measure $G^\star$.

Example 3.3 (Simulation example.)**.** We use a simulation setup from Dunson et al. (2007) to generate $n = 500$ observations from a mixture of two normal linear regression models,

$$y_i \mid x_i \overset{ind.}{\sim} e^{-2x_i} N(y_i | x_i, 0.01) + \left(1 - e^{-2x_i}\right) N(y_i | x_i^4, 0.04), \quad i = 1, \ldots, n,$$

and $x_i \overset{iid}{\sim} U(0, 1)$. Inference under DPM conditional regression is implemented in the `DPpackage` function `DPcdensity`. Below is the `R` code. See the package documentation for the interpretation of the hyperparameters in `prior`.

```
require(DPpackage)                                   ## cran.r-project.org/
nrec <- 500; x <- runif(nrec)                        ## generate the data
p <- exp(-2 * x)
y <- ifelse(runif(nrec) < p,
     x + rnorm(nrec, 0, sqrt(0.01)),
     x^4 + rnorm(nrec, 0, sqrt(0.04)))
w=cbind(y, x); wbar=apply(w, 2, mean); wcov=var(w) ## prior
prior <- list(a0 = 10, b0 = 1, nu1 = 4, nu2 = 4, s2 = 0.5 * wcov,
    m2 = wbar, psiinv2 = 2 * solve(wcov), tau1 = 6.01, tau2 = 3.01)
mcmc <- list(nburn=5000,nsave=5000,nskip=3,ndisplay=1000)    ## mcmc
```



(a) $F_x = p(y \mid x = 0.1, data)$ (b) $F_x = p(y \mid x = 0.88, data)$ (c) $\bar{f} = E(f \mid data)$

Figure 6: Panels (a) and (b) shows the estimated conditional $p(y \mid x, data)$ for a future data point with $x = 0.1$ (a) and $x = 0.88$ (b). The dashed black lines show pointwise 95% HPD intervals for the conditional density. For comparison the dotted red line shows the simulation truth. Panel (c) shows the estimated mean function $\bar{f}(x) = E(f(\boldsymbol{x} \mid F) \mid data)$. Again for comparison the dotted red curve shows the simulation truth.

```
xpred=seq(0,1,0.05)                               ## x-grid for fit
fit <- DPcdensity(y = y, x = x, xpred=xpred,      ## fit
    ngrid = 100, compute.band = TRUE, type.band = "HPD",
    prior = prior, mcmc = mcmc, state = NULL, status = TRUE)
    ## note, this might take a while.
plot(x, y, xlab = "x", ylab = "y",
     bty="l", pch=1, cex=0.5)                     ## E(f | dta)
lines(xpred, fit$meanfp.m, type = "l", lwd = 3, lty = 1)
lines(xpred, fit$meanfp.l, type = "l", lwd = 3, lty = 2)
lines(xpred, fit$meanfp.h, type = "l", lwd = 3, lty = 2)

j=6                ## plot cond p(y | x,data) for x=xpred[6]=0.1
plot(fit$grid, fit$densp.h[j,], ylim = c(0, 4),bty="l",
     lwd = 3, type = "l", lty = 2, xlab = "y", ylab = "f(y|x)")
lines(fit$grid, fit$densp.l[j,], lwd = 3, type = "l", lty = 2)
lines(fit$grid, fit$densp.m[j,], lwd = 3, type = "l", lty = 1)
```

Figure 6ab shows the estimated density $E(F_x \mid data)$ and pointwise 95% HPD intervals for $x = 0.1$ and $x = 0.88$. Panel (c) shows the data along with the estimated mean function $\bar{f}(\boldsymbol{x}) = E(f(\boldsymbol{x} \mid F) \mid data)$.

4 Classification

An interesting application of fully nonparametric regression arises when a categorical covariate x indexes different subpopulations of interest, and the aim of the study is to classify a new patient into one of these subpopulations. Without loss of generality assume $x \in \{0, 1\}$. Cruz-Mesía et al. (2007)

construct a BNP model that allows such classification. Let y_i denote the response for the i-th subject, and let x_i denote the classification into the two subpopulations. Assume that the classification x_i is known for $i = 1, \ldots, n$, and the unknown classification x_{n+1} for a future observation should be predicted on the basis of a partially observed response y_{n+1}. Cruz-Mesía et al. (2007) use an ANOVA-DDP for $p(y_i \mid x_i, \mathcal{F}) = F_x$, and augment the model by a simple additional assumption,

$$p(x_i = 1) = \pi.$$

That is, they add a prior probability model for the classification x_i. Under this simple augmentation the (marginal posterior) predictive probability $p(x_{n+1} = 1 \mid y_{n+1}, data)$ defines the desired classification for a future observation. In the following example $y_i = (y_{i1}, \ldots, y_{im_i})$ is a longitudinal response, allowing to update $p(x_{n+1} = 1 \mid y_{n+1,1\ldots,m}, data)$ with increasing number m of repeat observations.

Example 4.1 (Pregnancy classification)**.** De la Cruz et al. (2007) analyze hormone data y_{ij},for $n = 173$ pregnant women, $i = 1, \ldots, n$, for repeat measurements at times t_{ij}, $j = 1, \ldots, n_i$. The data include $n_0 = 124$ normal pregnancies ($x_i = 0$) and $n_1 = 49$ pregnancies that were classified as abnormal ($x_i = 1$). The data are modeled as a non-linear mixed-effects model

$$p(y_{ij} \mid x_i = x, \beta_x, \sigma_x^2, \theta_i) = N(m_{ij}, \sigma_x^2) \text{ with } m_{ij} = \theta_i \left[1 + \exp\left\{-(t_{ij} - \beta_{1x})/\beta_{2x}\right\}\right]^{-1},$$

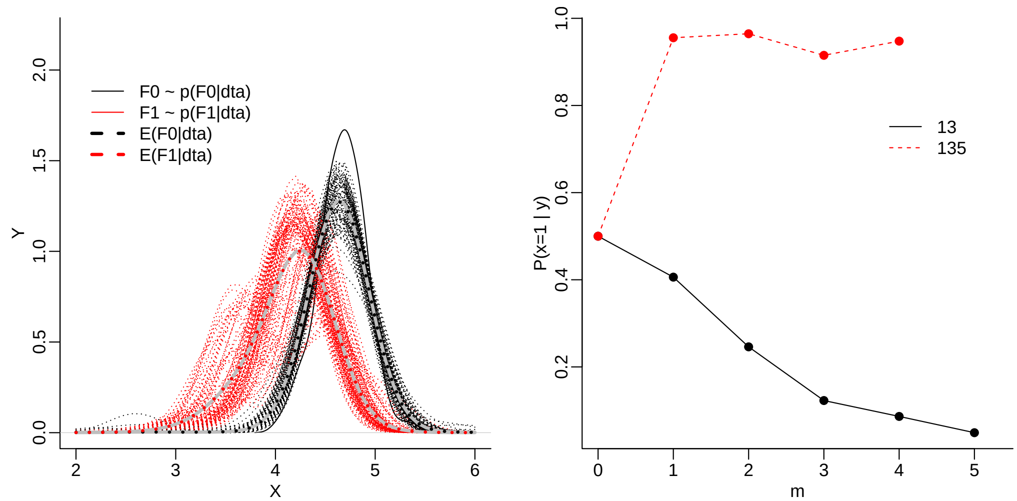
i.e., a logistic regression with coefficients β_x and scaled by random effects θ_i and with normal residuals. Fixed effects, β_x, σ_x^2 are group-specific. Let $\phi = (\beta_x, \sigma_x^2,\ x = 0, 1)$. The model includes a patient-specific random effect θ_i with $\theta_i \mid x_i = x \sim G_x(\theta_i)$, and an ANOVA DDP prior, $(G_0, G_1) \sim$ ANOVA-DDP$(M, G^\star, X, \tau^2)$ where X is a design matrix with i-th row $(1, 0)$ for normal pregnancies and $(1, 1)$ for abnormal pregnancies. The model is completed with a bivariate normal base measure $G^\star$ and conditionally conjugate priors for ϕ. Figure 7ab shows the estimated random effects distributions $E(F_x \mid data)$ (panel a) and the posterior classification probabilities $p(x_{n+1} = 1 \mid y_{n+1,1\ldots m}, data)$ as a function of m.

5 Conclusion

We have reviewed some popular BNP models, and showed how to implement inference for some of these models in R, using public domain software. BNP inference can be very useful when parametric models become too restrictive. However, while we tried to introduce a clear distinction between parametric and non-parametric inference by defining BNP as priors on infinite dimensional parameters, this distinction is not always as clear. Flexible parametric models, like finite mixture of normal models can be almost as flexible as BNP models, and often suffice for practical data analysis.

Acknowledgments

Peter Müller was partially funded by grant NIH R01 CA132891-06A1.



(a) $E(F_x \mid data)$ (b)$p(x_{n+1} = 1 \mid y_{n+1,1\ldots m}, data)$

Figure 7: Estimated random effects distributions F_x under $x = 0$ (thick black curve) and $x = 1$ (thick red) (panel a), and posterior probability $p(x_{n+1} = 1 \mid y_{n+1,1\ldots m}, data)$ for a future woman with unknown pregnancy status, as a function of hormone measurements $y_{n+1,j}$ over time. In panel (a) the think grey lines show posterior simulations $F_x \sim p(F_x \mid data)$. In panel (b), the red dashed line shows results for a simulated future woman with simulation truth of abnormal pregnancy, that is, $x_{n+1} = 1$ in the simulation ("$i = 135$"). The solid black curve shows the same for a woman who was simulated with a normal pregnancy ("$i = 13$").

References

Azzalini, A. and Bowman, A. W. (1990). A Look at Some Data on the Old Faithful Geyser. *Applied Statistics*, 39, 357–365.

Beadle, G., Harris, J., Silver, B., Botnick, L., and Hellman, S. (1984). Cosmetic results following primary radiation therapy for early breast cancer. *Cancer*, 54, 2911–2918.

Chung, Y. and Dunson, D. B. (2008). Nonparametric Bayes Conditional Distribution Modeling with Variable Selection. Tech. rep., Department of Statistical Science, Duke University.

Cruz-Mesía, R. D. l., Quintana, F. A., and Müller, P. (2007). Semiparametric Bayesian Classification with Longitudinal Markers. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 56, 119–137.

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2009). "Bayesian nonparametric non-proportional hazards survival modelling." *Biometrics*, 65, 762–771.

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

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

Escobar, M. D. (1988). Estimating the means of several normal populations by nonparametric estimation of the distributions of the means. Unpublished doctoral thesis, Deparment of Statistics, Yale University.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89, 268–277.

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

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

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of Normal distributions, in *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday*, eds. Siegmund, D., Rustage, J., and Rizvi, G. G., Bibliohound, pp. 287–302.

Ghoshal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

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.

Griffin, J. E. and Steel, M. F. J. (2006). Order-based Dependent Dirichlet Processes. *Journal of the American Statistical Association*, 101, 179–194.

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

Hanson, T. and Johnson, W. O. (2004). A Bayesian semiparametric AFT model for interval-censored data. *Journal of Computational and Graphical Statistics*, 13, 341–361.

Hjort, N. L. (2003). Topics in Nonparametric Bayesian Statistics, in *Highly structured stochastic systems*, eds. Green, P., Hjort, N., and Richardson, S., Oxford University Press, pp. 455–487.

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

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

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

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag Inc.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *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.

Lijoi, A., Nipoti, B., and Prünster, I. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli*, 20, 1260–1291.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates I: Density estimates. *The Annals of Statistics*, 12, 351–357.

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

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83, 67–79.

Müller, P. and Mitra, R. (2013). Bayesian nonparametric inference - why and how. *Bayesian Anal.*, 8, 269–302.

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

Müller, P., Quintana, F., and Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 66, 735–749.

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

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

Phadia, E. G. (2013). *Prior Processes and Their Applications*. Springer-Verlag.

Postman, M., Huchra, J. P., and Geller, M. J. (1986). Probes of large-scale structure in the Corona Borealis region. *Astronomical Journal*, 92, 1238–1247.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85, 617–624.

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

Walker, S. G., Damien, P., Laud, P. W., and Smith, A. F. M. (1999). Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 61, 485–509.