Handbook of Statistics, Vol. 25
ISSN: 0169-7161
© 2005 Elsevier B.V. All rights reserved.
DOI 10.1016/S0169-7161(05)25009-5

9

Bayesian Nonparametric Modeling and Data Analysis: An Introduction

Timothy E. Hanson, Adam J. Branscum and Wesley O. Johnson

Abstract

Statistical models are developed for the purpose of addressing scientific questions. For each scientific question for which data are collected, the truth is sought by developing statistical models that are useful in this regard. Despite the fact that restrictive parametric models have been shown to be extraordinarily effective in many instances, there is and has been much scope for developing statistical inferences for models that allow for greater flexibility. It would seem that just about any statistical modeling endeavor can be expanded and approached, at least conceptually, as a nonparametric problem. The purpose of this chapter is to give a brief discussion of, and introduction to, one of the two major approaches to the whole of statistics as it were, Bayesian nonparametrics.

1. Introduction to Bayesian nonparametrics

The term ‘nonparametric’ is somewhat of a misnomer. It literally connotes the absence of parameters. But it is usually the case that the goals of a data analysis include making inferences about functionals of an unknown probability measure, F, which are themselves parameters, regardless of whether the class of probability measures under consideration is quite broad (e.g., not indexed by parameters). Nonetheless, the spirit of the term ‘nonparametric’ is to be free of restrictive, inappropriate, or unrealistic constraints that are implied by particular parametric models. For example, it is often necessary to consider models that allow for unspecified multimodality, asymmetry and nonlinearity. This can be accomplished by considering a broad class of distributions and by making statistical inferences within that context. Semiparametric modeling involves incorporating parametric and nonparametric components into a single model, an example being a linear regression where the error distribution is allowed to be arbitrary subject to having median zero. Hundreds of frequentist nonparametric and semiparametric papers have been published. Classic methods were typically based on permutations and ranking, while with increases in computational capabilities, jackknifing and resampling methods have more recently played a major role. Bayesian and frequentist nonparametric regression modeling, density estimation, and smoothing remains an active area of research.

Parametric modeling has dominated the Bayesian landscape for many years. In the parametric setting, data are modeled according to a family of probability measures $\{F_\theta\colon \theta \in \Theta\}$ with corresponding probability density functions (pdf) $\{p(\cdot|\theta)\colon \theta \in \Theta\}$. Scientific evidence for θ, which is obtained independently of the current data, is used to construct a parametric "prior" pdf, $p(\mathrm{d}\theta)$. As a first step, the posterior pdf of θ, $p(\mathrm{d}\theta|\text{data})$, is obtained. The next steps usually involve finding various posterior characteristics such as medians or means, standard deviations, and probability intervals. Prediction is accomplished by integrating the sampling pdf for the future observation given the data against the posterior.

Nonparametric modeling begins with the specification of a broad class of models for the data at hand. For example, consider a single sample of data from an unknown distribution F. The goal is to make inferences about functionals of F, or possibly the pdf corresponding to F. We could simply assert that F belongs to $\mathcal{F}$, the class of all continuous distributions on the real line. Alternatively, standard regression data, $\{(y_i, x_i)\colon i = 1, \ldots, n\}$, can be modeled as $y_i|(x_i, f, \theta) \overset{\perp}{\sim} N(f(x_i), \theta)$, where $f \in \mathcal{F}^*$, a broad class of possible regression functions, and where $\theta \in (0, \infty)$. Bayesian approaches to these problems require specifying probability measures, $\mathcal{P}(\mathrm{d}F)$ and $\mathcal{P}^*(\mathrm{d}f)$ on $\mathcal{F}$ and $\mathcal{F}^*$, respectively, as well as a suitable parametric probability measure for θ. In general, constructing suitable $\mathcal{P}$'s on function spaces has been accomplished by a number of authors. Data analysis and applications involving these models were limited at first due to analytical intractability. However, the last fifteen years has seen a dramatic increase in nonparametric and semiparametric Bayesian modeling due to remarkable improvements in computational techniques and capabilities.

Müller and Quintana (2004) noted that Bayesian nonparametric models are also used to "robustify" parametric models and to perform sensitivity analyses. For example, the above regression problem includes standard parametric linear regression as a special case. Bayesian modeling can take specific account of this by constructing a prior $\mathcal{P}^*(\mathrm{d}f)$ that is centered on the parametric regression function. Along these lines, Ibrahim and Kleinman (1998) embedded the family of zero-mean normal models in a broader class of models for random effects in a generalized linear mixed model framework, and Berger and Gugliemi (1999) developed general Bayesian nonparametric (BNP) methodology for embedding a family of parametric models in a broader class for the purpose of determining the adequacy of parametric models.

In this chapter, we first discuss the basics of BNP modeling, e.g., the determination of suitable $\mathcal{P}$ to be defined on $\mathcal{F}$. This development begins with the Dirichlet process (DP) (Ferguson, 1973), the mixture of DP's (MDP) (Antoniak, 1974), the Dirichlet process mixture (DPM) (Antoniak, 1974; Escobar, 1994), the Polya tree (PT) (Lavine, 1992, 1994), mixtures of PT's (MPT) (Lavine, 1992) and the gamma process (GP) (Kalbfleisch, 1978). Special emphasis is given to the DPM, MPT and GP models so more details and/or illustrations are given for them. There are many other choices for $\mathcal{P}$, but we mainly focus on these. This material is like root stock, from which it is possible to grow more complex models and methods.

After this development, we present a variety of illustrations starting with an application to the independent two sample problem, and moving on to a variety of regression problems. The regression scenarios considered include (i) approaches to linear regression modeling with an unknown error distribution, which are illustrated in a survival analysis setting, (ii) nonlinear regression modeling with a parametric error distribution, which is illustrated on highly nonlinear data, and (iii) a fully nonparametric model where the regression function and the error distribution are modeled nonparametrically. Our presentation of nonparametric regression modeling of a mean function involves the representation of the mean function as an infinite linear combination of known basis functions (the coefficients are unknown). Bayesian modeling in this setting involves the truncation of the infinite series, resulting in a regression function specified as a finite linear combination. This can lead to a dimension varying linear model and requires specifying a joint prior probability distribution on the corresponding basis coefficients and (possibly) the number of basis functions to be included in the model. The resulting linear model is essentially a highly flexible parametric model so that standard parametric methods are applicable in fitting the semiparametric model. For this particular application, the fundamental background material is not needed.

We also discuss a variety of other modeling situations, but in less detail. We make no attempt to present an exhaustive discussion of Bayesian nonparametrics since it is possible to discuss *all* of inferential statistics from a BNP perspective, and this would be beyond the scope of any single article. We shall instead discuss basic ideas, provide some simple illustrations, and give the reader a taste of recent progress in a few important subfields.

The computing environment WinBUGS (Spiegelhalter et al., 2003) has made Bayesian modeling available to the masses. In our discussion, we indicate how this user-friendly software can be used to fit data to a number of non/semiparametric models. (Congdon 2001, Section 6.7) has examples of DPM and PT models fit in WinBUGS.

From here on we use the notation F to mean both a probability measure and its corresponding cumulative distribution function (CDF) where we trust that the context will make clear the distinction.

2. Probability measures on spaces of probability measures

In modeling a probability measure F as $F \sim \mathcal{P}(\mathrm{d}F)$, common choices of $\mathcal{P}$ are the DP, MDP, DPM, PT, MPT and GP (a primary application of the GP is in the area of survival analysis where the GP can be used to model the cumulative hazard function and thus induces a distribution on F). For many years, emphasis was placed on the DP due to its mathematical tractability in simple situations, however the DP prior was criticized because it places prior probability one on the class of discrete distributions. Although an MDP model can place mass on absolutely continuous distributions, the use of MDP's in data analysis was limited due to the complexity resulting from a computational explosion associated with possibilities for ties (Antoniak, 1974; Berry and Christensen, 1979; Johnson and Christensen, 1989).

The advent of modern BNP data analysis stems first from the development of Markov chain Monte Carlo (MCMC) technology starting with Gelfand and Smith (1990) and then from the observation by Escobar (1994) that these methods (in particular, Gibbs sampling) could be applied to DPM's after marginalization over the process F. There have been many papers that used DPM's for modeling and analyzing data since 1994 (for a small sampling see Dey et al., 1998). While PT priors have been discussed as early as Freedman (1963), Fabius (1964) and Ferguson (1974), the natural starting point for understanding their potential use in modeling data is Lavine (1992, 1994). The utility of using MPT's to generalize existing parametric families was illustrated by Berger and Gugliemi (1999) and Hanson and Johnson (2002). The GP model was used to model survival data in the context of the proportional hazards model by Kalbfleisch (1978) and we present a particular implementation of this model here.

Other general probability models for $\mathcal{P}$ have been developed by Freedman (1963) and Doksum (1974). In the particular area of survival analysis, there are a number of nonparametric and semiparametric models beyond the GP that have been developed that are based on modeling the hazard function and the cumulative hazard function (see, e.g., Dykstra and Laud, 1981; Ibrahim et al., 2001; Nieto-Barajas and Walker, 2002, 2004) but we do not discuss these here. Review articles by Müller and Quintana (2004), Walker et al. (1999), Gelfand (1999), Sinha and Dey (1997), the volume by Dey et al. (1998), the monograph by Ghosh and Ramamoorthi (2003), and the article in this volume by Choudhuri et al. (2004), all provide additional background and breadth beyond what we present here.

2.1. The Dirichlet process

Ferguson (1973) introduced the DP as a means to specify a (prior) probability measure $\mathcal{P}(\mathrm{d}F)$ on a probability measure F taking values in the space of all probability measures, $\mathcal{F}$, in the context of modeling statistical data. A random probability measure F is said to be a DP with parameter αF_0 if for all finite measurable partitions $\{A_j\}_{j=1}^J$ of the sample space, the vector $(F(A_1), F(A_2), \ldots, F(A_J))$ has a Dirichlet distribution with parameter $(\alpha F_0(A_1), \alpha F_0(A_2), \ldots, \alpha F_0(A_J))$. The parameter (αF_0) of a DP consists of a scalar precision parameter $\alpha > 0$ and a completely known base probability measure F_0. The DP is centered at F_0 in the sense that for any measurable set B, $E[F(B)] = F_0(B)$. The parameter α is referred to as a precision parameter because the prior variance for the probability of any measurable set, $\mathrm{Var}[F(B)] = \frac{F_0(B)[1-F_0(B)]}{\alpha+1}$, is small for large α. These results follow from the fact that $F(B) \sim \mathrm{Beta}(\alpha F_0(B), \alpha F_0(B^c))$. We write $F|\alpha, F_0 \sim \mathrm{DP}(\alpha F_0)$.

A key conjugacy result holds for the DP. Consider the model

$$\begin{aligned} y_1, y_2, \ldots, y_n | F &\overset{\text{i.i.d.}}{\sim} F \\ F|\alpha, F_0 &\sim \mathrm{DP}(\alpha F_0) \end{aligned}$$

and define $Y = (y_1, y_2, \ldots, y_n)$. Then the posterior distribution of F is $F|Y \sim \mathrm{DP}(\alpha^* F_0^*)$ where $\alpha^* = \alpha + n$ and $F_0^* = \frac{\alpha}{\alpha+n} F_0 + \frac{1}{\alpha+n} \sum_{i=1}^n \delta_{y_i}$; $\delta_y(\cdot)$ denotes point mass at y, e.g., $\delta_y(B) = 1$, if $y \in B$, and zero otherwise. Hence the posterior mean of the CDF $F(t)$ is given by

$$\widehat{F}(t) = E\big[F(t)|Y\big] = \frac{\alpha}{\alpha+n}F_0(t) + \frac{n}{\alpha+n}\widehat{F}_n(t),$$

where $\widehat{F}_n(t)$ is the empirical distribution function based on $(y_1, \dots, y_n)$. This is a common occurrence in Bayesian statistics that the estimate is a weighted average of the prior mean of F and an empirical estimate, in this instance the nonparametric maximum likelihood estimate.

In addition to estimating F, inferences for functionals, $T(F)$, are of interest. For instance, the mean functional is given by $E(y|F) = \int y\,\mathrm{d}F(y)$. Inferences for $T(F)$ can be obtained using the approach of Gelfand and Kottas (2002) where $\{F^j\colon j = 1, \dots, MC\}$ are simulated from the posterior distribution $F|Y$ and used to obtain the corresponding Monte Carlo sample of $T(F^j)$'s. We shall discuss this approach in detail for the DPM model.

Predictive inference for a future observation is also straightforward. The predictive distribution of a future observation y_f where $y_f|Y, F \sim F$ is F_0^*. This follows from the generalized Polya urn representation for the marginal distribution of Y (Blackwell and MacQueen, 1973).

There are two features of the DP that typically are viewed as its primary limitations. As previously indicated, the support of the DP distribution is the set of all discrete distributions (Ferguson, 1973; Blackwell, 1973). This can be visualized from the constructive definition of F (Sethuraman, 1994):

$$F = \sum_{j=1}^{\infty} V_j \delta_{\theta_j},$$

where with $W_i \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha)$, the V_j's are defined as $V_1 = W_1, \dots, V_j = W_j \prod_{r=1}^{j-1}(1 - W_r), \dots$, and $\theta_j \overset{\text{i.i.d.}}{\sim} F_0$. This is often referred to as the "stick-breaking" representation as the weights are defined in a way that the interval [0, 1] (the stick) is successively broken up or partitioned into pieces starting with the interval $[0, w_1]$, and then adding $[w_1, w_1 + (1 - w_1)w_2]$ etc. The lengths of each of the corresponding subintervals are the weights in the Sethuraman representation of F. The second drawback of the DP is that for any disjoint measurable sets B_1 and B_2, the correlation between $F(B_1)$ and $F(B_2)$ is negative, which for ("small") adjacent sets violates a belief that these two probabilities should be positively correlated.

2.2. Mixtures of Dirichlet processes

Centering the DP on a fixed F_0 may be appropriate for some applications but for the majority of applied problems centering the DP on a *family* of parametric distributions is preferable. The goal then is to embed a parametric family in the broad class of models $\mathcal{F}$.

The MDP model is specified as:

$$\begin{aligned} y_1, y_2, \dots, y_n|F &\overset{\text{i.i.d.}}{\sim} F \\ F|\alpha, F_\theta &\sim \text{DP}(\alpha F_\theta) \\ \theta &\sim p(\mathrm{d}\theta), \end{aligned}$$

where $\{F_\theta\colon \theta \in \Theta\}$ is a parametric family of probability models. The standard representation for the MDP is $F \sim \int \mathrm{DP}(\alpha F_\theta) p(\mathrm{d}\theta)$. This representation makes it clear that F is distributed as a literal mixture of DP's. Antoniak (1974) presented theoretical results for the MDP model and also gave a number of applications. In particular, Antoniak (1974) obtained the posterior pdf for θ, assuming absolutely continuous F_θ with pdf $p(\cdot|\theta)$, as:

$$p(\mathrm{d}\theta|Y) \propto p(\mathrm{d}\theta) \prod_{i=1}^{k} p\bigl(y_i^*|\theta\bigr),$$

where $\{y_i^*, i = 1, \ldots, k \leqslant n\}$ are the distinct y_j's. Also, for given θ, $F|Y, \theta \sim \mathrm{DP}(\alpha^* F_\theta^*)$ where $F_\theta^* = \frac{\alpha}{\alpha+n} F_\theta + \frac{1}{\alpha+n} \sum_{i=1}^n \delta_{y_i}$. Hence, inferences for functionals $T(F)$ can be obtained by first sampling $\theta^j \overset{\text{i.i.d.}}{\sim} p(\mathrm{d}\theta|Y)$, $j = 1, 2, \ldots, MC$, then (partially) sampling $F^j|Y, \theta^j$ from $\mathrm{DP}(\alpha^* F_{\theta^j}^*)$, and finally computing $T(F^j)$.

The posterior mean $E[F|Y] = \int F_\theta^* p(\mathrm{d}\theta|Y)$ provides an estimate of F and can be approximated by Monte Carlo integration, e.g.,

$$E[F|Y] \doteq \frac{1}{MC} \sum_{j=1}^{MC} F_{\theta^j}^*.$$

If $p(\mathrm{d}\theta)$ is conjugate to $p(y|\theta)$, $p(\mathrm{d}\theta|Y)$ is easily sampled. Otherwise, sampling from $p(\mathrm{d}\theta|Y)$ can be accomplished, for instance, using a Metropolis sampler (Tierney, 1994).

Briefly consider a BNP version of the classic empirical Bayes problem. Let $y_i|\theta_i \overset{\text{ind}}{\sim} F_{\theta_i}, \theta_i|G \overset{\text{i.i.d.}}{\sim} G$, $G \sim \mathrm{DP}(\alpha G_0)$, $i = 1, \ldots, n$. This model can be represented as $y_i|F \overset{\text{i.i.d.}}{\sim} F \equiv \int F_\theta G(\mathrm{d}\theta)$, $G \sim \mathrm{DP}(\alpha G_0)$. The definition of F here corresponds to the definition of a DPM in the next subsection. Antoniak (1974) established in his Corollary 3.1 that the posterior distribution of $F|y$ (for a single y) can be represented as an MDP, namely $F|y \sim \int \mathrm{DP}((\alpha + 1)(G_0 + \delta_\theta)) p(\mathrm{d}\theta|y)$. Thus there is a connection between the MDP and the DPM models. But aside from that, computational complexities arise using this model for the empirical Bayes problem as soon as one attempts to characterize the full posterior distribution. From Corollary 3.2 of Antoniak (1974), and with $\theta = (\theta_1, \ldots, \theta_n)$, we have $F|Y \sim \int \mathrm{DP}((\alpha + n)(wF_0 + (1 - w) \sum_{i=1}^n \delta_{\theta_i}/n)) p(\mathrm{d}\theta|Y)$, $w = \alpha/(\alpha + n)$. It is here where Berry and Christensen (1979) and Lo (1984) realized how complicated the problem is due to the discreteness of the distribution of $\theta|Y$. A brute force approach to the problem must consider all possible combinations of ties among the θ_i's. The Monte Carlo approach of Escobar (1994) made it possible to actually analyze data modeled as a DPM.

2.3. Dirichlet process mixture models

The DPM model has been very popular for use in BNP inference. A standard parametric model that strives to achieve flexibility is the finite mixture model

$$y_i \overset{\text{i.i.d.}}{\sim} \sum_{j=1}^{K} p_j F_{\theta_j},$$

where $\{F_\theta\colon \theta \in \Theta\}$ represents a standard parametric family, $\theta_j \in \Theta$ for $j = 1, \ldots, K$ are assumed to be distinct so the mixture is comprised of K distinct members of this family. The fixed unknown mixing probabilities $\{p_j,\ j = 1, \ldots, K\}$ add to one and there are additional constraints that insure identifiability (Titterington et al., 1985). Bayesian inference for this model is achieved by placing a prior distribution on K, $\{p_j, j = 1, \ldots, K\}$, and $\{\theta_j, j = 1, \ldots, K\}$. Such a model results in a varying dimensional parameter space and consequently specialized computational techniques, such as reversible jump MCMC (Green, 1995), are required.

The DPM model avoids such concerns as the data are modeled according to an infinite mixture model which, using the Sethuraman (1994) representation, is given by

$$y_i \overset{\text{i.i.d.}}{\sim} \sum_{j=1}^{\infty} V_j F_{\theta_j},$$

where the F_{θ_j}'s are parametric CDFs (the CDFs that would be used in a finite mixture model) with V_j and θ_j defined as in the DP. Here the (implied) induced prior on the θ_j's is that they are i.i.d. from the base measure (F_0) of the DP. This representation of the model makes clear that the DPM model is equivalent to selecting an infinite mixture and where the DP prior induces the specified distribution on the weights and the θ's. So while the DPM generalizes the Bayesian version of the finite mixture model above by allowing for an infinite mixture, it does so at the expense of having a particular prior for these inputs. With a small weight α selected for the DP, the DP places high probability on a few nonnegligible components. In this instance, the DPM model effectively results in a finite mixture model but where it is not necessary to specify the number of components of the mixture in advance. The data are allowed to determine the likely number of mixture components.

Alternatively, the DPM model is specified as

$$y_1, y_2, \ldots, y_n | F \overset{\text{i.i.d.}}{\sim} F(\cdot|G) = \int F_\theta(\cdot) G(\mathrm{d}\theta), \quad G|\alpha, G_0 \sim \mathrm{DP}(\alpha G_0).$$

Because G is a random probability measure, F is a random probability measure. Note that if F_θ is continuous, then $F(\cdot|G)$ is also continuous with probability one. Thus the DPM model does not suffer the same fate as the DP in this regard.

An equivalent (and more commonly used) DPM model specification introduces latent variables as discussed at the end of the previous section:

$$\begin{aligned} &y_i|\theta_i \overset{\perp}{\sim} F_{\theta_i} \\ &\theta_i|G \overset{\text{i.i.d.}}{\sim} G \\ &G|\alpha, G_0 \sim \mathrm{DP}(\alpha G_0). \end{aligned}$$

Contributions related to fitting DPM models include the work of Escobar (1994), MacEachern (1994), Escobar and West (1995), Bush and MacEachern (1996), MacEachern and Müller (1998), Walker and Damien (1998), MacEachern et al. (1999), and Neal (2000). Contributions related to obtaining inferences for F and functionals $T(F)$ for DPM models have been provided by Gelfand and Mukhopadhyay (1995), Mukhopadhyay and Gelfand (1997), Kleinman and Ibrahim (1998), Gelfand and Kottas (2002), and Regazzini et al. (2002) among many others.

We now proceed to discuss details of fitting the basic DPM model and some of its extensions since it is perhaps the single most important BNP model to date.

2.3.1. Fitting DPM models

A Monte Carlo approach to approximating the posterior distribution of $T(F)$ would involve sampling the infinite-dimensional parameter G. Such an approach cannot be implemented without introducing finite approximations. Escobar (1994) considered the DPM model obtained after marginalizing the DP. This reduces the problem to sampling only the finite-dimensional variables $(\theta_1, \dots, \theta_n)$ as will be seen below. Using the third characterization of the DPM, Escobar obtained a numerical approximation to the posterior of the vector $\theta = (\theta_1, \dots, \theta_n)$ using Gibbs sampling, e.g., by iteratively sampling $\theta_i|\theta_{-i}, y_i$ where θ_{-i} denotes the vector of all θ_j's excluding θ_i.

The marginalized DPM model is given by

$$\begin{aligned} &y_i|\theta_i \overset{\perp}{\sim} F_{\theta_i} \\ &p(\theta_1, \theta_2, \dots, \theta_n) = p(\theta_1)p(\theta_2|\theta_1)p(\theta_3|\theta_1, \theta_2)\cdots p(\theta_n|\theta_{1:n-1}), \end{aligned}$$

where $\theta_{1:i-1} = (\theta_1, \theta_2, \dots, \theta_{i-1})$, $i = 2, \dots, n$, and dependence of the distribution for θ on (α, G_0) has been suppressed. The generalized Polya urn scheme (Blackwell and MacQueen, 1973) is used to specify $p(\theta_1, \theta_2, \dots, \theta_n)$ as

$$\theta_1 \sim G_0$$

$$\theta_i|\theta_{1:i-1} \begin{cases} \sim G_0 & \text{with probability } \frac{\alpha}{\alpha+i-1}, \\ = \theta_j & \text{with probability } \frac{1}{\alpha+i-1},\ j = 1, 2, \dots, i-1. \end{cases}$$

This follows from the fact that, for an appropriate measurable set A, $\Pr(\theta_i \in A|\theta_{1:i-1}) = E[G(A)|\theta_{1:i-1}]$. For $i = 1$, we have $E[G(A)] = G_0(A)$. For $i > 1$, since $G|\theta_{1:i-1}$ is an updated DP, we have

$$G(A)|\theta_{1:i-1} \sim \text{Beta}\left(\alpha G_0(A) + \sum_{j=1}^{i-1} \delta_{\theta_j}(A), \alpha G_0\bigl(A^c\bigr) + \sum_{j=1}^{i-1} \delta_{\theta_j}\bigl(A^c\bigr)\right).$$

Hence $E[G(A)|\theta_{1:i-1}] = \frac{\alpha}{\alpha+i-1} G_0(A) + \frac{1}{\alpha+i-1} \sum_{j=1}^{i-1} \delta_{\theta_j}(A)$, which yields the above result.

Combining the pdf for $\theta_i|\theta_{1:i-1}$ with the contribution $p(y_i|\theta_i)$ and because the latent θ_j's are exchangeable, the full conditional for θ_i is:

$$\theta_i|\theta_{-i}, y_i \begin{cases} = \theta_j & \text{with probability } \frac{p(y_i|\theta_j)}{A(y_i) + \sum_{j\neq i} p(y_i|\theta_j)},\ j \neq i, \\ \sim p(\mathrm{d}\theta_i|y_i) & \text{with probability } \frac{A(y_i)}{A(y_i)+\sum_{j\neq i} p(y_i|\theta_j)}, \end{cases} \tag{1}$$

where $A(y_i) = \alpha \int p(y_i|\theta) G_0(\mathrm{d}\theta)$ and $p(\mathrm{d}\theta_i|y_i)$ is the conditional pdf for θ_i given the single observation y_i based on a parametric model with likelihood contribution $p(y_i|\theta_i)$

(the pdf corresponding to F_{θ_i}) and prior distribution G_0 on θ_i. Sampling these full conditional distributions will be straightforward if $p(y_i|\theta_i)$ and G_0 are a conjugate pair so that computing $A(y_i)$ and sampling $p(\mathrm{d}\theta_i|y_i)$ will be routine. Such models are referred to as conjugate DPM models. Escobar and West (1995) considered a generalization of this model with $y_i|\mu_i, \sigma_i^2 \sim N(\mu_i, \sigma_i^2)$ and $G_0(\mathrm{d}\mu, \mathrm{d}\sigma^2) = N(\mathrm{d}\mu|m, \tau\sigma^2)IG(\mathrm{d}\sigma^2|a, b)$, a normal/inverse gamma base measure.

Although fitting a conjugate DPM model using the Gibbs sampler above is straightforward, the Gibbs sampler will often exhibit slow convergence to the joint marginal posterior, and once convergence is achieved, subsequent sampling of the θ_i's may be very inefficient, as discussed by Neal (2000). This is due to the discreteness of the DP. The θ's will cluster at each iteration of the Gibbs sampler, namely there will be a vector of distinct values of $(\theta_1, \dots, \theta_n)$, say $\phi = (\phi_1, \dots, \phi_k)$ for $k \leqslant n$. The inefficiency results from ignoring this fact in the Gibbs sampler described above.

MacEachern and Müller (1998) overcome this problem by using the following sampling approach for conjugate DPM models. At a given iteration of the Gibbs sampler, let the vector $c = (c_1, c_2, \dots, c_n)$ denote the cluster membership of y_i so that $c_i = j$ if $\theta_i = \phi_j$ for $i = 1, 2, \dots, n$, and $j = 1, \dots, k$. The current state of the Markov chain is (c, ϕ). The actual sampling is accomplished in two steps: (i) Sample θ_i as previously described but only for the purpose of determining the cluster membership c_i of each y_i. This involves the possibility of adding a new value of θ or sampling one of the current values in the vector ϕ. If a new value is added, the vector ϕ is augmented to include the new value and $k \to k + 1$. It is also possible that in sampling a θ_i when the current value of θ_i has only multiplicity one (e.g., $c_i = j, \sum_l \delta_j(c_l) = 1$), the new value will be one of the θ_{-i} values so that the vector ϕ must be redefined to accommodate its removal from the collection and hence $k \to k - 1$ in this instance. (ii) Then generate ϕ_j by sampling from the posterior distribution of ϕ_j based on the parametric model with likelihood $p(\cdot|\phi_j)$ and prior G_0 on ϕ_j where the posterior distribution is computed using only the y_i's that belong to cluster j. With this approach, all the θ_i's associated with a given cluster will be updated to a new value simultaneously.

MacEachern and Müller (1998) and Neal (2000) developed and discussed methods for sampling nonconjugate DPM models. Such methods are necessary, for example, if the data are assumed to be normally distributed conditional on $\theta = (\mu, \sigma^2)$ but where the DP $G(\mathrm{d}\mu, \mathrm{d}\sigma^2)$ is centered on $G_0(\mathrm{d}\mu, \mathrm{d}\sigma^2) = N(\mathrm{d}\mu|a, b)\Gamma(\mathrm{d}\sigma^2|c, \mathrm{d})$ instead of the usual conjugate normal–gamma distribution. Alternatively, let F_θ denote a Poisson(θ) distribution and assume $G(\mathrm{d}\theta)$ is centered a log-normal distribution.

The issue that remains is how to use the MC samples from the marginal posterior of θ in order to make inferences. There are some inferences that can be made and some that cannot. For example, it is not possible to obtain interval inferences for the unknown CDF $F(\cdot|G)$, or the population mean $\int yF(\mathrm{d}y|G)$ based solely on an MC sample from $p(\mathrm{d}\theta|Y)$. In general, for the marginalized DPM model, full inferences are not available for arbitrary functionals of $F(\cdot|G)$ because G is not sampled. Subsection 2.3.3 addresses these issues. However, as pointed out by Gelfand and Mukhopadhyay (1995), it is possible to obtain posterior expectations of linear functionals. For example, let $p(\cdot|\theta^*)$ denote the pdf for a sampled observation were the value of θ^* to be known. Then the modeled sampling density is $p(\cdot|G) = \int p(\cdot|\theta^*)G(\mathrm{d}\theta^*)$. Let $T(p(\cdot))$ be a linear functional of an arbitrary pdf $p(\cdot)$. Then it is not difficult to show that (see Gelfand and Mukhopadhyay, 1995)

$$\int T\big(p(\cdot|G)\big)p(\mathrm{d}G|Y) = \int T\big(p\big(\cdot|\theta^*\big)\big)p\big(\mathrm{d}\theta^*|\theta\big)p(\mathrm{d}\theta|Y).$$

Having obtained a sample from the marginal posterior for (θ^*, θ) (using (1) to obtain the full conditional for θ^*), the above integral is easily approximated. So clearly it is possible to obtain MCMC approximations to the posterior mean of the conditional mean $E(y|\theta^*)$, and also the pdf $p(y|\theta^*)$, and corresponding CDF $F_{\theta^*}(y)$, for all y. West et al. (1994) catalogue very interesting applications of DPM's to multivariate multimodal density estimation and random coefficient growth curves. Kottas and Gelfand (2001a) modeled semiparametric survival data with DPM's and showed how to make inferences for the median time to survival functional.

2.3.2. Extensions

Three extensions of the basic DPM model include the incorporation of covariates for semiparametric regression, a prior distribution for α, and centering the DP on a family of parametric distributions G_η with a prior distribution specified for η.

Perhaps the most important extension involves the incorporation of covariates into the model. Gelfand (1999), Kleinman and Ibrahim (1998), Mukhopadhyay and Gelfand (1997), and Bush and MacEachern (1996) discussed semiparametric regression for the DPM model. The basic model is given by:

$$\begin{aligned} y_i|\theta_i, x_i, \beta &\overset{\perp}{\sim} p(y_i|\theta_i, x_i, \beta)\\ \theta_i|G &\overset{\text{i.i.d.}}{\sim} G\\ G|\alpha, G_0 &\sim \mathrm{DP}(\alpha G_0)\\ \beta &\sim p(\mathrm{d}\beta), \end{aligned}$$

where y_i denotes the response for subject i with covariate vector x_i, and the θ_i's are random effects. The model is fitted using Gibbs sampling where the θ_i's are sampled from the full (marginal) conditional distribution corresponding to $p(\mathrm{d}\theta|\beta, Y)$, which is obtained with only slight notational changes from what was previously described, and where β is sampled from the full (marginal) conditional distribution $p(\mathrm{d}\beta|\theta, Y) \propto p(\mathrm{d}\beta)\prod_{i=1}^n p(y_i|\theta_i, x_i, \beta)$. Sampling the full conditional distribution for θ will often require nonconjugate methods.

The precision parameter α can also be modeled thereby inducing a prior distribution on the number of distinct clusters. Escobar and West (1995) used a data augmentation approach to model α using a gamma prior, $\alpha|a, b \sim \Gamma(a, b)$, and introducing a clever latent variable that makes the Gibbs sampling easy. This same approach can be used in DP and MDP models.

Centering the DP on a parametric family of parametric $\{G_\eta\colon\ \eta \in \Omega\}$ with a prior $p(\mathrm{d}\eta)$ is also possible, e.g., $G \sim \int \mathrm{DP}(\alpha G_\eta)p(\mathrm{d}\eta)$. The full conditional distribution for η is obtained in the same way the marginal conditional was obtained in the MDP model. For the normal linear mixed model with simple random effects, centering the random effects distribution on the $N(0, \sigma^2)$ family with an inverse gamma prior on σ^2 results in an inverse gamma distribution for the full conditional for σ^2 (Bush and MacEachern, 1996). Modeling a random effects distribution with a DP prior centered on the zero-mean multivariate normal distribution with covariance matrix D, where D^{-1} is distributed Wishart, the full conditional of D^{-1} is distributed Wishart (Kleinman and Ibrahim, 1998).

2.3.3. *General inferences*

Inferences for the marginalized DPM model were discussed at the end of Section 2.3.1. The full DPM model is in the form $y_1, y_2, \ldots, y_n | F \overset{\text{i.i.d.}}{\sim} F(\cdot|G) = \int F_\theta(\cdot) G(\mathrm{d}\theta)$, where F_θ has corresponding pdf $p(\cdot|\theta)$. We first indicate how to obtain full inferences for linear functionals and then for arbitrary functionals of F.

In the first instance, run the Gibbs sampler for the marginalized DPM. Once convergence is achieved and the "burn-in" discarded, the Gibbs sampler yields the output $\{\theta^j = (\theta_1^j, \ldots, \theta_n^j)\colon\ j = 1, \ldots, MC\}$. Linear functionals of $F \equiv F(\cdot|G)$ are again given by $T \equiv T(F) = \int T[p(\cdot|\theta_*)] G(\mathrm{d}\theta_*)$. Then for each θ^j, we obtain T^j by first sampling from the updated DP for G, namely sample $G^j \sim G|\theta^j$ using the Sethuraman (1994) construct. Then for each j obtain a sample of B i.i.d. values from G^j, e.g., sample $\theta_*^i \overset{\text{i.i.d.}}{\sim} G^j$. Finally obtain,

$$T^j = \frac{1}{B} \sum_{i=1}^{B} T\bigl[p\bigl(\cdot|\theta_*^i\bigr)\bigr], \quad j = 1, \ldots, MC,$$

which yield (approximate) realizations from the posterior distribution of $T(F)|Y$. The sample $\{T^j\}_{j=1}^{MC}$ is used to obtain point and interval estimates of $T(F)$, as well as its posterior pdf.

Posterior inferences for nonlinear $T(F)$ are obtained as above by simply obtaining $F^j = \int p(\cdot|\theta_*) G^j(\mathrm{d}\theta_*)$, and the corresponding $T(F^j)$, $j = 1, \ldots, MC$. In each instance, G^j is obtained by sampling a truncated version of the Sethuraman (1994) representation for G. Gelfand and Kottas (2002) give details.

2.4. *Polya tree and mixtures of Polya tree models*

Polya tree models were first discussed by Freedman (1963), Fabius (1964) and Ferguson (1974). Use of PT models for complicated data was historically difficult due to mathematical intractability. However, as with DPM models, modern MCMC methods have allowed data analysts to once again consider PT's for modeling data nonparametrically. Lavine (1992, 1994) and Mauldin et al. (1992) have carefully developed and catalogued much of the current theory governing PT's.

The PT is a generalization of the DP. A particular general specification of the PT places probability one on absolutely continuous F's, thus avoiding the discreteness issues associated with the DP. Here, the sample space, Ω, is successively partitioned into finer-and-finer disjoint sets using binary partitioning. At the first level of the tree, a two set partition is constructed with a single pair of corresponding branch probabilities defining the marginal probabilities of these sets. The mth level partition has 2^m

sets and corresponding conditional branch probabilities (probability of being in a set in this partition, given that it is contained in the corresponding parent set in the $(m-1)$st level). Starting from the first level (i.e. the top of the tree), there is a unique path down the branches of the tree to each set at level m, and consequently to any real number in Ω if one continues as $m \to \infty$. The marginal probability of any level m set is simply the product of the corresponding conditional branch probabilities that lead to that set. Randomness is incorporated by specifying independent Dirichlet distributions on each of the pairs of conditional branch probabilities at each level of the tree.

To make this more precise, the first partition of Ω is $\{B_0, B_1\}$. Then further split B_0 into $\{B_{00}, B_{01}\}$, and split B_1 into $\{B_{10}, B_{11}\}$ yielding the 4 disjoint sets at level 2 of the tree. Continue by letting $\varepsilon = \varepsilon_1 \cdots \varepsilon_m$ be an arbitrary binary number, and split B_ε into $\{B_{\varepsilon 0}, B_{\varepsilon 1}\}$ for all ε, and continue *ad infinitum*. The schematic below conveys the splitting for $m = 2$.

B_0		B_1	
B_{00}	B_{01}	B_{10}	B_{11}

Then define the random marginal probabilities $Y_0 = F(B_0)$, $Y_1 = 1 - Y_0 = F(B_1)$, and the successive conditional probabilities $Y_{00} = F(B_{00}|B_0)$, $Y_{01} = 1 - Y_{00} = F(B_{01}|B_0)$, $Y_{10} = F(B_{10}|B_1)$, $Y_{11} = 1 - Y_{10} = F(B_{11}|B_1), \ldots, Y_{\varepsilon 0} = F(B_{\varepsilon 0}|B_\varepsilon)$, $Y_{\varepsilon 1} = 1 - Y_{\varepsilon 0} = F(B_{\varepsilon 1}|B_\varepsilon)$, etc. The marginal probability of a set in the mth partition is calculated as $F(B_{\varepsilon_1 \cdots \varepsilon_m}) = \prod_{j=1}^{m} Y_{\varepsilon_1 \cdots \varepsilon_j}$. The PT specification is completed by specifying $Y_{\varepsilon_1 \cdots \varepsilon_m 0} \overset{\text{ind}}{\sim} \text{Beta}(\alpha_{\varepsilon_1 \cdots \varepsilon_m 0}, \alpha_{\varepsilon_1 \cdots \varepsilon_m 1})$ (i.e. $(Y_{\varepsilon_1 \cdots \varepsilon_m 0}, Y_{\varepsilon_1 \cdots \varepsilon_m 1}) \sim \text{Dirichlet}(\alpha_{\varepsilon_1 \cdots \varepsilon_m 0}, \alpha_{\varepsilon_1 \cdots \varepsilon_m 1})$), for all sets in all partitions. The collection of partitions is denoted as Π and the collection of parameters of all the beta distributions is denoted $\mathcal{A}$. We write $F|\Pi, \mathcal{A} \sim \text{PT}(\Pi, \mathcal{A})$.

It is straightforward to establish conjugacy of the PT model, namely if $y|F \sim F$, $F \sim \text{PT}(\Pi, \mathcal{A})$, then $F|y, \Pi, \mathcal{A} \sim \text{PT}(\Pi, \mathcal{A}^*)$, $\mathcal{A}^* = \{\alpha_\varepsilon + I(y \in B_\varepsilon), \forall \varepsilon\}$.

The PT process can be centered on a particular F_0 by selecting $\Pi = \{F_0^{-1}((i - 1)/2^m), F_0^{-1}(i/2^m)): i = 1, \ldots, 2^m, m = 1, 2, \ldots\}$. Then setting $\alpha_{\varepsilon 0} = \alpha_{\varepsilon 1}$ for all ε, we obtain $E\{F(B_{\varepsilon_1 \cdots \varepsilon_m})\} = 2^{-m} = F_0(B_{\varepsilon_1 \cdots \varepsilon_m})$. Ferguson (1974) showed that for $\gamma > 0$ and $\alpha_{\varepsilon_1 \cdots \varepsilon_{m-1} 0} = \alpha_{\varepsilon_1 \cdots \varepsilon_{m-1} 1} = \gamma m^2$, F is absolutely continuous with probability one. This has become the "standard" parameterization for α_ε. The parameter γ determines how concentrated the prior specification is about the prior guess, F_0. Large γ results in the prior being more concentrated on F_0, e.g., random F's sampled from the PT will concentrate both in terms of similarity in shape and distance from the fixed F_0, while with γ near zero, simulated CDF's often will be considerably dispersed in terms of shape and distance from the fixed F_0. From here on, we choose this standard parametrization and denote the PT distribution as $\text{PT}(\Pi, \gamma)$.

A major criticism of the PT is that, unlike the DP, inferences are somewhat sensitive to the choice of a fixed partition Π. This led Paddock et al. (2003) to consider "jittered" partitions. Hanson and Johnson (2002) instead considered MPT's, wherein inferences are obtained having mixed over a random partition Π_θ thereby alleviating the influence of a fixed partition on inferences.

The MPT is simply defined by allowing the base probability measure to depend on an unknown $\theta \in \Theta$. Thus the base measure becomes a family of probability measures,

$\{F_\theta\colon \theta \in \Theta\}$. This leads to a family of partition families $\{\Pi_\theta\colon \theta \in \Theta\}$. A prior is placed on θ, $p(\mathrm{d}\theta)$. The basic MPT model is represented as

$$y_1, \ldots, y_n | F_\theta \overset{\text{i.i.d.}}{\sim} F_\theta, \qquad F_\theta \sim \mathrm{PT}(\Pi_\theta, \gamma), \quad \theta \sim p(\mathrm{d}\theta)$$

or equivalently $y_i \overset{\text{i.i.d.}}{\sim} F \sim \int \mathrm{PT}(\Pi_\theta, \gamma) p(\mathrm{d}\theta)$.

If we only specify Π or Π_θ to a finite level M, then we have defined a partially specified (or finite) PT or MPT. For a finite PT we write $F|\Pi_M, \gamma \sim \mathrm{PT}(\Pi_M, \gamma)$. Lavine (1994) detailed how such a level M can be chosen by placing bounds on the posterior predictive density at a point. Hanson and Johnson (2002) have recommended the rule of thumb $M \doteq \log_2 n$. On the sets that comprise level M of the tree, one may consider F to follow F_0 (or F_θ) restricted to this set.

Barron et al. (1999) note that the posterior predictive densities of future observations computed from Polya tree priors have noticeable jumps at the boundaries of partition sets and that a choice of centering distribution F_0 “that is particularly unlike the sample distribution of the data will make convergence of the posterior very slow.” The MPT appears to mitigate some of these problems (Hanson and Johnson, 2002). In particular, with a MPT, the predictive density in a regression problem was shown to be differentiable by Hanson and Johnson (2002).

Methods of fitting Polya trees to real data are discussed by Walker and Mallick (1997, 1999), and methods for MPT’s are discussed by Hanson and Johnson (2002). Berger and Gugliemi (1999) considered the problem of model fit by embedding a parametric family in a larger MPT family.

2.5. *The gamma process model*

The survival function for nonnegative data is defined as $S(t) = 1 - F(t)$, $t > 0$. For continuous data, the corresponding hazard function is defined to be $\lambda(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\ell n(S(t))$, and the cumulative hazard is defined to be $\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s$. It follows that $S(t) = \exp(-\Lambda(t))$. Thus in survival modeling for a continuous response, it is possible to place a probability distribution on the space of all probability models for nonnegative continuous data by placing a probability distribution on the family of all possible cumulative hazard functions. Kalbfleisch (1978) proposed using the gamma process (GP) to model the cumulative hazard function $\Lambda(\cdot)$ in the context of the proportional hazards model (Cox, 1972). We follow Ibrahim et al. (2001) and define the GP as follows.

On $[0, \infty)$ let $\Lambda_0(t)$ be an increasing, left-continuous function such that $\Lambda_0(0) = 0$. Let $\Lambda(\cdot)$ be a stochastic process such that (i) $\Lambda(0) = 0$, (ii) $\Lambda(\cdot)$ has independent increments in disjoint intervals, and (iii) $\Lambda(t_2) - \Lambda(t_1) \sim \Gamma(\alpha(\Lambda_0(t_2) - \Lambda_0(t_1)), \alpha)$ for $t_2 > t_1$. Then $\{\Lambda(t)\colon t \geqslant 0\}$ is said to be a GP with parameter (α, Λ_0) and denoted $\Lambda \sim \mathrm{GP}(\alpha, \Lambda_0)$.

Note that $\mathrm{E}(\Lambda(t)) = \Lambda_0(t)$ so that Λ is centered at Λ_0. Also, $\mathrm{Var}(\Lambda(t)) = \Lambda_0(t)/\alpha$ so that, similar to the DP and PT, α controls how “close” Λ is to Λ_0 and provides a measure of how certain we are that Λ is near Λ_0. It is interesting to note that Ferguson (1973, Section 4) recasts the DP as a scaled GP.

The posterior of the GP is characterized by Kalbfleisch (1978); his results for the PH model simplify when no covariates are specified. With probability one, the GP is a monotone nondecreasing step function, implying that the corresponding survivor function is a nonincreasing step function. Similar to the DP, matters are complicated by the presence of ties in the data with positive probability. When present in the observed data, such ties make the resulting computations awkward. Clayton (1991) described a Gibbs sampler for obtaining inferences in the proportional hazards model with a GP baseline.

3. Illustrations

In this section we discuss particular modeling applications and we analyze three data sets using a variety of BNP techniques. We first consider a two sample problem and apply BNP models to analyze these data. Next we discuss the rather large area of semiparametric regression modeling and illustrate with a number of fundamental survival analysis models for data. We analyze a classic data set on time to death from diagnosis with leukemia. We then discuss nonparametric regression function estimation using a variety of basis models for representing the regression function. These methods are illustrated on a data set involving the estimation of mean response of nitric oxide and nitric dioxide in engine exhaust (using ethanol as fuel) as a function of the air to fuel ratio. Methods for the two sample problem were implemented in S-Plus while the survival analysis and the function estimation analyses were done in WinBUGS and Mathematica.

3.1. Two sample problem

A randomized comparative study was conducted to assess the association between amount of calcium intake and reduction of systolic blood pressure (SBP) in black males. Of 21 healthy black men, 10 were randomly assigned to receive a calcium supplement (group 1) over a 12 week period. The other men received a placebo during the 12 week period (group 2). The response variable was amount of decrease in systolic blood pressure. Negative responses correspond to increases in SBP. The data appear in Moore (1995, p. 439). Summary statistics for both groups are given in Table 1.

Let F_1 and F_2 denote the population distributions for decrease in SBP for groups 1 and 2, respectively. The data were fitted to the DP, MDP, DPM, PT, and MPT models. Prior distributions were constructed assuming the range of decrease of SBP for the calcium group is between -20 and 30 and that the data for the placebo group would range between -20 and 20. The midpoints were used for prior estimates of the mean change in SBP, namely 5 and 0 for groups 1 and 2. Prior estimates for the standard deviation were computed as the range/6. Hence, for the calcium group we centered the DP and PT distributions on an $F_{10} = N(5, 70)$, and we centered the placebo group on an $F_{20} = N(0, 44)$.

For the MDP model, we assume $F_1|(\mu_1, \sigma_1^2) \sim \mathrm{DP}(\alpha N(\mu_1|\ 5, \sigma_1^2) IG(\sigma_1^2|\ 2, 70))$ and $F_2|(\mu_2, \sigma_2^2) \sim \mathrm{DP}(\alpha N(\mu_2|\ 0, \sigma_2^2) IG(\sigma_2^2|\ 2, 44))$. Therefore $E(\sigma_1^2) = 70$ and $E(\sigma_2^2) = 44$, and both prior variances are infinite.

Table 1
Blood pressure data: summary statistics for the decrease in systolic blood pressure data for the calcium and placebo groups

	n	Mean	Median	Std. Dev.	Min	Max
Calcium	10	5.0	4	8.7	-5	18
Placebo	11	-0.27	-1	5.9	-11	12

Table 2
Blood pressure data: prior and posterior medians and 95% probability intervals for functionals $T(F)$ for the two-sample problem. The mean and median functionals are denoted by $\mu(\cdot)$ and $\eta(\cdot)$, respectively

	DP		MDP		DPM	
$T(F)$	Prior	Posterior	Prior	Posterior	Prior	Posterior
$\mu(F_1)$	5.08	4.96	4.90	4.97	5.05	5.08
	$(-10.4, 20.3)$	$(0.5, 9.9)$	$(-14.7, 25.9)$	$(0.6, 10.0)$	$(-5.2, 16.5)$	$(0.3, 9.9)$
$\mu(F_2)$	-0.08	-0.31	0.02	-0.25	0.13	-0.30
	$(-9.5, 9.3)$	$(-3.3, 3.0)$	$(-16.2, 15.3)$	$(-3.2, 3.1)$	$(-8.8, 9.6)$	$(-3.3, 2.8)$
$\eta(F_1)$	5.01	5.17	4.93	5.27	5.14	4.89
	$(-10.3, 20.3)$	$(-3.0, 11.0)$	$(-16.4, 27.6)$	$(-3.0, 11.0)$	$(-4.6, 15.9)$	$(0.2, 9.9)$
$\eta(F_2)$	-0.10	-1.1	-0.10	-1.1	0.25	-0.35
	$(-12.4, 11.9)$	$(-3.1, 2.9)$	$(-17.8, 17.1)$	$(-3.1, 2.9)$	$(-8.1, 8.7)$	$(-3.3, 2.6)$
$\mu(F_1) - \mu(F_2)$	5.12	5.23	4.86	5.23	5.08	5.24
	$(-9.8, 20.5)$	$(-0.3, 11.1)$	$(-19.4, 31.5)$	$(-0.3, 10.8)$	$(-8.94, 20.4)$	$(0.0, 10.6)$
$\eta(F_1) - \eta(F_2)$	5.19	4.91	4.86	5.01	5.22	4.99
	$(-14.0, 24.7)$	$(-3.9, 14.1)$	$(-22.3, 34.2)$	$(-3.9, 14.1)$	$(-8.4, 18.9)$	$(-0.3, 10.8)$

The DPM model used was, for $k = 1, 2$, and $i = 1, \ldots, n_k$ with $n_1 = 10$, $n_2 = 11$,

$$\begin{aligned} x_{ki}|\bigl(\mu_{ki}, \sigma^2_{ki}\bigr) &\overset{\text{ind}}{\sim} N\bigl(\mu_{ki}, \tau\sigma^2_{ki}\bigr) \\ \bigl(\mu_{ki}, \sigma^2_{ki}\bigr)|G_k &\overset{\text{ind}}{\sim} G_k \\ G_k|\alpha, G_{k0} &\overset{\text{ind}}{\sim} \mathrm{DP}(\alpha G_{k0}). \end{aligned}$$

Escobar and West (1995) discuss the parameter τ, which for density estimation can be interpreted as a smoothing parameter. For the current problem, we selected $G_{10}(\mathrm{d}\mu_1, \mathrm{d}\sigma^2_1) = N(\mathrm{d}\mu_1|5, \tau\,\mathrm{d}\sigma^2_1)IG(\mathrm{d}\sigma^2_1|2, 70)$, and $G_{20}(\mathrm{d}\mu_2, \mathrm{d}\sigma^2_2) = N(\mathrm{d}\mu_2|5, \tau\,\mathrm{d}\sigma^2_2)IG(\mathrm{d}\sigma^2_2|2, 70)$, where τ was selected to be either 1 or 10 in the current analysis.

For the MPT model, we centered on the family $F_{10}(\mathrm{d}\mu_1, \mathrm{d}\sigma_1) = N(\mathrm{d}\mu_1|5, 5)$ $\Gamma(\mathrm{d}\sigma_1|0.64, 0.08)$ for the calcium group and $F_{20}(\mathrm{d}\mu_2, \mathrm{d}\sigma_2) = N(\mathrm{d}\mu_2|0, 5)\Gamma(\mathrm{d}\sigma_2|0.45, 0.067)$ for the placebo group. In all models with DP components, we set $\alpha = 1$ and we set $\gamma = 0.1$ for models involving Polya trees.

Table 2 contains prior and posterior medians and 95% probability intervals for functionals of F_1 and F_2 using the DP, MDP, and DPM models. The posterior estimates are similar for all 3 models, especially for the DP and MDP models. Estimates of these functionals using PT and MPT models are also readily available. For example, based on the MPT models, the population median change in SBP for the calcium group,

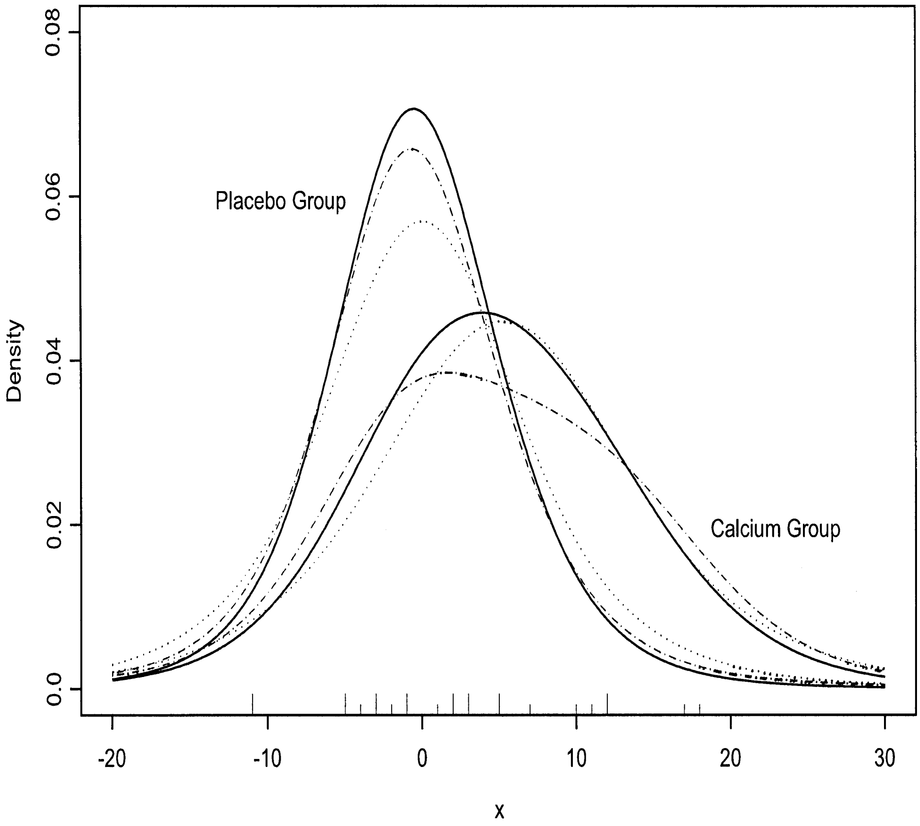

Fig. 1. Blood pressure data: prior (dotted) and posterior density estimates for both groups using the DPM model with $\tau = 1$ (solid) and $\tau = 10$ (dashed). The longer tick marks along the x-axis correspond to the observed data for the placebo group and the shorter tick marks to the observed data for the calcium group.

median(F_1), is estimated to be 3.86 (−2.45, 11.53) and for the placebo group, an estimate of *median*(F_2) is −1.0 (−3.27, 2.83). Inferences for the differences in means and medians are also given in Table 2. It appears that there would be a significant difference if 90% intervals had been considered. Observe that no attempt was made to guarantee that the priors were consistent across models and that this is clearly reflected in the induced priors for the functionals considered in Table 2.

Density estimates from DPM models with $\tau = 1$ and $\tau = 10$ are given in Figure 1. Also, the estimated CDF's for both groups using MDP, DPM, and MPT models are in Figure 2. The estimated CDF's using DP models (not shown) are essentially identical (for these data and for the given choices of α and γ) to those from the MDP models and the estimated CDF's from the PT models (not shown) were similar to those from the MPT models, but differ in that they were not as smooth due to partition effects. Finally note that the prior and posterior density estimates are quite similar. Since our prior was obtained independently of the data (from the second author), this is an indication that

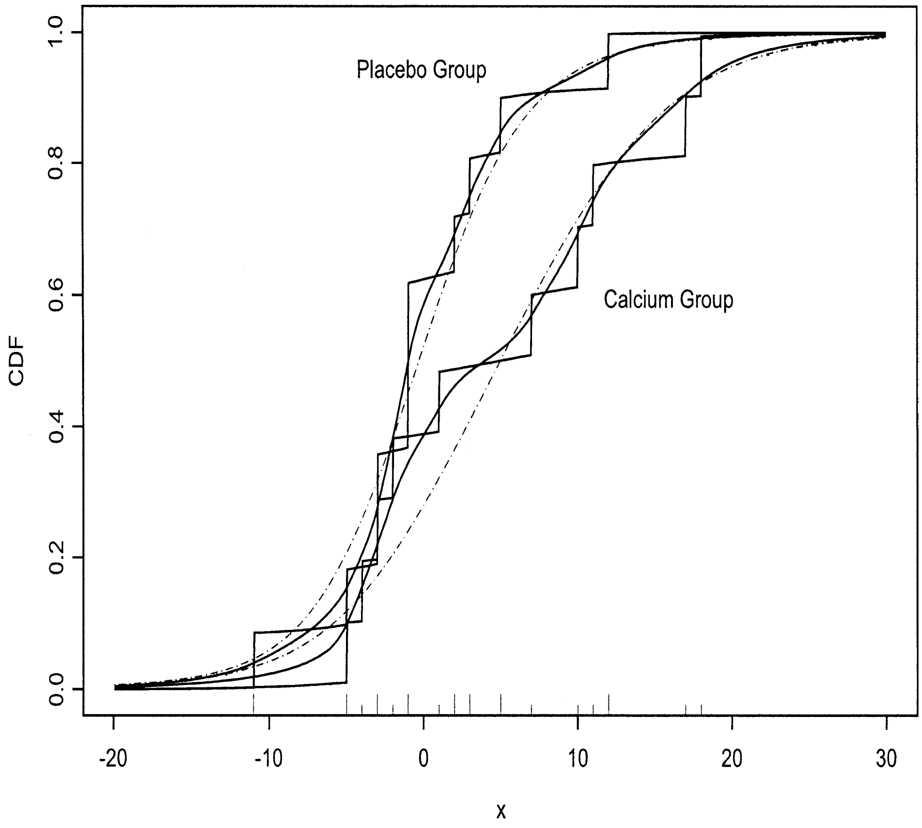

Fig. 2. Blood pressure data: posterior CDF estimates for both groups using the MDP (jagged), DPM (dashed), and MPT (solid) models. The longer tick marks along the x-axis correspond to the observed data for the placebo group and the shorter tick marks to the observed data for the calcium group.

the prior information was quite accurate. One final note is that the sample sizes are so small for this problem that the DPM model density estimates look parametric. If there were bumps in the true densities and with larger sample sizes, the DPM model would reflect this fact. Since the truth is unknown here, we are not in a position to say that any of the models are preferable.

3.2. Regression examples

Here, we mainly discuss two types of regression models. Both types can be expressed in the usual form $y = f(x) + \varepsilon$. In one instance, we consider $f(x) = x\beta$, and with $\varepsilon \sim F$, $F \in \mathcal{F}$ where $\mathcal{F}$ consists of continuous distributions with median zero, which results in $x\beta$ as the median of $y|x, \beta$ or what has been called median regression. In a second instance, we consider $f \in \mathcal{F}^*$ and where the distribution of the error is assumed to have been generated according to a parametric family. When the primary goal is estimation of the regression function, f, parametric error models may suffice, but when considering predictive inference or in estimation of certain survival models, it is desirable to estimate F nonparametrically. We discuss these two types of models in some detail and give illustrations. We also discuss the situation where both f and F are allowed to be flexible.

We want to emphasize at the outset that our purpose here is mainly to illustrate the fundamental ideas and methods. The published literature clearly goes well beyond what we present here and we make no claims to having used the best or most sophisticated method for any given problem. We emphasize the simplicity of the methods that are presented here as many of them are accomplished in WinBUGS.

Our main illustration of semiparametric regression with unknown error distribution is in the area of survival analysis, which is discussed next.

3.2.1. Regression for survival data

In this subsection, we first briefly discuss univariate survival data with censoring. We proceed to discuss semiparametric accelerated failure time (AFT) and proportional hazards (PH) models for censored survival data with covariates. We ultimately analyze a classic data set on leukemia remission times using BNP methodology applied to AFT and PH models. See Ibrahim et al. (2001) for descriptions of these models and for other analyses of these data. All of the modeling done here applies to uncensored data and thus to standard linear regression.

Denote survival times for n independently sampled individuals as $T_1, \ldots, T_n$. Right censored data are denoted $\{(t_i, \delta_i)\colon i = 1, \ldots, n\}$ where $\delta_i = 0$ implies that $T_i > t_i$, which corresponds to t_i being an observed censoring time, and $\delta_i = 1$ implies $T_i = t_i$. Censoring times are assumed independent of event times. With covariate information, we have data $\{(t_i, \delta_i, x_i)\colon i = 1, \ldots, n\}$.

Let T_0 be a random survival time from a baseline distribution. The AFT model specifies that an individual with covariate vector x has the survival time $T_x = g(x'\beta)T_0$, for regression coefficients β and a monotone function g. This is equivalent to $S(t|x) = S(t/g(x'\beta))$ where $S(t) = P(T_0 > t)$ is the baseline survival function and $S(t|x) = P(T_x > t)$.

Usually, g is taken to be the exponential function and the model is then equivalent to $\log(T_x) = x'\beta + \log(T_0)$, i.e. a standard linear regression model. Standard parametric analyses further assume that $\log(T_0) = \sigma\varepsilon$ where ε is standard normal, extreme value, or logistic. If ε has median zero, a median-zero regression model is obtained.

Christensen and Johnson (1988) obtain approximate, marginal inference in the AFT model with a DP baseline S while Johnson and Christensen (1989) show that obtaining full posterior inference from an AFT model with a DP baseline is infeasible. Kuo and Mallick (1997) circumvent this difficulty by considering a DPM for S. They interpret the baseline model as a “smoothed” DP. Walker and Mallick (1999), and Hanson and Johnson (2002) considered, respectively, PT and MPT baselines in the AFT model, whereas Kottas and Gelfand (2001a) described a DPM model for the baseline in the AFT model; these models are all median regression models. Hanson and Johnson (2004) extended the MDP model of Doss (1994) to an AFT model with a MDP baseline for interval censored data.

On the other hand, the PH model has by far enjoyed the greatest success of any other statistical model for survival data with covariates. Frequentist and Bayesian statistical literature on the topic far exceed that for any other survival model. The PH model is specified using the baseline hazard function $\lambda(t)$. For baseline survival T_0, the hazard is defined as

$$\lambda(t) = \lim_{\mathrm{d}t \to 0^+} \frac{P(t \leqslant T_0 < t + \mathrm{d}t)}{\mathrm{d}t},$$

or $\lambda(t)\,\mathrm{d}t \approx P(t \leqslant T_0 < t + \mathrm{d}t)$ for small $\mathrm{d}t$. If T_0 is absolutely continuous then $\lambda(t) = f(t)/S(t)$ where f and S the pdf and survivor function for T_0, respectively. Cox's PH model (Cox, 1972) assumes that for an individual with covariate x, $\lambda(t|x) = g(x'\beta)\lambda(t)$, where g and β are as before (except for no intercept here). Typically g is taken to be the exponential function yielding the interpretation of $\exp(x\beta)$ as a relative risk of "instantaneous failure" comparing an individual with covariates x to a baseline individual. Under the PH model $S(t|x) = \exp(-\mathrm{e}^{x\beta}\Lambda(t))$. The latter expression can be used to define the PH model when Λ has jump discontinuities, e.g., when T_0 is a mixture of continuous and discrete distributions.

The success of the PH model across a wide spectrum of disciplines is in part due to the interpretability of the regression parameters and in part due to the availability of easy to use software to fit the frequentist version of the model. In statistical packages the model is fit via *partial likelihood*, involving only β, which is not a proper likelihood but which does yield estimators with desirable properties such as asymptotic normality. The infinite-dimensional parameter Λ is treated as a nuisance parameter and, if needed, is estimated following the estimation of β. Bayesian approaches to the Cox model have considered both the use of the partial likelihood in inference and the consideration of a full probability model for (β, λ). We discuss only the latter and view the full, joint modeling of parameters, as well as nonasymptotic inference as a particular benefit of the Bayesian approach. Other BNP approaches have been discussed by Sinha and Dey (1997), Laud et al. (1998) and Ibrahim et al. (2001). It should be pointed out that, despite the flexibility of the PH model due to the baseline hazard being unspecified, the PH assumption is still quite restrictive and easily fails for many data sets. The semiparametric AFT model serves as a potential alternative when this is the case.

Given all of this background, we consider here a simple application of BNP methodology to a two sample survival analysis problem. Clearly there are many possible approaches but we only consider two here for the purpose of illustration.

Data on the remission times from two groups of leukemia patients are considered by Gehan (1965) and Kalbfleisch (1978) and are reproduced in Table 3. A PT AFT model was fitted to these data with a Weibull(1.47, 19.61) base measure, estimated from a parametric fit. We set $\gamma = 0.1$. The posterior median and equal-tailed 95% PI for β is 1.62 (0.70, 1.97). The Group 2 population has a median survival time estimated to be about $\mathrm{e}^{1.62} \approx 5$ times that of Group 1. In Figure 3, estimated survival curves are plotted for the two groups.

We now turn to the PH model. Although many stochastic processes have been used as priors for Λ in the Cox model, we focus attention on the first to be used in this context, the independent increments GP, which was discussed in Section 2.5. We now give a detailed discussion of the implementation of this model for use in WinBUGS, before discussing the BNP PH analysis of the leukemia data.

Table 3
Leukemia data: weeks of remission for leukemia patients

Group 1:	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
Group 2:	6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*,32*, 32*, 34*, 35*

*Right censored observation.

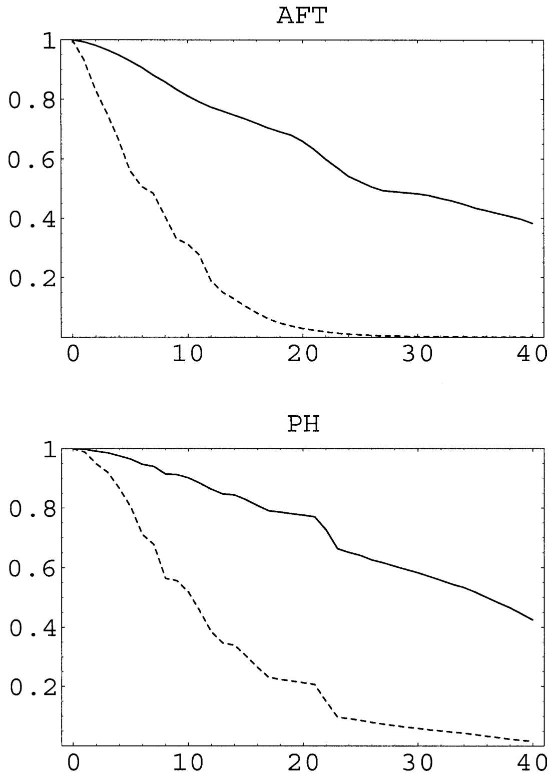

Fig. 3. Leukemia data: estimated survival curves for AFT and PH models; Group 2 (solid) and Group 1 (dashed).

Burridge (1981) and Ibrahim et al. (2001) suggest that the model as proposed by Kalbfleisch (1978) and extended by Clayton (1991) is best suited to grouped survival data. Walker and Mallick (1997) considered an approximation to the GP for continuous data that we describe here, in part because it is readily implemented in WinBUGS. Define a partition of $(0, \infty)$ by $\{(a_{j-1}, a_j]\}_{j=1}^{J} \cup (a_J, \infty)$ where $0 = a_0 < a_1 < a_2 < \cdots < a_{J+1} = \infty$. Here, a_J is taken to be equal to be $\max(\{t_i\}_{i=1}^{n})$. If $\Lambda \sim \mathrm{GP}(\alpha, \Lambda_0)$ then by definition $\Lambda(a_j) - \Lambda(a_{j-1}) \overset{\text{ind}}{\sim} \Gamma(\alpha(\Lambda_0(a_j) - \Lambda_0(a_{j-1})), \alpha)$. Walker and Mallick (1997) make this assumption *for the given partition* and further assume that $\lambda(t)$ is constant and equal to λ_j on each $(a_{j-1}, a_j]$ for $j = 1, \ldots, J$. This implies $\lambda_j \sim$

$\Gamma(\alpha\lambda_{0j}, \alpha)$ where $\lambda_{0j} = (\Lambda_0(a_j) - \Lambda_0(a_{j-1}))/(a_j - a_{j-1})$, and yields a particular piecewise exponential model.

Now given $\Lambda(\cdot)$, or equivalently $\{\lambda_j\}_{j=1}^J$, $S(t) = \exp(-\Lambda(t))$, $S(t|x) = \exp(-\mathrm{e}^{x'\beta}\Lambda(t))$ and $f(t|x) = \mathrm{e}^{x'\beta}\lambda(t)\exp(-\mathrm{e}^{x'\beta}\Lambda(t))$. Assume that the event times $\{t_i\}_{i=1}^n$ are included as some of the partition points $\{a_j\}_{j=1}^J$. Let $j(i)$ be such that $t_i = a_{j(i)}$. Then $\Lambda(t_i) = \sum_{j=1}^{j(i)} \lambda_j \Delta_j$ where $\Delta_j = a_j - a_{j-1}$ and $\lambda(t_i) = \lambda_{j(i)}$. The likelihood is given by

$$\begin{aligned}\mathcal{L}(\lambda, \beta) &= \prod_{i=1}^{n} \exp\bigl(-\mathrm{e}^{x_i\beta}\Lambda(t_i)\bigr)\bigl[\mathrm{e}^{x_i\beta}\lambda(t_i)\bigr]^{\delta_i} \\ &= \prod_{i}\prod_{j=1}^{j(i)} \exp\bigl(-\mathrm{e}^{x_i\beta}\lambda_j\Delta_j\bigr) \prod_{\{i:\delta_i=1\}} \mathrm{e}^{x_i\beta}\lambda_{j(i)}\end{aligned}$$

which is proportional to a product of Poisson kernels. Therefore, with independent gamma priors on $\{\lambda_j\}$, this model is readily fitted in WinBUGS. This likelihood is similar to that obtained by Clayton (1991) using a counting process argument (for example, see “Leuk: survival analysis using Cox regression” in Examples Volume I, WinBUGS 1.4). Clayton’s approach requires sampling $\Lambda(\cdot)$ only at the $\{t_i\}$ to obtain full inference for β. The piecewise exponential model has been used to accommodate approximations to a correlated prior process (Ibrahim et al., 2001, Section 3.6) and also used in joint models accommodating a latent longitudinal marker that affects survival (Wang and Taylor, 2001; Brown and Ibrahim, 2003) due to the simple structure of the model.

To get more of the flavor of the GP from this approximation, one might take the partition to be a fine mesh. Furthermore, a mixture of gamma processes can be induced by assuming $\Lambda \sim \mathrm{GP}(\alpha, \Lambda_\theta)$, $\theta \sim f(\theta)$. For example, one might center $\Lambda(\cdot)$ at $\Lambda_\theta = \theta t$, the exponential family, and place a hyperprior on θ. This results in a mixture of GP’s (MGP).

We adapted this approach and fit the MGP PH model to the leukemia data using vague hyperpriors in WinBUGS. The posterior median and 95% PI for β is 1.56 (0.84, 2.36). The hazard of expiring in Group 1 is about $\mathrm{e}^{1.56} \approx 4.8$ times as likely as Group 2 at any time t. Estimated survival curves are given in Figure 3.

Other prior processes used in PH survival models include the beta process (Hjort, 1990), and the extended gamma process (Dykstra and Laud, 1981), which smooths the GP with a known kernel. Ishwaran and James (2004) extend this work and the work of others (notably Lo and Weng, 1989, and Ibrahim et al., 1999) to a very general setting by capitalizing on a connection between the GP and the DP. Often Bayesian semiparametric survival models are fit by partitioning $[0, \infty)$ into a fine mesh and computing grouped data likelihoods; the approach of Ishwaran and James (2004) avoids this computationally intensive approach. Kim and Lee (2003) consider the PH model with left truncated and right censored data for very general neutral to the right priors.

Ibrahim et al. (2001) also discuss the implementation of frailty, cure rate, and joint survival and longitudinal marker models. Mallick and Walker (2003) develop a frailty model that uses PTs and includes proportional odds, AFT, and PH all as special cases. The model utilizes a PT error term and a monotone transformation function modeled with a mixture of incomplete beta functions. Prior elicitation for survival models are discussed by Ibrahim et al. (2001). Methods on prior elicitation for regression coefficients in parametric survival models developed by Bedrick et al. (2000) apply to Bayesian semiparametric AFT modeling. Ishwaran and James (2004) develop weighted GP's in the multiplicative intensity model. Huzurbazar (2004) provides an extensive introduction to the use of Bayesian flowgraph models for the modeling of survival data. Mallick et al. (1999) use multivariate adaptive regression splines in a highly flexible model allowing for time-dependent covariates. Space does not permit us to discuss the extensive literature on semiparametric cure rate models, competing risks models, multivariate models, and other important areas.

In the absence of covariates, Susarla and van Ryzin (1976) assumed a DP prior for F for right-censored data and established that the Kaplan and Meier (1958) estimator is obtained as $\alpha \to 0^+$. Johnson and Christensen (1986) extended the model to grouped survival data and similarly showed that Turnbull's (1974) estimator is the corresponding limiting form. Doss (1994) and Doss and Huffer (2004) discussed fitting the MDP model to censored data and compared various algorithms based on importance sampling and MCMC to obtain inferences. They also provided user-friendly software for the statistical packages R and S-Plus to fit these models. Other related approaches include Lavine (1992), who gave an example of density estimation for survival data via PT's. Wiper et al. (2001) used a mixture of Gamma densities in the spirit of Richardson and Green (1997) to model data with support on $[0, \infty)$. The DPM model of Escobar and West (1995) can also be used for survival data or log survival data.

3.2.2. Nonparametric regression with known error distribution

Estimation of an unknown regression function is a common and extensively researched area across many disciplines. The problem is typically to estimate the mean function f from data $\{(x_i, y_i)\}_{i=1}^n$ in the model

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d.}, \ E(\varepsilon_i) = 0,$$

but in some applications the shape of the error distribution ε_i is of interest as well. We initially assume x_i is univariate but later discuss the case when x_i is a vector of predictors.

Denison et al. (2002) provide an introduction to Bayesian semiparametric regression methods focusing primarily on splines. Müller and Quintana (2004) review advances in Bayesian regression and additionally discuss neural networks. Müller and Vidakovic (1999) discuss Bayesian models incorporating wavelets.

One successful approach borrows from the field of *harmonic analysis* and assumes f can be represented as a weighted sum of basis functions. For f sufficiently smooth, and given an orthonormal basis $\{\phi_j\}_{j=1}^\infty$ of the function space of square-integrable functions on some region R, $\mathcal{L}^2(R)$, one can write the Fourier representation of f as $f(x) = \sum_{j=0}^\infty \beta_j \phi_j(x)$ where $\beta_j = \int_R f(x)\phi_j(x)\,\mathrm{d}x$. The basis is said to be orthonormal if $\int_R \phi_i(x)\phi_j(x)\,\mathrm{d}x = \delta_{ij}$ where $\delta_{ij} = 1$ if $i = j$ and zero otherwise. Orthonormal bases make certain common calculations trivial in some problems, but are not required of this approach. Popular choices for $\{\phi_j\}$ are the Fourier series (sines and cosines), spline bases, polynomials, and wavelet bases.

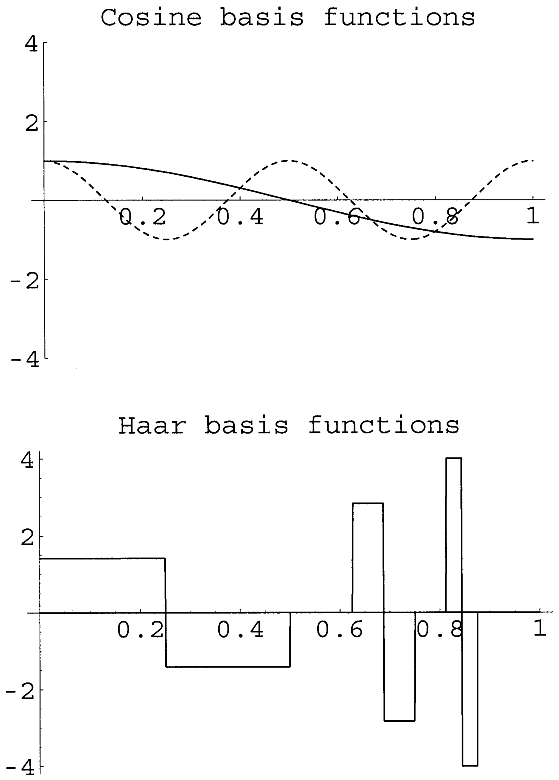

Fig. 4. Cosine basis functions $\cos(x\pi)$ (solid) and $\cos(x4\pi)$ (dashed); Haar basis functions $\phi_{2,1}$, $\phi_{4,6}$ and $\phi_{5,14}$ from left to right.

It is impossible to estimate $\{\beta_j\}_{j=0}^{\infty}$ with finite data. All but a finite number of these coefficients must be set to zero for estimation to proceed and therefore f is approximated by a finite number of basis functions $f(x) = \sum_{j=0}^{J} \beta_j \phi_j(x)$ in practice. The basis functions are often ordered in some fashion from broad functions that indicate a rough trend to functions that are highly oscillatory over R. A statistical problem is to determine at which point noise is essentially being modeled by the more oscillatory functions, or equivalently at which point J to "cutoff" the basis functions. In Figure 4 we see two of the cosine basis functions $\{\cos(xj\pi)\}_{j=0}^{\infty}$ and three Haar basis functions (described later) on $R = [0, 1]$.

Traditionally, the choice of J is an interesting problem with many reasonable, typically *ad hoc*, solutions. This choice deals intimately with the issue of separating signal from the noise. It is well-known that an $(n-1)$-degree polynomial fits data $\{(x_i, y_i)\}_{i=1}^{n}$ perfectly, an example of overfitting, or the inclusion of too many basis functions. Efromovich (1999) overviews common bases used in regression function estimation and addresses choosing J in small and large samples.

If one fixes J and assumes i.i.d. Gaussian errors then the standard linear model is obtained:

$$y_i = \beta_0 + \beta_1\phi_1(x_i) + \cdots + \beta_J\phi_J(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} N\bigl(0, \sigma^2\bigr).$$

Placing a prior on β and σ^{-2} yields the Bayesian linear model (Lindley and Smith, 1972), which is easily implemented in WinBUGS. In Figure 5 we examine orthonormal series fits to data on the amount of nitric oxide and nitric dioxide in the exhaust of a single-cylinder test engine using ethanol as fuel (Brinkman, 1981). The response is in μg per joules and the predictor is a measure of the air to fuel ratio. These data are part of a larger data set used throughout the S-Plus Guide to Statistics (MathSoft, 1999) to illustrate various smoothing techniques, including locally weighted regression smoothing, kernel smoothers, and smoothing splines. The cosine, $\phi_i(x) = \cos(i\pi(x - 0.5)/0.8)$, and Legendre polynomial bases were used for illustration with $R = [0.5, 1.3]$ and fixed $J = 5$. Independent $N(0, 1000)$ priors were placed on the regression coefficients and the precision σ^{-2} was assumed to be distributed $\Gamma(0.001, 0.001)$ as an approximation to Jeffreys' prior. The prior of Bedrick et al. (1996) can be used to develop an informative prior on β. The choice of basis functions, cutoff J, and region R will all affect posterior inference.

Multivariate predictors $x_i = (x_{i1}, \dots, x_{ip})$ can be accommodated via series expansions by considering products of univariate basis functions. For example, in the plane, simple products are formed as $\phi_{jk}(x_1, x_2) = \phi_j(x_1)\phi_k(x_2)$. The regression model is then $y_i = \sum_{j=1}^{J}\sum_{k=1}^{J}\beta_{jk}\phi_{jk}(x_{i1}, x_{i2}) + \varepsilon_i$. Additive models are an alternative where the mean response is the sum of curves in each predictor, e.g., $E(y_i) = \sum_{j=1}^{J_1}\beta_{j1}\phi_j(x_{i1}) + \sum_{j=1}^{J_2}\beta_{j2}\phi_j(x_{i2})$.

A popular Bayesian alternative to fixing the number of components is to place a prior on J and implement the reversible jump algorithm of Green (1995). Reversible jump MCMC approximates posterior inference over a model space where each model has a parameter vector of possibly different dimension. A prior probability is placed on each of $J = 1, 2, \dots, J_0$, where J_0 is some natural upper bound chosen such that consideration of $J > J_0$ would be superfluous. Reversible jump for the regression problem in the context of a spline basis is discussed in Denison et al. (2002) and used, for example, by Mallick et al. (1999) and Holmes and Mallick (2001). Many spline bases are built from truncated polynomials. For example $\{(x - a_j)_+^3\}_{j=1}^J$ is a subset of a cubic spline basis, where $\{a_j\}_{j=1}^J$ are termed *knots* and $(x)_+$ is equal to x when $x > 0$ and equal to zero otherwise.

Another approach is to fix J quite large and allow some of the $\{\beta_j\}_{j=1}^J$ to be zero with positive probability. This approach, advocated by Smith and Kohn (1996), can be formulated as $\beta_j \sim \gamma_j\beta_j^*$ where $\gamma_j \sim \text{Bernoulli}(\theta_j)$ independent of $\beta_j^* \sim N(b_j, \eta_j^2)$, and for moderate J and independent β_j priors can be programmed in WinBUGS. For the ethanol data using the cosine basis, we consider the rather naive, data-driven prior $\gamma_j \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(0.5)$, $\beta_j^*|\sigma^2 \sim N(b_j, 10\sigma^2 v_j)$. Where X is the design matrix from the model with all basis functions up to J, i.e., $\gamma_1 = \cdots = \gamma_J = 1$, v_j are the diagonal elements of $(X'X)^{-1}$ and $(b_1, \dots, b_J)$ are the least squares estimates taken from $(X'X)^{-1}X'y$. Figure 6 shows the resulting estimate of the regression function. Five of

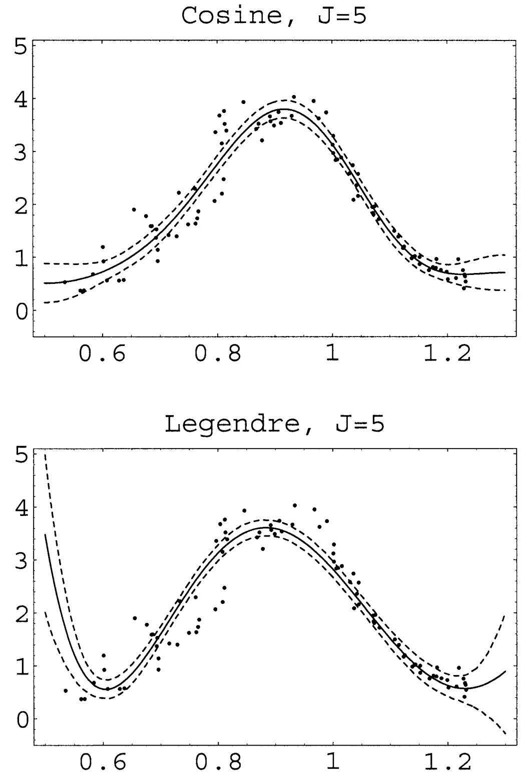

Fig. 5. Ethanol data: mean function estimate using cosine and Legendre bases, $J = 5$.

the ten basis functions have posterior probability $P(\gamma_j = 0|y)$ less than the prior value of 0.5. Clyde and George (2004) discuss priors of this type, specifically the g-prior, in more detail.

Crainiceanu et al. (2004) outline a strategy for fitting penalized spline models in WinBUGS. They capitalize on an equivalence between fitting penalized spline and mixed effect models and the resulting WinBUGS implementation is straightforward. They illustrate the possibilities by fitting nonparametric regression, binomial regression, and nonparametric longitudinal ANOVA models, all in WinBUGS. An advantage of the Bayesian approach over the frequentist approach is that it obviates the use of "plug-in" estimates when computing interval estimates. We apply the approach of Crainiceanu et al. (2004) by fitting a penalized quadratic spline model to the ethanol data. Specifically, the model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \sum_{k=1}^{10} b_k (x_i - \kappa_k)_+^2 + \varepsilon_i,$$

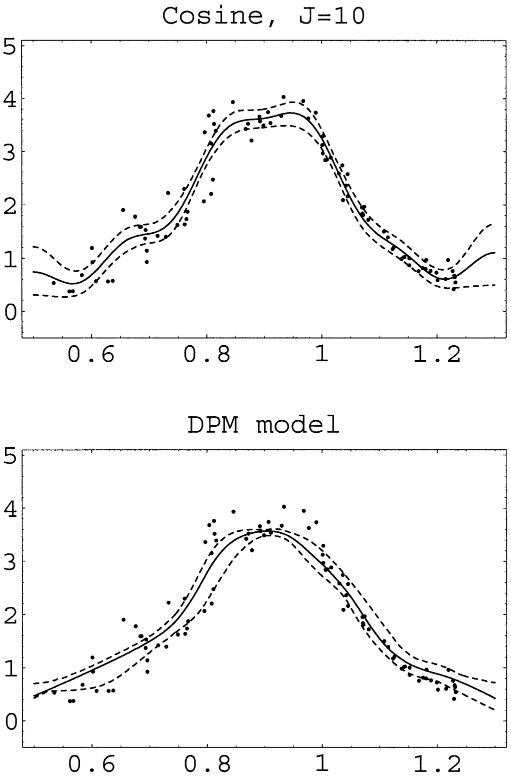

Fig. 6. Ethanol data: estimates of regression mean functions using a cosine basis, and a DPM model (see Section 3.2.3).

where $b_k|\sigma_b \overset{\text{i.i.d.}}{\sim} N(0, \sigma_b^2)$ independent of $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_\varepsilon^2)$. Here, the knots $\{\kappa_k\}_{k=1}^{10}$ are defined as $\kappa_i = 0.4 + 0.1i$, evenly spaced over the range of the predictor. In Figure 7 we see the penalized spline estimate along with 95% pointwise probability intervals.

A unique class of orthonormal bases are wavelet bases. Wavelets are useful for modeling functions whose behavior changes dramatically at different locations and scales, often termed “spatially inhomogeneous.” Think of a grayscale photograph of the Rocky mountains. Much of the photograph will be flat, rocky homogeneous areas where the grayscale changes little. At the edges of a mountain leading to sky, however, the scale changes abruptly. Also, foliage around the base of the mountain will have highly varying grayscale in a small area relative to the mountainous part. Wavelets can capture these sorts of phenomena and for this reason are extensively used in image processing.

The simplest wavelet basis is the Haar basis (Haar, 1910). The Haar basis is also the only wavelet basis with basis functions that have a closed form. On the interval $R = [0, 1]$ the Haar basis (as well as other wavelet bases) is managed conveniently by

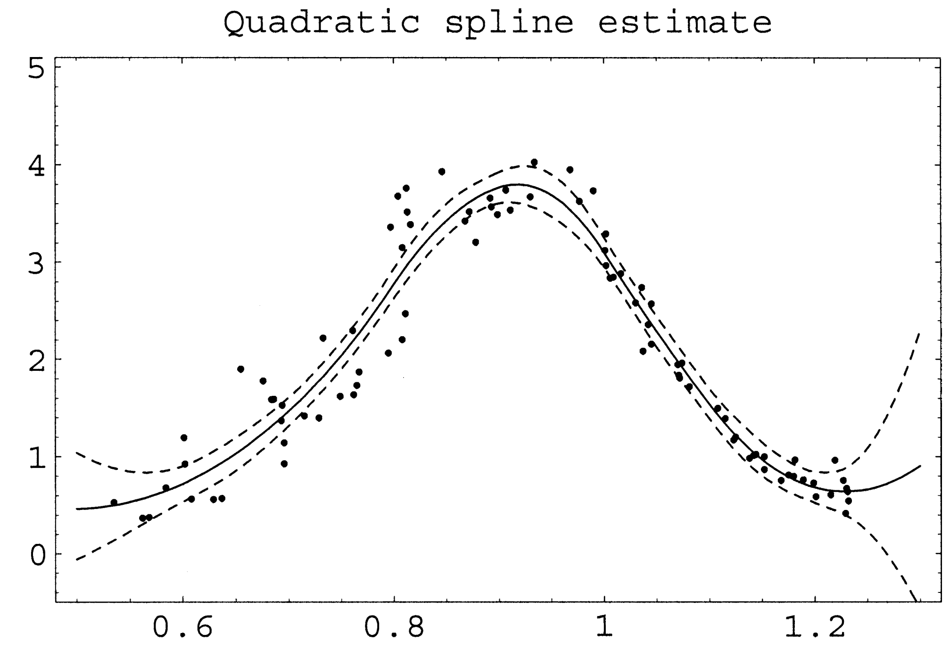

Fig. 7. Ethanol data: estimate of regression mean function using a penalized spline.

a double index and is derived from the Haar mother wavelet

$$\phi(x) = \left\{ \begin{array}{ll} 1, & 0 \leqslant x < 0.5, \\ -1, & 0.5 \leqslant x \leqslant 1, \\ 0, & \text{otherwise} \end{array} \right\}$$

through the relation $\phi_{ij}(x) = \phi(2^{(i-1)}x - j + 1)2^{(i-1)/2}$ for $i = 1, \ldots, \infty$, and $j = j(i) = 1, \ldots, 2^{(i-1)}$. The set $\{I_{[0,1]}(x)\} \cup \{\phi_{ij}\}$ forms an orthonormal basis of $[0, 1]$. Figure 4 shows three of the Haar basis functions; the i indexes the scale of the basis function whereas the j indexes location. For large i, wavelet basis functions can model very localized behavior. Contrast the Haar basis to the cosine basis where basis functions oscillate over the entire region R. For this reason wavelets can model highly inhomogeneous functions but also require special tools to ensure that mean estimates do not follow the data too closely. These tools, broadly termed “thresholding,” require that there is substantial data-driven evidence that a wavelet basis function belongs in the model, and more evidence is required for larger i. Bayesian thresholding typically places mixture priors on basis coefficients in the wavelet domain after transforming data using the discrete wavelet transform. These priors place positive probability that some coefficients are very small (or zero). Müller and Vidakovic (1999) discuss Bayesian wavelet modeling in detail. A nice, short introduction to Bayesian wavelets and thresholding is Vidakovic (1998).

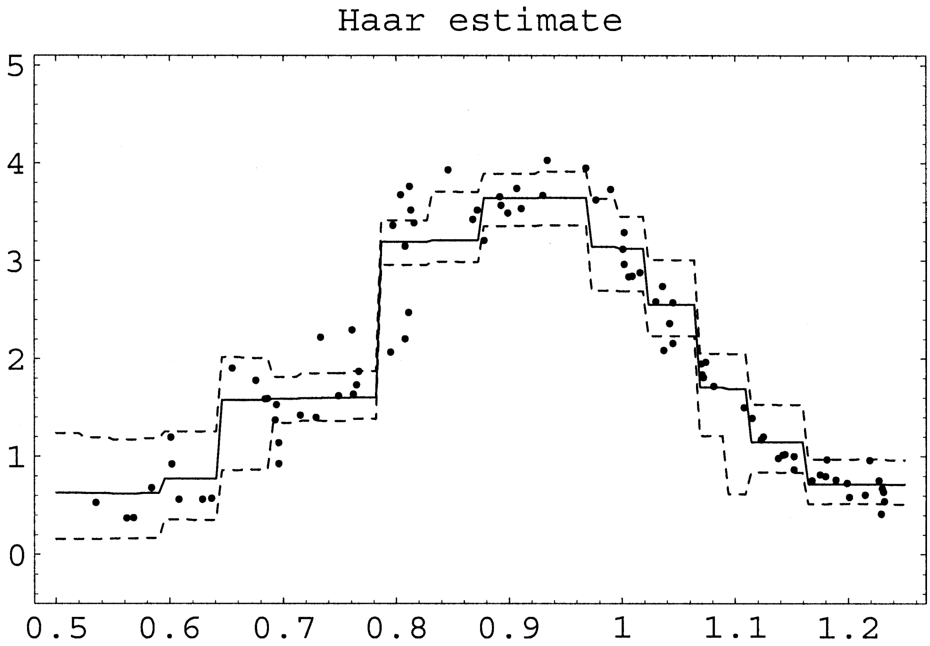

Fig. 8. Ethanol data: estimates of regression mean functions using Harr wavelets.

For illustrative purposes, we fit the following Haar wavelet model to the ethanol data in WinBUGS:

$$y_k = \beta_0 + \sum_{i=1}^{4} \sum_{j=1}^{2^{i-1}} \gamma_{ij} \beta^*_{ij} \phi_{ij}(x_k) + \varepsilon_k, \quad \varepsilon_k \overset{\text{i.i.d.}}{\sim} N\bigl(0, \sigma^2\bigr).$$

A simple data-driven prior was constructed in the same manner as for the cosine basis except $\gamma_{ij} \sim \text{Bernoulli}(2^{-i})$, ensuring that the prior probability of including a basis function decreases with how "localized" the function is. Figure 8 shows the resultant mean function estimate. Four of the 15 basis functions considered had posterior probabilities of being included in the model less than 0.1.

3.2.3. Nonparametric regression with unknown error distribution

Combinations of the approaches discussed so far yield very rich, highly flexible models for both the regression function f and the error ε. Alternatively, a highly flexible model that has ties to kernel regression and local linear regression is the use of DPMs for multivariate data.

To obtain inference in the general model $y(x) = f(x) + e(x)$, Müller et al. (1996) suggest modeling data $\{z_i = (x_i, y_i)\}_{i=1}^n$ as arising from a DPM of multivariate Gaussian densities. As is typical, inference is obtained with the DP integrated out and the model reduces to a particular finite mixture model. The model is given hierarchically by

$$z_i|\mu_i, \Sigma_i \overset{\text{ind}}{\sim} N(\mu_i, \Sigma_i), \quad (\mu_i, \Sigma_i)|G \overset{\text{i.i.d.}}{\sim} G, \quad G \sim \text{DP}(\alpha G_0).$$

The authors consider the prior $g_0(\mu, \Sigma^{-1}) = N_p(\mu|m, B)W_p(\Sigma^{-1}; \nu, (S\nu)^{-1})$ where p is the dimension of $z_i = (x_i, y_i)$, g_0 is the density of G_0, $N_p(x|\mu, \Sigma)$ is the pdf of a multivariate normal variate with mean μ and covariance Σ, and $W_q(\nu, \Sigma)$ is the pdf of a Wishart variate with degrees of freedom ν and mean $\nu\Sigma$. Hyperpriors can be further placed on m, B, S, and α.

An estimate of $f(x_0)$ is provided by $E(y_{n+1}|x_{n+1} = x_0, x, y)$ and is obtained using conditioning arguments. This estimate is essentially a locally-weighted piecewise linear estimate averaged over the MCMC iterates. We consider a simple version of this model for the ethanol data by taking m to be the sample mean $\bar{z}$, B as 10 times the sample covariance of $\{z_i\}_{i=1}^n$, $\alpha = 2$, $S = \text{diag}(0.05^2, 0.25^2)$, and $\nu = 2$.

The prior expected number of components is about 8. Let k denote the number of distinct components in the model. A posteriori, we find $P(k \leqslant 3|z) \approx 0$, $P(4 \leqslant k \leqslant 6|z) \approx 0.94$, and $P(k \geqslant 7|z) \approx 0.04$. The estimated regression function and pointwise 95% probability intervals are in Figure 6, assuming that the marginal finite mixture model (induced by the DPM) is the full probability model. Although the example illustrates regression with one predictor, an attractive feature of the DPM model is that it is readily extended to many predictors, as long as modeling assumptions are reasonable.

4. Concluding remarks

The field of Bayesian nonparametrics relies on an interesting combination of the (sometimes abstract) development of probability models on large spaces and modern Markov chain Monte Carlo technology. The former is necessary for the application of Bayes theorem and the latter for its implementation. Analysis of complex and interesting data using BNP methodology was made to wait for the recent development of MCMC methods. Our paper has attempted to give a flavor of what is now possible due to the merger of these areas. We remind the reader that our goal was to present fundamental ideas and to illustrate them with relatively simple methods, rather than the most sophisticated ones.

There is a long list of methods and models that have been left out, too long to mention all. We simply mention a few. First, we have not discussed nonparametric dependent data modeling. MacEachern (2000) invented the dependent Dirichlet Process (DDP), which builds in dependence among a collection of random probability measures. The DDP has recently been used by De Iorio et al. (2004), who used ANOVA structure in modeling dependence, and by Gelfand et al. (2004) for modeling spatial data. Longitudinal modeling using the DDP should be straightforward given their development for spatial data. Dependent nonparametric processes were also considered by Gelfand and Kottas (2001) and Kottas and Gelfand (2001b), and Hoff (2003) in the context of modeling stochastic order. Another area that is ripe for development is the application of BNP methods to bioinformatics and proteomics, see for example Do et al. (2004). Areas that, to our knowledge, still require attention are (i) the development of mixtures of Polya tree priors for multivariate data and (ii) methods for model selection and model fit, for example how can one formally choose between semiparametric PH and AFT models and also assess their goodness of fit.

Throughout this article, very little has been said about theory since our goal was to present basic modeling techniques and to give a flavor for their application to data. There are of course many articles that develop theoretical aspects of BNP models. See for example Diaconis and Freedman (1986) for a BNP model and method based on DP's that fails. However, there is much theoretical work that establishes that BNP methods are valid. For example, Ghosal et al. (1999) established consistency of density estimates based on an MPT. Regazzini et al. (2002) recently presented results for exact distributions of functionals of a DP. Choudhuri et al. (2004) discuss asymptotic properties of BNP function estimates and give many references. Also see the monograph of Ghosh and Ramamoorthi (2003) for additional theoretical background material and references.

References

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–1174.

Barron, A., Schervish, M., Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27**, 536–561.

Bedrick, E.J., Christensen, R., Johnson, W.O. (1996). A new perspective on priors for generalized linear models. *J. Amer. Statist. Assoc.* **91**, 1450–1460.

Bedrick, E.J., Christensen, R., Johnson, W.O. (2000). Bayesian accelerated failure time analysis with application to veterinary epidemiology. *Statist. in Med.* **19**, 221–237.

Berger, J.O., Gugliemi, A. (1999). Bayesian testing of a parametric model versus nonparametric alternatives. *J. Amer. Statist. Assoc.* **96**, 174–184.

Berry, D., Christensen, R. (1979). Empirical Bayes estimation of a binomial parameter via mixture of Dirichlet processes. *Ann. Statist.* **7**, 558–568.

Blackwell, D. (1973). Discreteness of Ferguson selections. *Ann. Statist.* **1**, 356–358.

Blackwell, D., MacQueen, J.B. (1973). Ferguson distributions via Polya urn schemes. *Ann. Statist.* **2**, 353–355.

Brinkman, N.D. (1981). Ethanol fuel – a single-cylinder engine study of efficiency and exhaust emissions. *SAE Transactions* **90**, 1410–1424.

Brown, E.R., Ibrahim, J.G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59**, 221–228.

Burridge, J. (1981). Empirical Bayes analysis of survival time data. *J. Roy. Statist. Soc., Ser. B* **43**, 65–75.

Bush, C.A., MacEachern, S.N. (1996). A semi-parametric Bayesian model for randomized block designs. *Biometrika* **83**, 275–286.

Choudhuri, N., Ghosal, S., Roy, A. (2004). Bayesian methods for function estimation. In: Dey, D.K., Rao, C.R. (Eds.), *Bayesian Thinking: Modeling and Computation, Handbook of Statistics*, vol. 25. Elsevier, Amsterdam. This volume.

Christensen, R., Johnson, W.O. (1988). Modeling accelerated failure time with a Dirichlet process. *Biometrika* **75**, 693–704.

Clayton, D.G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* **47**, 467–485.

Clyde, M., George, E.I. (2004). Model uncertainty. *Statist. Sci.* **19**, 81–94.

Congdon, P. (2001). *Bayesian Statistical Modeling*. Wiley, Chichester.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc., Ser. B Methodological* **34**, 187–220.

Crainiceanu, C.M., Ruppert, D., Wand, M.P. (2004). Bayesian analysis for penalized spline regression using WinBUGS. Johns Hopkins University, Dept. of Biostatistics Working Papers, Working Paper 40. http://www.bepress.com/jhubiostat/paper40.

Diaconis, P., Freedman, D. (1986). On inconsistent Bayes estimates of location. *Ann. Statist.* **14**, 68–87.

De Iorio, M., Müller, P., Rosner, G.L., MacEachern, S.N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.* **99**, 205–215.

Denison, D.G.T., Holmes, C.C., Mallick, B.K., Smith, A.F.M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley, Chichester.

Dey, D., Müller, P., Sinha, D. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*. Springer Lecture Notes, New York.

Do, K.-A., Müller, P., Tang, F. (2004). A Bayesian mixture model for differential gene expression. Preprint.

Doksum, K.A. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2**, 183–201.

Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Ann. Statist.* **22**, 1763–1786.

Doss, H., Huffer, F.W. (2004). Monte Carlo methods for Bayesian analysis of survival data using mixtures of Dirichlet process priors. *J. Comput. Graph. Statist.* **12**, 282–307.

Dykstra, R.L., Laud, P.W. (1981). A Bayesian nonparametric approach to reliability. *Ann. Statist.* **9**, 356–367.

Efromovich, S. (1999). *Nonparametric Curve Estimation*. Springer-Verlag, New York.

Escobar, M.D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89**, 268–277.

Escobar, M.D., West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**, 577–588.

Fabius, J. (1964). Asymptotic behavior of Bayes' estimates. *Ann. Math. Statist.* **35**, 846–856.

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

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615–629.

Freedman, D.A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34**, 1194–1216.

Gehan, E.A. (1965). A generalized Wilcoxin test for comparing arbitrarily single-censored samples. *Biometrika* **52**, 203–224.

Gelfand, A.E. (1999). Approaches for semiparametric Bayesian regression. In: Ghosh, S. (Ed.), *Asymptotics, Nonparametrics and Time Series*. Marcel Dekker, New York, pp. 615–638.

Gelfand, A.E., Kottas, A.A. (2001). Nonparametric Bayesian modeling for stochastic order. *Ann. Inst. Statist. Math.* **53**, 865–876.

Gelfand, A.E., Kottas, A.A. (2002). Computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *J. Comput. Graph. Statist.* **11**, 289–305.

Gelfand, A.E., Mukhopadhyay, S. (1995). On nonparametric Bayesian inference for the distribution of a random sample. *Canad. J. Statist.* **23**, 411–420.

Gelfand, A.E., Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398–409.

Gelfand, A.E., Kottas, A.A., MacEachern, S.N. (2004). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.*, submitted for publication.

Ghosh, J.K., Ramamoorthi, R. (2003). *Bayesian Nonparametrics*. Springer-Verlag, New York.

Ghosal, S., Ghosh, J.K., Ramamoorthi, R. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **17**, 143–158.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Haar, A. (1910). Zur theorie der orthogonalen funktionensysteme. *Math. Ann.* **69**, 331–371.

Hanson, T.E., Johnson, W.O. (2002). Modeling regression error with a mixture of Polya trees. *J. Amer. Statist. Assoc.* **97**, 1020–1033.

Hanson, T.E., Johnson, W.O. (2004). A Bayesian semiparametric AFT model for interval censored data. *J. Comput. Graph. Statist.* **13**, 341–361.

Hjort, N.L. (1990). Nonparametric Bayes estimators based on beta processes in models of life history data. *Ann. Statist.* **18**, 1259–1294.

Hoff, P.D. (2003). Bayesian methods for partial stochastic orderings. *Biometrika* **90**, 303–317.

Holmes, C.C., Mallick, B.K. (2001). Bayesian regression with multivariate linear splines. *J. Roy. Statist. Soc., Ser. B* **63**, 3–17.

Huzurbazar, A.V. (2004). *Flowgraph Models for Multistate Time-to-Event Data*. Wiley, New York.

Ibrahim, J.G., Kleinman, K.P. (1998). Semiparametric Bayesian methods for random effects models. In: *Practical Nonparametric and Semiparametric Bayesian Statistics*. In: *Lecture Notes in Statistics*, vol. 133. Springer-Verlag, New York.

Ibrahim, J.G., Chen, M.-H., MacEachern, S.N. (1999). Bayesian variable selection for proportional hazards models. *Canad. J. Statist.* **37**, 701–717.

Ibrahim, J.G., Chen, M.-H., Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag, New York.

Ishwaran, H., James, L.F. (2004). Computational methods for multiplicative intensity models using weighted gamma processes: Proportional hazards, marked point processes and panel count data. *J. Amer. Statist. Assoc.* **99**, 175–190.

Johnson, W.O., Christensen, R. (1986). Bayesian nonparametric survival analysis for grouped data. *Canad. J. Statist.* **14**, 307–314.

Johnson, W.O., Christensen, R. (1989). Nonparametric Bayesian analysis of the accelerated failure time model. *Statist. Probab. Lett.* **8**, 179–184.

Kalbfleisch, J.D. (1978). Non-parametric Bayesian analysis of survival time data. *J. Roy. Statist. Soc., Ser. B* **40**, 214–221.

Kaplan, E.L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481.

Kim, Y., Lee, J. (2003). Bayesian analysis of proportional hazard models. *Ann. Statist.* **31**, 493–511.

Kleinman, K., Ibrahim, J. (1998). A semi-parametric Bayesian approach to generalized linear mixed models. *Statist. in Med.* **17**, 2579–2596.

Kottas, A.A., Gelfand, A.E. (2001a). Bayesian semiparametric median regression modeling. *J. Amer. Statist. Assoc.* **95**, 1458–1468.

Kottas, A.A., Gelfand, A.E. (2001b). Modeling variability order: A semiparametric Bayesian approach. *Methodology and Computing in Applied Probability* **3**, 427–442.

Kuo, L., Mallick, B. (1997). Bayesian semiparametric inference for the accelerated failure-time model. *Canad. J. Statist.* **25**, 457–472.

Laud, P., Damien, P., Smith, A.F.M. (1998). Bayesian nonparametric and covariate analysis of failure time data. In: Dey, D., Müller, P., Sinha, D. (Eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*. Springer, New York, pp. 213–226.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *Ann. Statist.* **20**, 1222–1235.

Lavine, M. (1994). More aspects of Polya tree distributions for statistical modeling. *Ann. Statist.* **22**, 1161–1176.

Lindley, D.V., Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc., Ser. B* **34**, 1–41.

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

Lo, A.Y., Weng, C.S. (1989). On a class of Bayesian nonparametric estimates: II. Hazard rate estimates. *Ann. Inst. Statist. Math.* **41**, 227–245.

MacEachern, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Comm. Statist. Simulation Comput.* **23**, 727–741.

MacEachern, S.N. (2000). Dependent Dirichlet processes. Technical Report, Dept. of Statistics, The Ohio State University.

MacEachern, S.N., Clyde, M., Liu, J. (1999). Sequential importance sampling for nonparametric Bayes models: The next generation. *Canad. J. Statist.* **27**, 251–267.

MacEachern, S.N., Müller, P. (1998). Estimating mixture of Dirichlet process models. *J. Comput. Graph. Statist.* **7**, 223–238.

Mallick, B.K., Denison, D.G.T., Smith, A.F.M. (1999). Bayesian survival analysis using a MARS model. *Biometrics* **55**, 1071–1077.

Mallick, B.K., Walker, S.G. (2003). A Bayesian semiparametric transformation model incorporating frailties. *J. Statist. Plann. Inference* **112**, 159–174.

MathSoft, Inc. (1999). *S-Plus 5 for UNIX Guide to Statistics*. Data Analysis Products Division, MathSoft, Seattle.

Mauldin, R.D., Sudderth, W.D., Williams, S.C. (1992). Polya trees and random distributions. *Ann. Statist.* **20**, 1203–1221.

Moore, D.S. (1995). *The Basic Practice of Statistics*, first ed. W.H. Freeman and Company.

Mukhopadhyay, S., Gelfand, A.E. (1997). Dirichlet process mixed generalized linear models. *J. Amer. Statist. Assoc.* **92**, 633–639.

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

Müller, P., Quintana, F.A. (2004). Nonparametric Bayesian data analysis. *Statist. Sci.* **19**, 95–110.

Müller, P., Vidakovic, B. (1999). *Bayesian Inference in Wavelet-Based Models*. Springer-Verlag, New York.

Neal, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9**, 249–265.

Nieto-Barajas, L., Walker, S. (2002). Markov beta and gamma processes for modeling hazard rates. *Scand. J. Statist.* **29**, 413–424.

Nieto-Barajas, L., Walker, S. (2004). Bayesian nonparametric survival analysis via Lévy driven Markov processes. *Statistica Sinica* (submitted for publication).

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

Regazzini, E., Guglielmi, A., Di Nunno, G. (2002). Theory and numerical analysis for exact distributions of functionals of a Dirichlet process. *Ann. Statist.* **30**, 1376–1411.

Richardson, S., Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc., Ser. B* **59**, 731–792.

Sethuraman, J. (1994). A constructive definition of the Dirichlet prior. *Statistica Sinica* **4**, 639–650.

Sinha, D., Dey, D.K. (1997). Semiparametric Bayesian analysis of survival data. *J. Amer. Statist. Assoc.* **92**, 1195–1212.

Smith, M., Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75**, 317–343.

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D. (2003). WinBUGS 1.4 User Manual. MRC Biostatistics Unit, Cambridge. www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml.

Susarla, J., van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.* **71**, 897–902.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 1701–1762.

Titterington, D.M., Smith, A.F.M., Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

Turnbull, B.W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Soc.* **69**, 169–173.

Vidakovic, B. (1998). Wavelet-based nonparametric Bayes methods. In: Dey, D., Müller, P., Sinha, D. (Eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*. In: *Lecture Notes in Statistics*, vol. 133. Springer-Verlag, New York, pp. 133–155.

Walker, S., Damien, P. (1998). Sampling methods for Bayesian nonparametric inference involving stochastic processes. In: D. Dey, P. Mueller, D. Sinha, (Eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*, Springer Lecture Notes, pp. 243–254.

Walker, S.G., Mallick, B.K. (1997). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *J. Roy. Statist. Soc., Ser. B* **59**, 845–860.

Walker, S.G., Mallick, B.K. (1999). Semiparametric accelerated life time model. *Biometrics* **55**, 477–483.

Walker, S.G., Damien, P., Laud, P., Smith, A.F.M. (1999). Bayesian nonparametric inference for random distributions and related functions (with discussion). *J. Roy. Statist. Soc., Ser. B* **61**, 485–527.

Wang, Y., Taylor, J.M.G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *J. Amer. Statist. Assoc.* **96**, 895–905.

West, M., Müller, P., Escobar, M.D. (1994). Hierarchical priors and mixture models with applications in regression and density estimation. In: Smith, A.F.M., Freeman, P.R. (Eds.), *Aspects of Uncertainty: A Tribute D. Lindley*. Wiley, London, pp. 363–386.

Wiper, M.P., Ríos Insua, D., Ruggeri, F. (2001). Mixtures of gamma distributions with applications. *J. Comput. Graph. Statist.* **10**, 440–454.